Deep learning for diagnosing patients with rare genetic diseases

Emily Alsentzer[1,2,*], Michelle M. Li[1,3,*], Shilpa N. Kobren[1], Undiagnosed Diseases Network, Isaac S. Kohane[1], and Marinka Zitnik[1,‡]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[2]Program in Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA 02115, USA
‡Corresponding author. Email: marinka@hms.harvard.edu
*Equal contribution

There are more than 7,000 rare diseases, some of which affect 3,500 or fewer patients in the US. Due to clinicians' limited experience with such diseases and the considerable heterogeneity of their clinical presentations, many patients with rare genetic diseases remain undiagnosed. While artificial intelligence has demonstrated success in assisting diagnosis, its success is usually contingent on the availability of large labeled datasets. Here, we present SHEPHERD, a deep learning approach for multi-faceted rare disease diagnosis. SHEPHERD is guided by existing knowledge of diseases, phenotypes, and genes to learn novel connections between a patient's clinico-genetic information and phenotype and gene relationships. We train SHEPHERD exclusively on simulated patients and evaluate on a cohort of 465 patients representing 299 diseases (79% of genes and 83% of diseases are represented in only a single patient) in the Undiagnosed Diseases Network. SHEPHERD excels at several diagnostic facets: performing causal gene discovery (causal genes are predicted at rank = 3.52 on average), retrieving "patients-like-me" with the same gene or disease, and providing interpretable characterizations of novel disease presentations. SHEPHERD demonstrates the potential of artificial intelligence to accelerate the diagnosis of rare disease patients and has implications for the use of deep learning on medical datasets with very few labels.

Main

Rare diseases affect 300-400 million people worldwide, yet each disease has a very low prevalence, involving no more than 50 per 100,000 individuals [1, 2]. Due to their low prevalence, most front-line clinicians lack disease experience, resulting in numerous specialty referrals and expensive clinical workups for patients across multiple years and institutions. Furthermore, patients with the same disease can present variable symptoms, disease severity, and age of onset [3]. Such challenges make the task of rare disease diagnosis extremely difficult; approximately 70% of individuals seeking a diagnosis and up to 50% of the suspected Mendelian conditions remain undiagnosed [4, 5]. These diagnostic delays can lead to redundant testing or unnecessary medical procedures, inappropriate or delayed disease management, and irreversible disease progression if the time window for intervention is missed.

Machine-assisted diagnosis offers the opportunity to shorten diagnostic delays for rare disease patients. Advances in artificial intelligence (AI) and deep learning have considerably improved diagnostic accuracy [6–15]. Deep learning models that have been trained (via supervised learning) on labeled datasets can achieve near-expert clinical accuracy for common diseases, including diabetic retinopathy [16], skin cancers [17], and pediatric diseases [18]. In addition, these AI-assisted tools can augment physicians' decision-making in the clinic and support medical diagnosis, particularly in settings with limited or time-constrained resources.

However, existing AI-assisted diagnostic approaches require labeled datasets with thousands of diagnosed patients per disease in order to train deep learning models. Therefore, the applicability of these models to rare diseases is unclear — datasets are three orders of magnitude smaller than in other uses of AI for medical diagnosis. For example, a deep convolutional neural network developed for diagnosing diabetic retinopathy and diabetic macular edema was trained using a labeled dataset of retinal images from 128,175 patients [16]. Meanwhile, due to the heterogeneity and low prevalence of each rare disease, AI models are unlikely to have seen patients with the same—or similar—genetic disorders during training. Moreover, the low prevalence of rare diseases precludes the creation of datasets of sufficient size to use deep learning, even with manual expert curation. These reasons indicate that AI-assisted diagnosis of rare diseases encounters challenges distinct from other uses of AI for diagnosis. Concretely, approaches must be able to extrapolate beyond the training distribution to novel genetic conditions and atypical disease presentations (in other words, generalize to conditions or presentations of known conditions that were not observed during training). Further, given the lack of large labeled datasets, the successful use of AI depends

on the ability to learn from sparsely labeled datasets.

Here, we introduce SHEPHERD, a deep learning approach for multi-faceted diagnosis of patients with rare genetic conditions. SHEPHERD operates at multiple points throughout the rare disease diagnosis process to perform causal gene discovery, retrieve "patients-like-me" with similar conditions, and provide interpretable names for novel disease presentations. To overcome the limitations of supervised learning, SHEPHERD performs label-efficient training by (1) training exclusively on simulated rare disease patients without the use of any real-world labeled cases and (2) incorporating external knowledge of known phenotype, gene and disease associations via knowledge-guided deep learning. The simulated patients used for training are created using an adaptive simulation approach that provides realistic rare disease patients with varying numbers of phenotypes and candidate genes [19]. Knowledge-guided learning is achieved by training a graph neural network to represent a patient (specifically, the patient's presenting phenotypes) in relation to other phenotypes, genes, and diseases. When a new patient arrives, SHEPHERD produces an embedding (a dense vector of real numbers) of the patient such that the patient's embedding in the latent space is located close to the patient's candidate causal gene and disease as well as to other patients with the same gene or disease, and far away from irrelevant genes and diseases of other patients with different diseases. Using the embedding space optimized for rare disease diagnosis, SHEPHERD nominates genes and diseases for a patient even when no other patients are diagnosed with the same disease. Taken together, the two above mentioned components of SHEPHERD enable deep learning to diagnose rare genetic diseases, a medical problem defined by a small number of labeled cases.

We evaluate SHEPHERD on an external cohort of patients in the Undiagnosed Diseases Network (UDN) [20], a nationwide initiative with 12 clinical sites in the US tasked with diagnosing patients with rare, difficult to diagnose genetic conditions. In addition to the multi-site UDN cohort, our external evaluation includes a nationwide MyGene2 patient cohort. SHEPHERD performs granular, phenotype-based causal gene discovery by ranking candidate genes that are output from bioinformatics or expert pipelines. We find that SHEPHERD ranks the correct gene first in 40% of patients spanning 16 disease areas, improving diagnostic efficiency by at least twofold compared to a non-guided baseline. In addition, SHEPHERD nominates the correct diagnosis for patients with atypical presentations or novel genetic diseases, ranking the correct gene among the top five predictions for 75% of those hard-to-diagnose patients. By testing SHEPHERD on each disease area, clinical site, and year of diagnosis, we find that SHEPHERD has sustained performance over

time and across diseases and clinical sites in the UDN. Further, SHEPHERD generates meaningful patient representations that capture patient similarity (Adjusted Mutual Information = 0.304) and enable retrieval of "patients-like-me" with similar genetic conditions. Finally, SHEPHERD can provide interpretable characterizations of novel disease presentations. By describing never-before-seen diseases based on their similarity to known genetic diseases, SHEPHERD can point clinical researchers towards the most closely related diseases to investigate the novel disease in depth. For each use case, we illustrate SHEPHERD's capabilities on case studies from patients in the Undiagnosed Diseases Network and provide an interactive demo to explore SHEPHERD's predictions at https://huggingface.co/spaces/emilyalsentzer/SHEPHERD.

Results

Overview of the Undiagnosed Diseases Network patient cohort

We assemble a cohort of 465 patients in the Undiagnosed Diseases Network (UDN) with molecular diagnoses. Most patients are diagnosed with a single causal gene that explains their symptoms, 14 patients (3%) have two causal genes, and two patients (0.4%) have three causal genes. Most patients in the UDN receive an extensive clinical workup and whole genome or exome sequencing (Figure 1a). Sequencing data is analyzed with the involvement of clinicians and genetic counselors to identify candidate genes that harbor variants likely to explain the patient's symptoms. Once one to five strong candidates are identified, causality is assessed by searching for genotype- and phenotype-matched individuals in human and animal databases or by introducing candidates into model organisms to assess in vivo impact [21].

Through this diagnostic process, patients are annotated with a set of Human Phenotype Ontology (HPO) terms describing their symptoms or findings and a set of candidate genes that may explain the patient's syndrome. Clinical experts additionally annotate diagnosed patients with an Online Mendelian Inheritance in Man (OMIM) identifier describing their disease (if available). Each patient is characterized by 23.9 phenotypes on average (SD = 16.1; Figure 1b). The candidate genes are patient-specific and include genes in which the patient has a mutation. For each patient, the diagnostic process creates two sets of candidate gene lists, which are derived from genes considered at two different phases in the UDN diagnosis pipeline (Figure 1a): VARIANT-FILTERED, a gene list produced by initial variant-based filtering of candidate genes, and EXPERT-CURATED, a gene list that includes genes marked by clinical experts as strong candidates for the patient (Methods 2.1). The VARIANT-FILTERED gene lists are produced using Exomiser [22,23], a variant-based

tool that is used in parallel to existing pipelines at three UDN sites [21]. The two candidate gene lists contain 244.3 and 13.3 genes on average respectively (SD = 244.0 and SD = 8.0; Figure 1b). Each gene list is fed into SHEPHERD to nominate the causal gene, *i.e.*, the gene harboring variants that cause the patient's disease, from both a long list of candidates derived from automated filtering (VARIANT-FILTERED) and a short list of the strongest candidates that are more challenging to prioritize (EXPERT-CURATED).

Patients have heterogeneous disease presentations: 378 unique genes and 299 unique diseases are represented in the cohort, and 48% of phenotypes, 79% of genes, and 83% of diseases are represented in only a single patient (Figure 1c). On average, patients with the same disease have only 67% of phenotypes in common (SD = 43%). In addition, 7% of patients have novel genetic diseases, and only 28% of each patient's phenotypes are known to be associated with the causal gene on average (SD = 21%). The assembled cohort of UDN patients has been evaluated at 12 clinical sites across the United States (Figure 1d). While 75.9% of patients are less than five years old, patients can present to the UDN with suspected genetic diseases in their 40s or 50s (Figure 1e). Most patients present with neurological symptoms but can exhibit cardiac, musculoskeletal, rheumatic, and many other symptoms (Figure 1f). Due to the lag between starting the process at the UDN and receiving the diagnosis, most patients included in the analysis were evaluated by UDN clinicians in 2016-2018 (Figure 1g). The phenotypic heterogeneity and presence of novel and atypical diseases pose a challenge for diagnosis, requiring diagnostic technology that can accommodate previously unseen phenotypes, genes, and diseases and leverage knowledge beyond direct gene, phenotype, and disease associations (Supplementary Figure S1). The UDN patients represent a diverse, independent cohort that we use exclusively for model evaluation. Importantly, these patients are not used to train SHEPHERD.

Overview of SHEPHERD algorithm

Given a set of patient's clinical phenotypes and candidate disease(s) or candidate gene(s) harboring causal variants, SHEPHERD performs multi-faceted diagnosis of the patient to identify causal genes, retrieve "patients-like-me" with the same causal gene or disease, or provide interpretable characterizations of novel disease presentations (Figure 1h). SHEPHERD can integrate into the rare disease diagnostic process workflow at multiple points: (1) to find similar patients after the patient's clinical workup, (2) to identify strong candidate causal genes after the initial sequencing analysis or in conjunction with the clinical case review, and (3) to characterize the

patient's disease and/or find similar patients for experimental or cohort validation after candidate causal genes are identified.

SHEPHERD is a few-shot deep learning approach for rare disease diagnosis. Few-shot learning, which can make predictions when very few, if any, labeled data points are available, is central to rare disease diagnosis, given the low prevalence of each disease. Key to SHEPHERD's ability to provide diagnostic prediction when zero or at most a few labeled (diagnosed) patients per disease are available is to use a rare disease knowledge graph. SHEPHERD represents each patient as a subgraph of phenotypes in the knowledge graph of gene, phenotype, and disease associations (Methods 1). It jointly embeds each patient's phenotype subgraph and the candidate genes or diseases such that embeddings are informed by all of the existing biomedical knowledge (Figure 2a). The embedding function is a graph neural network that maps biomedical concepts and patient information to the embedding space such that patients embed nearby their causal gene(s), disease(s), and other similar patients. SHEPHERD leverages an attention mechanism to generate the aggregated embedding for each patient's set of phenotypes, which can be inspected to assess the contribution of each phenotype to the prediction.

Mathematically, SHEPHERD is first pretrained to embed genes, phenotypes, and diseases by learning to predict whether relations exist in the knowledge graph (Figure 2a). This step produces embeddings that satisfy three criteria: they are compact and amenable to further AI analyses, embeddings are biologically meaningful, and they are broadly generalizable by accounting for complementarity between diseases. Then, using the pretrained model as initialization, SHEPHERD is further trained for multi-faceted diagnosis of rare diseases through a novel objective function (Methods 3).

Due to the scarcity of data for patients with rare monogenic diseases, we leverage simulated but realistic rare disease patients for training SHEPHERD. We develop SHEPHERD on a cohort of over 40,000 simulated rare disease patients representing over 2,000 rare diseases in Orphanet (Figure 2b, Methods 2.3). The simulated data is critical for training a deep learning model for rare disease diagnosis. The simulated cohort is considerably larger, more diverse, and more representative of phenotype and genotype heterogeneity than any real-world dataset of rare disease patients. Furthermore, the trained models can be released without the risk of exposing any patient information [24]. While unable to represent all rare diseases, this dataset enables the training of deep learning models in conjunction with knowledge-guided few-shot approaches. Importantly, SHEPHERD learns how to generalize to novel diseases by being trained in a disease-stratified manner, in

which we assign patients with the same disease exclusively to the training or validation set. We externally evaluate SHEPHERD on the multi-site UDN cohort and nationwide MyGene2 cohorts of patient diagnoses.

SHEPHERD can perform causal gene discovery

A critical step in rare disease diagnosis is identifying the gene(s) that are strong candidates for causing the patient's syndrome (Figure 1a). Given a patient's set of phenotypes and a list of genes in which the patient has a mutation, SHEPHERD nominates genes harboring variants most likely to explain the patient's presenting symptoms. SHEPHERD produces a score for each candidate gene in the patient that fuses two complementary aspects of information: an embedding-based aspect that captures the global network topology and an aspect based on knowledge graph distance that captures local network information (Methods 3.3). We use SHEPHERD to prioritize genes found in both the EXPERT-CURATED and VARIANT-FILTERED candidate gene lists (Methods 2.1). In both instances, SHEPHERD performs granular prioritization by refining lists of patients' candidate genes outputted by bioinformatics pipelines. As such, SHEPHERD can complement existing variant-based approaches for gene prioritization while leveraging the extensive knowledge sources of gene-phenotype associations.

We report SHEPHERD's performance as the average recall at k, defined as the number of causal genes retrieved in the top k ranked genes on average for all patients in the cohort. SHEPHERD ranks the patient's causal gene first in 40% of UDN patients, achieving a recall of 0.69 when $k = 3$ and 0.85 when $k = 5$ on average (Figure 3a). On the much longer VARIANT-FILTERED gene lists, SHEPHERD achieves an average recall of 0.30, 0.60, and 0.73 for $k = 1$, 5, and 10, respectively (Supplementary Figure S2).

We evaluate SHEPHERD against six approaches, including an information-theoretic method, a network science baseline, shallow embedding lookup, two supervised learning strategies, and non-guided random reference (Methods 5.2). The baselines represent the different classes of methods that can perform the task. SHEPHERD significantly outperforms the second best approach in retrieving the causal gene first by 5% (p-value $= 9.30 \times 10^{-4}$) and outperforms other machine learning approaches by 24% or more (p-value $= 3.60 \times 10^{-6}$). Using SHEPHERD, clinicians would need to evaluate 1,012 genes from the EXPERT-CURATED lists or 4,019 genes from the VARIANT-FILTERED lists in order to arrive at the causal gene for all 465 UDN patients. In contrast, with non-guided ranking, clinicians would need to evaluate a total of 2,231 EXPERT-CURATED genes

or 27,727 VARIANT-FILTERED genes, suggesting that SHEPHERD has the potential to improve diagnostic efficiency by 2.2-times and 6.9-times, respectively. Furthermore, compared to the second best approach, SHEPHERD reduces the number of genes clinicians need to consider by 118 and 1,479 for the EXPERT-CURATED and VARIANT-FILTERED lists, achieving a 10% and a 40% reduction in the number of genes, respectively.

Stratified performance across UDN patients

We find no significant difference in performance across UDN sites throughout the US (p-value $= 0.235$; Kruskal-Wallis H-test Figure 3c), across the year of evaluation by UDN clinicians (p-value $= 0.789$; Figure 3d), and across patients with varying presenting symptoms (p-value $= 0.762$; Figure 3e). These results indicate that SHEPHERD is broadly generalizable across UDN sites and disease presentations over time. SHEPHERD's ability to generalize is essential because rare disease patients represent a heterogeneous group, and developing separate models that perform well for each subgroup is intractable due to the low prevalence of the disorders.

SHEPHERD can diagnose patients with atypical and novel genetic diseases

Patients in the UDN have atypical or novel disease presentations, which makes them challenging to diagnose because there are no direct associations between patients' genes, symptoms, and the correct diagnosis. This means that a lookup against medical knowledge bases is ineffective for diagnosis. We find that SHEPHERD can identify the causal gene even when the patient's presenting phenotypes are multiple hops away from the gene causing the disease in the knowledge graph. No strong correlation exists between SHEPHERD's performance and the distance between the patient's phenotypes and causal gene (Figure 3b; R2 = 0.166).

In the following, we demonstrate the use of SHEPHERD for patients diagnosed with atypical presentation of a known disease or a novel syndrome. Let us first illustrate how SHEPHERD nominates the correct diagnosis for a patient with atypical disease representation and considerable phenotypic heterogeneity. In particular, patient UDN-P1 (Figure 4a; SHEPHERD Demo Tab 1, Patient UDN-P1) received a diagnosis for POLR3-related leukodystrophy three years after acceptance into the UDN. While the involvement of gene *POLR3A* with leukodystrophy (MIM:607694) is known, the patient's case was challenging due to her atypical clinical presentation. Several of her presenting phenotypes, including lack of tear production, premature adrenarche, laryngeal cleft, hearing loss, and high blood pressure, are not typical of leukodystrophy. Further, only 28.3% (13 out of 46) of the patient's phenotypes are directly linked to *POLR3A* in the knowledge graph,

and the patient phenotypes are 1.98 hops away from the causal gene in the knowledge graph on average. The *POLR3A* gene is associated with five other diseases, and 93.7% (192 out of 205) phenotypes directly linked to *POLR3A* are not found in the patient, further complicating the diagnosis. Despite this atypical disease presentation, SHEPHERD identifies the patient's causal gene in the top 2 out of 17 and 86 candidate genes in the EXPERT-CURATED and VARIANT-FILTERED gene lists, respectively. Other genes placed high in the ranking by SHEPHERD include *KAT6A* and *UBE3A*, which cause Arboleda-Tham and Angelman syndromes and explain many symptoms the patient has had, but also indicate that symptoms alone cannot disambiguate diseases. Strikingly, SHEPHERD can disambiguate diseases by optimally up- and down-weighting phenotypes using an attention mechanism and correctly down-weights phenotypes that are atypical of leukodystrophy.

SHEPHERD can also identify strong candidate genes for patients with novel uncharacterized syndromes. Patient UDN-P2 (Figure 4b; SHEPHERD Demo Tab 1, Patient UDN-P2) was accepted into the UDN with several disparate presenting symptoms, including duodenal atresia, intestinal malrotation, vascular anomalies, pancreatic exocrine insufficiency, liver disease, and developmental delay. According to UDN clinicians, the most likely diagnosis for this patient's symptoms is a *GLYR1*-associated novel syndrome characterized by pancreatic insufficiency and malabsorption. The *GLYR1* gene is not associated with any known diseases. There are no phenotypes related to the gene in the knowledge graph, and the average shortest path length from the patient's phenotypes to the causal gene is 2.2. Nevertheless, SHEPHERD correctly identifies the suspected causal gene first in the EXPERT-CURATED candidate list and the top 9 of the 82 genes in the VARIANT-FILTERED candidate list, illustrating how SHEPHERD can assist in recognizing novel genetic diseases.

SHEPHERD finds rare disease patients with similar genetic and phenotypic features

Another key consideration for rare disease diagnosis is finding patients that share the same disease or causal gene, commonly referred to as "patients-like-me" [25] (Figure 1a). Starting from a set of patient phenotypes, SHEPHERD flags other patients in the cohort with similar genetic diseases suitable for follow-up diagnostic analysis. Concretely, SHEPHERD finds similar patients through a deep embedding scorer optimized to represent patients with the same causal genes or disease as nearby points in the embedding space (Figure 5a). For this analysis, we combine patients from three cohorts for a total of 43,235 patients: the simulated cohort, the UDN cohort, and another external MyGene2 cohort. The cohort from MyGene2 (part of the Matchmaker Exchange, a federated platform used by the UDN for validating strong candidates via case matching [26])

consists of patients with rare genetic diseases who decided to share their health information with other families, clinicians, and researchers (Methods 2.2).

SHEPHERD represents each patient as a point in the embedding space colored by the disease category of their diagnosed disease. The categories correspond to the 33 disease categories outlined in Orphanet (Methods 2). Robust clustering of patients by disease area (AMI = 0.304; p-value < 0.01) shows that SHEPHERD generates the embedding space that meaningfully captures patient relationships that can directly answer "patients-like-me" queries. Remarkably, even though SHEPHERD is trained on simulated patients, it generalizes to real-world UDN and MyGene2 cohorts, revealing disease-enriched regions in the embedding space where real-world patients are positioned nearby simulated patients with the same disease area (Supplementary Figure S3).

To further evaluate patient embeddings, we compare embedding distances between patients diagnosed with either the same or different disease (*i.e.*, comparing diagonal vs. off-diagonal entries, Figure 5b). We find that distances between patients of the same category are significantly smaller than between patients of different categories (p-value $< 1 \times 10^{-10}$ across all disease categories; Mann-Whitney test), which indicates that SHEPHERD captures the similarity between patients with similar disease presentations. We also observe several distinct clusters of disease categories in the embedding space (Figure 5b; Supplementary Figure S4). For example, patients with neoplastic diseases and gastroenterologic diseases cluster together. Similarly, patients with hematologic and hepatic diseases, and patients with odontologic and renal diseases cluster together in the embedding space. These clusters represent real co-occurrences of symptoms in disease presentations. For instance, patients with odontologic diseases, atypical dentin dysplasia, and orofaciodigital syndrome I, have both orofacial and renal disease presentations. Atypical dentin dysplasia is caused by a mutation in *SMOC2*, a matricellular protein involved in both craniofacial development and kidney fibrosis [27, 28]. Orofaciodigital syndrome I is caused by a mutation in *OFD1*, which is involved in organogenesis and plays an important role in the normal growth of orofacial and kidney tissues [29, 30]. These relationships reflect that diseases often involve multiple organ systems and indicate that the embedding space can capture the relationship between patients with similar symptoms even when their diagnoses are different.

SHEPHERD can identify "patients-like-me" with similar genetic diseases

We next examine SHEPHERD's ability to identify "patients-like-me" from a large cohort of rare disease patients. We either rank all simulated, UDN, and MyGene2 patients (UDN-P3 and

UDN-P4 cases) or all UDN and MyGene2 patients (UDN-P5 and UDN-P6 cases; Figure 5a; SHEPHERD Demo Tab 2) to identify patients most similar to the query UDN patient. We locate each query patient and all similar patients with the same causal gene in SHEPHERD's embedding space and find that patients with the same causal gene are embedded nearby in the space. In all four patient cases, SHEPHERD retrieves patients with the same causal gene and disease as the query patient among the top 5 predictions. Patients ranked above the patient with the same causal gene have very similar disease presentations to the query patient. For UDN-P4 and UDN-P5, the patients have a variant of the same disease caused by a different gene (Figure 5a). For UDN-P6, patients with Coffin-Siris syndrome 8 (ranked first) and GATAD2B-associated syndrome (ranked second) both exhibit impaired intellectual development, hypotonia, feeding difficulties, and hypertelorism, among other phenotypes. For UDN-P3, patients with X-linked intellectual disability due to *GRIA3* (ranked first) and Coffin-Lowry syndrome (ranked second) share impaired intellectual development, seizures, scoliosis, and other phenotypes.

The most similar patients identified by SHEPHERD do not necessarily have the most phenotypes in common with the query patient. This reflects SHEPHERD's ability to capture phenotypic similarity rather than just calculating a direct overlap in phenotypes, which is typical of some information-theoretic approaches used in practice. In particular, patients that share the same causal gene have two to four phenotypes in common. Only 10.0%, 9.0%, 26.6%, and 7.7% of phenotypes found in query patients UDN-P3, UDN-P4, UDN-P5, and UDN-P6 are also found in the most similar genotype-matched individual respectively. In contrast, patients who have the most phenotypes in common with the query are ranked at positions 366, 463, 41, and 16, respectively. For example, one patient shares 10 phenotypes with UDN-P6, which is 38.5% of UDN-P6's phenotypes, yet has a different causal gene and is ranked 16th. This capability of SHEPHERD to consider indirect, deep associations between genes and phenotypes makes SHEPHERD highly complementary to graph theoretic techniques and statistical tests that can only score direct associations, which can be ineffective for poorly characterized diseases.

Further, we compare SHEPHERD to two approaches that can calculate phenotypic similarity: an information theoretic approach, which uses information theory to calculate the similarity between two sets of phenotypes based on shared ancestors in the Human Phenotype Ontology, and a set-based approach that uses Jaccard distance defined as $1-(|P_i \cap P_j|)/(|P_i \cup P_j|)$ where P_i and P_j represent phenotype sets for two patients. We measure the phenotypic similarity between all pairs of UDN patients and MyGene2 patients with known disease categories, and we rank the 43,820

total comparisons according to their phenotypic similarity. Finally, we assess whether each method for calculating phenotypic similarity can differentiate between patients with the same versus distinct disease categories. Intuitively, patients with diseases in the same disease category should be more similar than those with diseases in different disease categories.

We find that SHEPHERD is best able to capture the similarity between patients with the same disease category (Figure 5c). Furthermore, SHEPHERD assigns higher phenotypic similarity to patients with diseases in the same disease category compared to patients with diseases in different disease categories (median rank of 13,296 and 25,015 for pairs of patients with the same versus different disease categories, respectively). In contrast, the information-theoretic approach can only differentiate these two groups to a lesser extent (median rank of 16,094 versus 24,035, respectively). The set-based distance metric fails to capture the similarity between patients whose diagnosed diseases are of the same category. This is unsurprising given the limited phenotypic overlap across rare disease patients, as the set-based distance only considers the overlap in phenotype terms between patients and cannot capture phenotypic similarity.

Finally, we evaluate whether SHEPHERD embeds patients with the same disease (rather than disease category) closer to each other than to patients with different diseases. We compare UDN patients to MyGene2 patients. Again, we find that embedding distances between patients diagnosed with the same disease are significantly smaller compared to patients with different diseases (p-value $= 2.42 \times 10^{-8}$; Kolmogorov-Smirnov test; Figure 5d), further strengthening the evidence that SHEPHERD can capture similarities between different diseases with similar presenting symptoms, but can nevertheless differentiate patients that have the same diagnosed disease.

SHEPHERD provides an interpretable characterization of novel diseases

In addition to supporting causal gene discovery and patients-like-me identification, SHEPHERD can help characterize novel clinical presentations through our current knowledge of rare diseases (Figure 1a). Given a patient's phenotypes, SHEPHERD provides an interpretable name for the patient's disease based on its similarity to each disease in the KG. Specifically, SHEPHERD produces a ranked list of all diseases using the embedding similarity between each disease and the patient's phenotypes. More concretely, SHEPHERD learns an embedding space in which the similarity between a patient and a disease is inversely proportional to the embedding distance between the patient and their diagnosed disease (Figure 6a). To enable additional interpretable characterization of the patient's disease, we aggregate SHEPHERD-generated similarities of individual diseases

by their disease category to generate a distribution of similarities to disease categories (where the distribution sums to 100%). For example, a patient's presenting syndrome may be $w_1\%$ similar to rare neurologic diseases, $w_2\%$ similar to rare bone diseases, $w_3\%$ similar to rare developmental defects during embryogenesis, etc. Overall, we find that SHEPHERD learns to embed patients near diseases of the same category; on average, 45.7% of the top 10 ranked diseases with a known disease category belong to the same category as the patient's disease, which is nearly three times more than random expectation alone (16.4%). We investigate such capabilities by separately exploring SHEPHERD's predictions for known and novel rare diseases.

To evaluate SHEPHERD's ability to provide interpretable disease names for patients with known rare diseases, we first calculate the similarity between UDN patients and all diseases. This allows us to assess whether the patients are most similar to diseases that share the same disease category as the patient's disease (Figure 6b). Concretely, for each patient, we stratify patients by their primary disease category and calculate the average similarity of a patient to all disease nodes under each disease category. As expected, we find that patients tend to be most similar to diseases of the same disease category as their own. For example, patients with a rare bone disease are most similar to diseases under the category of rare bone disease (13.0% similarity), followed by rare developmental defects during embryogenesis (10.2%), rare inborn errors of metabolism (9.6%), and rare odontology diseases (8.2%). Similarly, patients with a disease categorized as a rare developmental defect during embryogenesis, a rare inborn error of metabolism, or a rare neurologic disease tend to be most similar to other diseases of the same category.

We next examine two patients in depth to interrogate SHEPHERD's predictive capabilities for characterizing known rare diseases: UDN-P7 and UDN-P9. Patient UDN-P7 (red, top left corner of Figure 6a; SHEPHERD Demo Tab 3, Patient UDN-P7) received a diagnosis for limb-girdle muscular dystrophy 3 (sarcoglycanopathy; MIM:608099) due to variants in *SGCA*. SHEPHERD compares the patient's clinical presentation to diseases across 19 disease categories and finds that the patient is most similar to rare neurologic diseases, as expected. In fact, from SHEPHERD's interpretable name for the patient's disease, two of the top five most similar diseases are other types of AR limb-girdle muscular dystrophy, and all five are related to muscular dystrophy. Patient UDN-P9 (light blue, bottom right corner of Figure 6a; SHEPHERD Demo Tab 3, Patient UDN-P9) was diagnosed four years after acceptance to the UDN with the bone disease spondyloepimetaphyseal dysplasia caused by a mutation in *RPL13*. Again, SHEPHERD can ascertain in its predicted interpretable name that the patient's symptoms are similar to other bone diseases, and all top 5 ranked disorders are rare

bone diseases with overlapping phenotypes found in the query patient.

Finally, we investigate SHEPHERD's ability to provide interpretable names for two patients with novel genetic conditions, UDN-P8 and UDN-P2. The novelty of their genetic conditions is due to the lack of patients with disease-gene relationships. UDN-P8 (dark blue, bottom left corner of Figure 6a; SHEPHERD Demo Tab 3, Patient UDN-P8) was diagnosed with *ATP5PO*-related Leigh syndrome caused by a novel mutation in *ATP5PO*, a gene previously unassociated with any disease [31]. As Leigh syndrome is a metabolic disorder with neuropathological features, SHEPHERD produces the interpretable name that correctly identifies UDN-P8's disease as being most similar to diseases under the categories of rare inborn errors of metabolism and rare neurologic diseases. Three of the top five diseases—combined oxidative phosphorylation deficiency 39 (MIM:618397; ranked by SHEPHERD as #1), pyruvate dehydrogenase E3-binding protein deficiency (MIM:245349; ranked by SHEPHERD as #3), and combined oxidative phosphorylation defect type 26 (MIM:616672; ranked by SHEPHERD as #5)—are mitochondrial diseases affecting the same pathway as *ATP5PO* and result in a defect in the aerobic energy production. These diseases' causal genes co-localize with *ATP5PO* [32–35]. Combined oxidative phosphorylation deficiency 39 and combined oxidative phosphorylation defect type 26 are associated with neurological presentations of mitochondrial disease, including hypotonia, seizures, and features of Leigh syndrome [36]. The remaining two most similar diseases (ranked by SHEPHERD as #2 and #4) are rare neurologic diseases with phenotypes identical to UDN-P8's. The causal gene, *CNP*, for the second-ranked disease, hypomyelinating leukodystropy-20 (MIM:619071), is three hops away from *ATP5PO* in the physical protein interaction network [37, 38], suggesting that they may be functionally related [39–41] or operate together [42, 43] to mediate phenotypes associated with UDN-P8's disease and hypomyelinating leukodystropy-20.

Patient UDN-P2 (dark blue, top right corner of Figure 6a; SHEPHERD Demo Tab 3, Patient UDN-P2), previously described in Figure 4b), is characterized by SHEPHERD's interpretable name as most similar to diseases under the categories of rare inborn errors of metabolism, rare hepatic disease, rare gastroenterological disease, and rare endocrine disease. These top categories are aligned with many of the patient's symptoms, particularly duodenal atresia, intestinal malrotation, pancreatic exocrine insufficiency, liver disease, and developmental delay. Three of the top five most similar individual diseases from SHEPHERD's interpretable name—Methylmalonic acidemia with homocystinuria type cblF (MIM:277380; ranked by SHEPHERD as #1), Neonatal hemochromatosis (MIM:231100; ranked by SHEPHERD as #2), and ALG8-CDG (MIM:608104; ranked by

SHEPHERD as #4)—are also due to inborn errors of metabolism, and the diseases are associated with phenotypes that are similar to those seen in the patient, including abnormalities in liver and gastrointestinal function and developmental delay. Notably, the rare respiratory disease category is the third lowest-ranked category. UDN clinicians hypothesized that the patient's *GLYR1* variants cause a mislocalization of the cystic fibrosis conductance regulator (*CFTR*), which is associated with cystic fibrosis. While the patient has gastrointestinal and pancreatic symptoms similar to those in cystic fibrosis, the patient does not have any of the pulmonary features classic for that condition. Such granularity in SHEPHERD's predictions is a reflection of SHEPHERD's ability to differentiate between diseases despite partially overlapping phenotypes and causal genes sharing the same pathway.

Discussion

We present SHEPHERD, a deep learning approach for multi-faceted rare disease diagnosis. SHEPHERD overcomes limitations of traditional data-hungry AI approaches by (1) infusing external knowledge via deep learning on a knowledge graph, (2) leveraging label-efficient learning to align patient, gene, and disease representations, and (3) training on a large disease-split cohort of simulated patients. SHEPHERD generalizes to never-before-seen phenotypes, genes, and diseases, performing well in patients with heterogeneous clinical presentations and novel genetic conditions (Extended Data S1). The model leverages an attention mechanism to generate patient phenotype embeddings, whose weights provide insight into the contribution of each phenotype to the prediction. SHEPHERD is broadly applicable across multiple points in the diagnostic process, as shown by evaluations on two multi-site patient cohorts with varying disease presentations.

A unique feature of SHEPHERD is its ability to generate multi-modal representations of patients with rare genetic diseases. We model patient phenotypes as subgraphs and candidate genes and diseases as nodes in a large knowledge graph, which allows the representations to be infused with external biomedical knowledge. The multi-layer graph neural network enables indirect associations multiple hops away in the knowledge graph to influence the learned representations. While many existing approaches rely exclusively on known phenotype-gene-disease associations [44,45], leveraging indirect associations is essential for diagnosing patients with novel or atypical genetic conditions. Furthermore, subgraphs provide an elegant, flexible mathematical definition for modeling sets of patient phenotypes. Rather than model each phenotype individually [46], we can encode patients as a structured object (namely, a subgraph) and consider the co-occurrence of phenotypes

when diagnosing rare diseases. Genetic mutations can yield pleiotropic effects, and joint modeling of patient phenotypes is important for capturing co-morbidities, which can uniquely identify diseases [47–49].

SHEPHERD demonstrates the value of simulated data for training machine learning models. While simulated data is increasingly being leveraged to augment existing training sets to improve model robustness and generalizability [24, 50–54], here we exclusively use simulated patients to train SHEPHERD. Simulated data is not just an additional asset but a critical necessity for training deep learning models in the rare disease space, where data is extremely scarce. The synthetic patients are generated by a simulator [19] grounded in clinical-genetic knowledge. Furthermore, training on simulated data mitigates concerns regarding privacy breaches, in which specific individuals can be identified from the training data [55,56]. Hence, it is possible to publicly release the fully trained SHEPHERD without privacy concerns. We show that the models trained on simulated patients are readily applicable to diverse patients in the UDN and MyGene2 cohorts, providing evidence of the impact of synthetic data in a real-world clinical application.

There are several extensions to this work. Our method relies on a knowledge graph of disease, gene, and phenotype associations. Still, other sources of information, such as variant-level information or databases of model organism phenotype-gene associations, could be incorporated as well [57]. SHEPHERD's knowledge graph includes curated gene-phenotype-disease relationships and can be extended to include information from research literature [58]. The graph neural network underlying SHEPHERD is also readily extensible to multi-modal data types. For example, gene co-expression data or textual descriptions of diseases can be incorporated as node features. Furthermore, while efforts like the UDN are critical for establishing diagnoses for rare disease patients, they alone cannot address the rare disease burden. Approaches such as SHEPHERD can help identify and diagnose rare disease patients using claims data, electronic health records, and other data types. SHEPHERD's ability to characterize a patient's clinical presentation could be used to identify sub-specialists who should review the patient's case for the diagnostic recommendation.

Our study has a few limitations. First, our knowledge graph was constructed in June 2021. Additional associations of diseases, genes, and phenotypes since then may further improve SHEPHERD's performance. To this end, the knowledge graph curation and processing approaches are fully reproducible, and the graph can be automatically updated as data resources evolve and new data become available [59]. Second, the still-undiagnosed UDN patients may be more challenging than the already-diagnosed ones SHEPHERD was tested on. There are two categories of still-

undiagnosed patients: patients admitted to the UDN years ago who have yet to receive a diagnosis due to sequencing limitations (e.g., hard-to-detect variant types such as short tandem repeats or structural variants, missing second variants in recessive disorders, variants that lie in difficult-to-sequence regions or are masked due to biases in the human reference genome and ancestral genomes [60]), and patients recently admitted to the UDN. As sequencing approaches continue to advance, SHEPHERD can be evaluated on the still-undiagnosed patients whose causal variants will be detectable by the more accurate sequencing technologies. Moreover, the lack of an observed drop in SHEPHERD's performance for recently diagnosed patients indicates that data leakage has not occurred, evidently avoiding the bias that would otherwise cause overfitting of the model to the training data (e.g., information about older diagnoses being incorporated into the knowledge graph).

SHEPHERD demonstrates the use of AI for diagnosing rare disease patients. While AI-assisted diagnosis has focused on diseases for which large labeled datasets exist, this study shows how AI can be used for underserved rare diseases. Existing diagnostic processes require collaborations across bioinformaticians, clinicians, and genetic counselors. Reviewing even a single case can take many hours of a many-person team over days or weeks. SHEPHERD can substantially reduce the number of genes clinicians need to consider to provide a molecular diagnosis and identify patients with similar genetic conditions, even before they have undergone genetic sequencing. Scalable AI-based diagnostic strategies can enable efforts such as the UDN to shorten the diagnostic odyssey for rare disease patients.

Data availability. All data used in the paper, including the rare disease knowledge graph, simulated and MyGene2 cohorts, and the final and intermediate results of the analyses are shared with research community at https://zitniklab.hms.harvard.edu/projects/SHEPHERD. While the UDN dataset cannot be released in its entirety due to privacy concerns, anonymized UDN data has been deposited in dbGaP (accession phs001232) and PhenomeCentral. Phenotypes and causal variants and genes related to UDN diagnoses are also shared publicly in ClinVar at https://www.ncbi.nlm.nih.gov/clinvar/submitters/505999. The UDN study is approved by the NIH IRB Protocol 15HG0130. All patients accepted to the UDN provide written informed consent to share their data across the UDN.

Code availability. Python implementation of the methodology developed and used in the study is available via the project website at https://zitniklab.hms.harvard.edu/projects/SHEPHERD. The code to reproduce results, together with documentation and examples of usage, are at https://github.com/mims-harvard/SHEPHERD. We also provide an interactive demo for users to explore SHEPHERD's predictions at https://huggingface.co/spaces/emilyalsentzer/SHEPHERD.

Acknowledgements. We would like to thank Kimberly LeBlanc for reviewing our work to ensure that we are following UDN data privacy protocols. E.A. is supported by a Microsoft Research PhD Fellowship. M.L. is supported by T32HG002295 from the National Human Genome Research Institute and a National Science Foundation Graduate Research Fellowship. M.Z. gratefully acknowledges the support by NSF under Nos. IIS-2030459 and IIS-2033384, US Air Force Contract No. FA8702-15-D-0001, and awards from Harvard Data Science Initiative, Amazon Research, Bayer Early Excellence in Science, AstraZeneca Research, and Roche Alliance with Distinguished Scientists. UDN research reported in this manuscript was supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under the following award numbers: U01HG007709, U01HG010219, U01HG010230, U01HG010217, U01HG010233, U01HG010215, U01HG007672, U01HG007690, U01HG007708, U01HG007703, U01HG007674, U01HG007530, U01HG007942, U01HG007943, U01TR001395, U01TR002471, U54NS108251, and U54NS093793. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and other funders.

Authors contribution. E.A., M.L., and S.K. retrieved and processed the UDN, MyGene2, and simulated patient data. E.A. and M.L. developed, implemented, and benchmarked SHEPHERD and

performed detailed analyses of SHEPHERD's algorithm. E.A., M.L., I.K., and M.Z. designed the study. E.A., M.L., and M.Z. wrote the manuscript.

Competing interests. The authors declare no competing interests.

Undiagnosed Diseases Network Consortium. Undiagnosed Diseases Network: Maria T. Acosta, Margaret Adam, David R. Adams, Justin Alvey, Laura Amendola, Ashley Andrews, Euan A. Ashley, Mahshid S. Azamian, Carlos A. Bacino, Guney Bademci, Ashok Balasubramanyam, Dustin Baldridge, Jim Bale, Michael Bamshad, Deborah Barbouth, Pinar Bayrak-Toydemir, Anita Beck, Alan H. Beggs, Edward Behrens, Gill Bejerano, Hugo J. Bellen, Jimmy Bennet, Beverly Berg-Rood, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, Stephanie Bivona, Elizabeth Blue, John Bohnsack, Devon Bonner, Lorenzo Botto, Brenna Boyd, Lauren C. Briere, Elly Brokamp, Gabrielle Brown, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, Peter Byers, William E. Byrd, John Carey, Olveen Carrasquillo, Thomas Cassini, Ta Chen Peter Chang, Sirisak Chanprasert, Hsiao-Tuan Chao, Gary D. Clark, Terra R. Coakley, Laurel A. Cobban, Joy D. Cogan, Matthew Coggins, F. Sessions Cole, Heather A. Colley, Cynthia M. Cooper, Heidi Cope, William J. Craigen, Andrew B. Crouse, Michael Cunningham, Precilla D'Souza, Hongzheng Dai, Surendra Dasari, Joie Davis, Jyoti G. Dayal, Matthew Deardorff, Esteban C. Dell'Angelica, Katrina Dipple, Daniel Doherty, Naghmeh Dorrani, Argenia L. Doss, Emilie D. Douine, Laura Duncan, Dawn Earl, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Cecilia Esteves, Marni Falk, Liliana Fernandez, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Irman Forghani, William A. Gahl, Ian Glass, Bernadette Gochuico, Rena A. Godfrey, Katie Golden-Grant, Madison P. Goldrich, Alana Grajewski, Irma Gutierrez, Don Hadley, Sihoun Hahn, Rizwan Hamid, Kelly Hassey, Nichole Hayes, Frances High, Anne Hing, Fuki M. Hisama, Ingrid A. Holm, Jason Hom, Martha Horike-Pyne, Alden Huang, Yong Huang, Wendy Introne, Rosario Isasi, Kosuke Izumi, Fariha Jamal, Gail P. Jarvik, Jeffrey Jarvik, Suman Jayadev, Orpa Jean-Marie, Vaidehi Jobanputra, Lefkothea Karaviti, Jennifer Kennedy, Shamika Ketkar, Dana Kiley, Gonench Kilich, Shilpa N. Kobren, Isaac S. Kohane, Jennefer N. Kohler, Deborah Krakow, Donna M. Krasnewich, Elijah Kravets, Susan Korrick, Mary Koziura, Seema R. Lalani, Byron Lam, Christina Lam, Grace L. LaMoure, Brendan C. Lanpher, Ian R. Lanza, Kimberly LeBlanc, Brendan H. Lee, Roy Levitt, Richard A. Lewis, Pengfei Liu, Xue Zhong Liu, Nicola Longo, Sandra K. Loo, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A. MacRae, Valerie V. Maduro, Rachel Mahoney,

Bryan C. Mak, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Rong Mao, Kenneth Maravilla, Ronit Marom, Gabor Marth, Beth A. Martin, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Jacob McCauley, Allyn McConkie-Rosell, Alexa T. McCray, Elisabeth McGee, Heather Mefford, J. Lawrence Merritt, Matthew Might, Ghayda Mirzaa, Eva Morava, Paolo M. Moretti, Mariko Nakano-Okuno, Stan F. Nelson, John H. Newman, Sarah K. Nicholas, Deborah Nickerson, Shirley Nieves-Rodriguez, Donna Novacic, Devin Oglesbee, James P. Orengo, Laura Pace, Stephen Pak, J. Carl Pallais, Christina GS. Palmer, Jeanette C. Papp, Neil H. Parker, John A. Phillips III, Jennifer E. Posey, Lorraine Potocki, Barbara N. Pusey, Aaron Quinlan, Wendy Raskind, Archana N. Raja, Deepak A. Rao, Anna Raper, Genecee Renteria, Chloe M. Reuter, Lynette Rives, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Natalie Rosenwasser, Francis Rossignol, Maura Ruzhnikov, Ralph Sacco, Jacinda B. Sampson, Mario Saporta, Judy Schaechter, Timothy Schedl, Kelly Schoch, C. Ron Scott, Daryl A. Scott, Vandana Shashi, Jimann Shin, Edwin K. Silverman, Janet S. Sinsheimer, Kathy Sisco, Edward C. Smith, Kevin S. Smith, Emily Solem, Lilianna Solnica-Krezel, Ben Solomon, Rebecca C. Spillmann, Joan M. Stoler, Jennifer A. Sullivan, Kathleen Sullivan, Angela Sun, Shirley Sutton, David A. Sweetser, Virginia Sybert, Holly K. Tabor, Amelia L. M. Tan, Queenie K.-G. Tan, Mustafa Tekin, Fred Telischi, Willa Thorson, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Brianna M. Tucker, Tiina K. Urv, Adeline Vanderver, Matt Velinder, Dave Viskochil, Tiphanie P. Vogel, Colleen E. Wahl, Melissa Walker, Stephanie Wallace, Nicole M. Walley, Jennifer Wambach, Jijun Wan, Lee-kai Wang, Michael F. Wangler, Patricia A. Ward, Daniel Wegner, Monika Weisz-Hubshman, Mark Wener, Tara Wenger, Katherine Wesseling Perry, Monte Westerfield, Matthew T. Wheeler, Jordan Whitlock, Lynne A. Wolfe, Kim Worley, Changrui Xiao, Shinya Yamamoto, John Yang, Diane B. Zastrow, Zhe Zhang, Chunli Zhao, Stephan Zuchner

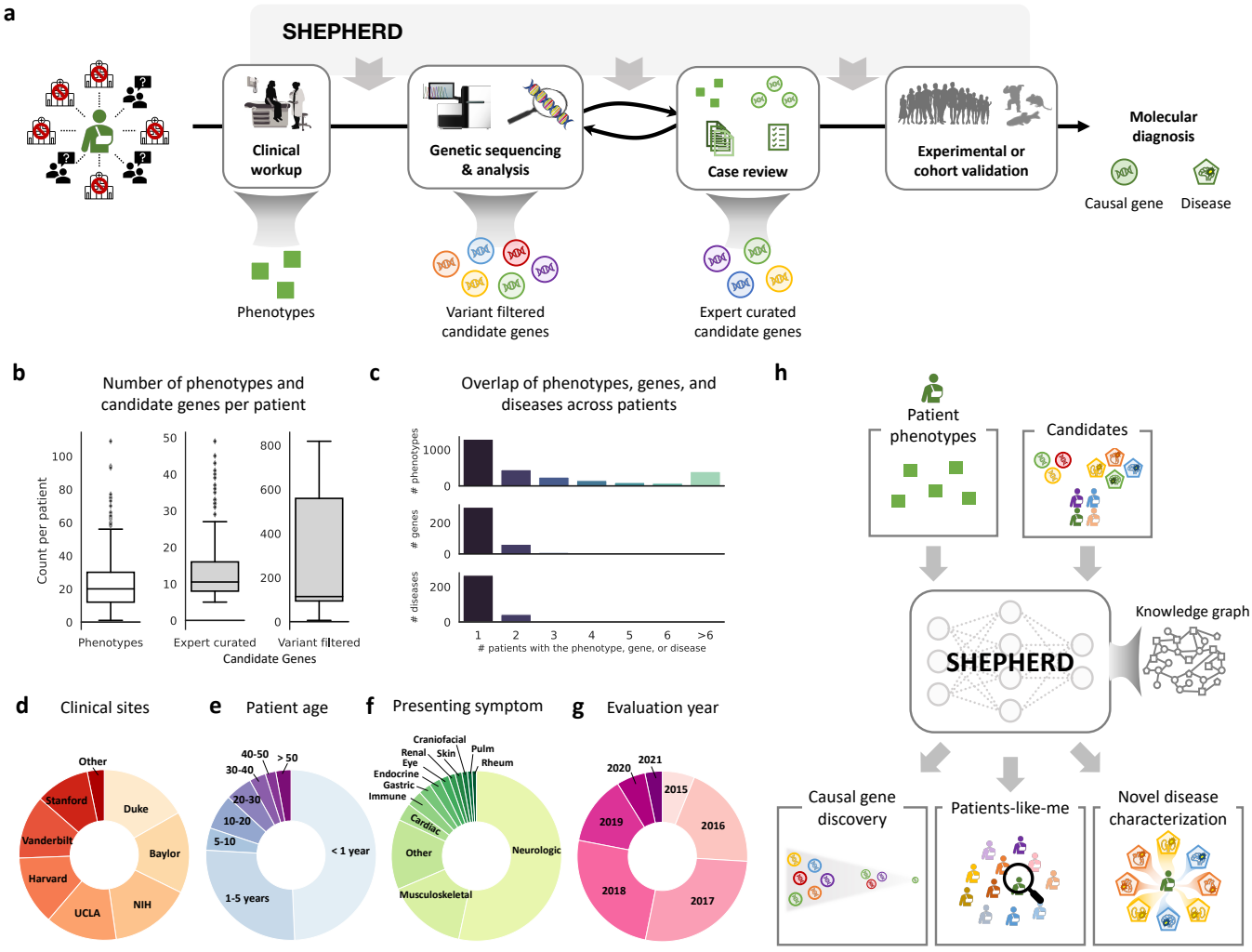


Figure 1: Overview of SHEPHERD in the rare disease diagnosis pipeline. **(a)** After years of failed diagnostic attempts, once a patient is accepted to the UDN, they receive a thorough clinical workup and genetic sequencing, and their case is analyzed in an iterative process to identify the candidate genes likely to explain the patient's symptoms. SHEPHERD can be used throughout the diagnostic process: after the clinical workup to find similar patients, after the sequencing analysis to identify strong candidate genes, and after the case review to further prioritize candidate genes, characterize the patient's disease, and/or validate candidate genes by finding phenotype and genotype-matched patients. **(b)** Number of phenotypes and candidate genes in each of the two candidate gene lists across patients in our UDN cohort. **(c)** Overlap of phenotypes, genes, and diseases across patients. Most phenotypes, genes, and diseases are found in only a single UDN patient. **(d-g)** Number of patients in each **(d)** UDN clinical site, **(e)** age category, **(f)** primary presenting symptom, and **(g)** evaluation year. **(h)** SHEPHERD takes in as input the patient's set of phenotypes as well a list of either candidate genes, patients, or diseases and leverages an external rare disease knowledge graph to perform multi-faceted rare disease diagnosis. For simplicity, the knowledge graph is depicted using three shapes: circles as genes, squares as phenotypes, and pentagons as diseases; refer to Methods for all node types.

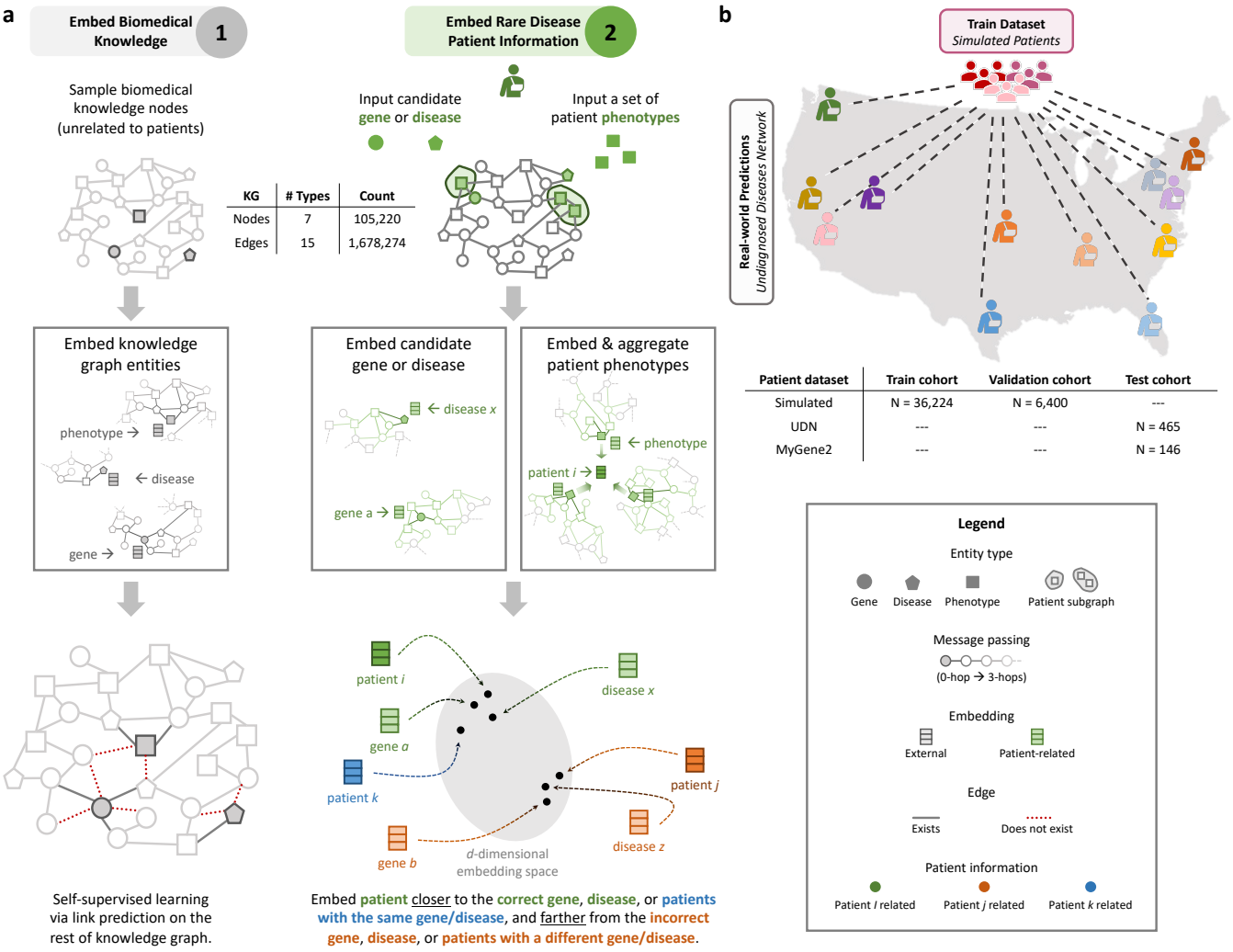


Figure 2: SHEPHERD architecture, training, and generalizability. (a) SHEPHERD is trained in a two-step process. First, the model is pretrained to embed the biomedical knowledge in the knowledge graph (left side). Then, the pretrained model is applied to the task of rare disease diagnosis (right side). Patient information is overlaid on the knowledge graph, and SHEPHERD generates an embedding for the patient phenotypes and each candidate gene, disease, or patient. The model is trained via a loss function that encourages patient embeddings to be close to the embeddings of their causal gene or disease or other patients with the same gene or disease. **(b)** SHEPHERD is trained on a large cohort of simulated patients and externally validated on patients across multiple sites in the Undiagnosed Diseases Network. For simplicity, the KG is depicted using three shapes: circles as genes, squares as phenotypes, and pentagons as diseases; refer to Methods for all node types.

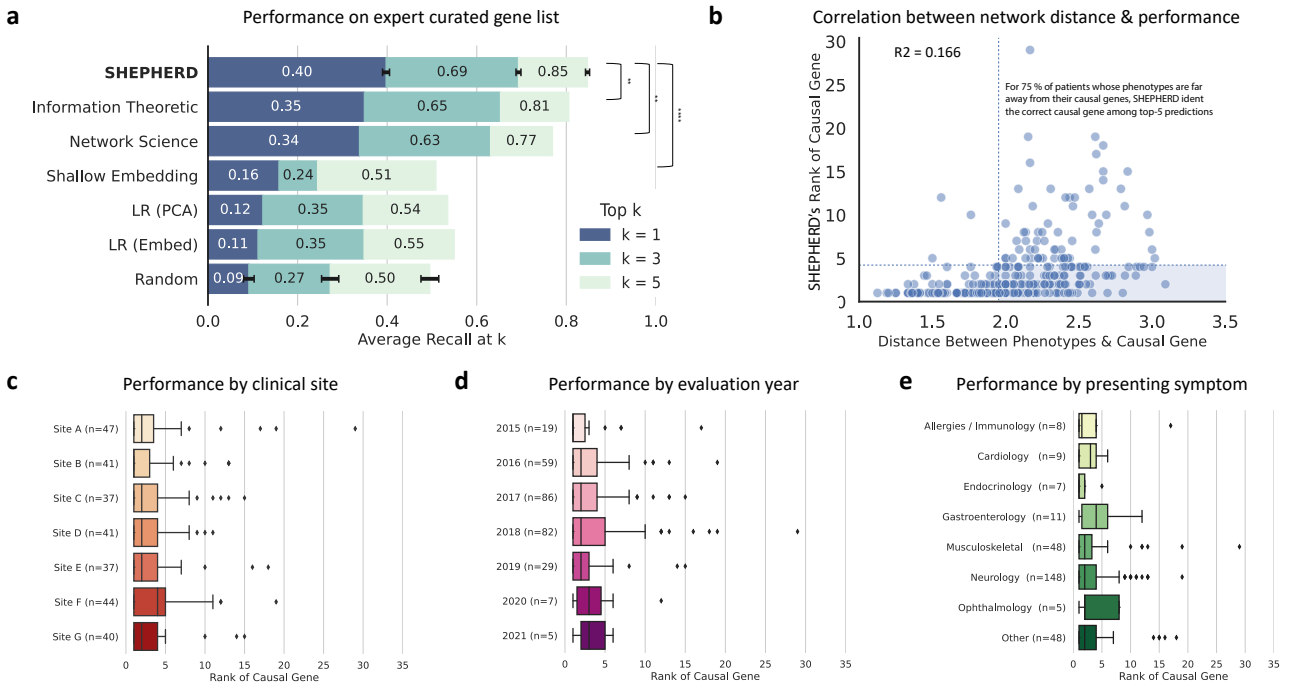
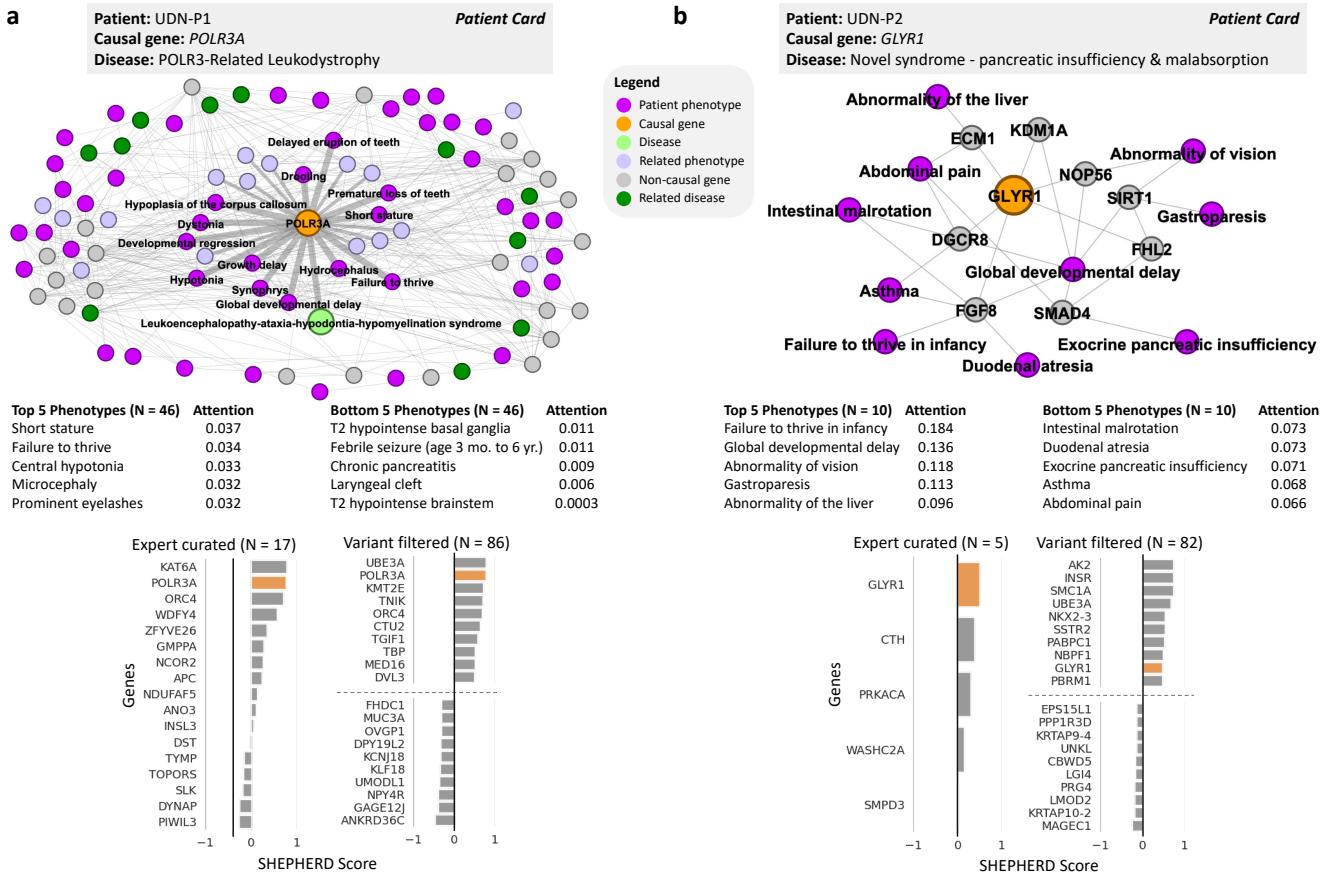


Figure 3: SHEPHERD performs generalizable causal gene discovery. **(a)** Performance of SHEPHERD and six baseline models evaluated via average recall at k for $k = 1$, 3, and 5. **(b)** Correlation between model performance (*i.e.*, rank of causal gene) and the average distance between a patient's phenotypes and causal gene in the knowledge graph. **(c-e)** Performance of SHEPHERD in ranking causal genes stratified by **(c)** clinical sites, **(d)** evaluation year, and **(e)** primary presenting symptom.



Top 5 Phenotypes (N = 46)	Attention	Bottom 5 Phenotypes (N = 46)	Attention
Short stature	0.037	T2 hypointense basal ganglia	0.011
Failure to thrive	0.034	Febrile seizure (age 3 mo. to 6 yr.)	0.011
Central hypotonia	0.033	Chronic pancreatitis	0.009
Microcephaly	0.032	Laryngeal cleft	0.006
Prominent eyelashes	0.032	T2 hypointense brainstem	0.0003

Top 5 Phenotypes (N = 10)	Attention	Bottom 5 Phenotypes (N = 10)	Attention
Failure to thrive in infancy	0.184	Intestinal malrotation	0.073
Global developmental delay	0.136	Duodenal atresia	0.073
Abnormality of vision	0.118	Exocrine pancreatic insufficiency	0.071
Gastroparesis	0.113	Asthma	0.068
Abnormality of the liver	0.096	Abdominal pain	0.066

Figure 4: Causal gene discovery case studies for patients with novel genetic conditions. SHEPHERD identifies the causal gene even in atypical or novel disease presentations. Each patient case study, shown in **(a)** and **(b)**, includes the subset of the knowledge graph containing all nodes in the shortest path between the patient's phenotypes, causal gene, and disease; a table of the patient's phenotypes and attention weights learned by SHEPHERD; and bar plots of scores SHEPHERD assigned to each candidate gene in the EXPERT-CURATED and VARIANT-FILTERED lists. The top and bottom 5 ranked genes in the VARIANT-FILTERED list are shown. The causal gene is highlighted in orange. In patient UDN-P1's network, the direct neighbors of the causal gene are emphasized. The patient's causal gene is directly connected to the disease in the knowledge graph. In patient UDN-P2's network, there is no disease node because the patient has a novel uncharacterized syndrome. All panels, except those labeled as a "Patient Card" (colored box with the information provided by the UDN), depict SHEPHERD's predictions or analyses performed on outputs of SHEPHERD.

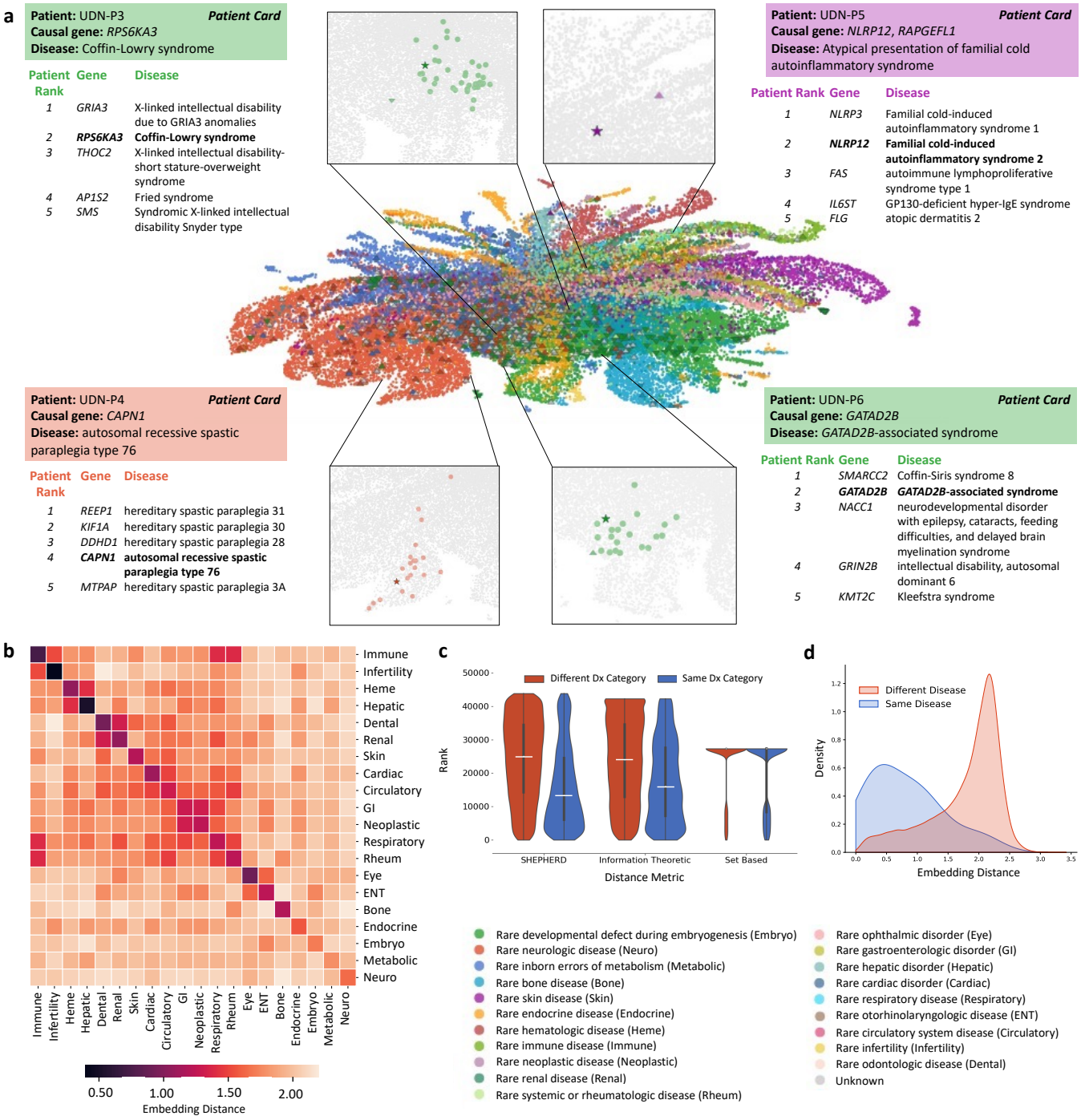


Figure 5: SHEPHERD identifies patients-like-me from simulated, UDN, and MyGene2 cohorts. **(a)** Two-dimensional UMAP plot of SHEPHERD's embedding space of all simulated (circle), UDN (up-facing triangle), and MyGene2 (down-facing triangle) patients colored by their Orphanet disease category. Each of the four case studies consists of a zoomed-in UMAP displaying the query patient (star) and all patients with the same causal gene as the query (colored circles) and a table containing information regarding the top five most similar patients retrieved by SHEPHERD. Patients are bolded in the table if they share the same causal gene. **(b)** Heatmap of the average distance between the phenotype embeddings of pairs of patients across disease categories. Darker colors indicate smaller distances and lighter colors indicate larger distances between patients of each pair of disease categories. **(c)** Violin plot comparing the ranks of patients with the same vs. different disease (dx) category using SHEPHERD, an INFORMATION THEORETIC approach, or a SET BASED approach. The white line in the violin plots represents the median value. **(d)** Distribution of SHEPHERD embedding distance between patients with the same vs. different diseases. All panels, except those labeled as a "Patient Card" (colored box with the information provided by the UDN), depict SHEPHERD's predictions or analyses performed on outputs of SHEPHERD.

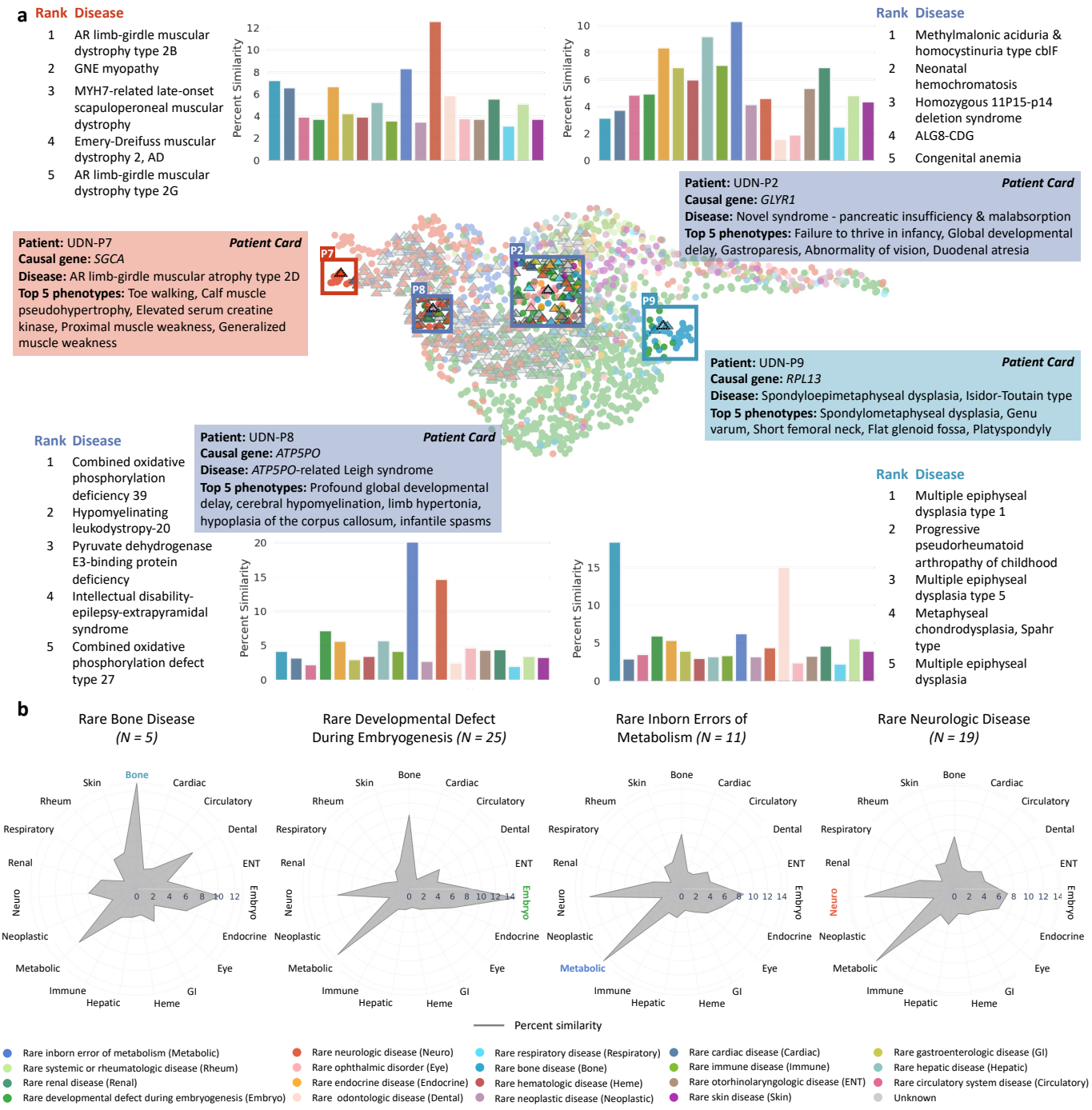


Figure 6: SHEPHERD performs novel disease characterization. (a) UMAP of SHEPHERD's embedding space of all UDN patients (up-facing triangle) and Orphanet disease nodes (circle), colored by disease category. Each of the four case studies consists of a list of the patient's five phenotypes that are most highly attended by SHEPHERD, distribution of average similarity of the patient to diseases in a particular disease category, and a table of the five most similar diseases according to SHEPHERD. **(b)** Radar plots of the similarity between UDN patients and diseases found in each disease category. We show group UDN patients by the disease category of their true disease and show plots for all categories with at least 5 patients. The disease category of each patient's true disease is bolded and colored. All panels, except the causal gene and disease listed in the "Patient Card" (colored box), depict SHEPHERD's predictions or analyses performed on outputs of SHEPHERD.

Online Methods

The Methods are structured as follows: 1) description of our rare disease knowledge graph, 2) description of our rare disease patient cohorts, 3) summary of our algorithmic approach for rare disease diagnosis, 4) details regarding model training, and 5) outline of our statistical analysis and evaluation setup.

1 Rare Disease Knowledge Graph Construction

We create a comprehensive knowledge graph (KG) for rare disease diagnosis. We start with PrimeKG [59] and adapt it to the rare disease setting by removing drug-specific entities and relations and adding additional sources of the gene, phenotype, and disease relationships. The resulting rare disease KG contains seven node types (i.e., phenotype, protein, disease, pathway, molecular function (MF), cellular component (CC), and biological process (BP)) and 15 unique relation types (i.e., phenotype-protein, disease-phenotype$^{(-)}$ (indicating that disease does not have phenotype), disease-phenotype$^{(+)}$ (indicating that disease has phenotype), protein-pathway, disease-protein, protein-MF, protein-CC, protein-BP, BP-BP, MF-MF, CC-CC, phenotype-phenotype, protein-protein, disease-disease, pathway-pathway).

1.1 Data Sources and Harmonization

Relationships are extracted from the following data sources: Gene Ontology (GO) [61], Reactome pathway knowledgebase [62], DisGeNET [63], NCBI [64], Human Phenotype Ontology (HPO) [65], MONDO disease ontology [66], and Orphanet [67]. PrimeKG contains disease-protein relationships from DisGeNET, and we add additional disease-protein and disease-phenotype relationships from Orphanet if they are not already present in the knowledge graph. All phenotypes are mapped to the Human Phenotype Ontology, all genes/proteins are mapped to Ensembl identifiers, and all diseases are mapped to MONDO identifiers. In instances where a concept is represented in both the HPO and MONDO ontologies, we remove the MONDO identifier. This differs from the original PrimeKG preprocessing, where conflicting identifiers are mapped to MONDO IDs. We perform all other preprocessing as in the original PrimeKG knowledge graph. For additional information about each data source and the harmonization process, refer to PrimeKG [59].

1.2 Knowledge Graph Pre-processing

We enforce homophily between genes and phenotypes by computing the triadic closure between gene-disease and disease-phenotype edges [68, 69]. We extract the largest connected component to ensure that the KG is fully connected. The largest connected component retains 99.91% of the nodes and 99.99% of the edges from the knowledge graph. Finally, we add reverse edges to ensure that the KG is represented as an undirected graph during model training.

1.3 Final Knowledge Graph Statistics

The final knowledge graph contains 105,220 nodes and 1,095,469 edges. Tables 1-2 outline the number of nodes and edges by node type and relation type, respectively.

Table 1: Rare disease knowledge graph. Reported is the number of nodes by node type. MF: molecular function, CC: cellular component, BP: biological process.

NODE TYPE	COUNT	AVERAGE DEGREE	VOCABULARY
PHENOTYPE	15,874	49.5 ± 190.6	HUMAN PHENOTYPE ONTOLOGY [65]
DISEASE	21,233	17.1 ± 37.4	MONDO [66]
PROTEIN	21,610	74.2 ± 119.8	ENSEMBL [70]
PATHWAY	2,516	19.0 ± 29.9	REACTOME [62]
MF	11,169	8.7 ± 127.4	GENE ONTOLOGY [61]
CC	4,176	22.3 ± 197.2	GENE ONTOLOGY [61]
BP	28,642	8.7 ± 28.1	GENE ONTOLOGY [61]

2 Rare Disease Patient Cohorts

We use three distinct rare disease patient cohorts for training and evaluating SHEPHERD: UDN (Section 2.1), a real-world cohort of hard-to-diagnose patients in the Undiagnosed Diseases Network, MYGENE2 (Section 2.2), a publicly available real-world cohort of patients with rare genetic conditions who have opted to share their information on the MyGene2 Portal, and SIMULATED (Section 2.3), a large diverse and realistic simulated patient cohort representing 2,134 unique rare diseases in Orphanet.

For every patient cohort, we categorize each patient's causal disease according to the 33 disease categories outlined in Orphanet. We map all diseases to Orphanet and leverage the Orphanet linearisation process (http://www.orphadata.org/cgi-bin/rare_free.html) to assign each disease to a single disease category based on a series of rules that consider the most severely affected body

Table 2: Rare disease knowledge graph. Reported is the number of edges by relation type. HPO: Human Phenotype Ontology.

RELATION TYPE	COUNT	SOURCES
PHENOTYPE-PHENOTYPE	21,925	HPO [65]
PHENOTYPE-PROTEIN	10,518	HPO [65], MONDO [66], ORPHANET [67]
DISEASE-DISEASE	35,167	DISGENET [63], MONDO [66], ORPHANET [67]
DISEASE-PHENOTYPE$^{(-)}$	1,483	HPO [65], DISGENET [63], MONDO [66], ORPHANET [67]
DISEASE-PHENOTYPE$^{(+)}$	204,779	HPO [65], DISGENET [63], MONDO [66], ORPHANET [67]
DISEASE-PROTEIN	86,299	DISGENET [63], MONDO [66], ORPHANET [67]
PROTEIN-PROTEIN	321,075	MENCHE *et al.* [71], BIOGRID [72], STRING [73], LUCK *et al.* [43]
PROTEIN-PATHWAY	42,646	REACTOME [62]
PATHWAY-PATHWAY	2,535	REACTOME [62]
PROTEIN-MF	69,530	GENE ONTOLOGY [61]
PROTEIN-CC	83,402	GENE ONTOLOGY [61]
PROTEIN-BP	144,805	GENE ONTOLOGY [61]
MF-MF	13,574	GENE ONTOLOGY [61]
CC-CC	4,845	GENE ONTOLOGY [61]
BP-BP	52,886	GENE ONTOLOGY [61]

system and the specialists most likely to be involved in treatment.

2.1 Patients in the Undiagnosed Diseases Network

Undiagnosed Diseases Network The Undiagnosed Diseases Network (UDN) is a nationwide research study supported by the National Institutes of Health Common Fund, which aims to bring together clinical and research experts around the United States to diagnose patients with rare genetic conditions [74]. The UDN consists of 12 clinical sites across the United States that evaluate patients, a sequencing core, a model organism screening center, a central biorepository, a metabolomics core, and a coordinating center. Patients are admitted to the UDN if they have objective findings and clinical testing has failed to produce a diagnosis. Most admitted patients receive either exome or full genome sequencing and an extensive clinical workup. The Undiagnosed Diseases Network study is approved by the National Institutes of Health institutional review board (IRB), which serves as the central IRB for the study (Protocol 15HG0130). All patients accepted to the UDN provide written informed consent to share their data across the UDN as part of a network-wide informed consent process.

Constructing patient subgraphs Deep phenotyping of patients during the clinical workup is a central component of the UDN process. Clinicians annotate each patient with a set of terms from the Human Phenotype Ontology (HPO) using PhenoTips, a tool integrated into the electronic health record that allows for structured phenotyping of patient symptoms [75]. We discard 406 unique prenatal phenotype terms related to the mother's pregnancy and use all remaining phenotype terms to construct patient subgraphs. Each patient subgraph is formed from the phenotype nodes in the rare disease knowledge graph that describe the patient's symptoms (Methods 3). In total, we construct phenotype subgraphs for the 465 UDN patients with annotated phenotypes who have received a molecular diagnosis as of January 5, 2022.

Obtaining EXPERT-CURATED candidate gene lists Genomic samples for each patient are sequenced at Baylor Genetics or Hudson Alpha. We construct an EXPERT-CURATED candidate gene list for each patient from the patient's sequencing data. Importantly, these gene lists are unique to each patient. The EXPERT-CURATED candidate gene list for each patient includes the union of both (1) disease-associated and other clinically-relevant genes listed on the patient's clinical sequencing reports from Baylor or Hudson Alpha per the UDN protocol and American College of Medical Genetics and Genomics (ACMG) guidelines [76–78] and (2) genes that were prioritized by UDN clinical teams who handled the patient's case. The genes in this list represent the strongest candidates identified by the UDN sequencing core or the clinical team. In addition, the list often includes known disease-causing genes, genes with suspected pathogenic variants, or genes expressed in tissues relevant to the patient's clinical presentation. While the EXPERT-CURATED gene list contains the strongest candidates, the list nevertheless requires further filtering to identify the ultimate causal gene(s) that explain the patient's condition. We exclude patients whose candidate gene lists have fewer than five candidate genes for the causal gene discovery task. The cohort contains 278 patients with at least five EXPERT-CURATED candidate genes.

Obtaining VARIANT-FILTERED candidate gene lists As part of the UDN analysis pipeline, the UDN performs the whole genome and/or exome sequencing for a patient and their immediate family members. Here, we use the patients' whole genome sequencing (WGS) data, which are aligned to the GRCh38.p13/hg38 human genome build and have undergone variant calling via the Genome Analysis Toolkit (GATK) best practices workflow [21]. Please refer to [21] for more details about the computational workflow across UDN sites. Access to the UDN patients' WGS

data allows us to construct for each patient a VARIANT-FILTERED candidate gene list consisting of genes that have at least one variant and that have been prioritized by a variant-level prioritization algorithm. We leverage the Exomiser algorithm, which considers variant frequency, predicted pathogenicity, and (if family members' sequencing data are available) mode of inheritance [79]. While Exomiser can leverage known associations between genes and phenotypes, we do not use it to construct our VARIANT-FILTERED candidate gene lists.

We analyze the patients' variant-called WGS data (i.e., variant call format, or VCF) using Exomiser under the following inheritance modes: autosomal dominant, autosomal recessive homozygous alternate, autosomal recessive compound heterozygous, X dominant, X recessive homozygous alternative, X recessive compound heterozygous, and mitochondrial. Their respective cutoff values (*i.e.*, the maximum minor allele frequency in percent (%) permitted for an allele to be considered a causative candidate under that mode of inheritance) are 0.1, 0.1, 2.0, 0.1, 0.1, 2.0, and 0.2. We remove variants with non-coding effects (*i.e.*, 5' and 3' UTR exon/intron variants, non-coding transcript exon/intron variants, coding transcript intron variants, up-/down-stream gene variants, intergenic variants, and regulatory region variants). We use the following pathogenicity sources, `POLYPHEN`, `MUTATION_TASTER`, and `SIFT`. We apply a frequency filter to remove variants with a frequency of at least 0.5% according to the variant frequency databases used. All variant frequency databases are used, as recommended by the Exomiser manual. We retain non-pathogenic variants in the output gene list. As with the EXPERT-CURATED gene lists, we filter out patients who do not have at least five candidate genes in their VARIANT-FILTERED gene list. The cohort includes 152 patients with at least five VARIANT-FILTERED candidate genes.

Preprocessing disease labels Diagnosed patients in the UDN are labeled with a disease identifier from the Online Mendelian Inheritance in Man (OMIM) database [80] when the patient is diagnosed with a known genetic disease. We map the OMIM disease identifiers to MONDO identifiers [66] using the MONDO ontology crosswalk in order to identify the diseases in the rare disease knowledge graph (Section 1).

Dataset statistics The final UDN cohort contains 465 patients representing 319 MONDO diseases and 378 unique causal genes. The EXPERT-CURATED and VARIANT-FILTERED candidate gene lists contain 244.3 and 13.3 genes on average, respectively (SD = 244.0 and SD = 8.0). Patients have 23.9 phenotypes on average (SD = 16.1).

2.2 Patients in the MyGene2 Portal

We assemble another cohort of real-world rare disease patients participating in the MyGene2 exchange [81]. MyGene2, developed by the University of Washington, is a portal through which families with rare genetic conditions can share their health information to connect with other families, clinicians, and researchers. MyGene2 contains information about 2,106 genes and the phenotypes of patients with mutations in the genes. MyGene2 is a member of the MatchMaker Exchange, a federated network designed to enable clinicians to find phenotype and genotype matches for rare disease patients [82]. The UDN leverages the MatchMaker exchange for validating patients' candidate genes by finding genotype-matched individuals.

Dataset preprocessing and patient subgraph construction We retrieved data containing the sets of phenotype terms and candidate genes for rare disease patients on MyGene2 as of May 7, 2022. We filter the patients to only include patients labeled with an OMIM disease identifier. This limits the cohort to patients that are likely already diagnosed. As with the other cohorts, we map all genes to Ensembl identifiers, diseases to MONDO identifiers, and construct patient subgraphs from the set of positive HPO terms associated with each patient.

Dataset statistics The final MyGene2 cohort contains 146 patients representing 55 MONDO diseases and 48 unique causal genes. Patients have 7.9 phenotypes on average (SD = 6.6). There are 14 unique causal genes and 12 diseases found in both the MyGene2 and UDN cohorts.

2.3 Simulated Patients with Rare Mendelian Disorders

We leverage simulated but realistic rare disease patients for training SHEPHERD [19]. The simulated patients closely resemble real-world patients found in the UDN. Each simulated patient is represented by an age range, a set of positive phenotypes they exhibit, a set of negative phenotypes they do not exhibit, and a set of challenging candidate genes that may cause the presenting symptoms. The patients are generated using a simulation framework that jointly samples candidate genes and phenotypes.

Generating realistic simulated rare disease patients To generate patients with rare Mendelian disorders, we adopt the pipeline described in [19]. Briefly, the simulation pipeline has two stages: phenotype generation and candidate gene generation. First, each patient is initialized with a set of phenotypes associated with a genetic disorder characterized in the rare genetic disease database Or-

phanet [67]. To reflect the imprecision of real-world diagnostic evaluations, the initial phenotypes undergo phenotype dropout and corruption (i.e. phenotypes are randomly removed or replaced with more general phenotype terms), and additional "noisy" phenotypes that are unrelated to the patient's disease are sampled from a large medical insurance claims database and added to the phenotype set. Next, candidate genes are sampled from "distractor" gene categories that do not cause the patient's disease yet would be plausible candidates during the diagnostic process. The challenging distractor genes and some of their associated phenotype terms are added. For additional details about the simulation process and validation of simulated patients, refer to [19]. To standardize across all patient cohorts, we ensure all genes are mapped to Ensembl identifiers, all diseases are mapped to MONDO identifiers, and we construct patient subgraphs from the phenotype terms associated with each patient.

Dataset statistics There are 42,624 simulated patients representing 2,132 unique Mendelian disorders and 2,396 unique causal genes in the simulated patient dataset. Each patient is characterized by an average of 18.4 positive phenotypes (SD = 7.7) and 14.0 candidate genes (SD = 3.5). Of the 378 unique causal genes and 319 unique MONDO diseases found in patients in the UDN cohort, 220 and 109 are represented in the simulated patient cohort, respectively. Furthermore, 81.8% of the phenotypes found across UDN patients are also found in the simulated patient cohort, and 29.7% of a single UDN patient's phenotypes are found in the most similar simulated patient on average. This indicates that the simulated patients have utility in training models that can apply to real-world UDN patients but also emphasizes the need for developing models that can generalize to genes, diseases, and phenotypes unseen at train time.

3 Few-shot Learning Framework for Rare Disease Diagnosis

We develop SHEPHERD, a geometric deep learning approach that uses few-shot capability and external biomedical knowledge for multi-faceted diagnosis of rare diseases. SHEPHERD learns to embed diseases, phenotypes, and genes and generate multi-modal representations of rare disease patients. It performs multi-faceted diagnosis, addressing the following challenges of rare disease diagnosis:

- **Causal Gene Discovery:** Each patient T_i in the dataset has P_i phenotypes and G_i candidate genes. The task is to identify the causal gene(s) $G_i^c \in G_i$ harboring the variants that explain the patient's presenting symptoms.

- **Identification of Similar Patients:** Given a cohort of rare disease patients C, the goal is to identify patients from the cohort that are similar to the query patient T_i, *i.e.*, patients that share a disease or causal gene. Mathematically, for each patient T_i, the task is to identify the set of patients $S_i^c = \{T_j \in C | G_i^c \cap G_j^c \neq \emptyset\}$. We leverage each patient's set of phenotypes P_i to perform patient matching.

- **Characterization of Novel Diseases:** The goal is to characterize novel diseases according to their similarity to a set of known genetic diseases D. We input the set of phenotypes P_i for each patient T_i and provide interpretable names for the patient's presenting syndrome.

Notation Let denote a heterogeneous knowledge graph comprised of a set of nodes V and a set of edges E. Each edge is defined by a triplet (u, r, v) where u is the source node, v is the target node, and $r \in R$ denotes the relationship between u and v. Each patient i is represented on the graph as a patient subgraph induced by a set of phenotype nodes P_i where $P_i \subseteq V$. The patient subgraph can contain any number of phenotypes and can consist of multiple connected components throughout $\mathcal{G}$. Each patient may also have a set of candidate genes $G_i \subseteq V$.

3.1 Encoding Biomedical Knowledge

The first step in SHEPHERD is to encode biomedical relationships found in the rare disease knowledge graph (KG). We pretrain SHEPHERD on millions of biomedical entity pairs across all entity and relation types in the KG to capture the topology of biomedical knowledge in the KG. To this end, we use a graph attention network [83], a type of graph neural network (GNN) model to generate embeddings $\mathbf{x}_v$ for every node v in the KG. Specifically, the choice of a graph attention network is necessary to achieve semantically-relevant mixing of biomedical entities in the embedding space, that is, to encourage distinct node types (*e.g.*, genes, diseases, and phenotypes) to be positioned near each other in the embedding space. GAT models, like most GNNs, can be formulated as message-passing networks, in which messages are propagated to a node v from all of the nodes in its neighborhood $\mathcal{N}_v$. The messages are aggregated and combined with the previous layer's representation of v to produce v's representation for the current layer. Concretely, each layer l in SHEPHERD's GNN encoder involves the following steps:

Step 1: Propagate neural messages: We define the message $\mathbf{m}_{v,k}^{(l)}$ for each node v as:

$$\mathbf{m}_{v,k}^{(l)} = \mathbf{W}_k^{(l)} \mathbf{h}_v^{(l-1)} \tag{1}$$

where k represents the attention head, $\mathbf{W}$ is a relation-specific trainable weight matrix, and $\mathbf{h}_v$ is the embedding of node v in the $(l-1)$-th hidden layer.

Step 2: Aggregate messages from local neighborhoods: We leverage the local neighborhood to generate a representation of each node v. Specifically, we aggregate messages of its neighboring nodes $u \in \mathcal{N}_v$ using an attention mechanism to generate $\mathbf{a}_{v,k}^{(l)}$:

$$\mathbf{a}_{v,k}^{(l)} = \sum_{u \in \mathcal{N}_v} \alpha_{v,u,k}^{(l)} \cdot \mathbf{m}_{u,k}^{(l)} \tag{2}$$

where $\alpha_{v,u,k}$ is the normalized attention weight on an edge from node v to node u computed by the k-th attention mechanism.

Step 3: Update node embeddings: To transform the messages into an order-invariant hidden representation $\mathbf{h}_v^{(l)}$, we apply a nonlinearity function σ and concatenate all of the aggregated messages as follows:

$$\mathbf{h}_v^{(l)} = \Big\Vert_{k=1}^{K} \sigma\left(\mathbf{a}_{v,k}^{(l)}\right) \tag{3}$$

In the final layer, we perform averaging instead of concatenation. We define the final embedding for each node v after L layers of neural message passing as $\mathbf{x}_v = \mathbf{h}_v^{(L)}$.

Objective function We frame pretraining as a binary classification task. SHEPHERD learns to perform link prediction, *i.e.*, predict whether a relationship exists between a pair of nodes for a given relation type. Formally, we compute the likelihood of an edge existing between node u and node v with relation r given their node embeddings $\mathbf{z_u}$ and $\mathbf{z_v}$ using a DistMult decoder [84]:

$$\text{LPSIM}(u, r, v) = \text{ACT}(\mathbf{x}_u^T \mathbf{W}_r \mathbf{x}_v) \tag{4}$$

where $\mathbf{W}_r$ is a relation-specific trainable weight matrix and ACT is a nonlinear function, here tanh. SHEPHERD is pretrained via a hinge loss objective. For any pair of nodes u and v connected by relation r, the loss function is defined as:

$$L_{\text{LP}} = \frac{1}{|E|} \sum_{(u,r,v) \in E} \max(0, \Delta - \text{LPSIM}(u, r, v) + \text{LPSIM}(u, r, v^-)), \tag{5}$$

where u and v are source and target nodes, v^- is a target node, representing a negative example that is not linked to u in the KG, LPSIM returns the score indicative of the knowledge relationship

existing between u and v, and Δ is a margin, which is set to 1 throughout all experiments in this study. For each triplet (u, r, v) in the KG, its contribution to the value of the loss function is 0 if the difference between the LPSIM 's score for the triplet and the LPSIM 's score for a negative example is at least as large as the margin.

3.2 Generating Rare Disease Patient Representations

We apply the pretrained SHEPHERD model to our multi-faceted rare disease diagnosis tasks. Starting with the pretrained GNN model, we learn patient embeddings that encode each patient's phenotype subgraph. Depending on the diagnostic task, we also learn embeddings for each patient's candidate genes, diseases, or other patients. Concretely, for every patient T_i, we generate an aggregated representation of all phenotypes $p \in P_i$ in the phenotype subgraph via an attention-weighted average of the individual phenotype embeddings:

$$\mathbf{x}_{P_i} = \sum_{p \in P_i} \alpha \cdot \mathbf{x}_p, \quad \text{where} \quad \alpha = \frac{\exp(\mathbf{x}_p \cdot \mathbf{a})}{\sum_{p \in P_i} \exp(\mathbf{x}_p \cdot \mathbf{a})} \tag{6}$$

where α denotes the attention weights, $\mathbf{x}_p$ denotes the embedding for phenotype p, and $\mathbf{a}$ is a trainable vector initialized via Xavier [85]. The aggregated phenotype representation $\mathbf{x}_{P_i}$, each candidate gene node embedding $\mathbf{x}_g$, and each candidate disease node embedding $\mathbf{x}_d$ are pushed through two nonlinear layers to produce the embeddings $\mathbf{z}_{P_i}$, $\mathbf{z}_g$, and $\mathbf{z}_d$, respectively, as:

$$\mathbf{z}_{P_i} = f(f(\mathbf{x}_{P_i} \cdot \mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2) \tag{7}$$

$$\mathbf{z}_g = f(f(\mathbf{x}_g \cdot \mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2) \tag{8}$$

$$\mathbf{z}_d = f(f(\mathbf{x}_d \cdot \mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2) \tag{9}$$

where f is a nonlinear function (here, leaky ReLU), and $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are trainable weights.

3.3 Discovering Causal Genes

SHEPHERD can prioritize candidate genes to assist clinicians in finding the causal gene(s) harboring the variants that best explain a patient's presenting symptoms. Candidate genes for each patient are scored by measuring the similarity $\text{SIM}(P, g)$ between a candidate gene g and a patient's set of phenotypes P. SHEPHERD is optimized such that the candidate gene embedded nearby the patient's set of phenotypes in the embedding space indicates that the gene is likely to cause the patient's

symptoms. SIM(P, g) consists of two components, EMBSIM(P, g) and SPLSIM(P, g). It is designed such that EMBSIM(P, g) captures global network topology (i.e., by leveraging SHEPHERD's rich low-dimensional embedding space) and SPLSIM(P, g) captures local network information (i.e., by calculating shortest path length distances). This approach is grounded in the observation that while methods that learn global network topology yield higher overall performance than local methods considering only local network information, the latter tends to rank true candidate genes higher when provided a short list of candidate genes [86].

Embedding-based similarity We calculate EMBSIM, an embedding-based similarity between aggregated embeddings of phenotypes P and an embedding of the candidate gene g as follows:

$$\text{EMBSIM}(P, g) = \text{ACT}(\mathbf{z}_P^T \mathbf{W} \mathbf{z}_g) \tag{10}$$

where ACT is a nonlinear function, here $tanh(x)$. EMBSIM values range between $[-1, 1]$.

Network-based similarity We calculate the shortest path length (SPL) similarity between aggregated phenotypes P and candidate gene g as follows:

$$\text{SPLSIM}(P, g) = \text{NORM}(\text{AGG}_{p \in P}(-d(p, g))) \tag{11}$$

where P is the patient's phenotypes and g is a candidate gene, AGG is some aggregation function (e.g., mean), $\text{NORM}(\mathbf{x}) = \frac{2(\mathbf{x} - \max(\mathbf{x}))}{\max(\mathbf{x}) - \min(\mathbf{x})} - 1$ is a normalization function to scale the values in the range $[-1, 1]$, and $d(p, g)$ is the minimum number of hops between p and g in the KG.

Overall similarity The final score between a patient's phenotypes P and candidate gene g is defined as:

$$\text{SIM}(P, g) = \eta \cdot \text{EMBSIM}(P, g) + (1 - \eta) \cdot \text{SPLSIM}(P, g) \tag{12}$$

where η is a hyperparameter ranging from $[0, 1]$ that represents the amount of weight to place on EMBSIM versus SPLSIM in the final gene prioritization scoring. SIM values range between $[-1, 1]$.

Objective function We leverage a multi-similarity loss to encourage patient phenotype embeddings to be near to their causal gene embedding and far away from the incorrect candidate gene

embeddings. The multi-similarity loss is defined as follows [87]:

$$L_{\text{G}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\alpha}\log(1+\sum_{g\in G_i^c}\exp(-\alpha(\text{SIM}(P_i,g)-\gamma))) + \frac{1}{\beta}\log(1+\sum_{g\in G_i^d}\exp(\beta(\text{SIM}(P_i,g)-\gamma)))\right),$$

where N is the number of patients, α, β, and γ are hyperparameters, and $\text{SIM}(P_i, g)$ denotes the similarity between the aggregated phenotype embedding for patient i and the gene embedding of either the patient's causal gene ($g \in G^c$) or distractor gene ($g \in G^d$) (Section 3.3). The optimized embedding space encodes patient information such that similarity between a patient's phenotypes and candidate genes – *i.e.*, how likely it is that a given gene explains the patient's symptoms – is inversely proportional to the distance between the patient embedding and the embedding of the candidate gene.

3.4 Finding Similar Patients

SHEPHERD can find similar patients from a cohort of rare disease patients. This is important for identifying molecular diagnoses and validating already prioritized candidate genes. To match rare disease patients, SHEPHERD generates an embedding space in which the similarity between two patients is inversely proportional to the distance between the two patient embeddings and uses the embedding space to answer "patients-like-me" queries. We define the similarity between two patients i and j as the L2 distance between their aggregated phenotype embeddings $\mathbf{z}_{P_i}$ and $\mathbf{z}_{P_j}$:

$$\text{SIM}(P_i, P_j) = -\|\mathbf{z}_{P_i} - \mathbf{z}_{P_j}\|_2^2 \tag{13}$$

When calculating patient similarity, importantly, we do not include any genotype information for the patients. This makes the model applicable in settings where the patient's genome has not been sequenced or when the analysis results are still pending.

Objective function SHEPHERD is trained to capture patient similarity using the neighborhood component analysis loss:

$$L_{\text{PH}} = \frac{-1}{|B_p|}\sum_{P_i\in B_p}\log\left(\frac{\sum_{P_j\in B_p\setminus P_i; P_j\in S_i^c}\exp(\text{SIM}(P_i,P_j))}{\sum_{P_j\in B_p\setminus P_i}\exp(\text{SIM}(P_i,P_j))}\right), \tag{14}$$

where B_p is a batch of patients sampled from the training set and S_i^c is the set of patients with the same causal gene as patient P_i. Optimizing the NCA loss [88] minimizes the distances between patient embeddings with the same causal gene and maximizes the distances between patient embeddings with different causal genes.

3.5 Estimating Patient-Disease Similarity

Finally, SHEPHERD can characterize a clinical presentation based on existing knowledge about other rare and common diseases. We analogously perform novel disease characterization by learning an embedding space such that the similarity between a patient and a disease - *i.e.*, how likely it is that a patient has that disease – is inversely proportional to the distance between the patient embedding and the disease embedding. We define the similarity between a patient's phenotypes P and disease d as the L2 distance between the aggregated phenotype embedding and the disease embedding:

$$\text{SIM}(P, d) = -\|\mathbf{z}_d - \mathbf{z}_P\|_2^2 \tag{15}$$

Objective function To optimize patient phenotype embeddings to be near their correct disease(s), we leverage a multi-modal version of the NCA loss, defined as:

$$L_{\text{D}} = \frac{-1}{|B_p|} \sum_{P_i \in B_p} \log \left(\frac{\sum\limits_{d_j \in B_d; d_j \in D_i^c} \exp(\text{SIM}(P_i, d_j))}{\sum\limits_{d_j \in B_d} \exp(\text{SIM}(P_i, d_j))} \right) \tag{16}$$

where B_p and B_d are batches of patients and candidate diseases, respectively, that are sampled from the training set, P_i corresponds to the phenotype set for patient i, and D_i^c is the set of correct diseases for patient i. While $|D_i^c| = 1$ for most patients in our cohorts, several patients with multiple diseases exist.

4 Training SHEPHERD Models

We first describe our approach for training SHEPHERD to perform multi-faceted diagnoses. We provide details about negative sampling strategies, patient-driven sampling, and disease-split training on simulated patient data. We conclude with details regarding hyperparameter tuning and implementation in Pytorch.

4.1 Overall Objective Function

We train SHEPHERD in two stages. First, we train the model to learn to capture the relationships between biomedical entities in the rare disease knowledge graph (Section 3.1). Then, we simultaneously train the model to perform patient-centric rare disease tasks and continue predicting knowledge relationships in the KG (Section 3.2-3.4). Concretely, the model is jointly trained to achieve two distinct objectives: (1) to capture the relationships in the underlying knowledge graph and (2) to match a patient's presenting symptoms with the patient's causal gene(s), disease(s), or other similar patients. We model these objectives with two separate loss functions, the pretraining link prediction loss, L_{LP}, and a diagnosis loss, $L_{\text{DX}} \in \{L_{\text{G}}, L_{\text{D}}, L_{\text{PH}}\}$, which aligns patient phenotypes to genes, diseases, or other patient phenotypes respectively. The overall loss is as follows:

$$L = \lambda L_{\text{DX}} + (1 - \lambda) L_{\text{LP}} \tag{17}$$

where λ is a hyperparameter controlling the weight of each loss. Whereas during pretraining, we train the model to capture generalizable biomedical knowledge by performing link prediction for all relation types, during fine-tuning, we focus on predicting gene, phenotype, and disease relations, which are most important for rare diseases. Training the model to perform link prediction is important for enabling the model to generalize to phenotypes and genes unseen in the training data.

4.2 Negative Sampling

To learn a meaningful representation space, we need negative examples, *i.e.*, edges that do not exist in the KG or candidate genes, diseases, or other patients that are not associated with a given patient. In the following, we outline the negative sampling strategies used for pretraining and each of the three rare disease diagnosis tasks.

Link prediction We construct negative examples of triplets (u, r, v^-) that do not exist in the KG by perturbing the target *nodes* while preserving the *types* of the source and target nodes and edge relation. For example, given a positive example of a triplet where the node and relation types are (*protein*, *has phenotype*, *phenotype*), a negative example is obtained by shuffling all phenotype *nodes* in the batch, thereby maintaining the node and relation *types* of the positive example.

Causal gene discovery Negative examples are constructed by taking the union of the candidate

genes for all patients in a given batch. As noted in Section 2.3 and 2.1, each simulated and UDN patient has a list of candidate genes that have been shortlisted as the most probable genes to cause the patient's symptoms, and identifying the true causal gene(s) among them is especially challenging. We ensure that these "hard" candidate genes are included in the candidate list for each patient during training, as using such "hard" examples tends to improve the efficiency of training [89]. Furthermore, to maximize the number and frequency of candidate genes seen during train time, we up-sample a subset of candidate genes that are under-represented across all patients. Concretely, we count the frequencies of candidate genes in the prior and current batches, select the k most infrequently seen candidate genes (*i.e.*, the k rarest candidate genes) in training batches, and add them to each patient's candidate gene list. Note that we only prioritize the "hard" candidate genes for each patient at inference time without any up-sampling.

Novel disease characterization Negative examples consist of all diseases that do not explain the patient's presenting symptoms. First, we randomly sample 1,000 diseases from all diseases in the KG to serve as negative examples for each batch. Then, we calculate a patient's similarity to all disease nodes in the KG at inference time.

Identifying similar patients Negative examples are simply all of the patients in the batch who do not have the same causal gene as the patient. We construct batches to ensure that there are at least two positive examples (*i.e.*, patients with the same gene) for each patient in the batch. All remaining patients serve as negative examples. At inference time, we calculate a patient's similarity to all patients in the cohort.

4.3 Disease-Split Training on Simulated Patients

We train our model exclusively on the simulated patient dataset. Training on simulated data alone offers several benefits: the simulated cohort is larger and more diverse than any real-world patient dataset, the trained models can be released without the risk of exposing any patient information, and the models can be evaluated on an independent real-world cohort to test how well a model can generalize to patients unseen during training. Further, and most importantly, we achieve generalizability to real-world cohorts by splitting patients into train and validation sets according to disease. Concretely, we first split the list of diseases represented by the simulated patient cohort into train and validation and then assign patients to training or validation sets such that patients with the same disease are either entirely in the training set or fully in the validation set. As a result,

the model is optimized such that its parameters are broadly transferable to patients with different diseases. The resulting train and validation cohorts contain 36,224 and 6,400 patients, respectively.

4.4 Additional Training Details

Patient-driven sampling We design a new approach for batch sampling that enables the model to perform patient gene prioritization while maintaining the topology of the KG. We first sample m patients and add their associated phenotypes and genes to the batch. Then, we add n nodes randomly sampled from the genes, phenotypes, and disease nodes in the KG. This allows for inductive generalization by maintaining the topology of nodes not found in the training data.

Normalization To help optimize model performance and convergence, we apply two normalization strategies to SHEPHERD. Specifically, we use LayerNorm [90] immediately after each convolutional layer and BatchNorm [91] following a nonlinear activation layer (here, leaky ReLU).

Hyperparameter tuning We leverage Weights and Biases [92] to select optimal hyperparameters via a random search over the hyperparameter space. We first select pretraining hyperparameters to optimize the micro F1 score on the pretraining validation set. Hyperparameters were selected via random search from the following values: learning rate $\in [0.0001, 0.0005, 0.001, 0.005]$, weight decay $\in [0, 0.005, 0.0005]$, dropout $\in [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$, and the number of GAT attention heads $\in [2, 4]$. We also perform a search over the dimension of the network layers: input size $\in [2048, 4096]$, hidden size $\in [256, 512, 1024]$, and output size $\in [64, 128]$. We then freeze the pretraining hyperparameters and perform a hyperparameter search independently for each rare disease task. We select task-specific hyperparameters to optimize the mean reciprocal rank of the correct genes, diseases, or patients on the disease-split simulated validation set. We consider the following hyperparameters: learning rate $\in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$, $\lambda \in [0.1, 0.9]$, $\eta \in [0.1, 0.9]$, k most infrequently seen genes $\in \{64, 128, 192\}$, and number of nodes n to sample per batch in $\in \{100, 200, 300, 400\}$. The code for hyperparameter selection and the optimal hyperparameters can be found at https://github.com/mims-harvard/SHEPHERD.

Implementation We implement SHEPHERD using Pytorch (Version 1.8.0) [93], Pytorch Lightning (Version 1.4.5) [94], and Pytorch Geometric (Version 1.7.2) [95]. We leverage Weights and Biases [92] for hyperparameter tuning and model training visualization, and we create interactive demos

of the model using Gradio [96]. Models are trained on a single NVIDIA Quadro RTX8000 GPU.

5 Further Details on Statistical Analysis

We outline the evaluation setup, baseline models, and statistical tests used to evaluate SHEPHERD.

5.1 Performance Stratified by Patient and Site Characteristics

We evaluate the trained model on the cohort of real-world UDN patients who have received a molecular diagnosis (Section 2.1). We measure the mean reciprocal rank of all of the patients' causal genes and calculate the percentage of causal genes that appear in the top k ranked genes for $k \in \{1, 3, 5\}$ for the EXPERT-CURATED candidate gene lists and $k \in \{1, 5, 10, 25, 50\}$ for the longer VARIANT-FILTERED candidate gene lists. We analyze the performance across each of the UDN clinical sites, disease categories, evaluation years, and diagnostic certainty, *i.e.*, the likelihood that the gene causes the patient's symptom.

5.2 Comparison to Alternative Approaches

We compare SHEPHERD to several diverse approaches for causal gene discovery: (1) NETWORK, a network-science approach that prioritizes genes according to their average shortest path in the KG to all of a patient's phenotypes; (2) INFORMATION THEORETIC, an information theory inspired approach that leverages a Bayesian network to calculate semantic similarity between sets of phenotype terms to prioritize genes or diseases [44]; (3) LR (EMBED) a logistic regression approach that frames prioritization as a binary prediction task for each candidate gene and represents each patient-gene option as the concatenation of the candidate gene pretrained node embedding and the patient's averaged phenotype node embeddings; (4) LR (PCA) a logistic regression approach similar to (3) that, instead of the KG node embeddings, utilizes a PCA-transformed shortest path length matrix from gene nodes to gene, phenotype, and disease nodes; (5) SHALLOW EMBEDDING, a shallow KG embedding approach that uses Node2Vec to learn node embeddings and ranks genes by computing the score associated with each individual patient phenotype [46]; (6) RANDOM, a simple shuffling approach of a gene list. These baselines constitute a diverse set of statistical, network, and machine-learning approaches for rare disease diagnosis.

We further compare SHEPHERD to other approaches that can be used to identify similar patients. SET BASED calculates distance between two sets of phenotypes P_i and P_j using Jaccard distance, defined as $J = 1 - \frac{|P_i \cap P_j|}{|P_i \cup P_j|}$. INFORMATION THEORETIC is the information theory approach

from above that calculates the semantic similarity between two sets of phenotypes based on the information content of their shared phenotype ancestors in the Human Phenotype Ontology [44].

5.3 Assessing Statistical Significance

We perform a one-sided Wilcoxon signed-rank test to assess whether there is a significant difference in causal gene performance between SHEPHERD and baseline methods. We evaluate whether there is a statistically significant difference in SHEPHERD's performance across sites, evaluation years, and primary presenting symptoms using a Kruskal-Wallis H-test after confirming that the data is not normally distributed. We calculate the Spearman correlation coefficient to measure the correlation between causal gene rank and the distance between a patient's phenotype and causal gene in the knowledge graph. To assess whether patients cluster by disease category, we perform K-means clustering with k set to the number of disease categories, and we evaluate the clusters according to an adjusted mutual information score from scikit-learn, which is designed to evaluate clusters of different sizes. We assess the significance of the resulting clustering via a permutation test with 100 random permutations of the true cluster labels. We perform a Mann-Whitney test to measure the difference in distances in embedding space for patients with the same versus different disease categories. Finally, we perform the two-sample Kolmogorov-Smirnov test to assess whether the distribution of embedding distances for patients with the same disease is identical to the distribution for patients with different diseases.

5.4 Visualization of Learned Embeddings

We visualize embeddings learned via SHEPHERD in a Uniform Manifold Approximation and Projection (UMAP) plot [97]. We use the `umap-learn` Python package [98] and perform a grid search over the `n_neighbors`, `min_dist`, and `spread` UMAP parameters. We select parameters that maintain global structure in the main panel of Figure 5a and Figure 6a.

5.5 Visualization of patient neighborhoods in the knowledge graph

To visualize the local neighborhood of patients' disease, phenotype, and gene nodes (Figure 4), we calculate the shortest paths between patient-relevant nodes and extract all nodes in those shortest paths. We visualize the resulting patient neighborhoods using Gephi 0.9.4 [99]. We apply Fruchterman Reingold, Noverlap, and Label Adjust layouts, as well as manual adjustment, to organize the nodes such that they are non-overlapping.

References

1. Haendel, M. *et al.* How many rare diseases are there? *Nature Reviews Drug Discovery* (2020).

2. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics* (2020).

3. Whicher, D., Philbin, S. & Aronson, N. An overview of the impact of rare disease characteristics on research methodology. *Orphanet Journal of Rare Diseases* (2018).

4. Gahl, W. A. *et al.* The NIH Undiagnosed Diseases Program: Insights into Rare Diseases. *Genetics in Medicine* (2012).

5. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American Journal of Human Genetics* (2015).

6. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* (2019).

7. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nature Biomedical Engineering* (2018).

8. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nature Medicine* (2020).

9. Saldanha, O. L. *et al.* Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nature Medicine* (2022).

10. Bulten, W. *et al.* Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine* (2022).

11. Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications* (2020).

12. Tang, A. S. *et al.* Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. *Nature Communications* (2022).

13. Qiu, S. *et al.* Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications* (2022).

14. Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. *Nature Medicine* (2020).

15. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* (2018).

16. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* (2016).

17. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* (2017).

18. Liang, H. *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine* (2019).

19. Alsentzer, E., Finlayson, S. G., Li, M. M., Kobren, S. N. & Kohane, I. S. Simulation of undiagnosed patients with novel genetic conditions. *medRxiv* (2022).

20. Ramoni, R. B. *et al.* The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *The American Journal of Human Genetics* (2017).

21. Kobren, S. N. *et al.* Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. *Genetics in Medicine* (2021).

22. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine* (2014).

23. Smedley, D. *et al.* Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature Protocols* (2015).

24. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* (2021).

25. Wicks, P. *et al.* Sharing health data for better outcomes on patientslikeme. *Journal of Medical Internet research* (2010).

26. Philippakis, A. A. *et al.* The matchmaker exchange: a platform for rare disease gene discovery. *Human Mutation* (2015).

27. Gerarduzzi, C. *et al.* Silencing SMOC2 ameliorates kidney fibrosis by inhibiting fibroblast to myofibroblast transformation. *JCI Insight* (2017).

28. Morkmued, S. *et al.* Deficiency of the SMOC2 matricellular protein impairs bone healing and produces age-dependent bone loss. *Scientific Reports* (2020).

29. Romio, L. *et al.* *OFD1* , the Gene Mutated in Oral-Facial-Digital Syndrome Type 1, Is Expressed in the Metanephros and in Human Embryonic Renal Mesenchymal Cells. *Journal of the American Society of Nephrology* (2003).

30. Saal, S. *et al.* Renal insufficiency, a frequent complication with age in oral-facial-digital syndrome type I. *Clinical Genetics* (2010).

31. Ganapathi, M. *et al.* A homozygous splice variant in atp5po, disrupts mitochondrial complex v function and causes leigh syndrome in two unrelated families. *Journal of Inherited Metabolic Disease* (2022).

32. Chen, H., Morris, M. A., Rossier, C., Blouin, J.-L. & Antonarakis, S. E. Cloning of the cdna for the human atp synthase oscp subunit (atp50) by exon trapping and mapping to chromosome 21q22. 1-q22. 2. *Genomics* (1995).

33. Aggeler, R. *et al.* A functionally active human f1f0 atpase can be purified by immunocapture from heart tissue and fibroblast cell lines: subunit structure and activity studies. *Journal of Biological Chemistry* (2002).

34. Brautigam, C. A., Wynn, R. M., Chuang, J. L. & Chuang, D. T. Subunit and catalytic component stoichiometries of an in vitro reconstituted human pyruvate dehydrogenase complex. *Journal of Biological Chemistry* (2009).

35. Jiang, Y. *et al.* Component co-expression and purification of recombinant human pyruvate dehydrogenase complex from baculovirus infected sf9 cells. *Protein Expression and Purification* (2014).

36. Glasgow, R. I. *et al.* Novel gfm2 variants associated with early-onset neurological presentations of mitochondrial disease and impaired expression of oxphos subunits. *Neurogenetics* (2017).

37. Warde-Farley, D. *et al.* The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research* (2010).

38. Franz, M. *et al.* Genemania update 2018. *Nucleic Acids Research* (2018).

39. Ispolatov, I., Yuryev, A., Mazo, I. & Maslov, S. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Research* **33**, 3629–3635 (2005).

40. Keskin, O., Tuncbag, N. & Gursoy, A. Predicting protein–protein interactions from the molecular to the proteome level. *Chemical Reviews* **116**, 4884–4909 (2016).

41. Zitnik, M., Sosič, R., Feldman, M. W. & Leskovec, J. Evolution of resilience in protein interactomes across the tree of life. *Proceedings of the National Academy of Sciences* **116**, 4426–4433 (2019).

42. Westermarck, J., Ivaska, J. & Corthals, G. L. Identification of protein interactions involved in cellular signaling. *Molecular & Cellular Proteomics* (2013).

43. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* (2020).

44. Jagadeesh, K. A. *et al.* Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genetics in Medicine* (2019).

45. Köhler, S. *et al.* Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *American Journal of Human Genetics* (2009).

46. Peng, C. *et al.* CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genomics and Bioinformatics* (2021).

47. Tyler, A. L., Asselbergs, F. W., Williams, S. M. & Moore, J. H. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays* (2009).

48. Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics* (2016).

49. Ried, J. S. *et al.* PSEA: Phenotype Set Enrichment Analysis—A New Method for Analysis of Multiple Phenotypes. *Genetic Epidemiology* (2012).

50. Li, J., Cairns, B. J., Li, J. & Zhu, T. Generating Synthetic Mixed-type Longitudinal Electronic Health Records for Artificial Intelligent Applications (2021).

51. Mahmood, F. *et al.* Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *IEEE Transactions on Medical Imaging* (2020).

52. Waheed, A. *et al.* CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access* (2020).

53. Jaipuria, N. *et al.* Deflating Dataset Bias Using Synthetic Data Augmentation. In *CVPR* (2020).

54. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J. & Greenspan, H. Synthetic data augmentation using GAN for improved liver lesion classification. In *ISBI* (2018).

55. Oprisanu, B., Ganev, G. & De Cristofaro, E. On Utility and Privacy in Synthetic Genomic Data. *arXiv:2102.03314* (2022).

56. Wang, Z., Myles, P. & Tucker, A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence* (2021).

57. Wang, J. *et al.* MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome. *The American Journal of Human Genetics* (2017).

58. Birgmeier, J. *et al.* AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Science Translational Medicine* (2020).

59. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *bioRxiv* (2022).

60. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Medicine* (2022).

61. Consortium, G. O. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Research* (2019).

62. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Research* (2020).

63. Piñero, J. *et al.* The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2020).

64. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research* (2005).

65. Köhler, S. *et al.* Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic Acids Research* (2019).

66. Vasilevsky, N. *et al.* Mondo disease ontology: harmonizing disease concepts across the world. In *CEUR-WS* (2020).

67. Pavan, S. *et al.* Clinical practice guidelines for rare diseases: the orphanet database. *PloS One* (2017).

68. Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K. & Kivelä, M. Cumulative effects of triadic closure and homophily in social networks. *Science Advances* (2020).

69. Kovács, I. A. *et al.* Network-based prediction of protein interactions. *Nature Communications* (2019).

70. Aken, B. L. *et al.* The ensembl gene annotation system. *Database* (2016).

71. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* (2015).

72. Oughtred, R. *et al.* The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* (2021).

73. Szklarczyk, D. *et al.* The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* (2021).

74. Gahl, W. A., Wise, A. L. & Ashley, E. A. The undiagnosed diseases network of the national institutes of health: a national extension. *JAMA* (2015).

75. Girdea, M. *et al.* Phenotips: Patient phenotyping software for clinical and research use. *Human Mutation* (2013).

76. Splinter, K. *et al.* Effect of genetic diagnosis on patients with previously undiagnosed disease. *New England Journal of Medicine* (2018).

77. Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* (2015).

78. UDN Manual of Operations (2022).

79. Robinson, P. N. *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research* (2014).

80. Hamosh, A. *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* (2002).

81. Genomics, U. o. W. C. f. M. MyGene2.

82. Philippakis, A. A. *et al.* The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Human Mutation* (2015).

83. Brody, S., Alon, U. & Yahav, E. How attentive are graph attention networks? (2021).

84. Yang, B., Yih, W.-t., He, X., Gao, J. & Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. *arXiv:1412.6575* (2015).

85. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS* (2010).

86. Zolotareva, O. & Kleine, M. A survey of gene prioritization tools for mendelian and complex human diseases. *Journal of Integrative Bioinformatics* (2019).

87. Wang, X., Han, X., Huang, W., Dong, D. & Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR* (2019).

88. Goldberger, J., Hinton, G. E., Roweis, S. & Salakhutdinov, R. R. Neighbourhood Components Analysis. In *NeurIPS* (2004).

89. Zhao, Z.-Q., Zheng, P., Xu, S.-t. & Wu, X. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems* (2019).

90. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. *arXiv:1607.06450* (2016).

91. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML* (2015).

92. Biewald, L. Experiment tracking with weights and biases (2020).

93. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS* (2019).

94. Falcon, W. & The PyTorch Lightning team. PyTorch Lightning (2019).

95. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).

96. Abid, A. *et al.* Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv:1906.02569* (2019).

97. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2020).

98. McInnes, L. Outlier detection using UMAP — umap 0.5 documentation (2018).

99. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, vol. 3, 361–362 (2009).

Supplementary Information for

Deep learning for diagnosing patients with rare genetic diseases

Emily Alsentzer[1,2,*], Michelle M. Li[1,3,*], Shilpa N. Kobren[1], Undiagnosed Diseases Network, Isaac S. Kohane[1], and Marinka Zitnik[1,‡]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

[2]Program in Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA 02115, USA

‡Corresponding author. Email: marinka@hms.harvard.edu

*Equal contribution

Supplementary Figures

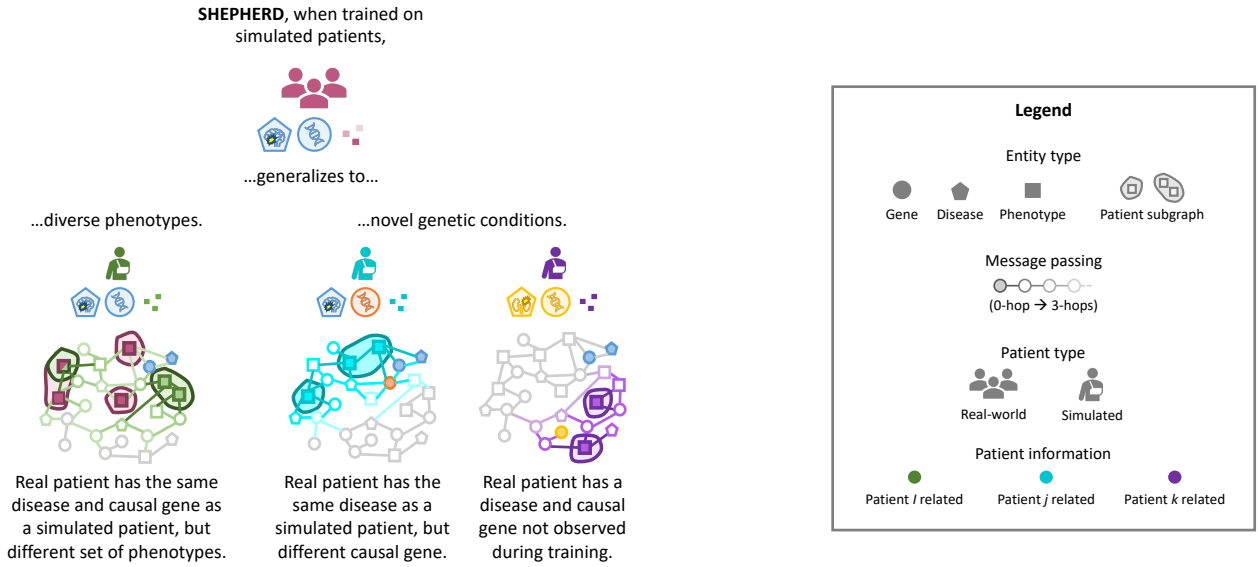


Figure S1: SHEPHERD can generalize to heterogeneous phenotypic presentations and novel genetic conditions. There are few patients with each rare disease, and patients with the same disease can have variable clinical presentations. SHEPHERD is trained on simulated rare disease patients and can generalize to real-world patients with unique, unseen phenotypes (left), with novel disease-causing genes (center), and with entirely novel diseases (right).

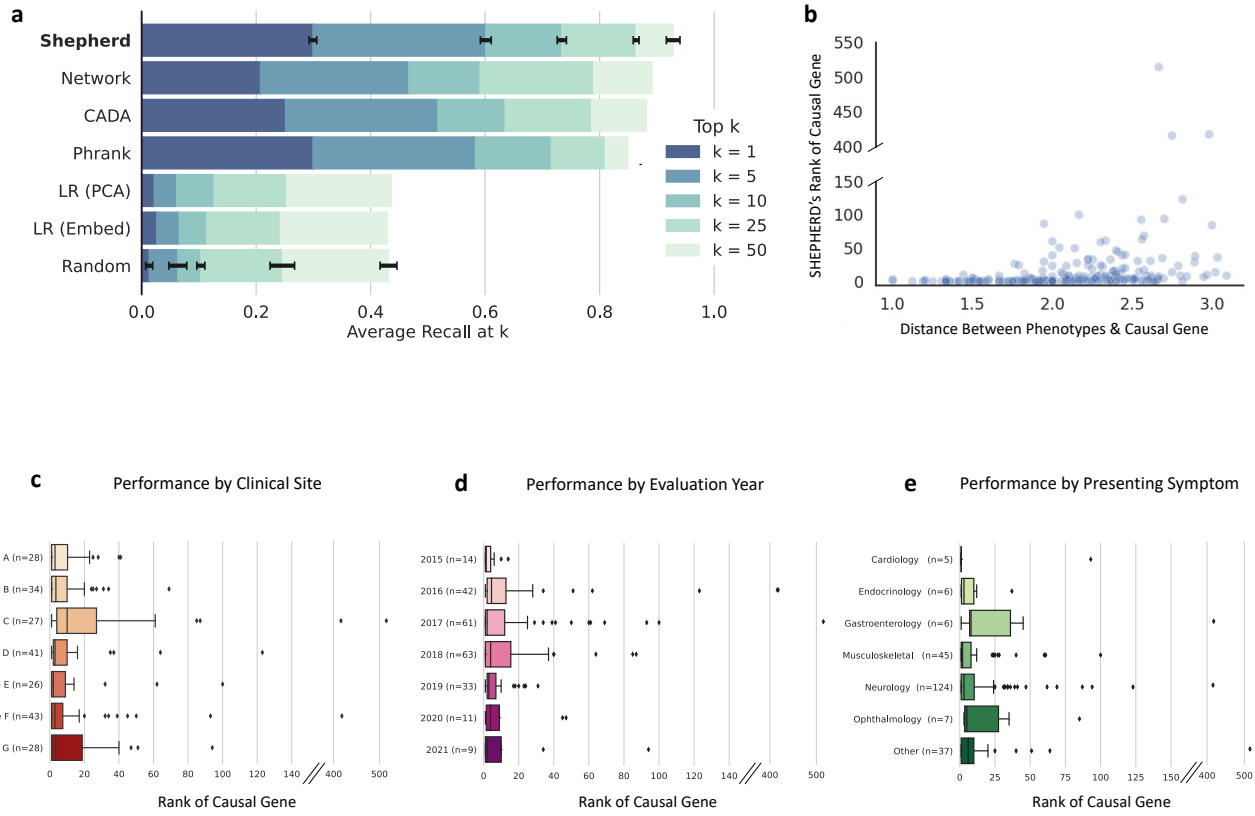


Figure S2: Causal gene discovery performance on VARIANT-FILTERED gene lists. a. Performance of SHEPHERD and six baseline models evaluated via average recall at k for k = 1, 5, 10, 25, and 50. **b.** Correlation between model performance (i.e., the rank of a disease-driving gene) and the average distance between a patient's phenotypes and causal genes in the knowledge graph. **c-e.** Performance of SHEPHERD in ranking causal genes stratified by **c.** clinical sites, **d.** evaluation year, and **e.** primary presenting symptom.

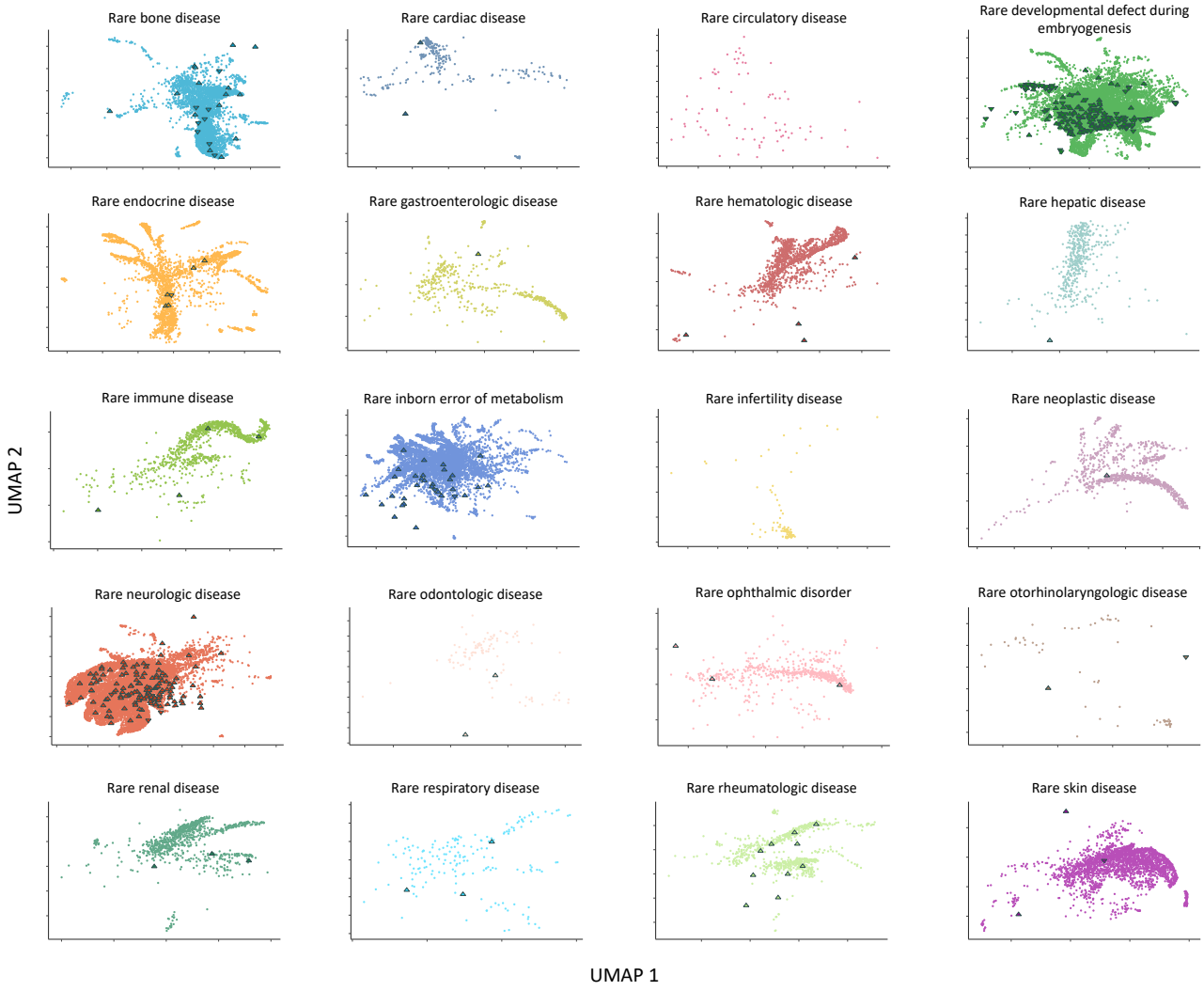


Figure S3: Visualization of rare disease patients by disease category. Two-dimensional UMAP plot of SHEPHERD's embedding space of all simulated patients (circles) and two real-world cohorts of UDN patients (up-facing triangles) and MyGene2 patients (down-facing triangles) grouped by the Orphanet disease category of medical diagnosis. Simulated, MyGene2 and UDN patients embed nearby other patients whose diagnoses belong to the same disease category.

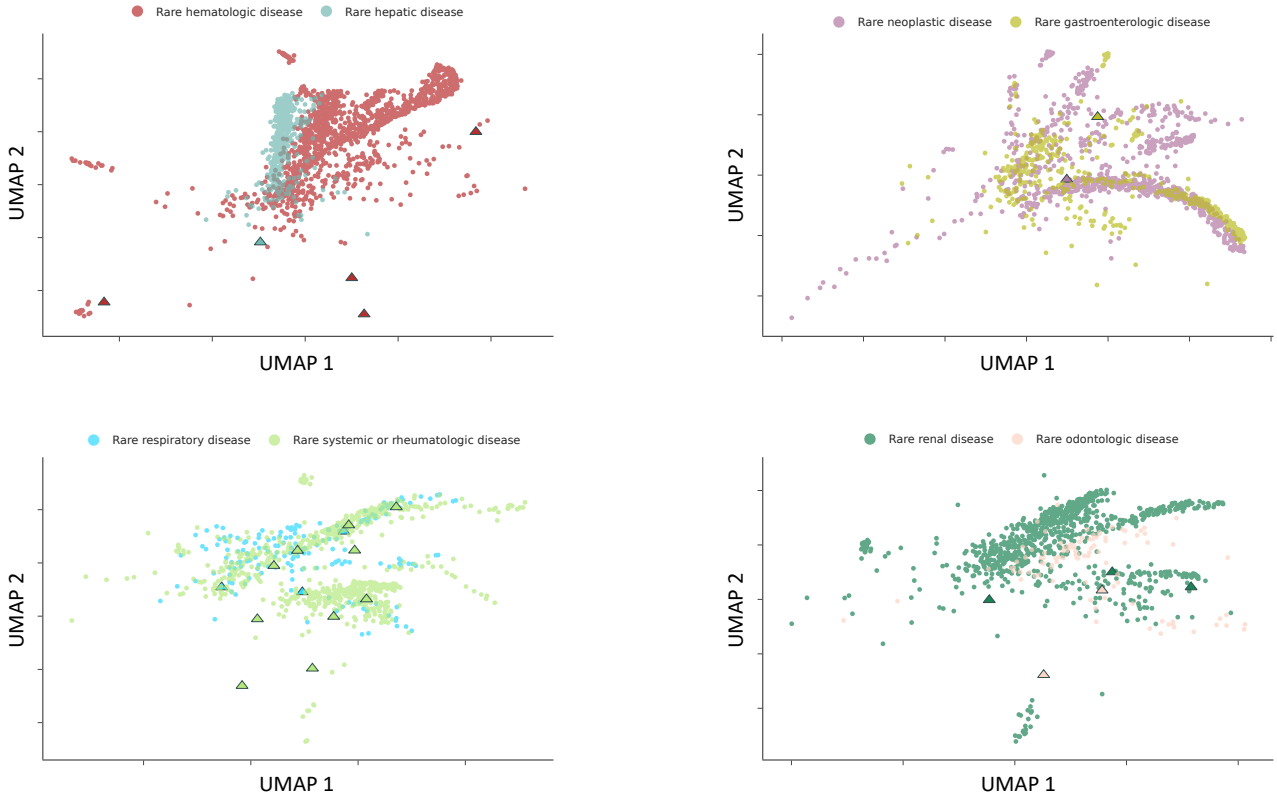


Figure S4: Visualization of the relationship between disease categories. Two-dimensional UMAP plot of SHEPHERD's embedding space for the most similar pairs of disease categories. Circles correspond to simulated patients, up-facing triangles to UDN patients, and down-facing triangles to MyGene2 patients.