Received: 7 January 2022 | Revised: 11 October 2022 | Accepted: 17 October 2022

DOI: 10.1111/exsy.13181

ORIGINAL ARTICLE

Expert Systems WILEY

Biomedical knowledge graph embeddings for personalized medicine: Predicting disease-gene associations

Joana Vilela[1,2] | Muhammad Asif[1,2,3] | Ana Rita Marques[1,2] | João Xavier Santos[1,2] | Célia Rasga[1,2] | Astrid Vicente[1,2] | Hugo Martiniano[1,2]

[1]Departamento de Promoção da Saúde e Prevenção de Doenças não Transmissíveis, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisbon, Portugal

[2]BioISI – Biosystems and Integrative Sciences Institute, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal

[3]Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Faisalabad, Pakistan

Correspondence
Hugo Martiniano, Instituto Nacional de Saúde Doutor Ricardo Jorge, Av. Padre Cruz, Lisbon, Portugal.
Email: hugo.martiniano@insa.min-saude.pt

Funding information
Fundação para a Ciência e a Tecnologia, Grant/Award Numbers: SAICTPAC/0010/2015, POCI-01-0145-FEDER-016428-PAC, EXPL/CCI-BIO/0126/2021, PTDC/MED-OUT/28937/2017, UIDP/04046/2020, UIDB/04046/2020; Fundo Europeu de Desenvolvimento Regional, Grant/Award Number: 022153

Abstract

Personalized medicine is a concept that has been subject of increasing interest in medical research and practice in the last few years. However, significant challenges stand in the way of practical implementations, namely in regard to extracting clinically valuable insights from the vast amount of biomedical knowledge generated in the last few years. Here, we describe an approach that uses Knowledge Graph Embedding (KGE) methods on a biomedical Knowledge Graph (KG) as a path to reasoning over the wealth of information stored in publicly accessible databases. We built a Knowledge Graph using data from DisGeNET and GO, containing relationships between genes, diseases and other biological entities. The KG contains 93,657 nodes of 5 types and 1,705,585 relationships of 59 types. We applied KGE methods to this KG, obtaining an excellent performance in predicting gene-disease associations (MR 0.13, MRR 0.96, HITS@1 0.93, HITS@3 0.99, and HITS@10 0.99). The optimal hyperparameter set was used to predict all possible novel gene-disease associations. An in-depth analysis of novel gene-disease predictions for disease terms related to Autism Spectrum Disorder (ASD) shows that this approach produces predictions consistent with known candidate genes and biological pathways and yields relevant insights into the biology of this paradigmatic complex disorder.

KEYWORDS
Autism Spectrum Disorder, gene–disease associations, Knowledge Graph Embedding, personalized medicine

1 | INTRODUCTION

Personalized medicine is a clinical approach that has been subject of increasing interest in medical research in the last years, and is grounded in the idea that each individual clinical context is unique and manifests in particular ways, that are driven by a specific clinical, physiological and molecular context, modulated by the exposure to environmental factors (Goetz & Schork, 2018).

Abbreviations: ASD, Autism Spectrum Disorder; KG, Knowledge Graph; KGE, Knowledge Graph Embedding; PM, Personalized Medicine.

It is widely known that a person's genetic background has an important contribution to several diseases and the advances in DNA sequencing technologies are contributing to understand how genetic variability determines the occurrence of diseases, promoting more accurate diagnosis and improving the development of personalized medicine directions (Vicente et al., 2020).

While advances in DNA sequencing methods are finally converging to a point where the integration of genomics into clinical practice is becoming a reality, the interpretation of sequencing results involves the identification of a relatively small number of disease-associated variants among the large number of common variants carried by an individual. This is often hampered by the lack of knowledge about the relationships between variants, genes and diseases, which precludes the identification of disease causing mutations, leading to low diagnostic yields.

The genetic architecture of complex diseases involves a large number of genes and it is hypothesised that this effect is present even in seemingly simple diseases (Boyle et al., 2017), and that the same genes can play a central role in several diseases, sometimes apparently unrelated. The former implies that mutations in several different genes can all contribute to a disease, while the later mean that the same mutation in the same gene can lead to different diseases (Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, 2017).

Personalized medicine is an approach in which patients are stratified based on their clinical profile. Patient stratification can take into account disease subtype, prognosis or treatment response, using diagnostic tests to support medical decisions, including molecular and behavioural biomarkers (Fröhlich et al., 2018). Recently several molecular level approaches are being developed to understand the genetic contribution to diseases. Complex diseases are a sum of genetic and environmental factors. A great proportion of the diseases follow this pattern, as congenital or adult-onset diseases, and several developmental disorders. Some examples of complex disorders include Autism Spectrum Disorder (ASD), Alzheimer, multiple sclerosis, autoimmune diseases, and others (Hunter, 2005).

Several approaches are being applied to define sets of candidate genes, which can be associated to human diseases. These range from manual curation efforts, including crowdsourcing approaches, such as the efforts of the communities contributing to PanelApp (Martin et al., 2019), Clinvar (Landrum et al., 2018) or Clingen (Rehm et al., 2015), or more traditional curation approaches such as those from OMIM (Amberger et al., 2015, 2019), to hybrid or fully automated approaches often using data mining and machine learning to derive insights from large amounts of structured or unstructured data (e.g., Alshahrani & Hoehndorf, 2018; Himmelstein et al., 2017; Hu et al., 2021; Liang et al., 2019; Luo, Li, et al., 2019; Luo, Xiao, et al., 2019; Nunes et al., 2021; Smaili et al., 2019; Wang et al., 2019; Yu et al., 2021). The latter rely more on data obtained with text-mining methods, while the former can include a multitude of approaches using data from one or more of several publicly available biomedical, clinical or biological databases, often containing data obtained by text-mining the scientific literature. Here, we concern ourselves with approaches dealing with graph or network data, more specifically heterogeneous multi-graphs.

Increasing amounts of biological and biomedical knowledge are produced everyday. Despite all the efforts to collect and organize this information, several challenges remain in integrating all the information scattered throughout different databases and obtaining meaningful insights from this wealth of data. In this work, we explore the use of Knowledge Graph Embedding (KGE) methods (Wang et al., 2017) as a tool to model the relationships between biological entities such as genes and diseases, and gain valuable insights into their associations that can be of use in the area of personalized medicine.

For this purpose, we built a large-scale Knowledge Graph (KG) combining data from publicly accessible curated biological and biomedical databases and applied Knowledge Graph Embedding (KGE) methods as a means to extract novel information from this KG. KGE methods have seen increased use in several areas. Some of the reasons for the success of these method lie in their broad applicability, scaling capabilities and good performance (Wang et al., 2017). In the past few years, KGs and KGE methods have seen broad application in various tasks in the biological and biomedical domains, such as drug repurposing, prediction of gene-disease associations and identification of drug side-effects (Himmelstein et al., 2017; Himmelstein & Baranzini, 2015; Liang et al., 2019; Mohamed et al., 2021; Nicholson & Greene, 2020; Nunes et al., 2021).

A Knowledge Graph (KG) is a directed heterogeneous multi-graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where each vertex ($v \in \mathcal{V}$) constitutes an entity (with a given entity type) and each edge ($e \in \mathcal{E}$) a relationship. Entities and relationships in a KG are organized in sets of triples (h, r, t), where h is the head entity, r is the relationship and t is the tail entity, h and t are vertices in the graph, while r is an edge connecting h and t. Each triplet in the KG represents a fact, where the head entity (or subject) is related to the tail entity (or object) through the relationship.

Knowledge Graph Embedding (KGE) methods learn a representation of entities in $\mathbb{R}^d$, termed an embedding, such that the representation in the embedding space reflects their relationships with other entities in the KG. This is done by optimizing a score function: $f(h, r, t)$. Several methods have been proposed for this task, with different score functions, such as ComplEx (Trouillon et al., 2016), DistMult (Yang et al., 2015) and TransE (Bordes et al., 2013). The resulting embedding vectors can be used for downstream supervised or unsupervised machine learning tasks.

To show the feasibility of our proposed approach in clinical settings, we applied it to a highly complex and heterogeneous disorder: Autism Spectrum Disorder (ASD). ASD is a neurodevelopmental disorder characterized by two main characteristics: communication deficits and repetitive behaviours (Diagnostic and Statistical Manual of Mental Disorders: Dsm-5, 2013). ASD segregates in families, has a strong genetic component and is clinically heterogeneous, often co-occurring with other conditions (Lord et al., 2020). Early and efficient intervention for children with ASD is fundamental, but pharmacological therapies can only be used to treat some of the associated symptoms or comorbidities, and do not target core symptoms. ASD can vary highly in the clinical presentation and in the associated symptoms. The underlying genetic causes of ASD are unclear, except when co-occurring with genetic syndromes. Given the diversity of biological mechanisms that can be affected, the development of

therapeutic approaches is challenging. In recent years, several groups, including ourselves, have developed, integrative approaches based on machine learning methods to obtain insights into the genetic and phenotypic complexity of ASD beyond what can be obtained with conventional analysis methods (Asif et al., 2018, 2020; Duda et al., 2018; Krishnan et al., 2016; Martiniano et al., 2020).

Here, we report an application of KGE methods to a custom-built biological KG, relating entities such as genes, biological processes and diseases, and showcase its application in the area of personalized medicine, namely for the prediction of gene-disease associations. As a use case of the applications of the gene-disease associations prediction algorithm developed, we identify and validate a set novel genes associated to ASD.

This paper is structured as follows, first we introduce the general area and the specific challenges we address, we then describe the methodology, including all data sources and software tools used. Afterwards we present our results and discuss them, focusing on the validation of genes and biological pathways predicted as candidates for implication in ASD. We conclude with an overview of the study, a discussion of the potential implications of our results and point out some future directions of this line of work.

2 | METHODS

2.1 | Data sources

Three main data sources were used to construct the KG: (a) Gene Ontology (https://geneontology.org); (b) DisGeNet (https://disgenet.org); and (c) Ensembl (https://ensembl.org).

2.1.1 | Gene Ontology

The Gene Ontology (GO) resource (http://geneontology.org/) develops structured controlled ontologies to characterize genes and their products (Ashburner et al., 2000). The GO ontology and the respective gene annotations were downloaded from the GO website (GO version: 2020-09-10) and GOA (GOA version: 2020-10-10), downloaded on October 12, 2020.

2.1.2 | DisGeNET

DisGeNET (V7, downloaded May 7, 2020) was obtained from the DisGeNET website (https://www.disgenet.org/). DisGeNET is a database containing publicly available collections of genes and variants associated to human diseases (Piñero et al., 2015, 2017, 2020), that integrates data from several sources, such as expert curated sources, Genome Wide Association Studies catalogues, animal models and the scientific literature. Data is annotated with controlled vocabularies and community-driven ontologies. The current version of DisGeNET (v7.0) contains 1,134,942 gene-disease associations, between 21,671 genes and 30,170 diseases, disorders, traits, and clinical or abnormal human phenotypes. This release also contains 369,554 variant-disease associations, between 194,515 variants and 14,155 diseases, traits, and phenotypes.

2.1.3 | Ensembl

Ensembl (https://www.ensembl.org/index.html) is a large-scale bioinformatics project to collect and organize information concerning genome sequencing of several species. The services provided by Ensembl include a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Among other activities Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data (Howe et al., 2021).

2.2 | Knowledge Graph

Using data obtained from the data sources described above, we built an integrated biomedical Knowledge Graph (KG). This KG is composed of a series of biological entities and their relationships. First, we obtained the full GO OBO file. For each relationship extracted from the GO, we kept the original semantics as much as possible. The annotation qualifier was used to build the relationship types, using the annotations files for human gene products, both for proteins and for RNA. Gene-disease associations from DisGeNET v7 were then merged. All gene names were converted to Ensembl Gene IDs. Conversion of gene symbols in DisGeNET and GO to Ensembl symbols was done with Ensembl Biomart, using the pybiomart python client.[1] The KG contains five unique entity types: genes, diseases, molecular functions, cellular components and biological

processes. Entities are represented by their codes in the various databases. Genes are represented by their Ensembl Gene IDs, diseases, phenotypes and disease groups are represented by Concept Unique Identifiers (CUI) from the Unified Medical System (UMLS), as obtained from DisGeNET. All GO terms for biological processes, molecular functions and cellular components, represented by their respective GO IDs.

2.3 | Knowledge Graph embeddings

We applied Knowledge Graph embedding methods to produce vector representations (embeddings) of the entities in the KG. In this study, we tested three KG embedding algorithms, ComplEx (Trouillon et al., 2016), DistMult (Yang et al., 2015) and TransE (Bordes et al., 2013), as implemented in the DGL-KE package (Zheng et al., 2020). Training is performed through negative sampling by corrupting triples (h, r, t) to create triples of the form (h', r, t) or (h, r, t'), where h' and t' are randomly sampled from the sets of h and t. We apply filtered sampling, whereby generated negative triples that are present in the KG are discarded from the set of negatives used in the training process. Table 1 contains a summary of all KGE methods used and their respective scoring functions.

2.3.1 | Training

We performed a 60/20/20 split of all the gene-disease associations in the KG into training, test, and validation sets, stratified to ensure that all genes and diseases are present in all sets in roughly equal amounts. For training and testing of the embedding step the set of go-go and go-gene triples was added to the training set only. The testing and validation sets consist solely of gene-disease associations. As the main objective is to predict gene-disease associations, this ensures that the method is explicitly trained to reproduce these as well as possible. Hyperparameter tuning was done using the training and test sets. For a more efficient exploration of the possible hyperparameter space we used the Optuna optimization framework (Akiba et al., 2019). Optuna is a hyperparameter optimization software package that implements several search strategies to achieve optimal coverage of high-dimensional hyperparameter spaces. The maximum number of evaluations was set to 30 and the default settings for Optuna were used. The following hyperparameter sets were sampled: max_step in {500, 1000, 2000, 5000, 10,000, 20,000, 50,000}, hidden_dim in {100, 200, 300, 400, 500}, neg_sample_size in {100, 200, 300, 400, 500, 1000, 2000, 5000}, batch_size in {1000, 2000, 5000, 10,000}, regularization_coef in {1e−5, 1e−6, 1e−7, 1e−8, 1e−9} and lr in {0.1, 0.01}. All other hyperparameters were left at their default values, with the exception of the gamma parameter, which was set to 12.

2.3.2 | Evaluation

To evaluate the performance of the KGE step we used standard ranking metrics, as calculated by the DGL-KE package: Mean Rank (MR), Mean Reciprocal Rank (MRR), HITS@1, HITS@3, and HITS@10 (the mean fraction of true results in the top 1, 3, and 10, respectively). These are defined as:

$$\text{HITS@}k = \frac{1}{|Q|} 1_{\text{rank}_i \le k}, \tag{1}$$

$$\text{MR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \text{rank}_i, \tag{2}$$

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}, \tag{3}$$

TABLE 1 Knowledge graph embedding methods used in this study and their respective scoring functions

Method	Scoring function
ComplEx	$-\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_{\frac{1}{2}}$
DistMult	$\mathbf{h}^T \text{diag}(\mathbf{r})\mathbf{t}$
TransE	$\text{Real}\left(\mathbf{h}^T \text{diag}(\mathbf{r})\mathbf{t}\right)$

where, Q is the number of elements in the ranked list and $1_{rank_i \leq k}$ is 1 if $rank_i < k$, otherwise is 0. MRR was used as the optimization target for Optuna. Performance evaluation was done with a negative sample size of 16 and a batch size of 2048.

To avoid test set contamination, evaluation was performed with the training and validation set, withholding the test set used for hyperparameter tuning.

2.4 | Prediction of disease-gene associations

The prediction of novel disease-gene associations can be framed as a link prediction problem on the KG. In link prediction, the aim is to learn a scoring function f, characteristic of the method being employed (see Table 1), The scoring function assigns scores $= f(h, r, t)$ to each input triple $(h, r, t) \in \mathcal{G}$, where $h,t \in \mathcal{V}$ are the head and tail entities and $r \in \mathcal{E}$ is the relationship. In this particular case, the head entities are genes, the tail entities are diseases and the relationship is the association of gene to diseases. This produces a ranking of genes for each disease, where the genes are ranked from higher to lower association to a given disease. Prediction of gene-disease associations was performed using the full KG with the best hyperparameters identified with the optimization procedure described above.

2.5 | Analysis of ASD-associated genes

From the set of predicted gene-disease associations produced as described above, we selected those involving autism-related disease terms. Genes in the first decile of novel associations were merged to create a list of ASD candidate genes. We used this gene list to produce a network of protein–protein interactions (PPI) with edge weights, using STRING (Franceschini et al., 2013, 2016; Snel et al., 2000; Szklarczyk et al., 2015, 2017, 2019, 2021) and applied the Leiden community detection algorithm (Traag et al., 2019) to the PPI to identify network functional modules (biological communities), as implemented in the CDlib python package[2] (Rossetti et al., 2019). The Leiden community detection algorithm is based on modularity optimization and is able to detect partitions in the whole dataset and identify the hierarchical community structure. Using this method with the default parameters, we decomposed the network into sub-units or communities. The identification of functional protein communities in the network may uncover a priori unknown functional biological modules.

Finally, to assess the enrichment in biological pathways of each community, we used Reactome pathways (Griss et al., 2020; Jassal et al., 2020; Wu & Haw, 2017). Reactome is a manually curated, peer-reviewed pathway database, widely used for clinical research purposes. Enrichment analysis was performed with the g:Profiler (Raudvere et al., 2019) python client.[3]

3 | RESULTS AND DISCUSSION

3.1 | Knowledge Graph

We created a Knowledge Graph (KG) by integrating data from two publicly accessible databases: GO and DisGeNET. Figure 1 depicts the metagraph of the KG. The KG has 1,705,585 triples, composed of 93,657 unique entities and 59 relationship types. These entities comprise 28,243 genes, 21,623 diseases, 11,170 molecular functions, 4183 cellular components and 28,438 biological processes. Regarding relationships, the KG contains 900,442 gene-disease associations, while the rest of the relationships comprise gene-GO annotations (715,550) and GO ontology relationships (89,593). Two major differences from other studies employing similar approaches are the number of genes and the proportion of RNA gene products. Most studies deal with smaller gene sets or solely with protein-coding genes and approaches similar in scope and size, such as hetionet (Himmelstein et al., 2017; Himmelstein & Baranzini, 2015), contain 73% of the genes present in our KG. This maximizes the predictive capabilities of our approach, and we expect to expand the KG in the future by increasing it is reach to a larger number of genetic features.

3.2 | Knowledge Graph embedding

A comparison of the performance of the KGE methods tested is displayed in Table 2. The performance metrics reported were calculated on the validation set using the optimum hyperparameter set for each algorithm, identified as described in the methods section. All methods exhibit good performance, with the TransE algorithm with l2-regularization exhibiting the best results. All subsequent analysis steps use embeddings trained using TransE with l2-regularization.

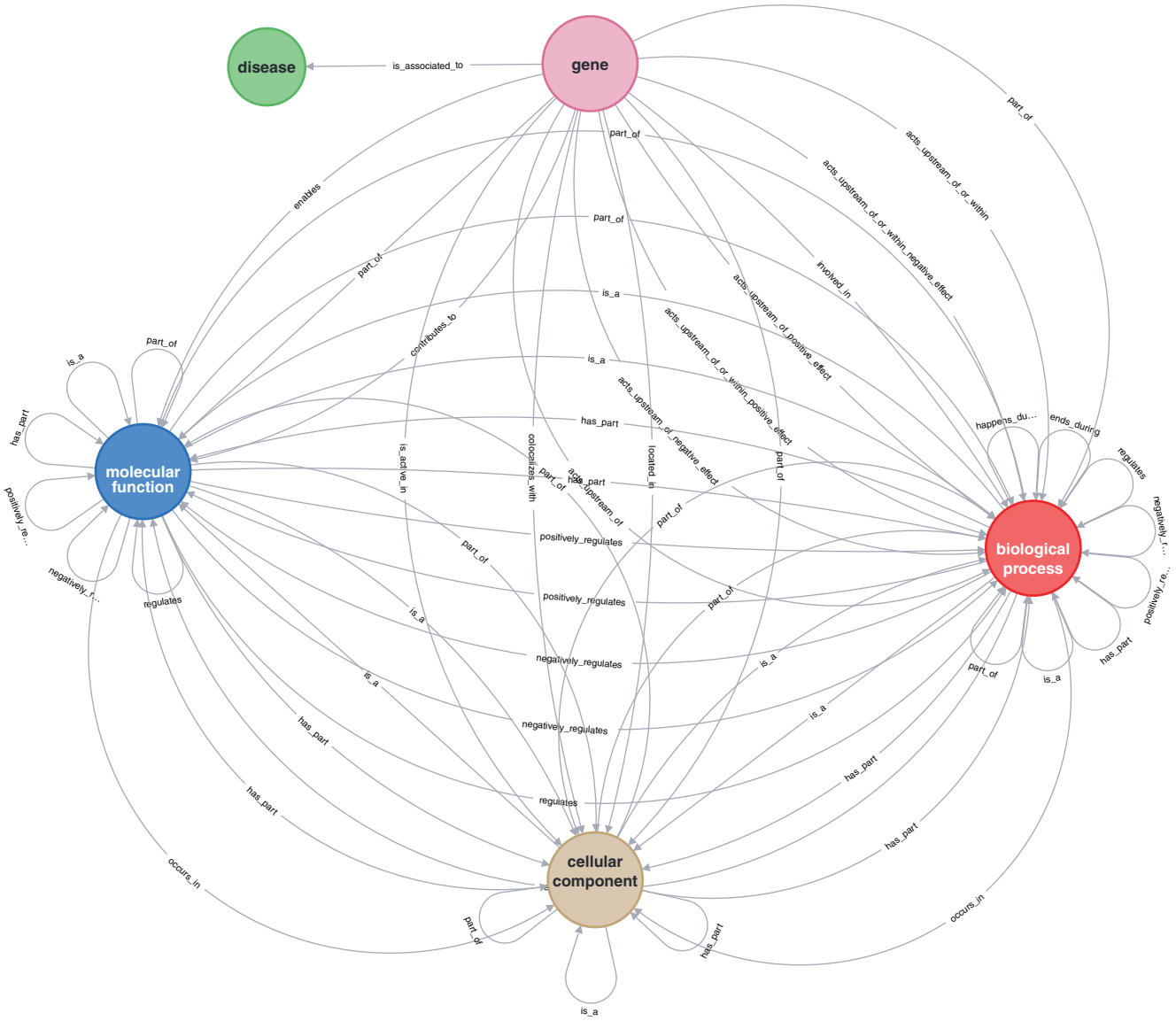

FIGURE 1 Metagraph representing all the node and relationship types in the KG. Circles (nodes) represent the KG different entities and lines represent the relationship types

TABLE 2 Performance metrics for each method

Method	MRR	MR	HITS@1	HITS@3	HITS@10
DistMult	0.93	1.24	0.88	0.98	**0.99**
ComplEx	0.95	1.17	0.92	0.98	**0.99**
TransE (l1)	0.95	1.17	0.91	**0.99**	**0.99**
TransE (l2)	**0.96**	**1.13**	**0.93**	**0.99**	**0.99**

Note: Highest values in each column are represented in bold.

3.3 | Prediction of gene-disease associations

Prediction of new genes associated to diseases is an important task in the context of personalized medicine approaches. New case–control studies can be designed taking these new associations into account, and analysis of genetic mutations in candidate genes resulting from these associations can lead to improved diagnosis and therapeutic interventions.

Using the TransE (l2) method with the optimum set of hyperparameters, we produced genome-wide predictions of gene-disease associations, that is, we predicted the scores of all possible gene disease-associations for all 28,243 genes and 21,623 diseases, producing a total of 610,698,389 predictions.

Other approaches for the prediction of disease-gene associations have explored the use of the GO as underlying source of data, either from gene semantic similarity or from embedding of GO and other ontologies (Alshahrani & Hoehndorf, 2018; Liang et al., 2019; Nunes et al., 2021; Smaili et al., 2019). Our approach offers an excellent performance and is easy to apply and to extend. To validate our approach from a biological and biomedical point of view, we apply it to the identification of novel candidate genes for ASD. The next section describes and discusses our results.

3.4 | Use case: Prediction of genes associated to Autism Spectrum Disorder

Autism Spectrum Disorder results of a combination of environmental and genetic factors, has a strong genetic component, segregates in families and there is an estimate of up to 1000 genes potentially implicated in the disease (Ramaswami & Geschwind, 2018). While several ASD-associated genes are present in the KG, this list is non-exhaustive, as the genetic diagnosis yields for ASD are usually low (Kreiman & Boles, 2020; Savatt & Myers, 2021), indicating that a larger number of genes is probably implicated. Here, we aimed to expand the list of ASD candidate genes by producing novel gene-disease association predictions for this disorder and use the produced genome-wide ranking to identify major biological communities.

3.4.1 | Prediction of ASD-associated genes

For the prediction of genes associated to ASD we selected two disease terms in the KG which correspond to general forms of ASD, 'Autism Spectrum Disorders' (C1510586) and 'Autistic Disorder' (C0004352). The scores of the associations of all genes in the KG for these two disease terms were extracted from the final prediction set, produced as described previously. Rankings for the association of all genes to these two terms were derived from the scores of the corresponding association, retaining only novel associations (i.e., those not present in the KG). For both ranked lists, we selected all genes in the first decile of the ranking (see supplementary file 1). The two gene sets were merged, resulting in a list composed of 3389 genes. This list was used for subsequent analyses.

3.4.2 | Identification of biological communities

To identify biological pathways that can be shared by people with mutations in ASD-associated genes, we created a network consisting of the genes from the ASD candidate gene list and gene–gene interactions obtained from STRING. For further analysis, we retained the largest connected component of this network, containing 3221 genes.

The interaction network containing these 3221 genes was used to perform network community detection using the Leiden algorithm. Six communities were identified (see supplementary file 2). Enrichment analysis of each community indicates that the PPI network is enriched in several biological pathways (Figure 2). From the results of enrichment analysis (supplementary file 3) we identified the communities as corresponding to six main pathways: Metabolism; Chemical synapse transmission mediated by G Protein Coupled Receptors (GPCRs); Cytokine signalling; Gene expression; Nervous system development and Signalling. All these biological pathways are likely to be affected in ASD in different patients as they are important to the nervous system and neuronal development at some stage of brain development. There is growing evidence linking these pathways to ASD and we discuss these connections and characterize the biological communities found in the next subsection.

3.4.3 | Characterization of biological communities

Chemical synapse transmission/GPCRs

There is strong genomic and functional evidence indicating that synaptic biological processes are altered in ASD (Abrahams & Geschwind, 2008; Lai et al., 2021; Leblond et al., 2014; Lionel et al., 2013; Tromp et al., 2021). Several studies suggest that mutations in genes that encode proteins that establish the connection between two neurons and the formation of a synapse such as the ones that encode neurexins, neuroligins or Shank proteins (genes that are also present in our gene-disease associations) share biological pathways including the synaptic pathways (Gong & Wang, 2015; Guang et al., 2018; Lai et al., 2021; Tromp et al., 2021). Synaptic transmission occurs between a presynaptic neuron and a postsynaptic cell. Neurotransmitters establish the communication between neurons and bind to ion channels on postsynaptic neurons to modulate voltage changes. Important modulators of neurotransmission are G protein coupled receptors (GPCRs), a superfamily of key proteins responsible for the signal transduction across cell membranes and that mediate diverse cellular responses. GPCRs mediate the regulation of synaptic transmission,

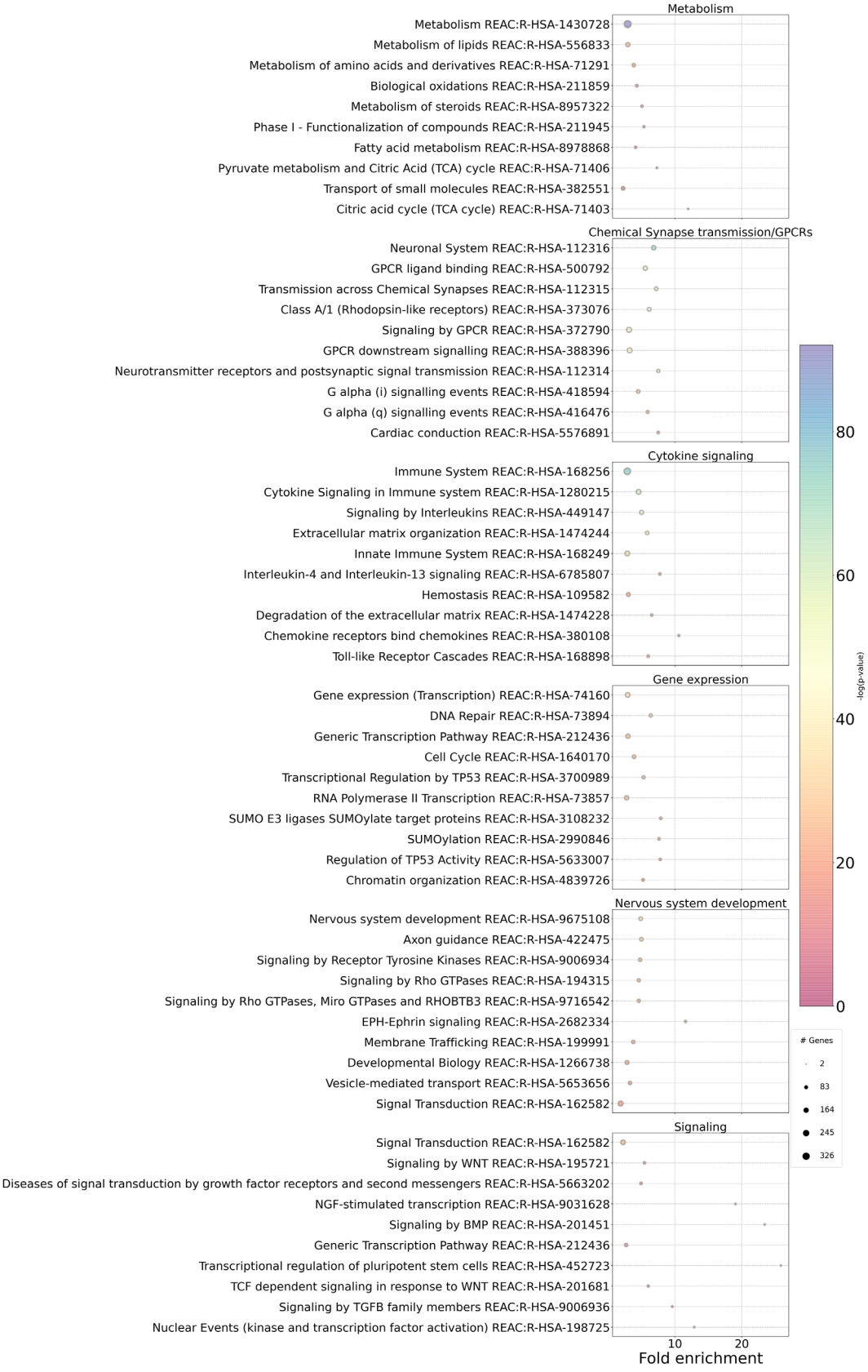

FIGURE 2 Reactome pathway enrichment of the PPI network in biological communities. The PPI network is enriched in several biological pathways with most represented being involved in six main pathways. The chart displays (*X* axis) the number of times each community is enriched in a term more than expected by chance, by (*Y* axis) the probability of obtaining the same result by chance; circle size represents the number of genes enriched in the term and circle colours the magnitude of the enrichment *p*-value

binding specifically to some neurotransmitters, modifying the structure of the receptor and regulating the mechanism of neurotransmission (Betke et al., 2012; Lutzu & Castillo, 2021). One of the difficulties in the pharmacotherapeutic research in ASD is the identification of effective pathophysiological targets. Most of the approaches developed target brain excitatory/inhibitory imbalance caused by alterations in gamma-aminobutyric acid (GABA) and glutamate receptors (DelaCuesta-Barrutia et al., 2020). However, there are other important neurotransmitter systems that are key for the proper establishment of brain excitatory/inhibitory balance as the ones regulating important neurotransmitters as oxytocin, serotonin or dopamine. These systems are primarily mediated through specific GPCRs (DelaCuesta-Barrutia et al., 2020; Gurevich, Gainetdinov, and Gurevich et al., 2016; McCorvy & Roth, 2015; Willets et al., 2009) and are also important for the brain excitatory/inhibitory balance, representing possible therapeutic targets (Marotta et al., 2020). The dysfunction of GPCRs potentially implicated in ASD, including the glutamatergic, dopaminergic, oxytocinergic or serotonergic systems, can contribute to the disorder, and new clinical directions taking these pathways into account can result in the discovery of noval treatments, as has been suggested for other brain disorders such as schizophrenia (DelaCuesta-Barrutia et al., 2020).

Gene expression

Autism research has long focused on genes involved in neuronal development and synaptic processes. Mutations in genes participating in these processes were the first to be linked to ASD and its symptomatology. However, in recent years, several studies have implicated other classes of genes and, often, the ones related with gene expression, chromatin organization and remodelling are mentioned. Genes involved in chromatin regulation determine whether other genes are turned off or not according to the need of being expressed or not. For a gene to be expressed at the right time, DNA needs to go through conformational changes from tightly to loosely packed coils. This process is controlled by chromatin remodelling complexes, and genes involved in such mechanisms are sometimes mutated in ASD and other neurodevelopmental disorders. Mutations in these complexes have been linked to ASD, Schizophrenia or Intellectual disability and other conditions (Gabriele et al., 2018).

Cytokine signalling

Although ASD pathophysiology is unclear, growing evidence also supports an important role of neuroinflammatory processes. The participation of astrocytes and microglia in ASD has been subject of study due to their roles in the regulation of immune and synaptic pathways. Elevated levels of reactive microglia and astrocytes in postmortem tissue in ASD has been reported (Matta et al., 2019). The immune system is interconnected to the nervous system and its dysfunction impacts several biological processes, including brain function and development, and behaviour (Filiano et al., 2015). Fever occurs as a body response to fight infection and is initiated by cytokines (Dantzer et al., 2008). The brain recognizes cytokines as signals of sickness (Dantzer, 2009). Cytokines are signalling molecules that mediate the communication among cells in the immune system, and are primary regulators of inflammation. Studies involving immune system alterations and ASD, including on the characterization of cytokine profiles, have been increasing in the last years (Masi et al., 2017).

Metabolism

The contribution of metabolic alterations to developmental disorders has been the subject of several studies. Metabolic alterations at different levels have also been reported in ASD, such as the ones involving biological oxidations (Bjørklund et al., 2020; Frye et al., 2013), alterations in the lipid metabolism (Luo et al., 2020; Tamiji & Crawford, 2010) and in Cytochrome P450 pathways. Oxidative stress is thought to be implicated in ASD, as shown by reports of increased levels of Reactive Oxygen Species (ROS) and increased lipid peroxidation (Bjørklund et al., 2020). Oxidative stress is an important cause of neuroinflammation and can contribute to ASD (Bjørklund et al., 2020). A significant portion of individuals diagnosed with ASD have elevated peripheral cytokines and chemokines and associated neuroinflammation (Bjorklund et al., 2016). People with ASD are considered more sensitive to oxidative stress due to glutathione imbalance (James et al., 2006), and the contribution of environmental exposure to heavy metals has also been discussed (Macedoni-Lukšič et al., 2015; Mostafa et al., 2016). Several studies have suggested that the oxidation–reduction imbalance and oxidative stress are important components of ASD pathophysiology (Yui et al., 2016). Regarding the relationship between lipid metabolism and ASD, it is known that the nervous system is enriched with important classes of lipids, thus the dysfunction of lipid metabolic pathways can play a role in the development of this disorder. Cholesterol and sphingolipids are signalling molecules with key roles in neuronal differentiation and in synaptogenesis. Cholesterol availability is essential to synapse development (Hussain et al., 2019). Several studies report abnormal levels of lipids in ASD, and some of these studies reported alterations in cholesterol and triglyceride levels in a subgroup of patients with ASD (Luo et al., 2020; Sikora et al., 2006). There is increasing evidence that alterations in fatty acid pathways may affect the nervous system leading to ASD. In line with these reports, there is evidence supporting the hypothesis that people with ASD have higher rates of lipid metabolism than controls, and that the dysregulation along the lipid metabolic pathway may contribute to ASD onset (Tamiji & Crawford, 2010).

Nervous system development

The 'Nervous system development' biological community is enriched in genes participating in mechanisms that are important for neuronal development and axon guidance, such as the Rho family of GTPases, which are proteins that act as molecular switches that regulate important cellular processes, such as growth, migration, differentiation or adhesion. These molecules are particularly important to the nervous system, as they regulate neuronal function and morphology. Recent studies suggests that Rho GTPase dysfunction has a role in ASD, as several genes encoding Rho

GTPases are candidate risk genes for ASD and are incorporated in the ASD candidate gene list of the Simons Foundation Autism Research Initiative (SFARI) (see Guo et al., 2020 for a review of Rho family of GTPases involved in ASD). The SFARI database (Abrahams et al., 2013; Banerjee-Basu & Packer, 2010; Wang et al., 2012; Yao et al., 2015) is an ASD dedicated database that integrates a gene scoring module which establishes a gene rank according to the strength of the evidence that associates a given gene to the disease, based on the analyses of several studies with ASD patients. Genes like *MYO9B*, *OPHN1*, *SRGAP3*, *OCRL* or *ITPR1* are genes from the Rho family of GTPases present in the SFARI gene list that are also associated to ASD in the gene-disease associations predicted with the methodology developed in this study.

Signalling

Signalling pathways are important to ASD at diverse levels, as a complex brain and neurodevelopmental disorder, and our algorithm also identifies genes associated to ASD terms as being enriched in several signalling pathways (Signalling biological community; Figure 2) such as the Wnt signalling pathway. Interestingly, the Wnt signalling pathway is evolutionarily conserved and regulates fundamental early developmental processes as cell determination and migration, cell polarity, neural patterning and organogenesis, during the stages of embryonic development (Komiya & Habas, 2008). ASD is an early-onset disorder mainly impacted by the embryonic development. The canonical Wnt pathway is thus fundamental for brain development and, consequently, for a proper synaptic function (Mulligan & Cheyette, 2016). Mutations in genes participating in the Wnt pathway have been suggested to contribute to ASD and to other psychiatric disorders (Kalkman, 2012; Mulligan & Cheyette, 2016).

3.5 | Relevance for personalized medicine

With the present work, we show that our approach, despite being of general application, identifies plausible gene-disease associations in ASD, from which useful biological insights can be derived. The top ranking genes associated to ASD in the KG identified in this study are involved in six main relevant biological communities for the nervous system and neuronal development, often referred in the scientific literature as candidate pathways for the aetiology of the disease. The methodology developed in this work can be useful for patient stratification into subtypes according to the biological pathways enriched in the biological communities implicated in the gene-disease associations identified, and can provide insights for the development of guidelines for personalized medicine approaches applied to ASD.

4 | CONCLUSIONS

We describe an approach to integrate biological information from several data sources and predict gene-disease associations. This is done through the construction of a KG containing biological and biomedical entities and the application of KGE techniques for link prediction of the relationships of interest in the KG.

To showcase a biological application, this methodology was applied and tested on a paradigmatic complex disorder: ASD. We showed that our approach allows for data-driven detection of sub-communities, which can be useful for patient stratification. Stratification of patients is a daunting task for complex diseases such as ASD. The identification of genes and biological communities involved in ASD could provide a possible way for effective patient stratification strategies.

The top decile of novel ASD-associated genes is enriched in six main relevant biological pathways (Metabolism; Chemical synapse transmission mediated by G Protein Coupled Receptors (GPCRs); Cytokine signalling; Gene expression; Nervous system development and Signalling), which are here reinforced as candidate pathways for ASD aetiology that can be important for the development of guidelines for personalized medicine approaches applied to ASD.

The major contributions of this work are, from a technical viewpoint, the use of a readily extensible and adaptable large-scale KG, with a considerable proportion of RNA gene products and, from an application viewpoint, a data-driven approach for the identification of genes and pathways relevant to human diseases, which we have shown to be reliable in the case of ASD. Most related approaches are disease-specific or deal with smaller gene sets or are aimed exclusively at protein-coding genes. In this study, we aimed to maximize the number of genes and we explicitly included RNA genes and the respective GO annotations. The later are much more numerous than protein-coding genes and, despite a growing body of evidence linking non-coding RNAs to human diseases, under-explored when compared to their protein-coding counterparts.

This approach has the potential for impact in several areas related with personalized medicine, namely in the analysis of genetic sequencing data, in patient stratification or in the development of novel therapeutic approaches or the identification on novel therapeutic targets.

Regarding the analysis of genetic sequencing data, one major hurdle in current practice is the establishment of reliable variant prioritization methods that can identify disease-causing genetic variants in the large amount of data generated by sequencing. Methods that associate the affected genes to a disease or phenotype have been used to address this issue and the gene rankings produced with our approach can be easily be used for this purpose. Patient stratification is one major goal of precision medicine and the characterization of subgroups of patients according to their shared clinical profiles is of major importance. The method developed in this study has direct applicability to patient stratification through the identification of shared pathogenic burden in biological pathways or gene communities. The identification of novel therapeutic targets or

therapeutic approaches is another area where we expect our approach to have an impact. The identification and ranking of disease-associated genes and pathways can be particularly helpful in prioritizing or expanding the range of targets for functional studies or for the development of gene therapy approaches.

We conclude by noting that, although we focus on ASD, this approach is applicable to all diseases in the KG, especially the ones with a strong genetic contribution and with complex genetic architectures. In future studies, we plan to expand the size and the scope of the KG by adding information from other biological and biomedical databases and explore the use of other embedding methods. Work is under way to apply this approach to develop tools for the identification of disease-associated genetic variants in sequencing datasets and to develop methods of patient stratification in cohorts of subjects diagnosed with ASD.

ACKNOWLEDGEMENTS

The authors would like to acknowledge support by the UIDB/04046/2020 and UIDP/04046/2020 Centre grants from Fundação para a Ciência e a Tecnologia (FCT), Portugal (to BioISI). This work was supported by FCT, through funding to the GEnvIA project (PTDC/MED-OUT/28937/2017), the Deeper project (EXPL/CCI-BIO/0126/2021) and the MedPerSyst project (POCI-01-0145-FEDER-016428-PAC; SAICTPAC/0010/2015). This work used the European Grid Infrastructure (EGI) with the support of NCG-INGRID-PT/INCD (Portugal). This work was produced with the support of INCD funded by FCT and Fundo Europeu de Desenvolvimento Regional (FEDER) under the project 01/SAICT/2016 no. 022153.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID


Hugo Martiniano https://orcid.org/0000-0003-2490-8913

ENDNOTES

[1] https://github.com/jrderuiter/pybiomart.

[2] https://cdlib.readthedocs.io/.

[3] https://pypi.org/project/gprofiler-official/.

REFERENCES

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., Menashe, I., Wadkins, T., Banerjee-Basu, S., & Packer, A. (2013). SFARI gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular Autism*, *4*(1), 36. https://doi.org/10.1186/2040-2392-4-36

Abrahams, B. S., & Geschwind, D. H. (2008). Advances in autism genetics: On the threshold of a new neurobiology. *Nature Reviews. Genetics*, *9*(5), 341–355. https://doi.org/10.1038/nrg2346

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation Hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623–2631). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330701

Alshahrani, M., & Hoehndorf, R. (2018). Semantic disease gene embeddings (SmuDGE): Phenotype-based disease gene prioritization without phenotypes. *Bioinformatics*, *34*(17), i901–i907. https://doi.org/10.1093/bioinformatics/bty559

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian inheritance in man (OMIM [textregistered]), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*(D1), D789–D798. https://doi.org/10.1093/nar/gku1205

Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, *47*(D1), D1038–D1043. https://doi.org/10.1093/nar/gky1151

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Asif, M., Martiniano, H. F., Marques, A. R., Santos, J. X., Vilela, J., Rasga, C., Oliveira, G., Couto, F. M., & Vicente, A. M. (2020). Identification of biological mechanisms underlying a multidimensional ASD phenotype using machine learning. *Translational Psychiatry*, *10*, 43. https://doi.org/10.1038/s41398-020-0721-1

Asif, M., Martiniano, H. F. M. C. M., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through gene ontology. *PLoS One*, *13*(12), 1–15. https://doi.org/10.1371/journal.pone.0208626

Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular Autism*, *8*, 21. https://doi.org/10.1186/s13229-017-0137-9

Banerjee-Basu, S., & Packer, A. (2010). SFARI gene: An evolving database for the autism research community. *Disease Models & Mechanisms*, *3*(3), 133–135. https://doi.org/10.1242/dmm.005439

Betke, K. M., Wells, C. A., & Hamm, H. E. (2012). GPCR mediated regulation of synaptic transmission. *Progress in Neurobiology*, *96*(3), 304–321. https://doi.org/10.1016/j.pneurobio.2012.01.009

Bjørklund, G., Meguid, N. A., El-Bana, M. A., Tinkov, A. A., Saad, K., Dadar, M., Hemimi, M., Skalny, A. V., Hosnedlová, B., Kizek, R., Osredkar, J., Urbina, M. A., Fabjan, T., El-Houfey, A. A., Kałużna-Czaplińska, J., Gątarek, P., & Chirumbolo, S. (2020). Oxidative stress in autism spectrum disorder. *Molecular Neurobiology*, *57*(5), 2314–2332. https://doi.org/10.1007/s12035-019-01742-2

Bjorklund, G., Saad, K., Chirumbolo, S., Kern, J. K., Geier, D. A., Geier, M. R., & Urbina, M. A. (2016). Immune dysfunction and neuroinflammation in autism spectrum disorder. *Acta Neurobiologiae Experimentalis (Wars)*, *76*(4), 257–268. https://doi.org/10.21307/ane-2017-025

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26, pp. 2787–2795). Curran Associates, Inc. http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Dantzer, R. (2009). Cytokine, sickness behavior, and depression. *Immunology and Allergy Clinics of North America*, *29*(2), 247–264. https://doi.org/10.1016/j.iac.2009.02.002

Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W., & Kelley, K. W. (2008). From inflammation to sickness and depression: When the immune system subjugates the brain. *Nature Reviews. Neuroscience*, *9*(1), 46–56. https://doi.org/10.1038/nrn2297

DelaCuesta-Barrutia, J., Peñagarikano, O., & Erdozain, A. M. (2020). G protein-coupled receptor heteromers as putative pharmacotherapeutic targets in autism. *Frontiers in Cellular Neuroscience*, *14*, 343. https://doi.org/10.3389/fncel.2020.588662

Diagnostic and Statistical Manual of Mental Disorders: Dsm-5. (2013). *Amer psychiatric pub incorporated*. Google-Books-ID: EIbMlwEACAAJ.

Duda, M., Zhang, H., Li, H.-D., Wall, D. P., Burmeister, M., & Guan, Y. (2018). Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. *Translational Psychiatry*, *8*(1), 1–9. https://doi.org/10.1038/s41398-018-0098-6

Filiano, A. J., Gadani, S. P., & Kipnis, J. (2015). Interactions of innate and adaptive immunity in brain development and function. *Brain Research*, *1617*, 18–27. https://doi.org/10.1016/j.brainres.2014.07.050

Franceschini, A., Lin, J., von Mering, C., & Jensen, L. J. (2016). SVD-phy: Improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, *32*(7), 1085–1087. https://doi.org/10.1093/bioinformatics/btv696

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., & Jensen, L. J. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, *41*(Database issue), D808–D815. https://doi.org/10.1093/nar/gks1094

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., Maathuis, M. H., Moreau, Y., Murphy, S. A., Przytycka, T. M., Rebhan, M., Röst, H., Schuppert, A., Schwab, M., Spang, R., Stekhoven, D., Sun, J., Weber, A., Ziemek, D., & Zupan, B. (2018). From hype to reality: Data science enabling personalized medicine. *BMC Medicine*, *16*(1), 150. https://doi.org/10.1186/s12916-018-1122-7

Frye, R. E., Delatorre, R., Taylor, H., Slattery, J., Melnyk, S., Chowdhury, N., & James, S. J. (2013). Redox metabolism abnormalities in autistic children associated with mitochondrial disease. *Translational Psychiatry*, *3*, e273. https://doi.org/10.1038/tp.2013.51

Gabriele, M., Lopez Tobon, A., D'Agostino, G., & Testa, G. (2018). The chromatin basis of neurodevelopmental disorders: Rethinking dysfunction along the molecular and temporal axes. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *84*(Pt B), 306–327. https://doi.org/10.1016/j.pnpbp.2017.12.013

Goetz, L. H., & Schork, N. J. (2018). Personalized medicine: Motivation, challenges, and progress. *Fertility and Sterility*, *109*(6), 952–963. https://doi.org/10.1016/j.fertnstert.2018.05.006

Gong, X., & Wang, H. (2015). SHANK1 and autism spectrum disorders. *Science China. Life Sciences*, *58*(10), 985–990. https://doi.org/10.1007/s11427-015-4892-6

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., & Hermjakob, H. (2020). ReactomeGSA – Efficient multi-omics comparative pathway analysis. *Molecular & Cellular Proteomics*, *19*(12), 2115–2125. https://doi.org/10.1074/mcp.TIR120.002155

Guang, S., Pang, N., Deng, X., Yang, L., He, F., Wu, L., Chen, C., Yin, F., & Peng, J. (2018). Synaptopathology involved in autism spectrum disorder. *Frontiers in Cellular Neuroscience*, *12*, 470. https://doi.org/10.3389/fncel.2018.00470

Guo, D., Yang, X., & Shi, L. (2020). Rho GTPase regulators and effectors in autism spectrum disorders: Animal models and insights for therapeutics. *Cell*, *9*(4), 835. https://doi.org/10.3390/cells9040835

Gurevich, E. V., Gainetdinov, R. R., & Gurevich, V. V. (2016). G protein-coupled receptor kinases as regulators of dopamine receptor functions. *Pharmacological Research*, *111*, 1–16. https://doi.org/10.1016/j.phrs.2016.05.010

Himmelstein, D. S., & Baranzini, S. E. (2015). Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes. *PLoS Computational Biology*, *11*(7), e1004259. https://doi.org/10.1371/journal.pcbi.1004259

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, *6*, e26726. https://doi.org/10.7554/eLife.26726

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Hu, J., Lepore, R., Dobson, R. J. B., Al-Chalabi, A., Bean, D. M., & Iacoangeli, A. (2021). DGLinker: Flexible knowledge-graph prediction of disease–gene associations. *Nucleic Acids Research*, *49*(W1), W153–W161. https://doi.org/10.1093/nar/gkab449

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Reviews. Genetics*, *6*(4), 287–298. https://doi.org/10.1038/nrg1578

Hussain, G., Wang, J., Rasul, A., Anwar, H., Imran, A., Qasim, M., Zafar, S., Kamran, S. K. S., Razzaq, A., Aziz, N., Ahmad, W., Shabbir, A., Iqbal, J., Baig, S. M., & Sun, T. (2019). Role of cholesterol and sphingolipids in brain development and neurological diseases. *Lipids in Health and Disease*, *18*(1), 26. https://doi.org/10.1186/s12944-019-0965-z

James, S. J., Melnyk, S., Jernigan, S., Cleves, M. A., Halsted, C. H., Wong, D. H., Cutler, P., Bock, K., Boris, M., Bradstreet, J. J., Baker, S. M., & Gaylor, D. W. (2006). Metabolic endophenotype and related genotypes are associated with oxidative stress in children with autism. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, *141B*(8), 947–956. https://doi.org/10.1002/ajmg.b.30366

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F. B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031

Kalkman, H. O. (2012). A review of the evidence for the canonical Wnt pathway in autism spectrum disorders. *Molecular Autism*, *3*(1), 10. https://doi.org/10.1186/2040-2392-3-10

Komiya, Y., & Habas, R. (2008). Wnt signal transduction pathways. *Organogenesis*, *4*(2), 68–75. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2634250/

Kreiman, B. L., & Boles, R. G. (2020). State of the art of genetic testing for patients with autism: A practical guide for clinicians. *Seminars in Pediatric Neurology*, *34*, 100804. https://doi.org/10.1016/j.spen.2020.100804

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., & Troyanskaya, O. G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, *19*, 1454–1462. https://doi.org/10.1038/nn.4353

Lai, E. S. K., Nakayama, H., Miyazaki, T., Nakazawa, T., Tabuchi, K., Hashimoto, K., Watanabe, M., & Kano, M. (2021). An autism-associated neuroligin-3 mutation affects developmental synapse elimination in the cerebellum. *Frontiers in Neural Circuits*, *15*, 58. https://doi.org/10.3389/fncir.2021.676891

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Leblond, C. S., Nava, C., Polge, A., Gauthier, J., Huguet, G., Lumbroso, S., Giuliano, F., Stordeur, C., Depienne, C., Mouzat, K., Pinto, D., Howe, J., Lemière, N., Durand, C. M., Guibert, J., Ey, E., Toro, R., Peyre, H., Mathieu, A., … Bourgeron, T. (2014). Meta-analysis of SHANK mutations in autism spectrum disorders: A gradient of severity in cognitive impairments. *PLoS Genetics*, *10*(9), e1004580. https://doi.org/10.1371/journal.pgen.1004580

Liang, X., Li, D., Song, M., Madden, A., Ding, Y., & Bu, Y. (2019). Predicting biomedical relationships using the knowledge and graph embedding cascade model. *PLoS One*, *14*(6), e0218264. https://doi.org/10.1371/journal.pone.0218264

Lionel, A. C., Vaags, A. K., Sato, D., Gazzellone, M. J., Mitchell, E. B., Chen, H. Y., Costain, G., Walker, S., Egger, G., Thiruvahindrapuram, B., Merico, D., Prasad, A., Anagnostou, E., Fombonne, E., Zwaigenbaum, L., Roberts, W., Szatmari, P., Fernandez, B. A., Georgieva, L., … Scherer, S. W. (2013). Rare exonic deletions implicate the synaptic organizer gephyrin (GPHN) in risk for autism, schizophrenia and seizures. *Human Molecular Genetics*, *22*(10), 2055–2066. https://doi.org/10.1093/hmg/ddt056

Lord, C., Brugha, T. S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E. J. H., Jones, R. M., Pickles, A., State, M. W., Taylor, J. L., & Veenstra-VanderWeele, J. (2020). Autism spectrum disorder. *Nature Reviews. Disease Primers*, *6*(1), 5. https://doi.org/10.1038/s41572-019-0138-4

Luo, P., Li, Y., Tian, L.-P., & Wu, F.-X. (2019). Enhancing the prediction of disease–gene associations with multimodal deep learning. *Bioinformatics*, *35*(19), 3735–3742. https://doi.org/10.1093/bioinformatics/btz155

Luo, P., Xiao, Q., Wei, P.-J., Liao, B., & Wu, F.-X. (2019). Identifying disease-gene associations with graph-regularized manifold learning. *Frontiers in Genetics*, *10*, 270. https://doi.org/10.3389/fgene.2019.00270

Luo, Y., Eran, A., Palmer, N., Avillach, P., Levy-Moonshine, A., Szolovits, P., & Kohane, I. S. (2020). A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nature Medicine*, *26*(9), 1375–1379. https://doi.org/10.1038/s41591-020-1007-0

Lutzu, S., & Castillo, P. E. (2021). Modulation of NMDA receptors by G-protein-coupled receptors: Role in synaptic transmission, plasticity and beyond. *Neuroscience*, *456*, 27–42. https://doi.org/10.1016/j.neuroscience.2020.02.019

Macedoni-Lukšič, M., Gosar, D., Bjørklund, G., Oražem, J., Kodrič, J., Lešnik-Musek, P., Zupančič, M., France-Štiglic, A., Sešek-Briški, A., Neubauer, D., & Osredkar, J. (2015). Levels of metals in the blood and specific porphyrins in the urine in children with autism spectrum disorders. *Biological Trace Element Research*, *163*(1), 2–10. https://doi.org/10.1007/s12011-014-0121-6

Marotta, R., Risoleo, M. C., Messina, G., Parisi, L., Carotenuto, M., Vetri, L., & Roccella, M. (2020). The neurochemistry of autism. *Brain Sciences*, *10*(3), E163. https://doi.org/10.3390/brainsci10030163

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., Leong, I. U. S., Smith, K. R., Gerasimenko, O., Haraldsdottir, E., Thomas, E., Scott, R. H., Baple, E., Tucci, A., Brittain, H., de Burca, A., Ibañez, K., Kasperaviciute, D., Smedley, D., … McDonagh, E. M. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics*, *51*(11), 1560–1565. https://doi.org/10.1038/s41588-019-0528-2

Martiniano, H. F. M. C., Asif, M., Vicente, A. M., & Correia, L. (2020). Network propagation-based semi-supervised identification of genes associated with autism spectrum disorder. In M. Raposo, P. Ribeiro, S. Sério, A. Staiano, & A. Ciaramella (Eds.), *Computational intelligence methods for bioinformatics and biostatistics* (pp. 239–248). Springer International Publishing. https://doi.org/10.1007/978-3-030-34585-3_21

Masi, A., Glozier, N., Dale, R., & Guastella, A. J. (2017). The immune system, cytokines, and biomarkers in autism spectrum disorder. *Neuroscience Bulletin*, *33*(2), 194–204. https://doi.org/10.1007/s12264-017-0103-8

Matta, S. M., Hill-Yardin, E. L., & Crack, P. J. (2019). The influence of neuroinflammation in autism spectrum disorder. *Brain, Behavior, and Immunity*, *79*, 75–90. https://doi.org/10.1016/j.bbi.2019.04.037

McCorvy, J. D., & Roth, B. L. (2015). Structure and function of serotonin G protein-coupled receptors. *Pharmacology & Therapeutics*, *150*, 129–142. https://doi.org/10.1016/j.pharmthera.2015.01.009

Mohamed, S. K., Nounu, A., & Nováček, V. (2021). Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics*, *22*(2), 1679–1693. https://doi.org/10.1093/bib/bbaa012

Mostafa, G. A., Bjørklund, G., Urbina, M. A., & Al-Ayadhi, L. Y. (2016). The levels of blood mercury and inflammatory-related neuropeptides in the serum are correlated in children with autism spectrum disorder. *Metabolic Brain Disease*, *31*(3), 593–599. https://doi.org/10.1007/s11011-015-9784-8

Mulligan, K. A., & Cheyette, B. N. R. (2016). Neurodevelopmental perspectives on Wnt signaling in psychiatry. *CXP*, *2*(4), 219–246. https://doi.org/10.1159/000453266

Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, *18*, 1414–1428. https://doi.org/10.1016/j.csbj.2020.05.017

Nunes, S., Sousa, R. T., & Pesquita, C. (2021). Predicting gene-disease associations with knowledge graph embeddings over multiple ontologies. *arXiv*: *2105.04944 [cs]*. http://arxiv.org/abs/2105.04944

Piñero, J., Bravo, l., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., & Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, *45*(D1), D833–D839. https://doi.org/10.1093/nar/gkw943

Piñero, J., Queralt-Rosinach, N., Bravo, L., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., & Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database: The Journal of Biological Databases and Curation*, *2015*, bav028. https://doi.org/10.1093/database/bav028

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, *48*(D1), D845–D855. https://doi.org/10.1093/nar/gkz1021

Ramaswami, G., & Geschwind, D. H. (2018). Genetics of autism spectrum disorder. *Handbook of Clinical Neurology*, *147*, 321–329. https://doi.org/10.1016/B978-0-444-63233-3.00021-X

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(W1), W191–W198. https://doi.org/10.1093/nar/gkz369

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L., Plon, S. E., Ramos, E. M., Sherry, S. T., & Watson, M. S. (2015). ClinGen – The clinical genome resource. *New England Journal of Medicine*, *372*(23), 2235–2242. https://doi.org/10.1056/NEJMsr1406261

Rossetti, G., Milli, L., & Cazabet, R. (2019). CDLIB: A python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, *4*(1), 1–26. https://doi.org/10.1007/s41109-019-0165-9

Savatt, J. M., & Myers, S. M. (2021). Genetic testing in neurodevelopmental disorders. *Frontiers in Pediatrics*, *9*, 52. https://doi.org/10.3389/fped.2021.526779

Sikora, D. M., Pettit-Kekel, K., Penfield, J., Merkens, L. S., & Steiner, R. D. (2006). The near universal presence of autism spectrum disorders in children with Smith-Lemli-Opitz syndrome. *American Journal of Medical Genetics. Part A*, *140*(14), 1511–1518. https://doi.org/10.1002/ajmg.a.31294

Smaili, F. Z., Gao, X., & Hoehndorf, R. (2019). OPA2Vec: Combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*, *35*(12), 2133–2140. https://doi.org/10.1093/bioinformatics/bty933

Snel, B., Lehmann, G., Bork, P., & Huynen, M. A. (2000). STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research*, *28*(18), 3442–3444. https://doi.org/10.1093/nar/28.18.3442

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & Von Mering, C. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*-(Database issue), D447–D452. https://doi.org/10.1093/nar/gku1003

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. V. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & Von Mering, C. (2021). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, *49*(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., & Von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, *45*(D1), D362–D368. https://doi.org/10.1093/nar/gkw937

Tamiji, J., & Crawford, D. A. (2010). The neurobiology of lipid metabolism in autism spectrum disorders. *Neurosignals*, *18*(2), 98–112. https://doi.org/10.1159/000323189

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Tromp, A., Mowry, B., & Giacomotto, J. (2021). Neurexins in autism and schizophrenia – A review of patient mutations, mouse models and potential future directions. *Molecular Psychiatry*, *26*(3), 747–760. https://doi.org/10.1038/s41380-020-00944-8

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 2071–2080). PMLR. http://proceedings.mlr.press/v48/trouillon16.html

Vicente, A. M., Ballensiefen, W., & Jönsson, J.-I. (2020). How personalised medicine will transform healthcare by 2030: The ICPerMed vision. *Journal of Translational Medicine*, *18*(1), 180. https://doi.org/10.1186/s12967-020-02316-w

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, *29*(12), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

Wang, T., Pan, Q., Lin, L., Szulwach, K. E., Song, C.-X., He, C., Wu, H., Warren, S. T., Jin, P., Duan, R., & Li, X. (2012). Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. *Human Molecular Genetics*, *21*(26), 5500–5510. https://doi.org/10.1093/hmg/dds394

Wang, X., Gong, Y., Yi, J., & Zhang, W. (2019). Predicting gene-disease associations from the heterogeneous network using graph embedding. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 504–511). https://doi.org/10.1109/BIBM47256.2019.8983134

Willets, J. M., Brighton, P. J., Mistry, R., Morris, G. E., Konje, J. C., & Challiss, R. A. J. (2009). Regulation of oxytocin receptor responsiveness by G protein-coupled receptor kinase 6 in human myometrial smooth muscle. *Molecular Endocrinology*, *23*(8), 1272–1280. https://doi.org/10.1210/me.2009-0047

Wu, G., & Haw, R. (2017). Functional interaction network construction and analysis for disease discovery. *Methods in Molecular Biology*, *1558*, 235–253. https://doi.org/10.1007/978-1-4939-6783-4_11

Yang, B., Yih, W.-T., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. *arXiv:1412.6575 [cs]*. http://arxiv.org/abs/1412.6575

Yao, P., Lin, P., Gokoolparsadh, A., Assareh, A., Thang, M. W. C., & Voineagu, I. (2015). Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nature Neuroscience*, *18*(8), 1168–1174. https://doi.org/10.1038/nn.4063

Yu, Z., Huang, F., Zhao, X., Xiao, W., & Zhang, W. (2021). Predicting drug–disease associations through layer attention graph convolutional network. *Briefings in Bioinformatics*, *22*(4), bbaa243. https://doi.org/10.1093/bib/bbaa243

Yui, K., Kawasaki, Y., Yamada, H., & Ogawa, S. (2016). Oxidative stress and nitric oxide in autism spectrum disorder and other neuropsychiatric disorders. *CNS & Neurological Disorders Drug Targets*, *15*(5), 587–596. https://doi.org/10.2174/1871527315666160413121751

Zheng, D., Song, X., Ma, C., Tan, Z., Ye, Z., Dong, J., Xiong, H., Zhang, Z., & Karypis, G. (2020). DGL-KE: Training knowledge graph embeddings at scale. In *Proceedings of the 43rd international ACM SIGIR conference on Research and Development in information retrieval* (pp. 739–748). Association for Computing Machinery.

AUTHOR BIOGRAPHIES

Joana Vilela has a Master degree in Biodiversity, Genetics and Evolution from the Faculty of Sciences of the University of Porto. Since 2017, she has been developing a PhD thesis at the Faculty of Sciences of the University of Lisbon, in the area of genetics of Autism Spectrum Disorder, carrying out her work in the Research and Development Unit (UID) of the Department of Health Promotion and Prevention of Non-Communicable Diseases from the National Institute of Health Doctor Ricardo Jorge. She is especially interested in the role of genes encoding synaptic proteins and neurotransmitters in neuronal development.

Muhammad Asif is a Bioinformatician and Data Analyst with a PhD in Systems Biology. His research is mainly focused on developing integrative and predictive machine learning based approaches. Currently, he is focusing on single cell omics and deep learning.

Ana Rita Marques received the B.Sc. degree in Applied Biology and M.Sc. degree in Health Sciences from University of Minho, Braga, Portugal, in 2009 and 2011, respectively. Currently, she is a Ph.D. candidate in Systems Biology at the Faculty of Sciences, University of Lisbon, Portugal and a research technician at the Biosystems and Integrative Sciences Institute (BioISI) and Instituto Nacional de Saúde Doutor Ricardo Jorge. Her main interests include genomics and bioinformatics, and her research involves the study of Autism Spectrum Disorder genetics, particularly focusing on regulatory mechanisms.

João Xavier Santos received the M.Sc. in Molecular Biology and Genetics from the Faculty of Sciences, University of Lisbon, Lisbon, Portugal in 2013. He is currently a PhD student in the Systems Biology program from BioISI and a technician at Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisbon, Portugal. His work focuses on how gene–environment interactions contribute to the onset and clinical heterogeneity of ASD.

Célia Rasga has a PhD in Cognitive Psychology from the Instituto de Psicologia Aplicada (ISPA) and Trinity College (Dublin, Ireland). Since 2015, she has been a postdoctoral at the Research and Development Unit, of the Department of Health Promotion and NCD Prevention of Instituto Nacional de Saúde Doutor Ricardo Jorge in the area of Neurodevelopment, with a special focus on interactions gene–environment in Autism Spectrum Disorder (ASD). She teaches the disciplines of Cognitive Psychology and Cognition and Affects, of the integrated master's program in Psychology, at ISPA.

Astrid Vicente is the Coordinator of the Department of Health Promotion and Prevention of Non-Transmissible Diseases (DPS) of Instituto Nacional de Saúde Doutor Ricardo Jorge, Senior Investigator in the field of Biomedicine and Public Health, Coordinator of the Research Group in Neurogenetics. She has a BcS in Biochemistry and PhD in Molecular Biology, is Associate Professor of the Faculty of Sciences of the University of Lisbon and coordinator of the Biomedical and Translation Research Group of the Institute of Biosystems and Integrative Sciences. Vice-Chair of the International Consortium for Personalized Medicine (ICPerMed), representative of Portugal and member of the Coordination Group of the 1M+ Genomes initiative. She is Principal Investigator or co-investigator of various national and international projects in the field of non-transmissible diseases and personalized medicine, member of the Psychiatric Genomics Consortium and the METASTROKE International Consortium. Her main areas of research are Personalized Medicine, Systems Medicine, Mental Health, Neurodevelopmental Disorders and Autism spectrum Disorder.

Hugo Martiniano is a researcher at Instituto Nacional de Saúde Doutor Ricardo Jorge. He received his degree in Chemistry from the Faculty of Sciences of the University of Lisbon in 2007 and his PhD degree in Physical Chemistry by the University of Lisbon in 2013. His research interests are centred on developing machine learning and data mining approaches to address problems in the areas of computational biology, bioinformatics, genomics and biomedicine. His current research focus lies in applying computational analysis techniques to understand the genetic and phenotypic heterogeneity in Autism Spectrum Disorder.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Vilela, J., Asif, M., Marques, A. R., Santos, J. X., Rasga, C., Vicente, A., & Martiniano, H. (2022). Biomedical knowledge graph embeddings for personalized medicine: Predicting disease-gene associations. *Expert Systems*, e13181. https://doi.org/10.1111/exsy.13181