



## Research and Applications

# Clustering rare diseases within an ontology-enriched knowledge graph

Jaleal Sanjak , PhD<sup>1,2</sup>, Jessica Binder, PhD<sup>1</sup>, Arjun Singh Yadaw , PhD<sup>1</sup>, Qian Zhu, PhD<sup>1</sup>, Ewy A. Mathé, PhD<sup>1,\*</sup>

<sup>1</sup>Division of Pre-Clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, MD, United States, <sup>2</sup>Chief Technology Office, Booz Allen Hamilton, Bethesda, MD, United States

\*Corresponding author: Ewy A. Mathé, PhD, Division of Pre-Clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), 9800 Medical Center Dr., Rockville, MD 20850 (ewy.mathe@nih.gov)

J. Sanjak and J. Binder contributed equally.

### Abstract

**Objective:** Identifying sets of rare diseases with shared aspects of etiology and pathophysiology may enable drug repurposing. Toward that aim, we utilized an integrative knowledge graph to construct clusters of rare diseases.

**Materials and Methods:** Data on 3242 rare diseases were extracted from the National Center for Advancing Translational Science Genetic and Rare Diseases Information center internal data resources. The rare disease data enriched with additional biomedical data, including gene and phenotype ontologies, biological pathway data, and small molecule-target activity data, to create a knowledge graph (KG). Node embeddings were trained and clustered. We validated the disease clusters through semantic similarity and feature enrichment analysis.

**Results:** Thirty-seven disease clusters were created with a mean size of 87 diseases. We validate the clusters quantitatively via semantic similarity based on the Orphanet Rare Disease Ontology. In addition, the clusters were analyzed for enrichment of associated genes, revealing that the enriched genes within clusters are highly related.

**Discussion:** We demonstrate that node embeddings are an effective method for clustering diseases within a heterogenous KG. Semantically similar diseases and relevant enriched genes have been uncovered within the clusters. Connections between disease clusters and drugs are enumerated for follow-up efforts.

**Conclusion:** We lay out a method for clustering rare diseases using graph node embeddings. We develop an easy-to-maintain pipeline that can be updated when new data on rare diseases emerges. The embeddings themselves can be paired with other representation learning methods for other data types, such as drugs, to address other predictive modeling problems.

**Key words:** rare disease; knowledge graph; ontology; drug repurposing.

### Background and significance

Rare diseases affect up to 25–30 million people in the United States<sup>1</sup> and more than 300 million worldwide,<sup>2</sup> making rare diseases common as a collective. The burden of rare diseases is disproportionately high because patients living with rare diseases tend to incur high healthcare costs along the course of long diagnostic odysseys and intensive treatment regimens.<sup>3,4</sup> Furthermore, the population of rare disease patients is distributed across 5000–10 000 distinct diseases,<sup>5</sup> yet the vast majority have no approved therapeutics.

Methods enabling research and development efforts to advance treatments for multiple diseases simultaneously may offer a path forward. Some such methods are already in practice, including therapeutic platforms like gene therapies,<sup>6</sup> basket clinical trials,<sup>7</sup> and drug repurposing.<sup>8</sup> Both basket clinical trials and drug repurposing require knowledge of the connections between diseases through their underlying causal factors.

Following similar efforts in the broader biomedical community,<sup>9</sup> data integration and harmonization efforts in the rare disease space have emerged to support research and

development aimed at multiple diseases at once. For example, the Encyclopedia of Rare Disease Annotations for Precision Medicine (eRAM)<sup>10</sup> was built using a text-mining approach from the biomedical literature, as well as integration of various data from open source databases (ie, Unified Medical Language System, Human Phenotype Ontology [HPO], Orphanet, Online Mendelian Inheritance in Man, and genome-wide association studies), to connect and annotate diseases, genes, and phenotypes. Another example is the RDMaP (a Rare Disease Map),<sup>11</sup> which was constructed based on only Orphanet<sup>12</sup> data (which uses HPO and Gene Ontology [GO] terms). Moreover, RDMaP measures the phenotypic and genetic distance between diseases and multidimensional scaling to convert the distance matrix into 2-dimensional (2D) points for visualization and the k-means clustering method to divide into several disease clusters. Both eRAM and RDMaP utilize methods for calculating the similarity of rare diseases using phenotype and pathogenetic gene annotations individually and then combining the similarity scores. More specifically, eRAM and RDMaP are helpful for

researchers and clinicians who want to explore similarities among diseases and seek clinical diagnosis assistance.

The National Center for Advancing Translational Sciences (NCATS) supports the Genetic and Rare Diseases (GARD) Information Center to maintain data on rare diseases within the United States. A preliminary attempt was made to harmonize data across the GARD diseases using multisource mappings across diseases and genes and phenotype annotations.<sup>13</sup> Here, we follow up on that study and use the similarity between diseases, with respect to their position within our KG, to perform disease clustering. Three factors differentiate our study from prior efforts: (1) the incorporation of explicit biological pathway and small molecule activity data, (2) the focus specifically on diseases tracked by GARD, and (3) the use of graph node embeddings.

DeepWalk<sup>14</sup> and Node2Vec<sup>15</sup> are notable graph node embedding algorithms that convert graph nodes into vectors by applying word2vec<sup>16</sup> embedding models to random walks taken from across the knowledge graph (or any graph-structured dataset). OPA2Vec<sup>17</sup> and, subsequently DL2Vec<sup>18</sup> were developed for the specific application of graph node embedding methods to the biomedical domain, with a particular emphasis on the use of semantic ontologies. A recent study utilizes the structure of GO to create gene and disease embeddings using only gene interaction data and gene-disease annotations.<sup>19</sup> In this study, we derive graph node embeddings for disease nodes within a KG containing diseases directly connected to associated genes and phenotypes and further enriched with small molecules (both drugs and metabolites), molecular pathway data, and biomedical ontologies. Our embeddings were generated using DL2Vec, modified to balance the probability of traversal from a disease node to either a gene-node or a phenotype-node. By using a variant of DL2Vec, we implicitly capture the semantic information contained within the ontology structures alongside the direct connections between diseases and genes/phenotypes.

The disease node embeddings are clustered, and the resulting clusters were analyzed for both validation and interpretation. We show that, indeed, graph node embeddings can be used to generate coherent, as measured through both quantitative and qualitative analyses, clusters of rare diseases within a heterogeneous knowledge graph. Further, several of our rare disease clusters show promising connections to drugs and investigational compounds.

## Materials and methods

### Data sources

The GARD internal data resources<sup>20</sup> were queried to obtain the overlap between GARD and Orphanet disease lists. We focused on GARD diseases to support additional efforts sponsored by NCATS and relied on Orphanet as an external source of validity for the disease list. Gene and phenotype annotations for each disease were obtained from Orphanet's ORPHADATA v4.0 resource.<sup>12</sup> The Gene Ontology<sup>21,22</sup> (GO release October 26, 2021) and The Human Phenotype Ontology<sup>23</sup> (HPO release October 10, 2021) were both obtained from The OBO Foundry.<sup>24</sup> GO annotations for genes were obtained from NIH-NCBI (<ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz>; accessed on April 25, 2022). Metabolic, gene regulatory, and physical interactions between genes, gene products, and small molecules were obtained

from the Pathway Commons (PTC) v12<sup>25</sup> database. Based on published bioassay results, additional connections between genes and small molecules were obtained from Pharos v3.8.0,<sup>26,27</sup> developed and sponsored by NCATS.

### Rare disease network construction

HPO and GO ontologies were extended to include additional logical connections with the ELK reasoner.<sup>28</sup> The HPO class of "HP: 000005: Mode of Inheritance" and its subclasses were pruned from the HPO ontology because it created an overly connected network and is irrelevant to our use cases. The PTC and Pharos data were ingested and harmonized through the ChEMBL-ChEBI and ChEBI-PubChem mapping files provided by UniChem (accessed on April 25, 2022).<sup>29</sup> In cases where a particular entity mapped to multiple HPO or GO classes within a subtree of the ontology, eg, a gene mapped to 2 GO terms that shared a parent-child relationship such as protein binding (GO: 0005515) and kinase binding (GO: 0019901), only the annotation to the lowest subclass, eg, kinase binding, was kept.

### Graph node embeddings

Random walks emanating from each disease node were generated following a modified DL2Vec approach. The random walks were compiled into a corpus with each walk sequence. The length of each random walk and the number of random walks generated per disease were varied for sensitivity analysis. Many diseases have far more HPO phenotype annotations than gene associations, yet the gene associations provide a very informative connection to the molecular processes involved in the disease. Therefore, to give more weight to the gene annotations, the probability of taking a random walk step was balanced between the gene and HPO annotations for diseases with both annotation types. In the absence of this bias toward genes, the random walks are dominated by HPO terms, and thus the clusters only reflect phenotypic similarity among diseases (data not shown). Word2Vec was used to ingest the random walks and generate word embedding models for various combinations of walk length and walk count. We used the skip-gram Word2Vec architecture and varied the vector embedding dimension and the context window size.

Because we do not have a gold standard-labeled dataset, we relied on internal clustering metrics to tune random walk and embedding model hyperparameters. We used 3 internal clustering metrics, each capturing a different aspect of clustering quality: the silhouette score, which is defined as the difference between the mean intracluster distance and the mean nearest-cluster distance divided by the larger of the 2 values; the Davies-Bouldin index, which is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances; and the Calinski-Harabasz index, which is defined as the ratio of the sum of between-cluster dispersion and of within-cluster dispersion. Higher scores indicate more coherent and separated clusters for the silhouette score and the Calinski-Harabasz index, while lower values are superior for the Davies-Bouldin index.

Embedding models of different dimensions are not directly comparable with these internal clustering metrics. Therefore, we relied on the heuristic rule as guidance coupled with empirical analysis of the complexity of the feature space captured in the embedding model. We started by assessing the fourth root of the total number of words in our corpus. In our

case, the number of words is the total number of unique nodes and unique edges traversed in the random walks and results in a suggested embedding dimension of 26. Using this heuristic, we selected a range of embedding dimensions between 4 and 128 to test.

We then performed a sensitivity analysis with the other 3 parameters (the number of walks, the length of the walks, and the embedding context window size) within each embedding dimension. The number of walks per disease varied between 5 and 250. The length of each walk varied between 25 and 250. The embedding context window varied between 6 and 20. We used the word2vec as embedding model architecture with the skip-gram algorithm, no minimum word count, and Gensim default values for all other parameters; the models were trained for 15 epochs. Final model parameters were chosen based on the visual observation that increasing model complexity, eg, longer walks, more walks, larger context window, lead to diminishing returns with respect to improvement of the internal clustering metrics.

### Rare disease node clustering

The disease node embedding vectors were extracted from the Word2Vec models and concatenated to form a matrix where each row represents a disease, and each column is treated as a feature. We then applied K-means clustering, with Euclidean distance metric, using the Python package scikit-learn<sup>30</sup> to cluster diseases. The number of clusters was selected using a variant of the elbow method entitled the kneedle algorithm<sup>31</sup> as implemented in the kneed v0.8.1 Python package, using polynomial interpolation setting.

### Feature enrichment analysis

The node embedding process captures complex information regarding the local context surrounding a disease node within the KG, including phenotypes, genes, small molecules (including drugs and metabolites), and more. Unfortunately, this approach makes it difficult to interpret in detail why a particular set of diseases ended up in a cluster together, yet this information is key for understanding the results and determining the next steps. Therefore, we pursued 2 different forms of feature enrichment analysis.

First, we tested for gene enrichment within each cluster, with the aim of determining which genes were represented more frequently than expected by chance in each cluster. We counted the number of diseases associated with each gene within each disease cluster. To test against the null hypothesis that diseases were assigned to clusters independently of their gene annotations, the disease-to-cluster assignments were permuted 500 000 times, keeping the gene-to-disease annotations the same (noting that those are derived from the graph annotations and not our embeddings). The distribution of the counts of gene-to-cluster assignments for each gene within each cluster in the permuted data represents the null hypothesis of random grouping of genes within the clusters. A *P* value was calculated based on the number of permutations in which the counts of each gene within each cluster were greater than observed. A false discovery rate threshold of .01 was then applied by adjusting the permuted *P* values using the Benjamini-Hochberg procedure. The resulting cluster gene enrichment table was used for downstream drug repurposing applications.

Random walk feature importance was estimated by calculating the term frequency-inverse document frequency

(TF-IDF)<sup>32</sup> of each feature within windows surrounding the occurrence of each disease within the random walk corpus. The window size was selected to be consistent with the window size used in training the vector embedding model. The feature occurrences for each disease were summed within each cluster to obtain the term frequency. The presence or absence of each feature within each cluster was treated as the “document” frequency. An empirical cumulative distribution function was constructed from the TF-IDF values, and various percentile cutoffs were used to summarize how informative each feature type was for each cluster.

We sought to further interpret the known relationships among clustered diseases with respect to their enriched gene annotations. We, therefore, used STRINGDB v11.5<sup>33</sup> to (1) assess whether the enriched genes within each cluster have more connections than expected by chance and (2) identify enriched GO terms first for each set of genes.

### Assessment of disease cluster separation

We created shuffled random walk corpuses that contained no real information about the structure of the graph. Each random walk was simply represented by a sequence of graph nodes and edges selected completely at random from the original graph. The shuffled random walk corpuses were then used to build embedding and clustering models as described above.

The 3 internal clustering metrics described above (Silhouette\_Euclidean, Davies-Bouldin, and Calinski-Harabaz) were calculated for each clustering model built off the shuffled walk corpuses. The metrics calculated on the randomized graphs were then compared to those obtained from the original dataset to assess the clarity of separation in the original clusters.

### Semantic similarity validation

The Orphanet Rare Disease Ontology (ORDO v4.0)<sup>34</sup> was used to calculate pairwise semantic similarity among diseases based on the Sanchez information criteria.<sup>35</sup> We sampled 100 random sets of diseases of size 87 (the average number of diseases per cluster) to obtain a sampling distribution of average pairwise semantic similarity. We then performed a one-sample Student *t*-test to assess the difference in the mean average semantic similarity between the disease clusters and the random samples by assuming that the mean and variance parameters from the random samples represent the sampling distribution under the null hypothesis.

### Drug-target association cross-validation

We mapped putative targets to disease association and/or indication data for targets JAK2 and MPL through their approved drugs using Pharos and cross-referencing with Inxight.<sup>36</sup>

### Pathway enrichment analysis

We used EnrichR<sup>37</sup> for cross-validation pathway enrichment analyses. We extracted gene sets from our cluster gene enrichment table results and ran them through the user interface provided by EnrichR for each cluster. After an input gene set is submitted, the analysis is divided into different categories of enrichment. We focused on the KEGG 2021 Pathway enrichment results. By clicking on the column header, we can sort the table or clustergram by the term, *P* value, *z* score, or combined score.

## Results

### Exploratory analysis of the rare disease network

The node degree distribution of the knowledge graph is shown in [Supplemental Figure S1](#). The overall degree distribution of the knowledge graph is nearly linear on a log-log scale, indicative of a power-law distribution. The degree distribution is largely determined by the preponderance of small molecule nodes ( $N=348\,395$ ). To evaluate the connectivity between diseases in the graph, we calculated the distribution of shortest path lengths between every pair of disease nodes, and the results are shown in [Supplemental Figure S2](#). The most common path lengths were 4 and 2 corresponding to having 3 and 1 intermediate nodes, respectively. The hierarchical tree structure of the ontologies included in the network results in an increase in the frequency of paths of length 4 compared to paths of length 3 or 5. For example, a common type of path between diseases goes as follows: disease, HPO phenotype, HPO class, HPO phenotype, disease; this type of path has an edge length of 4. The longest path between any 2 disease nodes in the graph was 6.

### Optimization of embedding dimension

We performed principal component analysis on the disease embedding vectors from models with different embedding dimensions. By plotting the explained variance as a function of the number of principal components, we can visually inspect the degree to which the dimensionality of the embedding space can be reduced. We built models with embedding dimensions of 4, 8, 16, 32, 64, and 128. We observed that for models with embedding dimensions of 4, 8, or 16, there was no drop-off in variance explained by successive PCs indicating that these feature spaces could not be significantly reduced. In contrast, models with a dimension of 32 or higher showed a large decrease in variance explained by higher PCs, suggesting

that those feature spaces could be reduced ([Supplemental Figure S3](#)). Therefore, an embedding dimension of 32 was selected for the final analysis. It is useful to note that this embedding dimension is roughly consistent with that recommended by the fourth root heuristic.

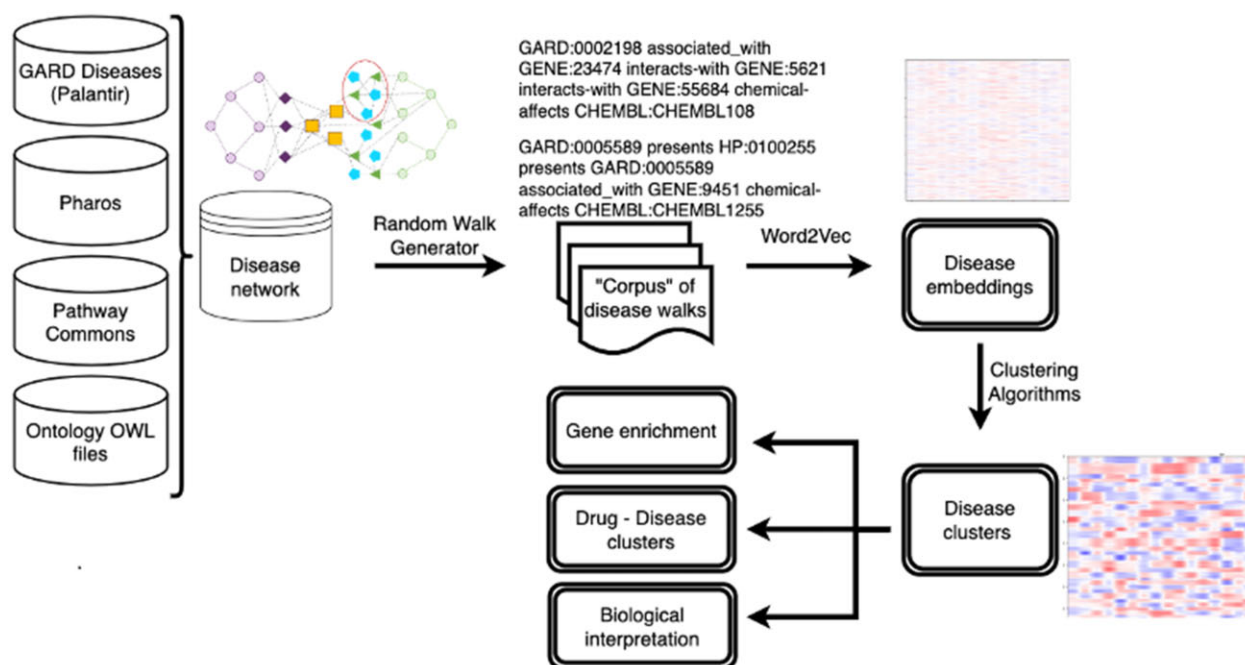
### Disease clusters

A total of 3242 diseases were used to construct a rare disease network that included data on genes, phenotypes, small molecules, biological pathways, and biomedical ontologies. The rare disease network contained 439 691 nodes and 2 716 895 edges (see the [Supplemental Material](#) for an exploratory data analysis of the network). [Figure 1](#) illustrates our workflow.

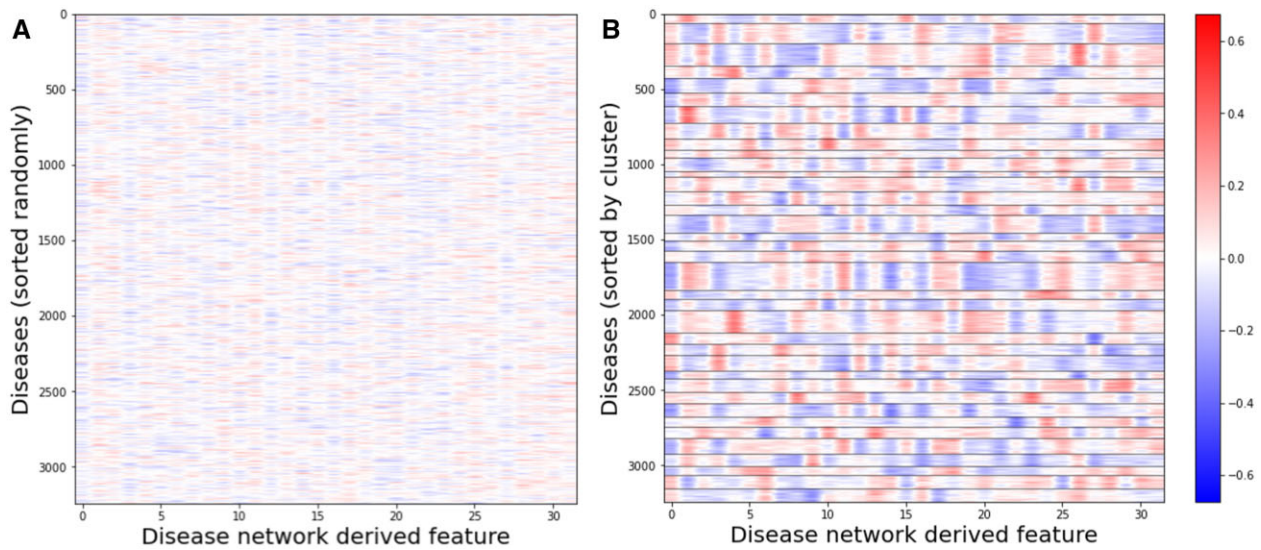
Several random walk and embedding model parameters, including the number of walks, the length of the walks, the embedding context window size, and the overall embedding dimension, were systematically varied to optimize the quality of the disease clusters produced. [Supplemental Figure S4](#) shows several internal clustering metrics as a function of the number of walks per disease, across the various walk lengths and context window sizes for an embedding dimension of 32. The number of walks per disease had the largest effect on all clustering metrics. Performance increased with the number of walks with a plateau reached beyond 250 walks per disease. Based on this sensitivity analysis, we selected the following model parameters: 250 walks, walk length of 250, and context size of 20.

### Final disease clustering model and its evaluation

Our final disease cluster model contained 37 clusters, with an average of 87 and a median of 83 diseases per cluster. The distribution of cluster sizes is shown in [Supplemental Figure S5](#). In [Figure 2](#), the disease embedding values are plotted as a heatmap with diseases sorted either randomly or by cluster



**Figure 1.** Rare disease clustering workflow. Input data resources are integrated into a single rare disease-focused network. Random walks are performed to create a corpus of surveying the local context around each disease. Node embeddings are created and clustered. Post hoc analyses are conducted to interpret and utilize the disease clusters.



**Figure 2.** Disease embedding vectors are plotted in heatmaps (A) randomly sorted and (B) sorted by cluster. Each column in the heatmap corresponds to a dimension within the embedding vector space, and each row corresponds to a disease. (B) The clusters are demarcated with horizontal black lines.

assignment, which shows that diseases with similar patterns across the embedding values tend to group together into clusters. To assess the degree of separation between the disease clusters, we projected the disease clusters into a 2D t-SNE map. The t-SNE projection of the disease clusters, colored by their cluster assignments, is depicted in [Figure 3A](#). The t-SNE map shows apparent separation among the disease clusters, especially when compared to an equivalent t-SNE map of embeddings constructed based on randomly shuffled walks ([Supplemental Figure S6a](#)).

### Evaluation of disease cluster separation

We evaluated the extent of our original cluster separation in 2 ways. First, we calculated a within-cluster semantic similarity index and compared that with randomly sampled disease sets. Specifically, we utilized the ORDO to calculate the Sanchez intrinsic information criterion within each disease cluster (see “Methods”). [Figure 3D](#) shows the distribution of the average semantic similarity both within each disease cluster and across a set of 100 randomly sampled diseases of size 87 (the average number of diseases per cluster). The *t*-test results suggest that the clustered diseases are significantly more semantically similar than a random selection of diseases ( $P = .00024$ ). Second, we compared internal clustering metrics calculated on clusters derived from randomized graphs to those calculated from the original data (see summary in [Supplemental Table S2](#)). [Figure 3C](#) clearly shows that the real clustering model metrics fall outside the range of metrics generated on the randomized graphs. Together, these results indicate that the disease clustering model has captured information contained within the knowledge graph that is useful for assessing similarity among diseases.

### Cluster feature enrichment

Our first feature enrichment analysis focused strictly on direct gene annotations. By permuting the disease cluster assignments, we identified 585 genes enriched in at least one cluster at an FDR (Benjamini-Hochberg) *q*-value cutoff of 0.01; all cluster gene enrichment results are presented in [Table S1](#). We found that 12 genes were enriched within more than one

cluster. The genes enriched in more than one cluster include genes associated with some major classes of diseases, such as (1) oncogenes: KRAS, PTEN, TP53, KIT, and FGFR1; (2) genes associated with musculoskeletal phenotypes: COL1A1, FKTN, GMPPB, POMT1, and POMT2; and (3) genes associated with blood disorders: HBB, NPM1. The remaining 573 genes were enriched within only one cluster each (totaling 33 clusters with at least one enriched gene).

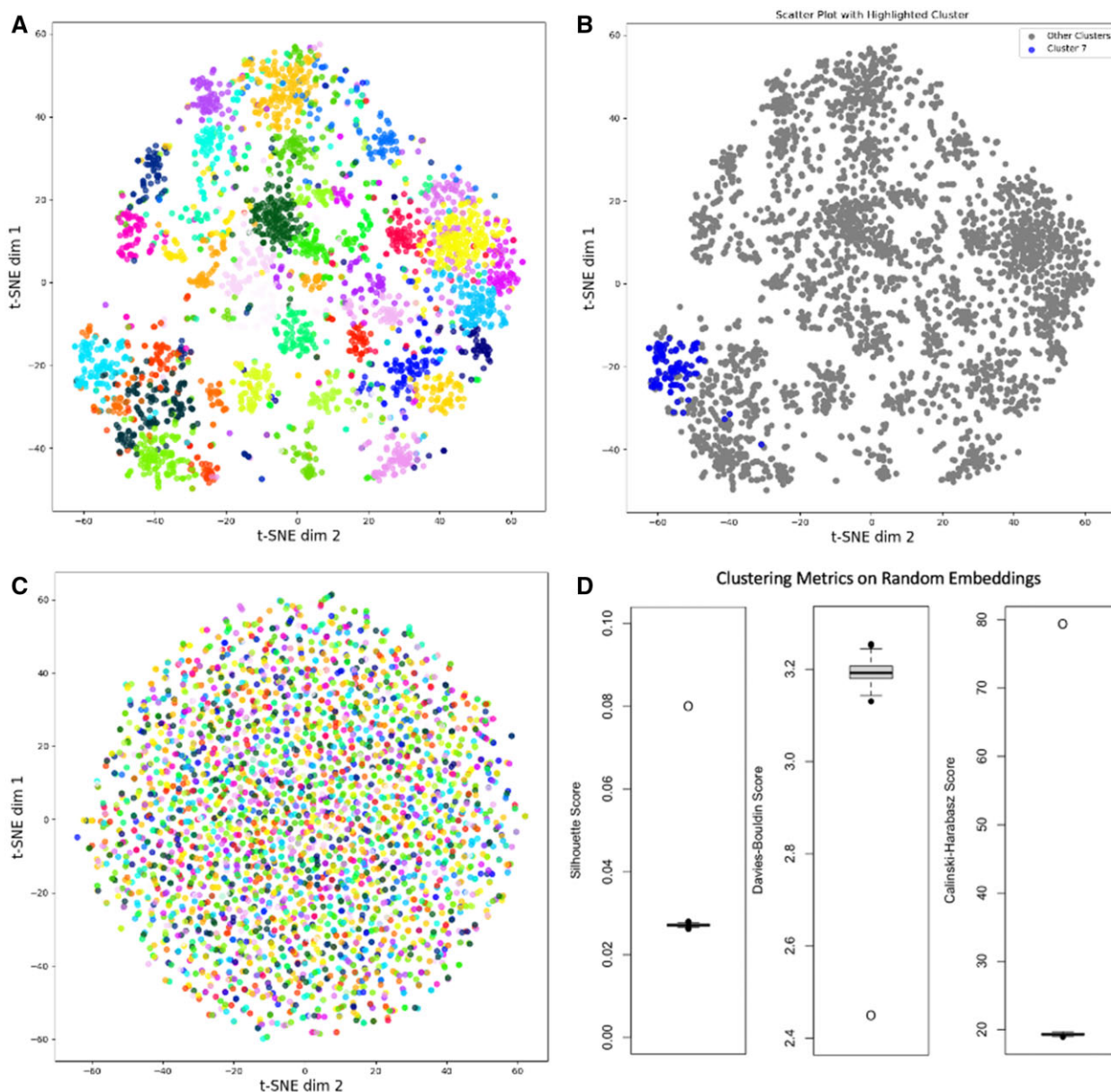
Second, we analyzed the random walks to identify context features (eg, genes, HPO terms, etc.) that comprise each disease cluster. [Figure 4](#) shows the number of nodes of each type as a function of the TF-IDF percentile threshold for a set of clusters selected to be exemplary of different cluster archetypes. The cluster archetypes we identified include those almost exclusively dominated by HPO terms (clusters 16 and 19), those where the highest end of the TF-IDF distribution is dominated by genes (clusters 22 and 24), clusters with important GO terms and genes (cluster 7) and clusters with no features among the higher TF-IDF percentiles (cluster 30).

### Interpreting the disease clusters

The sets of enriched gene annotations within each cluster were queried against STRINGDB. The enriched gene sets from 27 of the 33 clusters having at least one enriched gene had significantly more connections within STRINGDB than expected by chance. We manually reviewed the diseases, enriched genes, and enriched GO terms within each cluster with the goal of constructing concise descriptions of each cluster. [Table 1](#) describes 3 clusters (3, 7, and 28), 2 of which group diseases with globally similar clinical manifestations. To see all cluster-enriched gene sets to enriched GO (biological processes) terms, see [Supplemental Figure S7](#).

For example, cluster 7 ([Figure 3B](#)) is primarily composed of ocular diseases, such as Cone-Rod Dystrophy and Leber Congenital Amaurosis. Cluster 28 contains several cardiac and electrophysiological diseases caused by ion channel mutations (ie, channelopathies).

Cluster 3 though contains a mix of neurological, skeletal, and genetic disorders. More specifically, Charcot-Marie Tooth disease type 2C is a nerve damage disorder, while



**Figure 3.** Visualizing and quantifying similarity within disease clusters. (A) t-SNE projections of the disease embedding vectors were created and plotted, points are colored according to their cluster membership. (B) The same t-SNE projections as in (A) are plotted, but modifying the color coding to highlight a single cluster, in this case, cluster 7. To find all individual t-SNE single clusters highlighted, go to <https://doi.org/10.6084/m9.figshare.23748846>. (C) t-SNE map of embeddings constructed based on randomly shuffled walks. (D) The distribution of three clustering metrics (Silhouette, Davies-Bouldin, and Calinski-Harabasz) derived from the randomly shuffled embedding models with the real clustering model metrics plotted as single points.

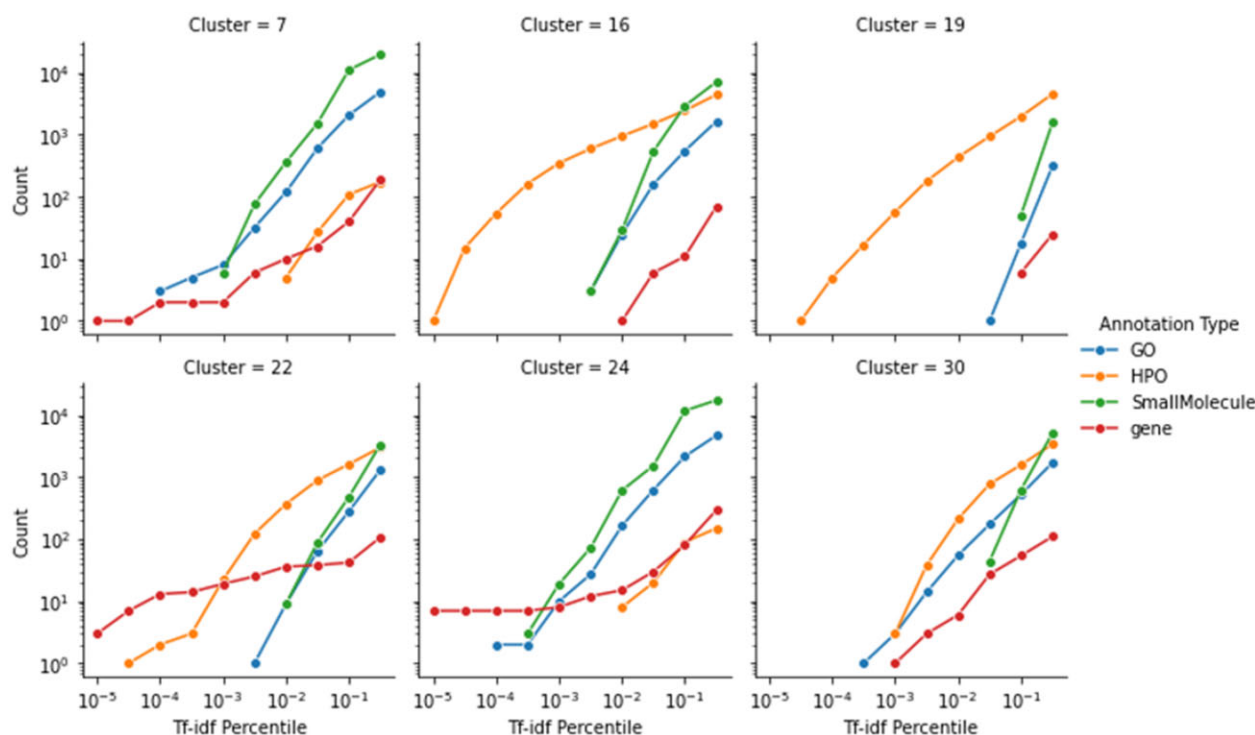
parastremmatic dwarfism and metatropic dysplasia are dwarfism-related diseases. This cluster also includes other skeletal and muscle degeneration disorders such as Duchenne Muscular Dystrophy and brachyolmia type. We note that many of these diseases are genetic disorders, and our approach correctly pulled out disease-causing genes such as TRPV4, which cause Charcot-Marie Tooth disease type 2C,<sup>38</sup> metatropic dysplasia,<sup>39</sup> brachyolmia type 3,<sup>40</sup> and parastremmatic dwarfism.<sup>41</sup> Others, like aggregate tubular myopathies, lack a clear genetic cause, yet a recent study<sup>42</sup> pointed to the involvement of potentially causative genes related to calcium signaling pathway. Notably, the calcium signaling pathway is highlighted as an enriched GO term for this cluster, and 2 other potential causative genes, ORAI1 and STIM1, are also strongly linked to this cluster, but not with sufficient causal

evidence to have appeared in our original knowledge graph (Table 1).

### Utility for drug repurposing

Drug repurposing is one important use case of our rare disease clusters. Identifying legitimate drug repurposing candidates will require an in-depth analysis of the clustered diseases and their connections to drugs. Here, we sought to describe the connections between known drugs (from ChEMBL database), gene targets (GO annotations), and disease clusters (GARD database).

Utilizing data from Pharos, Figure 5A shows the distribution of the number of gene targets by cluster and broken down by target druggability level (TDL). The TDL (consisting of 4 categories: Tclin, Tchem, Tbio, and Tdark) is a



**Figure 4.** Counts by annotation feature type as a function of TF-IDF percentile threshold for a selected set of exemplary clusters.

qualitative label assigned to targets based on what is known about their chemistry, biology, and whether an approved drug is available.<sup>26,27</sup> Moreover, Tclin targets have Food and Drug Administration (FDA)-approved drugs with known mechanisms of action that target them. Tchem targets have at least one compound with an activity cutoff of <30 nM and have been shown to have one or more active ligands. Tbio targets have published literature characterizing the target; however, no known drug or active ligands have been published. Lastly, Tdark targets are considered understudied targets.

At the time of writing this paper, the overall distribution of TDL values in Pharos was 11 867 Tbio, 5932 Tdark, 1930 Tchem, and 685 Tclin, totaling 20 412 proteins. [Figure 5A](#) shows that Tbio is the largest category of gene targets in every cluster. In contrast to the overall TDL, relatively few targets with associations to the clustered rare diseases are rated Tdark. Thirty-one out of 37 clusters have at least 1 gene target with a Tclin or Tchem rating, indicating that many of the clusters have putative drug repurposing candidates that could be mined in a more detailed analysis. Therefore, to explore the space of drug connections to our disease clusters, we filtered the original data from Pharos to include only approved drugs, ie, Tclin targets, in each cluster ([Figure 5B](#)). Among clusters with direct connections to drugs, the number of unique drugs per cluster ranged from 356 drugs in cluster 28 to just a single drug in cluster 6. In addition, clusters 2, 16, and 19 had no known drug connections.

We then examined a randomly selected cluster, 35, to find potential insights for drug repurposing via our disease cluster gene enrichment results.<sup>45</sup> Cluster 35 contains 29 Tclin targets, with CTLA4, JAK2, and MPL being the top 3 significantly enriched genes (see [Supplemental Table S1](#)). Interestingly, JAK2 and/or MPL mutations are well associated with myeloproliferative neoplasms, also known as myeloproliferative disorder (MPD).<sup>44</sup> Within cluster 35, JAK2

and MPL share an association with GARD-labeled diseases: polycythemia vera, essential thrombocythemia, and primary myelofibrosis. Notably, JAK2 and MPL are known to have direct protein-protein interactions<sup>34</sup> ([Supplemental Figure S8](#)). Further, JAK2 is known to participate in the JAK2/STAT cellular signaling cascade, which is important for the regulation of gene expression involved in basic cell processes.<sup>45</sup>

To cross-reference our results for drug use/indications, some examples of FDA-approved drugs for targeting JAK2 were queried from Inxight<sup>36</sup> (access date: July 24, 2023). Three of 5 approved drugs have indications for myelofibrosis, polycythemia vera, and MPD, all of which are present in our cluster 35 ([Supplemental Figure S6](#)). Whereas drugs that target MPL are indicated for immune thrombocytopenia and severe aplastic anemia indications, both of which are also in our cluster 35. It is worth noting that these drugs are used to treat nonrare diseases as well (eg, rheumatoid arthritis, atopic dermatitis, and chronic liver disease). Another approach for identifying drug repurposing candidates within these clusters is to look into pathway similarities between diseases within clusters. The top KEGG-enriched pathways derived from the entire target set for cluster 35 include, in order of statistical significance, primary immunodeficiency, cytokine-cytokine receptor interaction, inflammatory bowel disease, leishmaniasis, Th17 cell differentiation, tuberculosis, and JAK-STAT signaling pathway ([Supplemental Figure S7](#)). However, when inputting Tclin targets only, the most enriched pathway is JAK-STAT signaling, further supporting the relevance of JAK2/MPL targeting for other diseases in cluster 35 ([Supplemental Figure S8](#)). Overall, these results demonstrate how starting from a disease cluster and associated Tclin target genes, new insights for drug repurposing via adjacent diseases can readily be extracted using public resources.

**Table 1.** Example disease clusters with enriched genes and gene ontology terms.

Cluster	Exemplary diseases	Genes	Exemplary GO terms
3	Charcot-Marie-Tooth disease type 2C; Metatropic dysplasia; Brachyolmia type 3; Parastremmatic dwarfism; Spondylometaphyseal dysplasia; Centronuclear myopathy; King Denborough syndrome; Myopathy congenital; Cap myopathy; Congenital fiber type disproportion; Freeman-Sheldon syndrome; Distal arthrogryposis type 1; Spinocerebellar ataxia 15/29; Tubular aggregate myopathy; Rigid spine syndrome; Dysferlinopathy; Becker muscular dystrophy; Duchenne muscular dystrophy	TRPV4; RYR1; TPM3; NALCN; ITPR1; NEB; MYH3; TTN; AK9; CHRNA1; CHRNE; CHRNB1; CACNA1S; CHRNG; ORAI1; BIN1; TPM2; ACTA1; KLHL41; TCAP; CHRND; MYPN; STIM1; MYH7; SELENON; DYSF; DMD; LMOD3; RAPSIN	Muscle filament sliding; Muscle organ development; Myofibril assembly; Synaptic transmission, cholinergic; Skeletal muscle tissue development; Neuromuscular synaptic transmission; Skeletal muscle thin filament assembly; Sarcomere organization; Regulation of heart contraction; Calcium ion transport; Regulation of membrane potential; Ligand-gated cation channel activity
17	Leber congenital amaurosis; Cone-rod dystrophy; Achromatopsia 2/3; Stargardt disease; Usher syndrome type 1/2A/3A; Corneal dystrophy; Coats disease; Norrie disease; Retinal cone dystrophy 1	NMNAT1; SAG; RPGR; SPATA7; ZNF408; CDHR1; RHO; PDE6B; RPE65; CRX; TULP1; ABCA4; BEST1; USH2A; PROM1; TGFB1; IMPG2; NDP; OPN1MW; GUCA1A; PDE6H; GNAT2; CNGA3; PDE6C; CNGB3; ATF6; AIPL1; TIMP3; RDH12; IMPDH1; GNAQ; GRK1; PRPH2; LRAT; CLRN1; LRP5; CACNA1F; MYO7A; GUCY2D; TYR; CACNA2D4; FZD4; ADCY5; OPN1LW	Visual perception; Retina homeostasis; Phototransduction, visible light; Photoreceptor cell maintenance; Regulation of rhodopsin mediated signaling pathway
28	Autosomal recessive pseudohypoaldosteronism type 1; Liddle syndrome; Brugada syndrome; Thomsen and Becker disease; Familial hemiplegic migraine; Familial infantile convulsions and paroxysmal choreoathetosis; Benign familial infantile epilepsy; Paroxysmal kinesigenic choreoathetosis; Early Infantile Epileptic Encephalopathy; West syndrome; Familial primary hypomagnesemia; Dravet syndrome; Familial atrial fibrillation; Long QT syndrome 1; Progressive familial heart block type 1B/1A/2; Andersen-Tawil syndrome; Hyperkalemic periodic paralysis; Potassium aggravated myotonia; Rapid-onset dystonia-parkinsonism; Congenital insensitivity to pain; Paroxysmal extreme pain disorder; Erythromelalgia	SCNN1A; CLCN1; PRRT2; CACNA1A; SCNN1G; SIK1; KCNA1; SCN2A; SCN2B; SCN4B; SCN3B; EEF1A2; SCN1B; KCNJ2; SCN4A; PLCB1; TRPM4; ATP1A3; KCNE2; KCNQ2; SCN5A; SCN11A; NKX2-5; SCNN1B; ATP1A2; KCNQ3; SCN1A; KCNT1; SCN8A; SCN9A; SCN10A; AKAP9; ABCC9; DNM1; GRIN1; GRIN2B; KCND3; SYNGAP1; KCNJ8; KCNQ1; GABRA1; KCNJ10; GRIN2A; SLC25A22; KCNE3; CHD2; GABRG2; KCNE1; SLC4A11	Sodium ion transmembrane transport; Cardiac muscle contraction; Transmission of nerve impulse; Blood circulation; Regulation of ventricular cardiac muscle cell membrane repolarization; Ventricular cardiac muscle cell action potential; Sensory perception of pain; Ion channel binding; Ligand-gated cation channel activity; Potassium channel regulator activity; Calmodulin binding; Glutamate-gated calcium ion channel activity

Note: For list of all disease clusters and their (biological processes) gene-to-GO enrichment terms to go:

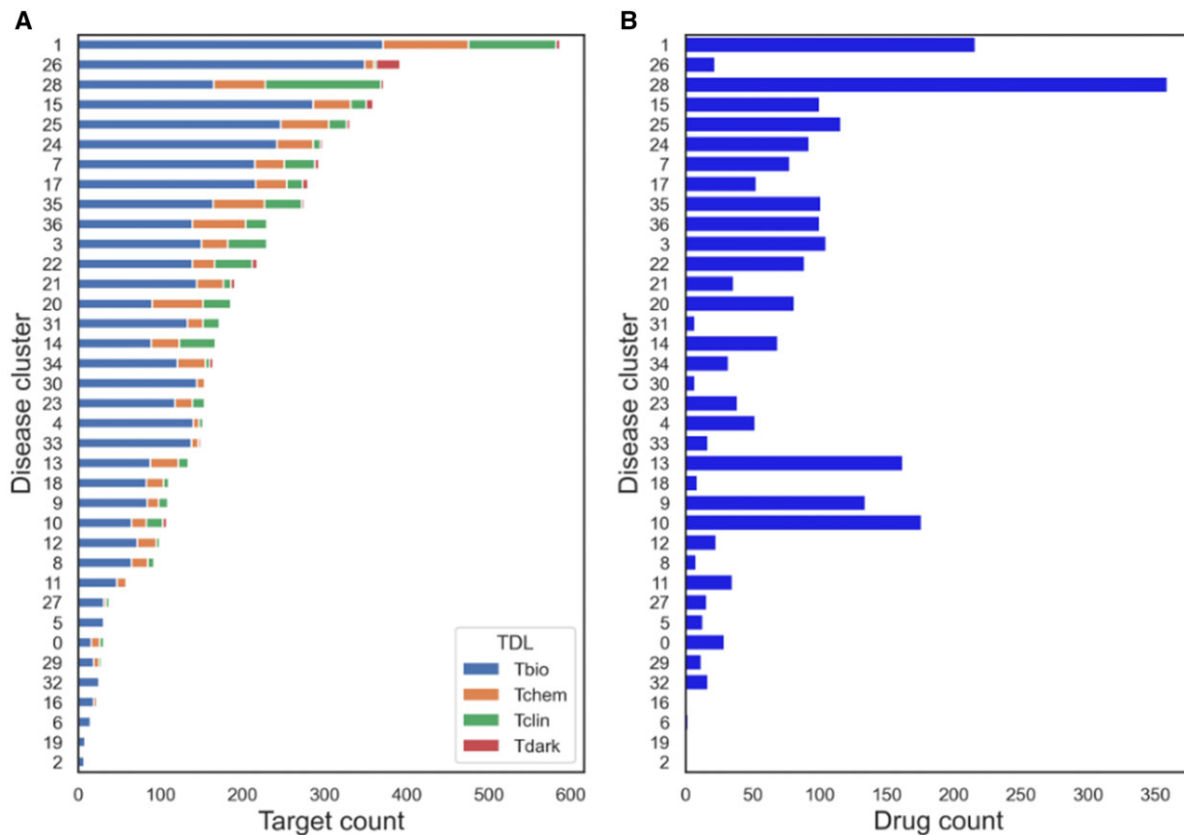
## Discussion

We constructed a knowledge graph based on the overlap between rare diseases tracked by GARD and Orphanet. The graph is enriched with additional information on small molecules and biological pathways. We used this enriched network to construct graph node embedding vectors for each disease. Those embedding vectors were used as a feature matrix in k-means clustering analysis. Hyperparameters of the embedding model and the k-means model were selected by a combination of heuristics, sensitivity analyses, and explicit tuning. Our method identified 37 disease clusters with an average of 87 diseases each. The quality of the resulting disease clustering model was validated by comparing semantic similarity within clusters to randomly selected disease sets based on the ORDO, which was not part of our disease network. The semantic similarity of the clustered diseases combined with a visual inspection of the cluster-sorted embedding vectors and feature enrichment analysis suggests that our method has identified groups of diseases with features in common. Furthermore, an in-depth review of clustered diseases, enriched genes, and enriched GO terms

showed that many clusters are clearly composed of related diseases based on their causal gene and pathophysiology.

The use of semantic similarity within clusters as a form of validation begs the question: Why was semantic similarity not the primary basis for clustering the diseases in the first place? Our aim was to expand beyond the semantic ontological organization of diseases by (1) directly using data related to the diseases and (2) using higher-order relationships between the diseases across data modalities. The former justification was taken up by both eRAM<sup>10</sup> and the RDMAP<sup>11</sup> projects, which developed similarity scores using both Phenotype-HPO and Gene-GO linkages to rare diseases. Our approach expands upon those methods by integrating the Phenotype-HPO, Gene-GO, Pathway Commons, and Pharos datasets into a single network. We capture higher-order relationships among diseases by building graph node embeddings. The graph node embeddings provided an integrated representation of the network context of each disease across the heterogeneous data types present in the network. However, our results are only as comprehensive





**Figure 5.** Summary of Pharos data by disease cluster. (A) The number of gene targets with a TDL label from Pharos within each cluster, and (B) the number of drugs connected to each cluster through gene targets. Drugs were obtained from the Pharos based on their "Tclin" ligand readiness level assignment.

as the underlying input data. Various sources of bias, including publication bias toward more prevalent diseases, limit the generalizability of our results.

Another key limitation of our study is the absence of common diseases. Most of the biomedical data pertain to common diseases. Therefore, expanding our knowledge graph to incorporate common disease information would greatly increase the scope and translational relevance of the work. However, expanding the graph would also create challenges surrounding data source selection and the overwhelming rare disease signal. Nevertheless, one goal for future work will be the incorporation of common disease data into the analysis.

Our analysis creates an additional layer of structure onto the large pool of rare diseases. This structure will help strengthen drug repurposing efforts by enabling focus on smaller disease sets. Yet it must be recognized that our analysis on its own does not directly yield translatable results. In-depth follow-ups, such as detailed subnetwork analysis or literature review, will be required to take full advantage of our work—a task that will be taken up in a related manuscript.

## Conclusion

Our approach expands upon prior efforts to identify similarities of rare diseases by integrating multiple data types and considering the higher-order structure of the rare disease network simultaneously. We show that diseases in the clusters are enriched for similar gene annotations and that there are many possible connections to approved and investigational

drugs. Future work will focus on expanding the knowledge graph with common disease data and detailed subnetwork analysis of the most promising clusters.

## Acknowledgments

We would like to thank Dac-Trung Nguyen, Yanji Xu, Eric Sid, and Andy Patt for their insights during the formation of this project.

## Author contributions

JS designed the study, executed the analysis, and wrote the manuscript. JB and AY carried out analysis and wrote the manuscript. QZ and EW designed the study and revised the manuscript.

## Supplementary material

[Supplementary material](#) is available at *Journal of the American Medical Informatics Association* online.

## Funding

This work was supported by the Intramural Research Program of the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, grant number ZIC TR000410-03.

## Conflicts of interest

None declared.

## Data availability

Publicly available data used in our workflow are referenced in scripts within the GitHub repository. GARD data used are presently not accessible to the public and are therefore provided as CSV files within the GitHub repository. All code for executing analysis and constructing figures is contained here: <https://github.com/ncats/RD-Clust>. Specific computational parameters for deploying the workflow on SLURM HPC cluster are provided in the GitHub repository; the slowest computational step is fitting the embedding model which will take up to 36 hours on a CPU node with 8GB of RAM.

## References

- Field MJ, Boat TF, eds. *Rare Diseases and Orphan Products: Accelerating Research and Development*, in *Rare Diseases and Orphan Products: Accelerating Research and Development*. Washington, DC: National Academies Press; 2010.
- Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*. 2020;28(2):165-173.
- Tisdale A, Cuttillo CM, Nathan R, et al. The IDEaS initiative: pilot study to assess the impact of rare diseases on patients and healthcare systems. *Orphanet J Rare Dis*. 2021;16(1):429.
- U.S. Government Accountability Office. *Rare Diseases: Although Limited, Available Evidence Suggests Medical and Other Costs Can Be Substantial*. Report # GAO-22-104235. 2021. <https://www.gao.gov/assets/gao-22-104235.pdf>
- Haendel M, Vasilevsky N, Unni D, et al. How many rare diseases are there? *Nat Rev Drug Discov*. 2020;19(2):77-78.
- Brooks PJ, Yang NN, Austin CP. Gene therapy: the view from NCATS. *Hum Gene Ther*. 2016;27(1):7-13.
- Park JJH, Siden E, Zoratti MJ, et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials*. 2019;20(1):572.
- Jarada TN, Rokne JG, Alhadj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Cheminform*. 2020;12(1):46.
- Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*. 2017;6:e26726.
- Jia J, An Z, Ming Y, et al. eRAM: encyclopedia of rare disease annotations for precision medicine. *Nucleic Acids Res*. 2018;46(D1):D937-D943.
- Yang J, Dong C, Duan H, et al. RDmap: a map for exploring rare diseases. *Orphanet J Rare Dis*. 2021;16(1):101-111.
- Orphanet: an online rare disease and orphan drug database. © INSERM. 1999. <http://www.orpha.net>. Accessed August 15, 2022.
- Zhu Q, Nguyen D-T, Aleya G, et al. Phenotypically similar rare disease identification from an integrative knowledge graph for data harmonization: preliminary study. *JMIR Med Inform*. 2020;8(10):e18395.
- Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; August 24-27, 2014: 701-710; New York, NY.
- Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. 2016. arXiv 1607.00653, preprint. <https://doi.org/10.48550/arXiv.1607.00653>
- Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*; May 2-4, 2013; Scottsdale, Arizona.
- Smaili FZ, Gao X, Hoehndorf R. OPA2Vec: Combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*. 2019;35(12):2133-2140.
- Chen J, Althagafi A, Hoehndorf R. Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics (Oxford, Engl)*. 2021;37(6):853-860.
- Chen Y, Hu Y, Hu X, et al. CoGO: a contrastive learning framework to predict disease similarity based on gene network and ontology structure. *Bioinformatics*. 2022;38(18):4380-4386.
- Zhu Q, Nguyen D-T, Grishagin I, et al. An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). *J Biomed Semantics*. 2020;11(1):13.
- Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29.
- Carbon S, Douglass E, Good BM, et al. The Gene Ontology resource: enriching a Gold mine. *Nucleic Acids Res*. 2021;49(D1):D325-D334.
- Köhler S, Gargano M, Matentzoglou N, et al. The human phenotype ontology in 2021. *Nucleic Acids Res*. 2021;49(D1):D1207-D1217.
- Jackson R, Matentzoglou N, Overton JA, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database (Oxford)*. 2021;2021:baab069.
- Rodchenkov I, Babur O, Luna A, et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic Acids Res*. 2019;48(D1):D489-D497.
- Sheils TK, Mathias SL, Kelleher KJ, et al. TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res*. 2021;49(D1):D1334-D1346.
- Kelleher KJ, Sheils TK, Mathias SL, et al. Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Res*. 2023;51(D1):D1405-D1416.
- Kazakov Y, Krötzsch M, Šimánčík F. The incredible ELK. *J Autom Reason*. 2014;53(1):1-61.
- Chambers J, Davies M, Gaulton A, et al. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J Cheminform*. 2013;5(1):3-9.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
- Satopaa V, Albrecht J, Irwin D, et al. Finding a “Kneedle” in a haystack: detecting knee points in system behavior. In: *2011 31st International Conference on Distributed Computing Systems Workshops*; June 20-24, 2011: 166-171; Minneapolis, MN.
- Sammur C, Webb GL, eds. TF-IDF. In: *Encyclopedia of Machine Learning*. Boston, MA: Springer; 2010: 986-987.
- Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605-D612.
- Vasant D, Chanas L, Malone J, Hanauer M. Ordo: an ontology connecting rare disease, epidemiology and genetic data. In: *Proceedings of ISMB*; July 13-15, 2014; Boston, MA.
- Sánchez D, Batet M, Isern D. Ontology-based information content computation. *Knowl Based Syst*. 2011;24(2):297-303.
- Siramshetty VB, Grishagin I, Nguyễn ĐT, et al. NCATS Inxight Drugs: a comprehensive and curated portal for translational research. *Nucleic Acids Res*. 2022;50(D1):D1307-D1316.
- Xie Z, Bailey A, Kuleshov MV, et al. Gene set knowledge discovery with Enrichr. *Curr Protoc*. 2021;1(3):e90.
- Landouré G, Zdebik AA, Martínez TL, et al. Mutations in TRPV4 cause Charcot-Marie-Tooth disease type 2C. *Nat Genet*. 2010;42(2):170-174.

39. Krakow D, Vriens J, Camacho N, et al. Mutations in the gene encoding the calcium-permeable ion channel TRPV4 produce spondylometaphyseal dysplasia, Kozlowski type and metatropic dysplasia. *Am J Hum Genet.* 2009;84(3):307-315.
40. Rock MJ, Prenen J, Funari VA, et al. Gain-of-function mutations in TRPV4 cause autosomal dominant brachyolmia. *Nat Genet.* 2008;40(8):999-1003.
41. Nishimura G, Dai J, Lausch E, et al. Spondylo-epiphyseal dysplasia, Maroteaux type (pseudo-Morquio syndrome type 2), and parastremmatic dysplasia are caused by TRPV4 mutations. *Am J Med Genet A.* 2010;152A(6):1443-1449.
42. Gang Q, Bettencourt C, Brady S, et al. Genetic defects are common in myopathies with tubular aggregates. *Ann Clin Transl Neurol.* 2022;9(1):4-15.
43. Sanjak J, Binder J, Mathe EA. GARD gene enrichment results. figshare. Dataset. 2023. [10.6084/m9.figshare.23748060.v1](https://doi.org/10.6084/m9.figshare.23748060.v1)
44. Passamonti F, Maffioli M, Caramazza D, Cazzola M. Myeloproliferative neoplasms: from JAK2 mutations discovery to JAK2 inhibitor therapies. *Oncotarget.* 2011;2(6):485-490.
45. Sopjani M, Morina R, Uka V, Xuan NT, Dërmaku-Sopjani M. JAK2-mediated Intracellular Signaling. *Curr Mol Med.* 2021;21(5):417-425.