

Automated drug repurposing workflow for rare diseases

Carmen A.T. Reep^{1,*}, Katherine Wolstencroft¹, Eleni Mina^{2,†} and Núria Queralt-Rosinach^{2,†}

¹The Leiden Institute of Advanced Computer Science (LIACS), Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

²Leiden University Medical Centre, Department of Human Genetics, Einthovenweg 20, 2333 ZC Leiden, The Netherlands

Abstract

There are over 7000 known rare diseases. Each one affects fewer than 1 in 2000 individuals, but collectively, they affect approximately 10% of the European and American populations. Developing treatment options for rare diseases is essential for those with such conditions, but as drug development is a time-consuming and costly process, developing new treatments is not often economically viable. The result is that fewer than 6% of rare disease have approved treatment options. The rare disease research community are adopting new approaches to this problem, where the focus is not on developing novel treatments, but on identifying approved drugs which could be repurposed to treat other conditions. These computational drug repurposing approaches require data and knowledge integration, to establish links between diseases, their symptoms, associated genes and drugs. Representing these concepts and relationships as a knowledge graph of machine-readable nodes and edges, enables predictions to be made about missing edges that may represent new drug target interactions.

In this study, we developed an automated computational drug repurposing workflow for rare diseases. The workflow integrates data mining and knowledge graph techniques, using the BioKnowledge Reviewer, together with state-of-the-art machine learning for link prediction, using graph embeddings and XGBoost. We demonstrate the utility of the workflow with a use-case in Huntington's disease, which is a rare neurodegenerative disorder of the central nervous system, caused by an elongated CAG repeat on the huntingtin gene. To evaluate the predictions made by the workflow, we manually explored the three top-ranked drug predictions for Huntington's disease. All three drugs are supported by evidence as plausible candidates. A similar analysis of Spinocerebellar ataxia type 1 (SCA1), a related neurodegenerative condition, yielded similarly promising results and showed the reproducibility of the method and workflow. The workflow is available at <https://github.com/carmenreep/DrugRepurposing>.

Keywords

drug repurposing, workflow, knowledge graph, rare diseases, Huntington's disease


SWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 26–29, 2024, Leiden, The Netherlands

✉ c.reep@erasmusmc.nl (C. A.T. Reep); k.j.wolstencroft@liacs.leidenuniv.nl (K. Wolstencroft); e.mina@lumc.nl (E. Mina); nqueralt.r@gmail.com (N. Queralt-Rosinach)

🆔 0000-0002-9408-7324 (C. A.T. Reep); 0000-0002-1279-5133 (K. Wolstencroft); 0000-0002-8972-9206 (E. Mina); 0000-0003-0169-8159 (N. Queralt-Rosinach)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

1. Background

Rare diseases are low-prevalent disorders caused by pathogenic mutations or harmful environmental factors that can have chronic, debilitating, or life-threatening effects [1]. Currently, there are over 7000 rare diseases that affect approximately 10% of the European and American populations, yet fewer than 6% have an approved treatment option [2]. This highlights the pressing need for developing therapies targeting rare diseases. However, the development of a new drug can be a time-consuming and costly process, taking up to 15 years and costing as much as US\$2.5 billion [2]. Consequently, the development of novel drugs for rare diseases, which affect only a small number of individuals, is not pursued frequently, as it is less likely to provide a return on investment for pharmaceutical companies [1]. A cost-efficient and faster way to provide drugs for rare diseases is via computational drug repurposing.

Drug repurposing is the process of identifying for an already approved or investigational drug a new use outside the scope of the original medical indication. For example, a drug could be repurposed for a different disease, based on the knowledge that drugs target particular pathways and disease mechanisms that may be shared by multiple diseases [3]. Computational drug repurposing aims to predict novel drug-disease associations, which can be achieved by predicting drug-target interactions (DTIs) [4]. The computational prediction of new DTIs can provide insights into potential pathological and drug mechanisms, as well as drug repurposing and design, helping researchers to generate testable hypotheses in the lab [5]. Network-based data integration and machine learning-based methods for DTIs prediction can mitigate costly and time consuming experimental verifications and are the current state of the art approaches in computational drug repurposing [6, 7, 8, 9, 10].

The landscape of biomedical information resources is heterogeneous and broad, yet most current methods for predicting DTIs are limited to homogeneous networks or bipartite models, failing to account for the intricate relationships among diverse data sources [6]. To fully exploit the potential of computational drug repurposing, we propose an automated workflow for predicting DTIs that does take complex relationships among diverse data sources into account. For our DTI prediction, we use the biophysical drug repurposing approach, which is based on the hypothesis that structurally similar drug molecules share similar targets. We extract biological data from multiple online databases using the BioKnowledge reviewer library (<https://github.com/SuLab/bioknowledge-reviewer>), a tool developed by Queralt-Rosinach et al. [11]. This library integrates heterogeneous knowledge and data into a knowledge graph, which is a machine-readable semantic representation of relational information, where concepts are encoded as nodes and relationships between them as edges. Now the prediction of DTIs can be framed as a link prediction problem, where the goal is to identify missing edges in the knowledge graph between drugs and targets that represent potential DTIs. To address this challenge, our proposed automated drug repurposing workflow leverages both network-based analysis and machine learning methods. Network-based methods help to identify potential interactions based on network topology and structural features, while machine learning methods can use more complex data features to make predictions. By combining these approaches, our automated workflow is able to discover new DTIs that can be further used by biologists to generate drug repurposing hypotheses that can be tested in the lab.

The drug repurposing workflow generalises to any rare disease, but as a proof-of-concept,

we focus on Huntington's disease (HD). HD is a rare neurodegenerative disorder of the central nervous system characterized by dementia, involuntary movements due to the movement disorder chorea and behavioural and psychiatric disturbances [12]. There are some symptomatic treatments available but because their effects are limited, there is a constant need for better, modifying drugs to treat symptoms of the disease [12].

We present the automated data mining workflow as a web application, using Flask, which makes the workflow accessible for researchers with no technical expertise. The Python code for running this app is accessible as a Docker container, available at <https://github.com/carmenreep/DrugRepurposing>, which runs on a laptop with minimal specifications (at least CPU 2.80 GHz).

2. Methods

2.1. Workflow steps

Figure 1 provides an overview of the proposed drug repurposing workflow. The workflow comprises four main steps: (1) creation of the knowledge graph, (2) embedding of the graph, (3) creation of edge representations, and (4) training of a supervised machine learning model. Besides finding missing edges (potential DTIs) in the knowledge graph, the machine learning model also predicts the interaction types of these missing edges. Good embeddings for this link prediction task were achieved through enriched information on the drugs and target sites.

2.2. Data sources

To obtain both human and animal biological data and metadata, the workflow uses the Monarch Biolink API version 1.1.14 (<https://api.monarchinitiative.org/api>). The Monarch Initiative, a collaborative, open science project, seeks to semantically integrate genotype-phenotype information from numerous sources and species [13]. To integrate drugs into the knowledge graph, the workflow utilizes the Drug-Gene Interaction Database (DGIdb), a web resource that aggregates information on drug-gene interactions and druggable genes from various sources, including publications, databases, and web-based resources [14]. We obtain drug-gene information from DGIdb using its API (version v2) available at <https://dgidb.org/api>.

2.3. Knowledge graph construction

The workflow leverages the BioKnowledge reviewer library to extract and integrate data from online sources into a knowledge graph. Starting with a list of seed nodes, a Monarch network is created by including the first layer of neighbours and relations from Monarch for each seed node, along with their ortholog-phenotype nodes. A seed node can take the form of a disease phenotype MIM number, such as '143100' representing Huntington's disease. The edges are formatted as triples, where each triple includes additional information such as the reference Uniform Resource Identifier (URI), the date when the information was obtained, and more information about the semantics of the relation. Nodes in the graph are identified using different biomedical ontologies in the OBO Foundry [<https://doi.org/10.1093/database/baab069>] maintained or used by Monarch and contain other attributes such as semantic group, URI, label,

Table 1

The three drug-gene interaction categories, with all interaction types that belong to each category. Source: https://www.dgldb.org/interaction_types.

category	ID	interaction types
inhibits	RO:0002408	antagonist, antibody, antisense oligonucleotide, blocker, cleavage, inhibitor, inhibitory allosteric modulator, inverse agonist, negative modulator, partial antagonist, suppressor
activates	RO:0002406	activator, agonist, chaperone, cofactor, inducer, partial agonist, positive modulator, stimulator, vaccine
regulates	RO:0011002	NA, None, n/a, other/unknown, adduct, allosteric modulator, binder, ligand, modulator, multitarget, potentiator, product of, substrate

name, synonyms, and description. The URI serves as a link to a web page that provides a more detailed description of the ontology term representing the node [11].

To obtain drug-target information, we use DGldb and take all genes (targets) in the Monarch graph as seeds. First, we need to map the Monarch genes to Entrez Gene identifiers (Entrez ID), which are used as a standard gene identifier system [15]. We accomplish this using the BioThings *MyGene.info* API, accessed with the Python wrapper *biothings_client* version v0.2.6 [16]. For each gene, we obtain a list of drugs (ID, name) that interact with the gene, along with the type of interaction and interaction source. The drug identifiers are from either the ChEMBL Database [17] or the Wikidata knowledge base [18]. There are various interaction types, such as 'activator', 'blocker', but to improve our predictions, we used the interaction direction (inhibits or activates) instead of the interaction types themselves [14]. Because some relations lack direction, we introduced a third category called 'regulates'. Table 1 shows each interaction direction category along with the interaction types belonging to that category. We mapped these three interaction groups to URIs using the OBO Relations Ontology (RO) <https://www.ebi.ac.uk/ols/ontologies/ro>, version 2022-05-23.

To enable the biophysical drug repurposing approach, it is necessary to identify structurally similar drugs. In our workflow, we use the Tanimoto coefficient [19] to measure the similarity between drugs. To achieve this, we first retrieve the SMILES chemical structure notation for each drug in our graph using the BioThings *MyGene.info* API accessed with the *biothings_client* version v0.2.6 [16]. Subsequently, we convert the SMILES structures into RDKit molecule objects using the *RDKit* Python package version 2022.3.2 with the *Chem* module [20]. The RDKit molecule objects are then transformed into Morgan fingerprints using the *GetMorganFingerprintAsBitVect()* function of the *AllChem* RDKit module. Using the *BulkTanimotoSimilarity()* function of the *DataStruct* module from *RDKit*, we can calculate Tanimoto coefficients between every possible pair of drugs in the graph. This results in a large number of weighted edges, which can lead to a complex network. To mitigate this, we adopt a method by Thafar et al. [21], where all similarity scores are ranked in descending order and only the top-10 most similar drugs are retained, similar to the k-nearest neighbours algorithm. Finally, we label all similarity edges with the 'CHEMINF:000481' ID, 'http://semanticscience.org/resource/CHEMINF_000481' URI, and the human readable string description 'similar to'.

The final graph is transformed into a Resource Description Framework (RDF) graph [22],

Table 2

All entity groups in the HD chorea KG graph with their SIO identifiers and description (source: Semanticscience Integrated Ontology (SIO) [23]). Column ‘count’ shows the number of nodes in each entity group.

entity	identifier	description	count
drug	SIO:010038	A drug is a chemical substance that contains one or more active ingredients that regulate one or more biological processes.	1352
gene	SIO:010035	A gene is part of a nucleic acid that contains all the necessary elements to encode a functional transcript.	284
disease	SIO:010299	A disease is the outward manifestation of one or more disorders.	194
genotype	SIO:001079	A genotype is a functional specification of a biological entity in terms of its genetic composition (or lack thereof).	127
variant	SIO:001381	A genomic sequence variant is part of a nucleic acid which is compositionally different than another reference genomic part.	106
phenotype	SIO:010056	A phenotype is an observable characteristic of an individual.	71
pathway	SIO:001107	A pathway is an effective specification that outlines a set of actions that forms a way to achieve an objective.	49

where each entity, relationship, and entity class (gene, drug, etc.) is represented as an ontology term by its URI. As Monarch did not provide identifiers for the entity classes, we manually mapped the entity class labels to terms in ontologies and used their URIs. To achieve this, we utilized the Semanticscience Integrated Ontology (SIO) [23] (<http://semanticscience.org/ontology/sio.owl>) version 1.53 and obtained URIs using the URI resolution service *identifiers.org* (<https://registry.identifiers.org/registry/sio>, accessed June 2022). Table 2 presents the specific URIs we used for the entity classes. To perform this transformation, we extended the BioKnowledge reviewer by using the RDFLib Python package version 6.1.1 [24] and ensured that the graph was stored in Turtle format [25].

2.4. Graph embedding

To prepare the knowledge graph for embedding, we first remove all known drug-gene interactions from the graph. This is important to prevent bias in the prediction task, as keeping these edges would make the embedding vectors of the drugs very similar to the embedding vectors of the genes they interact with. We therefore split the graph into two separate graphs, one for drugs and one for genes. Each graph is then embedded separately using a graph embedding algorithm. After the embedding, the drug vectors are fused with the gene vectors to obtain drug-gene edges, which are used for training the machine learning model, as explained in more detail in the *XGBoost prediction model* section below.

Celebi et al. [26] compared different knowledge graph embedding methods for drug-drug interaction prediction, and found that RDF2Vec with Skip-Gram generally outperforms other

methods. Therefore, this workflow employs RDF2Vec for graph embedding. RDF2Vec adapts the language modelling approach of Word2Vec to RDF graph embeddings [26]. First, random walks are performed over the graph to generate sequences of entities and relations. Then, the Skip-Gram model is used to learn one embedding for each entity/relationship in the graph. After training, semantically and syntactically similar entities/relationships have similar embeddings [26]. For the prediction task in this study, only the drug and gene vectors are of interest, and therefore, only these vectors were selected for further computation.

For RDF2Vec, the Python function *RDF2VecTransformer* from the *rdf2vec* module of the package *pyrdf2vec* version 0.2.3 is used [27]. The maximum depth of one walk is set to 4 and for each entity in the graph, the maximum number of walks is set to 10.

2.5. Fused embeddings for link prediction

To train a supervised machine learning model, it is essential to have both positive and negative samples of data [26]. The positive samples are all known interactions (regulates, inhibits, or activates). The negative samples can be obtained from unknown interactions between drugs and genes. Edge embeddings for positive and negative samples are generated by adopting a node embedding fusion approach. For every possible drug-gene combination, we obtain one embedding by fusing the drug embedding and the gene embedding with the Hadamard operator, which is a strong operator for learning edge features in link prediction tasks [28]. We then add the class of the interaction (inhibits, activates, or regulates) to the resulting embedding. For edges that do not exist in the graph, we assign the label "unknown" to represent the unknown interaction class.

The prediction data for our machine learning model includes all unknown drug-gene interactions involving genes that contribute to the disease phenotype of interest. The negative samples of the training data are all unknown interactions that are not the prediction data. However, the number of negative samples significantly outweighs the number of positive samples. Including all of these negative samples could result in data imbalance and affect the performance of our model [26]. To address this issue, we decided to downsample the negative cases by randomly selecting negative samples with a sample size equal to the class in the positive set with the largest number of interactions (regulates, inhibits, or activates).

2.6. XGBoost prediction model

For our machine learning model, we utilized the XGBoost classifier proposed by Thafar et al. in 2021 [21]. To implement the model, we used the *XGBClassifier()* function from the Python package *xgboost* (version 1.3.3) [29]. We set the learning objective to 'multi:softmax', which allows XGBoost to optimize the likelihood of each class label and assign a probability to each possible class.

To address minor class imbalance in our positive sample, we computed sample weights using the *compute_sample_weight()* function from *sklearn* [30] version 1.1.1. These weights are then used for training the model, which provides some bias towards the minority classes during training.

To optimize the hyperparameters of our model, we conducted a randomized search on

Table 3

The hyperparameter search space of the XGBoost classifier, with descriptions of each parameter. Source: <https://xgboost.readthedocs.io/en/stable/parameter.html>. `uniform(a,b)` indicates a uniform distribution on (a,b).

parameter	description	search space
<code>min_child_weight</code>	Minimum sum of instance weight (hessian) needed in a child.	2, 3, 5, 8, 13, 20, 30
<code>gamma</code>	Minimum loss reduction required to make a further partition on a leaf node of the tree.	0, 0.2, 0.5, 0.8, 1.2, 1.6, 2.5, 4, 6
<code>reg_alpha</code>	L2 regularization term on weights.	0, 0.5, 1, 3, 5, 10
<code>reg_lambda</code>	L1 regularization term on weights.	0, 0.5, 1, 3, 5, 10
<code>subsample</code>	Subsample ratio of the training instances.	<code>uniform(0.5, 1)</code>
<code>colsample_bytree</code>	Subsample ratio of columns when constructing each tree.	<code>uniform(0.2, 1)</code>
<code>max_depth</code>	Maximum depth of a tree.	4, 6, 8, 10, 12, 14, 16
<code>n_estimators</code>	Number of boosting rounds.	35, 45, 50, 70, 80, 90, 100
<code>learning_rate</code>	Step size shrinkage used after each boosting step to prevent overfitting.	<code>uniform(0, 0.3)</code>

the search space presented in Table 3 using the `RandomizedSearchCV()` function from the `model_selection` module of the `sklearn` Python package (version 1.1.1) [30]. We set the number of parameter setting combinations to be tested (`n_iter`) to 20.

Given the challenge of identifying negative examples of drug-target pairs, as unlinked drugs and targets may simply represent drug-target pairs that have not been identified yet, we opted against conducting an error analysis. Instead, the model performance is assessed using the repeated stratified k-fold cross-validation technique alongside the F1-score metric.. The number of subsets for the k-fold cross validation is set to 10 and the number of repeats is set to 5. During each iteration of the process, the F1-score is calculated and averaged for each class X. Finally, the average F1-score over all k iterations and number of repeats is computed to obtain the final evaluation metric.

The best hyperparameters are used to build the final model. The confidence of each prediction is obtained using the `predict_proba()` function of `xgboost` version 1.3.3, which returns the probability of an interaction belonging to its predicted class.

2.7. DTI ranking and validation

For every gene in the graph that is associated with the symptom of interest, an interaction type and score is predicted for every drug in the dataset, provided that this interaction does not exist in the graph. To prioritize the most promising drug candidates for further investigation, we perform a ranking step based on the predicted positive interactions. First, we remove all predictions with a confidence score lower than 0.9 to focus only on the most confident predictions. Next, we rank the drugs based on the number of positive interactions they have with the genes associated with the symptom. This ranking approach is based on the hypothesis

that drugs with more positive interactions with genes that cause a symptom are more likely to be effective in alleviating that symptom. In the case of drugs with the same number of interactions, we use the sum of prediction confidence scores as a secondary ranking criterion, with drugs having higher sums being ranked higher.

3. Results

Our workflow was initially run with the terms "huntington's disease" (OMIM number '143100') and "chorea" ('HP:0002072') as seeds representing the disease and symptom fields respectively, for constructing the knowledge graph. The graph was created on 2022-06-27 and has in total 2189 nodes and 17467 edges. Figure 2 provides an overview of all entities and relationships in the graph and Table 4 shows the identifiers and descriptions of each relation between nodes in this graph. The graph includes 1352 drugs and 284 genes, resulting in a total of 383,968 edge representations, of which 1753 are known (1301 regulates, 391 inhibits, and 61 activates). This graph has 200 genes that are associated with the symptom chorea, which are the genes of interest, and there are 1077 known drug-gene edges with these 200 genes, indicating that the prediction data consists of 269,323 unknown interactions of potential interest. To deal with the imbalance between the larger number of negative samples and the comparatively smaller number of positive samples, 1301 negative samples were randomly selected, to balance the number of the largest interactions in the positive class (regulates). The best XGBoost hyperparameters can be found in Table 5, and the F1 score with these hyperparameters is 0.867. The trained XGBoost model was used to predict the classes of the unknown drug-gene interactions of interest.

Table 6 shows the predicted top ten ranked drugs that interact with genes that are associated with the phenotype chorea. We manually explored the top ranked predictions for HD that are associated with chorea. Below we present the top three candidates. Table 7 presents the two highest ranked drugs and the genes that these drugs have a positive predicted interaction with. The top predicted ranked drug is CHEMBL29097. CHEMBL29097 (synonym MK-886) is an inhibitor of 5-lipoxygenase-activating protein activity, currently in preclinical phase. It has been found that 10 microM MK-886 can abolish the biosynthetic production of cysteinyl leukotrienes (CysLTs), which is suggested to be involved in brain inflammation and neurological diseases [33]. In addition to its anti-inflammatory activity, MK-886 has been shown to activate the proteasome which is known to have a causative role in HD [34]. Impaired function of the proteasome leads to the formation of intracellular aggregates in the nucleus as the proteasome cannot clear efficiently misfolded huntingtin proteins [35].

The second highest ranked drug is baicalein. Baicalein (CHEMBL8260) is a flavonoid isolated from the traditional Chinese medicinal herb *Scutellaria baicalensis* Georgi, currently on Phase 2. Baicalein has known anti-inflammatory and neuroprotective efficacy in neurodegenerative disease models [36]. Rui et al. [36] studied the effects of baicalein on inflammasome-induced neuroinflammation in Parkinson's disease (PD) and found that baicalein can suppresses MPTP-induced nigral dopaminergic neuron death, glial activation, and motor dysfunction in mice by suppressing the NLRP3/caspase-1/GSDMD pathway. In addition, several studies have demonstrated that baicalein protects neurons in animal models of Alzheimer's disease (AD) and

PD by inhibiting neuroinflammation [37].

Amphotericin b (CHEMBL267345) was another drug on our list that ranked very high. Amphotericin b is an approved antifungal drug used to treat serious fungal infections. Experimental evidence shows that some antibiotics and antifungal medication have neuroprotective action through anti-aggregating activity on disease-associated proteins [38]. Although this drug has been shown to cause a delay in the formation of amyloid- β , it was also found to induce toxicity [39]. However, Soler et al., [40] developed a derivative of amphotericin that has anti-aggregating action but lacks toxicity and antimicrobial activity [38].

3.1. Other rare diseases

To demonstrate the reusability of our approach, we also applied our methodology to another rare disease that currently lacks treatment; Spinocerebellar ataxia type 1 (SCA1). We used as seeds the terms "SCA1" (OMIM number '164400') and the symptom "hyporeflexia" (HP:0001265) to run our drug repurposing workflow and below we describe few of the top hits.

The first prioritized drug by our workflow was Dovitinib (CHEMBL522892) currently in phase 3. Dovitinib is a pan receptor tyrosine kinase (RTK) inhibitor that has anti-tumor activity in pre clinical models of several cancers [41]. It has been recently suggested as a candidate treatment for AD because it normalizes β amyloid mediated transcriptional responses by targeting the CREB3L2-ATF4 heterodimerization which is responsible for the majority of the transcriptional changes occurring in AD neurons [42]. Its well tolerated safety profile and the ability to cross the blood brain barrier [42] makes it an interesting candidate for AD but also potentially for other neurodegenerative disorders that exhibit similar disease mediated changes like AD.

The second predicted drug on the list was broquinolol (CHEMBL1394319), a small molecule that has antifungal and antibacterial activity. This is an investigational drug that was found to have activity against thyroid cancer in a high throughput screening experiment [43]. However, there is currently no evidence for being associated with neurodegenerative diseases.

Number three on the candidate drug list for SCA1 was an interesting compound, astemizole (CHEMBL296419). Astemizole is an approved second generation antihistamine drug [44] that has been found to rescue motor phenotype in a *Drosophila* model of PD [45].

4. Discussion

This work presents a novel disease-drug profiling approach to identifying potential candidate compounds that could alleviate the symptoms of a rare disease. It combines two established, and widely accepted approaches, of mining rare disease-specific data from multiple public databases into a knowledge graph [11], and graph-based machine learning approaches to identifying drug-target interactions [21]. The result is an automated workflow which makes disease-drug profile predictions targeted to specific rare diseases. We demonstrated its utility using predictions from Huntington's disease and SCA1.

The advance that this work provides to the field is the use of rare disease specific knowledge graphs. Using BioKnowledge reviewer in a drug repurposing automated workflow enables to learn from a comprehensive view of the underlying druggable rare disease biology and pathogenesis of interest. This is advantageous over current integration methods used in rare

disease research, which use information about thousands of complex disorders [2], because it leverages knowledge for precision medicine. Another advantage is that by comparing to existing solutions [8, 9, 10], our method harnesses heterogeneous and expressive semantic graphs for DTI prediction beyond bipartite networks. Integrating new types of entities with Semantic Web technologies enables us to represent more complex relations around drugs and targets, and it opens the possibility of learning from them and exploiting the semantics by means of methods such as RDF2Vec graph embedding methods.

Through sophisticated graph-based algorithms, we can traverse the knowledge graph to identify patterns in the data and predict potential new relationships between drugs and diseases. We demonstrated that knowledge graphs and graph machine learning used streamlined in an automated workflow gives testable DTI hypotheses for drug repurposing in the rare disease area. This can support researchers to systematically generate compound prioritization coupled with well-designed validation experiments to discover treatments for rare diseases in a timely and cost-effective manner. One limitation is that we did not integrate domain expert knowledge on disease pathobiology with patient data, which can be the basis for highly innovative drugs. While [11] gave a solution to include expert knowledge in graphs, access to patient data is a serious problem in health research. However, projects such as the EJP-RD¹ are providing Semantic Web based solutions for patient data sharing.

Our results provide some interesting candidates that could potentially be of great value for the rare disease community. Some of our prioritized drugs are already associated with other neurodegenerative disorders (AD and PD) targeting neuroinflammation, which is a hallmark of the HD pathology. Other candidates (Broquinadol and Amphotericin b) belong to the class of antifungal and antibacterial medication. These types of drugs have drawn a lot of attention and although they are mainly used to treat infections new applications are being discovered. It has been reported that antibiotics, for example Doxycycline and minocycline, have neuroprotective effects due to their anti-inflammatory properties [46].

The workflow is presented as a web application and yields promising results in a reusable and reproducible way for the rare diseases community. In the future, it could be extended and improved by the addition of experimentally validated negative interactions from reliable databases. Our current approach uses unknown drug-gene pairs as negative samples for classification. It is therefore possible that this set includes currently unknown positive interactions, which may adversely affect model training [47]. Additionally, the knowledge graph could be extended to include further information about each drug, such as side-effects and drug interactions, and additional input seeds could be obtained from the Monarch database, such as genes and metabolites associated with the particular disease, or related diseases. Moreover, integrating the predicted embeddings into the knowledge graph would enable the evaluation of performance across diverse prediction methods, offering valuable insights into model efficacy and versatility. Lastly, it is important to note that the RDF graphs are stored in turtle format at the location where the code is executed, to enable further use in code execution and analysis. The RDF graph is currently not served as a live RDF/SPARQL endpoint, which presents an opportunity for improvement in our approach. This aligns with broader efforts in the field to enhance the transparency and accessibility of machine learning outcomes.

¹<https://www.ejprarediseases.org/>

5. Conclusion

Integrating data into knowledge graphs with state-of-the-art graph-based machine learning methods results in a novel automated drug repurposing workflow, specifically suited for rare diseases, where data tends to be sparse and distributed. The workflow produces a ranked list of candidate compounds, which serve as new hypotheses for drug treatments. The drug repurposing workflow generalises to any rare disease, but as a proof-of-concept, we focused on Huntington's disease, and a related condition, SCA1. We identified several promising candidate drugs for Huntington's Disease for the symptom chorea, demonstrating the potential of our approach. With further testing and validation, these candidates could be explored as potential treatments for the disease. The workflow is provided as a web application, in a publicly available Docker container. It is therefore accessible for researchers with no technical expertise and is a reusable and reproducible application for the rare disease community.

References

- [1] S. Shah, M. M. Dooms, S. Amaral-Garcia, M. Igoillo-Esteve, Current drug repurposing strategies for rare neurodegenerative disorders, *Frontiers in Pharmacology* 12 (2021). doi:10.3389/fphar.2021.768023.
- [2] H. I. Roessler, N. V. Knoers, M. M. van Haelst, G. van Haften, Drug repurposing for rare diseases, *Trends in Pharmacological Sciences* 42 (2021) 255–267. doi:10.1016/j.tips.2021.01.003.
- [3] T. B. Malas, W. J. Vlietstra, R. Kudrin, S. Starikov, M. Charrouf, M. Roos, D. J. M. Peters, J. A. Kors, R. Vos, P. A. C. 't Hoen, E. M. van Mulligen, K. M. Hettne, Drug prioritization using the semantic properties of a knowledge graph, *Scientific Reports* 9 (2019). doi:10.1038/s41598-019-42806-6.
- [4] M. D. Paranjpe, A. Taubes, M. Sirota, Insights into computational drug repurposing for neurodegenerative disease, *Trends in Pharmacological Sciences* 40 (2019) 565–576. doi:10.1016/j.tips.2019.06.003.
- [5] W. Ba-alawi, O. Soufan, M. Essack, P. Kalnis, V. B. Bajic, Daspfind: new efficient method to predict drug–target interactions, *Journal of Cheminformatics* 8 (2016) 1758–2946. doi:10.1186/s13321-016-0128-4.
- [6] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, *Nature Communications* 8 (2017) 2041–1723. doi:10.1038/s41467-017-00680-8.
- [7] C. Chen, H. Shi, Z. Jiang, A. Salhi, R. Chen, X. Cui, B. Yu, Dnn-dtis: Improved drug-target interactions prediction using xgboost feature selection and deep neural network, *Computers in Biology and Medicine* 136 (2021) 104676. doi:https://doi.org/10.1016/j.combiomed.2021.104676.
- [8] K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao, J. Sun, DeepPurpose: a deep learning library for drug–target interaction prediction, *Bioinformatics* 36 (2020) 5545–5547. doi:10.1093/bioinformatics/btaa1005.

- [9] Y. Kalakoti, S. Yadav, D. Sundar, Deep neural network-assisted drug recommendation systems for identifying potential drug–target interactions, *American Chemical Society* 7 (2022) 12138–12146. doi:10.1021/acsomega.2c00424.
- [10] E. Amiri Souri, R. Laddach, S. N. Karagiannis, L. G. Papageorgiou, S. Tsoka, Novel drug–target interactions via link prediction and network embedding, *BMC Bioinformatics* 23 (2022) 1471–2105. doi:10.1186/s12859-022-04650-w.
- [11] N. Queralt-Rosinach, G. S. Stupp, T. S. Li, M. Mayers, M. E. Hoatlin, M. Might, B. M. Good, A. I. Su, Structured reviews for data and knowledge-driven research, *Database* 2020 (2020). doi:10.1093/database/baaa015.
- [12] R. A. Roos, Huntington's disease: a clinical review, *Orphanet Journal of Rare Diseases* 5 (2010). doi:10.1186/1750-1172-5-40.
- [13] C. J. Mungall, J. A. McMurry, S. Köhler, J. P. Balhoff, C. Borromeo, M. Brush, S. Carbon, T. Conlin, N. Dunn, M. Engelstad, E. Foster, J. Gourdine, J. O. Jacobsen, D. Keith, B. Laraway, S. E. Lewis, J. NguyenXuan, K. Shefchek, N. Vasilevsky, Z. Yuan, N. Washington, H. Hochheiser, T. Groza, D. Smedley, P. N. Robinson, M. A. Haendel, The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species, *Nucleic Acids Research* 45 (2016) D712–D722. doi:10.1093/nar/gkw1128.
- [14] S. L. Freshour, S. Kiwala, K. C. Cotto, A. C. Coffman, J. F. McMichael, J. J. Song, M. Griffith, O. L. Griffith, A. H. Wagner, Integration of the drug–gene interaction database (DGIdb 4.0) with open crowdsourcing efforts, *Nucleic Acids Research* 49 (2020) D1144–D1151. doi:10.1093/nar/gkaa1084.
- [15] D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova, Entrez gene: gene-centered information at NCBI, *Nucleic Acids Research* 35 (2007) D26–D31. doi:10.1093/nar/gkl1993.
- [16] C. Wu, biotings client, 2022. URL: https://github.com/biotings/biotings_client.py.
- [17] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, ChEMBL: towards direct deposition of bioassay data, *Nucleic Acids Res.* 47 (2019) D930–D940. doi:10.1093/nar/gky1075.
- [18] Wikimedia Foundation, Wikidata, 2022. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page.
- [19] A. Kumar, Chemical similarity methods : A tutorial review, *The Chemical Educator* (2011) 46–50. doi:10.1333/s00897112344a.
- [20] Rdkit: Open-source cheminformatics, 2022. URL: <http://www.rdkit.org>.
- [21] M. A. Thafar, R. S. Olayan, S. Albaradei, V. B. Bajic, T. Gojobori, M. Essack, X. Gao, DTi2vec: Drug–target interaction prediction using network embedding and ensemble learning, *Journal of Cheminformatics* 13 (2021). doi:10.1186/s13321-021-00552-w.
- [22] World Wide Web Consortium, Resource Description Framework (RDF): Concepts and Abstract Syntax, 2014. URL: <https://www.w3.org/TR/rdf11-concepts/>, w3C Recommendation.
- [23] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. P. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, R. Hoehndorf, The semantic science integrated ontology (SIO) for biomedical research and knowledge discovery, *J. Biomed. Semantics* 5 (2014) 14. doi:10.1186/2041-1480-5-14.

- [24] I. Aucamp, RdfLib, 2021. URL: <https://github.com/RDFLib/rdfLib>.
- [25] RDF 1.1 turtle, 2014. URL: <https://www.w3.org/TR/turtle/>.
- [26] R. Celebi, H. Uyar, E. Yasar, O. Gumus, O. Dikenelli, M. Dumontier, Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings, *BMC Bioinformatics* 20 (2019). doi:10.1186/s12859-019-3284-5.
- [27] G. Vandewiele, B. Steenwinckel, T. Agozzino, M. Weyns, P. Bonte, F. Ongenaes, F. D. Turck, pyRDF2Vec: Python Implementation and Extension of RDF2Vec, IDLab, 2020. URL: <https://github.com/IBCNServices/pyRDF2Vec>.
- [28] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, 2016. doi:10.48550/ARXIV.1607.00653.
- [29] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, Scikit-learn: Machine learning in python, *Journal of Machine Learning Research* 12 (2011) 2825–2830. URL: <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [31] C. Mungall, J. A. Overton, D. Osumi-Sutherland, M. Haendel, Mbrush, Obo-relations: 2015-10-29 release, 2015. doi:10.5281/ZENODO.32899.
- [32] M. Brush, N. Matentzoglou, M. Haendel, Geno-ontology, 2022. URL: <https://github.com/monarch-initiative/GENO-ontology>.
- [33] P. Ballerini, P. D. Iorio, R. Ciccarelli, F. Caciagli, A. Polp, A. Beraudi, S. Buccella, I. D'Alimonte, M. D'Auro, E. Nargi, P. Patricelli, D. Visini, U. Traversa, P2y₁ and cysteinyl leukotriene receptors mediate purine and cysteinyl leukotriene co-release in primary cultures of rat microglia, *International Journal of Immunopathology and Pharmacology* 18 (2005) 255–268. doi:10.1177/039463200501800208.
- [34] E. E. Liao, M. Yang, N. Nathan Kochen, N. Vunnam, A. R. Braun, D. M. Ferguson, J. N. Sachs, Proteasomal stimulation by mk886 and its derivatives can rescue tau-induced neurite pathology, *Molecular neurobiology* 60 (2023) 6133–6144. doi:10.1007/s12035-023-03417-5.
- [35] T. R. Soares, S. D. Reis, B. R. Pinho, M. R. Duchen, J. M. Oliveira, Targeting the proteostasis network in huntington's disease, *Ageing Research Reviews* 49 (2019) 92–103. doi:10.1016/j.arr.2018.11.006.
- [36] W. Rui, S. Li, H. Xiao, M. Xiao, J. Shi, Baicalein attenuates neuroinflammation by inhibiting NLRP3/caspase-1/GSDMD pathway in MPTP induced mice model of parkinson's disease, *Int. J. Neuropsychopharmacol.* 23 (2020) 762–773. doi:10.1093/ijnp/pyaa060.
- [37] Y. Li, J. Zhao, C. Hölscher, Therapeutic potential of baicalein in alzheimer's disease and parkinson's disease, *CNS drugs* 31 (2017) 639–652. doi:10.1007/s40263-017-0451-y.
- [38] S. B. Socias, F. González-Lizárraga, C. L. Avila, C. Vera, L. Acuña, J. E. Sepulveda-Diaz, E. Del-Bel, R. Raisman-Vozari, R. N. Chehin, Exploiting the therapeutic potential of ready-to-use drugs: Repurposing antibiotics against amyloid aggregation in neurodegenerative diseases, *Progress in Neurobiology* 162 (2018) 17–36. doi:10.1016/j.pneurobio.2017.12.002.
- [39] F. Durães, M. Pinto, E. Sousa, Old drugs as new treatments for neurodegenerative diseases,

Pharmaceuticals 11 (2018). doi:10.3390/ph11020044.

- [40] L. Soler, P. Caffrey, H. E. McMahon, Effects of new amphotericin analogues on the scrapie isoform of the prion protein, *Biochimica et Biophysica Acta (BBA) - General Subjects* 1780 (2008) 1162–1167. doi:<https://doi.org/10.1016/j.bbagen.2008.07.005>.
- [41] S. S. Yadav, J. Li, J. A. Stockert, B. Herzog, J. O'Connor, L. Garzon-Manco, R. Parsons, A. K. Tewari, K. K. Yadav, Induction of neuroendocrine differentiation in prostate cancer cells by dovitinib (tki-258) and its therapeutic implications, *Translational Oncology* 10 (2017) 357–366. doi:<https://doi.org/10.1016/j.tranon.2017.01.011>.
- [42] C. G. Roque, K. M. Chung, E. P. McCurdy, R. Jagannathan, L. K. Randolph, K. Herline-Killian, J. Baleriola, U. Hengst, Creb3l2-atf4 heterodimerization defines a transcriptional hub of alzheimer's disease gene expression linked to neuropathology, *Science Advances* 9 (2023) eadd2671. doi:10.1126/sciadv.add2671.
- [43] L. Zhang, M. He, Y. Zhang, N. Nilubol, M. Shen, E. Kebebew, Quantitative High-Throughput Drug Screening Identifies Novel Classes of Drugs with Anticancer Activity in Thyroid Cancer Cells: Opportunities for Repurposing, *The Journal of Clinical Endocrinology Metabolism* 97 (2012) E319–E328. doi:10.1210/jc.2011-2671. arXiv:<https://academic.oup.com/jcem/article-pdf/97/3/E319/10416587/jcemE319.pdf>.
- [44] P. M. Krstenansky, J. Robert J. Cluxton, Astemizole: A long-acting, non-sedating antihistamine, *Drug Intelligence & Clinical Pharmacy* 21 (1987) 947–953. doi:10.1177/106002808702101202.
- [45] K. Styczyńska-Soczka, L. Zechini, L. Zografos, Validating the predicted effect of astemizole and ketoconazole using a drosophila model of parkinson's disease, *Assay and drug development technologies* 15 (2017) 106–112.
- [46] A. Dominguez-Meijide, V. Parrales, E. Vasili, F. González-Lizárraga, A. König, D. F. Lázaro, A. Lannuzel, S. Haik, E. Del Bel, R. Chehín, R. Raisman-Vozari, P. P. Michel, N. Bizat, T. F. Outeiro, Doxycycline inhibits α -synuclein-associated pathologies in vitro and in vivo, *Neurobiology of Disease* 151 (2021) 105256. doi:<https://doi.org/10.1016/j.nbd.2021.105256>.
- [47] L. Xu, X. Ru, R. Song, Application of machine learning for drug–target interaction prediction, *Frontiers in Genetics* 12 (2021). URL: <https://doi.org/10.3389/fgene.2021.680117>. doi:10.3389/fgene.2021.680117.

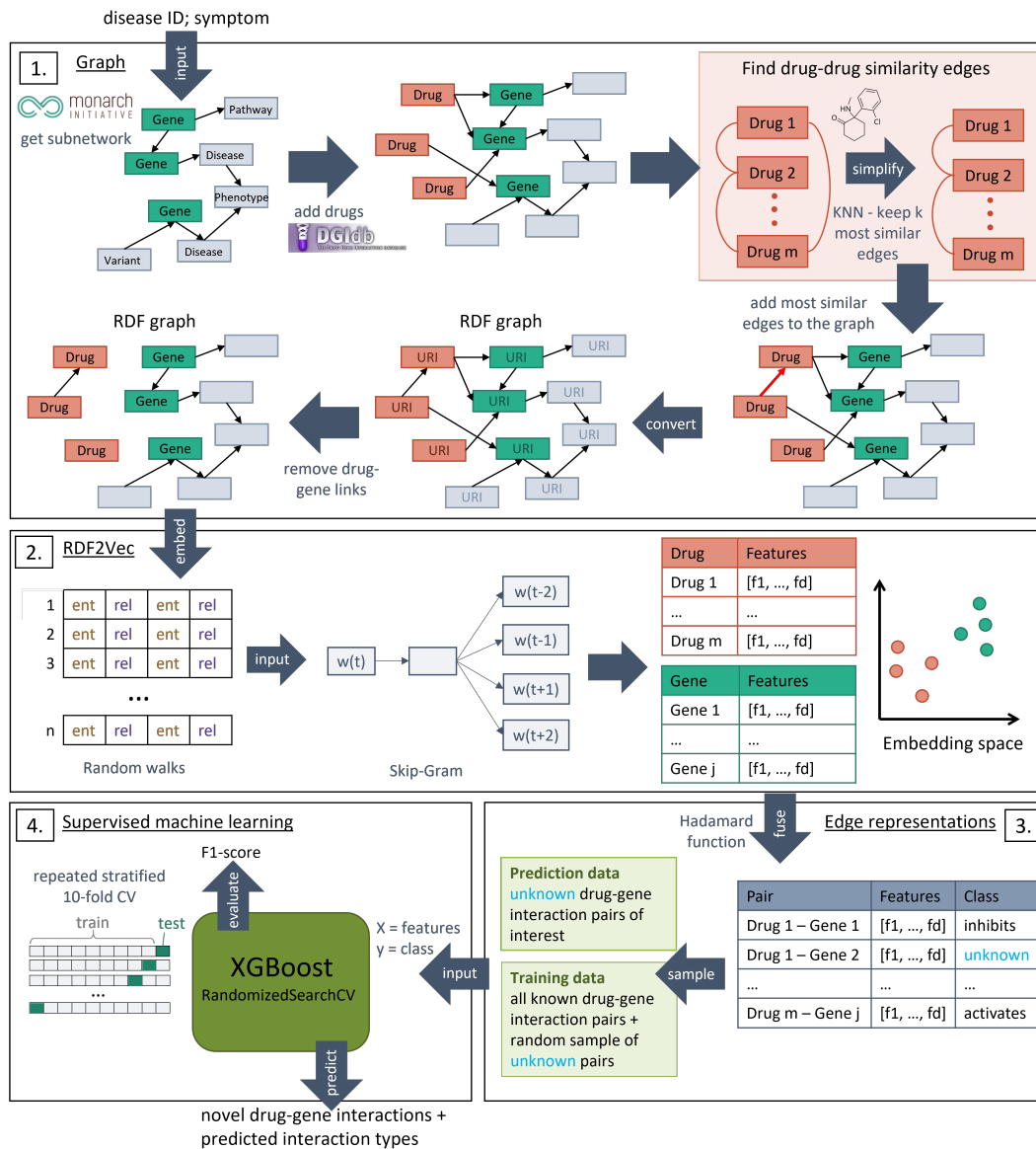


Figure 1: The drug repurposing workflow. (1.) It takes as input a disease and symptom of interest, then constructs the knowledge graph using Monarch and DGIdb, adds drug-drug similarity edges based on SMILES compound structure, turns the graph into an RDF graph and removes drug-gene links ready for embedding. (2.) It then applies RDF2Vec, which creates random walks over the graph for each entity, trains a skip-gram model and outputs one feature vector for each entity in the graph. (3.) Next, it generates edge representations for each drug-gene pair and turns this into prediction and training data. (4.) Then it trains an XGBoost classification model using the prediction data, finds the best model by hyperparameter tuning using randomized search, evaluates using repeated stratified 10-fold CV and uses the best found model to predict the interactions of interest.

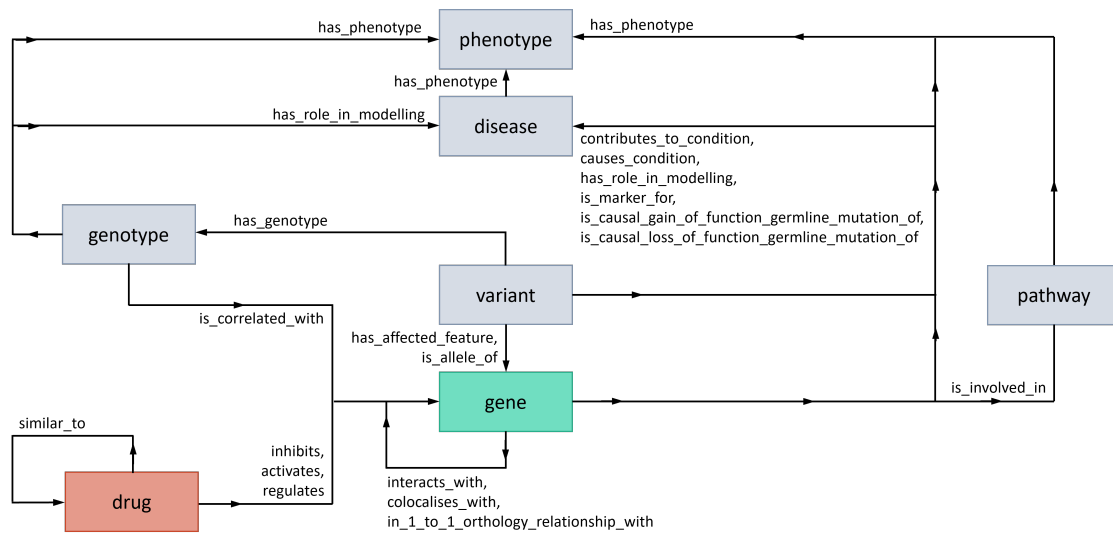


Figure 2: Data model of the HD chorea graph. Overview of all entities and relations in the HD chorea knowledge graph.

Table 4

All relations in the HD chorea KG with their identifiers and description (sources: OBO Relations Ontology [31]; GENO ontology [32]). Column 'count' shows the number of edges in each relation group.

relation	identifier	description	count
similar to	CHEM-INF:000481	Connects a molecular entity that is deemed similar to another according to some algorithm.	13023
regulates	RO:0011002	The entity x has an activity that regulates an activity of the entity y .	1301
has phenotype	RO:0002200	A relationship that holds between a biological entity and a phenotype. Here a phenotype is construed broadly as any kind of quality of an organism part, a collection of these qualities, or a change in quality or qualities. The subject of this relationship can be an organism, a genomic entity such as a gene or genotype, or a condition such as a disease.	1016
interacts with	RO:0002434	A relationship that holds between two entities in which the processes executed by the two entities are causally connected.	900
inhibits	RO:0002408	Directly negatively regulates.	391
causes condition	RO:0003303	A relationship between an entity (e.g. a genotype, genetic variation, chemical, or environmental exposure) and a condition (a phenotype or disease), where the entity has some causal role for the condition.	212
contributes to condition	RO:0003304	A relationship between an entity (e.g. a genotype, genetic variation, chemical, or environmental exposure) and a condition (a phenotype or disease), where the entity has some contributing role that influences the condition.	107
has genotype	GENO:0000222	A relationship that holds between a biological entity and some level of genetic variation present in its genome.	106
has role in modelling	RO:0003301	A relation between a biological, experimental, or computational artefact and an entity it is used to study, in virtue of its replicating or approximating features of the studied entity.	103
correlated with	RO:0002610	A relationship that holds between two entities, where the entities exhibit a statistical dependence relationship. The entities may be statistical variables, or they may be other kinds of entities such as diseases, chemical entities or processes.	72
activates	RO:0002406	Directly positively regulates.	61
involved in	RO:0002331	x is involved in y if and only if x enables some process y' , and y' is part of y .	
enables	RO:0002327	catalyses.	49
colocalises with	RO:0002325	x colocalises with y if and only if x is transiently or peripherally associated with y .	38
is allele of	GENO:0000408	A relation linking an instance of a variable feature (aka an allele) to a genomic location/locus it occupies. This is typically a gene locus, but a feature may be an allele of other types of named loci such as QTLs, or alleles of some unnamed locus of arbitrary size.	39
has affected feature	GENO:0000418	A relation that holds between an instance of a genetic variation and a genomic feature (typically a gene class) that is affected in its sequence or expression.	22
is causal loss of function germline mutation of	RO:0004012	Relates a gene to a condition, such that a mutation in this gene in a germ cell impairs the function of the corresponding product and that is sufficient to produce the condition and that can be passed on to offspring.	15
in 1 to 1 orthology relationship with	RO:HOM0000020	Orthology that involves two genes that did not experience any duplication after the speciation event that created them.	10
is marker for	RO:0002607	x is marker for y if the presence or occurrence of y is correlated with the presence or occurrence of x , and the observation of x is used to infer the presence or occurrence of y . Note that this does not imply that x and y are in a direct causal relationship, as it may be the case that there is a third entity z that stands in a direct causal relationship with x and y .	1
is causal gain of function germline mutation of	RO:0004011	Relates a gene to a condition, such that a mutation in this gene in a germ cell provides a new function of the corresponding product and that is sufficient to produce the condition and that can be passed on to offspring.	1

Table 5

The best found hyperparameters for the XGBoost model for the HD chorea graph.

parameter	best
min_child_weight	5
gamma	0.5
reg_alpha	0.5
reg_lambda	3
colsample_bytree	0.8053
max_depth	10
n_estimators	50
learning_rate	0.1258

Table 6

The top ten ranked drugs for HD chorea.

URI	name
https://identifiers.org/chembl:CHEMBL29097	CHEMBL29097
https://identifiers.org/chembl:CHEMBL8260	BAICALEIN
https://identifiers.org/chembl:CHEMBL221137	EMBELIN
https://identifiers.org/chembl:CHEMBL267345	AMPHOTERICIN B
https://identifiers.org/chembl:CHEMBL308688	5,7-DIMETHOXYISOFLAVONE
https://identifiers.org/chembl:CHEMBL2110660	IGMESINE
https://identifiers.org/chembl:CHEMBL275809	FR-122047
https://identifiers.org/chembl:CHEMBL161343	ARACHIDONOYL GLYCINE
https://identifiers.org/chembl:CHEMBL585	TRIAMTERENE
https://identifiers.org/chembl:CHEMBL1269845	CHEMBL1269845

Table 7

The two highest ranked drugs for HD chorea with the genes they interact with, the interaction types and prediction confidence.

drug ID	drug label	gene ID	gene label	interaction type	confidence
chembl:CHEMBL29097	CHEMBL29097	HGNC:10555	ATXN2	regulates	0.990
		HGNC:10596	SCN8A	inhibits	0.963
		HGNC:4572	GRIA2	inhibits	0.934
		HGNC:29259	TAOK1	regulates	0.955
		HGNC:1461	CAMK2B	regulates	0.935
		HGNC:4235	GFAP (human)	regulates	0.935
		HGNC:713	ARSA	regulates	0.969
		HGNC:4076	GABRA2	activates	0.908
		HGNC:4580	GRIK2	inhibits	0.974
		HGNC:11005	SLC2A1	regulates	0.914
		HGNC:30035	PIK3R5	inhibits	0.981
chembl:CHEMBL8260	BAICALEIN	HGNC:10555	ATXN2	regulates	0.988
		HGNC:10596	SCN8	inhibits	0.979
		HGNC:4572	GRIA2	inhibits	0.946
		HGNC:29259	TAOK1	regulates	0.963
		HGNC:4584	GRIN1	inhibits	0.911
		HGNC:1461	CAMK2B	regulates	0.918
		HGNC:4235	GFAP (human)	regulates	0.905
		HGNC:713	ARSA	regulates	0.960
		HGNC:4580	GRIK2	inhibits	0.971
		HGNC:2295	CP (human)	regulates	0.915
		HGNC:30035	PIK3R5	inhibits	0.984