Uncovering Latent Biological Function Associations through Gene Set Embeddings

Yuhang Huang[1], Fan Zhong*

[1]Institutes of Biomedical Sciences,Shanghai Medical College, Fudan University, 131 Dongan Road, Shanghai, Shanghai 200032, China

*Corresponding authors. Intelligent Medicine Institute, Shanghai Medical College, Fudan University, Shanghai 200032, China. E-mails: zonefan@163.com (F.Z.)

Abstract

The complexity of biological systems has increasingly been unraveled through computational methods, with biological network analysis now focusing on the construction and exploration of well-defined interaction networks. Traditional graph-theoretical approaches have been instrumental in mapping key biological processes using high-confidence interaction data. However, these methods often struggle with incomplete or/and heterogeneous datasets. In this study, we extend beyond conventional bipartite models by integrating attribute-driven knowledge from the Molecular Signatures Database (MSigDB) using the node2vec algorithm. Our approach explores unsupervised biological relationships and uncovers potential associations between genes and biological terms through network connectivity analysis. By embedding both human and mouse data into a shared vector space, we validate our findings cross-species, further strengthening the robustness of our method. This integrative framework reveals both expected and novel biological insights, offering a comprehensive perspective that complements traditional biological network analysis and paves the way for deeper understanding of complex biological processes and diseases.

Keywords: Biological network analysis, MSigDB, Gene-term associations, Cross-species analysis, Network embedding

Introduction

The intricate architecture of biological systems has been progressively demystified through advanced computational methodologies. Currently, biological network analysis predominantly focuses on constructing and interrogating well-defined interaction networks. These approaches often employ graph-theoretical methods to identify key nodes and critical pathways, utilizing high-confidence interaction data to map out precise biological processes. Databases such as KEGG (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017) and STRING (Szklarczyk et al., 2019) are commonly used for mapping and analyzing these interactions.

Initially, researchers employed network-based methods primarily to analyze biological data(Dybowski & Gant, 1995). Subsequent advancements have profoundly deepened our understanding of biological networks through unsupervised analysis. Jeong et al. (Jeong, Tombor, Albert, Oltvai, & Barabási, 2000) unveiled complex structures within metabolic networks, Barabási and Oltvai (Barabási & Oltvai, 2004) elucidated the functional organization of cellular networks, and Stuart et al. (Stuart, Segal, Koller, & Kim, 2003) were pioneers in identifying conserved gene modules across species. Cumulatively, these seminal studies underscore the evolution and significant impact of network analysis in bioinformatics. Within the past decades, researchers have been applying machine learning on graphs and utilizing various methods to represent (W. Hamilton, Z. Ying, & J. Leskovec, 2017): from shallow embedding approaches like DeepWalk (Perozzi, Al-Rfou, & Skiena, 2014), node2vec (Grover & Leskovec, 2016), and HARP (Chen, Perozzi, Hu, & Skiena, 2018), to methods based on deep learning like neighborhood autoencoder methods DNGR (Cao, Lu, & Xu, 2016), SDNE (Wang, Cui, & Zhu, 2016), and convolutional like GraphSAGE (W. L. Hamilton, R. Ying, & J. Leskovec, 2017).

Among these methods, node2vec stands out due to its simplicity and strong interpretability in an unsupervised learning context. In the field of computational biology, graph representation learning has gained popularity in the analysis of biological relationships. Numerous studies have reviewed and summarized the applications of graph representation learning in the biomedical field (Nelson et al., 2019; Yi, You, Huang, & Kwoh, 2021), demonstrating its versatility. These approaches typically model biological systems through bipartite (Zhang et al., 2020), multirelational (Y. Liu, Zeng, He, &

Zou, 2017), or knowledge-based graphs (Rotmensch, Halpern, Tlimat, Horng, & Sontag, 2017), which capture intricate relationships between biological entities, molecules, and biomedical concepts, providing powerful tools for tasks such as drug discovery (Zong et al., 2021), disease treatment, and molecular interaction analysis (Fachet et al., 2020).

Building on advancements in graph-based representation learning, this study extends traditional bipartite models by incorporating attribute-driven knowledge database, Molecular Signatures Database ((Liberzon et al., 2015)). Traditional biological network analyses, while effective with specific, well-curated datasets, often falter when faced with incomplete or ambiguous data from diverse sources (R. Liu, Hirn, & Krishnan, 2023). To bridge this gap, we introduce a novel network-based framework designed to explore complex relationships within biological networks. Rather than relying solely on annotated data, our approach uncovers potential gene associations with biological terms (e.g., pathways, functions) through network connectivity. By analyzing network structure and topology, we aim to reveal hidden patterns and novel associations, offering deeper insights, particularly in exploratory studies where direct evidence is limited or mechanisms are unclear. This integrative approach enhances traditional data mining by offering a more comprehensive perspective.

Results

Pipeline

Our experimental findings, which reveal both anticipated and unforeseen potential biological connections, open up broader discussions on how non-obvious relationships contribute to our understanding of biological processes and diseases (Figure 1). Crucially, beyond human datasets, we also utilized mice collections as an external validation set. We achieved this by embedding both human and mouse data into a shared vector space using node2vec. By embedding homologous genes from mice into the same space, we were able to validate our findings cross-species, uncovering similar patterns and biological insights from both human and mice collections. This cross-species validation strengthens our results and demonstrates the robustness of our approach in identifying meaningful biological associations.

Overview

Our analysis revealed that both the *Jaccard index* and node2vec embeddings exhibited extreme value distributions across 10^8 pairwise comparisons. These distributions (Figure 2A) could not be modeled by classical statistical methods. To assess significant biological associations, we normalized both metrics and selected representative pairs at distance thresholds of 0.1, 0.3, 0.5, 0.7, and 1 for further manual evaluation and literature-based validation. For each threshold, 10 pairs were analyzed in-depth to determine their biological relevance. To enhance the rigor of biological relevance assessment, we introduced a reliability score ranging from 0 to 5, reflecting levels of confidence from speculative associations to those strongly supported by extensive literature (Supplemental Table 1).

A higher *Jaccard index* generally signifies a substantial overlap between gene sets, indicating a robust biological connection. However, lower *Jaccard index* should not be overlooked, particularly when the gene sets involved contain fewer genes, where even modest overlap might indicate meaningful associations. To refine the analysis, we applied a hypergeometric distribution test with a p value threshold of 0.05 to assess the significance of shared gene counts. All selected pairs had a hypergeometric p value < 0.05. From the analysis, it is evident that higher Jaccard values are generally associated with higher reliability scores (Figure 2B), reflecting a trend that aligns with our expectations.

The distribution of credible associations, especially those with the highest biological significance (top 5% as determined by the cumulative distribution function (CDF)), was most prominent in the 64-dimensional space (Figure 2C). Although the same distance thresholding approach was applied consistently, the 64-dimensional node2vec embeddings outperformed their 32-dimensional counterparts in identifying biologically meaningful results, both in the number of validated pairs and their distribution.

We employed node2vec embeddings in both 32- and 64-dimensional spaces. Although the correlation between

node2vec distances and Jaccard indices was relatively weak, pairs with smaller spatial distances, particularly in the 64-dimensional space, were more likely to exhibit biological credibility, as confirmed by manual validation and literature review (Figure 2D). Overall, the 64-dimensional embeddings consistently outperformed the 32-dimensional ones in identifying biologically relevant associations, particularly for pairs with distances below 0.1.

Interpreting Biological Relevance of Jaccard Index

The *Jaccard index* is an effective measure of biological relevance, with any value greater than zero indicating a potential relationship. In all cases, the hypergeometric distribution suggests that gene overlaps are not random. We observe that as the *Jaccard index* increases, the number of supporting references rises accordingly. This is logical: larger overlaps between gene sets often lead to stronger evidence, particularly when the index nears 1, where terms often represent the same biological pathways sourced from different databases (Figure 2B).

For example, the association between GOBP_UBIQUITIN_DEPENDENT_GLYCOPROTEIN_ERAD_PATHWAY and GOMF_MANNOSYL_OLIGOSACCHARIDE_1_2_ALPHA_MANNOSIDASE_ACTIVITY (with a *Jaccard index* of 1) is biologically intuitive. Although this connection is not explicitly indicated in MSigDB, these terms describe closely related processes. The ubiquitin-dependent glycoprotein ERAD pathway is responsible for recognizing and degrading misfolded glycoproteins in the endoplasmic reticulum (ER) via the ER-associated degradation (ERAD) system (Lopata, Kniss, Löhr, Rogov, & Dötsch, 2020). During this process, mannosyl oligosaccharide 1,2-alpha mannosidase enzymes trim misfolded glycoproteins, preparing them for ubiquitination and degradation. The inclusion of this enzyme activity within the pathway underscores the functional cohesion, highlighting the expected biological relationship between these terms.

As the Jaccard values decrease, a turning point emerges around 0.5, where reliability scores typically range between 3 and 4 (Figure 2B). Although direct literature support becomes sparse, these associations remain biologically plausible. A notable example is the interaction between FU_INTERACT_WITH_ALKBH8 and GOCC_CHAPERONIN_CONTAINING_T_COMPLEX (CCT). While no explicit link has been documented between the CCT complex and ALKBH8, both share a key gene, *TCP1*, suggesting a potential functional connection. CCT, a type II molecular chaperone in eukaryotes, plays a pivotal role in folding nascent proteins, particularly during T-cell activation and in response to external stressors. Its critical function in maintaining protein homeostasis means that CCT deficiency can lead to immune dysfunction and impaired stress response. Conversely, ALKBH8, an enzyme responsible for tRNA modification, ensures translation accuracy and regulates oxidative stress by modifying the wobble position (e.g., mcm5U). Loss of ALKBH8 can result in neurological disorders and dysregulated oxidative stress responses. In a stress environment, accurate protein folding and translation are crucial. Although CCT and ALKBH8 operate at different molecular levels—CCT ensuring proper protein folding and ALKBH8 ensuring translation fidelity—they converge on the broader roles of stress management and protein quality control. Deficiencies in tRNA modification, such as those caused by ALKBH8 dysfunction, could increase the burden of misfolded proteins, thereby heightening the reliance on the CCT complex for correction. Thus, despite the moderate *Jaccard index*, the inferred connection suggests a meaningful biological interplay between these components.

At very low *Jaccard index* values, most of the observed terms relate to chromosomal locations. While these are technically valid associations, they provide limited phenotypic insight. Consequently, these low-index pairs may be regarded as noise in the context of this analysis. However, they can still serve a role in refining models like node2vec, aligning them more closely with the phenotypic relationships we aim to uncover.

Identifying Novel Biological Associations through Node2vec-Based Distances

Compared to traditional methods like the *Jaccard index*, which resemble differential gene analysis, node2vec results tend to have lower confidence scores based on literature surveys. However, they hold greater research value, as node2vec-derived spatial proximity often correlates with phenotypic synergy. Notably, the 64-dimensional embeddings outperform the 32-dimensional ones (Figure 2C). This is evident in the reduced number of chromosome location-related terms when distances approach zero. Additionally, comparisons between the *Jaccard index* and node2vec spatial distances reveal that

the 64-dimensional embeddings (d64) exhibit more extreme values (less than 5%) in the CDF, indicating higher gene-level reliability. Furthermore, d64 scores consistently surpass d32 in terms of reliability (Figure 2D), underscoring its superior performance in capturing meaningful biological relationships. we specifically focused on capturing those from the top 5% of the CDF in the 64-dimensional embeddings (d64). To further elucidate the relationships among biological terms, we developed a minimum spanning tree (MST, Supplemental Material 2) (Eisner, 1997) and an interactive web interface

In this analysis, we focus on cases where the *Jaccard index* equals 1, indicating identical gene sets between pathways. Both KEGG_MEDICUS_REFERENCE_TYROSINE_DEGRADATION and WP_TYROSINE_METABOLISM_AND_RELATED_DISORDERS are pathways associated with tyrosine metabolism and degradation. The former emphasizes the overall metabolic process, while the latter focuses on the pathological mechanisms resulting from metabolic disorders. Although these pathways share an identical gene set, as indicated by a *Jaccard index* of 1, their spatial distance in the node2vec analysis falls within the extreme values of the CDF. This suggests a tight biological network connection between these pathways and hints at potential clinical phenotypic commonalities. In contrast, despite also sharing a *Jaccard index* of 1, KEGG_MEDICUS_REFERENCE_TRANSPORT_OF_CALCIUM and GOMF_GLUTAMATE_GATED_CALCIUM_ION_CHANNEL_ACTIVITY demonstrate a much larger spatial distance in the node2vec analysis and hold a reliability score 4 points lower than the others.. This highlights node2vec's ability to distinguish between pathways based on their functional roles and pathological relevance, rather than being solely driven by gene set overlap. The KEGG pathway describes the broad physiological process of calcium ion transport across membranes, involving various cell types and tissues. In comparison, the GO molecular function pathway is more specific, focusing on the activity of glutamate-gated calcium ion channels within the nervous system, a more narrowly defined molecular mechanism. Consequently, node2vec separates these pathways in the biological network, reflecting differences in their functional and contextual roles despite their identical gene sets. This demonstrates node2vec's capacity to capture topological distinctions in the network, offering a more nuanced understanding of functional relationships beyond simple gene overlap.

Interestingly, our analysis demonstrates that a low *Jaccard index* does not necessarily imply a lack of functional or biological relevance. For instance, we observed a *Jaccard index* of merely 0.083 between the GNF2_PRDX2 gene set from the GNF2 expression atlas and the HP_ANISOCYTOSIS gene set. Despite this minimal overlap, the node2vec spatial distance between these pathways falls within the extreme values of the CDF, suggesting a potential biological connection. This link may be driven by oxidative stress mechanisms, as PRDX2, an antioxidant enzyme, plays a pivotal role in regulating cellular redox balance. Oxidative stress is a key factor in various hematological disorders, and its modulation could influence red blood cell (RBC) morphology, specifically in conditions like anisocytosis, which is characterized by variations in RBC size. Altered PRDX2 expression has been associated with hematological conditions such as anemia, which frequently presents with anisocytosis. We hypothesize that PRDX2 may regulate RBC morphology by mitigating oxidative damage, thereby establishing a mechanistic link to anisocytosis. This potential connection between PRDX2 and RBC size variability could pave the way for further research into its role in hematological disorders, particularly those involving oxidative stress (Sadowska-Bartosz & Bartosz, 2023; Stresing et al., 2013).

Cross-Species Discovery Through Node2Vec Embedding Reveals Shared Biological Insights

Due to ethical constraints prohibiting functional perturbation experiments within human subjects, alongside the limitations of cell line experiments that diverge from the true in vivo environment, a substantial amount of research is conducted within murine models. The central aim of these animal experiments is to bridge murine data with human disease and physiological mechanisms, rendering cross-species association studies indispensable.

We extended our analysis to include cross-species validation, utilizing node2vec unsupervised embeddings at 64 dimensions. By embedding both human and mouse gene datasets into a shared vector space, we aimed to identify biologically relevant patterns that are conserved across species. Specifically, we converted mouse genes to their homologous human counterparts and embedded the mice collections based on the pre-trained human model. Following this, we searched for the nearest human term-related nodes corresponding to the embedded mouse terms. This approach allowed us to discover and

strengthen our findings across species (Supplementary Material 1). We selected mouse collections that include curated cluster markers for cell types identified in single-cell sequencing studies of mouse tissues, including markers related to immune cells in the lungs (Almanzar et al., 2020).

The observed connection between mice_TABULA_MURIS_SENIS_LUNG_B_CELL_AGEING and human_CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP suggests potential links between ageing lung B cells and distinct cellular states in breast cancer. The mesenchymal transition in cancer is frequently associated with immune responses and cellular ageing, which might explain this correlation. Ageing B cells, particularly in tissues like the lung, may undergo significant functional changes, such as a reduction in antigen-specific responses or a shift towards pro-inflammatory profiles. These changes are characteristic of immunosenescence, which is known to both impair immune surveillance and enhance a pro-tumorigenic environment. Chronic inflammation, often seen with ageing immune cells, creates a microenvironment conducive to cancer progression through the secretion of inflammatory cytokines and growth factors that support tumor cell proliferation, survival, and metastasis. This is especially relevant to the mesenchymal transition in cancer, as inflammatory signals are critical drivers of epithelial-mesenchymal transition (EMT), facilitating cancer cell invasion and dissemination.

Discussion

In this study, we utilized node2vec to explore complex biological relationships within diverse datasets. The key strength of this approach lies in its ability to embed terms from various biological datasets into a shared vector space, enabling researchers to discover relationships between similar terms across different datasets. This provides a flexible tool for uncovering hidden patterns, particularly in contexts where data is heterogeneous and often incomplete. The broad applicability of this method makes it a valuable resource for exploratory research into environmental stimuli, developmental processes, immune responses, and all relevant molecular mechanisms.

However, one of the significant challenges of this approach is the reliance on pre-existing datasets, which introduces a level of uncertainty. Our term embeddings are drawn from various sources, each with varying levels of curation and confidence, and we did not assign explicit weights to these terms. This results in a broad, unsupervised analysis where both input and output lack precise control. While this introduces potential noise and ambiguity, it also allows for serendipitous discoveries, as seen in some of the unexpected relationships we uncovered in our results. These findings suggest that even in a loosely structured exploration, novel biological insights can emerge, particularly when the direct evidence is limited or the underlying mechanisms are not well understood.

The node2vec algorithm itself alleviates some of the concerns mentioned above. Its simplicity and ease of parameter tuning make it well-suited for exploratory analyses. Unlike more complex graph neural networks, node2vec allows for a higher degree of control and interpretability, making it possible to identify broad trends and patterns in large datasets. Moreover, its unsupervised nature is particularly advantageous when working with heterogeneous and incomplete data, which is common in biological research. This flexibility allows us to capture potential relationships that more rigid, supervised methods might overlook.

Despite these strengths, it is important to acknowledge that our study primarily serves as a proof-of-concept. The current framework demonstrates the potential of node2vec in biological research, but more concrete case studies with well-curated datasets are necessary to fully exploit its capabilities. Additionally, while node2vec offers simplicity and control, future work should explore the use of more sophisticated models, such as graph neural networks, which may offer greater precision and sensitivity when applied to specific, focused research questions.

Code Availability

The code utilized in this study is available on GitHub at git@github.com:YuHang-aw /MsigDB_mining.git. It includes all

necessary scripts for data processing, model training, and evaluation. The repository is licensed under the MIT license, facilitating reproducibility and supporting further research

Reference

Almanzar, N., Antony, J., Baghel, A. S., Bakerman, I., Bansal, I., Barres, B. A., . . . The Tabula Muris, C. (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature, 583*(7817), 590-595. doi:10.1038/s41586-020-2496-1

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics, 5*(2), 101-113. doi:10.1038/nrg1272

Cao, S., Lu, W., & Xu, Q. (2016). Deep Neural Networks for Learning Graph Representations. *Proceedings of the AAAI Conference on Artificial Intelligence, 30*(1). doi:10.1609/aaai.v30i1.10179

Chen, H., Perozzi, B., Hu, Y., & Skiena, S. (2018). *Harp: Hierarchical representation learning for networks.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*(2), 224-227.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*(8), 1184-1191. doi:10.1038/nprot.2009.97

Dybowski, R., & Gant, V. (1995). Artificial neural networks in pathology and medical laboratories. *Lancet, 346*(8984), 1203-1207. doi:10.1016/s0140-6736(95)92904-5

Eisner, J. (1997). *State-of-the-Art Algorithms for Minimum Spanning Trees - A Tutorial Discussion*.

Fachet, M., Witte, C., Flassig, R. J., Rihko-Struckmann, L. K., McKie-Krisberg, Z., Polle, J. E. W., & Sundmacher, K. (2020). Reconstruction and analysis of a carbon-core metabolic network for Dunaliella salina. *BMC Bioinformatics, 21*(1), 1. doi:10.1186/s12859-019-3325-0

Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*.

Grover, A., & Leskovec, J. (2016). *node2vec: Scalable feature learning for networks.* Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems, 30*.

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature, 407*(6804), 651-654. doi:10.1038/35036627

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res, 45*(D1), D353-d361. doi:10.1093/nar/gkw1092

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst, 1*(6), 417-425. doi:10.1016/j.cels.2015.12.004

Liu, R., Hirn, M., & Krishnan, A. (2023). Accurately modeling biased random walks on weighted networks using node2vec+. *Bioinformatics, 39*(1). doi:10.1093/bioinformatics/btad047

Liu, Y., Zeng, X., He, Z., & Zou, Q. (2017). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans Comput Biol Bioinform, 14*(4), 905-915. doi:10.1109/tcbb.2016.2550432

Lopata, A., Kniss, A., Löhr, F., Rogov, V. V., & Dötsch, V. (2020). Ubiquitination in the ERAD Process. *Int J Mol Sci, 21*(15). doi:10.3390/ijms21155369

Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., & Sharan, R. (2019). To Embed or Not: Network Embedding as a Paradigm in Computational Biology. *Frontiers in Genetics, 10*. doi:10.3389/fgene.2019.00381

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). *Deepwalk: Online learning of social representations.* Paper presented at the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports, 7*(1), 5994.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53-65.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science, 302*(5643), 249-255. doi:10.1126/science.1087447

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., . . . Mering, C. V. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res, 47*(D1), D607-d613. doi:10.1093/nar/gky1131

Wang, D., Cui, P., & Zhu, W. (2016). Structural Deep Network Embedding. *Kdd '16*, 1225–1234. doi:10.1145/2939672.2939753

Yi, H.-C., You, Z.-H., Huang, D.-S., & Kwoh, C. K. (2021). Graph representation learning in bioinformatics: trends, methods and applications. *Briefings in Bioinformatics, 23*(1). doi:10.1093/bib/bbab340

Zhang, Z.-C., Zhang, X.-F., Wu, M., Ou-Yang, L., Zhao, X.-M., & Li, X.-L. (2020). A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics, 36*(11), 3474-3481.

Zong, N., Wong, R. S. N., Yu, Y., Wen, A., Huang, M., & Li, N. (2021). Drug-target prediction utilizing heterogeneous bio-linked network embeddings. *Brief Bioinform, 22*(1), 568-580. doi:10.1093/bib/bbz147

Method

Data Integration

Our dataset incorporates a comprehensive set of terms that encapsulate a wide array of biological phenomena, we have opted to focus on a single species to mitigate these potential confounders and to ensure the integrity of our analysis. Within this scope, we define two essential sets: the set of **genes**, denoted as $G = \{g_1, g_2, \ldots, g_m\}$, and the set of **terms**, denoted as $T = \{t_1, t_2, \ldots, t_n\}$. Data for these sets were sourced from the MSigDB (https://www.gsea-msigdb.org/gsea/msigdb). To bridge species-specific differences in gene nomenclature, we utilized the biomaRt (Durinck, Spellman, Birney, & Huber, 2009) R package, which facilitated the conversion of murine genes to their human counterparts, ensuring cross-species comparability in our analysis. This conversion is critical for applying insights gained from model organisms to human biology.

Following data preprocessing, we construct an index dictionary to facilitate the creation of a PyG (Fey & Lenssen, 2019) (PyTorch Geometric) data object. This entails defining two mappings: one from genes to indices, termed gene_to_index: $G \rightarrow \{0,1,\ldots,m-1\}$, and another from terms to indices, termed term_to_index: $T \rightarrow \{m, m+1, \ldots, m+n-1\}$.

We then generate the list of edges E by adding an edge for each gene-term pair (g,t) to form the edge set $E = \{(\text{gene_to_index}(g), \text{term_to_index}(t)\}$. The PyG data object, represented as $Graph$, encompasses the node set $V = G \cup T$, and the edge set E: $Graph = (V, E)$, where the number of nodes $|V|$ is equal to $m + n$.

Distance Calculation Methodology

In this study, we utilized the node2vec algorithm to embed nodes within our graph-structured data, capturing their contextual relationships. The node2vec algorithm learns these embeddings by performing biased random walks on the graph, balancing between breadth-first search (BFS) and depth-first search (DFS) through two parameters: p (return parameter) and q (in-out parameter).

The random walks are controlled by the following transition probability formula:

$$P(c_i = x | c_{i-1} = v) = \begin{cases} \frac{1}{p} & \text{if } d(v,x) = 0 \\ 1 & \text{if } d(v,x) = 1 \\ \frac{1}{q} & \text{if } d(v,x) = 2 \end{cases}$$

where c_i is the candidate note at step i, $c_{i-1} = v$ is the previous node in the walk, $d(v,x)$ is the shortest path distance between nodes v and x,

The Jaccard index measures the overlap between two sets, and in our context, it represents the similarity between terms based on the overlap of their corresponding gene sets. Consider two terms, $T_A = \{gene_1, gene_2, \ldots, gene_m\}$ and $T_B = \{gene'_1, gene'_2, \ldots, gene'_m\}$, each associated with a set of genes based on their adjacency in the graph structure.

The *Jaccard Index* between these two terms is calculated using the formula:

$$J(term_A, term_B) = \frac{|T_A \cap T_B|}{|T_A \cup T_B|}$$

Node2Vec Training and Hyperparameter Optimization

To uncover the underlying patterns within our data, which involve complex many-to-many relationships without precise numerical values, we have adopted an unsupervised learning approach. Central to this approach is the node2vec algorithm, as implemented in the torch_geometric library, which we integrated into a machine learning pipeline. This pipeline enables the sequential execution of node embeddings and subsequent clustering.

Clustering is essential in unsupervised learning as it allows us to organize data points into groups based on their inherent similarities, without relying on pre-labeled classes. The hyperparameter tuning for the node2vec model involved varying several parameters. Embedding dimensions were set at either 32 or 64. Walk lengths were tested at values of 10, 20, and 30. The return parameter p and in-out parameter q were each explored across the values 0.25, 0.5, 1, 2, and 4. Context size was evaluated at 10 and 20, while the number of negative samples was tested at 1, 5, and 10. Finally, the number of walks per node was adjusted to 10, 20, or 30. This comprehensive hyperparameter search ensured an optimized configuration for the model.

Following the optimization of hyperparameters, we applied *k*-means clustering to evaluate the quality of the embeddings. The cluster labels assigned by *k*-means were used to calculate two internal clustering validation metrics, which assess the coherence and separation of the clusters.

The Silhouette score (Rousseeuw, 1987) ranges from -1 to 1, with higher values indicating better clustering performance. It is defined as:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{\mathbf{max}\,(a_i, b_i)}$$

Where a_i is the average distance between the i-th data point and other points in the same cluster, b_i is the minimum average distance from the i-th data point to points in a different cluster, n is the total number of data points.

A high Silhouette score suggests that clusters are compact and well-separated from each other.

The Davies-Bouldin score (Davies & Bouldin, 1979) measures the average similarity between clusters, with lower values indicating better clustering performance. It is calculated by:

$$DB = \frac{1}{k}\sum_{i}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Where σ_i is the average distance of all points in cluster i to the cluster centroid, $d(c_i, c_j)$ is the distance between centroids of clusters i and j, k is the number of clusters.

A lower Davies-Bouldin score implies that the clusters are well-separated and distinct.

In addition to these metrics, we developed a custom scoring function that integrates the linear independence of embedding dimensions and the clustering structure within the data. This function operates as follows:

Embedding Transformation and Covariance: Let X represent the dataset, and $E(X)$ be the embeddings obtained using the estimator's transformation function. The covariance matrix C of the embeddings is calculated as:

$$C = \mathrm{Cov}(E(X))$$

The rank of the covariance matrix $\mathrm{rank}(C)$ provides the number of linearly independent dimensions.
Clustering Analysis: Using the DBSCAN algorithm on the embeddings $E(X)$, we identify k significant clusters, excluding noise points. Let $Nclusters$ represent the number of clusters:

$$N_{clusters} = \mathrm{DBSCAN}(E(X)) - \text{noise points}$$

Weighted Score Calculation: The final weighted score S is calculated using a weighted sum of the matrix rank and the number of clusters, with weights $w_1 = 0.7$ for the rank and $w_2 = 0.3$ for the clusters:

$$S = w_1 \cdot \mathrm{rank}(C) + w_2 \cdot N_{clusters}$$

By combining these metrics and our custom score, we create a robust framework for assessing the embedding quality. This combination allows for a balanced evaluation that considers both individual point coherence within clusters and the overall structural insights provided by the clustering and dimensionality analysis. This holistic approach ensures a thorough understanding of how well the embeddings capture the inherent complexities of the data.

The model was trained on a server equipped with an NVIDIA GeForce RTX 3080 Ti GPU, using CUDA for parallel computations. The server ran Ubuntu 20.04 LTS and had 256 GB of RAM. We used Python 3.8 and the torch_geometric library for our implementation. The hyperparameter search was conducted over the course of one week, with each parameter combination undergoing 40 epochs of training. The duration of training for each parameter set varied depending on the specific parameters chosen, with some combinations taking more time to complete.

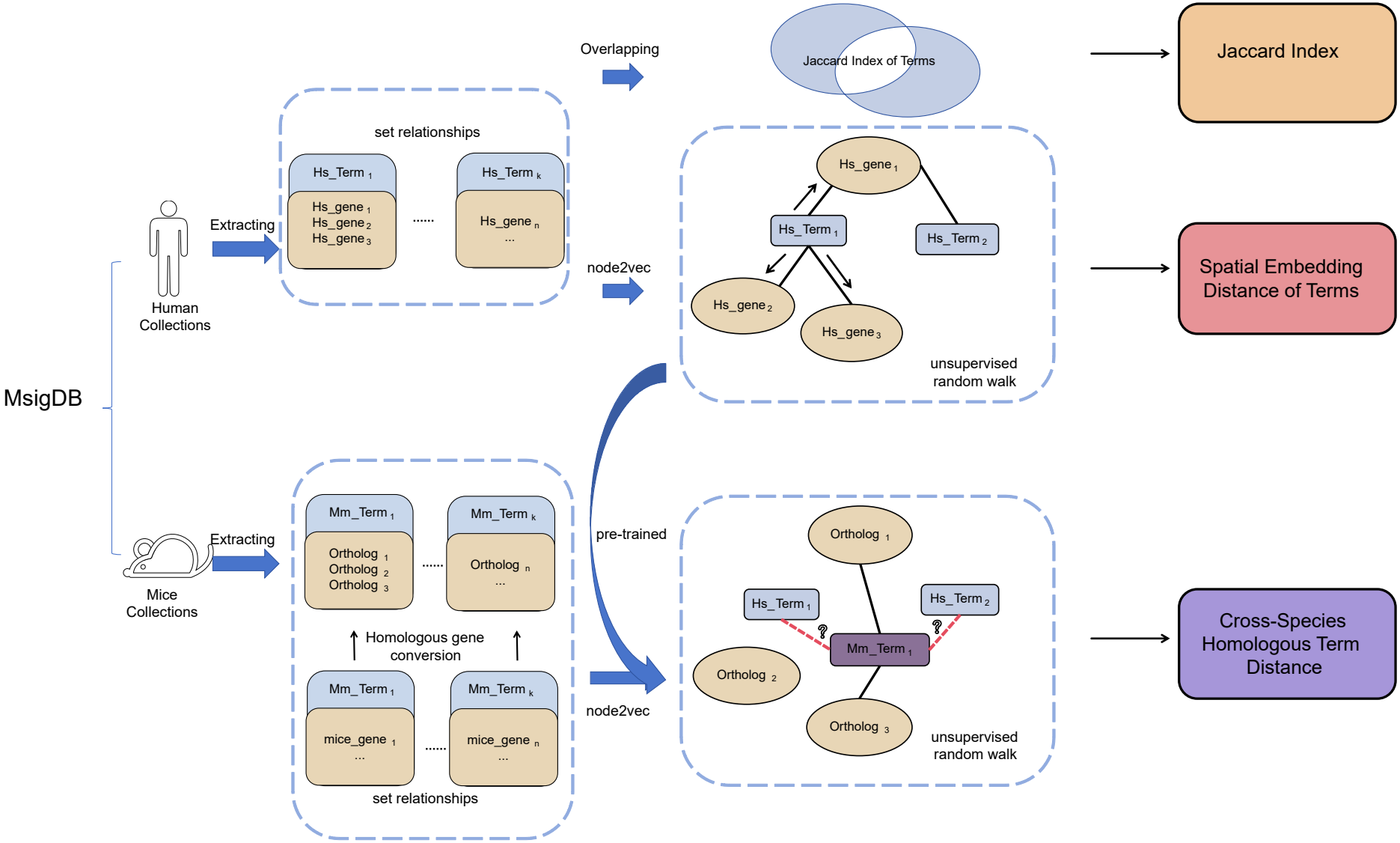

Figure 1. Cross-species validation pipeline: We utilized the Jaccard index and node2vec-based distance for analysis, embedding both human and mouse datasets into a shared vector space. Aligning homologous genes across species revealed consistent patterns and reinforced the robustness of our findings.

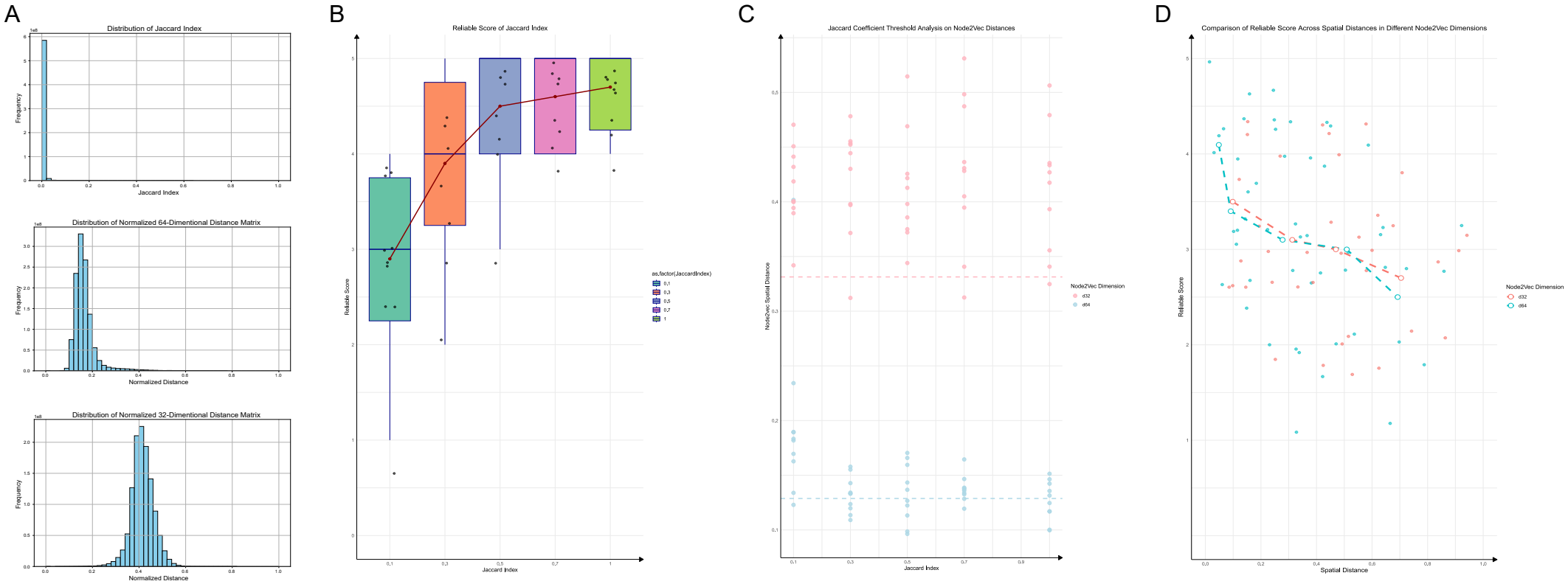

Figure 2. (A) Three vertically aligned plots illustrating the distributions of the Jaccard index, 64-dimensional node2vec distances, and 32-dimensional node2vec distances, respectively, across pairwise comparisons. (B) Box plot showing the relationship between the Jaccard index and reliability scores, where higher Jaccard values tend to correspond to higher reliability scores. (C) Correlation plot between the Jaccard index and both 64-dimensional and 32-dimensional node2vec distances, with dashed lines marking the top 5% of the cumulative distribution function (CDF) areas: d64 = 0.13 and d32 = 0.33. Despite a relatively weak overall correlation, the plot supports the observation that smaller spatial distances in the 64-dimensional space are more likely to indicate biologically credible associations. (D) Linear correlation between reliability scores and node2vec distances in both 64- and 32-dimensional spaces, highlighting the superior performance of the 64-dimensional embeddings, particularly for pairs with distances below 0.1.