BMC Bioinformatics

SOFTWARE Open Access

A clinical knowledge graph-based framework to prioritize candidate genes for facilitating diagnosis of Mendelian diseases and rare genetic conditions

Rohan Gnanaolivu[1], Gavin Oliver[1], Garrett Jenkinson[1], Emily Blake[1], Wenan Chen[1], Nicholas Chia[2], Eric W. Klee[1] and Chen Wang[1*]

*Correspondence:
wang.chen@mayo.edu

[1] Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA
[2] Department of Surgery, Mayo Clinic, Rochester, MN 55905, USA

Abstract

Background: Diagnosing Mendelian and rare genetic conditions requires identifying phenotype-associated genetic findings and prioritizing likely disease-causing genes. This task is labor-intensive for molecular and clinical geneticists, who must review extensive literature and databases to link patient phenotypes with causal genotypes. The challenge is further complicated by the large number of genetic variants detected through next-generation sequencing, which impacts both diagnosis timelines and patient care strategies. To address this, in silico methods that prioritize causal genes based on patient-derived phenotypes offer an effective solution, reducing the time involved in diagnostic case reviews and enhancing the efficiency of clinical diagnosis.

Results: We developed the phenotype prioritization and analysis for rare diseases (PPAR) to rank genes based on human phenotype ontology (HPO) terms, with the specific goal of aiding the interpretation of genetic testing for Mendelian and rare diseases. PPAR leverages embeddings from a knowledge graph and incorporates knowledge from connections between genes, HPO terms, and gene ontology annotations. When applied on a clinical rare disease cohort and the publicly available deciphering developmental disorders (DDD) dataset. PPAR ranked the causal gene in the top 10 for 27% of cases in the clinical cohort and for 85% of cases in the DDD dataset, outperforming other established HPO-based methods.

Conclusion: Our findings demonstrate that PPAR, a method developed from the clinical knowledge graph, effectively ranks causal genes based on patient-derived HPO terms in rare and Mendelian disease contexts. PPAR has shown superior performance compared to other well-established HPO-only methods and provides an efficient, accessible solution for clinical geneticists. The Python-based tool is publicly available at https://github.com/dimi-lab/PPAR, offering a user-friendly platform for gene prioritization.

Keywords: Knowledge graph, Rare disease, Gene prioritization, Human Phenotype Ontology

Background

Diagnosing rare and Mendelian diseases is often challenging and requires labor-intensive genetic testing and extensive literature research. Molecular and clinical geneticists use in silico methods to facilitate rapid and accurate diagnoses. One such approach involves predicting the causal relationship between genes and disease phenotypes. To date, approximately 6466 phenotypes are mapped to 4544 single gene disorders, and 4909 have been identified with phenotype-causing mutation [1]. Exome and genome data are playing an increasingly vital role in the diagnosis of rare diseases, with the genotype–phenotype associations further elucidating the study of complex biological processes that involve the coordinated expression and interaction of proteins [2, 3]. Despite comprehensive testing with these techniques, the reported diagnostic rates are only 24–34% [4–6]. Review of the large number of variants identified in a patient using exome and whole genome sequencing methods is often time-consuming, and the number of variants identified can hinder experts from identifying causative genes. Additionally, identifying causal events, such as splicing alterations or gene fusions, often necessitates the integration of multiomic technologies for a more comprehensive analysis. To improve diagnostic rates and reduce the time required, information from multiple omics technologies and knowledge from biological databases are often used to identify causative genes related to disease phenotypes.

To help prioritize and identify the causal genes that impact the phenotypic representation of the patient, several genotype–phenotype prediction tools have been developed that rely on the Human Phenotype Ontology (HPO). Notably, several of these tools do not require a patient-derived Variant Call Format (VCF) file as input, including PCAN [7], Phenolyzer [8], Phen2Gene [9], GADO [10] and CADA [11] which are referred to as "HPO-only" methods [12]. However, these tools are limited by the knowledge used to build their *in-silico* prediction model, and many have been shown to perform poorly on independent datasets [12]. Each year, new knowledge is being added to variant and rare-disease databases, such as ClinVar [13], the Online Mendelian in Man (OMIM) [1], Orphanet [14], the Developmental Disorders Genotype-to-Phenotype (DDD) [15, 16] and other databases. These resources enhance our understanding of the genotype–phenotype relationship by elucidating the underlying biological activities. Despite this wealth of information, integrating continuous knowledge from multiple sources remains challenging due to factors such as data heterogeneity, storage, and integration difficulties. Additionally, knowledge sources are often created for different purposes and may not cover the same subjects, thereby complicating the interpretation.

In recent years, knowledge graphs have demonstrated their ability to handle complex, heterogeneous information, allowing new knowledge and insights that can be gleaned from the connections and directions of nodes generated from diverse resources. The edges between nodes represent binary links and are typically oriented in meaningful directions. Many groups have worked on creating extensive pre-built knowledge graphs that incorporate information from multiple biological databases. One such knowledge graph of interest is the Clinical Knowledge Graph (CKG), which comprises 20 million nodes and 220 million relationships, sourced from 26 biomedical databases and ten ontologies [6]. Among these, 50 million relationships involve publication nodes that

link scientific publications, creating connections for proteins, drugs, diseases, functional regions, and tissues, thus representing a comprehensive biological knowledge network.

A knowledge graph of this magnitude can only be hosted on a graph database platform like Neo4J. Neo4J's plugins enable visualization of the graph database, as well as modification, updating, deletion of nodes and relationships. These plugins also enable the generation of embeddings for every node and relationship type. Neo4J's graph data science library offers several embedding algorithms, with Fast Random Projection (FastRP) [17] being the most efficient. The combination of the CKG and Neo4J provides a unique opportunity to utilize harmonized knowledge from multiple data sources and analyze the genotype–phenotype relationships critical for diagnosing rare genetic disorders.

In this study, we developed PPAR, a prediction algorithm designed to use patient-derived HPO terms to predict and rank causal genes. The algorithm utilizes FastRP embeddings of gene and HPO nodes derived from a modified version of the CKG, hosted in a Neo4J graph database. Gene-HPO relevance is ranked based on the cosine similarity between the embeddings of the nodes. The algorithm incorporates predicted links and considers common parent genes associated with the given set of HPO terms provided as input, weighted by their information content (IC) to enhance the scoring. PPAR generates a rank-ordered list of the top K genes based on user-defined input for K and a set of HPO terms that represent the patient's phenotype. The predicted output list of rank-ordered genes would assist clinicians in evaluating and diagnosing patients. The Python-based tool is publicly available at https://github.com/dimi-lab/PPAR for easy access and use.

Implementation

PPAR workflow

Here, we describe the PPAR workflow, designed to prioritize causal genes for a given set of HPO terms (Fig. 1) and (Supplementary 1: Fig. S1). Initially, the pre-built CKG was set up on a Neo4J graph database platform using a CKG dump file. To reduce noise, we pruned the database by removing node types: "Experiment," "Units," "GWAS study," "Analytical sample," "Biological sample," "Subject," "Project," and "User," as they were not relevant to our study. We updated the CKG with the 2023–09-01 release of the HPO, which contains gene-to-HPO and HPO-to-disease mappings by using the cypher query language within Neo4J. This data was sourced from the Open Biological and Biomedical Ontologies (OBO) library (http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt). Using Cypher query language, we linked HPO terms to gene nodes in the direction from gene to HPO. Additionally, Gene Ontology (GO) information version 2024–06–17, sourced from protein annotation through evolutionary relationship (PANTHER) database v.17 [18], was incorporated into the CKG, establishing connections from GO terms to genes through Cypher queries. In total, the modified CKG contained 24 different node types and 27 different relation types (Supplementary 2: Table S1). Next, we generated gene and HPO node embeddings using the FastRP algorithm via Neo4J's graph data science library v2.13. The dataset included 19,231 protein-coding genes and 8897 HPO terms from the OBO library. In developing PPAR, we integrated multiple factors, including the IC of each HPO term, the probabilities from the link prediction task between genes and HPO terms, the cosine similarity between the embeddings from gene and

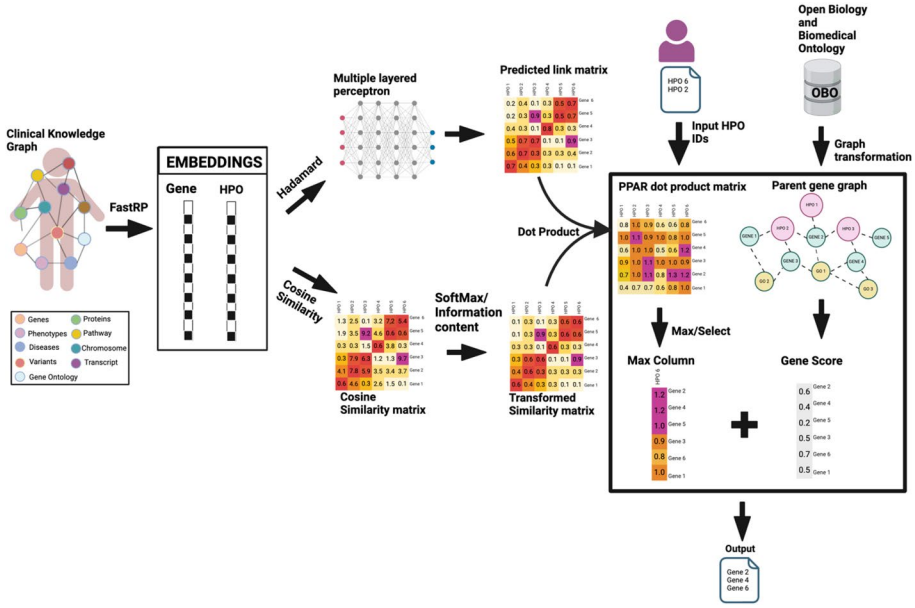

Fig. 1 Illustration figure of the development of the PPAR algorithm for gene-HPO prioritization. Gene and HPO embeddings are generated using the FastRP algorithm on a Clinical Knowledge Graph (CKG) that integrates data from multiple biological resources. These embeddings are processed through a multiple-layer perceptron (MLP) with Hadamard link operations, resulting in a predicted link matrix. Cosine similarity is computed between embeddings, followed by a normalization step, producing a transformed similarity matrix. The PPAR dot product matrix is then generated by combining these matrices via a dot product. A parent gene graph is constructed using information from the Open Biology and Biomedical Ontology (OBO) database, with user-provided HPO terms guiding the extraction and computation of gene scores. The PPAR dot product matrix is filtered and sorted with the input HPO terms, followed by the summation of the ordered genes matrix values with the gene scores. The final output is a prioritized list of genes, ranked based on their relevance to the input HPO terms

HPO nodes, and the scoring of parent genes connected to the patient-identified HPO terms. This approach resulted in creating a PPAR relationship matrix, a $19{,}231 \times 8897$ matrix representing comprehensive gene-HPO relationships, and an independent graph network that incorporates the connections between the genes, HPO terms, and GO terms using publicly available information [19, 20]. Using this static matrix and the graph network, the PPAR model generated a rank-ordered list of genes based on a given set of HPO terms. PPAR demonstrated efficient performance, making it well-suited for clinical and research applications that require timely results.

Neo4J and CKG integration

To set up the CKG, we used Neo4j Desktop, which was downloaded from www.neo4j.com. We built the CKG database using Neo4j database version 4.2.3 from the dump file, following the instructions available at CKG Builder. The system requirements for building a single instance of the CKG database included a memory capacity of 16 GB and disk space of at least 200 GB. To accommodate the large CKG database, we modified the heap memory settings in Neo4j to 180 GB (dbms.memory.heap.initial_size = 180 g, dbms.memory.heap.max_size = 180 g) while keeping the other settings at their defaults. Additionally, we installed two essential plugins: APOC version 4.2.0.5 and the Graph

Data Science Library version 1.5.1. These plugins facilitated data queries, visualization, and the generation of node embeddings. Using Cypher query language, we modified the CKG database, including deleting nodes that added noise and added new relationships to refine and update the database for our study, which resulted in 24 node types and 27 relationship types.

FastRP embeddings

To generate embeddings from the CKG, we used the FastRP algorithm, available in the Neo4j Graph Data Science library. FastRP initializes each node with a random vector and iteratively refines these embeddings by incorporating information from neighboring nodes, effectively capturing both local and extended neighborhood structures. Given the size of the CKG, FastRP was the most suitable choice over other embedding algorithms, such as GraphSage [21] and Node2Vec [22]. We generated embeddings for all gene and phenotype node types by configuring the hyperparameter "embeddingDimension" to 1024, the maximum available dimension, as this enabled the capture of more detailed and comprehensive information. We also modified the "iterationWeights" hyperparameter to (2, 1.5, 1.2, 1, 0.8, 0.6, 0.4, 0.2), assigning larger weights to closer neighbors and applying these weights to up to eight neighboring nodes. Additionally, we set the "normalizationStrength" to −0.75, downplaying the influence of higher-degree neighboring nodes. Using Cypher queries, we extracted all the embeddings from the gene and phenotype nodes and saved them in feather files for further analysis.

FastRP fine-tuning

We selected five cases from an internal Mayo Clinic rare disease cohort (MCRD) cohort that were representative of both direct and indirect associations between HPO terms and their causative genes in the CKG. These cases were derived from separate genetically diagnosed patients and were chosen to illustrate both direct and indirect gene-phenotype associations, providing a robust framework for evaluating and refining the embeddings. The first scenario demonstrates a direct association between a causative gene and an HPO term, representing the simplest connectivity pattern. The second scenario highlights an indirect association, where the relationship is mediated through an intermediary node, showcasing more complex connectivity. The third scenario involves a single HPO term connected to a causal gene that is further associated with multiple additional HPO terms, reflecting a broader phenotype spectrum. The fourth scenario focuses on cases where the input HPO terms have low IC, which presents challenges for accurately resolving gene-phenotype relationships. Finally, the fifth scenario examines cases with multiple input HPO terms spanning a range of IC values, including both low and high IC HPO terms, to assess the impact of IC heterogeneity. Together, these cases offered a comprehensive examination of the connectivity, while utilizing the minimum number of datapoints to evaluate and fine-tune the embeddings (Supplementary 2: Table S2). Based on the ranking of these cases, the CKG nodes were pruned and the FastRP hyperparameters, such as the "iterationWeights" and "normalizationStrength" were tuned. We manually evaluated various degrees lengths of the "iterationWeights" and evaluated values ranging from −1 to +1 for the "normalizationStrength".

Information content (IC)

The IC was calculated for every HPO term to estimate the degree of connectivity. To derive the IC, we used the equation listed below.

Let n be the total number genes found in the OBO library and H_d the degree of connected genes to a HPO term. We define the IC as follows:

$$IC_h = \frac{H_d}{n}$$

Here IC_h is the information content for HPO term (h), where H_d is the degree of connected genes to the HPO term.

Link prediction

To evaluate the connectivity and quality of the connections between the gene and phenotype nodes, a link prediction task was performed to predict the probability of a gene node being connected to a HPO node. We evaluated four different link operation methods and coupled them with four different ML/Deep learning algorithms to derive the most accurate probabilities for these links. The link operation methods we selected for evaluation were Hadamard, Average, L1, and L2, as they measure various aspects of similarity. The Hadamard product measures the similarity by element-wise multiplication of two vectors, the L1 distance is the Manhattan distance that measures dissimilarity by calculating the sum of absolute differences between the two vectors. The L2 distance, or Euclidean distance, calculates the straight-line distance between two vectors, and the Average method computes the midpoint between the two vectors to evaluate their similarity.

Let E_g and E_p represent the embedding matrices for the gene and HPO nodes, respectively. We define Hadamard, Average, L1 and L2 as follows:

$$Hadamard(E_g, E_p) = f(E_g).f(E_p)$$

$$Average(E_g, E_p) = \frac{f(E_g) + f(E_p)}{2}$$

$$L1(E_g, E_p) = |f(E_g) - f(E_p)|$$

$$L2(E_g, E_p) = |f(E_g) - f(E_p)|^2$$

Eight trials were conducted to evaluate the links between genes and HPO nodes. In each trial, the total number of non-linked gene-HPO pairs were increased to simulate the approximate imbalanced nature of the connections in the knowledge graph. The dilution of the non-linked nodes varied incrementally, from 1x to 8x, where $1\times$ represented a balanced scenario with equal number of connected and non-linked nodes between the genes and HPO terms. Each subsequent trial increased the imbalance, with 2x, 3x, and higher dilutions introducing progressively more non-linked pairs (Supplementary 2: Table S3). A binary classification problem was set up, with a class of "1" indicating

a gene-HPO connection and a class of "0" indicating no connection. The ground truth labels for the connections were based on the gene to HPO connections listed in the OBO library. Using the embeddings from the gene and HPO nodes, we applied the four link operation methods Hadamard, Average, L1, and L2 to every gene-HPO pair across all trials. In each trial, the data was split into training and validation sets in an 80:20 ratio, with the validation set further split 50:50 to generate the test set.

To establish the prediction probability for the link operation methods, we compared the prediction outcomes from four different models: XGBoost, Naïve Bayes, Random Forest, and Multi-Layer Perceptron (MLP). The metrics used for comparing the classification performance were accuracy and area under the precision-recall curve (AUCPR). The model with the highest accuracy and AUC-PR, was chosen to predict the links between all possible gene and HPO nodes, which resulted in a matrix (L), with dimensions19,231 × 8897.

Similarity measure

To evaluate the similarity of genes and HPO nodes within the embedding space, we employed a cosine similarity measure. Let E_g and E_h represent the embedding terms for a gene and HPO node, respectively. The cosine similarity between these vectors is defined as follows:

$$Cosine\ Similarity = \frac{\mathrm{E_g}.\ \mathrm{E_h^T}}{|\mathrm{E_g}||\mathrm{E_h^T}|}$$

Parent gene

To evaluate the parent genes connected to a provided list of HPO terms, we initially created an undirected graph comprising of genes, HPO terms and GO terms as nodes. Let $G = (V, E)$ represent an undirected graph, where V is the set of nodes and E is the set of edges. Let $D = \left\{ \left(g_i, h_i, go1_i, go2_i\right) \right\}_{i=1}^{n}$ be the dataset D from the OBO library containing the relationship between the HPO term h_i and genes g_i, along with the interaction of the GO terms $(go1_i, go2_i)$. We create the graph G, from each row of the dataset, by adding edge E, between the gene g_i and the HPO term h_i, defined as follows:

$$E = EU\left\{ \left(g_i, h_i\right) \right\} \forall_i$$

We then add an edge E between the gene g_i and $go1_i$, defined as follows:

$$E = EU\left\{ \left(g_i, go1_1\right) \right\} \forall_i$$

Finally, we add another edge E between $go1_i and go2_i$, only if there exist a connection. Defined as follows:

$$E = EU\left\{ \left(go1_i, go2_i\right) \right\}\ if\ go\ 2_i \neq NaN$$

Utilizing the created graph (G), we then derive a gene score for all genes connected to a HPO term (g_h) in the graph, based on the in input HPO terms $H = \left\{h_1, h_2, \ldots, h_p\right\}$

From graph (G), the count of connected genes (C_g) was initially calculated from H, defined below:

$$C_g = \sum_{h \in H} 1\{gG_h\}, else0$$

The equation defines a function $\{gG_h\}$ that returns 1 if gene (g) is connected to a HPO term (h).

We utilize the IC to weight the connected HPO term. We created a weighted sum (W_g) for all connected genes to the HPO terms (g_h). We define (W_g) as follows:

$$W_g = \sum_{h \in H} 1 - IC_h$$

Finally, to create the score (S_g) for each gene (g) connected to HP terms (H). We define (G_{common}) as the intersection of gene sets connected to each phenotype in H.

For each $(g \varepsilon G_{common})$, the gene score (S_g) is computed as follows:

$$\mathrm{S_g} = C_g \times W_g$$

PPAR model

In summary, the final algorithm we developed to prioritize and rank genes is as follows:

Let E_g and E_p be the embedding matrices for the gene and HPO nodes, respectively. We initially compute the cosine similarity matrix X as defined below:

$$Cosine\ Similarity(X) = \frac{\mathrm{E_g}.\ \mathrm{E_h^T}}{|\mathrm{E_g}||\mathrm{E_h^T}|}$$

It is in this step, we measure the semantic similarity X. where (X) comprises of a set of HPO terms h_i, for gene g_i from the corresponding row.

We then calculate the stable SoftMax (S) for matrix X in a two-step process, as follows:

We first scale the matrix and then calculate the SoftMax (S) as defined below:

$$S = \frac{\exp\left(X_{ij}\right)}{\sum_k \exp(X_{ik})}$$

where i and j represent the row and column of the matrix (X), and k represents the index that iterates over all the columns. Here, we normalize the similarity to bring it to a probability scale.

To account for the connection between genes to common representative HP terms, we weighted the matrix (S) with IC_p.

Let IC_p be the information content for a HPO term, and we define a new matrix F as follows:

$$F = \frac{\mathrm{S}}{\mathrm{IC_p}}$$

In this step, we assign greater importance to phenotypes with lower degrees, prioritizing those with fewer connections.

Then we create another matrix (Q_{gh}) by incorporating the link prediction matrix as defined in the methods, and by calculating the dot product of the matrix F and link

prediction matrix (L). Q_{gh} represents a matrix where each row corresponds to a gene, each column a HPO term. Q_{gh} represents the similarity score between the gene and HPO term and is defined as follows:

$$Q_{gh} = \mathrm{F} \times \mathrm{L}$$

In this step, we enhance the matrix F with the probabilities from link prediction task to complement the predictive inference of potential novel links between the genes and phenotypes. The matrix $\left(Q_{gh}\right)$ comprises 19,231 genes and 8897 HPO terms, in a 19,231 × 8897 matrix. Q_{gh} is also referred to as the PPAR dot product matrix.

To rank order the genes for a given set of HPO terms $H = \left\{h_1, h_2, \ldots, h_p\right\}$, matrix Q_{gh} is initially filtered to the columns containing only the HPO terms from the input list. This is described as follows:

$$H_f = H \cap columns\left(Q_{gh}\right)$$

Next, we ranked the genes (rows) $\left(g\right)$ on the maximum values found across the HP columns, which represent the highest value for each gene. This is described as follows:

$$MaxScore\left(g\right) = \max_{\mathrm{h} \in \mathrm{h_f}} Q_{gh}$$

As described in the methods, we simultaneously calculate the parent genes score (S_g) with H_f provided as the input. To equally weight the (S_g) on a scale that comparable to $MaxScore\left(g\right)$, we compute the standard deviation (σ) of $MaxScore\left(g\right)$, denoted as σ_{Max}

Finally, we create the PPAR algorithm, described as follows:

$$PPAR = MaxScore\left(g\right) + S_g \times \sigma_{Max}$$

PPAR validation

The internal MCRD cohort of 229 genetically diagnosed rare Mendelian disease cases was used to build and validate the model (Table 1 and Supplementary 2: Table S4). Each case included an average of 8 HPO terms. This cohort was collected over many years from the Center for Individualized Medicine division at Mayo Clinic. It includes 125 males and 104 females, with ages ranging from 0 to 81. Most causal genes were uniquely identified in the cohort, with the exceptions of *CACNA1A*, identified in 4 cases, and *CHD2*, *CLN6*, *CTCF*, *SETD5*, and *MECP2* each identified in 3 cases. For fine tuning the parameters from FastRP, we used 5 test cases to fine tune the FastRP algorithm, carefully selected to ensure diversity across five distinct scenarios. These scenarios were designed to capture variations in the network connections between genes and HPO terms within the CKG and the IC of the connected HPO terms. This approach allowed us to optimize the algorithm for a wide range of connectivity patterns and semantic complexities. These cases included both high (> 8) and low (< 3) numbers of HPO terms, along with causal genes that had either direct or indirect links to the HPO terms, providing a broad representation for effective optimization.

We also evaluated PPAR on a separate cohort containing cases of developmental disorder in the DDD study [15]. This cohort includes 1133 cases with an average of 22 HPO

Table 1 Description of the cohorts used in the validation of PPAR

Cohort		MCRD	DDD
Basic information	Size	229	1133
	Average HPO count	8	22
Gender	Male	125	550
	Female	104	583
Age (years)	0–18	148	1133
	18–65	78	NA
	65 +	3	NA
Gene (frequency)	> 5	0	*COL2A1, FLNA, GDF5, FGFR3, FGFR2, PAX6, COL1A1, TP63, FBN1, PTEN*
	5		*COL11A2, CEP290, GJA1, NOG, HOXD13, ARX, FLNB, LRP5, GJB2*
	4	*CACNA1A*	*FKTN, GLB1, SLC26A2, ERCC6, FGFR1, TMEM67, MECP2, L1CAM, IKBKG, PITX2, PRPS1, SHH, NF1, PTH1R, SOX10*
	3	*CHD2, CLN6, CTCF, SETD5, MECP2*	*NDUFS4, CRYBB2, ATP7A, CASK, VSX2, CRYGD, CHD7, TRPS1, SCN4A, GNAS, FKRP, GATA6, TGFBR1, OFD1, POMT1, ERCC2, DMD, FOXC1, RECQL4, RPGRIP1L, CTNS, NKX2-5, CC2D2A, NPHP1, MITF*

terms per case. It includes 550 males and 583 females, primarily comprised of children with a median age of 5.5 years. This cohort exhibits a wide range of phenotypes such as developmental delay, hearing impairment, seizures, visual impairment, scoliosis, autism spectrum disorder, oral cleft, congenital heart defects, and polydactyly. Several genes were identified as causal in more than 5 cases, with *COL2A1*, *FLNA*, and *GDF5* appearing with a frequency of 8 or greater.

Statistical analysis

We utilized Python packages Scikit-learn (version 1.0.2) [23], Pandas (version 1.3.5), and NumPy (version 1.21.6) for statistical analysis. For the embedding generation step, we used the FastRP algorithm from the Neo4j data science library and a Python plugin built into Neo4j to save the embeddings of the gene and phenotype nodes via Cypher queries. During the link prediction step in the development of PPAR, we evaluated three prediction methods using scikit-learn implementations of the Random Forest classifier, Naïve Bayes classifier, and MLP classifier. Additionally, we evaluated the classifications from the XGBoost model using Python package XGBoost (version 1.6.2). To identify the best classifier for each link prediction operation, we calculated precision, recall, area under the precision-recall curve (AUCPR), and accuracy metrics using scikit-learn. For determining the similarity between gene and phenotype nodes, we used the cosine similarity method from scikit-learn. The plots were generated using the Python packages seaborn (version 0.11.2) and matplotlib (version 3.5.3).

Results

The aim of our method was to prioritize and rank genes based on a given set of patient-derived phenotypes. To address this, we introduce PPAR, to assess gene-phenotype relationships and rank genes based on specified HPO terms. Our approach involves generating scores to predict these associations, by leveraging FastRP-generated embeddings

of genes and HPO nodes from the CKG. Additionally, our method includes the probabilities for predicting links between every gene and HPO node. To do this, we evaluated four linking methods (Hadamard, Average, L1, and L2), followed by comparing the prediction probabilities derived from Random Forest, XGBoost, Naïve Bayes, and MLP models. Given the CKG dataset consisted of an imbalance between linked and non-linked node pairs, we used the metric AUCPR and accuracy to evaluate performance. Our results revealed that in general, the performance from these predictive models, coupled with a link operation method were comparable in terms of the AUCPR and accuracy metrics. From our analysis (Fig. 2A), the Hadamard product link method combined with MLP-generated probabilities achieved the highest mean accuracy of 0.88. Conversely, when considering AUCPR (Fig. 2B), the L1 link method combined with the Naïve Bayes probabilities achieved a mean AUCPR of 0.75 and Hadamard product method with MLP-generated probabilities resulted in mean AUCPR of 0.73. Based on these findings, for developing PPAR, we opted for MLP-generated probabilities with Hadamard product link operations due to their superior mean accuracy (0.88) and competitive mean AUCPR (0.73) compared to the other methods.

PPAR performance

To evaluate the performance of PPAR, we compared its predictions with other well-established HPO-only methods, PCAN, Phen2Gene, GADO and CADA, using 1133 cases from the DDD dataset and 229 cases from the MCRD cohort, all of which had confirmed genetic diagnoses in the clinical reports. Phen2Gene v1.2.3 was utilized for analysis, installed via Miniconda, with results generated in accordance with the official documentation provided on its GitHub repository. GADO v2.0 was employed to predict the top 100 genes for each case through Application Programming Interface (API) calls. Custom python code was developed to automate these API requests, with the API returning the top 100 genes for each case. CADA was installed on a Linux platform using pip, and following the documentation provided on its GitHub repository, causal ranked gene prediction was performed.

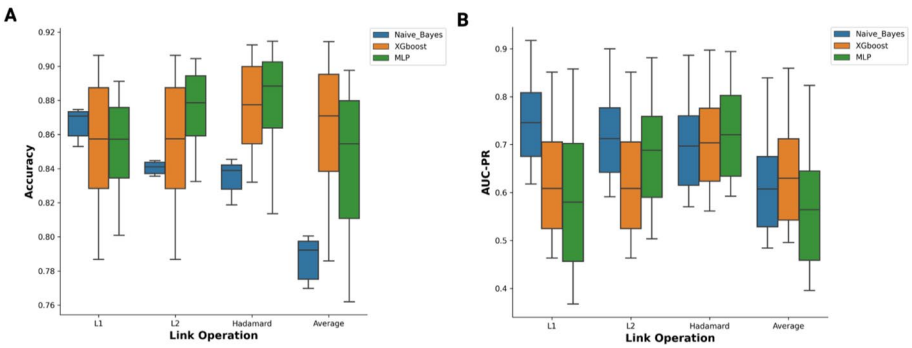

Fig. 2 Link prediction task results of predicting links between genes and HPO nodes. **A** Accuracy metric of (Hadamard, Average, L2 and L2) link operation method with the predicted probabilities from XGboost, Naive Bayes and MLP across several iterations by varying the total number of non-linked nodes. **B** AUCPR metric of (Hadamard, Average, L2 and L2) link operation method with the predicted probabilities from XGboost, Naive Bayes and MLP across several iterations by the varying the total number of non-linked nodes

In the DDD dataset, 37 cases were excluded due to missing HPO terms in the dataset or the absence of the causal gene in the CKG. Additionally, 47 cases were removed as it contained exact duplicates of the causal gene and HPO terms. This resulted in a total of 1049 cases available for evaluation. The distribution of the total number of phenotypes per case revealed a wide range of complex phenotypes associated with each case (Fig. 3A, B), with some cases having more than 100 phenotypes with the DDD dataset, and 25 in MCRD cohort, each case resulting in single identified causal gene.

A prioritized gene list was successfully generated for all 1049 cases in the DDD cohort and 229 cases in the MCRD cohort using PPAR. However, prioritized lists could not be generated for all cases using PCAN, Phen2Gene, GADO, and CADA due to limitations in HPO term recognition or the absence of the causal gene in the respective tool vocabularies. In the DDD cohort, 11 cases contained causal genes that were not recognized by PCAN. Similarly, Phen2Gene failed to rank the causal gene for 18 cases in the DDD cohort and 1 case from the MCRD cohort due to the absence of these genes in its vocabulary. GADO failed to recognize the HPO terms for 2 cases in the DDD cohort and 3 cases in the MCRD cohort, instead suggesting alternative HPO terms. CADA exhibited similar limitations, with 287 cases in the DDD cohort and 10 cases in the MCRD cohort contained HPO terms that fell outside its recognized vocabulary.

In the DDD dataset, PPAR identified the top causative gene in 70% of the cohort, outperforming PCAN, which ranked the top gene in 57% of the cases, followed by CADA (47%), Phen2Gene (46%) and GADO (1%). Within the top 10 ranked genes, PPAR ranked 82% of the cases, compared to PCAN at 76%, CADA and Phen2gene at 70% each,

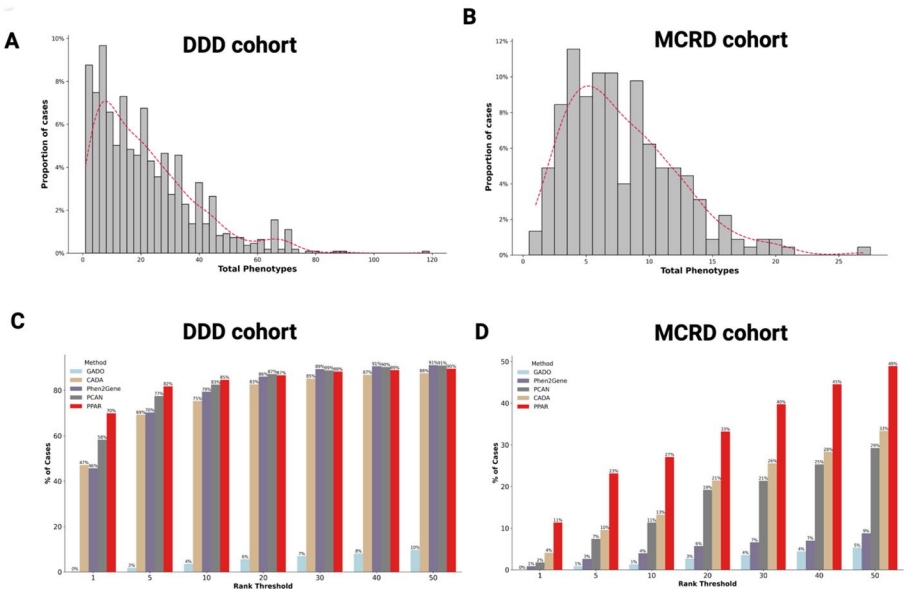

Fig. 3 Performance of PPAR from the DDD and MCRD cohort. **A** Distribution of the HPO term counts found in the DDD cohort, with the red distribution curve highlighting the mean count to be 22. **B** Distribution of the HPO term counts found in the Mayo Clinic rare disease cohort, with the red distribution curve highlighting the mean count to be 8. **C** Ranked bins illustrating the percentage of cases ranked in bins from 1–50 for PPAR, PCAN and Phen2Gene across all 1049 cases from the DDD cohort. **D** Ranked bins illustrating the percentage of cases ranked in bins from 1–50 for PPAR, PCAN and Phen2Gene across all 229 cases from the cases from the Mayo Clinic rare disease cohort

and GADO at just 2% of the cases (Fig. 3C). A comparison of the cumulative distribution of ranks demonstrated that PPAR ranked a larger portion of the cases at lower rank (< 20), although Phen2Gene ranked 16% cases better than PPAR at a higher rank threshold (Supplementary 2: Table S5).

In the MCRD cohort, PPAR similarly outperformed Phen2gene, CADA, PCAN, Phen-2gene and GADO, by identifying the top causal gene in 11% of the cohort, compared to CADA (4%), PCAN (2%), Phen2Gene (1%) and GADO, which identified none. Within the top 10 ranks, PPAR ranked 27% of the cohort, followed by CADA (13%), PCAN (11%), Phen2Gene (4%) and GADO (1%) (Fig. 3D). The cumulative distribution function indicated that PPAR ranked majority of the cases higher than PCAN, CADA, PCAN, Phen2Gene and GADO (Supplementary 2: Table S6).

PPAR case review

We highlight a case where a patient exhibited 15 different phenotypes, including *Abnormality of chromosome stability* (HP:0003220), *Abnormal hair morphology* (HP:0001595), *Arachnodactyly* (HP:0001166), *Deeply set eye* (HP:0000490), *Intellectual disability profound* (HP:0002187), *Global developmental delay* (HP:0001263), *Hearing impairment* (HP:0000400), *Optic nerve hypoplasia* (HP:0000609), *Pectus carinatum* (HP:0000768), *Periventricular leukomalacia* (HP:0006970), *Pes planus* (HP:0001763), *Postauricular skin tag* (HP:0004451), *Seizure* (HP:0001250), *Short stature* (HP:0004322) and *Thoracic scoliosis* (HP:0002943). An evaluation of the distribution of these HPO terms associated with top ranked genes from the PPAR model revealed the most influential HPO term contributing to the overall prediction of the rank (Fig. 4A). Notably, the HPO term HP:0004451 corresponding to phenotype "Postauricular skin tag" was highlighted as the significant contributor to the rank. Evaluation of the top ranked genes by PPAR revealed that gene *NR2F1* was ranked as the top causal gene. Evaluating the connectivity from the custom graph to estimate the parent gene connection revealed that four HPO terms were connected to *NR2F1*, thus generating a high score for that gene (Fig. 4B). Finally, based on the review of the patient's clinical report, *NR2F1* was indeed identified as the causative gene contributing to the patient's disease.

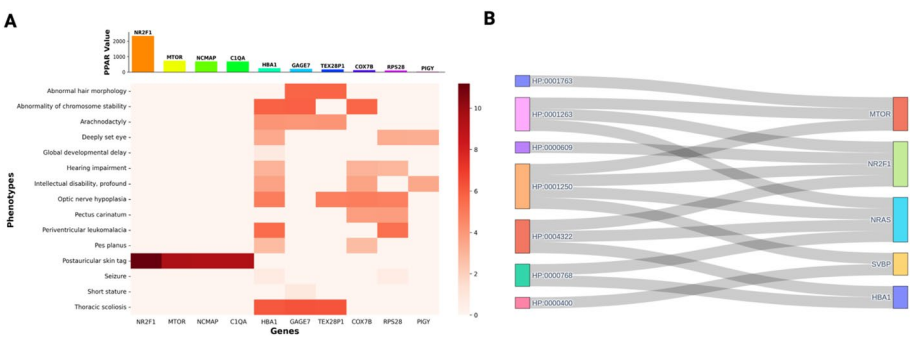

Fig. 4 PPAR results from a single case in the MCRD cohort. **A** PPAR scores between the phenotypes and the top-ranked genes, highlighting the strong association of the phenotype "Postauricular skin tag", with gene NR2F1 causality compared to other genes. **B** Illustration of the connectivity between the HP terms and the genes derived from the custom graph. The figure highlights the association of NR2F1 with four out of the seven HP terms provided as the input, demonstrating its central role in this case

PPAR functionality

We developed PPAR to efficiently predict top ranked genes by utilizing the pre-built static PPAR matrix containing the similarity scores for 19,231 protein coding genes and 8897 HPO terms, along with the pre-built graph that contains genes, HPO terms and GO terms as nodes (Supplementary 1: Fig. S2). The PPAR model, implemented in Python version 3.7, takes a list of HPO terms from a single case as the input. It then subsets the matrix based on the provided HPO terms and computes the parent gene scores from the pre-built graph. The model then generates the PPAR scores from all genes, which are subsequently ranked from highest to lowest. By default, PPAR outputs a ranked list of 100 genes, although this number can be altered by a user defined input for the parameter k.

Discussion

Here we describe PPAR, a novel HPO-only gene prioritization algorithm that utilizes the embeddings generated from a clinical knowledge graph to identify and rank causal genes in the embedding space for a given set of HPO terms. PPAR exploits the knowledge within the CKG, built in a Neo4J graph data platform and uses the FastRP embeddings to understand the neighborhood within the gene and phenotype nodes. FastRP is a dimensionality reduction method, that generates its embedding based on its construction of a sparse random projection matrix between each data point, followed by a similarity matrix construction. In addition to the embeddings, PPAR utilizes a custom constructed graph with genes, HPO terms and GO terms in the OBO library, as nodes. The relationship between the nodes is defined by the data within the OBO library, enabling PPAR to integrate structured knowledge into its predictions.

Many existing phenotype-to-gene prioritization tools rely on APIs that often require institutional licenses and may necessitate the use of patient-derived VCF files to return ranked predictions of causal genes. This dependency can raise concerns about sharing sensitive patient data externally, particularly in regard to Health Insurance Portability and Accountability Act (HIPAA) regulations, especially when participation is voluntary. In contrast, PPAR offers a Python-based HPO-only method for gene prioritization that operates independently of external APIs, providing a more secure and accessible solution for identifying causal genes. The performance of PPAR has shown to be superior to other well-known HPO-only methods, namely, Phen2Gene, PCAN, CADA, and GADO, as evidenced by evaluations conducted on two independent cohorts in this study. The substantial variation in the predictive performance among these methods between the two datasets can likely be attributed to the extensive gene-to-phenotype associations from the 2015 DDD study. These associations were incorporated into multiple biological databases, which served as the foundation to build these predictive models.

Clinical and molecular geneticists often spend hours diagnosing a single rare or Mendelian disorder case due to the complexity and limitations in understanding the underlying genetics of presenting phenotypes. PPAR is an HPO-only method designed to generate a ranked list of candidate genes to aid genetic experts in rapidly diagnosing patients. This rank-ordered list can aid in results prioritization when interpretating data generated from clinical exome or genome sequencing tests. PPAR could also be used

for pre-test evaluation, leading to targeted panel testing instead of exploratory exome or genome sequencing in some cases. Other HPO-only methods, like PCAN, rely solely on pathway and Protein–Protein Interaction (PPI) network information to compute semantic similarity scores for their predictions. PCAN is also confined to the total number of genes found in the ClinVar database, whereas PPAR utilizes all protein coding genes. Similarly, methods such as Phen2Gene and CADA also utilize knowledge resources, however they are less comprehensive compared to PPAR. A notable limitation of CADA and GADO is their restricted vocabulary of HPO terms, which constrains their applicability for certain use cases.

A limitation to the PPAR model is that incorporating novel information or integrating updates from external databases into the CKG necessitates regeneration of the FastRP embedding and retraining of the PPAR model. This process requires large computational resources, and it is a time-consuming process. The current implementation of the PPAR model utilizes precomputed static embeddings to ensure efficient result generation, and this computation does not require GPU resources. Additionally, PPAR's reliance on the knowledge databases within the CKG means it does not have a comprehensive list of all novel HPO terms, and much of the information in this database has not been updated since its initial release. This may result in the model missing critical updated or newly discovered gene-phenotype associations.

In summary, the developed PPAR method has demonstrated high accuracy in ranking genes, offering valuable support to clinicians in diagnosing rare and Mendelian diseases. PPAR successfully utilizes a knowledge graph that integrates information from 24 different biological databases and 10 ontology databases which includes hundreds of thousands of variants and several thousand peer reviewed publications to benefit patients with rare or Mendelian disorders by predicting the causal genes from a set of disease-associated phenotypes [24].

Conclusions

PPAR is an open-source gene-HPO prioritization algorithm developed in Python. We developed PPAR to be VCF-agnostic, requiring only HPO terms as input. It leverages graph embeddings from the CKG, integrating databases that represent information for 19,231 protein-coding genes and 8897 HPO terms. PPAR outperforms other HPO-only methods in ranking causal genes for rare or Mendelian diseases. Optimized for efficiency and speed, it is freely available at https://github.com/dimi-lab/PPAR.

Availability and requirements

- Project name: PPAR
- Project home page: https://github.com/dimi-lab/PPAR
- Operating system: All operating systems supporting Python
- Programming language: Python
- License: MIT
- Any restriction to use by non-academics: None

Abbreviations

API	Application programming interface
AUCPR	Area under the precision-recall curve
CKG	Clinical knowledge graph
DDD	Developmental disorders genotype-to-phenotype
FastRP	Fast random projection
HPO	Human phenotype ontology
HIPPA	Health insurance portability and accountability act
IC	Information content
MCRD	Mayo clinic rare disease
OBO	Open biological and biomedical ontology
OMIM	Online Mendelian inheritance in man
PCAN	Phenotype consensus analysis
PPAR	Phenotype prioritization and analysis of rare disease
PPI	Protein–protein interaction
VCF	Variant call format

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06096-2.

Supplementary material 1

Supplementary material 2

Acknowledgements
The DDD study presents independent research commissioned by the Health Innovation Challenge Fund- grant number HICF-1009-003. This study makes use of DECIPHER (http://www.deciphergenomics.org), which is funded by Wellcome [grant number WT223718/Z/21/Z]

Author contributions
All authors conceptualized the study and reviewed the manuscript. RG, GO, GJ and CW designed the methodology. RG developed the python program. RG and EB designed and performed the validation experiments. RG wrote the manuscript. EK, NC, WC and CW provided guidance and supervised manuscript writing. All authors read and approved the final manuscript.

Funding
This research is supported by the Mayo Clinic Center for Individualized Medicine. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The knowledge graph dataset used in this study is sourced from MannLabs and can be accessed https://datashare.biochem.mpg.de/s/kCW7uKZYTfN8mwg/download. The gene-phenotype associations utilized are available via the Human Phenotype Ontology at http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt. The PPAR model is accessible at https://github.com/dimi-lab/PPAR. The validation data, which includes the DDD and MCRD cohort used in the model's development, is publicly available and can be accessed at https://github.com/dimi-lab/PPAR/tree/main/data. The DDD cohort dataset used in this study was frozen on 7th November 2013 and is freely available as a supplementary file in https://doi.org/10.1016/S0140-6736(14)61705-0 on behalf of the DDD study (https://www.ddduk.org/). The MCRD cohort utilized in this study is an internally curated collection of rare disease patient data. Access to this data is regulated under Mayo Clinic Institutional Review Board #12–009346.

Declarations

Ethics approval and consent to participate
All the participants provided written informed consent to participate in this study. All protocols were approved by the Mayo Clinic Institutional Review Board. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 31 October 2024 Accepted: 25 February 2025
Published online: 14 March 2025

References

1. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 2009;37(Database issue):D793-6.
2. Ewans LJ, Minoche AE, Schofield D, Shrestha R, Puttick C, Zhu Y, et al. Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis. Eur J Hum Genet. 2022;30(10):1121–31.
3. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, Kingsmore SF. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med. 2018;3:16.
4. Helman G, Lajoie BR, Crawford J, Takanohashi A, Walkiewicz M, Dolzhenko E, et al. Genome sequencing in persistently unsolved white matter disorders. Ann Clin Transl Neurol. 2020;7(1):144–52.
5. Alfares A, Aloraini T, Subaie LA, Alissa A, Qudsi AA, Alahmad A, et al. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. Genet Med. 2018;20(11):1328–33.
6. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;19(5):253–68.
7. Godard P, Page M. PCAN: phenotype consensus analysis to support disease-gene association. BMC Bioinformatics. 2016;17(1):518.
8. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015;12(9):841–3.
9. Zhao M, Havrilla JM, Fang L, Chen Y, Peng J, Liu C, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. NAR Genom Bioinform. 2020;2(2):lqaa032.
10. Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. Nat Commun. 2019;10(1):2837.
11. Peng C, Dieck S, Schmid A, Ahmad A, Knaus A, Wenzel M, et al. CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genom Bioinform. 2021;3(3):lqa078.
12. Yuan X, Wang J, Dai B, Sun Y, Zhang K, Chen F, et al. Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. Brief Bioinform. 2022;23(2).
13. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.
14. Aymé S, Dallapiccola B, Donnai D. Orphanet journal of rare diseases: launch editorial. Orphanet J Rare Diseases. 2006;1(1):1.
15. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015;385(9975):1305–14.
16. Study TDDD. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2014;519(7542):223–8.
17. Chen H, Sultan SF, Tian Y, Chen M, Skiena S. Fast and accurate network embeddings via very sparse random projection. In: Proceedings of the 28th ACM international conference on information and knowledge management 2019. p. 399–408.
18. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8(8):1551–66.
19. Kohler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207–17.
20. Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. Inferring gene ontologies from pairwise similarity data. Bioinformatics. 2014;30(12):i34-42.
21. Hamiltion WLYRLJ. Inductive representation learning on large graphs. 2018.
22. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. KDD. 2016;2016:855–64.
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Duchesnay É. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825–30.
24. Santos A, Colaco AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol. 2022;40(5):692–702.

Publisher's Note