Bioinformatics, 2024, **40(5)**, btae301
https://doi.org/10.1093/bioinformatics/btae301
Advance Access Publication Date: 2 May 2024
Original Paper

Databases and ontologies

Prioritizing genomic variants through neuro-symbolic, knowledge-enhanced learning

Azza Althagafi [1,2,3,*], **Fernando Zhapa-Camacho** [1,2], **Robert Hoehndorf** [1,2,4,*]

[1]Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), 4700 KAUST, Thuwal 23955, Saudi Arabia
[2]Computer Science Program, Computer, Electrical and Mathematical Sciences & Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), 4700 KAUST, Thuwal 23955, Saudi Arabia
[3]Computer Science Department, College of Computers and Information Technology, Taif University, Taif 26571, Saudi Arabia
[4]SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence, King Abdullah University of Science and Technology (KAUST), 4700 KAUST, Thuwal 23955, Saudi Arabia

*Corresponding authors. E-mails: robert.hoehndorf@kaust.edu.sa (R.H.) and azza.althagafi@kaust.edu.sa (A.A.)

Associate Editor: Peter Robinson

Abstract

Motivation: Whole-exome and genome sequencing have become common tools in diagnosing patients with rare diseases. Despite their success, this approach leaves many patients undiagnosed. A common argument is that more disease variants still await discovery, or the novelty of disease phenotypes results from a combination of variants in multiple disease-related genes. Interpreting the phenotypic consequences of genomic variants relies on information about gene functions, gene expression, physiology, and other genomic features. Phenotype-based methods to identify variants involved in genetic diseases combine molecular features with prior knowledge about the phenotypic consequences of altering gene functions. While phenotype-based methods have been successfully applied to prioritizing variants, such methods are based on known gene–disease or gene–phenotype associations as training data and are applicable to genes that have phenotypes associated, thereby limiting their scope. In addition, phenotypes are not assigned uniformly by different clinicians, and phenotype-based methods need to account for this variability.

Results: We developed an Embedding-based Phenotype Variant Predictor (EmbedPVP), a computational method to prioritize variants involved in genetic diseases by combining genomic information and clinical phenotypes. EmbedPVP leverages a large amount of background knowledge from human and model organisms about molecular mechanisms through which abnormal phenotypes may arise. Specifically, EmbedPVP incorporates phenotypes linked to genes, functions of gene products, and the anatomical site of gene expression, and systematically relates them to their phenotypic effects through neuro-symbolic, knowledge-enhanced machine learning. We demonstrate EmbedPVP's efficacy on a large set of synthetic genomes and genomes matched with clinical information.

Availability and implementation: EmbedPVP and all evaluation experiments are freely available at https://github.com/bio-ontology-research-group/EmbedPVP.

1 Introduction

The contribution of genetics to human diseases ranges from almost 100% for monogenic, Mendelian disorders to much smaller percentages for complex diseases, including infectious disease (Hyman 2000). Understanding how variation in an individual's genome relates to disease risk is important, as it allows us to prevent and predict negative health effects in individuals, generate better diagnoses and prognoses for disease, and enable new approaches for treatment and development of new drugs (Bloss *et al.* 2011). Predicting possible health effects from genome sequences is a significant emerging challenge and is important to support genetic counseling and prevent major health problems. Whole-exome and genome sequencing (WGS/WES) has become a common tool in the diagnosis of patients with rare diseases as it has improved diagnostic yields and enables efficient identification of novel gene–disease associations. The interpretation of WGS/WES data linked to individuals is increasingly being used to identify causal variants that may lead to an abnormal phenotype or a disease (Krier *et al.* 2016). Despite its success, these approaches leave many patients undiagnosed, with estimated diagnostic yields of 25%–50% (Clark *et al.* 2018).

While there have been several efforts to predict and prioritize pathogenic genomic variants, in particular, single-nucleotide polymorphisms (SNPs) and small Insertion or Deletion (InDels) (Eilbeck *et al.* 2017), predicting the functional impact of variants discovered through genome sequencing studies remains challenging. This is due to the limited gene–phenotype information available; also, variants may cover multiple coding, noncoding, or intergenic regions and overlap several genes (Shameer *et al.* 2016). Existing methods for predicting the pathogenicity of genomic variants may be based on the impact of variants on protein structure, measures of sequence conservation, or function by relying only on the genomic sequence information (Eilbeck *et al.* 2017). While several methods exist to identify disease-associated variants in patient cohorts, it is more challenging

Received: 24 October 2023; **Revised:** 5 April 2024; **Editorial Decision:** 24 April 2024; **Accepted:** 30 April 2024

to discover disease-associated variants that exist in a single sample or pedigree, in particular in rare Mendelian disorders (Sanchis-Juan *et al.* 2018).

Another group of methods for finding variants causing abnormal phenotypes predicts variant pathogenicity and prioritizes damaging variants using the relation between the phenotypes of a patient and the phenotypes in a database of genotype–phenotype associations (Köhler *et al.* 2014). Phenotype-driven variant prioritization methods aim to link variants to the phenotypes observed in individuals using prior knowledge (Eilbeck *et al.* 2017). Commonly, the link is established using a similarity measure between phenotypes associated with a variant or gene and the phenotypes observed in a patient (Smedley *et al.* 2015). Phenotype-based methods are successful in finding disease-associated variants (Shefchek *et al.* 2020) but suffer from the limited information about variant– or gene–phenotype associations. One way to overcome this limitation is to utilize and link the phenotypes observed in model organisms to human phenotypes (Shefchek *et al.* 2020). However, even when including phenotypes from model organisms, a large number of human protein-coding genes remain without associations, thereby limiting the success of phenotype-based methods to variants or genes that have previously been studied either in human or animal models or relying on guilt-by-association approaches in which information about phenotypes is propagated through associations such as interaction networks (Smedley *et al.* 2014).

Several deep learning and machine learning methods are now available that can predict phenotypes from genotype (Zhou *et al.* 2019, Kulmanov and Hoehndorf 2020) or associate phenotypes with different types of information available for genes, including the functions of gene products and anatomical sites of expression (Smaili *et al.* 2019, Chen *et al.* 2021). These methods use machine learning to relate information through background knowledge contained in formalized knowledge bases, or ontologies, and can accurately identify phenotype-associated genes without prior knowledge about phenotypes, often significantly improving over the use of semantic similarity measures (Kulmanov *et al.* 2020). A limitation of these methods is that they are usually transductive instead of inductive (Kulmanov *et al.* 2020), i.e. the diseases or disorders for which associated genes are predicted should already be available at the time of training the model. As these methods require information about disease-associated phenotypes during training, they cannot generalize to entirely new cases, thereby limiting their application in identifying phenotype-associated genomic variants. Another limitation can be biases introduced by the neural network and the phenotypes annotations (Alghamdi *et al.* 2022) or similarity measure (Kulmanov and Hoehndorf 2017).

We developed Embedding Pathogenicity Variant Predictor (EmbedPVP), a computational method to prioritize variants that are pathogenic and involved in the development of specific phenotypes or genetic diseases. EmbedPVP prioritizes single nucleotide variants or small insertions or deletions involved in genetic diseases. Our method combines genomic information and clinical phenotypes and leverages a large knowledge base derived from human and model organisms for knowledge-enhanced learning. We use different neuro-symbolic embedding-based methods to learn from the background knowledge and combine the information from embedding and pathogenicity prediction to predict the variant that most likely causes the phenotypes observed in the patients. We demonstrate that our method improves over the state-of-the-art in detecting disease-associated variants in multiple benchmark datasets. We have made EmbedPVP freely available as a Python package at https://github.com/bio-ontology-research-group/EmbedPVP.

2 Materials and methods

2.1 Genotype and clinical phenotype datasets

We performed all of our experiments on a set of pathogenic and disease-causing variants for diseases collected from different databases. We inserted the variants we obtained into synthetic genomes with a set of benign, pathogenic, and unknown variants from the 1000 Genome Project Consortium. We use three different datasets of variants to generate synthetic patients and evaluate the performance of EmbedPVP. The first dataset, the Phenotype-Associated Variants in Saudi Arabia (PAVS)-synthetic dataset, covers clinically validated Saudi variants from an in-house database, the PAVS database (http://pavs.phenomebrowser.net) representing 1528 individuals. PAVS is a database that combines a set of clinically validated pathogenic variants with a set of manually curated pathogenic variants observed in the genomes of the Saudi population and their associated phenotypes. All phenotypes are mapped to their Human Phenotype Ontology (HPO) identifiers. The second dataset, Phenopackets (Jacobsen *et al.* 2022), represents 384 individuals described in published case reports with HPO terms and their causal genetic variants. As the final dataset, we selected 1082 newly inserted pathogenic variants (between 4 January 2022 and 31 October 2022) from the ClinVar database (Landrum *et al.* 2020). We further subsetted these datasets to cover other evaluations, such as exonic versus nonexonic variants (Supplementary Materials Section S2.2), variants in overlapping and intergenic regions (Supplementary Section S2.3), variants in genes with no phenotype annotations (Supplementary Section S2.4), newly discovered genes, and diseases observed or not observed during the training.

2.1.1 Clinical phenotypes versus OMIM phenotypes

To distinctly differentiate the ranking of variants using clinical phenotypes from those linked with OMIM identifiers, we performed the experiments on the PAVS and Phenopackets datasets twice using the same genotype data. In the first run, we utilized the phenotypes assigned by clinicians. In the second run, we utilized a set of phenotypes associated with OMIM identifiers. Clinicians reported these OMIM-linked phenotypes for the PAVS dataset, while for the Phenopackets dataset, they were provided as additional annotations for the variants. For the ClinVar benchmark dataset, we conducted experiments only once using the phenotypes associated with the reported disease in the HPO database.

2.2 Resources for ontologies and annotation phenotypes

We use four primary ontologies: HPO, Mammalian Phenotype Ontology (MP), Gene Ontology (GO), and the Uberon cross-species integrated anatomy ontology (UBERON). First, we downloaded the phenotypes associated with human genes from the HPO database on 30 May 2022. We obtained the phenotype annotations for 4318 human genes, including 205 429 associations between genes and

HPO. Second, the phenotypes associated with mouse genes and the orthologous gene mappings from mouse genes to human genes were obtained from the Mouse Genome Informatics (MGI) database (Smith and Eppig 2009), downloaded on 7 June 2022. We obtained phenotype annotations for 13 529 mouse genes, including 228 214 associations between genes and MP classes. We mapped each mouse gene to its human ortholog using the file `HMD_HumanPhenotype.rpt` available at the MGI database, resulting in 9879 human genes for which the mouse ortholog has phenotype associations. Third, we used biological function (GO) annotations from the GO website (Ashburner *et al.* 2000) downloaded on 14 March 2022. We collected 18 495 human gene products (495 719 annotations in total). We mapped the UniProt accessions to Entrez gene identifiers using the mappings provided by the Entrez database (Maglott *et al.* 2010), and we obtained 17 786 Entrez genes for which the gene product has GO annotations. Fourth, for the anatomical location of gene expression, we downloaded the Tissue Expression Profiles (GTEx) dataset (GTEx Consortium 2015) from the Gene Expression Atlas (Papatheodorou *et al.* 2020), which characterizes gene expression across 53 tissues. We mapped the Ensembl protein identifiers to Entrez gene identifiers using the mapping provided by the Entrez database (Maglott *et al.* 2010). We obtained 20 538 Entrez genes, which have expression levels above the 4.0 threshold in one or more tissue. We mapped each tissue to the UBERON ontology, excluding the expression in *EBV-transformed lymphocyte* and *transformed skin fibroblast* since these two tissues are not available in the UBERON ontology.

Finally, because these annotations are available for different numbers of genes, we also used the phenotypes based on the union of all genes and their annotations (i.e. for genes that have annotations from one, two, or all four datasets, HPO, MP, GO, and Uberon). We used the integrated phenotype ontology uPheno (Shefchek *et al.* 2020) as our phenotype ontology to add background knowledge from biomedical ontologies, as it integrates human and model organism phenotypes and allows them to be compared.

To evaluate gene–disease associations, we used the phenotypes available in the HPO database (Köhler *et al.* 2019) to associate diseases from the Online Mendelian Inheritance in Men (OMIM) database (Amberger *et al.* 2011) to their phenotypes. In total, we have 4431 OMIM diseases and 3418 genes in our knowledge base, representing 7405 associations; we used 80% of these associations during supervised training to generate the representations, 15% for the validation, and 5% for testing.

2.3 Generation of synthetic patients and synthetic phenotypes

We created synthetic genotypes in Variant Call Format (VCF) format using the reference genome from the 1000 Genomes Project. The use of synthetic genomes allows us to systematically evaluate the performance of our prioritization method under controlled conditions. By introducing known causative variants into synthetic genomes, we can assess how well our approach identifies these variants among other genomic variants. We simulated a more realistic genome by randomly selecting 100 000 variants such that 90% are in intronic regions and 10% as exonic within regions.

In our experimental design, we set the threshold of MAF to be $<1\%$ which aims to exclude the common variants and prioritize rare and potentially pathogenic variants (Evans *et al.* 2013). We filtered the variants to select variants with MAF $<1\%$ in the 1000 Genomes Project, The Exome Aggregation Consortium (ExAC) (Karczewski *et al.* 2017), and Genome Aggregation Database (gnomAD) (Collins *et al.* 2020) databases for all the population, and as a result, we obtained 98 194 variants to represent our synthetic genome. We then inserted the causative pathogenic variants from our evaluation cohorts (PAVS, Phenopackets, ClinVar Time-based split) into the synthetic genome, which, together with the associated phenotypes, represents the synthetic patients.

For the phenotypes linked to each patient, we evaluated the phenotypes reported for each patient (using PAVS and Phenopackets cohorts). In addition, since we have the OMIM diseases reported for each causative variant, we performed the same experiments using the phenotypes linked with the disease in HPO, which represents more phenotypic variability compared to the reported phenotypes. We used the VCF files together with the HPO phenotypes, either clinical or from OMIM, to run the different models. We then ranked the inserted variants using EmbedPVP models and other prioritization methods for the OMIM diseases set of phenotypes and reported clinical phenotypes.

2.4 Generation of ontology annotation-based embeddings

Formally, we define an ontology using a signature $\Sigma = (\mathbf{C}, \mathbf{R}, \mathbf{I})$, where $\mathbf{C}, \mathbf{R}, \mathbf{I}$ are sets corresponding to concept names, role names, and individual names, respectively. An embedding is a structure-preserving mapping between two mathematical structures. To generate embeddings from ontology entities into vector representations we followed different approaches identified and categorized in (Kulmanov *et al.* 2020) such as (i) graph-embeddings with random walks, (ii) graph embeddings with knowledge graph embeddings methods, and (iii) model-theoretic embeddings. To predict gene–disease associations, we used a scoring function s given by the embedding method. For a gene g and a disease d, $s(g, d)$ will output a value in the range $[0, 1]$ indicating the plausibility of the association to hold true. The following subsections summarize each category of embeddings, all of which we implemented using the mOWL library (Zhapa-Camacho *et al.* 2023); parameters are reported in Supplementary Section S1.

2.4.1 Graph-based + random walk embeddings

A relational graph is a tuple $G = (V, E, L)$ with sets V of vertices, L of edge labels, and $E \subseteq V \times L \times V$ of edges represented as triples (h, r, t), where h, t are nodes, and r is an edge label. Graph-based embedding methods require the generation of a graph out of the ontology axioms. This process is called *graph projection* (Zhapa-Camacho and Hoehndorf 2023). The graph projection methods we chose are the ones found in DL2Vec (Chen *et al.* 2021) and OWL2Vec* (Chen *et al.* 2021). These methods complement each other to enhance the overall robustness of our approach, capturing semantic relationships of different entities by leveraging ontological information.

Traditionally, given a graph, a random walk $w = \{v_0, v_1, v_2, \dots, v_n\}$ of length n is constructed iteratively by choosing an initial node $v_0 \in V$ and obtaining nodes $v_{i+1} = next(v_i)$ given by the function *next*. For example, in DeepWalk, the function $next(v_i)$ generates the element v_{i+1} by choosing randomly from the neighbors of v_i. However, to include edge label information, we used a variation from DeepWalk that takes not only neighboring vertices but also the edge label between them. Therefore, a random walk with n nodes will contain $2n-1$ elements. After generating a graph using the projection function in DL2Vec, we used DeepWalk (Perozzi *et al.* 2014) to create k walks of size $2n-1$ for each node in the graph.

To capture the co-occurrence of ontology entities, we trained a Word2Vec model, where the input is the collection of $k \cdot |V|$ random walks. The Word2Vec model, under the Skip-gram architecture, is optimized to find word representations that are useful to predict surrounding words (Mikolov *et al.* 2013). Thus, given a sequence $v_0, v_1, \ldots, v_n$, the training objective is $\frac{1}{n}\sum_{t=1}^{n}\sum_{-c \le j \le c, j \ne 0} \log(p(w_{t+j}|w_t))$, where p is the softmax function. Given that Word2Vec can capture the co-occurrence of entities, we chose a similarity-based scoring function defined as $s_{gda}(g,d) = \sigma(\tilde{g} \cdot \tilde{d})$, where $\tilde{g}, \tilde{d}$ are the vector representations obtained by training the Word2Vec model, $(\cdot)$ correspond to the dot product, and σ is the sigmoid function.

2.4.2 Knowledge graph embeddings

Graph embeddings using random walks generate embeddings that are useful for computing similarity between nodes, but they neglect the relation labels in testing phase. To incorporate relations information, we generated embeddings by using Knowledge Graph Embedding (KGE) methods (Wang *et al.* 2017), which use a function $s(h,r,t)$ to score triples and can be optimized using an objective function of the form $\mathcal{L} = \sum_{(h,r,t)\in E}\sum_{(h',r,t')\in E'} [s(\tilde{h},\tilde{r},\tilde{t}) - s(\tilde{h'},\tilde{r},\tilde{t'}) + \gamma]_+$, where the set E' is the set of negative triples (i.e. triples not existing in the graph) generated by either corrupting the head or tail of a positive triple in E and γ is a margin between positive and negative scores. The training objective minimizes the score of a positive triple while maximizing the scores of negative ones. KGE methods have been categorized into (i) translational-based, (ii) similarity-based, and (iii) neural-network-based (Wang *et al.* 2021). We used representative methods from each category: (i) TransE (Bordes *et al.* 2013), TransR (Lin *et al.* 2015), TransD (Ji *et al.* 2015), (ii) DistMult (Yang *et al.* 2015), and (iii) ConvE (Dettmers *et al.* 2018). All KGE methods implement a scoring function $s(h,r,t)$ indicating the plausibility of the triple (h, r, t) to exist in the graph. To predict gene–disease associations for a gene g and a disease d, we compute $s_{gda}(g,d) = s(\tilde{g}, \text{is\_associated\_with}, \tilde{d})$. We used the PyKEEN library (Ali *et al.* 2021b) to provide implementations of the chosen KGE methods.

2.4.3 Model-theoretic embedding

Graph-based methods ignore semantic information of ontology axioms. Concept descriptions $\mathbf{C}$ in the Description Logic $\mathcal{EL}$ can be constructed as any of the normal forms $C \sqsubseteq D$, $C \sqcap D \sqsubseteq E$, $C \sqsubseteq \exists R.C$ and $\exists R.C \sqsubseteq D$. An interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ is given by an nonempty domain $\Delta^{\mathcal{I}}$ and an interpretation function mapping every concept $C \in \mathbf{C}$ to a set $C^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and every role $R \in \mathbf{R}$ to a set $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. Moreover, the interpretation function maps complex concept descriptions as follows: $\bot^{\mathcal{I}} = \emptyset$, $\top^{\mathcal{I}} = \Delta^{\mathcal{I}}$, $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$, $(\exists R.C)^{\mathcal{I}} = \{a \in \Delta^{\mathcal{I}} | \exists b \in \Delta^{\mathcal{I}} : (a,b) \in R^{\mathcal{I}} \;\; b \in C^{\mathcal{I}}\}$ An interpretation $\mathcal{I}$ is a *model* if for every axiom $C \sqsubseteq D$ the inclusion $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ holds.

In order to incorporate semantic information, we used two geometric-based embedding methods: ELEmbeddings (Kulmanov *et al.* 2019) and ELBoxEmbeddings (Peng *et al.* 2022). These methods represent ontology concepts as geometric bodies such as n-dimensional balls and n-dimensional boxes, respectively. For every axiom $C \sqsubseteq D$, the training objective minimizes the inclusion loss of the geometric representation of C within the geometric representation of D. Therefore, the scoring method for every axiom is $s(C \sqsubseteq D) = \mathit{inclusion}(C, D)$.

The *inclussion* function is defined for each normal form in ELEmbeddings and ELBoxEmbeddings. The training objective follows a similar approach as in Equation (3), where the positive samples are the axioms in the ontology, and the negative samples are generated by corrupting the concept names on the right-hand side of the axiom. To predict gene–disease associations, we compute the score of the axiom: $s_{gda}(g,d) = s(\tilde{g} \sqsubseteq \exists \text{is\_associated\_with}.\tilde{d})$.

2.4.4 Training procedure

To train our models, we optimized hyperparameters (Supplementary Section S1) for each embedding method. We trained all models using the annotations information for GO, MP, HP, UBERON, and the Union. We used 80% of gene–disease associations during supervised training to generate the representations, 15% for the validation, and 5% for testing. We used Adam (Kingma and Ba 2014) optimizer and adapted the learning rate.

2.4.5 Updating embedding models to handle new phenotypes

While EmbedPVP primarily uses a transductive approach, we also implemented an inductive approach to embedding generation. To achieve this, we first trained each EmbedPVP model (using different embeddings and different ontology) with an initial set of diseases from OMIM until a convergence criterion is reached; we utilized the validation loss as the convergence criterion. During the training process, we continuously monitor the loss calculated on a separate validation dataset. We stop the training when the validation loss no longer decreases or starts to increase, indicating that the model's performance on unseen data is not improving.

The resulting trained models, along with their corresponding embeddings, are saved. We then added information of the phenotypes of each new disease d' separately into the model and trained the model for a small number of iterations to update the embedding representations with the new set of phenotypes. Supplementary Algorithm S1, shows the details of updating the trained model.

2.5 Functional variant features

We annotate variants with a set of genomic features using public databases. We use Annovar (Wang *et al.* 2010), which uses data from multiple external databases. From the annotations provided by Annovar, we use the type of variants and the gene information. While not used as a feature of our prediction model, we also use Annovar to identify the allele frequency of variants using the 1000 Genomes allele frequency (Sudmant *et al.* 2015), ExAC (Karczewski *et al.* 2017), gnomAD (Collins *et al.* 2020). We use this information to filter out common variants before applying our predictions. For the pathogenicity prediction, we rely on the Combined Annotation Dependent Depletion (CADD) (Rentzsch *et al.* 2019) score. CADD is a tool for scoring the deleteriousness of single nucleotide variants and insertion/deletion variants in the human genome.

2.6 Performance evaluation and comparison

We compare EmbedPVP with variant prioritization tools based on genotype information, specifically CADD (Kleinert and Kircher 2021), SIFT (Ng and Henikoff 2003), PolyPhen2 (Adzhubei *et al.* 2013), MetaSVM (Sun and Yu 2019), and DANN (Quang *et al.* 2015), to determine whether the addition of phenotype information can improve over sequence-based methods alone. We also evaluated and compared EmbedPVP to different phenotype-based methods, PhenIX (Zemojtel *et al.* 2014), Exomiser-hiPHIVE (Robinson *et al.* 2014), PHIVE (Robinson *et al.* 2014), and DeepPVP (Boudellioua *et al.* 2019). We assessed their effectiveness in the different benchmark datasets. We evaluated the performance of our models and baseline methods by calculating the recall at different ranks, i.e. finding the rank of the inserted variants and then reporting the top hits, top 10, top 30, and top 50 hits. In addition, for a more comprehensive evaluation and to provide further insights into the interpretation of the results, we incorporated the Receiver Operating Characteristic Area Under the Curve (AUC) and the area under the precision–recall curve (AUPR) metrics.

3 Results

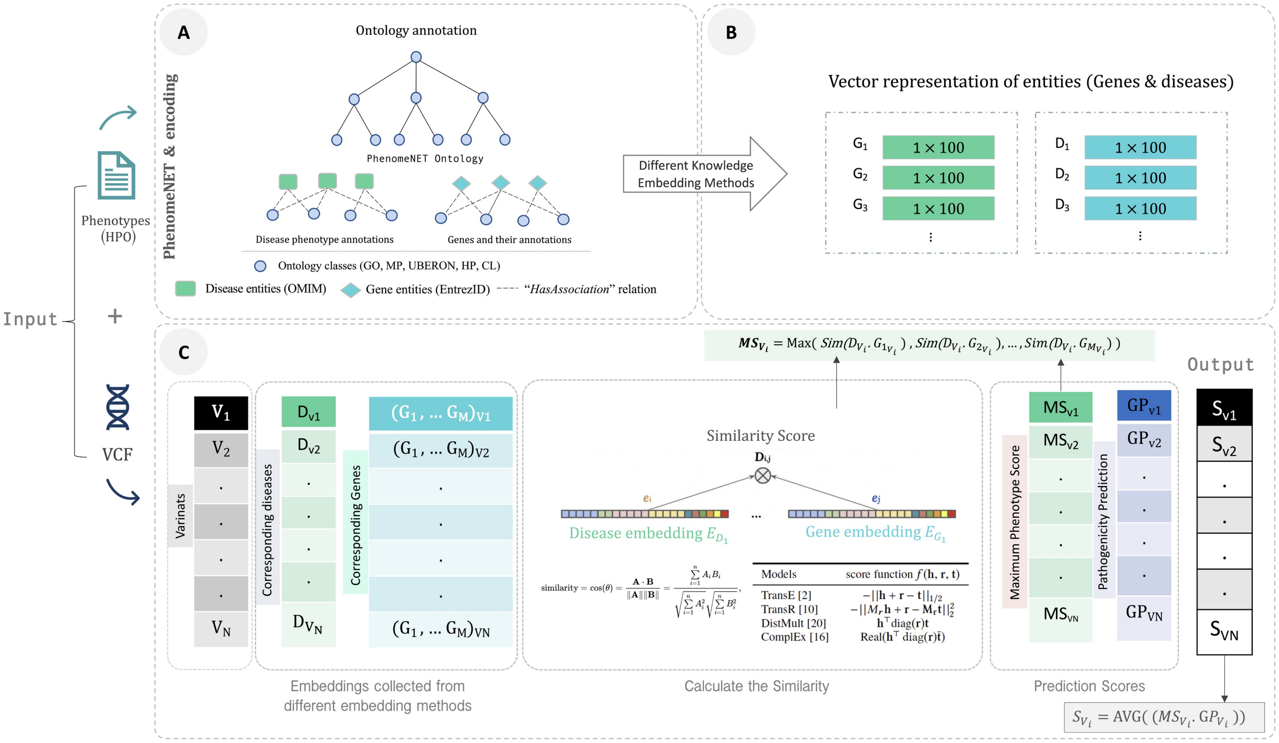
3.1 Overview of the EmbedPVP model

EmbedPVP workflow contains a systematic process that integrates genotypic and phenotypic information, utilizing embeddings and ontologies to prioritize variants based on their potential associations with given phenotypes. Specifically, the workflow (Fig. 1) takes a VCF file as input, which contains a set of SNPs or InDels, and phenotypes encoded using HPO. Using this input, EmbedPVP generates a prioritized list of variants from the input VCF file based on their likelihood of being associated with the input set of phenotypes.

To achieve this goal, EmbedPVP leverages a knowledge base featuring different ontologies and their annotations (Fig. 1A). This knowledge base facilitates the connection between genotypes and phenotypes. Subsequently, EmbedPVP utilized different embedding methods to generate embedding representations for both the input phenotypes and the genes (Fig. 1B). Furthermore, for the given set of variants, EmbedPVP collects genomic features for each variant based on its association with the gene or set of overlapping genes. Specifically, coding variants are linked to the gene in whose coding region they reside, while intergenic variants connect to their nearest genes. In cases where a variant lies in the coding region of multiple genes, it relates to all of them (Fig. 1C). EmbedPVP then calculates the similarity between the input set of phenotypes and the ontology-based embedding for the genes using the scoring function associated with the selected embedding method. The pathogenicity prediction method is a parameter of EmbedPVP; we have used CADD because it is used by other phenotype-based methods and provides genome-wide predictions. Finally, EmbedPVP computes the final prediction score for the variant by using the weighted averages of the similarity score with the pathogenicity prediction.

3.2 EmbedPVP evaluation: clinical phenotypes and OMIM phenotypes

EmbedPVP combines two sources of information to evaluate variants, variant pathogenicity and relevance to observed


A
Ontology annotation
PhenomeNET & encoding
PhenomeNET Ontology
Disease phenotype annotations
Genes and their annotations
Ontology classes (GO, MP, UBERON, HP, CL)
Disease entities (OMIM)
Gene entities (EntrezID)
"HasAssociation" relation
Phenotypes (HPO)
Input
+
VCF
Different Knowledge Embedding Methods
B
Vector representation of entities (Genes & diseases)
G_1 1 × 100
G_2 1 × 100
G_3 1 × 100
D_1 1 × 100
D_2 1 × 100
D_3 1 × 100
C
$MS_{V_i} = \text{Max}(\, Sim(D_{V_i}.G_{1_{V_i}}), Sim(D_{V_i}.G_{2_{V_i}}), \dots, Sim(D_{V_i}.G_{M_{V_i}}))$
Output
Varinats
V_1
V_2
V_N
Corresponding diseases
D_{v1}
D_{v2}
D_{V_N}
Corresponding Genes
$(G_1, \dots G_M)_{V1}$
$(G_1, \dots G_M)_{V2}$
$(G_1, \dots G_M)_{VN}$
Similarity Score
$D_{i,j}$
e_i
e_j
Disease embedding E_{D_1}
Gene embedding E_{G_1}
$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$
Maximum Phenotype Score
MS_{v1}
MS_{v2}
MS_{VN}
Pathogenicity Prediction
GP_{v1}
GP_{v2}
GP_{VN}
S_{v1}
S_{v2}
S_{VN}
Embeddings collected from different embedding methods
Calculate the Similarity
Prediction Scores
$S_{V_i} = \text{AVG}(\,(MS_{V_i}.GP_{V_i}))$

Models	score function $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$
TransE [2]	$-\|\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|\|_{1/2}$
TransR [10]	$-\|\|M_r\mathbf{h}+\mathbf{r}-\mathbf{M}_r\mathbf{t}\|\|_2^2$
DistMult [20]	$\mathbf{h}^\top \text{diag}(\mathbf{r})\mathbf{t}$
ComplEx [16]	$\text{Real}(\mathbf{h}^\top \text{diag}(\mathbf{r})\bar{\mathbf{t}})$

Figure 1. EmbedPVP Model Workflow. (**A**) Generates background knowledge from different ontologies. (**B**) Generates embeddings for diseases (D_i) and genes (G_i) using various embedding methods. (**C**) Calculates phenotype–genotype similarity using the scoring function associated with the selected embedding method, considering the maximum similarity score for multiple genes associated with the phenotype, and then averages this similarity with pathogenicity prediction. V_i represents variant *i*, MS_{vi} is max phenotype similarity for variants, GP_{vi} is genotype prediction (CADD), and S_{vi} is the final weighted score of phenotypes and genotypes for the variants.

phenotype. We include a comparison to pathogenicity prediction methods to demonstrate whether and how much additional information is provided by the phenotype-matching component of EmbedPVP. We conducted evaluations on different benchmark datasets, including synthetic datasets using clinical phenotypes and OMIM phenotypes. For this purpose, we trained EmbedPVP with a unique representation for each sample based on its phenotypes, i.e. all samples (patients with their phenotypes) were already known during embedding generation. In other words, we perform transductive inference. The evaluations aimed to assess the performance of different embedding methods on these datasets. Table 1 provides a comparison of the performance of EmbedPVP against other state-of-the-art methods using the PAVS dataset (refer to Supplementary Table S1 for the results of all other methods). Additionally, Supplementary Fig. S2 shows the average ranks for hits at different ranks.

Based on the results, we observed that EmbedPVP using the TransD model with HP ontology achieved the highest performance among the phenotype-based prediction tools using clinical phenotypes. However, when using OMIM phenotypes, OWL2Vec* demonstrated slightly better performance. DL2Vec and OWL2Vec* performed similarly in both clinical and OMIM phenotypes compared to other phenotype-based models. These findings suggest that TransD captures the complex representations of relationships more effectively. TransD utilizes a translation-based approach to model relationships, which enables it to capture multi-relational relationships between entities. On the other hand, OWL2Vec* and DL2Vec primarily focus on representing hierarchical relationships using the ontology's structure. Although they excel at capturing hierarchical relationships, they may struggle to represent more intricate relationships involving multiple entities or more complex representations, in contrast to the TransD model.

When evaluating the Phenopackets dataset (Supplementary Table S3), we observed that the EmbedPVP (DL2Vec) method performed better in terms of top hits. However, among the phenotype-based methods, Phenix demonstrated better performance for the remaining metrics.

Furthermore, we evaluated our method using ClinVar time-split variants, and the results of the different methods are presented in Supplementary Table S4. In this dataset, EmbedPVP using the TransD method outperforms other methods using the HP model for the top hits, the Union model for the H@10, and the GO model for H@30 and H@50. To further assess the model's performance and remove potential biases due to partial information about gene–disease tuples being present during training, we conducted additional evaluations by splitting the dataset based on novel genes and diseases that were not present during training. We created different subsets, as follows: (A) novel genes and diseases (454 variants), (B) novel genes and known diseases (31 variants), (C) novel diseases and known genes (111 variants), and (D) known genes and diseases (484 variants). The results for these different subsets are shown in Supplementary Table S5 for A and B, and Table S6 for C and D. We also noticed the EmbedPVP models performed better compared to other phenotype- and sequence-based methods.

To assess the impact of ontology axioms compared to annotations, we conducted an additional ablation study. In this study, we included only the phenotype annotations without the axioms from the uPheno ontologies (i.e. we removed all axioms including subclass axioms). The results, shown in Supplementary Table S10, demonstrate a drop in performance for all the metrics when axioms are removed, indicating that EmbedPVP can effectively utilize ontology structural information in addition to the phenotype annotations.

3.3 Improved generalization to new phenotypes with inductive inferences

We evaluate the performance of inductive inference using PAVS with clinical phenotypes, with selected models based on the best-performing transductive approach, including the OWL2Vec*, DL2Vec, TransE, and TransD embedding models. Table 2 presents the results comparing the inductive and transductive approaches. The results show a drop in performance, with slight differences in terms of ROCAUC and AUPR (∼2%), as a consequence, the inductive model does not perform better than other methods. This result demonstrates that the additional time required for retraining EmbedPVP in the presence of new individuals to analyze is necessary for its performance.

4 Discussion

We developed a method for prioritizing candidate causative variants when given a set of disease-associated phenotypes and genotypes. Our approach utilizes various features characterized through ontologies and employs neuro-symbolic embedding methods to exploit the information in ontologies and their annotations. As a result, EmbedPVP can improve phenotype-based prediction of disease-causing variants. Moreover, we also explored the impact of clinical phenotype descriptions and could demonstrate that the embeddings we utilize are robust to noisy phenotype descriptions.

Knowledge-enhanced learning involves the utilization of background knowledge to enhance predictive models. Knowledge-enhanced learning is especially useful when too little training data are available to apply supervised learning directly, and where structured knowledge is available that can constrain search (Feigenbaum *et al.* 1977). The large number of biomedical ontologies and the knowledge they contain has been used deductively to generate additional knowledge that could then be used to improve machine learning tasks (Hoehndorf *et al.* 2011, Köhler *et al.* 2013, Matentzoglu *et al.* 2019); in our work, we use the background knowledge in ontologies not deductively but rather as part of a neuro-symbolic method (Hitzler and Sarker 2022) where a form of inference happens in a latent space (Hitzler *et al.* 2023).

In our application, we rely on axioms from the GO (Ashburner *et al.* 2000), phenotype ontologies, and anatomy ontologies (Smith *et al.* 2007). We use these ontologies to integrate information about pathways, interactions between genes, anatomical site of gene expression, and protein functions, and ontologies already link all this information to phenotypes using formal axioms. In particular, phenotype ontologies have long been constructed using the entity–quality (EQ) method where phenotypes are decomposed into an affected entity (an anatomical site, or a biological function) and a quality (using the PATO ontology of qualities) (Mungall *et al.* 2010, Gkoutos *et al.* 2018). Using these axioms now proves useful not only for data integration (which was one of the original intentions in developing these axioms)

Table 1. EmbedPVP variant prediction results across several ontologies with different neuro-symbolic knowledge embedding methods.

		Using the clinical phenotypes						Using OMIM phenotypes					
		H@1	H@10	H@30	H@50	ROCAUC	AUPR	H@1	H@10	H@30	H@50	ROCAUC	AUPR
Genotype-based prediction tools	CADD	116 (0.0759)	266 (0.1741)	467 (0.3056)	591 (0.3868)	0.9778	0.0494	116 (0.0759)	266 (0.1741)	467 (0.3056)	591 (0.3868)	0.9778	0.0494
	MCAP	4 (0.0026)	261 (0.1708)	442 (0.2893)	511 (0.3344)	0.6389	0.0076	4 (0.0026)	261 (0.1708)	442 (0.2893)	511 (0.3344)	0.6389	0.0076
	SIFT	201 (0.1315)	201 (0.1315)	201 (0.1315)	201 (0.1315)	0.6436	0.0736	201 (0.1315)	201 (0.1315)	201 (0.1315)	201 (0.1315)	0.6436	0.0736
	PolyPhen2	127 (0.0831)	127 (0.0831)	127 (0.0831)	226 (0.1479)	0.6465	0.0481	127 (0.0831)	127 (0.0831)	127 (0.0831)	226 (0.1479)	0.6465	0.0481
	DANN	21 (0.0137)	263 (0.1721)	263 (0.1721)	263 (0.1721)	0.8422	0.0115	21 (0.0137)	263 (0.1721)	263 (0.1721)	263 (0.1721)	0.8422	0.0115
	MetaSVM	20 (0.0131)	111 (0.0726)	318 (0.2081)	406 (0.2657)	0.6510	0.0108	20 (0.0131)	111 (0.0726)	318 (0.2081)	406 (0.2657)	0.6510	0.0108
Phenotype-based prediction tools	PHIVE	181 (0.1185)	325 (0.2127)	364 (0.2382)	380 (0.2487)	0.8047	0.0709	346 (0.2264)	496 (0.3246)	518 (0.3390)	523 (0.3423)	0.8151	0.1477
	DeepPVP	221 (0.1446)	661 (0.4326)	762 (0.4987)	795 (0.5203)	0.7662	0.1389	449 (0.2938)	858 (0.5615)	905 (0.5923)	924 (0.6047)	0.8041	0.2853
	Phenix	472 (0.3089)	628 (0.4110)	746 (0.4882)	788 (0.5157)	0.8148	0.2154	**1104 (0.7225)**	1130 (0.7395)	1153 (0.7546)	1159 (0.7585)	0.8206	0.6275
	hiPHIVE	431 (0.2821)	653 (0.4274)	768 (0.5026)	809 (0.5295)	0.8098	0.1982	868 (0.5681)	1025 (0.6708)	1149 (0.7520)	1184 (0.7749)	0.8151	0.4693
EmbedPVP (TransD)	GO	307 (0.2009)	563 (0.3685)	726 (0.4751)	829 (0.5425)	0.9524	0.1386	670 (0.4385)	894 (0.5851)	1006 (0.6584)	1042 (0.6819)	0.9795	0.3464
	HP	**482 (0.3154)**	**846 (0.5537)**	**1007 (0.659)**	**1056 (0.6911)**	**0.9895**	**0.2507**	996 (0.6518)	1230 (0.805)	1352 (0.8848)	**1391 (0.9103)**	**0.9960**	0.5865
	MP	396 (0.2592)	675 (0.4418)	868 (0.5681)	947 (0.6198)	0.9587	0.1869	779 (0.5098)	922 (0.6034)	1031 (0.6747)	1072 (0.7016)	0.9822	0.4120
	UBERON	287 (0.1878)	509 (0.3331)	674 (0.4411)	800 (0.5236)	0.9493	0.1278	699 (0.4575)	892 (0.5838)	995 (0.6512)	1023 (0.6695)	0.9775	0.3594
	Union	409 (0.2677)	639 (0.4182)	833 (0.5452)	928 (0.6073)	0.9581	0.1934	899 (0.5884)	1086 (0.7107)	1158 (0.7579)	1245 (0.8148)	0.9933	0.5087
EmbedPVP (DL2Vec)	GO	152 (0.0995)	382 (0.2500)	554 (0.3626)	614 (0.4018)	0.9282	0.0659	491 (0.3213)	804 (0.5262)	944 (0.6178)	1010 (0.6610)	0.9787	0.2485
	HP	362 (0.2369)	666 (0.4359)	787 (0.5151)	826 (0.5406)	0.9867	0.1758	1011 (0.6616)	1300 (0.8508)	1366 (0.8940)	1384 (0.9058)	0.9942	0.6168
	MP	255 (0.1669)	491 (0.3213)	639 (0.4182)	701 (0.4588)	0.9501	0.1128	639 (0.4182)	914 (0.5982)	1043 (0.6826)	1106 (0.7238)	0.9804	0.3386
	UBERON	174 (0.1139)	390 (0.2552)	498 (0.3259)	556 (0.3639)	0.8928	0.0751	539 (0.3527)	801 (0.5242)	904 (0.5916)	940 (0.6152)	0.9271	0.2713
	Union	358 (0.2343)	636 (0.4162)	771 (0.5046)	824 (0.5393)	0.9605	0.1673	950 (0.6217)	1216 (0.7958)	1310 (0.8573)	1353 (0.8855)	0.9936	0.5625
EmbedPVP (OWL2Vec*)	GO	188 (0.1230)	385 (0.2520)	525 (0.3436)	592 (0.3874)	0.9190	0.0797	557 (0.3645)	876 (0.5733)	1011 (0.6616)	1059 (0.6931)	0.9780	0.2935
	HP	409 (0.2677)	685 (0.4483)	783 (0.5124)	842 (0.5510)	0.9874	0.1987	1026 (0.6715)	**1313 (0.8593)**	**1373 (0.8986)**	**1391 (0.9103)**	0.9940	**0.6304**
	MP	222 (0.1453)	470 (0.3076)	618 (0.4045)	677 (0.4431)	0.9508	0.0992	665 (0.4352)	965 (0.6315)	1068 (0.6990)	1116 (0.7304)	0.9785	0.3582
	UBERON	158 (0.1034)	379 (0.2480)	474 (0.3102)	525 (0.3436)	0.8866	0.0673	577 (0.3776)	800 (0.5236)	888 (0.5812)	937 (0.6132)	0.9291	0.2937
	Union	375 (0.2454)	650 (0.4254)	787 (0.5151)	835 (0.5465)	0.9563	0.1774	959 (0.6276)	1253 (0.8200)	1325 (0.8671)	1368 (0.8953)	0.9939	0.5775

Bold values indicate the highest scores achieved among different models.

Table 2. EmbedPVP variant prediction results for inductive versus transductive approach for the HP model using the clinical phenotypes and selected models based on the best-performing models using the transductive approach.

		H@1	H@10	H@30	H@50	ROCAUC	AUPR
EmbedPVP (TransE)	**Inductive**	204 (0.1335)	388 (0.2539)	535 (0.3501)	678 (0.4437)	0.8918	0.0840
	Transductive	218 (0.1427)	415 (0.2716)	589 (0.3855)	710 (0.4647)	0.9144	0.0908
EmbedPVP (TransD)	**Inductive**	168 (0.1099)	491 (0.3213)	706 (0.4620)	847 (0.5543)	0.9631	0.0790
	Transductive	482 (0.3154)	846 (0.5537)	1007 (0.6590)	1056 (0.6911)	0.9895	0.2507
EmbedPVP (DL2Vec)	**Inductive**	202 (0.1322)	409 (0.2677)	558 (0.3652)	613 (0.4012)	0.9776	0.0868
	Transductive	362 (0.2369)	666 (0.4359)	787 (0.5151)	826 (0.5406)	0.9867	0.1758
EmbedPVP (OWL2Vec*)	**Inductive**	179 (0.1171)	386 (0.2526)	543 (0.3554)	612 (0.4005)	0.9776	0.0764
	Transductive	409 (0.2677)	685 (0.4483)	783 (0.5124)	842 (0.5510)	0.9874	0.1987

but also enables knowledge-enhanced learning in these domains.

EmbedPVP is not the first approach that uses ontology semantics in detecting genotype–phenotype relations; in particular semantic similarity measures have been used for a long time to predict gene–disease associations (Köhler *et al.* 2009), and semantic similarity measures have also been incorporated in variant prioritization tools such as Exomiser (Robinson *et al.* 2014). While semantic similarity measures are able to compare sets of classes from a single ontology, our neuro-symbolic approach is able to "learn" a similarity measure within a latent space, and determine the similarity between classes that are related through complex and heterogeneous axioms. This property allows us not only to improve predictive performance over approaches that rely on semantic similarity (such as the Exomiser tool), but, maybe more importantly, extends the scope of phenotype-similarity methods for finding candidate disease genes to genes for which no phenotypes are known. Previously, a major advance has been the use of model organism phenotypes to expand the scope of methods that find disease-associated genes or variants through comparison to patient phenotypes (Hoehndorf *et al.* 2011, Chen *et al.* 2012); the combination of mouse and zebrafish phenotypes spans a large part of the human genome, but still there are gaps where no phenotypes are associated with a gene. EmbedPVP can apply phenotype similarity for any gene for which a site of expression or gene function is known. We also perform an ablation study where we remove ontology axioms, and find that all methods we tested can effectively utilize the ontology axioms.

We investigated the influence of different ontology embeddings methods on variant prioritization performance, comparing different approaches to ontology embedding. Similarly to KGE methods (Ali *et al.* 2021a), we find the different approaches to be quite variable and sensitive to parameter choices. Nevertheless, based on our results, we can identify some general trends from which we can derive recommendations. When comparing different approaches to ontology embeddings, we find that approaches that first project ontologies onto graphs and then use KGE work better in our case than model-based approaches like ELEmbedding. Among the KGE approaches, methods that explicitly optimize for link prediction (as a training objective) perform better than approaches that only capture similarity (usually based on random walks); and among the link prediction approaches, we find that TransD generally performs better than other methods we evaluated.

One main limitation of EmbedPVP is that it uses a transductive method which requires retraining parts of the model when a new case or set of cases is analyzed. This is mainly a limitation of time as retraining is part of applying EmbedPVP to a new case; however, in particular, when analyzing larger number of cases, it may still be reasonable to retrain and then predict. In the future, however, we intend to focus our efforts on designing novel strategies for inductive inference.

5 Conclusion

We developed EmbedPVP, a method for prioritizing candidate causative variants given a set of abnormal phenotypes. Our method applies machine learning to background knowledge integrated through ontologies and not only improves the phenotype-based prediction of disease-associated variants, but also extends phenotype-based variant prioritization to variants in genes for which no phenotypes are available; instead, EmbedPVP can use knowledge about gene functions, sites of expression, interactions, and also phenotypes in humans or model organisms to prioritize variants. We implemented and evaluated different embedding-based methods for learning from biomedical knowledge bases, applying graph-based as well as model-based methods. EmbedPVP is an end-to-end model and is applicable not only to single nucleotide variants in coding regions, but also to noncoding variants and small insertions and deletions. EmbedPVP has been designed to prioritize variants even when phenotype information is missing or noisy, and EmbedPVP could improve the prediction of causative variants even in the presence of noise. EmbedPVP improves over state-of-the-art methods for phenotype-based variant prioritization, particularly in improving the recall in finding phenotype-associated variants across various benchmark datasets. EmbedPVP is freely available at https://github.com/bio-ontology-research-group/EmbedPVP.

Acknowledgements

We acknowledge the use of computational resources from the KAUST Supercomputing Core Laboratory.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

The authors declare that no conflicts of interest exist.

Funding

This work has been supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award Numbers REI/1/5659-01-01, REI/1/5235-01-01, URF/1/4355-01-01, URF/1/4675-01-01, and URF/1/5041-01-01, FCC/1/1976-46-01. This work was supported by the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

Data availability

All data underlying this article are freely available at https://github.com/bio-ontology-research-group/EmbedPVP.

References

Adzhubei I, Jordan DM, Sunyaev SR *et al.* Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013;Chapter 7:Unit 7.20.

Alghamdi SM, Schofield PN, Hoehndorf R *et al.* Contribution of model organism phenotypes to the computational identification of human disease genes. *Dis Model Mech* 2022;**15**:dmm049441.

Ali M, Berrendorf M, Hoyt CT *et al.* Bringing light into the dark: a large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Trans Pattern Anal Mach Intell* 2021a; **44**:8825–45.

Ali M, Berrendorf M, Hoyt CT *et al.* PyKEEN 1.0: a Python library for training and evaluating knowledge graph embeddings. *J Mach Learn Res* 2021b;**22**:3723–8.

Amberger J, Bocchini C, Hamosh A *et al.* A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum Mutat* 2011;**32**:564–7.

Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

Bishan Y, Wen-tau Y, Xiaodong H, *et al.* Embedding entities and relations for learning and inference in knowledge bases. In: Proceedings of the International Conference on Learning Representations (ICLR) 2015, *San Diego, CA, USA, May 7–9, 2015*, *Conference Track Proceedings*. 2015.

Bloss CS, Jeste DV, Schork NJ *et al.* Genomics for disease treatment and prevention. *Psychiatr Clin North Am* 2011;**34**:147–66.

Bordes, A., Usunier, N., Garcia-Duran, A., *et al.* Translating embeddings for modeling multi-relational data. *Adv Neural Inform Process Systems* 2013;**26**:2787–95.

Boudellioua I, Kulmanov M, Schofield PN *et al.* DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinformatics* 2019;**20**:65–8.

Chen C-K, Mungall CJ, Gkoutos GV *et al.* MouseFinder: candidate disease genes from mouse phenotype data. *Hum Mutat* 2012; **33**:858–66.

Chen J, Azza A. and Robert H. Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics* 2021; 37:853–860.

Chen J, Hu P, Jimenez-Ruiz E *et al.* OWL2Vec: embedding of owl ontologies. *Mach Learn* 2021;**110**:1813–45.

Clark MM, Stark Z, Farnaes L *et al.* 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.

Collins RL, Brand H, Karczewski KJ *et al.*; Genome Aggregation Database Consortium. A structural variation reference for medical and population genetics. *Nature* 2020;**581**:444–51.

Dettmers T, Minervini P, Stenetorp P *et al.* Convolutional 2D knowledge graph embeddings. In: *AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. Washington, DC: Association for the Advancement of Artificial Intelligence, 2018, 1811–1818.

Eilbeck K, Quinlan A, Yandell M *et al.* Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet* 2017;**18**:599–612.

Evans DM, Brion MJA, Paternoster L *et al.*; TAG Consortium. Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genet* 2013;**9**:e1003919.

Felgenbaum, E.A. The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering. In *IJCAI'77: Proceedings of the 5th international joint conference on Artificial Intelligence* Volume 2. Burlington, Massachusetts: Morgan Kaufmann Publishers, **1977**, 1014–1029.

Gkoutos GV, Schofield PN, Hoehndorf R *et al.* The anatomy of phenotype ontologies: principles, properties and applications. *Brief Bioinform* 2018;**19**:1008–21.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**:648–60.

Guoliang J., Shizhu H. and Liheng X., *et al.* Knowledge graph embedding via dynamic mapping matrix. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Kerrville, TX, USA: Association for Computational Linguistics, 2015, 687–96.

Hitzler, P., Sarker, M.K. (Eds.) Neuro-Symbolic Artificial Intelligence: The State of the Art. In: *Frontiers in Artificial Intelligence and Applications*, Vol 342. Amsterdam: IOS Press, 2022.

Hitzler, P., Sarker, M.K. and Eberhart, A. (eds), *Compendium of Neurosymbolic Artificial Intelligence, Frontiers in Artificial Intelligence and Applications / Faia* Vol. 369. Amsterdam: IOS Press, 2023.

Hoehndorf R, Schofield PN, Gkoutos GV *et al.* PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res* 2011;**39**:e119.

Hyman SE. The genetics of mental illness: implications for practice. *Bull World Health Organ* 2000;**78**:455–63.

Irene P, Pablo M, Jonathan M, *et al.* Expression Atlas update: from tissues to single cells. *Nucleic AcidsResearch* 2020;**48**:D77–83.

Jacobsen JOB, Baudis M, Baynam GS *et al.*; GAGH Phenopacket Modeling Consortium. The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat Biotechnol* 2022; **40**:817–20.

Karczewski KJ, Weisburd B, Thomas B *et al.*; The Exome Aggregation Consortium. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* 2017;**45**:D840–5.

Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint* arXiv:1412.6980, 2014.

Kleinert P., and Martin K. A framework to score the effects of structural variants in health and disease. *Genome Research* 32, no. 4 2022: 766-777. doi: 10.1101/gr.275995.121

Köhler S, Carmody L, Vasilevsky N *et al.* Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res* 2019;**47**:D1018–27.

Köhler S, Doelken SC, Ruef BJ *et al.* Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Res* 2013;**2**:30.

Köhler S, Schoeneberg U, Czeschik JC *et al.* Clinical interpretation of CNVs with cross-species phenotype data. *J Med Genet* 2014; **51**:766–72.

Köhler S, Schulz MH, Krawitz P *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* 2009;**85**:457–64.

Krier JB, Kalia SS, Green RC *et al.* Genomic sequencing in clinical practice: applications, challenges, and opportunities. *Dialogues Clin Neurosci* 2016;**18**:299–312.

Kulmanov M, Hoehndorf R. Evaluating the effect of annotation size on measures of semantic similarity. *J Biomed Semantics* 2017;**8**:7–10.

Kulmanov M, Hoehndorf R. DeepPheno: predicting single gene loss-of-function phenotypes using an ontology-aware hierarchical classifier. *PLoS Comput Biol* 2020;**16**:e1008453.

Kulmanov, M, Wang, L.-W., Yuan, Y, and Robert, H. EL Embeddings: geometric construction of models for the description logic EL + +. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence,* Red Hook, NY, USA: AAAI Press, 2019, 6103–9.

Kulmanov, Maxat, Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics* 2020;**22**:bbaa199.

Landrum MJ, Chitipiralla S, Brown GR *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* 2020;**48**:D835–44.

Maglott D, Ostell J, Pruitt KD *et al.* Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2010;**39**:D52–7.

Matentzoglu, N., Osumi-Sutherland, D., Balhoff, J. P., Bello, S., Bradford, Y., Cardmody, L., Grove, C., Harris, M. A., Harris, N. and Köhler, S. uPheno 2: framework for standardised representation of phenotypes across species. *F1000Res* 2019;**8**:403.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (Lake Tahoe, Nevada) (NIPS 2013)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

Mungall CJ, Gkoutos GV, Smith CL *et al.* Integrating phenotype ontologies across multiple species. *Genome Biol* 2010;**11**:R2–16.

Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812–4.

Perozzi, B, Rami, A.-R., and Steven, S. Deepwalk: Online learning of social representations. In: KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery, 2014; 701–710.

Quang D, Chen Y, Xie X *et al.* DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015; **31**:761–3.

Rentzsch P, Witten D, Cooper GM *et al.* CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–94.

Robinson PN, Köhler S, Oellrich A *et al.*; Sanger Mouse Genetics Project. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* 2014;**24**:340–8.

Sanchis-Juan A, Stephens J, French CE *et al.* Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med* 2018;**10**:95.

Shameer K, Tripathi LP, Kalari KR *et al.* Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Brief Bioinform* 2016;**17**:841–62.

Shefchek KA, Harris NL, Gargano M *et al.* The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 2020;**48**:D704–15.

Smaili FZ, Gao X, Hoehndorf R *et al.* OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics* 2019;**35**:2133–40.

Smedley D, Jacobsen JOB, Jäger M *et al.* Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc* 2015; **10**:2004–15.

Smedley D, Köhler S, Czeschik JC *et al.* Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics* 2014;**30**:3215–22.

Smith B, Ashburner M, Rosse C *et al.*; OBI Consortium. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.

Smith CL, Eppig JT. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med* 2009;**1**:390–9.

Sudmant PH, Rausch T, Gardner EJ *et al.*; 1000 Genomes Project Consortium. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**:75–81.

Sun H, Yu G. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci Rep* 2019;**9**:1667–711.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**,56–65.

Wang K, Li M, Hakonarson H *et al.* ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.

Wang M, Qiu L, Wang X *et al.* A survey on knowledge graph embeddings for link prediction. *Symmetry* 2021;**13**:485.

Wang Q, Mao Z, Wang B *et al.* Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans Knowl Data Eng* 2017;**29**:2724–43.

Xi P, Zhenwei T, Maxat K, *et al.* Description logic EL + + embeddings with intersectional closure. arXiv preprint, CoRR abs/2202.14018 (2022) 2202.14018.

Yankai L, Zhiyuan L, Maosong S, *et al.* Learning entity and relation embeddings for knowledge graph completion. In: AAAI'15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Red Hook, NY, USA: AAAI Press, 2015. 2181–2187

Zemojtel T, Köhler S, Mackenroth L *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* 2014;**6**:252ra123.

Zhapa-Camacho F, Hoehndorf R. From axioms over graphs to vectors, and back again: evaluating the properties of graph-based ontology embeddings. In: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, July 3–5, 2023, Volume 3432 of CEUR Workshop Proceedings*, p. 85–102. Aachen, Germany: CEUR-WS.org, 2023.

Zhapa-Camacho F, Kulmanov M, Hoehndorf R *et al.* mOWL: Python library for machine learning with biomedical ontologies. *Bioinformatics* 2023;**39**:btac811.

Zhou N, Jiang Y, Bergquist TR *et al.* The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol* 2019;**20**:244–23.

Bioinformatics, 2024, 40, 1–10
https://doi.org/10.1093/bioinformatics/btae301
Original Paper