Journal of Biomedical Informatics 96 (2019) 103246

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology

Feichen Shen[a,*], Suyuan Peng[a,b], Yadan Fan[a,c], Andrew Wen[a], Sijia Liu[a], Yanshan Wang[a], Liwei Wang[a], Hongfang Liu[a,*]

[a] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[b] The Second Clinical College Guangzhou University of Chinese Medicine, China
[c] Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA

ARTICLE INFO

Keywords:
Enriched node embeddings
Human Phenotype Ontology
Heterogeneous knowledge resources
Phenotypic relevance detection
Deep phenotyping

ABSTRACT

Background: In precision medicine, deep phenotyping is defined as the precise and comprehensive analysis of phenotypic abnormalities, aiming to acquire a better understanding of the natural history of a disease and its genotype-phenotype associations. Detecting phenotypic relevance is an important task when translating precision medicine into clinical practice, especially for patient stratification tasks based on deep phenotyping. In our previous work, we developed node embeddings for the Human Phenotype Ontology (HPO) to assist in phenotypic relevance measurement incorporating distributed semantic representations. However, the derived HPO embeddings hold only distributed representations for IS-A relationships among nodes, hampering the ability to fully explore the graph.

Methods: In this study, we developed a framework, HPO2Vec+, to enrich the produced HPO embeddings with heterogeneous knowledge resources (i.e., DECIPHER, OMIM, and Orphanet) for detecting phenotypic relevance. Specifically, we parsed disease-phenotype associations contained in these three resources to enrich non-inheritance relationships among phenotypic nodes in the HPO. To generate node embeddings for the HPO, node2vec was applied to perform node sampling on the enriched HPO graphs based on random walk followed by feature learning over the sampled nodes to generate enriched node embeddings. Four HPO embeddings were generated based on different graph structures, which we hereafter label as HPOEmb-Original, HPOEmb-DECIPHER, HPOEmb-OMIM, and HPOEmb-Orphanet. We evaluated the derived embeddings quantitatively through an HPO link prediction task with four edge embeddings operations and six machine learning algorithms. The resulting best embeddings were then evaluated for patient stratification of 10 rare diseases using electronic health records (EHR) collected at Mayo Clinic. We assessed our framework qualitatively by visualizing phenotypic clusters and conducting a use case study on *primary hyperoxaluria* (PH), a rare disease, on the task of inferring relevant phenotypes given 22 annotated PH related phenotypes.

Results: The quantitative link prediction task shows that HPOEmb-Orphanet achieved an optimal AUROC of 0.92 and an average precision of 0.94. In addition, HPOEmb-Orphanet achieved an optimal F1 score of 0.86. The quantitative patient similarity measurement task indicates that HPOEmb-Orphanet achieved the highest average detection rate for similar patients over 10 rare diseases and performed better than other similarity measures implemented by an existing tool, HPOSim, especially for pairwise patients with fewer shared common phenotypes. The qualitative evaluation shows that the enriched HPO embeddings are generally able to detect relationships among nodes with fine granularity and HPOEmb-Orphanet is particularly good at associating phenotypes across different disease systems. For the use case of detecting relevant phenotypic characterizations for given PH related phenotypes, HPOEmb-Orphanet outperformed the other three HPO embeddings by achieving the highest average P@5 of 0.81 and the highest P@10 of 0.79. Compared to seven conventional similarity measurements provided by HPOSim, HPOEmb-Orphanet is able to detect more relevant phenotypic pairs, especially for pairs not in inheritance relationships.

Conclusion: We drew the following conclusions based on the evaluation results. First, with additional non-inheritance edges, enriched HPO embeddings can detect more associations between fine granularity phenotypic nodes regardless of their topological structures in the HPO graph. Second, HPOEmb-Orphanet not only can achieve the optimal performance through link prediction and patient stratification based on phenotypic

* Corresponding authors at: Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55901, USA.
E-mail addresses: shen.feichen@mayo.edu (F. Shen), liu.hongfang@mayo.edu (H. Liu).

https://doi.org/10.1016/j.jbi.2019.103246
Received 1 March 2019; Received in revised form 25 June 2019; Accepted 26 June 2019
Available online 27 June 2019

similarity, but is also able to detect relevant phenotypes closer to domain expert's judgments than other embeddings and conventional similarity measurements. Third, incorporating heterogeneous knowledge resources do not necessarily result in better performance for detecting relevant phenotypes. From a clinical perspective, in our use case study, clinical-oriented knowledge resources (e.g., Orphanet) can achieve better performance in detecting relevant phenotypic characterizations compared to biomedical-oriented knowledge resources (e.g., DECIPHER and OMIM).

1. Introduction

The Human Phenotype Ontology (HPO) is commonly used as a resource to provide a controlled vocabulary of phenotypic characterizations related to human diseases [1]. Characterizations maintained by the HPO are collected from heterogeneous knowledge resources, including biomedical literature, the database of chromosomal imbalance and phenotype in humans using ensemble resources (DECIPHER) [2], the Online Mendelian Inheritance in Man (OMIM) [3], and the Orphanet [4]. Deep phenotyping is defined as the precise and comprehensive analysis of phenotypic abnormalities, aiming to obtain a better understanding of the natural history of a disease and its genotype-phenotype associations [5,6]. Therefore, the HPO is able to play an important role in deep phenotyping by translating precision medicine into clinical practice, especially for rare disease differential diagnostic support. Patient stratification is an essential step involved in differential diagnosis, aiming to group patients with similar clinical phenotypic characterizations into the same subgroups so as to accelerate the diagnostic process. Hence, methodology to detect the relevance between clinical phenotypes has become an important piece in deep phenotyping research. A majority of existing studies (see related work in Section 2) identify phenotypes as relevant solely based on topological and inheritance relationships of any two phenotypic nodes in the HPO graph. Those studies do not take into account feature learning of different nodes for detecting associations that cannot be inferred directly through the graph structure.

Meanwhile, inspired by the success of word embeddings in building distributed semantic representations for each word given a corpus, node embeddings provide a solution to map nodes to distributional representations and translate nodes' relationships from graph space to embedding space. Node2vec is one of the commonly adopted models used to build node embeddings [7]. It uses a biased random walk algorithm [8] to perform a flexible neighborhood sampling strategy and feeds the sampling data as input to a word2vec model [9]. In previous work, we constructed node embeddings for the HPO using the node2vec model [10]. The resulting HPO embeddings can quantify the relevance between any two phenotypic characterizations, which is considered to be an important factor for patient stratification on rare disease differential diagnosis. However, the lack of non-inheritance relationships in the HPO hampers the ability to fully explore the graph. For example, according to the HPO, *chronic kidney disease* is a subclass phenotype of *renal insufficiency*, so it is not difficult to detect their relevance by just checking for an inheritance relationship. On the other hand, *Chronic kidney disease* is also related to *synovitis* by contributing to the same rare disease *primary hyperoxaluria* but does not share any direct inheritance relationship with *renal insufficiency* in the HPO. Therefore, it would be difficult to detect the link between those two relevant phenotypes through the HPO.

In this study, we sought to improve our prior developed HPO embeddings and developed HPO2Vec+, a framework to enrich HPO embeddings with information gained from heterogeneous knowledge resources. To the best of our knowledge, it is the first study to enhance HPO embeddings by incorporating knowledge insights from heterogeneous knowledge resources. Specifically, three knowledge resources, namely DECIPHER, OMIM and Orphanet, were used in this study. Based on the annotations provided by these three knowledge resources for HPO, we first designed an algorithm to extract disease-phenotype associations and used these associations to enhance the connectivity of the HPO graph. We then generated different node embeddings using HPO2Vec+. We conducted the evaluation quantitatively and qualitatively. For the quantitative evaluation, we first generated different HPO embeddings using a downstream application on graph link prediction and measured the performance with different machine learning algorithms. The resulting best embeddings were then evaluated for patient stratification of 10 rare diseases using electronic health records (EHR) collected at Mayo Clinic with five [0, 1] bounded conventional HPO based semantic similarity measurements implemented by an existing tool, HPOSim. For the qualitative evaluation, we visualized selected clusters generated by different HPO embeddings through a two-dimensional visualization plot. We then conducted a use case study on *primary hyperoxaluria* (PH), a rare disease. Given some annotated PH related phenotypes, we analyzed relevant phenotypes inferred by different HPO embeddings and conventional graph based semantic similarity measurements.

2. Related work

In the general domain, several metrics are commonly used to measure the similarity amongst ontology annotations. For example, information content (IC) [11] assigns the probability of information gained for each node based on graph structure. The Resnik [12] metric quantifies the similarity between any two nodes as the IC of their most informative common ancestor (MICA) using IS-A relationships. The Jiang-Conrath [13] and Lin [14] metrics measure IC similarity of two terms without considering their MICA score. The information coefficient [15] and relevance [16] metrics are variations of the Lin measure. The graph IC [17] metric takes into account all of the shared ancestor nodes given any two nodes. The Wang [18] metric assigns a weight to each edge representing its semantic contribution, with the limitation of only being applicable to directed acyclic graphs (DAG).

In the clinical domain, several existing studies have built applications specifically leveraging the HPO graph structure with the aforementioned similarity measurements to check the relevance of any two phenotypic terms. For example, Phenomizer [19] is a clinical diagnostic tool that uses the Resnik metric for phenotypic similarity measurements leveraging the hierarchical structure of the HPO to provide differential diagnostic suggestions. Masino et al. also proposed a clinical phenotypic-based gene prioritization system using semantic similarity derived from the combination of IS-A relationships and IC scores provided by the HPO [20]. OWLSim [21] applied the Jaccard similarity [22] and the Resnik-based approach to support pairwise term-term similarity based on users' manual inputs. Built on top of OWLSim, PhenoDigm [23] applied the mean of the Jaccard and Resnik-based similarities for cross-species phenotype comparisons. PhenomeNET [24] uses the graph IC metric to calculate nodes' similarity derived from different ontologies (e.g., HPO, Mammalian Phenotype Ontology (MPO) [25], Worm Phenotype Ontology (WPO) [26] etc.) for constructing a disease network. The PhenoHM [27] tool is used to integrate human diseases and mouse models according to phenotypic characterizations annotated in both the HPO and MPO using MetaMap [28] similarity scores. PhenoSim [29] leveraged both path-constrained IC and PageRank-based noise reduction methods to measure similarity among the HPO terms. Gong et al. proposed the RelativeBestPair [30] approach utilizing hierarchical IC and the best pair method to measure semantic similarity for terms

contained in the HPO. HPOSim [31] is an HPO-based R package for phenotypic similarity measurement using seven similarity measurements, including the Resnik, Jiang-Conrath, Lin, information coefficient, relevance, graph IC, and Wang metrics.

On the topic of incorporating the HPO with other biomedical resources for knowledge enrichment, several related studies do exist. HPO2GO [32] predicted associations between phenotypic terms using co-occurrences recorded in cross ontology annotation (the HPO and Gene Ontology [33]). OntoFUNC [34] integrated pharmacogenomics databases and biomedical ontologies to identify disease pathway through multi-ontology enrichment. STOP [35] uses over 650,000 k annotations derived from the Gene Ontology, HPO, Disease Ontology [36], and Pathway Ontology [37] for gene and protein annotations. Shen et al. developed a framework using association rule mining [38] to enrich Orphanet annotations for the HPO through data mining from electronic medical records [39]. However, none of these studies have investigated how to build HPO-based distributed semantic representations leveraging heterogeneous resources.

3. Materials

3.1. HPO annotations on disease-phenotype from heterogeneous resources

The HPO team prepared a file "phenotype_annotation.tab" (version released on May 2018) [40] to annotate diseases with relevant HPO phenotypes derived from three knowledge resources: DECIPHER, OMIM, and Orphanet. We leveraged this file to identify relationships amongst HPO phenotypes and enrich relationships utilizing the different resources. DECIPHER aims to facilitate data sharing for phenotypes and genotypes, which stores data for over 28,000 patients. The HPO annotation file incorporates 285 disease-phenotype associations annotated by DECIPHER. The OMIM maintains up-to-date and comprehensive knowledge about human genes and genetic phenotypes that relate to Mendelian disorders. We leveraged 88,169 disease-phenotype associations annotated by the OMIM in this study. The Orphanet is a knowledge resource specifically designed for rare diseases. We found 58,968 disease-phenotype associations annotated by the Orphanet in the HPO annotation file.

3.2. Node2Vec

The node2vec model is designed based on the word2vec model. Each node in the graph is analogous to a single word in text, and a group of neighborhood nodes are similar to the context around said word. The difference is that graph data is represented in a non-linear manner and it is non-trivial to identify a "context" for any single node in the graph. Therefore, node2vec first prepares input data through a random walk based sampling strategy. Specifically, for graph data, node2vec takes two forms of equivalences into consideration: homophily [41] and structural equivalence [42–44]. For homophily equivalence, node2vec uses a breadth first search (BFS) algorithm to reach neighborhood nodes based on homogeneous clusters. For structural equivalence, node2vec leverages a depth first search (DFS) strategy to identify neighborhood nodes based on their structural roles (e.g., hub node or peripheral node). Node2vec makes a mixture of both equivalence and customizes a random walk algorithm to switch between BFS and DFS in a moderate way, in order to balance graph searching. After the input data is ready, the node2vec model applies word2vec to generate node embeddings. There also exist some algorithms that are able to generate node embeddings, such as DeepWalk [45] and Line [46]. DeepWalk is a random walk based algorithm with a fixed return parameter and in-out parameter of 1. The node2vec model extends DeepWalk by providing a more flexible way to train different combinations of these parameters. Therefore, the sampling strategy provided by DeepWalk can be considered as a special case of node2vec. Line provides an efficient way to learn graph embedding leveraging both first-order and second-order proximities. However, it lacks the ability to learn feature representation with a balance between BFS and DFS in a graph network. According to a graph embedding survey paper [47], node2vec is the only work that takes into account both BFS and DFS in a biased random walk algorithm, which is important for learning an enriched embedding with relationships from both intra-communities and inter-communities. Therefore, we chose node2vec in this study on generating node embeddings.

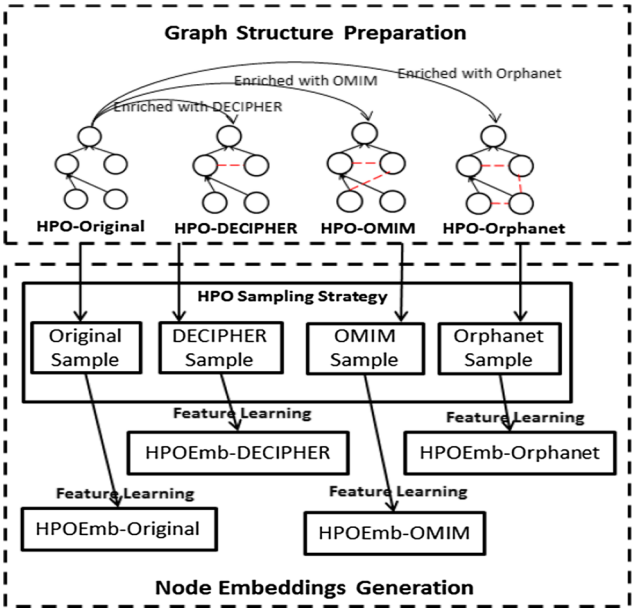

Fig. 1. The workflow of HPO2Vec+.

4. Methods

The workflow of HPO2Vec+ consists of two components as shown in Fig. 1: graph structure preparation and node embeddings generation. The graph structure preparation component first enriches the original HPO graph with phenotype-phenotype associations from knowledge resources and builds the structure of different graphs. The node2vec model then samples network neighborhoods for nodes and performs feature learning to generate node embeddings for each specific graph.

4.1. Preparation of graph structure

In our previous work, we only represented the HPO graph by parsing the IS-A relationships that connected between superclass and subclass nodes. In this study, we extended this work by incorporating three different knowledge resources (DECIPHER, OMIM, and Orphanet) and made updated HPO graphs with enriched non-inheritance relationships amongst the different nodes. For each knowledge resource, we constructed a bipartite graph $G = (D, P, E)$, where D represents disease set, P stands for phenotype sets and E denotes edges between diseases and phenotypes. For any two phenotypes $p_i \in P$ and $p_j \in P$, if there exists a common disease $d \in D$ that could be connected to both p_i and p_j through edges $e_i \in E$ and $e_j \in E$ respectively, we linked a sibling edge between p_i and p_j in the original HPO graph. As shown in Fig. 1, four different graphs were generated for further processing: the original graph HPO-Original generated in our previous work and three new graphs HPO-DECIPHER, HPO-OMIM, and HPO-Orphanet generated by HPO2Vec+ in this study.

4.2. Generation of node embeddings

4.2.1. HPO sampling strategy

Node2vec first analyzed each specific HPO graph and adopted a biased random walk algorithm to sample neighborhood sets for each individual node.

Node2vec adopts a 2nd order random walk based on the topology of any graph, indicating that three nodes will be involved in a walk, namely source node, intermediate node, and target node. We defined any source node as ph_i, any target node as ph_j, and any intermediate node that exists on the path between ph_i and ph_j as ph_x, the distribution of phenotypic node ph_j with a fixed length of random walk can be represented as:

$$P(ph_j|ph_x) = \begin{cases} \frac{\pi(ph_x, ph_j)}{Z} & if \ < ph_x, ph_j > forms \ an \ edge \\ 0 & otherwise \end{cases} \quad (1)$$

where Z indicates the normalization constant. A transition probability $\pi(ph_x, ph_j)$ between phenotypic nodes ph_x and ph_j can be calculated as:

$$\pi(ph_x, ph_j) = \alpha(ph_i, ph_j) \bullet w(ph_x, ph_j) \quad (2)$$

Here $w(ph_x, ph_j)$ indicates the weight assigned to the edge between ph_x and ph_j. Specifically, in this study, we set $(ph_x, ph_j) = 1$.

To determine the search bias term α with respect to ph_i and ph_j, Node2vec used a re-visitation parameter p and an in-out parameter q to control the balance between BFS and DFS to implement homophily and structural equivalence simultaneously. α for phenotypic nodes ph_i and ph_j is computed based on p and q:

$$\alpha(ph_i, ph_j) = \begin{cases} \frac{1}{p} & if \ sp(ph_i, ph_j) = 0 \\ 1 & if \ sp(ph_i, ph_j) = 1 \\ \frac{1}{q} & if \ sp(ph_i, ph_j) = 2 \end{cases} \quad (3)$$

where $sp(ph_i, ph_j)$ indicates the shortest path between ph_i and ph_j.

4.2.2. Feature learning

A feature learning component was then used to construct different HPO embeddings based on different sample strategies. Let HPO graph $HG = (PH, E)$ be a specified network, where PH denotes phenotypic nodes and E indicates edges between nodes. Let f: $PH \rightarrow \mathbb{R}^d$ denote the mapping function from any HPO phenotypic nodes to their corresponding feature representation, where d specifies dimensions and f is a matrix of size $|PH| \times d$.

Node2vec extends the Skip-gram model provided by word2vec and applies it on each sampled neighborhood node using the aforementioned biased random walk algorithm. For each phenotypic node $ph_i \in PH$, we used $\mathrm{N}(ph_i)$ to represent the neighbors of a phenotype ph_i, with an objective function for feature learning described as:

$$\max_f \sum_{ph_i \in PH} \log P(\mathrm{N}(ph_i)|f(ph_i)) \quad (4)$$

We used a softmax function to produce a vector of normalized probabilities for each neighbor n_k and node feature $f(ph_i)$:

$$P(n_k|f(ph_i)) = \frac{\exp(f(n_k) \bullet f(ph_i))}{\sum_{ph_j \in PH} \exp(f(ph_j) \bullet f(ph_i))} \quad (5)$$

Two assumptions are made by node2vec. First, it assumes that the likelihood of observing any neighborhood from one source phenotypic node is conditionally independent. In addition, it assumes that the feature space between any phenotypic node and its neighborhood node is symmetric. Based on these two assumptions, Eqs. (4) and (5) were combined and an objective function was simplified as shown in Eq. (6), where $T(ph_i) = \sum_{ph_j \in PH} \exp(f(ph_j) \bullet f(ph_i))$. Given a constant window size, the Stochastic gradient descent algorithm was applied to optimize this objective function.

$$\max_f \sum_{ph_i \in PH} \left[-\log T(ph_i) + \sum_{n_k \in N(ph_i)} f(n_k) \bullet f(ph_i) \right] \quad (6)$$

Fig. 2 shows an example of learning node embeddings for four graphs given an input node *nephrolithilasis* with the Skip-gram model. Since various graphs provide their own insight to define neighbors for *nephrolithilasis* using biased random walk, node2vec was able to leverage the sampled neighbors within a certain window size to train embeddings. As a result, four node embeddings for the HPO were generated based on differing graph structures which we hereafter label as HPOEmb-Original (the basic embeddings done in previous work), HPOEmb-DECIPHER, HPOEmb-OMIM, and HPOEmb-Orphanet.

5. Graph characterizations

We first characterized details of different HPO graphs that were used in the evaluation. In addition to using the size of nodes and edges for graph characterizations, the average degree and density were also

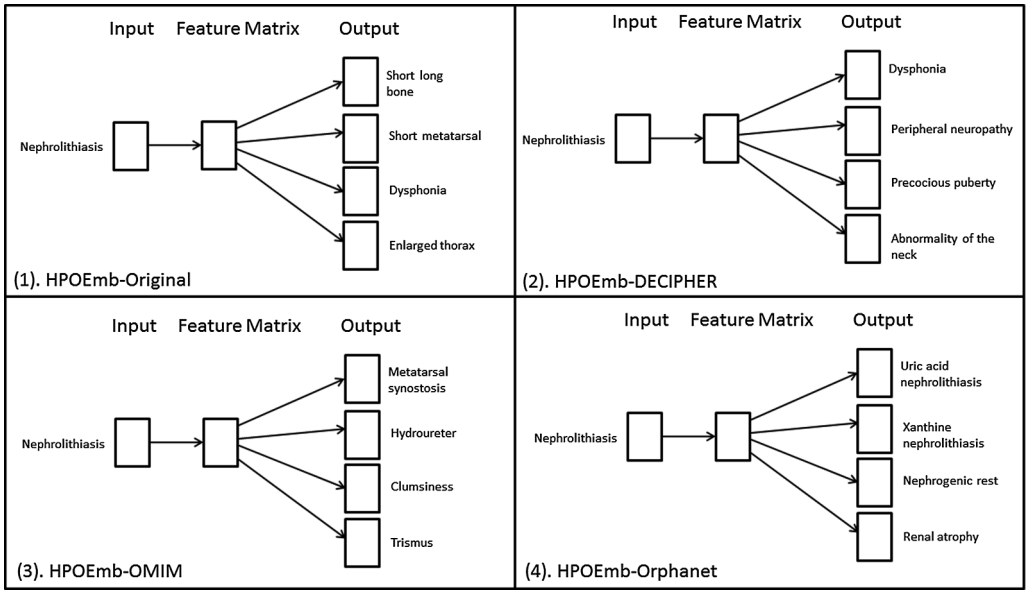

Fig. 2. An example of feature learning process for different graphs.

introduced here. For any given graph $G = (V, E)$, average degree and density for G are defined as shown in Eqs. (7) and (8) respectively, where $|V|$ indicates the number of vertices and $|E|$ denotes the number of edges in the graph.

$$\bar{\Delta}(G) = \frac{2|E|}{|V|} \tag{7}$$

$$D(G) = \frac{2\ |E|}{|V|(|V|-1)} \tag{8}$$

HPO-Original is a subgraph of the HPO containing 7,258 nodes, a size consistent to that reported in our previous work [10]. We enriched the phenotypic relationships extracted from HPO-Original using 87, 20,881, and 21,171 disease-phenotype associations that have phenotypic terms overlapping with those contained in HPO-Original from DECIPHER, OMIM, and Orphanet respectively. The HPO-DECIPHER, HPO-OMIM, and HPO-Orphanet graphs all hold a consistent node count of 7,258 but contain differing numbers of edges. We made specific characterizations for different HPO graphs as shown in Table 1. In the experiment, we used a subset of the entire edges for each graph that formed among these 7,258 nodes to learn embeddings. We observed that without any enrichment, HPO-Original has the lowest average degree and density. HPO-DECIPHER does not show much increment on average degree and density due to a fewer number of enriched associations. Although the OMIM holds fewer disease-phenotype associations than the Orphanet, HPO-OMIM has more new edges created than HPO-Orphanet, which leads to the highest average degree and density.

6. Evaluation approaches

6.1. Quantitative evaluation

We first generated the optimal embeddings by conducting a link prediction task. The aim of the link prediction task is to predict the relationships between any two HPO nodes (positive or negative) and use the performance of prediction to evaluate the quality of the four generated embeddings. Edge embeddings were used in this task to investigate relationships between nodes leveraging distributional features provided by node embeddings. Given any two phenotypic nodes ph_i and ph_j and their corresponding feature representations $f(ph_i)$ and $f(ph_j)$, edge embeddings can be calculated using four operations as shown in Table 2.

For any given phenotypic nodes ph_i and ph_j, a boolean function $L(ph_i, ph_j)$ was used to indicate the existence of edge(s) between these two nodes, where $L(ph_i, ph_j) = 1$ indicates positive links and $L(ph_i, ph_j) = 0$ denotes negative links. During the training process, we fit features provided by edge embeddings along with labels provided by $L(ph_i, ph_j)$ to build the model. For positive examples, for each of the four graphs, we randomly used 60%, 10%, and 30% of all their edges for training, validation, and testing purposes respectively. For negative examples, we randomly sampled an equal number of node pairs and kept the same ratio amongst training, validation, and testing sets as 60%, 10%, and 30% respectively. Specifically, we used the Decision Tree (DT) [51], Logistic Regression (LR) [52], Support Vector Machine (SVM) [53], Random Forest (RF) [54], Naïve Bayes (NB) [55], and Multi-Layer Perceptron (MLP) [56] machine learning algorithms to perform the evaluation. We plotted the receiver operating characteristic (ROC) curve and computed the area under the ROC curve (AUROC) to report link prediction performance. In addition, as shown in Eqs. (9)–(12), precision, recall, F1 score, and average precision were used to quantify the performance on link prediction for the different HPO embeddings.

$$Precision = \frac{|\{TrueRelations\} \cap \{PredictedRelations\}|}{|\{PredictedRelations\}|} \tag{9}$$

$$Recall = \frac{|\{TrueRelations\} \cap \{PredictedRelations\}|}{|\{TrueRelations\}|} \tag{10}$$

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

$$AP = \sum_n (Recall_n - Recall_{n-1}) Precision_n \tag{12}$$

In order to evaluate if the generated HPO embeddings had the capability to assist patient stratification in clinical practice, in this section, we conducted another task on detecting similar patients for a list of rare diseases using the optimal HPO embeddings and other semantic similarity measurements. Specifically, we first selected 10 rare diseases with high prevalence in electronic health records (EHR) generated at Mayo Clinic from 2010 to 2015. These rare diseases are *ovarian cancer*, *myelofibrosis*, *primary sclerosing cholangitis*, *b-cell lymphoma*, *dilated cardiomyopathy*, *laryngomalacia*, *cryoglobulinemia*, *esophageal cancer*, *papillary thyroid carcinoma*, and *clear cell renal cell carcinoma*.

We first filtered out patients who have more than one rare disease. For each aforementioned rare disease, we then randomly selected 5 patients who already received a final diagnosis. We applied our previously developed HPO annotation pipeline [57] on the diagnosis section of each patients' EHR within 12 months of their confirmed diagnosis of a rare disease, in order to extract annotated phenotypes and assembled them together for each patient as a phenotype vector. For each rare disease, we compared each patient with 4 other patients that had the same diagnosis (each disease has $5 * 4 = 20$ pairwise comparisons) by applying different measurements solely based on patients' phenotypic characterizations. This experiment was done with the aim of investigating which measurement can achieve a better performance for detecting similar patients for each specific rare disease.

As shown in eq. (13), for each patient p, we applied the optimal HPO embeddings over the phenotype vector phv_p and calculated the average embeddings (AE), where f is a feature matrix to map any phenotype ph_p to its embeddings. A cosine similarity calculation (PatientSim) was followed to compare similarity between any two patients using average embeddings (Eq. (14)).

$$AE(p) = \frac{\sum_{ph_p \in phv_p}^{|phv_p|} f(ph_p)}{|phv_p|} \tag{13}$$

$$PatientSim(p_i, p_j) = \frac{AE(p_i) \bullet AE(p_j)}{\|AE(p_i)\| \|AE(p_j)\|} \tag{14}$$

We further compared the optimal embedding generated by HPO2Vec+ with conventional phenotypic similarity measurements. Specifically, we selected HPOSim as it implements seven different ontology-based similarity measurements and it uses the HPO as the backbone ontology. These similarity measurements are Resnik (HPOSim-Resnik), Jiang-Conrath (HPOSim-JC), Lin (HPOSim-Lin), information coefficient (HPOSim-ICE), relevance (HPOSim-RL), graph IC (HPOSim-GIC), and Wang (HPOSim-Wang). In this task, since we only used the positive space for cosine similarity bounded in [0,1], to make a fair comparison, we adopted five out of seven conventional similarity measurements with the same value range: HPOSim-Lin, HPOSim-ICE,

Table 1
Graph characterizations for four graphs.

Graphs	Number of nodes	Number of edges	Average degree	Density
HPO-Original	7,258	16,250	4.48	6.17E-04
HPO-DECIPHER	7,258	16,534	4.56	6.28E-04
HPO-OMIM	7,258	130,916	36.07	0.005
HPO-Orphanet	7,258	93,944	25.89	0.004

Table 2
Operations to generate edge embeddings.

Operations	Definition
Hadamard [48]	$f(ph_i) * f(ph_j)$
Average	$\frac{f(ph_i)+f(ph_j)}{2}$
L1 [49]	$\lvert f(ph_i) - f(ph_j)\rvert$
L2 [50]	$\lvert f(ph_i) - f(ph_j)\rvert^2$

HPOSim-RL, HPOSim-GIC, and HPOSim-Wang.

6.2. Qualitative evaluation

We first visualized phenotypic clusters generated by different HPO embeddings with a two-dimensional plot leveraging t-distributed stochastic neighbor embedding (t-SNE) [58]. We then conducted a use case study on *primary hyperoxaluria* (PH). PH is a rare heterogeneous disease but is an important cause of variable progression into kidney failure from childhood through adolescence [59]. Differential diagnosis among *recurrent nephrolithiasis*, *nephrocalcinosis*, or *end-stage renal disease* (ESRD) must be provided before making a confirmed diagnosis for PH [60]. Since delayed diagnosis of PH is extremely common, given its rarity, it is meaningful if phenotypic characterizations can accelerate the differential diagnosis process. Specifically, for each selected phenotypes related to PH, we identified the most relevant phenotypes using cosine similarity with four HPO embeddings as shown in Eq. (15), where ph_i denote annotated PH related phenotypes and ph_j indicate each phenotype inferred by node embeddings, and V_i and V_j denote the embeddings for ph_i and ph_j respectively.

$$similarity(ph_i, ph_j) = \frac{V_i \bullet V_j}{\|V_i\|\|V_j\|} \tag{15}$$

In addition, we invited a nephrologist as a domain expert to validate the usefulness on phenotypic relevance detection with different HPO embeddings. Specifically, the nephrologist considered whether the inferred phenotypes are relevant if they are comorbidities of the given phenotypes, or holds a similar meaning to that of the given phenotypes, or shares the same risk factors (risk equivalent [61]) as the given phenotypes. As shown in Eq. (16), we used information retrieval metric precision at k (P@K) to quantify the domain expert's evaluation, where K inferred phenotypes indicate top K phenotypes recommended by the embeddings with the descending order of cosine similarity. We compared the top inferred phenotypes for each of the PH related phenotypes respectively given by HPO2Vec + and seven conventional similarity measurements (HPOSim-Resnik, HPOSim-JC, HPOSim-Lin, HPOSim-ICE, HPOSim-RL, HPOSim-GIC, and HPOSim-Wang). Relevance for each inferred phenotype was also evaluated by the domain expert.

$$P@K = \frac{|\{RelevantPhenotypes\} \cap \{KInferredPhenotypes\}|}{|\{KInferredPhenotypes\}|} \tag{16}$$

7. Results

7.1. Embeddings generation with link prediction task

We used a neighbor size of 10, number of walks of 10, walk length of 5, and dimensions of 128 to generate four different embeddings. We leveraged the validation set to tune p and q. By heuristics, we chose $p, q \in \{0.05, 0.25, 1\}$, and the optimal combinations for p and q for HPOEmb-Original, HPOEmb-DECIPHER, HPOEmb-OMIM, and HPOEmb-Orphanet are (1, 0.05), (1, 0.25), (1, 0.05), and (1, 0.05), respectively.

Table 3 illustrates the link prediction performance (AUC) for four HPO embeddings with different edge embeddings operations. In general, we found that HPOEmb-DECIPHER did not show much difference on link prediction compared to HPOEmb-Original. HPOEmb-OMIM achieved its optimal performance (AUC = 0.9) when using the LR and RF algorithms with either the Hadamard or Average operations, while the performance was even worse than HPOEmb-Original and HPOEmb-DECIPHER with the L1 and L2 operations. With the Hadamard and Average opoerations, HPOEmb-Orphanet achieved an optimal AUC of 0.92 with both of the RF and LR algorithms. HPOEmb-Orphanet did not perform well with the L1 and L2 ooperations. ROC curves for different machine learning algorithms with the optimal HPOEmb-Orphanet embeddings using Hadamard is shown in Fig. 3. A comprehensive ROC analysis for all the HPO embeddings with different

Table 3
AUCs for link prediction performance amongst four HPO embeddings with six machine learning algorithms using four edge embeddings operations. (The best AUC for each edge embeddings in bold.)

Operations	Algorithms	HPOEmb-Original	HPOEmb-DECIPHER	HPOEmb-OMIM	HPOEmb-Orphanet
Hadamard	DT	0.69	0.69	0.77	0.8
	LR	0.76	0.76	0.9	**0.92**
	SVM	0.72	0.73	0.84	0.85
	RF	0.79	0.79	0.9	**0.92**
	NB	0.73	0.73	0.82	0.85
	MLP	0.72	0.72	0.84	0.86
Average	DT	0.7	0.69	0.76	0.8
	LR	0.79	0.79	0.9	**0.92**
	SVM	0.75	0.75	0.84	0.86
	RF	0.79	0.8	0.9	**0.92**
	NB	0.75	0.75	0.84	0.86
	MLP	0.75	0.74	0.84	0.86
L1	DT	0.67	0.66	0.58	0.57
	LR	**0.81**	0.8	0.71	0.67
	SVM	0.75	0.74	0.66	0.64
	RF	0.8	0.8	0.73	0.71
	NB	0.75	0.75	0.65	0.62
	MLP	0.75	0.75	0.67	0.63
L2	DT	0.67	0.66	0.58	0.57
	LR	**0.8**	**0.8**	0.7	0.66
	SVM	0.74	0.74	0.64	0.61
	RF	**0.8**	**0.8**	0.73	0.72
	NB	0.74	0.74	0.63	0.6
	MLP	0.75	0.74	0.65	0.63

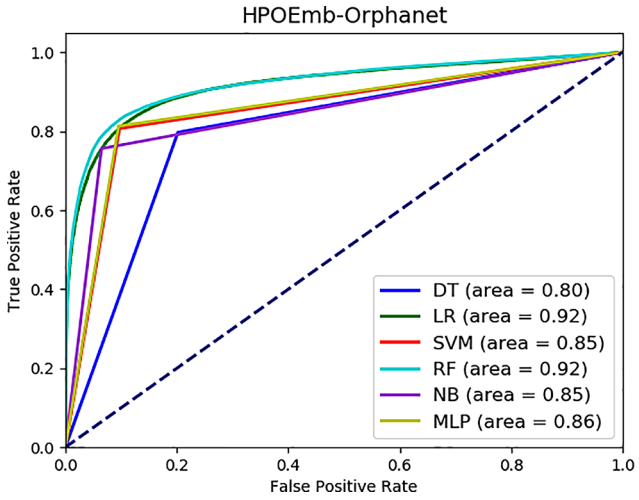

Fig. 3. ROC curves and AUCs on link prediction performance for HPOEmb-Orphanet with six machine learning algorithms using Hadamard.

experiment settings are illustrated in Supplementary File 1.

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103246.

We selected HPOEmb-Orphanet as the optimal embeddings and calculated average precision for four edge embeddings operations and six machine learning algorithms as shown in Table 4. Average (optimal: 0.94) and Hadamard (suboptimal: 0.93) performed better than the other two operations with RF (0.74 for both). In addition, the highest F1 score (0.86) for HPOEmb-Orphanet was achieved with Hadamard and Average. Comprehensive evaluation results on the precision, recall, and F1 scores for all HPO embeddings is provided in Supplementary File 2.

7.2. Patient similarity measurement

As shown in Table 5, HPOEmb-Orphanet with cosine similarity can achieve the highest average patient similarity score across ten rare diseases (0.71) as well as the highest similarity score for 6 out of 10 rare diseases, including *ovarian cancer*, *myelofibrosis*, *primary sclerosing cholangitis*, *b-cell lymphoma*, *cryoglobulinemia*, and *esophageal cancer*.

By checking the percentage of shared phenotypes by each pair of patients, we observed that HPOEmb-Orphanet performed better than others especially for patient pairs with fewer shared phenotypes ($<=10\%$ in our experiment). For example, two patients with *esophageal cancer* had phenotype vectors of [*dehydration*, *nausea*, *vomiting*] and [*gastrointestinal stroma tumor*, *neoplasm*] respectively. One patient's observable phenotypes were all about digestive problem and abnormal fluid regulation, and the other patient's phenotypic characterizations were about neoplasm of the gastrointestinal tract. Since phenotypes from these two patients are neither directly connected nor within short reachable distance according to the HPO graph, all five conventional similarity measurements can hardly consider they are relevant. HPOEmb-Orphanet increased the chance of detecting these patients as similar even though the phenotypic nodes in the graph were separate. For each disease for which any two patients share many common phenotypes, HPOEmb-Orphanet was also able to detect these similar patients in a manner consistent with HPO-Lin, HPO-GIC, and HPO-Wang (e.g., 1 for *b-cell lymphoma*, 0.95 for *dilated cardiomyopathy*, 0.9 for *laryngomalacia*, 0.97 for *papillary thyroid carcinoma*, and 0.96 for *clear cell renal cell carcinoma*). Although some of the conventional similarity measurements achieved a higher patient similarity based on commonly shared phenotypes, it is more interesting and meaningful to consider any two patients as similar even if they held a large number of phenotypic characterizations that seem to be different but are actually relevant, which is the motivation of constructing HPOEmb-Orphanet.

7.3. Phenotypic cluster visualization

We picked one cluster from each subset of randomly generated HPO embeddings as shown in Fig. 4 (all the subsets of randomly generated clusters could be found in Supplementary File 3). Fig. 4.1 depicts a cluster generated by HPOEmb-Original including the greatest number of phenotypic terms with coarse granularity in the HPO graph, such as *abnormal epiphyseal ossification*, *abnormal lung morphology*, *abnormality of muscle morphology*, *abnormal proportion of double-negative alpha–beta regulatory T cell*. The other three HPO embeddings were able to generate more clusters with phenotypic terms in finer granularity. As shown in Fig. 4.2, a cluster generated by HPOEmb-DECIPHER contains six phenotypes that are inherited from different root phenotypic terms. For example, *compensatory chin elevation* belongs to *abnormality of the eye*, *high forehead* is a subtype of *abnormality of head or neck*, *increased hematocrit* belongs to *abnormality of blood and blood-forming tissues*, *hamartoma* is a *neoplasm*, *emphysema* is an *abnormality of the respiratory system*, and *progressive alopecia* is inherited from *abnormality of the integument*. *Progressive alopecia* and *increased hematocrit* appeared in clusters generated by both HPOEmb-DECIPHER and HPOEmb-OMIM, but the latter assigned different neighbors to these two phenotypes. For example, as shown in Fig. 4.3, *skeletal muscle atrophy* (a subclass of *abnormality of the musculature*), *low back pain* (a subclass of *abnormality of the skeletal system*), and *micromelia* (a subclass of *abnormality of limbs*) are also included in the selected cluster. As shown in Fig. 4.4, the cluster generated by HPOEmb-Orphanet shared *hamartoma* with HPOEmb-DECIPHER and shared *skeletal muscle atrophy* with HPOEmb-OMIM. In addition, this cluster also included *elevated right atrial pressure* (a subclass of *abnormality of the cardiovascular system*), *increased serum bile acid concentration during pregnancy* (a subclass of *abnormality of metabolism/homeostasis*), and *pericarditis* (a subclass of *abnormality of the cardiovascular system*).

7.4. Use case study

We selected 22 phenotypes annotated by the HPO that are proved to be related to PH and also overlapped with our generated graphs. As shown in Table 6, the average P@5 and P@10 were calculated for these 22 phenotypes based on the domain expert's comments (Detailed information for domain expert's evaluation and cosine similarity scores can be found in Supplementary File 4). For both P@5 and P@10, HPOEmb-Orphanet achieved the optimal overall performance amongst all HPO embeddings, with a P@5 slightly higher than P@10. HPOEmb-Original yielded suboptimal precision for the top 5 as well as for the top 10 phenotypes. HPOEmb-DECIPHER has the same performance as HPOEmb-Original for P@5, and HPOEmb-DECIPHER outperformed HPOEmb-OMIM for both P@5 and P@10. The performance for each of the 22 selected phenotypes can also be found in Supplementary File 5.

We further picked *nephrocalcinosis*, *stage 5 chronic kidney disease*, and *calcium oxalate nephrolithiasis* out of the 22 phenotypes as significant phenotypes for PH patient stratification, as suggested by the domain expert, in order to compare qualitative performance between optimal HPOEmb-Orphanet and HPOSim with seven similarity measurements.

Table 4
Average precision for link prediction performance on HPOEmb-Orphanet with six machine learning algorithms using four edge embeddings operations. (The best average precision in bold.)

Operations	DT	LR	SVM	RF	NB	MLP
Hadamard	0.74	**0.93**	0.82	**0.93**	0.82	0.82
Average	0.74	0.93	0.82	**0.94**	0.82	0.82
L1	0.54	0.69	0.6	**0.74**	0.59	0.59
L2	0.54	0.66	0.58	**0.74**	0.58	0.59

Table 5
Patient similarity measurement between HPOEmb-Orphanet and five conventional similarity measurements (The highest similarity score in bold.)

Rare diseases	Shared same phenotypes per pairwise patients (average percentage)	HPOSim-Lin	HPOSim-ICE	HPOSim-RL	HPOSim-GIC	HPOSim-Wang	HPOEmb-Orphanet
Ovarian cancer	6.67%	0.42	0.31	0.38	0	0.41	**0.55**
Myelofibrosis	2.5%	0.25	0.17	0.28	0.11	0.23	**0.33**
Primary sclerosing cholangitis	10%	0.42	0.29	0.36	0.05	0.46	**0.56**
b-cell lymphoma	50%	**1**	0.83	0.99	**1**	**1**	**1**
Dilated cardiomyopathy	43.2%	**0.97**	0.72	0.91	0	0.96	0.95
Laryngomalacia	43.3%	**1**	0.82	0.99	**1**	**1**	0.9
Cryoglobulinemia	9.64%	0.09	0.06	0.08	0.04	0.4	**0.56**
Esophageal cancer	5%	0.23	0.17	0.22	0	0.25	**0.36**
Papillary thyroid carcinoma	43.3%	**1**	0.84	0.99	**1**	**1**	0.97
Clear cell renal cell carcinoma	46%	**1**	0.82	0.99	**1**	**1**	0.96
Average	26%	0.64	0.5	0.62	0.42	0.67	**0.71**

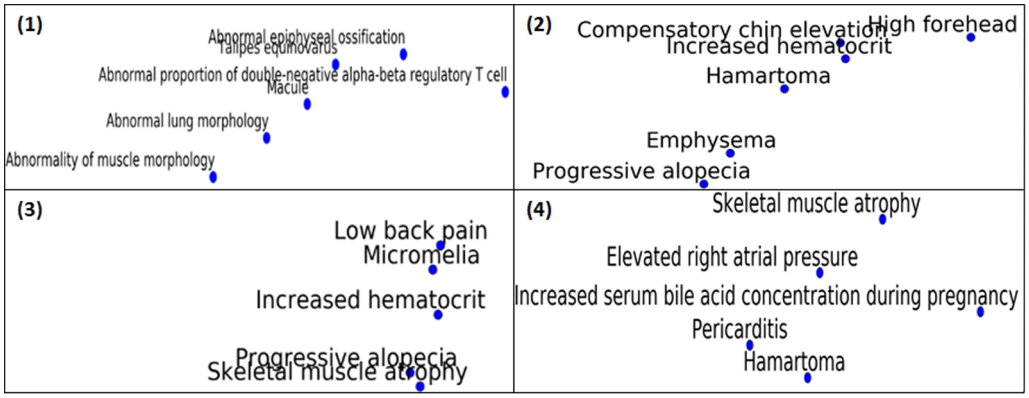

Fig. 4. Examples of phenotypic clusters in the visualization of four HPO embeddings using t-SNE.

Specifically, we applied HPOSim on the HPO-Orphanet graph for a fair comparison. Table 7 depicts top 5 inferred phenotypes for each of the selected three phenotypes using different metrics. For *nephrocalcinosis*, we found that only HPOEmb-Orphanet inferred 4 out of 5 relevant phenotypes but all the other similarity measurements failed to generate relevant phenotypes. For *stage 5 chronic kidney disease*, all the measurements detected relevant phenotypes. However, seven conventional similarity measurements implemented in HPOSim all inferred the exact same term "*stage 5 chronic kidney disease*" as one of the top answers. In addition, *chronic kidney disease* and *renal insufficiency* are phenotypes that were highly recommended for all different measurements. We found that all the inferred phenotypes provided by the seven conventional similarity measurements belong to the urinary system. In other words, those inferred phenotypes were detected via solely traversing the hierarchical structure of the HPO graph. HPOEmb-Orphanet was, however, able to infer relevant signs and symptoms from different organ systems that are not predecessor or descendent nodes of the urinary system in the HPO. For example, *secondary hyperparathyroidism* belongs to *abnormality of the parathyroid physiology* and *polyarticular arthritis* belongs to *abnormality of the skeletal system*. For *calcium oxalate nephrolithiasis*, HPOSim-Resnik detected the greatest number of relevant phenotypes, but all of the inferred phenotypes were either exactly the same phenotype or a subclass/superclass of *calcium oxalate nephrolithiasis*. HPOSim-RL inferred *increased urinary sulfite*, which is a urinary phenotype in fine granularity. Other conventional similarity measurements inferred either non-relevant phenotypes or superclass term (*nephrolithiasis*). Although HPOEmb-Orphanet only inferred 3 out of 5 relevant phenotypes, it provided insights to link diseases in the urinary system (*calcium oxalate nephrolithiasis*, *renal calcium wasting*, and *hypercalciuria*) to diseases in the endocrine system (*parathyroid hyperplasia*).

8. Conclusion and discussion

In this study, we proposed a framework, HPO2Vec+, to enrich node embeddings for the HPO leveraging heterogeneous biomedical and clinical knowledge resources. We used DECIPHER, OMIM, and Orphanet to generate enriched HPO embeddings, namely HPOEmb-DECIPHER, HPOEmb-OMIM, and HPOEmb-Orphanet. For the quantitative evaluation, we evaluated performance on link prediction amongst three aforementioned embeddings and our previously constructed HPO embeddings (HPOEmb-Original). For the qualitative evaluation, we analyzed different node embeddings through clustering visualization. We then conducted a use case study on PH and invited a domain expert to evaluate the quality of the inferred phenotypes generated by the different embeddings as well as by conventional graph-based similarity measurements, given any annotated PH related phenotypes. Results indicated that HPOEmb-Orphanet outperformed both the other HPO embeddings and conventional similarity measurements in all evaluations. This study showed that by combining enriched non-inheritance edges derived from appropriate knowledge resources with HPO hierarchy and gene annotations, HPO2Vec+ is able to provide a

Table 6
Average of P@5 and P@10 amongst four HPO embeddings. (The best score in bold.)

Metrics	HPOEmb-Original	HPOEmb-DECIPHER	HPOEmb-OMIM	HPOEmb-Orphanet
Average of P@5	0.25	0.25	0.21	**0.81**
Average of P@10	0.26	0.23	0.16	**0.79**

Table 7
Inferred relevant phenotypes for nephrocalcinosis, stage 5 chronic kidney disease, and calcium oxalate nephrolithiasis generated by HPOEmb-Orphanet and HPOSim with seven conventional similarity measurements. (Phenotypes in bold indicate they are relevant to a corresponding PH-related phenotype according to the domain expert's feedback.)

Methods	Nephrocalcinosis	Stage 5 chronic kidney disease	Calcium oxalate nephrolithiasis
HPOEmb-Orphanet	1. **Medullary nephrocalcinosis**	1. **Chronic kidney disease**	1. **Renal calcium wasting**
	2. **Decreased numbers of nephrons**	2. **Secondary hyperparathyroidism**	2. **Hypercalciuria**
	3. **Hyperechogenic kidneys**	3. **Renal insufficiency**	3. Parietal bossing
	4. **Renal atrophy**	4. **Polyarticular arthritis**	4. Obtuse angle of mandible
	5. Nephrosclerosis	5. **Synovitis**	5. **Parathyroid hyperplasia**
HPOSim-Resnik	1. Overbite	1. **Stage 5 chronic kidney disease**	1. **Calcium oxalate nephrolithiasis**
	2. Abnormal glomerular filtration rate	2. **Chronic kidney disease**	2. **Calcium nephrolithiasis**
	3. Severe periodontitis	3. **Renal insufficiency**	3. **Nephrolithiasis**
	4. Premature loss of permanent teeth	4. **Acute kidney injury**	4. **Uric acid nephrolithiasis**
	5. Agenesis of incisor	5. **Glomerulonephritis**	5. **Xanthine nephrolithiasis**
HPOSim-JC	1. Ulnar claw	1. **Renal insufficiency**	1. Delayed calcaneal ossification
	2. Central hypothyroidism	2. **Stage 5 chronic kidney disease**	2. Mask-like facies
	3. Mucopolysaccharidurina	3. **Abnormal renal physiology**	3. Abnormal atrial septum morphology
	4. Abnormality of the 5th metacarpal	4. **Abnormality of the urinary system physiology**	4. **Nephrosclerosis**
	5. Percussion myotonia	5. **Abnormality of the kidney**	5. Schizencephaly
HPOSim-Lin	1. Orthostatic tachycardia	1. **Stage 5 chronic kidney disease**	1. Impaired pursuit initiation and maintenance
	2. Childhood onset short-limb short stature	2. **Chronic kidney disease**	2. Thenar muscle atrophy
	3. Odontoma	3. **Renal insufficiency**	3. **Decreased urinary urate**
	4. Finger symphalangism	4. **Abnormal renal physiology**	4. **Urocanic aciduria**
	5. Acute necrotizing encephalopathy	5. **Abnormality of the urinary system physiology**	5. **Elevated urinary catecholamines**
HPOSim-ICE	1. Foam cells	1. **Stage 5 chronic kidney disease**	1. Irregular myelin loops
	2. Abnormality of the fallopian tube	2. **Chronic kidney disease**	2. Fused cervical vertebrae
	3. Increased mean platelet volume	3. **Renal insufficiency**	3. Precocious puberty in males
	4. Aplasia/Hypoplasia of the fibula	4. **Abnormal renal physiology**	4. Hemiparesis
	5. Congenital nonbullous ichthyosiform erythroderma	5. **Nephrotic syndrome**	5. Abnormality of forebrain morphology
HPOSim-RL	1. Colon cancer	1. **Stage 5 chronic kidney disease**	1. Rib segmentation abnormalities
	2. Abnormality of the seventh cranial nerve	2. **Chronic kidney disease**	2. Cerebral calcification
	3. Facial hypertrichosis	3. **Renal insufficiency**	3. Vacuolated lymphocytes
	4. Progressive flexion contractures	4. **Abnormal renal physiology**	4. **Increased urinary sulfite**
	5. Mandibular aplasia	5. **Nephrotic syndrome**	5. Abnormality of brainstem morphology
HPOSim-GIC	1. Short nail	1. **Stage 5 chronic kidney disease**	1. Hypermobility of distal interphalangeal joints
	2. Abnormal autonomic nervous system physiology	2. **Acute kidney injury**	2. Duodenal atresia
	3. Leiomyosarcoma	3. **Chronic kidney disease**	3. Delayed peripheral myelination
	4. Abnormal platelet membrane protein expression	4. **Renal insufficiency**	4. Curved linear dimple below the lower lip
	5. Dilated vestibule of the inner ear	5. **Nephrotic syndrome**	5. Spastic tetraplegia
HPOSim-Wang	1. Ulnar claw	1. **Stage 5 chronic kidney disease**	1. Delayed calcaneal ossification
	2. Corneal stromal edema	2. **Chronic kidney disease**	2. Radial metaphyseal irregularity
	3. Central hypothyroidism	3. **Renal insufficiency**	3. Mask-like facies
	4. Mucopolysaccharidurina	4. **Abnormal renal physiology**	4. Abnormal atrial septum morphology
	5. Abnormal blistering of the skin	5. **Acute kidney injury**	5. **Nephrosclerosis**

generalized way to enhance HPO embeddings and better assist in detecting phenotypic relevance for deep phenotyping. Resources could be found at: https://github.com/shenfc/HPO2Vec.

For the link prediction task in quantitative evaluation, although the HPO-DECIPHER graph had a higher average degree and density than the HPO-Original graph, the performance did not show much improvement over HPOEmb-Original, which may be due to the relatively small number of enriched phenotype-phenotype relationships derived from DECIPHER with only 316 pairs. The HPO-OMIM graph had the largest average degree and density but HPOEmb-OMIM did not achieve the best performance, which may be caused by the lack of differentiation of significance based on random walk due to the large number of phenotype-phenotype relationships detected from the OMIM with a total of 116,751 pairs. The top five phenotypic nodes in the HPO-OMIM graph ranked by the number of phenotype-phenotype relationships are *hypertelorism*, *depressed nasal bridge*, *malar flattening*, *frontal bossing*, and *downslanted palpebral fissures*. Specifically, *hypertelorism* has 1,423 connections, *depressed nasal bridge* has 1,152 connections, *malar flattening* has 1,030 connections, *frontal bossing* has 1,005 connections, and *downslanted palpebral fissures* has 984 connections. Such a large number of enriched connections resulted in high in-/out-degrees at the point of choosing next step during random walk algorithm, which could weaken the importance of existing associations. The large number of connections in HPO-OMIM is primarily due to the genetic oriented characteristics of OMIM. For example, *osteogenesis imperfecta (type X)* is recorded to have a strong relationship with gene SERPINH1 in OMIM. Therefore, HPO-OMIM associates this disease with *nephrocalcinosis* and another 29 phenotypes solely based on this specific gene. As genetic information could be inherited by hierarchical superclasses or subclasses, more phenotypes could be considered to be related, which hampers the capability of detecting phenotypic relevance. As shown in Supplementary File 6, we found that with dimension size 128, HPOEmb-Orphanet achieved the similar performance compared to the usage of other options (i.e., 32, 64, and 256). In HPO-Orphanet, *hypertelorism* has the largest number of connections with other phenotypic nodes, a total of 943 connections, smaller than the 1,423 connections in HPO-OMIM. Since Orphanet tends to provide more descriptions from the perspective of clinical observations, there are relatively fewer overlaps between phenotypic nodes, which highlight significant phenotype-phenotype connections.

For the patient similarity measurement task, we also tested HPOEmb-Original, HPOEmb-DECIPHER, and HPOEmb-OMIM on the same 10 rare diseases mentioned in Section 7.2 and found that the similarity scores generated by those three HPO embeddings were all

above 0.99. However, similarities between patient groups across many combinations of totally different diseases (e.g., *idiopathic pulmonary fibrosis* patients and PH patients) were also above 0.99. This showed that the phenotypic embeddings generated by HPOEmb-Original, HPOEmb-DECIPHER, and HPOEmb-OMIM could not make a differentiation. Compared to these three embeddings, HPOEmb-Orphanet can represent appropriate distributed semantic representations among phenotypes leveraging both the entire HPO topological structure as well as disease-phenotype associations maintained in Orphanet.

Through our quantitative evaluation experiments, we learned that it is essential to understand the art of balancing the size of the enriched phenotype-phenotype associations for generating optimal HPO embeddings. One of the future studies would be to investigate how to filter out redundant noise from heterogeneous knowledge resources to extract phenotype-phenotype associations with high information gain. We also found HPOEmb-Orphanet did not outperform other conventional similarity measurements especially for patients with highly similar phenotypic characterizations. This may be because, after enriching HPO embedding with Orphanet, HPOEmb-Orphanet tends to find phenotypic similarity across multiple disease systems so that it is easier to detect inter-community patients (heterogeneous) rather than intra-community patients (homogeneous). Therefore, it is also important to tune the re-visitation parameter p and the in-out parameter q to switch between BFS and DFS based on different use cases. We plan to investigate further on how to better utilize HPOEmb-Orphanet for a more balanced capability to detect patient similarity.

With respect to the qualitative evaluation, for cluster visualization, the reason behind the frequent appearance of nodes with coarse granularity in HPOEmb-Original might be related to fewer non-inheritance edges amongst phenotypic nodes, which increases the chance of random walking through IS-A links. All the relationships between nodes maintained by the HPO-Original graph are represented as IS-A and different phenotypic nodes are only connected through subclass or superclass relationships [62]. Therefore, HPOEmb-Original only applied a biased random walk algorithm on IS-A relationships and the only way to find a neighbor is through the HPO hierarchical edges. Such a "vertical" searching strategy is more likely to include neighbor nodes with coarse granularity and thus infer that these nodes are relevant. For example, *nephrocalcinosis* is a subclass of *abnormal renal morphology* according to the HPO hierarchical structure and the cosine similarity provided by HPOEmb-Original is also high (0.99). However, it would be more interesting to discover two phenotypes as relevant if they contribute to the same diseases rather than an obvious hierarchical inheritance relationship (e.g., *nephrocalcinosi* and *decreased numbers of nephrons* inferred by HPOEmb-Orphanet), which will increase the chance to detect more phenotypic relevance.

The same influence was observed in the use case study. For example, 7 of 10 inferred phenotypes for *stage 5 chronic kidney diseases* were high level nodes with coarse granularity: *abnormality of the musculature of the lower limbs*, *abnormality of the phalanges of the 5th finger*, *abnormal macular morphology*, *abnormality of epiphysis morphology*, *abnormal tongue morphology*, *abnormality of phalanx of finger*, and *abnormality of muscle morphology*. By adding more non-inheritance links between nodes, HPOEmb-DECIPHER inferred 5 out of 10 phenotypic nodes with coarse granularity, among which *behavioral abnormality* and *abnormality of hair pigmentation* were two phenotypes that were not inferred by HPOEmb-Original. With many non-inheritance edges being enriched, HPOEmb-OMIM was able to infer 8 out of 10 leaf nodes with fine granularity, such as *malar flattening*, *hydronephrosis*, *progressive intellectual disability*, *inguinal hernia*, *gingival overgrowth*, *short foot*, *childhood onset* and *deeply set eye*. HPOEmb-Orphanet achieved the optimal performance in both quantitative and qualitative evaluations. It inferred 6 out of 10 leaf nodes with fine granularity, such as *secondary hyperparathyroidism*, *polyarticular arthritis*, *synovitis*, *cellulitis*, *gangrene*, and *isothenuria*, and 3 out of 6 were considered relevant for *stage 5 chronic kidney disease*.

According to the domain expert's feedback, HPOEmb-Orphanet achieved the highest P@5 and P@10, indicating that disease-phenotype associations in Orphanet best complement relationships in the HPO graph and enhance the clinical predictive power on detecting relevant phenotypes for embeddings. Although HPOEmb-DECIPHER and HPOEmb-OMIM were also able to enrich the HPO graph with new relationships and achieve better performance on quantitative link prediction task, they did not show much performance difference on inferring relevant phenotypes compared to HPOEmb-Original. Compared to the other three HPO embeddings, since the Orphanet annotates disease-phenotype associations based on comorbidities and signs/symptoms of rare diseases, it is easier to link different phenotypic characterizations from a clinical perspective. This shows that, distributed representations coming from clinical-oriented knowledge resources (e.g., Orphanet) can provide more evidence to assist in detecting similar phenotypic characterizations compared to biomedical-oriented knowledge resources (e.g., DECIPHER and OMIM). In addition, compared to HPOSim with seven conventional similarity measurements, since HPOEmb-Orphanet enriched the original HPO hierarchical relations with more clinical insights from Orphanet, most of the inferred relevant phenotypes are more than synonyms or in inheritance relationships, showing an increasing chance of detecting relevant phenotypes.

There are some limitations in our study. First, due to the limited size of annotations and a desire to maximize enrichment of the HPO, associations derived from different annotation files were solely based on co-occurrences between diseases and phenotypes. Although the annotations used were all manually curated, it may be important to limit to a smaller set of disease-phenotype associations. For example, there exist some studies utilizing association rule mining and odds ratio to further extract significant disease-phenotype associations from a large amount of EHR [39,57]. Secondly, we used an old version of the HPO as well as the phenotype annotation file released in May 2018 to construct the HPO embeddings. We used a subset of 7,258 nodes and their relationships in the old version to be consistent with our prior work. We plan to apply our generalized framework HPO2Vec+ on the latest heterogeneous knowledge resources to update the HPO embeddings and assess the impact of versioning for our experiments as even 7,258 nodes and their relationships remain the same in the latest HPO version, annotations for HPO do increase in the latest phenotype annotation file (accessed May 2019) with 609 and 761 more diseases for OMIM and Orphanet, respectively.

In the future, we would like to leverage the large volume of clinical notes available at Mayo Clinic and mine essential disease-phenotype associations from clinical narratives to further enrich connectivity among the HPO nodes and thus enhance embeddings. In addition, we will upgrade HPO2Vec+ to integrate normalized entity nodes from different biomedical ontologies into the HPO to increase the size of the graph and to allow for larger HPO-based heterogeneous embeddings to be generated and tested. Moreover, associations between phenotypes and common diseases will also be investigated, and phenotypic relevance measurements will be combined with our previously developed collaborative filtering rare disease recommendation system [63] to further transform phenotypic level analysis into patient similarity and disease similarity measurements. Regarding methodology, to generalize the 2-step random walk algorithm implemented by node2vec, we will seek to develop a k-step based biased random walk algorithm based on distances between any pair of nodes contained in the graph leveraging some previously developed distance-based knowledge discovery algorithms [64].

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgement

This work has been supported by the National Institute of Health (NIH) grants U01TR0062-1 and Rare Kidney Stone Consortium (U54DK083908). The Rare Kidney Stone Consortium (U54DK083908) is part of Rare Diseases Clinical Research Network (RDCRN), an initiative of the Office of Rare Diseases Research (ORDR), NCATS. This consortium is funded through collaboration between NCATS, and the National Institute of Diabetes and Digestive and Kidney Diseases.

References

[1] P.N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, S. Mundlos, The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease, Am. J. Hum. Genet. 83 (2008) 610–615.
[2] H.V. Firth, S.M. Richards, A.P. Bevan, S. Clayton, M. Corpas, D. Rajan, et al., DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources, Am. J. Hum. Genet. 84 (2009) 524–533.
[3] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucl. Acids Res. 33 (2005) D514–D517.
[4] S. Aymé, J. Schmidtke, Networking for rare diseases: a necessity for Europe, Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz. 50 (2007) 1477–1483.
[5] P.N. Robinson, Deep phenotyping for precision medicine, Hum. Mutat. 33 (2012) 777–780.
[6] J.H. Son, G. Xie, C. Yuan, L. Ena, Z. Li, A. Goldstein, et al., Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes, Am. J. Hum. Genet. 103 (2018) 58–73.
[7] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2016, pp. 855–864.
[8] K. Pearson, The problem of the random walk, Nature 72 (1905) 342.
[9] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.
[10] F. Shen, S. Liu, Y. Wang, L. Wang, A. Wen, A.H. Limper, et al., Constructing node embeddings for human phenotype ontology to assist phenotypic similarity, Measurement. 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W): IEEE, 2018, pp. 29–33.
[11] D. McMahon, Quantum computing explained, John Wiley & Sons, 2007.
[12] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007. 1995.
[13] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008. 1997.
[14] D. Lin, An information-theoretic definition of similarity, Citeseer, Icml, 1998, pp. 296–304.
[15] B. Li, J.Z. Wang, F.A. Feltus, J. Zhou, F. Luo, Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. arXiv preprint arXiv:10010958. 2010.
[16] A. Schlicker, F.S. Domingues, J. Rahnenführer, T. Lengauer, A new measure for functional similarity of gene products based on Gene Ontology, BMC Bioinf. 7 (2006) 302.
[17] C. Pesquita, D. Faria, H. Bastos, A. Falcao, F. Couto, Evaluating GO-based semantic similarity measures, Proc 10th Annual Bio-Ontologies Meeting, 2007, p. 38.
[18] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, C.-F. Chen, A new method to measure the semantic similarity of GO terms, Bioinformatics 23 (2007) 1274–1281.
[19] S. Köhler, M.H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C.E. Ott, et al., Clinical diagnostics in human genetics with semantic similarity searches in ontologies, Am. J. Hum. Genet. 85 (2009) 457–464.
[20] A.J. Masino, E.T. Dechene, M.C. Dulik, A. Wilkens, N.B. Spinner, I.D. Krantz, et al., Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology, BMC Bioinf. 15 (2014) 248.
[21] N.L. Washington, M.A. Haendel, C.J. Mungall, M. Ashburner, M. Westerfield, S.E. Lewis, Linking human diseases to animal models using ontology-based phenotype annotation, PLoS Biol. 7 (2009) e1000247.
[22] S. Mathur, D. Dinakarpandian, Finding disease similarity based on implicit semantic similarity, J. Biomed. Inform. 45 (2012) 363–371.
[23] D. Smedley, A. Oellrich, S. Köhler, B. Ruef, M. Westerfield, P. Robinson, et al., PhenoDigm: analyzing curated annotations to associate animal models with human diseases, Database 2013 (2013).
[24] R. Hoehndorf, P.N. Schofield, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery, Nucl. Acids Res. 39 (2011) e119-e.
[25] C.L. Smith, C.-A.W. Goldsmith, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information, Genome Biol. 6 (2005) R7.
[26] G. Schindelman, J.S. Fernandes, C.A. Bastiani, K. Yook, Sternberg PW. Worm, Phenotype Ontology: integrating phenotype data within and beyond the C. elegans community, BMC Bioinf. 12 (2011) 32.
[27] D. Sardana, S. Vasa, N. Vepachedu, J. Chen, R.C. Gudivada, B.J. Aronow, et al., PhenoHM: human–mouse comparative phenome–genome server, Nucl. Acids Res. 38 (2010) W165–W174.
[28] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Proceedings of the AMIA Symposium: American Medical Informatics Association; 2001. p. 17.
[29] J. Peng, H. Xue, Y. Shao, X. Shang, Y. Wang, J. Chen, A novel method to measure the semantic similarity of HPO terms, IJDMB. 17 (2017) 173–188.
[30] X. Gong, J. Jiang, Z. Duan, H. Lu, A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology, BMC Bioinf. 19 (2018) 162.
[31] Y. Deng, L. Gao, B. Wang, X. Guo, HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology, PLoS ONE 10 (2015) e0115692.
[32] T. Doğan, HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences, PeerJ 6 (2018) e5298.
[33] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, et al., Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25.
[34] R. Hoehndorf, M. Dumontier, G.V. Gkoutos, Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics, Bioinformatics 28 (2012) 2169–2175.
[35] T. Wittkop, E. TerAvest, U.S. Evani, K.M. Fleisch, A.E. Berman, C. Powell, et al., STOP using just GO: a multi-ontology hypothesis generation tool for high throughput experimentation, BMC Bioinf. 14 (2013) 53.
[36] J.D. Osborne, J. Flatow, M. Holko, S.M. Lin, W.A. Kibbe, L.J. Zhu, et al., Annotating the human genome with Disease Ontology, BMC Genomics 10 (2009) S6.
[37] M.R. Dwinell, E.A. Worthey, M. Shimoyama, B. Bakir-Gungor, J. DePons, S. Lauledernkind, et al., The Rat Genome Database 2009: variation, ontologies and pathways, Nucl. Acids Res. 37 (2008) D744–D749.
[38] R. Agarwal, R. Srikant, Fast algorithms for mining association rules. Proc of the 20th VLDB, Conference (1994) 487–499.
[39] F. Shen, Y. Zhao, L. Wang, M.R. Mojarad, Y. Wang, S. Liu, et al., Rare disease knowledge enrichment through a data-driven approach, BMC Med. Inf. Decis. Making 19 (2019) 32.
[40] HPO Disease Annotation Repository. Available at: http://compbio.charite.de/jenkins/job/hpo.annotations/lastStableBuild/. Accessed in Feb 2019.
[41] L. Tang, H. Liu, Leveraging social media networks for classification, Data Min. Knowl. Disc. 23 (2011) 447–478.
[42] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[43] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, et al., Rolx: structural role extraction & mining in large graphs, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 1231–1239.
[44] J. Yang, J. Leskovec, Overlapping Communities Explain Core-Periphery Organization of Networks, Proc. IEEE 102 (2014) 1892–1902.
[45] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 701–710.
[46] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, Line: Large-scale information network embedding, Proceedings of the 24th international conference on world wide web: International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.
[47] H. Cai, V.W. Zheng, K.C.-C. Chang, A comprehensive survey of graph embedding: Problems, techniques, and applications, IEEE Trans. Knowledge Data Eng. 30 (2018) 1616–1637.
[48] C. Davis, The norm of the Schur product operation, Numer. Math. 4 (1962) 343–344.
[49] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc.: Ser. B (Methodol.) (1996) 267–288.
[50] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[51] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[52] S.H. Walker, D.B. Duncan, Estimation of the probability of an event as a function of several independent variables, Biometrika 54 (1967) 167–179.
[53] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[54] T.K. Ho, Random decision forests. Document analysis, and recognition, 1995, proceedings of the third international conference on: IEEE, 1995, pp. 278–282.
[55] I. Rish, An empirical study of the naive Bayes classifier, IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM, New York, 2001, pp. 41–46.
[56] F. Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, Cornell Aeronautical Lab Inc Buffalo NY (1961).
[57] F. Shen, L. Wang, H. Liu, Phenotypic analysis of clinical narratives using human phenotype ontology, Stud. Health Technol. Informat. 245 (2017) 581–585.
[58] Maaten Lvd, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[59] V.O. Edvardsson, D.S. Goldfarb, J.C. Lieske, L. Beara-Lasic, F. Anglani, D.S. Milliner, et al., Hereditary causes of kidney stones and chronic kidney disease, Pediatric Nephrol. 28 (2013) 1923–1942.
[60] D.L. Raju, M. Cantarovich, M.-L. Brisson, J. Tchervenkov, M.L. Lipman, Primary hyperoxaluria: Clinical course, diagnosis, and treatment after kidney failure, Am. J. Kidney Dis. 51 (2008) e1–e5.
[61] R. Hajar, Diabetes as “coronary artery disease risk equivalent”: A historical perspective, Heart views: Off. J. Gulf Heart Assoc. 18 (2017) 34.
[62] Introduction to Human Phenotype Ontology. Available at: https://hpo.jax.org/app/help/introduction. Accessed in Feb 2019.
[63] F. Shen, S. Liu, Y. Wang, L. Wang, N. Afzal, H. Liu, Leveraging collaborative filtering to accelerate rare disease diagnosis, AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2017, p. 1554.
[64] F. Shen, Y. Lee, Knowledge discovery from biomedical ontologies in cross domains, PLoS ONE 11 (2016) e0160005.