Bioinformatics, 37(19), 2021, 3311–3318
doi: 10.1093/bioinformatics/btab340
Advance Access Publication Date: 8 May 2021
Original Paper

Systems biology

CLEP: a hybrid data- and knowledge-driven framework for generating patient representations

Vinay Srinivas Bharadhwaj[1,2], Mehdi Ali [3,4], Colin Birkenbihl[1,2], Sarah Mubeen[1,2,5], Jens Lehmann[3,4], Martin Hofmann-Apitius[1,2], Charles Tapley Hoyt [1,3,5] and Daniel Domingo-Fernández [1,3,5,*]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, 53757 Sankt Augustin, Germany, [2]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53115 Bonn, Germany, [3]Rheinische Friedrich-Wilhelms-Universität Bonn, 53113 Bonn, Germany, [4]Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Dresden and Sankt Augustin, Germany and [5]Fraunhofer Center for Machine Learning, Bonn, Germany

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol
Received on November 5, 2020; revised on March 29, 2021; editorial decision on April 29, 2021; accepted on May 3, 2021

Abstract

Summary: As machine learning and artificial intelligence increasingly attain a larger number of applications in the biomedical domain, at their core, their utility depends on the data used to train them. Due to the complexity and high dimensionality of biomedical data, there is a need for approaches that combine prior knowledge around known biological interactions with patient data. Here, we present CLinical Embedding of Patients (CLEP), a novel approach that generates new patient representations by leveraging both prior knowledge and patient-level data. First, given a patient-level dataset and a knowledge graph containing relations across features that can be mapped to the dataset, CLEP incorporates patients into the knowledge graph as new nodes connected to their most characteristic features. Next, CLEP employs knowledge graph embedding models to generate new patient representations that can ultimately be used for a variety of downstream tasks, ranging from clustering to classification. We demonstrate how using new patient representations generated by CLEP significantly improves performance in classifying between patients and healthy controls for a variety of machine learning models, as compared to the use of the original transcriptomics data. Furthermore, we also show how incorporating patients into a knowledge graph can foster the interpretation and identification of biological features characteristic of a specific disease or patient subgroup. Finally, we released CLEP as an open source Python package together with examples and documentation.

Availability and implementation: CLEP is available to the bioinformatics community as an open source Python package at https://github.com/hybrid-kg/clep under the Apache 2.0 License.

Contact: daniel.domingo.fernandez@scai.fraunhofer.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent advancements in machine learning (ML) and artificial intelligence (AI) methodologies have initiated a paradigm shift in bioinformatics. As new technologies have steadily generated large volumes of -omics data, AI methods have garnered great insights into human health and biology. With the availability of large-scale biological datasets, ML/AI are becoming highly relevant for biomedical applications, such as predictive modeling, patient stratification and simulation (Fröhlich *et al.*, 2018). However, despite the successful application of ML/AI in the biomedical domain, the datasets underlying the generation of models can play a far more crucial role in a given application than the complexity of the model itself. For example, in some cases, if data is predictive enough, simpler methods can outperform state-of-the-art ML/AI methods in prediction tasks (Lynam *et al.*, 2020; Smith *et al.*, 2020).

The development of novel high-throughput experimental techniques has led to a broad availability of biological data from multiple entity types (Zitnik *et al.*, 2019). In practice, integrating multiple data types can be advantageous, particularly in the context of complex diseases, where no single type of data can effectively explain the cause of dysfunction. However, biological datasets tend to be both inherently complex and noisy (Fan *et al.*, 2014), making their integration challenging. Furthermore, biological data typically

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023

contain a far greater number of features than samples due to several factors, including a lack of available resources and obstacles in sample collection (Xu and Jackson, 2019). In failing to address these challenges and generating comprehensive representations of biological data, novel techniques can suffer in a range of analytic tasks.

One approach adopted by the systems biology community is in representing biological data in the form of networks. By doing so, multiple scales of biology can be represented, as can the relationships within and across these scales. These networks are also advantaged by their ability to integrate heterogeneous biological data types (Hu *et al.*, 2017). Generally, one can classify biological networks into two categories depending on the source of information used to generate them (Yu *et al.*, 2013).

The first of these two classes of networks are constructed from biological data by using a variety of methodologies. For instance, co-expression networks can be generated from transcriptomics data to represent pairwise correlations between genes (Langfelder and Horvath, 2008). Another example is Bayesian networks which can be used to model conditional interdependencies across heterogeneous biological entities (Khanna *et al.*, 2018). While the majority of these methods transform biological data into networks comprising relations between the biological entities under study, other methods directly translate patient-level data into networks through correlation techniques (e.g. Pearson correlation) as an indicator of similarity between patients. These patient similarity networks closely link patients that are more similar to each other while less similar patients contain fewer close connections. Patient similarity networks have been successfully used to represent multimodal patient level-data for clustering tasks (Cavalli *et al.*, 2017; Pai *et al.*, 2019; Pai and Bader, 2018; Raphael *et al.*, 2017; Wang *et al.*, 2014a). While so far, these methods have been mostly employed for clustering tasks based on patient similarity networks, (Pai *et al.*, 2019) recently expanded the concept to incorporate gene-level measurements into pathways (via gene sets) in order to reduce dimensionality and ultimately use pathway features for patient classification.

A second class of networks can be generated from prior knowledge of known interactions between biological entities. When sets of these interactions are assembled, they can be used to represent discrete biological networks. These networks can be referred to as knowledge graphs (KGs) when they comprise entities from various biological modalities and the complex interactions between them. However, despite their advantages, networks cannot be directly represented in vector space; thus, impeding their direct use by ML/AI techniques. This impediment has led to the development of methodologies (i.e. Knowledge Graph Embedding Models; KGEMs) designed to encode entities and relations in a KG into a latent feature space while preserving its structure. While these new representations can be used for entity disambiguation, clustering and several downstream ML/AI tasks, they have been primarily used in the biomedical domain for link prediction tasks, including the prediction of side effects (Zitnik *et al.*, 2018), disease-gene associations (Himmelstein and Baranzini, 2015) and novel therapeutic targets (Muslu *et al.*, 2020). Recently, the notion of adding patients into KGs has been proposed from Electronic Health Records (EHRs) (i.e. MIMIC-III dataset) with the ultimate goal of generating patient embeddings through KGEMs that can be used for clustering (Gong *et al.*, 2021) and medicine recommendation tasks (Lin *et al.*, 2020). However, until now, there have yet to be integrative approaches which incorporate both patient-similarity networks and KGs from non-textual patient-level data, such as multi-omics, and prior knowledge.

In this work, we introduce CLinical Embedding of Patients (CLEP), a hybrid data- and knowledge-driven framework that exploits -omics patient-level data and incorporates this information into a KG. Once the KG has been generated by CLEP, the framework then drives the generation of novel patient representations through various KGEMs. In building upon previous data-driven approaches (Pai *et al.*, 2019; Wang *et al.*, 2014a) which have demonstrated the advantages of representing patients as nodes in a network, we show that additionally integrating prior knowledge can make for more robusts analyses. We showcase our approach by generating new patient representations derived from two different transcriptomics datasets and a KG comprising heterogenous protein–protein interactions from multiple biological databases. Using the new patient representations, we find that performance on a panel of ML models trained to classify between patients and controls is significantly improved with respect to the use of original data. Furthermore, the flexibility of our approach makes it applicable to heterogenous multimodal biological datasets, enabling researchers to generate new patient representations by combining patient-level data with the prior knowledge contained in a KG. Finally, we have made CLEP available to the bioinformatics community as an open source Python package (https://github.com/hybrid-kg/clep) under the Apache 2.0 License.

2 Materials and methods

2.1 Framework description

Figure 1 illustrates each step of the presented framework. The methodology requires a patient-level dataset and a KG. Patients are incorporated into the KG as new nodes and connected to features that most closely characterize a given patient. Once patients are embedded into the KG, KGEMs are then used to generate new patient representations, learnt from both prior knowledge underlying the KG and the patient-feature connections that have been generated. Finally, these novel patient representations can subsequently be employed for a variety of downstream applications, as we demonstrate in the case scenarios section.

2.1.1 Input data

Our framework requires two inputs: a patient-level dataset and a KG (Fig. 1a). It can be applied to any dataset and KG so long as the dataset features can be mapped to nodes in the KG. In other words, if we intend to use CLEP on a transcriptomics dataset such as RNA-Seq, the KG must contain relationships between genes/proteins that are mappable to the transcripts that have been measured. We would like to note that although the framework does not require that all features in the dataset are also present in the KG, it is recommended to maximize this overlap.

2.1.2 Incorporating patients into the knowledge graph

The first step of the methodology consists of incorporating patients as nodes in the KG in order to subsequently generate novel individualized feature representations for each patient (i.e. embeddings) through the use of KGEMs (Fig. 1b). These models generate the embeddings by exploiting the topology of the KG. Generating comprehensive embeddings requires an approach that can position a patient node in the KG according to that patient's most informative features (i.e. features that distinguish them from patients who possess clinically different features). This way, patient nodes with similar features would be close together in a network and would thus have similar embeddings, while patients with dissimilar features would be farther away (see example in Fig. 4b). On the other hand, connecting every patient to a large number of nodes in the KG instead of a smaller node set that represents a patient's most relevant features, would result in poor patient representations due to large overlaps of irrelevant features. The positioning of the patient nodes in the KG can be seen as a feature selection technique in which our method identifies the most characteristic features of each patient in order to generate edges between the patient and these features, thus, embedding the patient as a node in the KG. The rationale behind this method is that by connecting patients with their most characteristic features, we are able to generate a KG that integrates patient-specific information (i.e. relations between patients and features) together with prior knowledge (i.e. relations among these features).

To identify the most characteristic features and incorporate patients into the KG, our methodology leverages the empirical cumulative distribution function (eCDF) based on the quantitative measurements for each feature in the dataset which can be mapped to the KG. Using all samples in the dataset or exclusively the healthy

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023

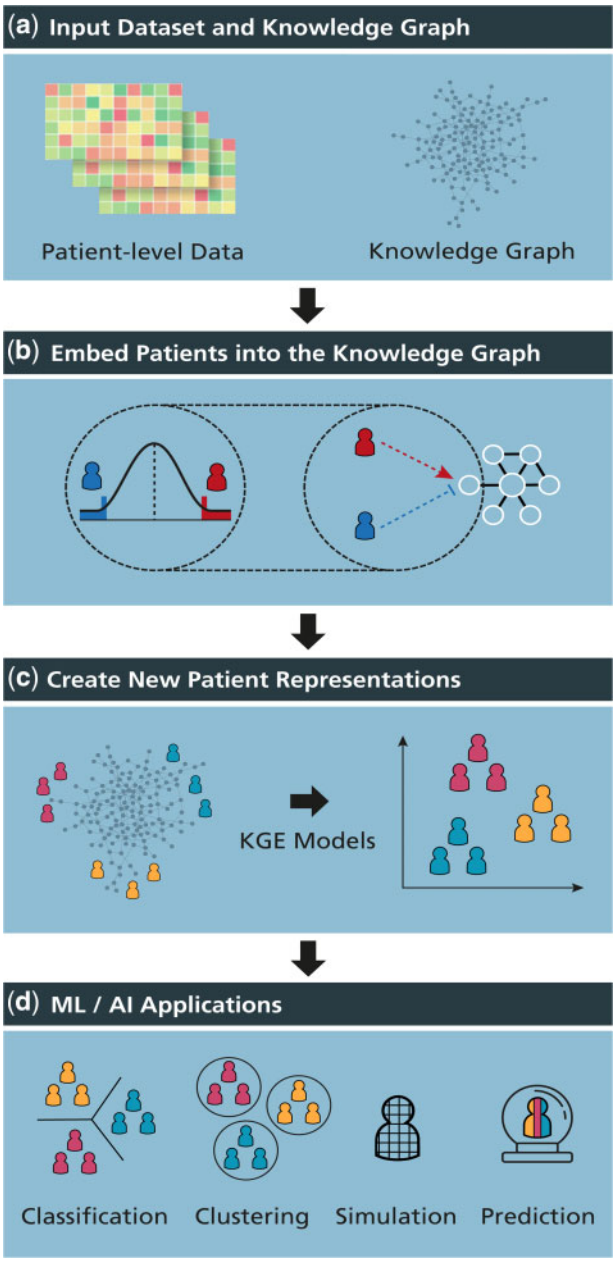

Fig. 1. Schematic illustration of the framework. (**a**) CLEP requires two inputs: (i) a patient-level dataset such as multi-omics, and (ii) a KG comprising relations between features measured in the previously mentioned dataset. (**b**) Using one of the proposed methods, CLEP incorporates patients into the KG by connecting them to their most distinctive features in the dataset. (**c**) KGEMs are then used to generate new patient representations based on both data- and knowledge-driven features. (**d**) These patient representations can subsequently be used for several downstream tasks, such as patient classification and stratification. A high quality version of this figure is available at https://doi.org/10.6084/m9.figshare.12834605

controls data, we generate a reference distribution with data points that represent the measurements of each feature. We then identify patients that fall at the extreme ends of this distribution based on a predetermined threshold (e.g. 5% of the eCDF) (Fig. 1b). Finally, these patients are connected to the node in the KG that represents that particular feature with an edge (i.e. -1 or +1) depending on the extreme in which the patient falls. The sign of the edge indicates whether a particular patient has a higher or lower measurement for that given feature to differentiate the patient from other patients in the distribution.

In the Supplementary Text, we present alternative methodologies that can similarly be used to incorporate patients to the KG. However, we have focused here on one specific method as it can be applied to any type of patient-level data and it does not make any assumptions on the feature distribution as opposed to other methods which either require pathway information or assume that the features are normally distributed.

2.1.3 Generating new patient representations

Once patients have been incorporated into the KG, the next step involves generating new patient representations through KGEMs (Fig. 1c). We restricted ourselves to KGEMs since they consider both directionality and edge types, and our KG contains several directed edge types. Incorporating edge types during the learning process can help the model to differentiate between node types (e.g. patients and biological entities) and node sub-types (e.g. diseased patients, and controls). CLEP has adopted PyKEEN (Ali et al., 2021) as the KGEM-software due to its wide range of functionalities (e.g. a large number of KGEMs, hyperparameter optimization functionalities).

2.1.4 Applications of patient representations

Once novel patient representations have been generated with CLEP, they can be used for a vast number of applications including classification and clustering tasks (Fig. 1d). Because our approach leverages prior knowledge of known interactions, we hypothesize that the use of these representations can yield superior performances on these tasks with respect to the use of the original patient-level datasets from which they were generated. Furthermore, our methodology facilitates biologically meaningful interpretations of patient-level data by positioning patients into different KG neighborhoods which may potentially correspond to biological processes that are characteristic for specific patient subgroups.

2.2 Software implementation

CLEP is implemented as a Python package to facilitate its usage within the scientific community. It contains several workflows corresponding to each of the steps presented in the methods from generating new patient representations to conducting downstream applications presented in the case scenario. Each workflow is both accessible through a command line interface (CLI) as well as programmatically, allowing users to input their own patient-level datasets and custom KGs. In total, CLEP offers three different methods for incorporating patients into the KG, all KGEMs available through PyKEEN (Ali *et al.*, 2021), and five ML classifiers. Furthermore, thanks to its flexible implementation, users can independently use each of its modules as well as incorporate classifiers tasks into the framework (Supplementary Fig. S2). Finally, the source code of the CLEP Python package is available at https://github.com/hybrid-kg/clep under the Apache 2.0 License, its latest documentation can be found at https://clep.readthedocs.io, and it is distributed via PyPI at https://pypi.org/project/clep.

2.3 Case scenarios

2.3.1 Patient-level data

The first dataset, the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller *et al.*, 2005), is one of the world's largest dementia cohorts and, with more than 1300 citations, the most referential resource for data-driven dementia research. In this work, we used the blood plasma transcriptomic data collected in the study [we refer to Saykin *et al.* (2015) and http://adni.loni.usc.edu for details]. The dataset is already preprocessed and contains a total of 260 cognitively healthy control participants, 215 patients with early mild cognitive impairment, 225 patients with late mild cognitive impairment and 44 patients with Alzheimer's disease. To conduct the binary classification task, the latter three (i.e. all cognitively impaired patients) were grouped together into a single class ($n = 494$). Preprocessed gene expression data [Robust Multichip Average (RMA) normalized data] was directly used as a baseline for the

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023

benchmarking of CLEP. To incorporate the patients into the KG, the expression of genes whose transcripts appear multiple times were considered individually.

The second dataset is a transcriptomics dataset containing samples from three psychiatric disorders (i.e. major depressive disorder, schizophrenia and bipolar disorder) and healthy controls (Hagenauer *et al.*, 2018). In total, this dataset contains 172 samples (41 major depressive disorder, 22 schizophrenia, 26 bipolar disorder and 83 control samples). Similar to the ADNI dataset, the preprocessed gene expression data (RMA normalized data) from the second dataset was directly used as a baseline for the validation of CLEP. Similar to the previous dataset, we grouped the three psychiatric indications together to conduct a binary classification task.

2.3.2 Knowledge graph

For the case scenario, we employ a KG referred to as PPI-KG, which comprises protein–protein interactions from six resources: PathMe (Domingo-Fernández *et al.*, 2019) [which includes KEGG (Kanehisa *et al.*, 2017) Reactome (Jassal *et al.*, 2020) and WikiPathways (Slenter *et al.*, 2018)], BioGrid (Oughtred *et al.*, 2019), IntAct (Orchard *et al.*, 2014) and Pathway Commons (Rodchenkov *et al.*, 2020) (Supplementary Text).

2.3.3 Generating representations of patients

We used different thresholds (i.e. the 1%, 1.5%, 2.5%, 5% 10% and 20% quantiles of the eCDF) to define the tail of the distribution (i.e. extreme) which is required to incorporate ADNI patients and the patients of the GSE92538 dataset to the PPI-KG using the eCDF-based method described in Section 4.1. By applying this method on each threshold, we generated different KGs (i.e. one for each threshold) by connecting the patients that fall in the extremes of the reference distributions (Supplementary Table S4). Reference distributions were generated based on the expression values of healthy controls for each gene that was present in both the transcriptomics data and the PPI-KG (Supplementary Fig. S3). The vast majority of nodes in the PPI-KG were among the proteins measured in the transcriptomics dataset, resulting in 8085 mapped proteins in total.

We would like to note that the resulting KG (i.e. PPI-KG after incorporating the ADNI patients) must be split into three triple sets (i.e. training, validation and testing set) in order to train a KGEM. Thus, we developed an algorithm that splits a given KG in a way that relations and nodes are balanced across the three different splits. The pseudocode for this algorithm can be found in the Supplementary Figure S1 and is implemented in the CLEP Python package.

2.3.4 Selected knowledge graph embedding models

We selected RotatE (Sun *et al.*, 2019), TransE (Bordes *et al.*, 2013), ComplEx (Trouillon *et al.*, 2016) and HolE (Nickel *et al.*, 2016) to learn the patient embeddings, because they have been shown to be effective in a large-scale benchmarking study for KGEMs (Ali *et al.*, 2020). To train the final models, we used the best hyperparameters obtained by the hyperparameter optimization (Supplementary Table S3).

2.3.5 Classifying between cognitively impaired and healthy controls

The new representations generated from KGEMs were used to classify between normal and cognitively impaired patients (i.e. AD and MCI) using five different statistical modeling and ML methods (Table 1).

Prediction performance was evaluated via 5 times repeated 5-fold cross-validation in which the hyperparameters of the model were tuned within the cross-validation loop via a grid search (Fig. 2). The cross-validation prediction performance for the binary classification task was evaluated using the area under the receiver operating characteristic curve (AUC-ROC) or the area under the precision–recall curve (AUC-PR) as a metric.

We then conducted three experiments to evaluate the robustness of our results. First, we trained the ML models using the new representations generated from RotatE with the 5% threshold while permuting patient labels (i.e. y-scrambling). Second, we trained the ML models on patient representations again generated using RotatE with a 5% threshold but using instead a permuted version of the KG generated using the XSwap algorithm (Hanhijärvi *et al.*, 2009) in order to investigate if a KG's topology makes meaningful improvements to performance. By using this algorithm, we ensured that the permuted versions preserved the original structure of the original network (i.e. each node has the same number of in- and out-edges) but edges were randomly generated. Third, we trained the ML models using a subset of the PPI-KG corresponding to a single database (i.e. either WikiPathways or KEGG) in order to investigate the importance of KG size and completeness.

2.3.6 Classifying between psychiatric disorders and healthy controls

Using the same setting used in the previous dataset (Fig. 2), we trained the same five ML models to classify between normal samples and patients with a psychiatric disorder (i.e. bipolar disorder, major depressive disorder and schizophrenia) using both the raw data and the new representations of the GSE92538 data generated by CLEP. Similar to the previous dataset, we used RotatE KGEM and a threshold value of 2.5% on either side of the control distribution for generating the new representations.

3 Results

3.1 CLEP's representations outperform raw data in classifying cognitively impaired patients and healthy controls

Here, we present the results of our methodology and demonstrate how CLEP can improve the performance of ML models on patient classification tasks within the context of Alzheimer's disease (AD). We incorporated ADNI patients (Mueller *et al.*, 2005) into a protein–protein interaction KG (i.e. PPI-KG) in order to generate novel patient representations using various KGEMs (see Section 2). We then compared the performance of several ML models in distinguishing between cognitively impaired patients from controls depending on whether the input data was the original transcriptomics data or novel patient representations. We summarize the performance of each of the five ML models in Figure 3a and b.

Our results show that using original transcriptomics data as input for the binary classifier leads to relatively low prediction power (Fig. 3b), as opposed to the new representations generated by CLEP which substantially increase prediction performance (Fig. 3a). As an illustration, the SVM model, which was the highest performing, yielded area under the receiver operator characteristic curve (AUC-ROC) values ranging from 0.48 to 0.64 when trained on the transcriptomics data. In comparison, the same model, when trained on the new patient representations, yielded AUC-ROC values between 0.83 and 0.91. This difference was consistent across each of the five ML models, demonstrating how CLEP generates patient representations that significantly outperform the original data for this particular binary classification task. However, it is worth noting that the performance of the two tree-based methods (i.e. random forest and XGBoost) is significantly worse than the remaining models, yet they still outperformed their corresponding counterpart models trained on the raw data. Furthermore, to validate that the results are not an effect of the class imbalance between cognitively impaired patients and healthy controls (Saito and Rehmsmeier, 2015), we reevaluated the ML models using the area under the precision–recall curve (AUC-PR) as a metric, which resulted in similar results (Supplementary Fig. S6). Finally, we would like to note that through the usage of the KGEMs, the dimension of the patient representations could significantly be compressed since the input dimension of the raw transcriptomics was larger than 40 000 features, while the representations generated by the KGEMs had a dimension of 256.

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023

Table 1. List of statistical modeling and machine learning methods available at CLEP for conducting classification tasks

Model	Reference	Implementation
Logistic regression with L2 regularization	–	scikit-learn (Pedregosa *et al.*, 2011)
Logistic regression with elastic net regularization	Zou and Hastie (2005)	scikit-learn (Pedregosa *et al.*, 2011)
Support Vector Machines	Cortes and Vapnik (1995)	scikit-learn (Pedregosa *et al.*, 2011)
Random Forest	Ho (1995)	scikit-learn (Pedregosa *et al.*, 2011)
XGBoost	Chen and Guestrin (2016)	XGBoost (Chen and Guestrin, 2016)

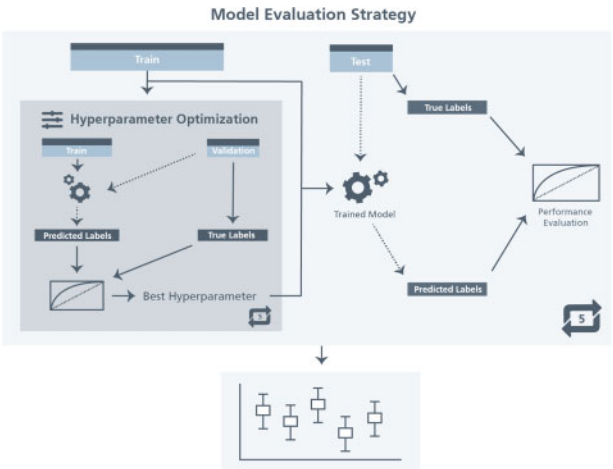

Fig. 2. Schematic representation of the model evaluation strategy. The first step is to split the data into training (80%) and hold-out test set (20%). Next, we performed 5-fold stratified cross-validation by further repeatedly splitting the training data into 80% training and 20% validation in order to identify the best hyperparameter settings. Once these five cross-validation rounds were performed, the best hyperparameters were used to train the model on the full training data and the trained model was evaluated on the held-out test set. This entire strategy was repeated 5 times to generate 5 AUC-ROC scores that were used to determine the performance of each of the five models (Table 1). We would like to note that the first split (training-test) ensured maintaining the overall class distribution and classes have been shuffled beforehand to avoid having the same data used in training/test. Furthermore, we chose AUC-ROC scores as evaluation metrics for both best hyperparameter selection and model evaluation. A high quality version of this figure is available at https://doi.org/10.6084/m9.figshare.12834608

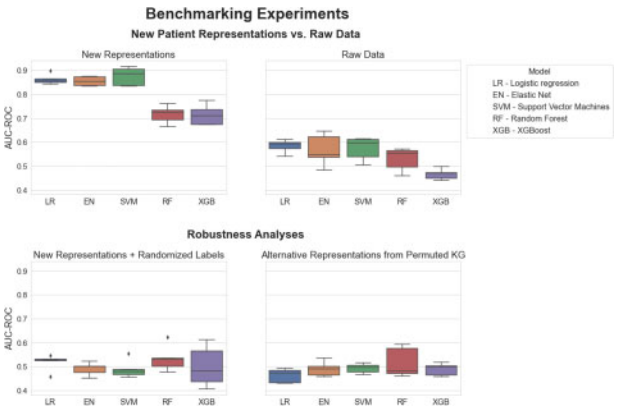

Fig. 3. Benchmarking of five ML models trained to classify between cognitively impared patients and healthy controls. Each boxplot shows the distribution of the AUC-ROC values over 5 repeats of the 5-fold nested cross-validation procedure. Statistical modeling and ML methods are listed in Table 1. The new patient representations were generated by incorporating ADNI patients into the PPI-KG using a threshold of 2.5% on the eCDF of the control distribution for each mapped feature. The RotatE KGEM was trained on the KG using PyKEEN. The patient representations were used to train the five ML models with the original patient labels (**a**) and compared against the raw transcriptomics data (**b**). To investigate the robustness of our results, we randomized patient labels and trained the five ML models using the new representations (**c**). Furthermore, we trained the five ML models using the new representations generated from a permuted version of the original KG (**d**). Both robustness analyses yielded AUC-ROC values on all machine learning models equivalent to that of a random classifier (i.e. AUC-ROC values ∼0.5)

To investigate the robustness of our approach, we conducted two independent experiments. The first experiment consisted of training each of the five classifiers using new representations with randomized patient labels (Fig. 3c) in order to confirm that the models did not fit to arbitrary artifacts in the data. In contrast, in the second experiment, patient representations were generated on a permuted version of the original KG (see Section 2 for details) to ensure that new representations reflect the information encoded in the KG by generating patient representations on a permuted version of the original KG (Fig. 3d). The results of these experiments yielded models with a performance equivalent to a random classifier (i.e. AUC-ROC values ∼ 0.5); thus, confirming (i) the robustness of our model evaluation strategy and (ii) that the new representations are driven by information in the KG.

While we were able to show how CLEP successfully generated novel representations that yield superior prediction performance, the process of generating these representations is non-trivial. As an initial step, a threshold that determines which patient-measurable edges will be incorporated into the KG must first be selected as this parameter influences the KG generated (see Section 4.1). We demonstrate the effect of the threshold on the previous binary classification task by training a single classifier using KGs derived from different thresholds (Supplementary Fig. S4). Unsurprisingly, while lower threshold settings (i.e. from 1% to 5%) resulted in increased predictive power, higher thresholds (i.e. 10% and 20%) penalized the performance of the model. This can be attributed to the fact that a low

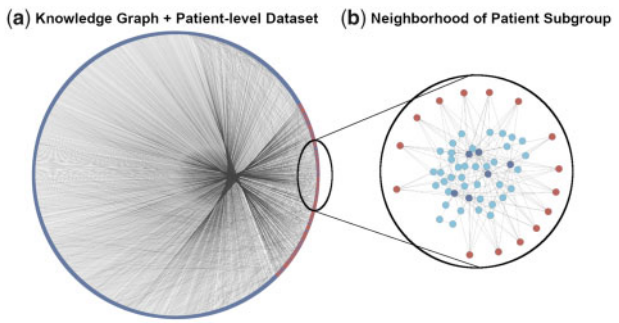

Fig. 4. Incorporating patients into a KG fosters biological interpretation and identification of patient subgroups. (**a**) PPI-KG subgraph after incorporating ADNI participants visualized using a circular layout. Red nodes represent ADNI patients and blue nodes represent proteins present in the PPI-KG. The majority of connections are between ADNI patients and their most characteristic proteins, which is represented by the large number of outgoing edges between the right part of the circle (where patients are located) and the rest of the KG. (**b**) Local neighborhood around the subgroup of cognitively impaired patients investigated in the case scenario. Red nodes represent patients, dark blue the six proteins linked to the patient subgroup and light blue other related proteins present in the neighborhood

threshold exclusively creates connections between KG nodes and the patients falling at the extreme ends of a feature distribution, thus capturing patients that can best characterize a particular feature. On the other hand, higher thresholds generate a larger number of edges

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023

between patients and features, resulting both in a loss of specificity and the ability to distinguish between different patient groups.

The right choice of the KGEM is essential for generating patient representations that are useful for the classifiers. While the results of the case scenario are based on RotatE, we also generated patient representations based on further KGEMs (Supplementary Fig. S5). We can observe that patient representations generated by RotatE outperforms the remaining KGEMs by a large margin. It is known that specific relational patterns that can be modeled by RotatE cannot be modeled by TransE (Bordes *et al.*, 2013), ComplEx (Trouillon *et al.*, 2016) and HolE (Nickel *et al.*, 2016). However, to obtain a clearer picture, relational patterns around patient nodes could be investigated. Finally, we also investigated the effect the size of a KG can have on results by evaluating the performance of ML models on smaller versions of the PPI-KG. As expected, smaller PPI-KGs resulted in lower performance for each of the ML models (Supplementary Fig. S7).

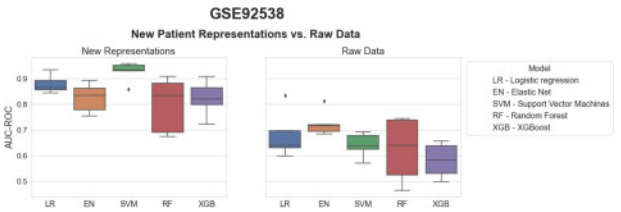

Fig. 5. Benchmarking of five ML models trained to classify between psychiatric patients and healthy controls. Each boxplot shows the distribution of the AUC-ROC values over 5 repeats of the 5-fold nested cross-validation procedure. Statistical modeling and ML methods are listed in Table 1. The new patient representations were generated by incorporating the patient data from GSE92538 into the PPI-KG using a threshold of 2.5% on the eCDF of the control distribution for each mapped feature. The RotatE KGEM was trained on the KG using PyKEEN. The patient representations were used to train the five ML models with the original patient labels (**a**) and compared against the raw transcriptomics data (**b**)

3.2 Biological interpretation and patient subgroup identification through CLEP

Though the advent of machine learning methods have led to a new range of applications in the biomedical field, these methods inherently come with a tradeoff with regards to interpretability. The so-called black-box models lack transparency, providing no explanation as to what accounts for the predictions they generate. In the biomedical field in particular, this can often translate to a lack of success in clinical practice (Fröhlich *et al.*, 2018); while discrete patterns that arise from ML techniques can be discerned, without an understanding of the fundamental cause and the characteristic features of a disease, clinicians may not have an adequate level of knowledge to make a confident diagnosis. In this section, we demonstrate how the hybrid KG generated by CLEP can drive the identification of patient-specific mechanisms and pathways made possible by the incorporation of patients into a KG comprising mechanistic knowledge (Fig. 4a).

In the previous section, we demonstrated how novel representations generated on the ADNI dataset outperformed raw data in a diagnosis prediction task. However, these representations cannot be directly interpreted as they are low-dimensional vector representations embedded in a latent space. Nonetheless, because the original KG comprises biological knowledge, features and/or mechanisms that are connected to a given embedded patient or patient subgroup can be easily pinpointed by studying their local neighborhood in the KG. Thus, using the KG derived from the ADNI dataset, we identified sets of genes with connections to cognitively impaired patients (groups) but without any connections to control participants. Of these gene sets, we focused on a particular set of genes (HS2ST1, ESR2, IKBKG, UBE2D3, PCGF5 and NFIC), all of which were connected to each other, and were also identified in a subgroup of cognitively impaired patients ($n = 15$). We then investigated the interplay of genes in the local neighborhood of the KG around this subset of patients in order to identify and deconvolute common pathways that could be responsible for their phenotypic make-up (Fig. 4b).

To identify the biological pathways these genes participate in, pathway enrichment analysis was run on this gene set (Supplementary Table S2). Of the nine enriched pathways we identified (q-value < 0.05), six were related to Toll-like receptor signaling (specifically, Toll-like receptor 4 signaling) due to the involvement of IKBKG and UBE2D3 in this pathway. Interestingly, this inflammatory pathway is often noted for its association to AD (Tahara *et al.*, 2006; Walter *et al.*, 2007), while UBE2D3 has been proposed as a potential biomarker for Alzheimer's disease (Wu et al., 2019). Furthermore, HS2ST1 is a member of the heparan sulfate biosynthetic enzyme family and responsible for the synthesis of heparan sulfate (HS), the latter of which is known to cause protein aggregation and lead to neurodegenerative disease (Maïza *et al.*, 2018). Additionally, ESR2 polymorphisms have been linked to cognitive impairment and an increased risk for AD, predominantly in women (Ulhaq and Garcia, 2020; Zhao *et al.*, 2015). It is worth noting that by investigating patient subgroups characterized by particular features, we can also assess whether they share common alleles. Finally, the remaining two genes, HFIC and PCGF5, are both present in the enriched 'Gene expression (transcription)' pathway, and are associated with the regulation of transcriptional factors.

3.3 CLEP's representations outperform raw data in classifying psychiatric conditions and healthy controls

We also reproduced our methodology on an additional dataset containing transcriptomics experiments on psychiatric conditions and healthy controls (see Section 2). The results for this dataset resembled the ones observed in the previous case scenario (Fig. 5). The patient representations generated by CLEP outperformed the original data in the binary classification task by a large margin. By investigating the individual ML models, we observed that the difference in performance between the two tree-based methods (i.e. random forest and XGBoost) and the rest was smaller for this dataset. Similarly to the previous case scenario, SVM yielded the highest performance. However, the variability in performance for the elastic net, random forest, and XGBoost was significantly larger compared to the ADNI dataset.

4 Discussion

In this work, we have presented CLEP, a novel hybrid data- and knowledge-driven framework which leverages patient-level data and KGs for generating personalized patient representations. In the case scenarios, we demonstrated the utility of our framework on two independent datasets by employing transcriptomics data and an integrative KG containing knowledge from several protein–protein interaction databases. When compared to the raw transcriptomics data, we have shown how these representations yield superior performance in a binary classification task using a broad panel of machine-learning models. Furthermore, we have illustrated how incorporating knowledge from the KG for the generation of patient representations not only improves performance in classification tasks, but also facilitates the interpretation of the biological mechanisms that uniquely characterize patients and/or patient subgroups. In summary, we have shown the utility of a hybrid approach for the generation of new patient representations that can then be used in a broad range of applications, ranging from predictive modeling to personalized medicine. We have also made CLEP available as an extensible and reproducible software package, enabling researchers to conduct these experiments on various kinds of datasets and networks.

Our framework shares several limitations inherent to many ML methods. The first is the computational cost and time associated with training and optimizing KGEMs, which could be improved with the help of libraries focused on using multiple GPUs and distributing computation across a cluster. The second is that our

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023

methodology does not compensate for a lack of training data, as is often the case with clinical datasets, nor for poor quality data. The third is that this approach relies on the ability to meaningfully integrate two or more datasets with a KG. If there are too few mappable features between the clinical data and the KG, then the resulting patient representations may only have limited quality and utility, and some patients may be excluded entirely. Fourthly, the methodology to incorporate patients into the KG has been exclusively devised for continuous features. Lastly, despite our framework being generalizable to any dataset, there may be cases where CLEP fails to improve the performance as the original dataset possesses sufficient informative power for a particular application. In these cases, however, one could always employ the KG to interpret and identify biological mechanisms (e.g. pathways in the KG) associated with an individual or group of patients, as demonstrated in our case scenario.

While we generated two types of edges (i.e. -1 or +1) between ADNI patients and proteins in the PPI-KG, we also implicitly generated negative edges between each ADNI patient and all other proteins to which neither a -1 nor +1 edge was inferred. Unfortunately, we were unable to explicitly incorporate them in the training algorithm because PyKEEN, as well as most KGEM packages, generate negative edges through uniform negative sampling or Bernoulli negative sampling (Wang *et al.*, 2014b) with respect to the given positive KG. Thus, we are interested to make improvements to the upstream package itself to enable these kinds of explicit inclusions, as there are a growing number of negative edges available in various biological knowledge sources.

There are a number of applications in precision medicine for the representations generated by CLEP. For instance, in the AD area, patient representations of cognitively impared patients could be systematically used to stratify patients in order to identify shared mechanisms that explain their observed phenotype. Furthermore, our approach can be generalized such that different types of patient-level data can be integrated and mapped to heterogeneous biological KGs, thus enabling scientists to combine their datasets with context-specific knowledge. In the future, we plan to extend this work as well as to adapt it to well-known network representation learning methods such as node2vec (Grover and Leskovec, 2016) and LINE (Tang *et al.*, 2015). Finally, we ambition that our framework could serve as an integration platform for multimodal datasets including clinical, imaging and -omics data which could lead to more comprehensive patient representations.

Acknowledgements

The authors thank André Gemünd for his technical assistance running the experiments. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Author Contributions

D.D.-F. and C.T.H. conceived and designed the study. V.S.B. implemented CLEP and ran the experiments with supervision and support from D.D.-F. M.A. guided and supported the implementation of the knowledge graph embedding generation and classification tasks. C.B. assisted with data and method development. S.M. processed the knowledge graph. M.H.-A. and J.L. acquired the funding. All the authors contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

Funding

This work was funded by the Fraunhofer Cluster of Excellence 'Cognitive Internet Technologies' as well as by the German Federal Ministry of Education and Research (BMBF) [01IS18050D] (project 'MLWin').

Conflict of Interest: none declared.

References

Ali,M. *et al.* (2021) PyKEEN 1.0: a Python library for training and evaluating knowledge graph embeddings. *J. Mach. Learn. Res.*, **22**, 1–6.

Ali,M. *et al.* (2020) Bringing light into the dark: a large-scale evaluation of knowledge graph embedding models under a unified framework. *arXiv* preprint *arXiv*:2006.13365.

Bordes,A. *et al.* (2013) Translating embeddings for modeling multi-relational data. *Adv. Neural Inf. Process. Syst.*, 2787–2795.

Cavalli,F.M. *et al.* (2017) Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell*, **31**, 737–754.

Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Domingo-Fernández,D. *et al.* (2019) PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, **20**, 243.

Fan,J. *et al.* (2014) Challenges of big data analysis. *Natl. Sci. Rev.*, **1**, 293–314.

Fröhlich,H. *et al.* (2018) From hype to reality: data science enabling personalized medicine. *BMC Medicine*, **16**, 150.

Gong,F. *et al.* (2021) SMR: medical knowledge graph embedding for safe medicine recommendation. *Big Data Res.*, **23**, 100174.

Grover,A. and Leskovec,J. (2016) Node2Vec: scalable feature learning for networks. In *KDD: Proceedings. International Conference on Knowledge Discovery and Data Mining, 2016*, pp. 855–864.

Hagenauer,M.H. *et al.* (2018) Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. *PLoS One*, **13**, e0200003.

Hanhijärvi,S. *et al.* (2009) Randomization techniques for graphs. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 780–791.

Himmelstein,D.S. and Baranzini,S.E. (2015) Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS Comput. Biol.*, **11**, e1004259.

Ho,T.K. (1995) Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*. pp. 278–282.

Hu,W. *et al.* (2017) BioSearch: a semantic search engine for Bio2RDF. *Database*, **2017**, bax059.

Jassal,B. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

Khanna,S. *et al.* (2018) Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci. Rep.*, **8**, 1–13.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023

Lin,Z. *et al.* (2020) Patient similarity via joint embeddings of medical knowledge graph and medical entity descriptions. *IEEE Access*, **8**, 156663–156676.

Lynam,A.L. *et al.* (2020) Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn. Prognostic Res.*, **4**, 1–10.

Maïza,A. *et al.* (2018) The role of heparan sulfates in protein aggregation and their potential impact on neurodegeneration. *FEBS Lett.*, **592**, 3806–3818.

Mueller,S. *et al.* (2005) Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's Dementia*, **1**, 55–66.

Muslu,O. *et al.* (2020) GuiltyTargets: prioritization of novel therapeutic targets with deep network representation learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, doi: 10.1109/TCBB.2020.3003830.

Nickel,M. *et al.* (2016) Holographic embeddings of knowledge graphs. In: *Thirtieth AAAI Conference on Artificial Intelligence*.

Orchard,S. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.

Oughtred,R. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.

Pai,S. and Bader,G.D. (2018) Patient similarity networks for precision medicine. *J. Mol. Biol.*, **430**, 2924–2938.

Pai,S. *et al.* (2019) netDx: interpretable patient classification using integrated patient similarity networks. *Mol. Syst. Biol.*, **15**, e8497.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Raphael,B.J. *et al.* (2017) Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*, **32**, 185–203.

Rodchenkov,I. *et al.* (2020) Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.*, **48**, D489–D497.

Saito,T. and Rehmsmeier,M. (2015) The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, **10**, e0118432.

Saykin,A.J. *et al.*; Alzheimer's Disease Neuroimaging Initiative. (2015) Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans. *Alzheimer's Dementia*, **11**, 792–814.

Slenter,D.N. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.

Smith,A.M. *et al.* (2020) Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, **21**, 1–18.

Sun,Z. *et al.* (2019) Rotate: knowledge graph embedding by relational rotation in complex space. *arXiv* preprint *arXiv*:1902.10197.

Tahara,K. *et al.* (2006) Role of toll-like receptor signalling in Aβ uptake and clearance. *Brain*, **129**, 3006–3019.

Tang,J. *et al.* (2015) Line: large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067–1077.

Trouillon,T. *et al.* (2016) Complex embeddings for simple link prediction. In: *International Conference on Machine Learning*, pp. 2071–2080.

Ulhaq,Z.S. and Garcia,C.P. (2020) Estrogen receptor beta (ESR2) gene polymorphism and susceptibility to dementia. *Acta Neurol. Belgica*, doi: 10.1007/s13760-020-01360-z.

Walter,S. *et al.* (2007) Role of the toll-like receptor 4 in neuroinflammation in Alzheimer's disease. *Cell Physiol. Biochem.*, **20**, 947–956.

Wang,B. *et al.* (2014a) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Wang,Z. *et al.* (2014b) Knowledge graph embedding by translating on hyperplanes. *AAAI*, **14**, 1112–1119.

Wu,M. *et al.* (2019) Identification of key genes and pathways for Alzheimer's disease via combined analysis of genome-wide expression profiling in the hippocampus. *Biophys. Rep.*, **5**, 98–109.

Xu,C. and Jackson,S.A. (2019) Machine learning and complex biological data. *Genome Biol.*, **20**, 76.

Yu,D. *et al.* (2013) Review of biological network data and its applications. *Genomics Inf.*, **11**, 200–210.

Zhao,L. *et al.* (2015) Estrogen receptor β in Alzheimer's disease: from mechanisms to therapeutics. *Ageing Res. Rev.*, **24**, 178–190.

Zitnik,M. *et al.* (2018) Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **34**, i457–i466.

Zitnik,M. *et al.* (2019) Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf. Fusion*, **50**, 71–91.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **67**, 301–320.

Downloaded from https://academic.oup.com/bioinformatics/article/37/19/3311/6272574 by guest on 17 November 2023