PHENOTYPE DRIVEN GENE PRIORITIZATION AND FAST PATIENT RETRIEVAL WITH ROTATIONAL KNOWLEDGE GRAPH EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

The central challenge in variant interpretation for rare diseases is identifying the causal gene from a pool of candidates based on the patient's phenotypes. It is heavily dependent on subject domain expertise and is constrained to associations already documented in curated databases and the literature. In this study, we introduce a novel method - Patient Retrieval, Interpretation, and gene Scoring Model (PRISM), which is geometry aware, permutation invariant Set Transformer encoder that aligns variable size sets of patient phenotypes or knowledge graph (KG) entities, to candidate entities in rotational KG embedding (KGE) space. The KGE was built with the RotatE model over gene-phenotype and phenotype-phenotype relations from the Human Phenotype Ontology (HPO). Entity text descriptions were encoded with the pretrained biomedical language model, conferring inductive capability on PRISM. PRISM was trained on phenotypic profiles of nearly 22,000 rare disease patients, drawn from an in-house cohort of $\sim$30,000, to prioritize causal genes with a listwise multi-positive softmax loss function. On a held out cohort of more than 400 samples, integrated into the proprietary variant prioritization pipeline, PRISM recovered the causative pathogenic/likely pathogenic gene and variant within the Top-20 in 92.2% of cases and ranked the causative variant Top-1 in 57.8% of cases. Beyond the causal gene prioritization score, the same encoder embeds each patient's phenotypic profile in under half a second per inference, enabling retrieval of prior cases with similar phenotypes to aid variant interpretation.

1 INTRODUCTION

Next Generation Sequencing (NGS) has dramatically accelerated rare disease gene discovery and diagnosis (Dawood et al., 2025). However, variant interpretation remains a complex time consuming task riddled with differences in expert opinions (Costa et al., 2025). A central bottleneck of clinical genomics operation is phenotype driven gene prioritization to speed up variant interpretation. Given a patient's observed phenotypes, typically encoded as a set of Human Phenotype Ontology or HPO terms (Gargano et al., 2024), rank the genes most likely to harbor the causal variant. Because a patient is described by an unordered set of phenotypes and the label is one or more causal genes drawn from a catalog of several thousand, the task is naturally cast as listwise retrieval over the full gene catalog. The defining machine learning challenge is generalization. A clinically useful model must score genes it never observed as causal during training, since the diagnostic value lies precisely in the long tail of rarely or newly implicated disease genes.

Classical prioritization tools address this with hand designed semantic similarity between the patient's HPO set and curated gene phenotype annotations, for example information content (IC) based sets similarity (Jagadeesh et al., 2019), aggregated HPO to gene knowledge (Zhao et al., 2020; Zhai et al., 2023), and the phenotype plus variant scoring of Exomiser (Smedley et al., 2015; Cipriani et al., 2020). A broader class of clinical diagnostic systems augments phenotype matching with genotype and curated knowledge from literature evidence (Birgmeier et al., 2020), candidate gene network neighbors (Yoo et al., 2021), and machine learning integration of phenotypic and variant features (Li et al., 2019; Mao et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Classical systems work well in practice but rely on hand engineered features and curated gene-disease or literature knowledge. They target joint variant plus phenotype diagnosis rather than learning a phenotype–gene representation that transfers to unseen genes, so their scores are bounded by existing annotations and cannot place a novel gene in a shared phenotypic space. More recently, large language models and agentic pipelines have been prompted to rank candidate genes and variants (Kafkas et al., 2025; Kara et al., 2026). While flexible, they hallucinate genes, degrade sharply beyond a few dozen candidates, and are too costly to rank a full gene catalog, making them ill suited to retrieval. Another approach like Multiscore (Ustach et al., 2026) instead trains a random forest ranker on real clinical data, but it still operates over engineered similarity subscores rather than a learned representation of phenotypes and genes.

The most relevant advances learn representations directly on a biomedical knowledge graph (KG). Graph convolutional (Rao et al., 2018) and KG embedding (Peng et al., 2021; Gnanaolivu et al., 2025) methods embed genes and phenotypes and score by proximity, and SHEPHERD (Alsentzer et al., 2025) trains a graph attention network over a huge KG comprising of phenotypes, genes, pathways, diseases, proteins etc. explicitly targeting generalization to novel diseases. Related graph variants, such as the hypergraph traversal of Natarajan et al. (2025), likewise rank genes from existing phenotype-gene associations and do not score genes absent from the curated graph.

Moreover, SHEPHERD scores KG edges with a DistMult decoder (Yang et al., 2015), a symmetric bilinear form that assigns identical scores to a triple and its reverse and therefore cannot represent antisymmetric or inverse relations. This is a poor fit for an ontology whose backbone is the directed, antisymmetric HPO "is_a" hierarchy and whose gene-phenotype edges are materialized together with their inverses. To overcome this Sun et al. (2019) developed RotatE that capture each relation as a complex rotation, capturing symmetric, antisymmetric, inverse, and compositional patterns, the structure needed to encode a directed ontology and its inverse edges faithfully.

These approaches establish that KG representations help, but they leave three gaps open. First, their scorers do not preserve the relational geometry of the knowledge graph. Proximity in a learned embedding or message-passing space is not the same as applying the KG's phenotype to gene relation as a first class operation, so the geometry that made the KG informative is discarded at scoring time. Second, generalization is typically achieved through disease stratified training or transductive embeddings rather than a scorer that is architecturally gene agnostic. When any component of the score is tied to gene identity, transfer to unseen genes is not guaranteed by construction. Third, the strongest such model is trained predominantly on simulated patients. SHEPHERD uses 40k+ simulated rare disease cases, leaving open whether these gains hold for real clinical cohorts.

To overcome the existing gaps, we propose PRISM - Patient Retrieval, Interpretation, and gene Scoring Model. As shown in Figure 1, given a patient's set of HPO terms, PRISM ranks the full catalog of disease genes by their likelihood of being the causal gene, and simultaneously exposes: (i) a per phenotype attention profile explaining which observed phenotypes drove the ranking, and (ii) a dense patient embedding that supports patient-to-patient retrieval. It does all of that by learning to align a patient's phenotypic profile with its causal gene(s) by applying the KG's phenotype-gene relation as a genuine rotation in RotatE space (preserving relational geometry), trained on a cohort of $\sim$22,000 real diagnostic cases. By streamlining the discovery of gene-phenotype associations for causal gene prioritization, PRISM reduces reliance on manual interpretation of information latent in KGs and offers a path to improved diagnostic yield and faster retrieval of patients with similar phenotypic profiles.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

A patient p is represented as an unordered set of observed phenotypes $H_p = \{h_1, h_2, \ldots, h_{n_p}\}$, where each h_j is an HPO term. The label is one or more causal genes $G_p \subseteq G$, where G is a fixed gene catalog of 5,250 genes with known phenotype associations in the Human Phenotype Ontology Annotations (HPOA) data. PRISM learns a scoring function

$$s_\theta : (H_p, g) \to \mathbb{R}, \quad g \in G,$$

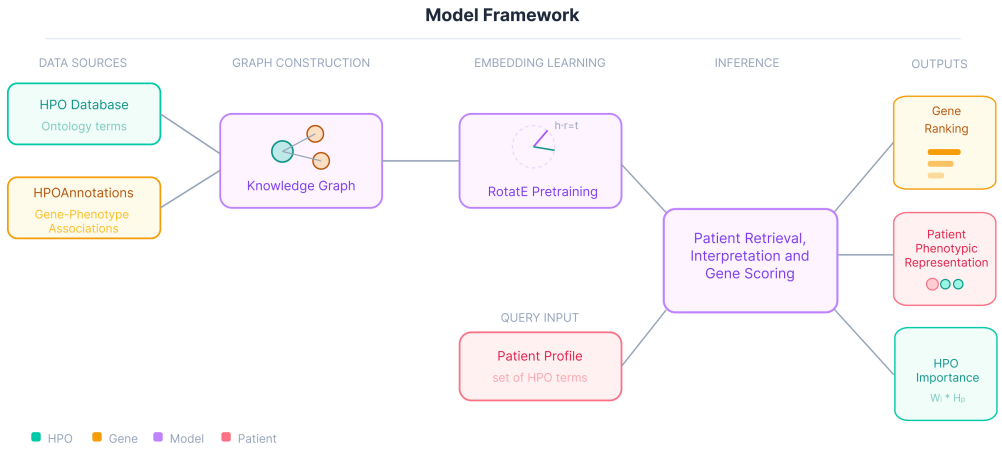


Figure 1: PRISM model framework outlining the data sources, knowledge graph construction, pre-trained rotational embedding and outputs produced by PRISM during inference.

and ranks all genes by $s_\theta(H_p, g)$. The score is learned by optimizing a multi-positive listwise ranking loss. At inference, the model scores every gene in G and returns a ranked list along with attention weights for each patient phenotype.

2.2 PRISM Overview: A two-stage pipeline

PRISM is trained in two decoupled stages:

(i) *Stage 1: Knowledge Graph Embedding (KGE).* A RotatE model is trained on a heterogeneous biomedical knowledge graph that couples the HPO ontology with gene to phenotype associations under Human Phenotype Ontology Annotations (HPOA). This yields: (a) node embeddings for every gene and every HPO term and (b) relation embeddings for each edge type.

(ii) *Stage 2: PRISM patient encoder and a gene scorer.* A Set Transformer encoder maps a patient's HPO set to a pooled embedding, weighting terms by learned attention and by ontology derived information content (IC) (Figure 2). A scoring head then compares the patient embedding against the frozen gene embeddings, aligning the patient embeddings with its respective causal gene embeddings.

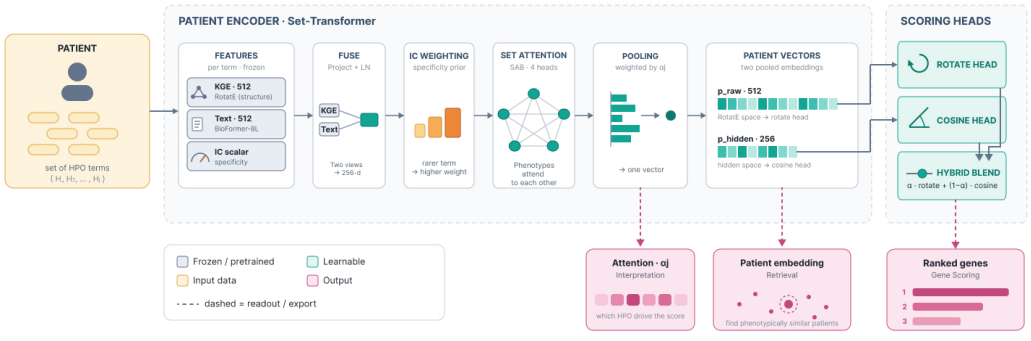


Figure 2: PRISM architecture details with a Set Transformer at the center to combine patient HPO terms and create patient phenotypic representation aligned with the most likely causal gene.

2.2.1 STAGE 1: ROTATE KNOWLEDGE GRAPH EMBEDDINGS

The knowledge graph unifies two sources into a single entity and relation vocabulary:

1. Ontology backbone: The HPO "is_a" directed acyclic graph (DAG) v2026-02-16, (Gargano et al., 2024) restricted to the Phenotypic abnormality subtree rooted at HP:0000118. The subtree is obtained by keeping the root together with all of its ontology descendants (the terms that reach the root through is_a edges), which removes non-phenotype branches such as clinical modifiers, frequency, and mode of inheritance terms.

2. Gene-phenotype associations are curated gene to phenotype links from the HPO Annotations (HPOA).

Each undirected biological relationship is materialized as two directed relations, phenotype-child to phenotype-parent ("is_a"), phenotype-parent to phenotype-child ("rev_is_a"), gene to phenotype ("has_phenotype"), and phenotype to gene ("rev_has_phenotype"), so that scoring can traverse it either way. Triples are deduplicated before training, resulting in an entity set containing 5,250 genes and 18,690 HPO terms with 557,420 edges including phenotype-phenotype and gene-phenotype relationships.

RotatE models each relation as an element wise rotation in complex space. For a triple (h, r, t), where h is the head entity, t is the tail entity, and r is the relation linking them (e.g., (gene, has_phenotype, phenotype)), complex entity embeddings $\mathbf{e}_h, \mathbf{e}_t \in \mathbb{C}^k$ and a unit-modulus relation phase $\mathbf{r} \in \mathbb{C}^k$ (i.e., $|r_i| = 1$ for all i), the score is computed as

$$f(h, r, t) = -\left\|\mathbf{e}_h \odot \mathbf{r} - \mathbf{e}_t\right\|,$$

where $\odot$ is the Hadamard (per-coordinate) complex product. The model was trained with the standard margin-based ranking loss with uniform negative sampling, as implemented in the PyTorch Geometric framework (Fey & Lenssen, 2019), with $k = 256$ complex dimensions, margin $\gamma = 9.0$, the Adam optimizer (learning rate 10^{-3}), a batch size of 4096, and 100 epochs.

With the trained RotatE KGE, node embeddings were exported as the concatenation of the real and imaginary parts,

$$\mathbf{h}_{\text{rot}} = [\Re(\mathbf{e}) \,\|\, \Im(\mathbf{e})] \in \mathbb{R}^{2k} = \mathbb{R}^{512},$$

and L2 normalized. Because a RotatE rotation acts per coordinate and preserves each coordinate's modulus, it preserves the global L2 norm, so rotation and normalization commute, and the downstream scoring in Stage 2 is well defined on normalized vectors. Relation embeddings are exported as $[\cos\boldsymbol{\theta} \,\|\, \sin\boldsymbol{\theta}] \in \mathbb{R}^{2k}$ and are left unnormalized, since normalizing would destroy the per-coordinate phase that encodes the rotation. The phenotype-to-gene relation is the one applied to the patient in Stage 2, because PRISM builds a patient vector from HPO embeddings and scores it against learned gene embeddings.

2.2.2 STAGE 2: PATIENT PHENOTYPES SET ENCODER

PRISM consumes three frozen inputs per HPO term: a RotatE node KGE vector $\mathbf{h}^{\text{rot}} \in \mathbb{R}^{512}$, a text embedding vector $\mathbf{h}^{\text{txt}} \in \mathbb{R}^{512}$, and the information content $\text{IC} \in \mathbb{R}$. Text embeddings were created with the bioformer-8L (Fang et al., 2023) language model for each HPO term, encoding the term's title, definition, and synonyms. The information content of HPO term h_j was computed directly from the HPO DAG as the negative log of the fraction of all phenotype terms that are descendants of h_j (inclusive):

$$\text{IC}(j) = -\log \frac{|\text{desc}(j)| + 1}{T},$$

where $\text{desc}(j)$ is the set of descendant terms of j in the HPO DAG and T is the total number of phenotype terms, so broad, high-level terms score low and specific leaf terms score high. The IC was then min-max normalized to $[0, 1]$ across terms before use. This makes the IC topology based rather than corpus frequency based like Resnik's IC. In the model, IC acts as a fixed, interpretable prior and enters the encoder only through monotone, sign-constrained gates, so more specific phenotypes carry more influence.

The patient encoder is a lightweight Set Transformer. Because a patient is a set of phenotypes with no meaningful order, no positional encoding is used and every operation is permutation equivariant.

In the initial step, the two views, KGE and text vectors are linearly projected into a shared hidden space ($H = 256$), summed, and layer-normalized:

$$\mathbf{x}_j = \mathrm{LayerNorm}(\mathbf{W}_{\mathrm{rot}}\mathbf{h}_{\mathrm{rot}} + \mathbf{W}_{\mathrm{txt}}\mathbf{h}_{\mathrm{txt}}) \in \mathbb{R}^H.$$

After passing $\mathbf{x}_j$ through an IC gate and bias, each patient's set is packed into a dense padded tensor and passed through a Set Attention Block (SAB) (Lee et al., 2019) consisting of prenorm multi head self attention module with 4 heads (implemented as in PyTorch (Ansel et al., 2024)), followed by a residual connection and a pre-norm feed-forward residual layer with masked padded position. This contextualizes each phenotype against the others in the same patient and gives a contextualized output $\mathbf{x}_j^{\mathrm{ctx}}$.

In order to construct the patient representation from the contextualized phenotype vectors, a gated attention pooling (Ilse et al., 2018) was applied where per term pooling logits used a gated ($\tanh \odot \sigma$) attention head plus a monotone IC bias, and are normalized with a segmented softmax within each patient:

$$e_j = \mathbf{w}^\top \big(\tanh(V\mathbf{x}_j^{\mathrm{ctx}}) \odot \sigma(U\mathbf{x}_j^{\mathrm{ctx}})\big) + \mathrm{softplus}(c) \cdot \mathrm{IC}_j,$$
$$\alpha_j = \frac{\exp(e_j)}{\sum_{k \in H_p} \exp(e_k)}.$$

The $\tanh$ captures the direction of each feature's contribution (bounded in $[-1, 1]$), while the sigmoid, a multiplicative relevance gate (bounded in $[0, 1]$), independently suppresses or passes each dimension. The weights α_j are the interpretability weights returned at inference to expose patient phenotype contributions, whereas $\mathrm{softplus}(c)$ is a learnable, sign-constrained coefficient. Now, the encoder returns two attention-pooled representations that share the same weights α_j but live in different spaces:

$$\mathbf{p}_{\mathrm{hidden}} = \sum_j \alpha_j \mathbf{x}_j^{\mathrm{ctx}} \in \mathbb{R}^H \qquad \text{(for the cosine head)},$$
$$\mathbf{p}_{\mathrm{raw}} = \sum_j \alpha_j \mathbf{r}_j \in \mathbb{R}^{512} \qquad \text{(for the rotate head)}.$$

where $\mathbf{r}_j = \mathbf{h}_{\mathrm{rot}}$. Keeping $\mathbf{p}_{\mathrm{raw}}$ in the original RotatE space is what lets the downstream rotate scoring head apply the KG relation as a genuine complex rotation.

Taking the learned phenotypic profile, PRISM tries to align the patient representation with the gene representation using three scores: a Rotate head (RotatE-consistent), a Cosine head, and a Hybrid head. All are gene agnostic, which is what makes zero-shot scoring of novel genes possible. The Rotate head applies the frozen "rev_has_phenotype" rotation $\mathbf{r}_{\mathrm{assoc}}$ to the pooled patient vector, L2-normalizes the result, and measures cosine similarity against the frozen, normalized gene embeddings:

$$S_{\mathrm{rot}}(p, g) = \tau \cdot \Big\langle \widehat{\mathrm{Rotate}(p_{\mathrm{raw}},\, r_{\mathrm{assoc}})},\ \hat{g} \Big\rangle \tag{1}$$

The logits are scaled by a single learnable temperature parameter τ, initialized to $\log(1/0.07)$ and clamped to $[10^{-3}, 10^2]$, following the contrastive scaling scheme of CLIP (Radford et al., 2021). This controls the sharpness of the ranking distribution. S_{rot} is consistent with the RotatE scoring geometry and introduces no new gene parameters. The Cosine head on the other hand uses the learnable projection of $\mathbf{p}_{\mathrm{hidden}}$ which is compared against the frozen PCA gene tower. The gene projection is initialized from the top-d principal directions of the gene KGE matrix (computed once via SVD, mean-centered), and its parameters are frozen. It is an unsupervised, gene agnostic map over all genes, so it too is safe for novel genes. Thus, all learnable capacity resides in the gene agnostic patient encoder. This preserves a single, globally consistent RotatE coordinate system across the full gene catalog and enables scoring of genes unseen during training.

The two heads are combined by a convex blend, where the mixing weight $\beta = \sigma(b)$ is obtained by passing a single learnable, gene agnostic scalar b through a sigmoid, constraining $\beta \in (0, 1)$:

$$S_{\mathrm{hybrid}}(p, g) = \beta\, S_{\mathrm{rot}}(p, g) + (1 - \beta)\, S_{\cos}(p, g).$$

2.3 TRAINING OBJECTIVE AND OPTIMIZATION DETAILS

The primary loss is a listwise multi-positive softmax over the entire gene catalog, which matches full-catalog inference. For a patient with positive gene set G_p

$$\mathcal{L}_{\text{listwise}} = \log \sum_{g \in G} e^{s_g} \; - \; \log \sum_{g \in G_p} e^{s_g},$$

i.e., it drives probability mass onto the true causal gene(s) relative to all genes. PRISM was trained by minimizing $\mathcal{L}_{\text{listwise}}$ with AdamW (Loshchilov & Hutter, 2019). Because the different components of the patient encoder operate at different scales, we grouped parameters and assigned each group its own learning rate. The Set Transformer encoder (projections, self-attention, and gated pooling) was trained at 1×10^{-4}, the interpretable IC gate and bias parameters at 2×10^{-5}, the learnable cosine-head projection at 2×10^{-4}, and the hybrid blend weight b at 1×10^{-3}. A weight decay of 1×10^{-4} was applied to the encoder and cosine projection only, with none on the IC, temperature, and blend parameters. The frozen PCA gene tower and all gene embeddings receive no gradients. We used a batch size of 512 and clipped gradients to a global norm of 1.0. The learning rate is reduced by a factor of 0.5 whenever validation mean reciprocal rank (MRR) plateaus for 3 epochs (down to 1×10^{-5}), and training was stopped early once validation MRR failed to improve for 5 consecutive epochs, after which the checkpoint with the best validation MRR was restored for evaluation.

2.4 GENE DISJOINT SPLIT (NOVEL GENE EVALUATION)

Patients were split into train, validation and test set by causal gene. The set of causal genes was partitioned (80/10/10) under a fixed random seed (42), and a patient was routed to a split only if all of its causal genes fell within that split's gene set and patients straddling splits were dropped. Consequently, no gene seen as a positive in training appeared as a positive in validation or test. This makes the reported metrics a direct measure of generalization to previously unseen disease genes, the setting the frozen, gene agnostic heads are designed for, rather than memorization.

3 EXPERIMENTS AND RESULTS

3.1 DATASETS

PRISM was trained and evaluated on an internal cohort of ∼30,000 patients collected between August 2017 and October 2025, comprising cases in which causal gene variants were reported during routine clinical genomics operations. All samples were analyzed using either the whole exome or whole genome clinical sequencing pipeline and in-house variant annotation. Each patient was annotated with a median of 4.0 HPO terms (IQR 2.0-7.0; range 1–54). After mapping phenotypes and causal genes to the knowledge graph vocabulary and applying a gene disjoint split (Methods), 26,721 patients remained giving 21,907 for training, 2,354 for validation, and 2,460 for test during the training process.

From the same cohort, we reserved an additional held out test set of 410 patients, distinct from the split level test set, whose 123 causal genes never appeared as positives during training, directly measuring generalization to unseen genes. This led to the training set containing 2,484 genes seen as positive during training compared to 5,250 total genes in the KG. Across the split, we observe 6,111 unique HPO terms in training and 726 in the held out test set, with 716 overlapping terms. Because SHEPHERD, unlike PRISM, requires a candidate gene list as input, we applied hard variant filtering to the held out set (Appendix A.1), yielding an average of 275 candidate genes per patient.

For external validation we additionally used a simulated rare disease dataset of 42,224 simulated patients (Alsentzer et al., 2022), generated by an adaptive simulation approach (Alsentzer et al., 2023) that produces realistic rare disease patients from curated disease-gene-phenotype knowledge using the real patient cohort drawn from the Undiagnosed Disease Network(Ramoni & Mulvihill, 2017) dataset. PRISM was benchmarked against the state-of-the-art method SHEPHERD on the held out test set and the simulated rare disease dataset.

3.2 RESULTS

3.2.1 PERFORMANCE EVALUATION CRITERIA

Gene ranking quality was measured with average recall@k ($k \in \{1, 5, 10, 20, 50\}$) as described by (Alsentzer et al., 2025). As each patient can have multiple causal genes, for each case we compute the proportion of its true causal genes that appear in the model's top-k ranked list, then average this per case recall over all evaluable cases.

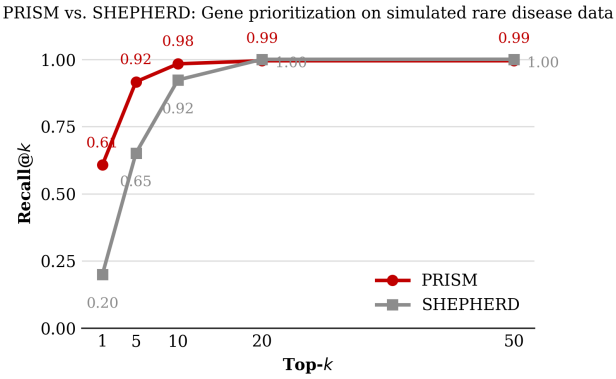


Figure 3: Validation on external cohort: causal gene prioritization performance of PRISM measured against state-of-the-art SHEPHERD on a simulated rare disease dataset.

3.2.2 GENE PRIORITIZATION IN SIMULATED RARE DISEASE PATIENTS

On the simulated rare disease cohort ($N = 42{,}224$ patients), PRISM applied zero-shot, i.e., trained only on our in-house cohort with no fine-tuning on these data, achieves average recall of 0.61, 0.92, and 0.98 at $k = 1, 5$, and 10, respectively, outperforming SHEPHERD (0.20 / 0.65 / 0.92; Figure 3), which we ran from the authors' released checkpoints without retraining. Notably, PRISM exceeds SHEPHERD, even though this cohort is drawn from the distribution SHEPHERD was trained on, indicating that PRISM generalizes to an independent, externally generated cohort. For a fair comparison, both models were scored over the same candidate gene set per patient as provided in the simulated rare disease cohort.

3.2.3 GENE AND VARIANT PRIORITIZATION IN IN-HOUSE RARE DISEASE PATIENTS

In the real clinical environment, as shown in Figure 4 (a), PRISM scored an average recall of 0.89 at k = 50 for gene prioritization, while SHEPHERD underperformed by 0.4. Both methods, however, underperformed at k=1. Since the held out cohort consists of patients with causal genes that were never seen as positive during training, performance on it directly measures generalization to unseen genes. On this real clinical cohort, PRISM outperformed the state-of-the-art method SHEPHERD, although its absolute performance was lower than on the simulated rare disease dataset. The only difference from their setup is candidate gene list construction, which we tried to replicated as faithfully as possible (Appendix A.1).

Because clinical variant interpretation ultimately requires ranking variants, not just genes, we evaluated PRISM end to end by integrating it into an in-house variant prioritization pipeline (Baseline; Figure 4 (b)). For a fair comparison, the only change to the Baseline was the addition of PRISM or SHEPHERD scores for the candidate genes. PRISM again outperformed SHEPHERD, reaching a mean recall of 0.895 at k=10 and ranking the causal variant first in 57.8% of cases, compared with 30.2% for SHEPHERD. In this end-to-end setting the causative pathogenic/likely pathogenic gene and variant were recovered within the Top-20 in 92.2% of cases (Figure 4 (b)).

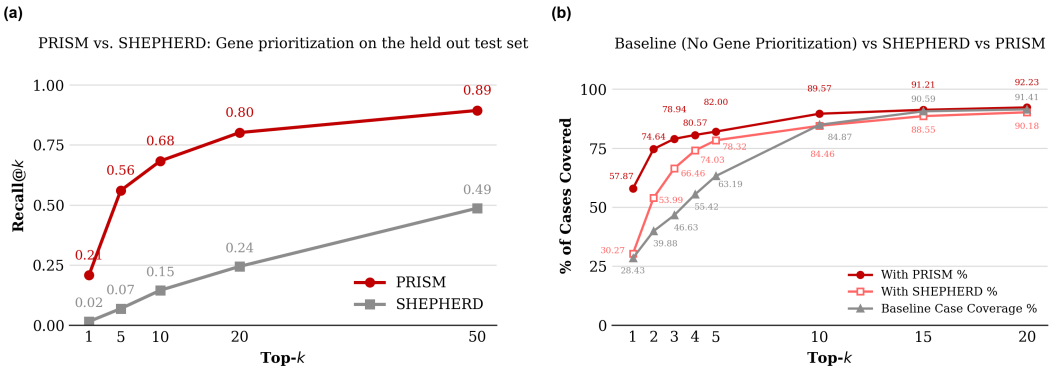


Figure 4: (a) Causal gene prioritization performance of PRISM measured against SHEPHERD on the held out test dataset. (b) Impact of phenotype driven gene ranking on downstream variant prioritization. Two phenotype aware gene ranking methods, PRISM and SHEPHERD, are compared against a baseline that performs variant prioritization without any phenotype-gene association signal.

3.2.4 Clustering of patients' learned phenotypic representation

To assess whether PRISM's patient embeddings capture clinically meaningful structure, we applied unsupervised agglomerative hierarchical clustering (Ward linkage, Euclidean distance) to the training cohort embeddings, selecting k = 15 where cluster-causal gene agreement (adjusted mutual information) plateaued without producing degenerate clusters (minimum cluster size 673; Appendix Figure 6). The clustered patient-patient distance matrix (Figure 5) shows clear low distance blocks along the diagonal, indicating cohesive groups of phenotypically similar patients. Although the model received no disease category supervision, every cluster was enriched for a coherent clinical entity and its canonical causal genes e.g., glomerular/nephrotic disease (COL4A3/4/5, NPHS1/2), neuromuscular disease (DMD, LAMA2), skeletal dysplasia (FGFR3, FBN1), and developmental/epileptic encephalopathy (SCN1A, KCNQ2, STXBP1), indicating that phenotype-only embeddings recover molecular etiology.

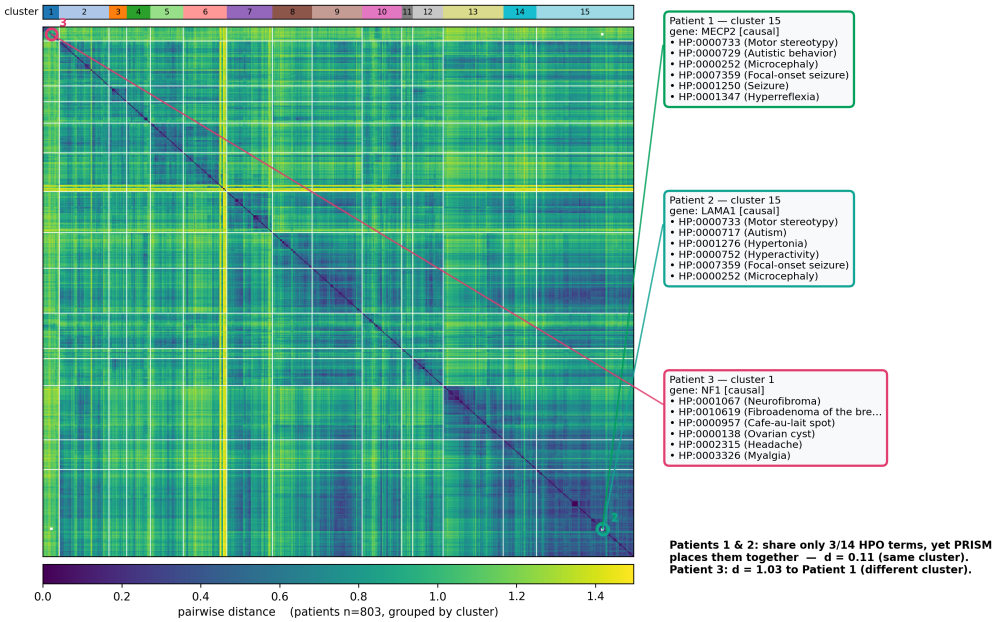


Figure 5: Hierarchical clustering of patients using the learned phenotypic representations from PRISM (n=803 randomly subsampled for visualization; clustering used full training set).

4 DISCUSSION

PRISM demonstrates that preserving the relational geometry of a knowledge graph and making the scoring head architecturally gene agnostic together yield a phenotype driven ranker that generalizes to genes never seen in training. Under the gene disjoint split, where the 123 test genes never appear as positives, PRISM recovers the causal gene with substantially higher recall than SHEPHERD on a real clinical cohort. Applied zero shot to SHEPHERD's own simulated distribution, PRISM, trained on only $\sim$22k real cases, still outperforms it, indicating that the gains stem from representation and scoring design rather than cohort overlap.

The generalization traces to a single design principle, the gene representations are frozen. Supervision exists for only a small, non random subset of causal genes, so fine tuning gene embeddings would deform the globally trained RotatE geometry and misalign exactly the held out genes that matter most, the long tail of rare disease genes. Freezing confines all learnable capacity to the gene agnostic patient encoder and keeps every gene, seen or unseen, in one consistent coordinate system, so a patient representation calibrated on training genes transfers to novel ones. Applying the phenotype to gene relation as a genuine RotatE rotation, rather than scoring by proximity in a learned space, is what exploits that geometry at scoring time.

Beyond ranking, PRISM is interpretable and supports retrieval, capabilities clinical tools require. Per phenotype attention weights expose which phenotypes drove each prediction, and the learned patient embeddings organize cases into an etiologically structured space. Unsupervised clustering recovers coherent disease groups enriched for their canonical causal genes Figure 5, and patients sharing a causal gene form tight, well-separated neighborhoods in the embedding (Appendix Figure 7), despite no disease category supervision. In hard to diagnose variant interpretation, retrieving previously reported patients with similar phenotypes is often decisive, a use case the etiological structure of PRISM's embedding space is well suited to support.

Limitations remain. The underlying HPO/HPOA resources have uneven coverage across genes, diseases, and ancestries, so performance may degrade for under-annotated or under-represented cases. Both PRISM and the baseline remain weak at Top-1 on real data, underscoring that gene ranking is a decision support signal rather than a stand alone diagnosis. The two stage design also ties the model to a fixed ontology release, requiring retraining as HPO evolves. Additionally, we demonstrate PRISM's retrieval capability qualitatively, through the structure of the learned embedding space, a quantitative retrieval evaluation (e.g., precision of same causal gene neighbors at k) is left to future work. Future work will jointly train the KGE with the ranking objective under a multi-objective loss, incorporate variant-level evidence, pursue ontology version agnostic representations, and evaluate fairness across diverse populations within clinical diagnostic workflows.

5 CONCLUSION

We introduced PRISM, a two stage, phenotype driven framework for causal gene prioritization. PRISM first learns a shared representation of genes and phenotypes by training a RotatE knowledge graph embedding over the HPO ontology and known phenotype-gene associations, and then learns to create a patient's phenotypic representation with a lightweight, permutation invariant Set Transformer whose attention is modulated by an interpretable, ontology derived information content prior. By scoring a patients' phenotypic representations against gene representations through frozen, gene agnostic heads, PRISM ranks the genes most likely to be disease causing, including genes never seen as causal during training. Our approach shows that coupling a structured biomedical knowledge graph with a permutation invariant, information content aware patient encoder in a rotational space is an effective and interpretable route to zero shot gene prioritization.

AI USE STATEMENT

We used an AI assistant (a large language model) solely to aid and polish the writing of this paper. Its use was limited to proofreading, improving grammar and clarity, checking the consistency of the method description against our implementation, and formatting citations. All technical content, experimental design, results, and claims are the authors' own. The AI assistant or Agent was not

used to generate research ideas, produce experimental results. The authors reviewed and verified all text and take full responsibility for the content of the paper.

ETHICS STATEMENT

PRISM is intended as a research and clinical decision support tool that prioritizes genes linked with patient phenotypes to assist expert interpretation, not as an autonomous diagnostic system. Its outputs should be reviewed by qualified clinicians or genetic analysts before informing any clinical decision. The model operates on Human Phenotype Ontology (HPO) term, standardized, nonidentifiable phenotypic abstractions, rather than raw patient records, and all patient data used in this work were deidentified and handled in accordance with the terms under which they were obtained. We note that the model's knowledge graph including gene-phenotype associations are derived from curated resources (HPO and HPOA) whose coverage is uneven across genes, diseases, and populations. Consequently, predictions may be less reliable for underannotated genes or for phenotypes and ancestries that are underrepresented in these resources, and users should be aware of this potential for unequal performance. The gene-disjoint evaluation is designed to measure generalization to previously unseen disease genes, but reported metrics do not guarantee performance in any specific clinical setting. We encourage users to treat gene rankings as hypotheses to be validated through established diagnostic workflows, and we caution against over reliance on model scores in isolation.

REPRODUCIBILITY STATEMENT

The Methods section and experimental setup fully specifies the model architecture, the two stage training procedure, all hyperparameters, the knowledge graph construction, and the gene disjoint evaluation technique. These details are sufficient to reimplement PRISM. Our external validation uses the openly available SHEPHERD simulated rare disease cohort, so those results can be reproduced against public data. The primary training and evaluation cohort is proprietary clinical data and cannot be shared due to patient privacy and data sharing restrictions. Source code can be made available for review upon request, subject to approval, and we can provide access to a hosted PRISM inference endpoint to reproduce the reported figures following publication.

REFERENCES

Emily* Alsentzer, Michelle M. Li*, Shilpa N. Kobren, Undiagnosed Diseases Network, Isaac S. Kohane, and Marinka Zitnik. Deep Learning for diagnosing patients with rare genetic diseases, 2022. URL https://doi.org/10.7910/DVN/TZTPFL.

Emily Alsentzer, Samuel G. Finlayson, Michelle M. Li, Shilpa N. Kobren, Isaac S. Kohane, and Undiagnosed Diseases Network. Simulation of undiagnosed patients with novel genetic conditions. *Nature Communications*, 14(1):6403, Oct 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-41980-6. URL https://doi.org/10.1038/s41467-023-41980-6.

Emily Alsentzer, Michelle M. Li, Shilpa N. Kobren, Ayush Noori, Isaac S. Kohane, Marinka Zitnik, and Undiagnosed Diseases Network. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8(1):380, Jun 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01749-1. URL https://doi.org/10.1038/s41746-025-01749-1.

Jason Ansel, Edward Yang, He, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, pp. 929–947, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703850. doi: 10.1145/3620665.3640366. URL https://doi.org/10.1145/3620665.3640366.

Johannes Birgmeier, Maximilian Haeussler, Cole A. Deisseroth, Ethan H. Steinberg, Karthik A. Jagadeesh, Alexander J. Ratner, Harendra Guturu, Aaron M. Wenger, Mark E. Diekhans, Peter D. Stenson, David N. Cooper, Christopher Ré, Alan H. Beggs, Jonathan A. Bernstein, and Gill Bejerano. Amelie speeds mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Science Translational Medicine*, 12(544):eaau9113, 2020. doi: 10.1126/scitranslmed.

aau9113. URL https://www.science.org/doi/abs/10.1126/scitranslmed.aau9113.

Valentina Cipriani, Nikolas Pontikos, Gavin Arno, Panagiotis I. Sergouniotis, Eva Lenassi, Penpitcha Thawong, Daniel Danis, Michel Michaelides, Andrew R. Webster, Anthony T. Moore, Peter N. Robinson, Julius O.B. Jacobsen, and Damian Smedley. An improved phenotype-driven tool for rare mendelian variant prioritization: Benchmarking exomiser on real patient whole-exome data. *Genes*, 11(4), 2020. ISSN 2073-4425. doi: 10.3390/genes11040460. URL https://www.mdpi.com/2073-4425/11/4/460.

Mireia Costa, Alberto García S, Ana León, and Oscar Pastor. The promises and pitfalls of automated variant interpretation: a comprehensive review. *Briefings in Bioinformatics*, 26(5):bbaf545, 09 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf545. URL https://doi.org/10.1093/bib/bbaf545.

Moez Dawood, Heavner, Ben, Wheeler, GREGoR Partner Marsha M., Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) Consortium, et al. Gregor: accelerating genomics for rare diseases. *Nature*, 647(8089):331–342, Nov 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09613-8. URL https://doi.org/10.1038/s41586-025-09613-8.

Li Fang, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang. Bioformer: an efficient transformer language model for biomedical text mining, 2023. URL https://arxiv.org/abs/2302.01588.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric, 2019. URL https://arxiv.org/abs/1903.02428.

Michael A Gargano, Nicolas Matentzoglu, Ben Coleman, et al. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, 01 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1005. URL https://doi.org/10.1093/nar/gkad1005.

Rohan Gnanaolivu, Gavin Oliver, Garrett Jenkinson, Emily Blake, Wenan Chen, Nicholas Chia, Eric W. Klee, and Chen Wang. A clinical knowledge graph-based framework to prioritize candidate genes for facilitating diagnosis of mendelian diseases and rare genetic conditions. *BMC Bioinformatics*, 26(1):82, Mar 2025. ISSN 1471-2105. doi: 10.1186/s12859-025-06096-2. URL https://doi.org/10.1186/s12859-025-06096-2.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Karthik A. Jagadeesh, Johannes Birgmeier, Harendra Guturu, Cole A. Deisseroth, Aaron M. Wenger, Jonathan A. Bernstein, and Gill Bejerano. Phrank measures phenotype sets similarity to greatly improve mendelian diagnostic disease prioritization. *Genetics in Medicine*, 21(2):464–470, Feb 2019. ISSN 1098-3600. doi: 10.1038/s41436-018-0072-y. URL https://doi.org/10.1038/s41436-018-0072-y.

Şenay Kafkas, Marwa Abdelhakim, Azza Althagafi, Sumyyah Toonsi, Malak Alghamdi, Paul N. Schofield, and Robert Hoehndorf. The application of large language models to the phenotype-based prioritization of causative genes in rare disease patients. *Scientific Reports*, 15(1):15093, Apr 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-99539-y. URL https://doi.org/10.1038/s41598-025-99539-y.

Muti Kara, Ahmet Furkan Gungor, Safa Eren Kuday, Oguzhan Ozcelik, Serdar Ceylaner, Fatma Dereli, and Furkan Ozden. Deep agentic variant prioritisation for expert level genetic diagnosis. *medRxiv*, 2026. doi: 10.64898/2026.02.17.26346421. URL https://www.medrxiv.org/content/early/2026/05/25/2026.02.17.26346421.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Qigang Li, Keyan Zhao, Carlos D. Bustamante, Xin Ma, and Wing H. Wong. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genetics in Medicine*, 21(9):2126–2134, Sep 2019. ISSN 1098-3600. doi: 10.1038/s41436-019-0439-8. URL https://doi.org/10.1038/s41436-019-0439-8.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Dongxue Mao, Chaozhong Liu, Linhua Wang, Rami AI-Ouran, Cole Deisseroth, Sasidhar Pasupuleti, Seon Young Kim, Lucian Li, Jill A. Rosenfeld, Linyan Meng, Lindsay C. Burrage, Michael F. Wangler, Shinya Yamamoto, Michael Santana, Victor Perez, Priyank Shukla, Christine M. Eng, Brendan Lee, Bo Yuan, Fan Xia, Hugo J. Bellen, Pengfei Liu, and Zhandong Liu. Ai-marrvel — a knowledge-driven ai system for diagnosing mendelian disorders. *NEJM AI*, 1 (5):AIoa2300009, 2024. doi: 10.1056/AIoa2300009. URL https://ai.nejm.org/doi/full/10.1056/AIoa2300009.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Shrinithi Natarajan, Niveditha Kundapuram, Nisarga Bhaskar, Sai Sailaja Policharla, and Bhaskarjyoti Das. A hypergraph powered approach to phenotype-driven gene prioritization and rare disease prediction. *Scientific Reports*, 15(1):23780, Jul 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-04428-z. URL https://doi.org/10.1038/s41598-025-04428-z.

Chengyao Peng, Simon Dieck, Alexander Schmid, Ashar Ahmad, Alexej Knaus, Maren Wenzel, Laura Mehnert, Birgit Zirn, Tobias Haack, Stephan Ossowski, Matias Wagner, Theresa Brunet, Nadja Ehmke, Magdalena Danyel, Stanislav Rosnev, Tom Kamphans, Guy Nadav, Nicole Fleischer, Holger Fröhlich, and Peter Krawitz. Cada: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genomics and Bioinformatics*, 3(3):lqab078, 09 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab078. URL https://doi.org/10.1093/nargab/lqab078.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rachel B. Ramoni and John J.and others Mulvihill. The undiagnosed diseases network: Accelerating discovery about health and disease. *The American Journal of Human Genetics*, 100(2):185–192, Feb 2017. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.01.006. URL https://doi.org/10.1016/j.ajhg.2017.01.006.

Aditya Rao, Saipradeep VG, Thomas Joseph, Sujatha Kotte, Naveen Sivadasan, and Rajgopal Srinivasan. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Medical Genomics*, 11(1):57, Jul 2018. ISSN 1755-8794. doi: 10.1186/s12920-018-0372-8. URL https://doi.org/10.1186/s12920-018-0372-8.

Damian Smedley, Julius O. B. Jacobsen, Marten Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, Tomasz Zemojtel, Orion J. Buske, Nicole L. Washington, William P. Bone, Melissa A. Haendel, and Peter N. Robinson. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature Protocols*, 10(12):2004–2015, Dec 2015. ISSN 1750-2799. doi: 10.1038/nprot.2015.124. URL https://doi.org/10.1038/nprot.2015.124.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space, 2019. URL https://arxiv.org/abs/1902.10197.

Vincent D. Ustach, Maria J. Guillen Sacoto, Stephen McGee, et al. Multiscore, a gene ranker powered by artificial intelligence and real-world clinical data, shows high sensitivity for the molecular diagnosis of mendelian disorders in nearly 10,000 exomes and genomes. *Human Genetics*, 145(1):22, Feb 2026. ISSN 1432-1203. doi: 10.1007/s00439-025-02797-z. URL https://doi.org/10.1007/s00439-025-02797-z.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases, 2015. URL https://arxiv.org/abs/1412.6575.

Boyoung Yoo, Johannes Birgmeier, Jonathan A. Bernstein, and Gill Bejerano. Inphernet accelerates monogenic disease diagnosis using patients’ candidate genes’ neighbors. *Genetics in Medicine*, 23(10):1984–1992, Oct 2021. ISSN 1098-3600. doi: 10.1038/s41436-021-01238-2. URL https://doi.org/10.1038/s41436-021-01238-2.

Weiqi Zhai, Xiaodi Huang, Nan Shen, and Shanfeng Zhu. Phen2disease: a phenotype-driven model for disease and gene prioritization by bidirectional maximum matching semantic similarities. *Briefings in Bioinformatics*, 24(4):bbad172, 07 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad172. URL https://doi.org/10.1093/bib/bbad172.

Mengge Zhao, James M Havrilla, Li Fang, Ying Chen, Jacqueline Peng, Cong Liu, Chao Wu, Mahdi Sarmady, Pablo Botas, Julián Isla, Gholson J Lyon, Chunhua Weng, and Kai Wang. Phen2gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genomics and Bioinformatics*, 2(2):lqaa032, 06 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa032. URL https://doi.org/10.1093/nargab/lqaa032.

A Additional experimental details

A.1 Candidate gene filtering

To filter candidate genes for SHEPHERD re-run input we performed variant filtering for each sample by removing variants in genes on an internal exclusion list of recurrent false positive genes, selecting variants present in MANE or MANE Plus Clinical transcripts, and passing a population frequency cut-offs by considering the maximum of allele frequencies from gnomAD, G1000 and Genome Asia 100K and their subpopulations which was kept at 0.05. We avoid using the foundation population while calculating the max Minor Allele Frequency (MAF). We take special considerations for some intergenic variants and have a separate max MAF. We retain variants which have been previously reported as Pathogenic/Likely Pathogenic in disease databases such as ClinVar which also includes drug response variants, HGMD, LOVD, OMIM and SwissVar.

B Additional Results

B.1 Selecting the number of clusters

We swept the number of Ward clusters from $k = 4$ to 30 and tracked silhouette score (geometric separation) and gene AMI (chance-corrected agreement between cluster assignments and ground-truth causal genes). Silhouette peaks at $k = 8$ but remains low across the range (0.07–0.12), consistent with a continuous rather than sharply partitioned phenotype space, whereas gene AMI rises steeply up to $k \approx 15$ ($\geq 92\%$ of its maximum) and gains little thereafter; since our goal is etiological coherence rather than geometric separation, we selected $k = 15$, which also avoids degenerate clusters (minimum size 673).

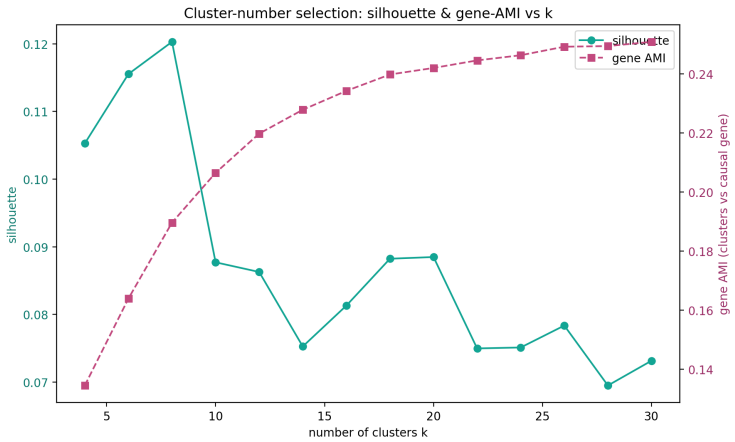


Figure 6: Silhouette score (left axis) and gene AMI (right axis) versus number of clusters k for Ward clustering of PRISM patient embeddings (training cohort).

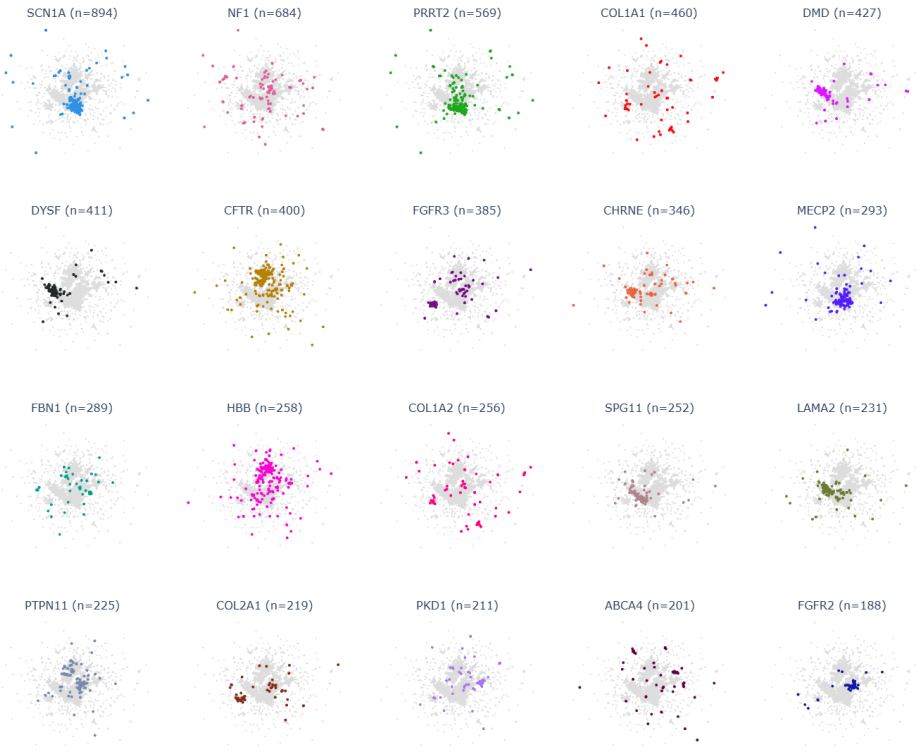


Figure 7: Patient clusters highlighted by their respective causal genes, sorted by most frequent causal genes in the training cohort, visualized using UMAP. Patients sharing a causal gene form tight, well-separated neighborhoods for most.

B.2 GENE LEVEL NEIGHBORHOODS IN THE EMBEDDING SPACE

In figure 7 each panel highlights the training cohort patients carrying one of the 20 most frequent causal genes on a shared UMAP projection (McInnes et al., 2020) of PRISM patient embeddings. Genes with distinctive phenotypes form compact neighborhoods (e.g., SCN1A, PRRT2, MECP2, DMD, DYSF, LAMA2, CHRNE, FGFR2), and genes with related phenotypes occupy adjacent regions (the epilepsy/neurodevelopmental genes SCN1A, PRRT2 and MECP2; the neuromuscular genes DMD, DYSF and LAMA2). Phenotypically variable genes such as NF1, COL1A1/COL1A2, CFTR, PKD1 and ABCA4 are more dispersed, reflecting the heterogeneity of their clinical presentations.