Representation Learning of Human Disease Mechanisms for a Foundation Model in Rare and Common Diseases

Babak Ravandi [1,2,3,†], William R. Mowrey[1,*], Ayan Chatterjee[1,2,*]
Piero Ricchiuto[1], Parham Haddadi[1], Mario Abdelmessih[1], Wei Ding[1],
Simon Lambden[1], Michaël Ughetto[4], Ian Barrett[5], Tom Diethe[6],
Guillermo Del Angel[7], Tina Eliassi-Rad[2,8,9,10]

[1]Alexion AstraZeneca Rare Disease, Boston, USA
[2]Network Science Institute, Northeastern University, Boston, USA
[3]Department of Physics, Northeastern University, Boston, USA
[4]Biological Insight Knowledge Graph (BIKG), AI Engineering, AstraZeneca, Gothenburg, Sweden
[5]Data Sciences and Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[6]Centre for AI, Data Science & Artificial Intelligence, AstraZeneca, Cambridge, UK
[7]Center for Genomics Research, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[8]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[9]Santa Fe Institute, Santa Fe, NM, USA
[10]The Institute for Experiential AI, Northeastern University, Boston, MA, USA

[†]Corresponding author e-mail: Babak.Ravandi@Alexion.com
[*]William R. Mowrey and Ayan Chatterjee equally contributed.

The limited amount of data available renders it challenging to characterize which biological processes are relevant to a rare disease. Hence, there is a need to leverage the knowledge of disease pathogenesis and treatment from the wider disease landscape to understand rare disease mechanisms. Furthermore, it is well understood that rare disease discoveries can inform the our knowledge of common diseases. In this paper, we introduce Dis2Vec (Disease to Vector), a new representation learning method for characterizing diseases with a focus on learning the underlying biological mechanisms, which is a step toward developing a foundation model for disease-association learning. Dis2Vec is trained on human genetic evidence and observed symptoms, and then evaluated through cross-modal transfer-learning scenarios based on a proposed drug association learning benchmark with drug targets (positive controls) and Orphanet Rare Disease Ontology (negative controls). Finally, we argue that clustering diseases in the Dis2Vec space, which captures biological mechanisms instead of drug-repurposing information, could increase the efficiency of translational research in rare and common diseases, and ultimately improve treatment strategies for patients.

INTRODUCTION

There are may pathophysiological pathways that are shared in both common and rare disease. However, the number of rare diseases is currently estimated to be over 10,000 [1, 2]. Globally, more than 300 million people are living with rare diseases and there are no approved therapies for over 90% of these disorders [3, 4]. It is estimated that 39% of rare diseases are genetic in etiology [5], meaning that they are directly caused by specific genetic variants. Unfortunately, for the remaining 60% of rare diseases genetic insights are generally lacking, as small patient populations preclude population genetic studies. As a result, 88% of rare diseases have less than 10 gene associations (Figure S2). The lack of disease-gene associations creates a scarcity of potential therapeutic targets (seed genes), and hence makes it difficult to use established bioinformatics tools for target identification, such as disease module identification based on Protein-Protein Interaction (PPI) networks [6]. For instance, Buphamalai et al. applied various network analysis methods on rare diseases and concluded that more than 20 disease-gene associations are needed to build a disease module [7], due to the incompleteness of data in the PPI networks [8].

Knowledge repurposing within and across both common and rare diseases provides a unique opportunity to develop comprehensive and mechanistic representations of diseases and their relationships to each other, which we refer to here as "disease associations". Historically, rare disease research enhanced our core understanding of human biology and drug discovery for common diseases [9–13] as rare diseases offer a unique window to human biology through their strong genetic links [14]. However, this repurposing can also occur in the reverse direction, with weak effect genetic signals serving to complete the mechanistic understanding of rare disease biology (where population genetics studies are often infeasible). These two knowledge repurposing opportunities are illustrated in Figure 1A-B. Finally, disease association learning can help to address inefficiencies in rare disease research created by data silos built around individual diseases and traditional areas of clinical practice, rather than shared biological mechanisms [15–19].

Clustering diseases based on shared biological mechanism could potentially accelerate knowledge repurposing and translation to new therapies. For example, basket trials on rare diseases shows the effectiveness of grouping multiple diseases in accelerating translation [20], referred to as the Same-Target Multiple-Diseases strategy. Further, aggregating rare diseases with shared biology can help to overcome the paucity of gene associations by incorporating overall biological pathways and uncovering novel therapeutic targets. This is a key need as many of the 39% rare disorders with a gene association are linked only to their causal gene, which is often not realistically druggable with current drug modalities [5]. Hence, the ability to identify shared biology

among disorders with and without gene associations would help to accelerate the identification of novel drug targets in the rare space.

Representation learning of diseases and phenotypes, to the best of our knowledge, has only been approached based on Natural Language Processing (NLP) and Knowledge Graphs (KG). The NLP-based approaches are susceptible to corpus selection [21–23], hindering biological interpretations and insights. On the other hand, KGs combined with geometric deep learning [24] are proposed as an omni-solution for representation learning in biomedicine and healthcare [25, 26]. Yet, we still lack a foundation model for geometric learning in biomedicine and healthcare [27, 28], partially because KGs combine noise and inductive biases from multiple modalities in the training data, which the Graph Neural Networks (GNN) such as heterogeneous graph attention networks [29] are unable to disentangle. This discourages the model from learning the mechanisms underlying the pathogenesis of diseases, and ultimately learning the generative processes behind the formation of disease-centric biological networks [30–32]. Notably, TxGNN, a state-of-the-art disease-drug association prediction model [33], learns disease-drug associations using a zero-shot learning approach on biologically-grounded meta-pathways linking drugs to diseases. However, zero-shot learning on a dataset derived from clinical trials, which have a known bias toward targetable proteins [34], creates a disease representation that ultimately fails to capture the biological mechanisms of disease pathophysiology. These pitfalls in the state-of-the-art approaches hinder the development of a foundation model for diseases, which can be used for multiple downstream tasks on diseases including studying disease association learning and disease clustering for investment decisions.

A disease foundation model may instead be achievable via tasks that explicitly demand generalization [35,36], knowledge transfer [37], and interpretability [38]. Knowledge repurposing across rare and common diseases meets these criteria. From a machine learning perspective, rare diseases require operating in the data-scarce regime as illustrated in Figure 1D. The data-scarce regime encourages enhancing the generalizability and transferability of models by conducting stringent performance evaluation scenarios such as inductive tests [8,39,40] and transfer learning setting [36, 37], building confidence that the model can overcome the lack of data through the ability to learn the generative-mechanisms and biology behind the formation of biological networks. The lack of data and strict inductive tests reduce the risk of artificially inflating a model's test performance, an issue that is common in the data-rich regime and hinders the generalizability of many state-of-the-art models because of biases such as shortcut learning and class imbalance [39, 41, 42].

The contributions of this work are summarized below:

- We develop Dis2Vec, a continuous and stable representation space of diseases that captures the underlying biological mechanisms for both rare and common diseases, enabling knowledge repurposing between them.
- We show the effectiveness of Dis2Vec in multiple downstream tasks on diseases including disease-association prediction and disease clustering, paving the path towards building a foundation model for studying diseases.
- For a disease-association prediction task, we trained on strong genetic evidence and tested on investment decisions made in clinical trials (drug targets), bolstering the transferability aspect of the proposed model.
- We show that Dis2Vec enables clustering diseases based on shared pathways and mechanisms, hence addressing a major cause of inefficiencies in drug discovery, i.e., the tendency to define diseases in isolation, causing significant inefficiencies in rare disease research (refer to Figure 1C).

Results

Disease to Vector

Genetic associations are still undiscovered for more than 39% of rare diseases (See Methods and Figure S2A). Similarly, in Orphanet 38% of 6,353 rare diseases do not have any curated symptoms (see Methods). Hence, to learn a representation of disease mechanism of actions, we need to capture the highly non-linear dependencies between Biological Processes and symptoms, and combine them with experimentally validated disease-gene relationships and the established knowledge of disease symptoms. In order to achieve this, we created Dis2Vec representation learning pipeline as illustrated in Figure 2, enabling to combine and repurpose disease-specific biological knowledge to learn disease representations. The main innovation of Dis2Vec pipeline is the method to build the OntoDist (Ontology Distance) network by minimizing the LCA distances in Gene Ontology (GO) and Human Phenotype Ontology (HPO) for each disease pair (see Methods and Figure S4). KGs rely on simple linear direct association between biological processes and symptoms, which we argue that is not sufficient for disease associations learning. Yet, strategic data engineering combined with the linear direct association approach could be powerful for link prediction for disease-BiologicalProcesses and disease-symptoms (Section S5).

Next, to obtain disease representations any graph embedding function capable of aggregating both local (breadth first) and far neighborhood (depth first) structures can be used. We used

Node2Vec [43] on OntoDist network to obtain Dis2Vec embeddings, enabling us to learn the higher-order mechanistic dependencies between diseases through both depth-first and breadth-first strategies for learning higher-order disease neighborhoods. Lastly, Figure 2B demonstrate our strategy to benchmark Dis2Vec embeddings based on data modalities independent of the inputs we used to for Dis2Vec representation learning.

Disease Association Benchmark

We build a benchmark dataset for disease association which captures the underlying biological processes that are conserved across diseases and provides a transfer learning test dataset for evaluating disease association prediction models. We provide both positive and negative examples for our benchmark. In the existing literature, the established proven effectiveness of drug repurposing allows us to obtain positive controls for disease association learning [44]. Drug-target information is suitable for benchmarking of predictive models, but not suitable for training models due to their highly biased representation in biological pathways. Only 10% of druggable human proteins are targeted in trials [34], capturing a very small fraction of human biology that is not ideal for learning disease mechanisms. Yet, clinical trials offer a picture of investment decisions in drug development. We utilize a drug-disease bipartite network to create the Investment-Decision network as illustrated in Figure 3A, serving as the positive control in our benchmark. Similarly, we require negative controls to quantify the rate of false negatives for disease association learning. Negative controls are obtained based on Orphanet Rare Disease Ontology (ORDO) [45] by maximizing the Lowest Common Ancestor (LCA) distances and creating the ORDO-Disparity network as illustrated in Figure 3B. The known knowledge of causal genes on rare diseases provides a surgical insight into human biology that can be learned with geometric machine learning, and benchmark transfer learning to our Investment-Decision and ORDO-Disparity networks. Figure 3C-D demonstrates that the euclidean distance of Dis2Vec embeddings distinguish the positive and negative transfer learning benchmarks, indicating Dis2Vec offers rich disease representations.

Downstream Tasks for a Foundation Model on Diseases

We propose two downstream tasks for Dis2Vec that set the expectations towards a foundation model for disease association learning. These tasks require the coherent representation of diseases at the level of disease-disease dyads (Task 1), as well as in the higher-order structure of disease-disease relationships into mechanistically distinct clusters (Task 2). A foundation model for disease association learning should both be able to perform these tasks:

- **[Downstream Task 1 - Link Prediction] Disease Association Prediction**: We currently lack a foundation model for disease association learning [27,28]. We demonstrate that providing stable pre-trained disease representations (Dis2Vec) combined with a link prediction decoder trained on the OntoDist network is capable of predicting the drug-target investment decisions and disease mechanism of actions in a transfer learning setting, paving the path towards a foundation model for disease association learning.

- **[Downstream Task 2 - Node Labeling] Unsupervised Clustering of Diseases in Dis2Vec Space**: A major inefficiency in research on common diseases and especially rare diseases is caused because of compartmentalizing diseases. We show that instead of treating diseases individually, considering clusters of diseases that share mechanisms of action is a key to improve the efficiency of translational sciences, particularly for rare diseases (see Figure 1C-D).

In what follows, we will assess the performance and biological explainability of Dis2Vec on these tasks and thus its suitability as a foundation model for disease association learning.

Dis2Vec Demonstrates High Performance on Disease-Disease Link Prediction

We first assess whether the disease representations learned by Dis2Vec is able to accurately predict whether a pair of diseases will be evaluated for response to the same drug (Investment Decision) or whether they would fall into our class of ORDO ontological negatives. We constructed this task as follows. We have a disease association graph G. V is the set of disease nodes. The graph has 4 different types of edges:

- E_1: Edges based on Gene Ontology (G), derived based on minimizing LCA (Figures S3 and S4).

- E_2: Edges derived based on (Human Phenotype Ontology) HPO.

- E_3: Edges from investment decisions.

- E_4: Non-edges based on ORDO disparity.

$G_{OntoDist} = (V, crank(E_1, E_2) - E_3)$ forms the OntoDist network (Methods), which we use for training a disease association prediction decoder on Dis2Vec embeddings. We used the *crank* [46] to combine the information from two modalities (GO and HPO) into a comprehensive score (Methods). Note that we removed E_3 from $G_{OntoDist}$ as the purpose of investment decision links is benchmarking. Next, we construct a transfer learning task on disease association prediction

(link prediction). Thus, the train and validation of our model is based on the disease graph $G_{train} = G_{OntoDist}$. We use for transfer learning test the graph $G_{test} = G_{ID} = (V, E_3)$. For negative sampling during training on G_{train}, we use random negative sampling from G_{train} for training the decoder. On the other hand, we use the non-edges E_4 from $G_{ORDO} = (V, E_4)$ for calculating the transfer link prediction performance serving as negative control for investment decisions.

Figure 4A presents the model architecture for disease association learning under the following transfer learning scenario: (a) obtain Dis2Vec space using experimentally validated human genetic links and symptoms, and (b) test if Dis2Vec space is able transfer learning and predict Investment-Decision and ORDO-Disparity links. The model presents a strong predictive power with AUROC = 0.89 and AUPRC = 0.84, demonstrating the ability of Dis2Vec to transfer learning, capturing the biological reasoning behind the formation of Investment-Decision and ORDO-Disparity links. Moreover, the predicted association probabilities are well separated between the positive and negative controls (Figure 4B). For instance, Ravulizumab is an FDA-approved drug for both aHUS (Atypical hemolytic uremic syndrome) and PNH (Paroxysmal nocturnal hemoglobinuria) [47]. Dis2Vec confidently predicts an investment decision link between aHUS and PNH, despite the fact that these disorders do not share causal genes and there is no link between aHUS and PNH in OntoDist network. Interestingly, in many cases, drugs do not target the causal genes, indicating that Dis2Vec is able to learn the biological pathways and disease mechanisms shared between individual diseases, i.e., capturing the biological rationale used by pharmaceutical companies to invest in drug repurposing.

We next examined the ability of Dis2Vec to capture common-to-rare repurposing decisions. Rheumatoid arthritis (RA) represents a common disorder known to impinge on the same biological pathways and share similar symptomology with rare disorders, such as Idiopathic Inflammatory Myopathies (IIM, ORPHA:98482) [48] and Overlapping Connective Tissue Diseases (ORPHA:251312) [49]. Dis2Vec predicts strong association probabilities between RA and most disorders belonging to these groups. For example, Antisynthetase Syndrome (association score 0.95) is an IIM associated with auto-antibodies reactive to aminoacyl-transfer RNA synthetases. This disorder presents with joint manifestations that can, in some cases, be considered a rheumatoid arthritis overlap syndrome [50], and both disorders share genetic risk associated with specific haplotypes in the HLA region [51]. Most other IIMs and Overlap disorders also showed strong association scores for the disorders, including Polymyositis (0.94), Dermatomyositis (0.90), Mixed Connective Tissue Disease (0.94) and Overlap Myositis (0.89). It is interesting to note that one IBM, Inclusion Body Myositis, has a strikingly lower association score (0.34)

than the others, whereas some less well-known autoimmune disorders, such as Autoerythrocyte Sensitization Syndrome, showed higher association probability than all IIMs (0.97). Such cases may present interesting directions for research on shared disease mechanisms, as well as suggest novel repurposing opportunities.

While novel associations, such as that between RA and Autoerythrocyte Sensitization Syndrome, suggest potential new repurposing opportunities, we also noted cases where the model failed to support repurposing efforts already enacted in clinical trials. As expected, therapeutics that do not target a biological pathway or disease phenotype are not identified by Dis2Vec. Only 21/55 indications for Ataluren which enables read-through of early stop codons created by nonsense mutations [52], are identified because those indications share a disease mechanism only. In contrast, mechanistically conserved therapies like investment decisions for setmelanotide (the MC4R peptide) are strongly supported.

Dis2Vec supports unsupervised disease clustering that is supported by biological knowledge

We next asked whether the disease representation of Dis2Vec enables a mechanistically-supported unsupervised clustering of diseases (Downstream Task 2 – Node Labeling). Unsupervised clustering was performed on the 128-dimensional vectors associated with each disease using a Gaussian Mixture Model (see Methods). Within a two-dimensional UMAP of the clusters (Figure 5A), we observe some well-delineated islands that is dominated by diseases affecting specific anatomical structures. Cluster 14 and 25 forms an island of disorders with prominent cardiovascular manifestations, and Cluster 39 forms an island of disorders with dental presentations. The predominance of specific anatomies in these clusters may reflect clustering that is dominated by HPO-proximity, as anatomical features are often implicitly in phenotype terms. However, we also find clusters that appear to be defined more by biological process than anatomy. Cluster 0 comprises disorders featuring dysregulation of the immune system, including deficiencies in specific immune cell types (e.g., Autosomal agammaglobulinemia, Cyclic neutropenia), abnormalities of immune cell function (e.g., Leukocyte adhesion syndrome) and combined immunodeficiencies (e.g., T-B+ severe combined immunodeficiency). In some cases these clustering show remarkable insight, such as the grouping of cancers that are associated with autoimmunity (e.g., Thymoma [53] Cluster 31) with paraneoplastic autoimmune syndromes (e.g., Myasthenia gravis, Cluster 38), although in the cluster 38 scatter in the UMAP space we observe that Thymoma and Myasthenia gravis remain adjacent. Finally, some highly related clusters appear as close neighbors in the UMAP, showing higher order structure between clusters

of diseases. For example, the immunodeficiency-dominated neighbors the cluster dominated by immune-system cancers and autoimmune diseases (Cluster 0 and Cluster 31), which in turn is neighbor to a cluster of disorders with autoimmune destruction of blood cells, like thromocytopenia and hemolytic anemias (Clusters 5 and 47). Moreover, Liposomal Acid Lipase Deficiency (LAL-D, ORPHA:275761) and its subtypes, Cholesteryl Ester Storage Disease (ORPHA:75234) and Wolman Disease (ORPHA:75233) are located in cluster 40. These disorders and all linked solely to the gene LIPA. Thus, the unsupervised clustering exhibits structure with clear correspondence to biological mechanisms, which may favor phenotype-based or biological process-based features in specific instances. These disorders are differentially annotated in Orphanet, but are nonetheless clustered together by Dis2Vec.

Importantly, rare and common disorders are also co-located within the UMAP in a mechanistically-interpretable way (Figure 5B). Common neurological disorders like Parkinson Disease (PD; MONDO:0005180) and Essential tremor (EFO:0003108) are adjacent in the UMAP, and are surrounded by rare forms of dystonia and hereditary forms of PD in Cluster 37. Common cancers such as Melanoma (EFO:0000756) and Prostate carcinoma (EFO:0001663) co-locate with the with rare, hereditary cancers in Cluster 44, e.g., Li-Fraumeni syndrome (ORPHA:524), Hereditary Mixed Polyposis Syndrome (ORPHA:157794). Dis2Vec's enabling of information from common-to-rare disorders (and vice versa) through mechanistically-interpretable disease associations is thus supported by structure of unsupervised clustering as illustrated in Figure 5B.

To further assess the quality of mechanistic disease representations produced by Dis2Vec, we also assessed hard negative examples: disease pairs that are superficially similar, but mechanistically distinct. NF1 and NF2 are both Neurofibromatosis disorders, but are related to distinct biological pathways and manifest in different phenotypes. NF1 is caused by loss of the NF1 gene, resulting in upregulated Ras-signaling and hyperproliferation of cells in several different tissues [54]. NF2 is caused by the loss of the Schwann cell protein Merlin that coordinates growth factor signaling and cell adhesion, resulting in the development of vestibular Schwannomas [55]. We find that these disorders are appropriately separated in the Dis2Vec UMAP. NF1 clusters together with other cancer-predisposing syndromes of multiple tissue overgrowth linked to RAS mutations, including Linear Nevus Sebaceous Syndrome (LNSS; ORPHA:2612) and specific subtypes of NF1 like 17q11 microdeletion syndrome (ORPHA:97685) in Cluster 6. NF2 clusters together with other neurological cancers, such as Medulloblastoma (ORPHA:616) and Hemangioblastoma (ORPHA:252054) in Cluster 2. From this case, it can be seen that Dis2Vec correctly identifies the two neurofibromatosis types as distinct.

However, deeper exploration of these same examples shows ways in which the Dis2Vec representation is inconsistent with our understanding of disease. While NF1 and LNSS cluster together, they are separated from other related RASopathies [56] in Cluster 32, like Noonan Syndrome, Legius Syndrome, and Costello Syndrome, for reasons that are unclear. Even some subtypes of NF1 are located in different clusters from NF1, such as "Neurofibromatosis type 1 due to NF1 mutation or intragenic deletion" (ORPHA:363700), also located in Cluster 32. As both NF1 and its subtypes are linked solely to the same single gene (NF1), these differences in clustering must be due to the relatively modest differences in the phenotypic annotations for these disorders. Thus, while Dis2Vec features a disease representation that is globally coherent in terms of currently understood disease mechanisms, the placement of individual disorders can exhibit sensitivity to the exact annotation.

DISCUSSION

Dis2Vec is a continuous and stable disease representation that supports biologically meaningful disease association learning and information repurposing. The Dis2Vec representation was constructed using expert-curated evidence on disease-gene and disease-phenotype associations and its performance was assessed on two key downstream tasks: 1) predicting disease associations (e.g., to enable drug repurposing decisions), and 2) unsupervised clustering of diseases (e.g., to facilitate information repurposing). Dis2Vec demonstrated excellent performance on Task 1, showing effective transfer learning to real-world drug repurposing investment decisions. Furthermore, Dis2Vec showed biologically-meaningful unsupervised clustering of diseases in broad agreement with current knowledge, supporting its effectiveness on Task 2. Together, these results demonstrate the utility of the Dis2Vec representation in drug-repurposing and the facilitation of translational research. Moreover, the generalization of the Dis2Vec representation to these two key tasks supports the utility of this representation as a foundation model for common and rare diseases.

Downstream Task 1

Drug repurposing (our Task 1) is a critical function in the expensive and risk-laden process of pharmaceutical development [44]. We show that the Dis2Vec representation can be decoded to identify such opportunities with high precision and recall through pure transfer learning; that is, without any drug or repurposing data included in the development of the representation. The success of this transfer learning demonstrates that the Dis2Vec representation, built solely from expert knowledge of disease-gene and disease-phenotype associations, captures the essential

mechanistic information that is leveraged by pharmaceutical companies in making life cycle management investment decisions. The investment decisions linked to the complement pathway inhibitor Ravulizumab are notable examples of this success, as Dis2Vec association probability agreed with the investment in 30 of 36 cases. The only Ravulizumab investment decision to receive consistent low association scores from Dis2Vec, Amyotrophic Lateral Sclerosis, was also a notable Phase 3 failure (the reason for failure is unknown) [57]. We attribute this success in part to the coarse-grained disease associations to biological processes that are used to build the Dis2Vec representation, which enables generalization beyond specific genes to biological mechanisms. However, we note that this approach also has the disadvantage of limiting the ability to identify molecular-level repurposing opportunities, represented by drugs like Ataluren that address specific types of molecular lesions.

Downstream Task 2

Our Task 2, unsupervised clustering of diseases, enables broader knowledge transfer between disorders and is particularly important to translational efficiency in the rare disease space. We find that unsupervised clustering using the Dis2Vec disease representation reveals meaningful mechanistic relationships between diseases. Examples of this include the both the general co-clustering of disorders of similar pathogenesis, such as autoimmune disorders or dystonias, to more specific examples of cancers co-clustering with their associated paraneoplastic syndromes. These relationships extend to the degree that we see evidence for linear compositionality, in that the position of known overlap disorders correspond roughly to the vector barycenter of the disorders whose symptoms they share (Section S7). This mechanistic coherence obtains even when examining challenging examples, but does breakdown in some cases. Our findings suggest a mechanistically-grounded global structure of the Dis2Vec embedding space that will enable disease-knowledge repurposing: a key facilitator of translational research in rare disease.

Common to rare (and back again)

Importantly, we find that information repurposing extends from rare to common diseases. In the context of drug repurposing investment decisions, Dis2Vec captures key mechanistic similarities between common and rare disorders that can be used to identify repurposing opportunities. An example of this is Dis2Vec's uncovering of similarities between the common autoimmune disorder Rheumatoid Arthritis and closely related rare Idiopathic Inflammatory Myopathies (cluster 7). In the context of disease clustering, we show that common disorders co-cluster with related rare disorders. An example of this is the co-clustering of Parkinson Disease with related

Dystonias (including rare familiar forms of PD in Cluster 37). These observations demonstrate the utility of the Dis2Vec in common-to-rare drug and knowledge repurposing efforts.

Contrast with other studies

Notably, TxGNN, a state-of-the-art disease-drug association foundation model [33], learns disease-drug associations using a zero-shot learning approach on meta-pathways involving drugs, anatomies, pathways, cellular components, molecular functions, biological processes, phenotypes, and disease nodes. Hence, deriving disease similarities using TxGNN involves finding longer meta-paths through drug, cellular components, and anatomy (organ) nodes. However, drugs and clinical trials are biased toward targetable proteins resulting in the current drug-discovery winter where new targets are rarely introduced [34], unable to capture the similarity between diseases based on biological mechanisms. Furthermore, the zero-shot approach creates a disease representation amenable to drug-disease associations, but not resulting in universal disease representations capable of capturing the true mechanisms under disease pathophysiology which can be used in diverse downstream tasks on diseases. Modalities like anatomy (organs or tissues) and cellular components can predict biologically infeasible meta-paths, leading to poor disease-drug association predictions. For example, TxGNN provides biological explanations for drug repurposing on the basis of the liver being targeted by two drugs, there are over 10,000 distinct proteins in the liver involved in various pathways [58]. Hence, the underlying biological assumptions behind the predictions from TxGNN are often ambiguous and biologically infeasible. These pitfalls in the state-of-the-art approaches hinder the development of a foundation model for diseases, which can be used for multiple downstream tasks on diseases including studying disease association learning and disease clustering for investment decisions.

Limitations and Future directions

Dis2Vec is an important first step toward a foundation model for common and rare diseases, but many important advances remain for this vision to be fully realized. Firstly, it is important to demonstrate generalization beyond our first two downstream tasks. While these are indisputably important tasks to rare disease research and drug development, additional knowledge repurposing-dependent tasks remain to be tacked. It remains possible that these will expose weaknesses of the current representation that need to be addressed. For example, patient HPO terms and sequencing be mapped to the Dis2Vec representation to uncover the most likely diagnosis. Such individual patient data could also be mapped to Dis2Vec to support potential lumping or splitting of patient populations. The performance of Dis2Vec on such tasks remains

to be evaluated. Second, the development of quantitative, robust benchmarks will be essential to judging the performance of Dis2Vec relative to other algorithms. The assessment of unsupervised clustering in particular is currently qualitative and non-comprehensive, making it a weak point in our evaluation. Finally, we find some evidence that the Dis2Vec representation itself may need improvement. Example of disorder subtypes that fail to co-cluster, and investment decisions that are missed by our model, warrant deeper investigation. These investigations may highlight general failure modes of the approach, such as that encountered with Ataluren, which may be remedied by enhancements to how the representation is constructed. Alternatively, some of these apparent discrepancies may yield new insights into the nature of disorders and their subtypes, reinforcing the utility of the current model.

METHODS

Data Preparations

Selection of Gene Associations with Strong Genetic Evidence

We used Orphanet [45] to obtain disease gene associations. Orphanet disease-gene relationship inventory is built based on clinical evidence, and organized into the following categories: causal genes (disease-causing germline and disease-causing somatic mutations), major susceptibility factor, modifying germline mutation, part of a fusion, have a role in the phenotype, candidate gene tested in clinical settings, and biomarker tested to monitor disease activity and/or patient outcome. Therefore, Orphanet enabled us to work to build OntoDist and GODS (GeneOntology-Disease-Symptom tripartite network, Section S5) networks based on clinically validated disease-gene relationships with strong genetic effects on diseases.

After merging disease subtypes in Orphanet, 46% (2937 from 6353) of diseases do not have any genetic associations. Moreover, we used HPO ontology without any filters for obtaining disease symptom associations, in Orphanet 38% (2444 from 6351) of diseases do not have any symptoms. Our aim is to design generalizable ML to tackle this incompleteness in disease-gene and disease-symptom associations.

OntoDist Network

We defined ontological similarities based on minimizing the LCA distances from ontology polyarchichal structures [59] (see Figure S3 and Figure S4). OntoDist is created by calculating LCA distances for each pair of diseases on two data modalities, GO Biological Processes and the HPO. We also enforce a threshold (0.2) to filter pair with high distance (Figure S5). Next,

we used crank [46] to combine scores from the two modalities and used the top 200,000 diseases pairs according to crank to build OntoDist. The rationale for using crank [46] tackling the issue of data incompleteness, as 46% (2937 from 6353) of diseases do no have any genetic associations in Orphanet, and 38% (2444 from 6351) of diseases do not have HPO associations. In multiple downstream applications, we showed that OntoDist is able to recover from the lack of HPO and GO associations by leveraging disease neighborhoods. The main hypothesis is that diseases that share biology and some symptoms may lead to the same molecular target if the underlying mechanisms of the diseases is the same. An example is the case of Chron Disease (CD) and Ulcerative Colitis (UC) which are two types of Inflammatory Bowel Diseases (IBD) [60]. Despite their very different pathological characteristics, they may present with common symptoms and share disease-gene associations [60].

Investment Decision Network as Positive Control

The investment decision network is created as described in Figure 3A. This network serves as our primary benchmark dataset for disease association learning since it recapitulates the drug repurposing investment decisions conducted in the pharmaceutical industry.

ORDO Disparity Network as Negative Control

To quantify the rate of false negatives, we introduce a new benchmark network for disease association learning. This disparity network is created based on maximizing the LCA distanced in the Orphanet Rare Disease Ontology network (Figure S6).

Transfer Learning Setup

The transfer learning is based on the different natures of OntoDist, Investment-Decision, and ORDO-Disparity networks. The disease association decoder is trained using Dis2Vec embeddings that are learned from the OntoDist network. Next, the transfer learning is evaluated by predicting Investment-Decision and ORDO-Disparity links (Figure 4). Since the network properties of the OntoDist, Investment-Decision, and ORDO-Disparity networks are significantly different (see Figure S8), the transfer learning task ensures the stability of our model under distribution shifts [61] and the ability to circumvent topological shortcuts [39, 62, 63].

Model Architecture for Disease Association Learning

Similar to the two-shot learning approach proposed by Chatterjee et al. c [64], we feed the Node2Vec [43] embeddings into a 3-layer multi-layer perception (MLP) [65, 66]. The decoder is

trained and validated in a transductive setting on the OntoDist network for the disease association prediction task and on the GODS network for the explanation task. For our experiments, we use Node2vec dimension of 128 since it provided the best performance in downstream tasks, learning rate 10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ (default Node2Vec parameters). In Node2vec, we use a walk length of 10 for $G_{OntoDist}$, the number of walks of 100, a window size of 10, $p = 1$, and $q = 1$.

Unsupervised Clustering of Diseases

We used the Gaussian Mixture model from sklearn directly on disease vectors (obtained form Node2Vec) to identify 50 clusters. The number of clusters is empirically obtained by assessing positive and negative controls such as LALD (Lysosomal acid lipase deficiency) and CESD (Cholesteryl ester storage disease) as a positive control, and NF1 (Neurofibromatosis type 1) and NF2 (Full NF2-related schwannomatosis) as a negative control. Next, we used UMAP to reduce dimension and visualize the clusters. We observe that the reduced dimensions by UMAP largely agrees with the clusters obtained directly from the 128 dimensions of Dis2Vec.

Data availability

All data used in this study is from public sources such as Orphanet, OpenTargets, Gene Ontology, and Human Phenotype Ontology. The engineered data from thee aforementioned sources is also available in the project GitHub as well as the disease vectors and clusters.

Code availability

The codes that support the findings of this study are openly available on our GitHub at https://github.com/alxndgb/Dis2Vec.

Acknowledgements

We thank the Bioinformatics and Data Science group in Alexion AstraZeneca Rare Disease for funding this project. We are grateful to AstraZeneca's Scientific Computing Platform (SCP) for providing the computing resources. We thank Gyan Srivastava, Associate Director of Bioinformatics and Data Science group in Alexion for fruitful discussions on rare diseases.

Author Contributions

B.R., W.R.M., A.C., and P.R. conceived and designed the project. B.R., W.R.M., and A.C. wrote the manuscript and conducted data curation and preparation. B.R. conducted the data engineering and integration to produce OntoDist, Investment Decision benchmark (positive controls), negative controls, conducted network-based modeling and analysis. W.R.M. performed in-depth evaluation of clusters and the investment decision benchmark. A.C. designed the machine learning architecture and tune the hyper-parameters. M.A. and W.D. provided biological feedback and guidance. W.D. helped with designing the negative control benchmark. S.L. offered clinical feedback on positive and negative controls. M.U., I.B., T.D., and G.D.A. interpreted the results, offered biological insights, and contributed with writing the manuscript. T.E.R. provided guidance on designing the machine-learning experiments and writing the manuscript.

References

[1] Genetic and rare diseases (GARD). https://rarediseases.info.nih.gov/ (2021).

[2] Elliott, E. J. & Zurynski, Y. A. Rare diseases are a 'common' problem for clinicians. *Australian family physician* **44**, 630–633 (2015).

[3] Rare diseases, common challenges. *Nature Genetics* **54**, 215–215 (2022). URL https://doi.org/10.1038/s41588-022-01037-8.

[4] Rare diseases: Although limited, available evidence suggests medical and other costs can be substantial. US government accountability office. https://www.gao.gov/products/gao-22-104235 (2021).

[5] Ferreira, C. R. The burden of rare diseases. *American journal of medical genetics Part A* **179**, 885–892 (2019).

[6] Ghiassian, S. D., Menche, J. & Barabási, A.-L. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS computational biology* **11**, e1004120 (2015).

[7] Buphamalai, P., Kokotovic, T., Nagy, V. & Menche, J. Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nature communications* **12**, 1–15 (2021).

[8] Chatterjee, A. *et al.* Topology-driven negative sampling enhances generalizability in protein-protein interaction prediction. *In Prep* (2024).

[9] Tabor, H. K. & Goldenberg, A. What precision medicine can learn from rare genetic disease research and translation. *AMA Journal of Ethics* **20**, 834–840 (2018).

[10] Brinkman, R. R., Dubé, M.-P., Rouleau, G. A., Orr, A. C. & Samuels, M. E. Human monogenic disorders - a source of novel drug targets. *Nature Reviews. Genetics* **7**, 249–260 (2006).

[11] Lalakgas, P. N. & Melamed, R. D. Mendelian disease comorbidity as a resource for novel therapeutic targets (2023). URL https://www.biorxiv.org/content/10.1101/2023.07.23.550190v1. Pages: 2023.07.23.550190 Section: New Results.

[12] Hardy, J. The discovery of Alzheimer-causing mutations in the APP gene and the formulation of the "amyloid cascade hypothesis". *The FEBS journal* **284**, 1040–1044 (2017).

[13] Goldstein, J. L. & Brown, M. S. History of Discovery: The LDL Receptor. *Arteriosclerosis, thrombosis, and vascular biology* **29**, 431–438 (2009). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2740366/.

[14] Guthrie, J. *et al.* Autocore: A network-based definition of the core module of human autoimmunity and autoinflammation. *Science Advances* **9**, eadg6375 (2023). URL https://www.science.org/doi/abs/10.1126/sciadv.adg6375. https://www.science.org/doi/pdf/10.1126/sciadv.adg6375.

[15] Halley, M. C., Smith, H. S., Ashley, E. A., Goldenberg, A. J. & Tabor, H. K. A call for an integrated approach to improve efficiency, equity and sustainability in rare disease research in the united states. *Nature genetics* **54**, 219–222 (2022).

[16] Denton, N. *et al.* Data silos are undermining drug development and failing rare disease patients. *Orphanet Journal of Rare Diseases* **16**, 1–4 (2021).

[17] Groft, S. C. & Gopal-Srivastava, R. A model for collaborative clinical research in rare diseases: experience from the rare disease clinical research network program. (2013).

[18] Boulanger, V., Schlemmer, M., Rossov, S., Seebald, A. & Gavin, P. Establishing patient registries for rare diseases: rationale and challenges. *Pharmaceutical Medicine* **34**, 185–190 (2020).

[19] Pariser, A. R. & Gahl, W. A. Important role of translational science in rare disease innovation, discovery, and drug development. *Journal of general internal medicine* **29**, 804–807 (2014).

[20] Zanello, G. *et al.* Targeting shared molecular etiologies to accelerate drug development for rare diseases. *EMBO Molecular Medicine* **15**, e17159 (2023).

[21] Ghosh, S., Chakraborty, P., Cohn, E., Brownstein, J. S. & Ramakrishnan, N. Characterizing diseases from unstructured text: A vocabulary driven word2vec approach. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, 1129–1138 (2016).

[22] Miftahutdinov, Z., Kadurin, A., Kudrin, R. & Tutubalina, E. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics* **37**, 3856–3864 (2021).

[23] Daniali, M. *et al.* Enriching representation learning using 53 million patient notes through human phenotype ontology embedding. *Artificial intelligence in medicine* **139**, 102523 (2023).

[24] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine* **34**, 18–42 (2017).

[25] Li, M. M., Huang, K. & Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering* **6**, 1353–1369 (2022).

[26] Zhao, B.-W., Hu, L., You, Z.-H., Wang, L. & Su, X.-R. Hingrl: predicting drug–disease associations with graph representation learning on heterogeneous information networks. *Briefings in bioinformatics* **23**, bbab515 (2022).

[27] Mao, H. *et al.* Graph foundation models. *arXiv preprint arXiv:2402.02216* (2024).

[28] Huang, K. *et al.* Artificial intelligence foundation for therapeutic science. *Nature chemical biology* **18**, 1033–1036 (2022).

[29] Wang, X. *et al.* Heterogeneous graph attention network. In *The world wide web conference*, 2022–2032 (2019).

[30] Bonner, S., Kirik, U., Engkvist, O., Tang, J. & Barrett, I. P. Implications of topological imbalance for representation learning on biomedical knowledge graphs. *Briefings in Bioinformatics* **23**, bbac279 (2022).

[31] Bonner, S. *et al.* Understanding the performance of knowledge graph embeddings in drug discovery. *Artificial Intelligence in the Life Sciences* **2**, 100036 (2022).

[32] Bonner, S. *et al.* A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics* **23**, bbac404 (2022).

[33] Huang, K. *et al.* A foundation model for clinician-centered drug repurposing. *Nature Medicine* 1–13 (2024).

[34] Vasan, K., Gysi, D. M. & Barabási, A.-L. The clinical trials puzzle: How network effects limit drug discovery. *Iscience* **26** (2023).

[35] Tang, H. & Liu, Y. Towards understanding generalization of graph neural networks. In Krause, A. *et al.* (eds.) *Proceedings of the 40th International Conference on Machine Learn-*

ing, vol. 202 of *Proceedings of Machine Learning Research*, 33674–33719 (PMLR, 2023). URL https://proceedings.mlr.press/v202/tang23f.html.

[36] Chuang, C.-Y., Torralba, A. & Jegelka, S. Estimating generalization under distribution shifts via domain-invariant representations. In III, H. D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 1984–1994 (PMLR, 2020). URL https://proceedings.mlr.press/v119/chuang20a.html.

[37] Zhu, Q. *et al.* Transfer learning of graph neural networks with ego-graph information maximization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, 1766–1779 (Curran Associates, Inc., 2021). URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0dd6049f5fa537d41753be6d37859430-Paper.pdf.

[38] Miao, S., Luo, Y., Liu, M. & Li, P. Interpretable geometric deep learning via learnable randomness injection. In *The Eleventh International Conference on Learning Representations* (2023). URL https://openreview.net/forum?id=6u7mf9s2A9.

[39] Chatterjee, A. *et al.* Improving the generalizability of protein-ligand binding predictions with AI-bind. *Nature Communications* **14** (2023). URL https://doi.org/10.1038/s41467-023-37572-z.

[40] Chatterjee, A., Walters, R., Menichetti, G. & Eliassi-Rad, T. Disentangling node attributes from graph topology for improved generalizability in link prediction (2023). **In Preparations** for Submission at the Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

[41] Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020). URL https://doi.org/10.1038/s42256-020-00257-z.

[42] Puli, A., Zhang, L., Wald, Y. & Ranganath, R. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. *arXiv preprint arXiv:2308.12553* (2023).

[43] Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864 (2016).

[44] Kulkarni, V. S., Alagarsamy, V., Solomon, V. R., Jose, P. A. & Murugesan, S. Drug Repurposing: An Effective Tool in Modern Drug Discovery. *Russian Journal of Bioorganic*

Chemistry **49**, 157–166 (2023). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9945820/.

[45] Orphanet rare disease ontology. https://www.orpha.net/.

[46] Zitnik, M., Sosič, R. & Leskovec, J. Prioritizing network communities. *Nature communications* **9**, 2544 (2018).

[47] FDA approved drugs. https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm.

[48] Klein, M., Mann, H. & Vencovský, J. Arthritis in Idiopathic Inflammatory Myopathies. *Current Rheumatology Reports* **21**, 70 (2019).

[49] Maddison, P. J. Overlap syndromes and mixed connective tissue disease. *Current Opinion in Rheumatology* **3**, 995 (1991). URL https://journals.lww.com/co-rheumatology/abstract/1991/12000/overlap_syndromes_and_mixed_connective_tissue.16.aspx.

[50] Weerasinghe, W. S. & Jayasinghe, C. Overlapping rheumatoid arthritis and antisynthetase syndrome with secondary Sjögren's syndrome: a case report and review of the literature. *Journal of Medical Case Reports* **16**, 132 (2022). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8978378/.

[51] Lamb, J. A. The Genetics of Autoimmune Myositis. *Frontiers in Immunology* **13** (2022). URL https://www.frontiersin.org/articles/10.3389/fimmu.2022.886290.

[52] Roy, B. *et al.* Ataluren stimulates ribosomal selection of near-cognate tRNAs to promote nonsense suppression. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 12508–12513 (2016).

[53] Blum, T. G., Misch, D., Kollmeier, J., Thiel, S. & Bauer, T. T. Autoimmune disorders and paraneoplastic syndromes in thymoma. *Journal of Thoracic Disease* **12**, 7571–7590 (2020). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7797875/.

[54] Friedman, J. M. Neurofibromatosis 1. In Adam, M. P. *et al.* (eds.) *GeneReviews®* (University of Washington, Seattle, Seattle (WA), 1993). URL http://www.ncbi.nlm.nih.gov/books/NBK1109/.

[55] Evans, D. G. NF2-Related Schwannomatosis. In Adam, M. P. *et al.* (eds.) *GeneReviews®* (University of Washington, Seattle, Seattle (WA), 1993). URL http://www.ncbi.nlm.nih.gov/books/NBK1201/.

[56] Rauen, K. A. Defining RASopathy. *Disease Models & Mechanisms* **15** (2022). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8821523/. Publisher: Company of Biologists.

[57] Genge, A. *et al.* Efficacy and Safety of Ravulizumab, a Complement C5 Inhibitor, in Adults With Amyotrophic Lateral Sclerosis: A Randomized Clinical Trial. *JAMA Neurology* **80**, 1089–1097 (2023). URL https://doi.org/10.1001/jamaneurol.2023.2851.

[58] Niu, L. *et al.* Dynamic human liver proteome atlas reveals functional insights into disease pathways. *Molecular Systems Biology* **18**, e10947 (2022). URL https://www.embopress.org/doi/abs/10.15252/msb.202210947. https://www.embopress.org/doi/pdf/10.15252/msb.202210947.

[59] Yang, J., Dong, C., Duan, H., Shu, Q. & Li, H. Rdmap: a map for exploring rare diseases. *Orphanet journal of rare diseases* **16**, 1–11 (2021).

[60] Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms–disease network. *Nature communications* **5**, 4212 (2014).

[61] Chuang, C.-Y., Torralba, A. & Jegelka, S. Estimating generalization under distribution shifts via domain-invariant representations (2020). 2007.03511.

[62] Chatterjee, A., Walters, R., Menichetti, G. & Eliassi-Rad, T. Disentangling node attributes from graph topology for improved generalizability in link prediction (2023). 2307.08877.

[63] Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airoldi, E. M. & Clauset, A. Stacking models for nearly optimal link prediction in complex networks. *Proceedings of the National Academy of Sciences* **117**, 23393–23400 (2020). URL http://dx.doi.org/10.1073/pnas.1914950117.

[64] Chatterjee, A., Walters, R., Menichetti, G. & Eliassi-Rad, T. Inductive link prediction in static and temporal graphs for isolated nodes. In *Temporal Graph Learning Workshop @ NeurIPS 2023* (2023). URL https://openreview.net/forum?id=DRrSYKNhD1.

[65] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika* **71**, 599–607 (1986).

[66] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65**, 386 (1958).

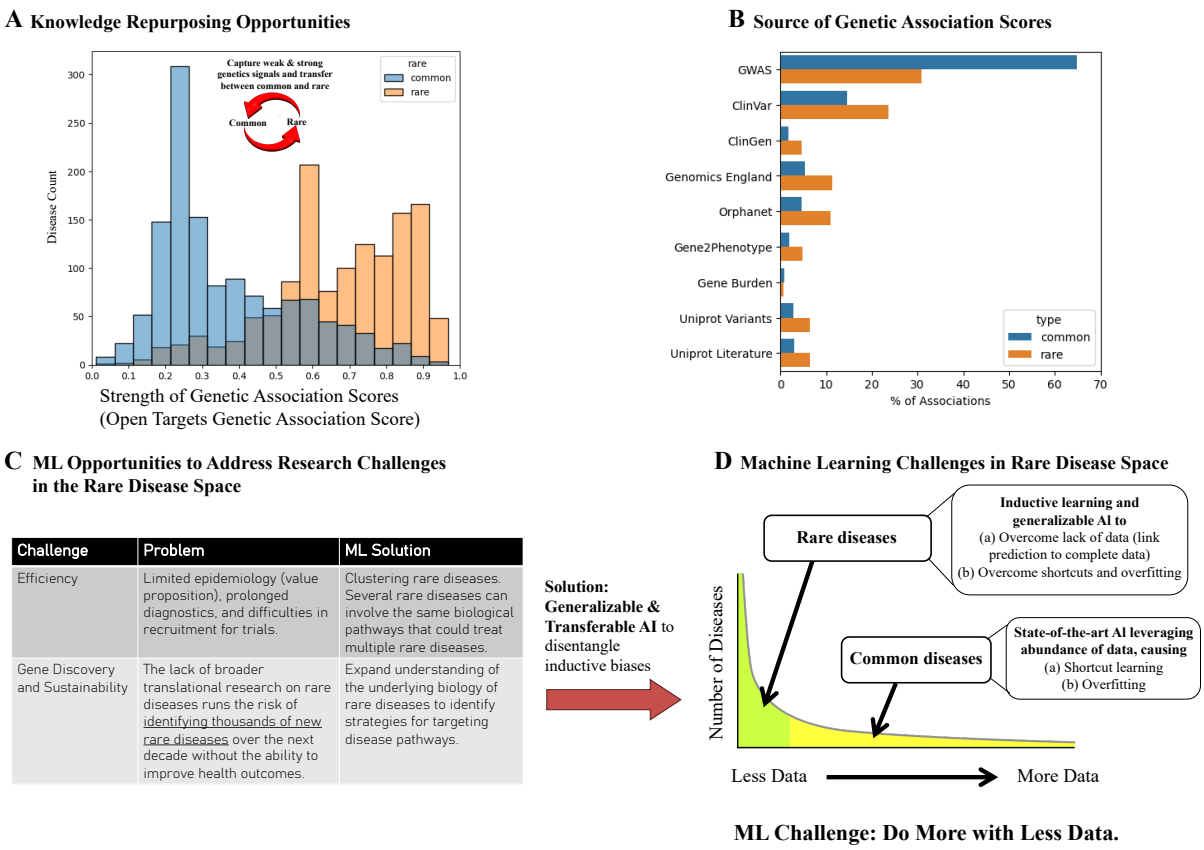

Figure 1

Figure 1: **Repurposing Opportunities for ML to address Research Challenges in the Rare and Common Diseases. (A-B)** Both rare and common diseases are driven by Human biology. Rare diseases have a stronger genetic association score (obtained from OpenTargets that combines various databases), driven by strong effect genetic variants that are casual for diseases. In contrast, common diseases have a weaker genetic association score, driven by GWAS which has limited power to uncover causal links. Hence there is a knowledge repurposing opportunity to capture weak and strong genetics signals and transfer between common and rare diseases. **(C-D)** Generalizable and transferable AI offers opportunities to address efficiency and discovery challenges in the data scarce regime.

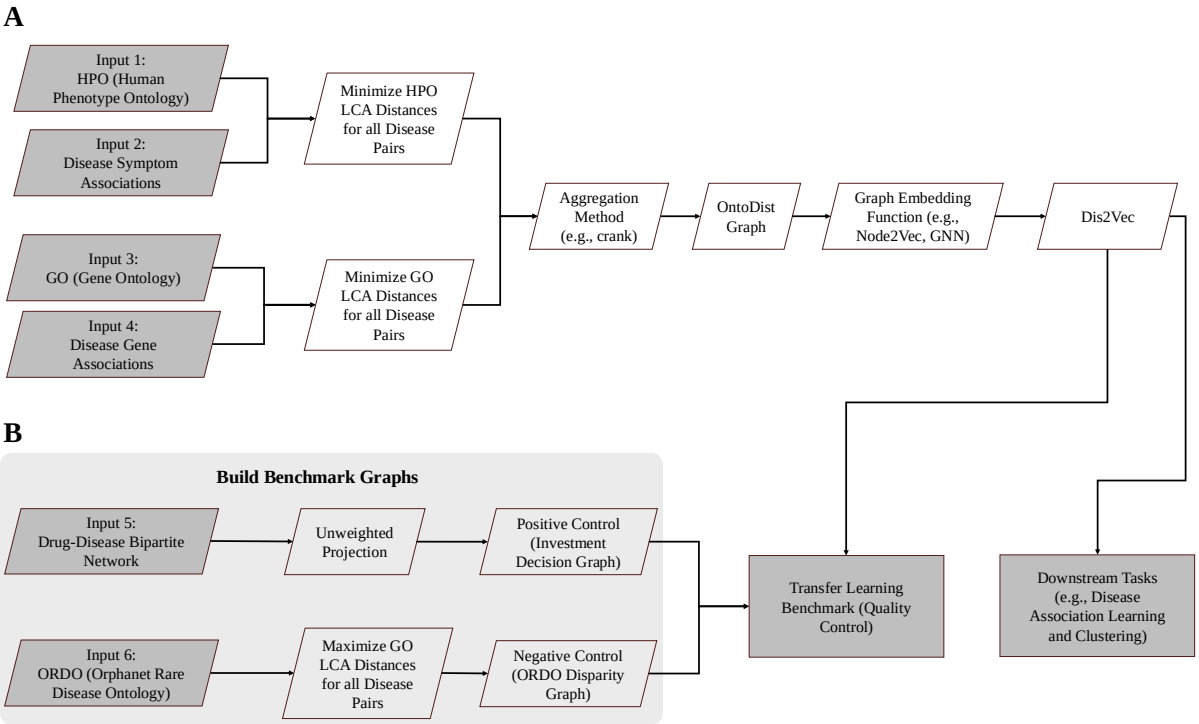

Figure 2: **Dis2Vec Pipeline. (A)** Dis2Vec is a multi-modal representation learning pipeline, enabling to build a disease-disease association network, namely OntoDist, by combining biological knowledge from various sources. We incorporate non-linear knowledge from ontologies by minimizing the LCA distances for all diseases pairs (see Methods). Next, we used *crank* [46] for combining and prioritizing LCA distances. Lastly, any graph embedding function capable of aggregating both local (breadth first) and far neighborhoods (depth first) can be used on OntoDist network to obtain Dis2Vec (we used Node2Vec). **(B)** We construct positive and negative controls utilizing independent data modalities not used in learning Dis2Vec embeddings. For positive controls, we utilize drug repurposing investment decisions from clinical trials, obtaining disease-disease relationships where two diseases entered trials with the same drug. For negative controls, we used ORDO by maximizing LCA distances to identify disease pairs that are significantly different in terms of underlying biology.

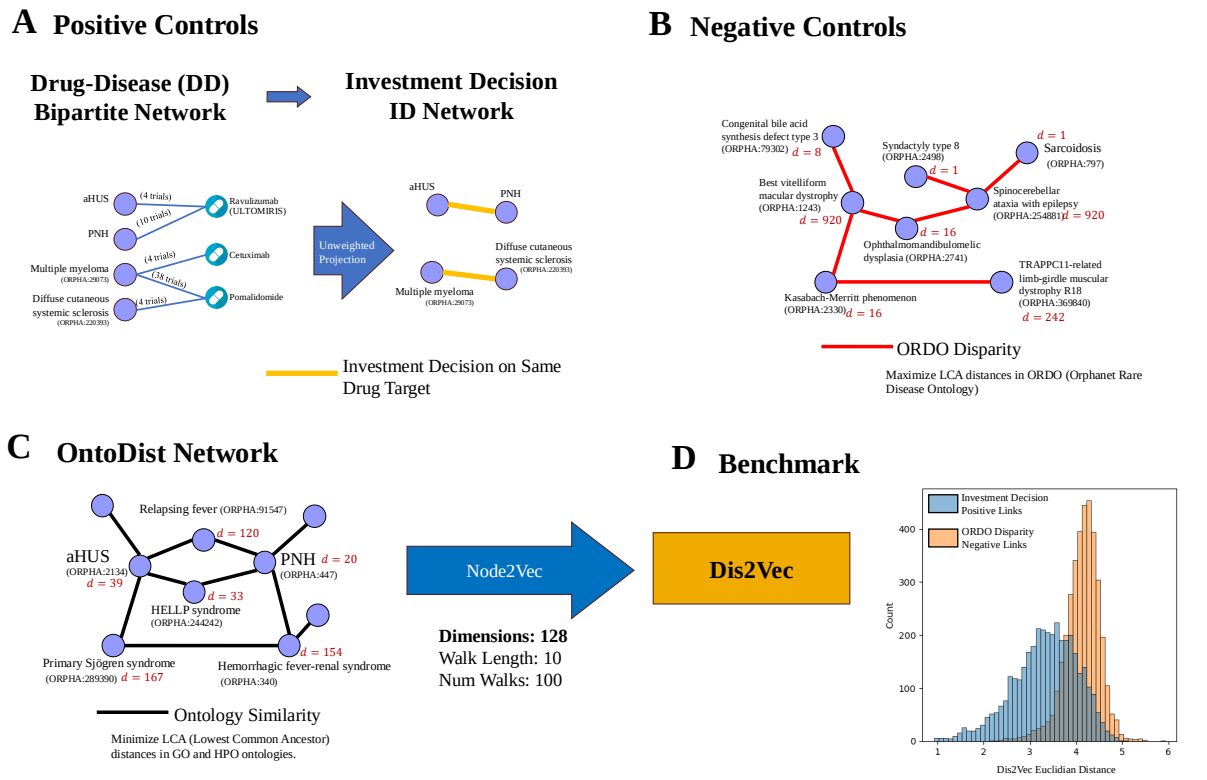

Figure 3: **Disease Representation Learning (Disease to Vector) and Transfer Learning Benchmark. (A-B)** We establish transfer learning benchmark by utilizing drug repurposing investment decisions from clinical trials as positive controls. For negative controls, we created transfer learning benchmark by identifying rare disease with maximum LCA distances in Orphanet Rare Disease Ontology. **(C)** We created OntoDist network by combining non-linear dependencies in ontologies (GO and HPO) with disease genetics and symptoms associations (Methods). OntoDist is an unweighted network where an edge represents two disease share biology according to the knowledge enriched in GO and HPO. **(D)** We observe that indeed OntoDist's structure captures higher order biological relationships between diseases by utilizing Node2Vec to learn disease representation. Remarkably, the euclidean distance of Dis2Vec embeddings distinguish the positive and negative transfer learning benchmarks, indicating Dis2Vec offers rich disease representations that can be utilized for downstream ML tasks such as disease association learning and clustering.

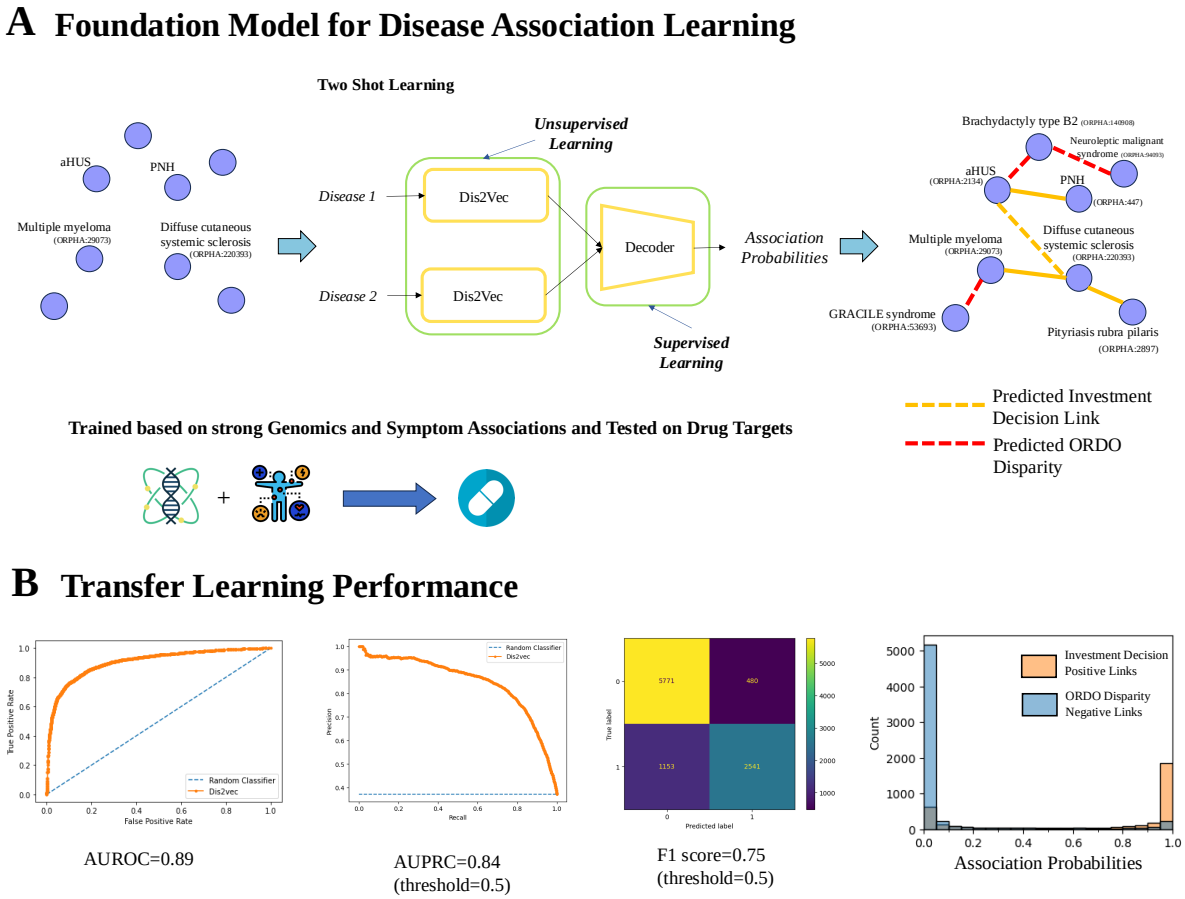

Figure 4: **Foundation Model for Disease Association Learning.** **(A)** The first downstream task of Dis2Vec is disease-disease representation learning, paving the way toward developing a foundation model for disease association learning. **(B)** Dis2Vec's disease representations are learned from Human genetic evidence and observed symptoms. We setup a transfer learning scenario on drug targets (clinical trials) that are often do not target the casual genetic links. Dis2Vec present a remarkable transfer learning scenario, being able to recapitulate the majority of drug repurposing investment decisions made in clinical trials.

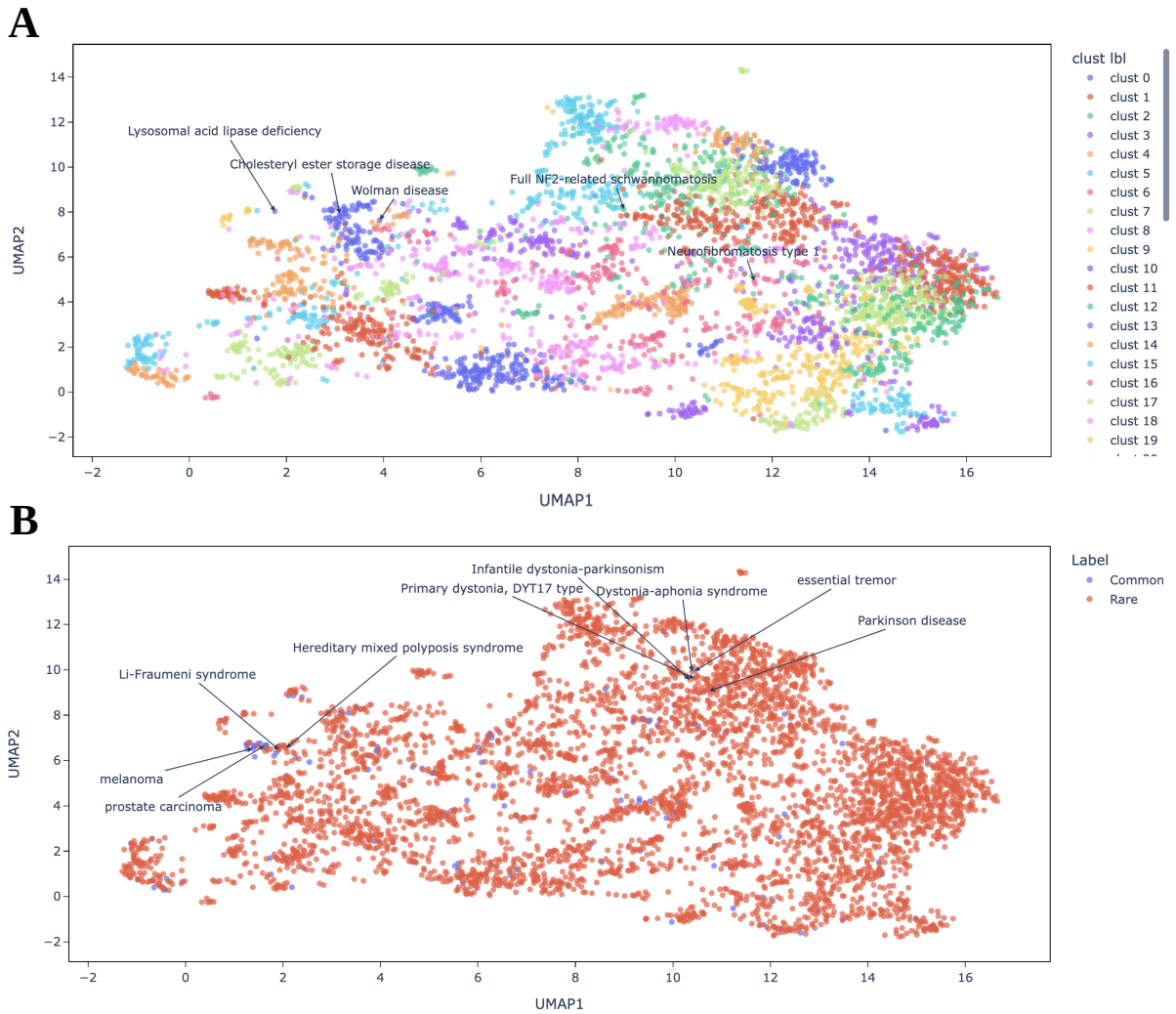

Figure 5: **Unsupervised Diseases Clustering.** We used the 128-dimensions vectors of Dis2Vec to cluster rare and common diseases (Methods). **(A)** Next, we utilised UMAP to reduce Dis2Vec 128-dimensions, interestingly there is a considerable alignment between UMAP 2-dimension and unsupervised clusters. We assessed the quality of clusters with multiple positive and negative controls. For instance, Liposomal Acid Lipase Deficiency (LAL-D) and its subtypes, Cholesteryl Ester Storage Disease and Wolman Disease are located in cluster 40, although UMAP did locate them slightly separated reflecting differential symptom annotations from Orphanet. Despite this differential annotations, Dis2Vec still appropriately clustered these two disease subtypes. In contrast, the unsupervised clustering locate NF1 and NF2 in different clusters and far from each other. This is supported by pathophysiology of these diseases, NF1 and NF2 are both Neurofibromatosis disorders, but are related to distinct biological pathways and manifest in different phenotypes. **(B)** Both common and rare diseases are based on the same Human biology. We see this hypothesis supported by the unsupervised clustering, where related rare and common diseases are clustered together. Common neurological disorders like Parkinson Disease (PD) and Essential tremor are adjacent are clustered with rare forms of dystonia and hereditary forms of PD. Moreover, common cancers such as Melanoma and Prostate carcinoma co-locate with the with rare, hereditary cancers including Li-Fraumeni syndrome and Hereditary Mixed Polyposis Syndrome.

Supplementary Information

Representation Learning of Disease Mechanisms for a Foundation Model in Rare and Common Diseases

Contents

1 Prevalence of Rare Diseases **2**

2 OpenTargets Disease Gene Asosciations **3**

3 Disease Similarity from Polyarchichal Structure of Ontologies **4**

4 Negative Sampling from Orphanet Rare Disease Ontology by Maximizing LCA Distances **5**

5 Tripartite Network of Biological Processes, Diseases, and Symptoms **6**

6 Degree Distribution of Disease and Benchmark Networks **7**

7 Compositionality of Dis2Vec Space **7**

1 Prevalence of Rare Diseases

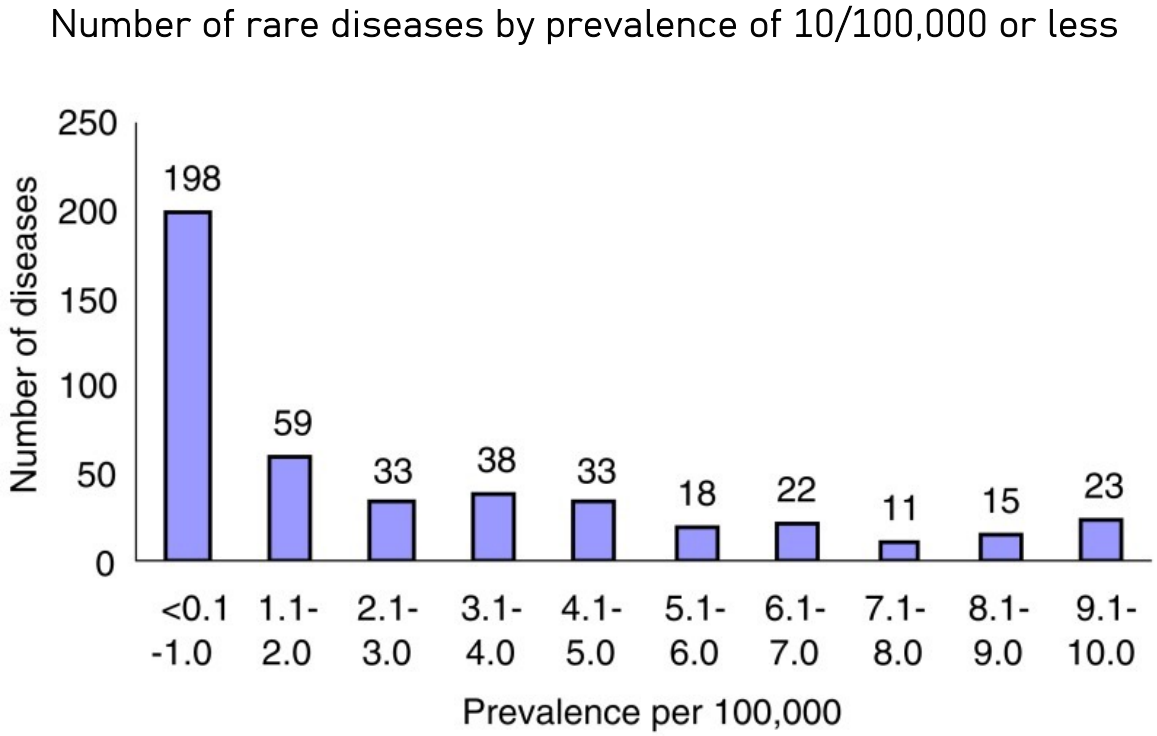

Figure S1: **Prevalence of Rare Diseases.** The distribution of prevalence of rare diseases follow a long tail, where majority of rare disease are ultra-rare, hence various phenotype are observed in the low prevalence rare diseases. Only a few rare diseases have comparably higher prevalence, hence are more studied. Lastly, rare diseases are heavily under diagnosed, making it difficult to obtain accurate prevalence of majority rare diseases. Source of the distribution plot: Courtesy of Boat and Field [1].

2 OpenTargets Disease Gene Asosciations

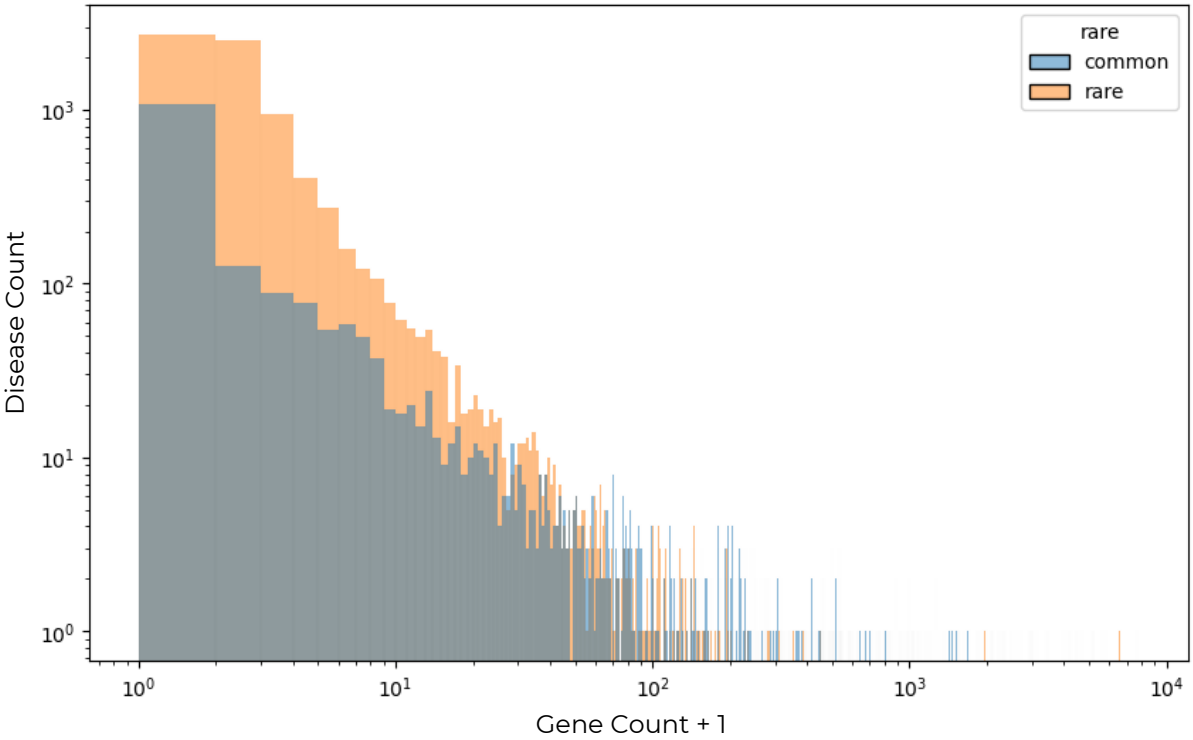

Figure S2: **Distribution of Number of Disease-Gene Associations from OpenTargets for Common and Rare Diseases.** No filters are applied, for example diseases without genetic score are included if they have disease-gene associations with the OpenTargets overall score, which includes extended sources such NLP-based disease-gene associations.

3 Disease Similarity from Polyarchichal Structure of Ontologies

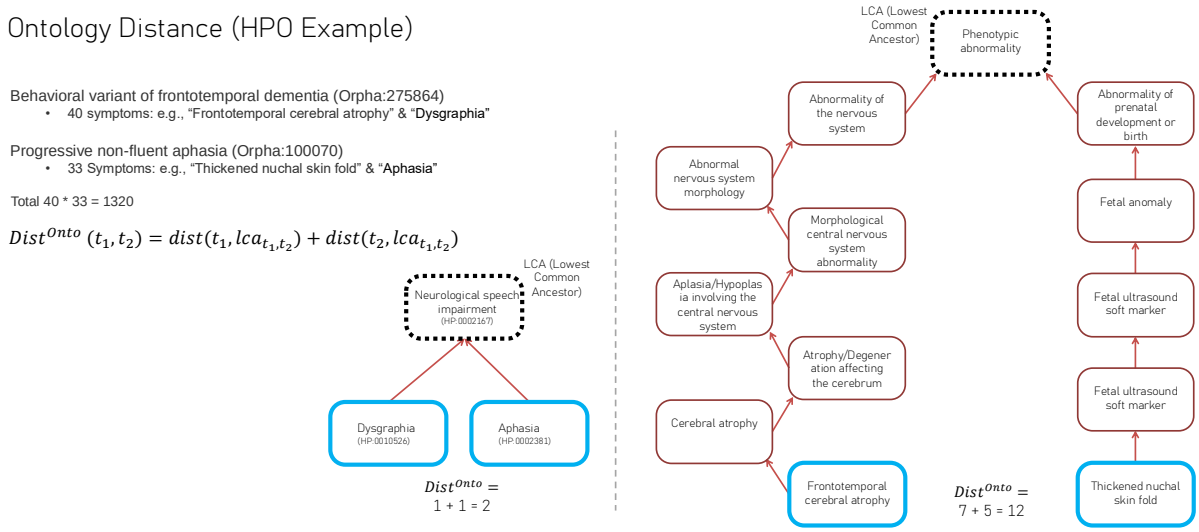

Figure S3: **Distance to Lowest Common Ancestor.**

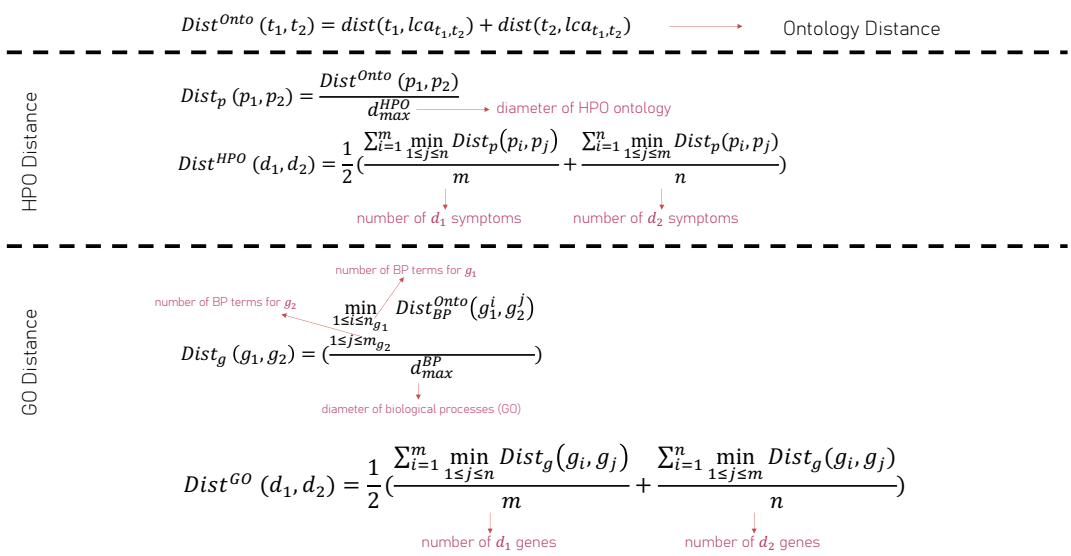

Figure S4: **Disease Similarity by Minimizing LCA Distances.** The diameter of a network is obtained by finding the longest shortest path in the network.

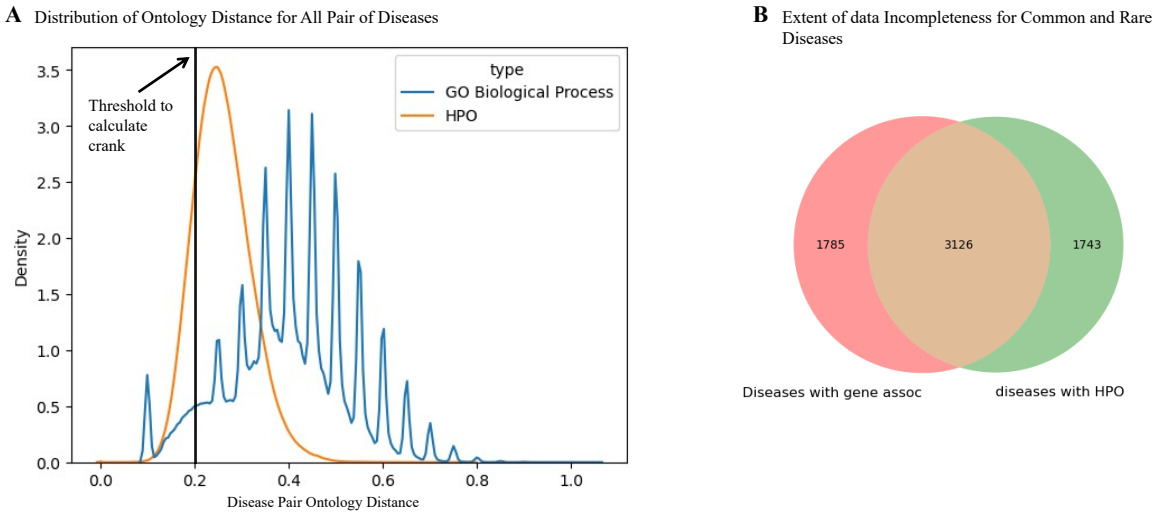

Figure S5: **Distributions of Disease-Pair Similarity Distances from HPO and GO. (A)** Due to minimizing LCA distances for genes then projecting a harmonic mean to diseases (Figure S4), we observed oscillation in disease similarity distances for GO Biological Processes hierarchy. Based on the distributions, We selected 0.2 threshold for obtaining disease-disease association. **(C)** Not all disease have both symptom and gene associations, hence it is necessary to combine both modalities to obtain a comprehensive disease representation.

4 Negative Sampling from Orphanet Rare Disease Ontology by Maximizing LCA Distances

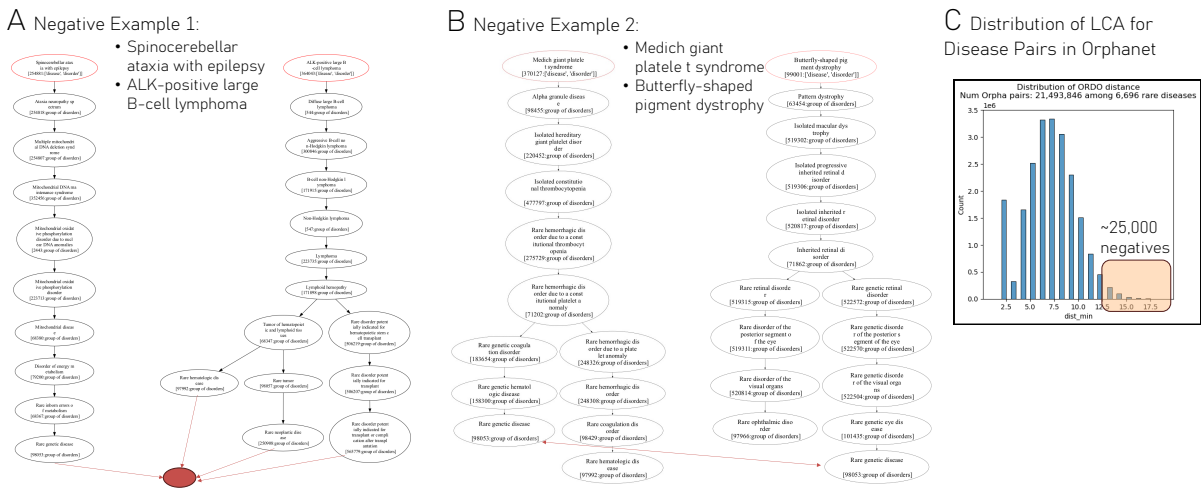

Figure S6: **Maximize LCA Distance to Obtain ORDO Disparity Network.**

5 Tripartite Network of Biological Processes, Diseases, and Symptoms

OntoDist network is created by calculating LCA distances in the structure of ontologies (GO and HPO). Hence, OntoDist captures the non-linear dependencies in ontologies in relation with disease pairs to form a uni-unipartite network of disease-disease relationships. The structural characteristics of Dis2Vec enables learning disease-disease associations using a two-shot learning approach (Figure 4).

Similarly, we can combine disease-BiologicalProcesses and disease-symptoms to build the GODS (GO-Disease-Symptom) tripartite network (Figure S7). with two edge types. Similar to the model for disease-disease learning (Figure 4), we utilized a two-shot learning approach to conduct link prediction over the GODS network. This link prediction task enables to extend disease association learning to predicting disease-BiologicalProcesses and disease-symptoms. Further exploration of GODS network and its potential down stream tasks are beyond scope of this work.

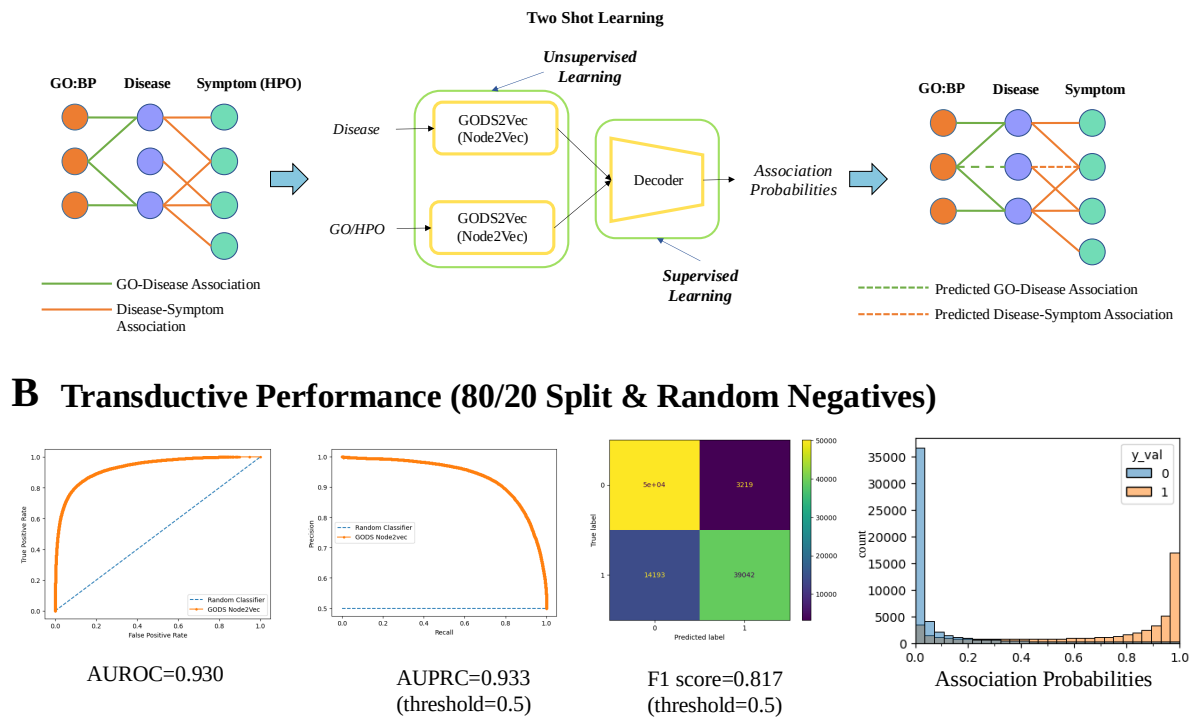

Figure S7: **A Tripartite Network for Expanding Disease Association learning to Disease-Biological Processes and Disease-Symptom Link Prediction.**

6 Degree Distribution of Disease and Benchmark Networks

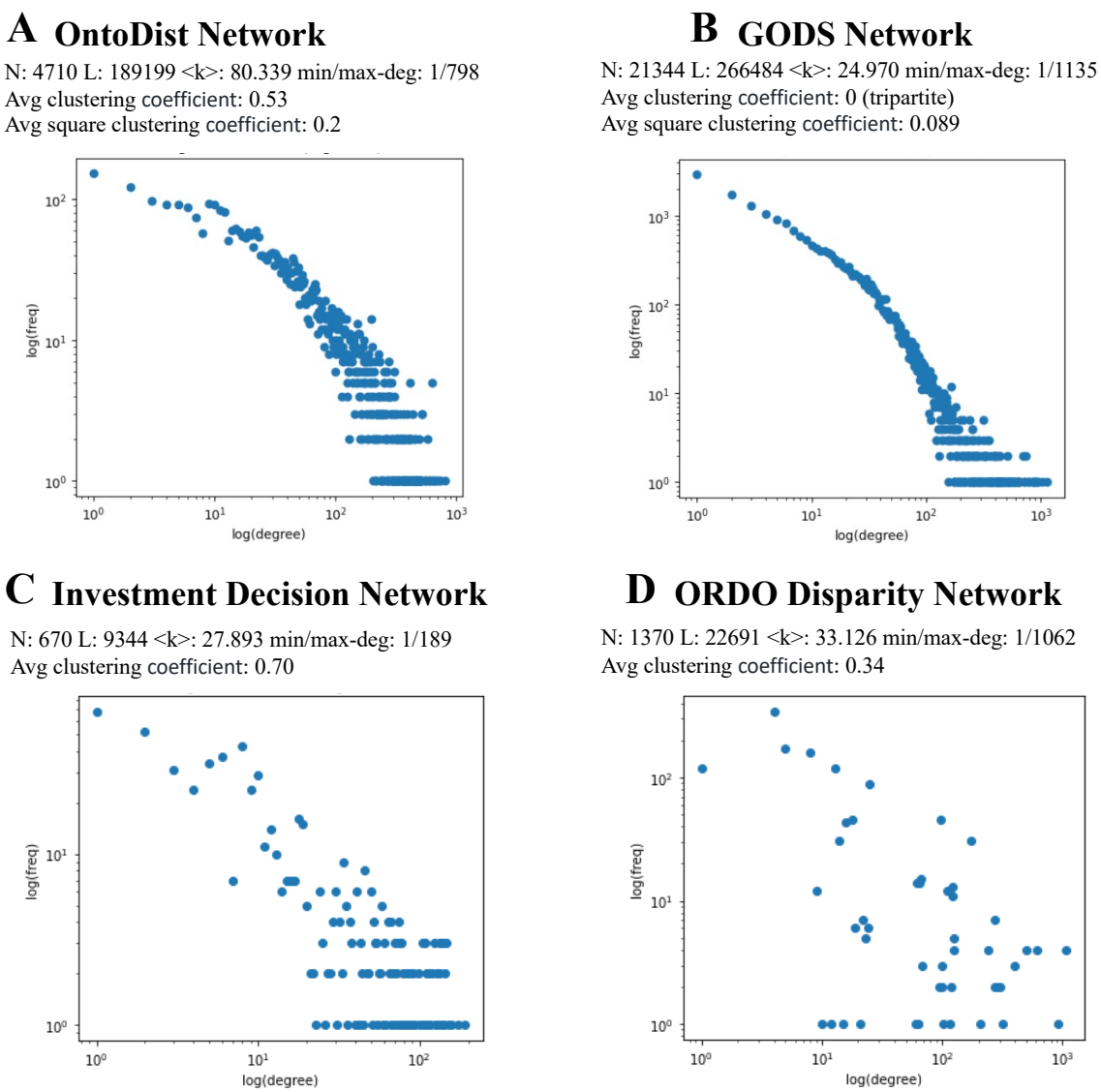

Figure S8: **Degree Distributions of Four Networks used in Dis2Vec.** The four networks exhibits very different patterns of interactions and structure, making the task of transfer learning from Dis2Vec to Investment Decision and ORDO Disparity networks more difficult.

7 Compositionality of Dis2Vec Space

Compositionality is a characteristic of some well-structured vector representations [2] wherein analogical reasoning can be implemented as simple linear operations, like vector addition. In word representations for example, the closest word to the sum of "Russia" + "capital" might be "Moscow". Some degree of compositionality is also evident in the Dis2Vec representations of diseases. Take the example of Mixed Connective Tissue Disease (ORPHA:809), which has overlapping features of the following 4 diseases in varying combinations: Systemic Lupus Ery-

thromatous (ORPHA:536), Systemic Sclerosis (ORPHA:90291), Primary Sjogren Syndrome (ORPHA:289390), Systemic-onset juvenile idiopathic arthritis (ORPHA:85414), and Polymyositis (ORPHA:732), which are all in Cluster 31 although Polymyositis is located farther in the UMAP space. Another example are the many paraneoplastic syndromes associated with Thymoma (ORPHA:99867), including Myasthenia gravis (ORPHA:589), and Dermatomyositis (ORPHA:221). The barycenter of these 2 diseases is most proximal Thymoma (after Myasthenia Gravis itself). These examples suggest that the positioning of overlap syndromes in the Dis2Vec representation corresponds roughly with linear combinations of disorders representing their overlapping features.

References

[1] Boat, T. F. & Field, M. J. Rare diseases and orphan products: Accelerating research and development (2011).

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality (2013). URL `https://arxiv.org/abs/1310.4546v1`.