future internet

MDPI

Article

Drug Repurposing for Parkinson's Disease by Integrating Knowledge Graph Completion Model and Knowledge Fusion of Medical Literature

Xiaolin Zhang and Chao Che *

Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, Dalian University, Dalian 116622, China; zhangxiaolin@s.dlu.edu.cn

* Correspondence: chechao@dlu.edu.cn; Tel.: +86-0411-8740-2046

Abstract: The prevalence of Parkinson's disease increases a tremendous medical and economic burden to society. Therefore, the effective drugs are urgently required. However, the traditional development of effective drugs is costly and risky. Drug repurposing, which identifies new applications for existing drugs, is a feasible strategy for discovering new drugs for Parkinson's disease. Drug repurposing is based on sufficient medical knowledge. The local medical knowledge base with manually labeled data contains a large number of accurate, but not novel, medical knowledge, while the medical literature containing the latest knowledge is difficult to utilize, because of unstructured data. This paper proposes a framework, named **D**rug **R**epurposing for Parkinson's disease by integrating **K**nowledge Graph Completion method and Knowledge **F**usion of medical literature data (DRKF) in order to make full use of a local medical knowledge base containing accurate knowledge and medical literature with novel knowledge. DRKF first extracts the relations that are related to Parkinson's disease from medical literature and builds a medical literature knowledge graph. After that, the literature knowledge graph is fused with a local medical knowledge base that integrates several specific medical knowledge sources in order to construct a fused medical knowledge graph. Subsequently, knowledge graph completion methods are leveraged to predict the drug candidates for Parkinson's disease by using the fused knowledge graph. Finally, we employ classic machine learning methods to repurpose the drug for Parkinson's disease and compare the results with the method only using the literature-based knowledge graph in order to confirm the effectiveness of knowledge fusion. The experiment results demonstrate that our framework can achieve competitive performance, which confirms the effectiveness of our proposed DRKF for drug repurposing against Parkinson's disease. It could be a supplement to traditional drug discovery methods.

Keywords: drug repurposing; Parkinson's disease; knowledge graph completion method; knowledge fusion

Citation: Zhang, X.; Che, C. Drug Repurposing for Parkinson's Disease by Integrating Knowledge Graph Completion Model and Knowledge Fusion of Medical Literature. *Future Internet* **2021**, *13*, 14. https://doi.org/10.3390/fi13010014

Received: 19 November 2020
Accepted: 5 January 2021
Published: 8 January 2021

1. Introduction

Parkinson's disease is a neurodegenerative disease that occurs more frequently in the elderly. According to the latest statistics from the authoritative institution, The Michael J. Fox Foundation for Parkinson's Research (MJFF), only in the United States, the total cost of Parkinson's disease to individuals, families, and the government is $51.9 billion each year, which fully shows the large medical and economic burden of Parkinson's disease on society. Therefore, it is very urgent to develop effective drugs to treat Parkinson's disease. In view of the good effects of dopaminergic medications on Parkinson's disease and the slowing of the pathogenesis of Parkinson's disease [1], the exploration of levodopa drugs is the mainstream research direction for the treatment of Parkinson's disease. However, this method also has the disadvantage that the long-term use of this drug will cause motor complications of patients [2]. Therefore, the exploration of levodopa drugs is risky, and it is difficult to achieve huge breakthroughs. In recent years, drug repurposing has been playing an increasingly important role in drug development research [3], such as amantadine [4],

which was originally used to treat influenza infections, was found to have a very positive effect on Parkinson's disease. Therefore, it is very promising to explore the treatment of Parkinson's disease by drug repurposing.

Drug repurposing is a new exploration and development of the application of existing drugs. The traditional development of new drugs is a time-consuming, costly, and high-risk process [5]. The great cost of traditional pharmaceutical development can be significantly reduced by drug repurposing, since drug repurposing could explore new indications of existing drugs and bypass several stages of de novo [6]. Compared with traditional pharmaceutical methods, drug repurposing is more efficient, low-cost, and riskless. Recently, drug repurposing has become the focus of the attention of major research institutions in view of the advantageous features of drug repurposing. Drug repurposing has greatly reduced the time cost of the drug development process due to the rapid growth of biomedical knowledge and related big data. Researchers can determine new drug targets, on average, in 1–2 years [7]. In addition, the R&D investment that is required for drug repurposing is lower than traditional drug R&D methods. Increasing drug repurposing approaches have been proposed in recent years due to the advantages and huge potential of drug repurposing in the medical field. The review conducted by Xue and Li [5] provides a detailed introduction of the drug repurposing approach.

In the field of drug repurposing against Parkinson's disease, most studies often obtain data from medical literature that contains large novel knowledge to construct a medical knowledge graph. However, the knowledge extracted from medical literature through data mining tool without manually labeling is not accurate and complete. The data of a medical knowledge bases, such as DrugBank, which combines detailed drug data with comprehensive drug target information has high accuracy, but these knowledge bases do not contain the latest medical knowledge. In order to improve this problem, this study proposed a drug repurposing framework for Parkinson's disease by integrating medical literature data and knowledge base. Firstly, we extract novel medical information in the medical literature and integrate them with the data in the local medical base to build a fused knowledge graph that combines novel medical knowledge with accurate medical knowledge. Subsequently, we employ knowledge graph completion methods utilizing fused knowledge graph to predict the drug candidates for Parkinson's disease. Finally, we employ machine learning methods to make classification on the Parkinson's disease-drug pair data sets to predict the drug candidates for Parkinson's disease and make a comparative experiment in order to confirm the effectiveness of knowledge fusion. Our proposed DRKF mainly includes literature retrieval and acquisition, extraction of entities and their relationships in medical literature, construction and fusion of medical knowledge graph, knowledge graph completion methods to make prediction of drugs, and machine learning-based classification for repurposing the drug candidates. This paper further explores the drug repurposing research of Parkinson's disease integrating knowledge fusion and knowledge graph completion method.

The main contributions of our work are as follows:

- we combine novel knowledge and accurate knowledge by integrating the literature-based knowledge graph with a local medical knowledge base;
- we apply relatively effective knowledge graph completion methods to predict the drug candidates for Parkinson's disease and discover that ConvTransE get a better prediction results;
- we employ classic machine learning methods to repurpose the drug candidates against Parkinson's disease and compare the results with the method only using literature-based knowledge graph to confirm the effectiveness of knowledge fusion.

The paper is organized, as follows. Section 2 introduces the related work of computational drug repurposing. Section 3 introduces the data sets and the method for repurposing drug candidates for Parkinson's disease. Section 4 introduces the experimental process and analyzes the results. Section 5 summarizes the conclusion.

2. Related Work

In recent years, drug repurposing combining with computational methods has been developing rapidly. Increasing computational methods have been proposed to explore the drug-disease relationship for implementing drug repurposing [8]. Li et al. [9] developed a combination of network mining and text mining to extract the relationship between diseases and proteins in the molecular interaction network and retrieve drugs that are indirectly related to certain diseases in the PubMed abstract, and then combine the drugs, proteins, and disease relationships to construct a disease-specific drug–protein network. Rastegar et al. [10] obtained the drug-disease relationship by extracting the drug–gene and gene–disease relationship and rank score from the medical abstract and then verified the performance of the method by comparing with Comparative Toxicogenomics Database. Wu et al. [11] determined the modules that are closely related to diseases and drugs by establishing a weighted network model of disease and drug heterogeneity, and then obtained information regarding potential drug-disease candidates for drug repurposing. Napolitano et al. [12] proposed a drug-based drug repurposing prediction method that is based on machine learning algorithms. It integrated drug chemical structure, protein–protein, and gene–gene similarity information, and then utilized classification methods in order to classify drugs to implement drug repurposing. To the best of our knowledge, the drug repurposing method integrated literature-based knowledge and the medical base has rarely been explored. Therefore, there is great room to discovery drug repurposing by employing these methods.

Knowledge graph is the data basis for achieving drug repurposing, which could organize, manage, and utilize massive amounts of information. The information in the knowledge graph is generally organized in the form of triple (h, r, t), where h, r, t represent the head entity, relationship, and tail entity, respectively. This form is very intuitive and widely used in knowledge graph completion tasks. However, the knowledge graph, in reality, is often sparse and incomplete. Therefore, it is of great necessity to expand the knowledge graph through knowledge fusion. Knowledge fusion is to merge knowledge graphs from multiple sources. Entities are the basic units of knowledge graphs. These data have diversity and heterogeneity in different knowledge graphs. The basic problem is to study how to integrate descriptive information about the same entity or concept from multiple sources. Knowledge fusion can make the knowledge graph contain richer and more accurate information. Combining it with the knowledge graph completion learning method can greatly improve the effect of drug repurposing.

Knowledge graph completion method employing knowledge graph to achieve knowledge reasoning has broad prospects. It learns to project entities and relations of knowledge graph into computable low-dimensional vectors, and then achieves knowledge inference and reasoning. The dependent relations between entities of knowledge graph make these entities and relations contain rich structural information. We can obtain these latent information of entities and relations through those knowledge graph completion methods, and then implement the tasks of node classification [13] and link prediction [14]. The classic knowledge graph completion methods include TransE [15], DistMult [16], ComplEx [17], etc. The combination of knowledge graph completion method and knowledge fusion can provide a new idea for the research of drug repurposing.

The research of drug repurposing for Parkinson's disease just recently appeared. In recent research of drug repurposing for Parkinson's disease, the relative representative of which is the following one. Zhu et al. [18] proposed a framework that includes biomedical entities and their relationship extraction, the construction of knowledge graph, knowledge representation learning, and machine learning-based prediction to implement drug repurposing for Parkinson's disease. This article integrates the drug–disease, drug–gene, and medicine-related relation information in the literature into a knowledge graph, represents the entity information by employing knowledge graph embedding methods for the medical knowledge graph, and, finally, achieves the drug repurposing of Parkinson's disease through machine learning classification methods. The idea of this framework is very novel

and valuable in studying the drug repurposing against Parkinson's disease. However, the paper employs a smaller data sets in the PubMed literature, which leads to a relatively sparse knowledge graph, and the employed knowledge graph embedding methods were classic, but not novel enough. There is room for improvement in expanding data sets by knowledge fusion and employing more effective and novel computational methods.

Here, this paper aims to solve the problems in Zhu et al.'s paper, we construct a relatively complete medical knowledge graph that is based on the Zhu et al.'s PubMed literature data and integrate it with a local medical knowledge base. The knowledge fusion not only expands the amount of information, but also builds it into a highly accurate and novel data sets. In terms of methods, we employ relatively novel knowledge graph completion models to predict the drug candidates for Parkinson's disease. In addition, we employ machine learning methods to repurpose the drug candidates against Parkinson's disease and confirm the effectiveness of knowledge fusion. The experimental results prove that the framework has implemented relatively good results.

3. Materials and Methods

3.1. Data Sets

The data we use consists of two parts. The specific data description is as follows.

3.1.1. Literature Data

The knowledge that is related to Parkinson's disease in the literature. The medical literature contains the latest research results in this field. So, the information that is extracted in the literature represents the latest information on medical development. These rich and latest knowledge largely exists in unstructured data in medical literature databases such as PubMed. PubMed is a huge corpus that contains citations to biomedical literature from MEDLINE and life science journals and it can be used for drug discovery [16]. Currently, it contains more than 26 million biomedical abstracts. The source of the literature-based data that we used here were obtained in the PubMed database. The relevant literature is retrieved through Medical Subject Heading (MeSH) and we then download the abstract of the articles that are related to Parkinson's disease in the PubMed. 54,100 published articles are extracted in PubMed from 1945 to 2018.

3.1.2. The Data in the Local Medical Knowledge Base

Another source of the data we used is a local medical knowledge base, which integrated several medical knowledge bases that are freely available: DrugBank [19], PharmGKB [20], KEGG DRUG [21], TTD [22], DID, and SIDER [23] It is a medical knowledge network that is composed of entity types, such as Drug, Disease, Gene, Side effect, and Pathway with their relationships. Figure 1 shows the data schema of the local medical knowledge base.

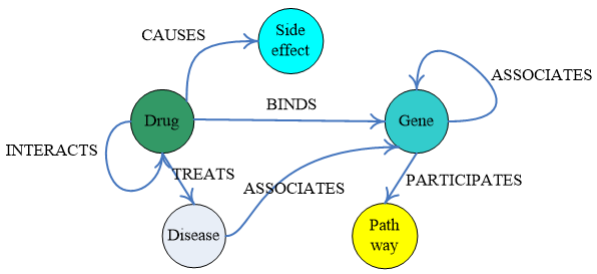


Figure 1. The schema of the local medical knowledge base.

3.2. Method

In this paper, our proposed DRKF aims to predict drug candidates that can treat Parkinson's disease. The framework includes the extraction of medical entities and relations, knowledge graph construction, knowledge fusion, knowledge graph completion

method, and machine learning classification for drug repurposing. As shown in Figure 2, the specific process is as follows.

- extracting and prerocessing medical data in the literature;
- constructing medical entities and their relationships into a literature-based knowledge graph and integrating it with local medical base;
- employing the knowledge graph completion methods to predict the drug candidates for Parkinson's disease; and,
- using the machine learning methods to repurpose the drug candidates against Parkinson's disease.

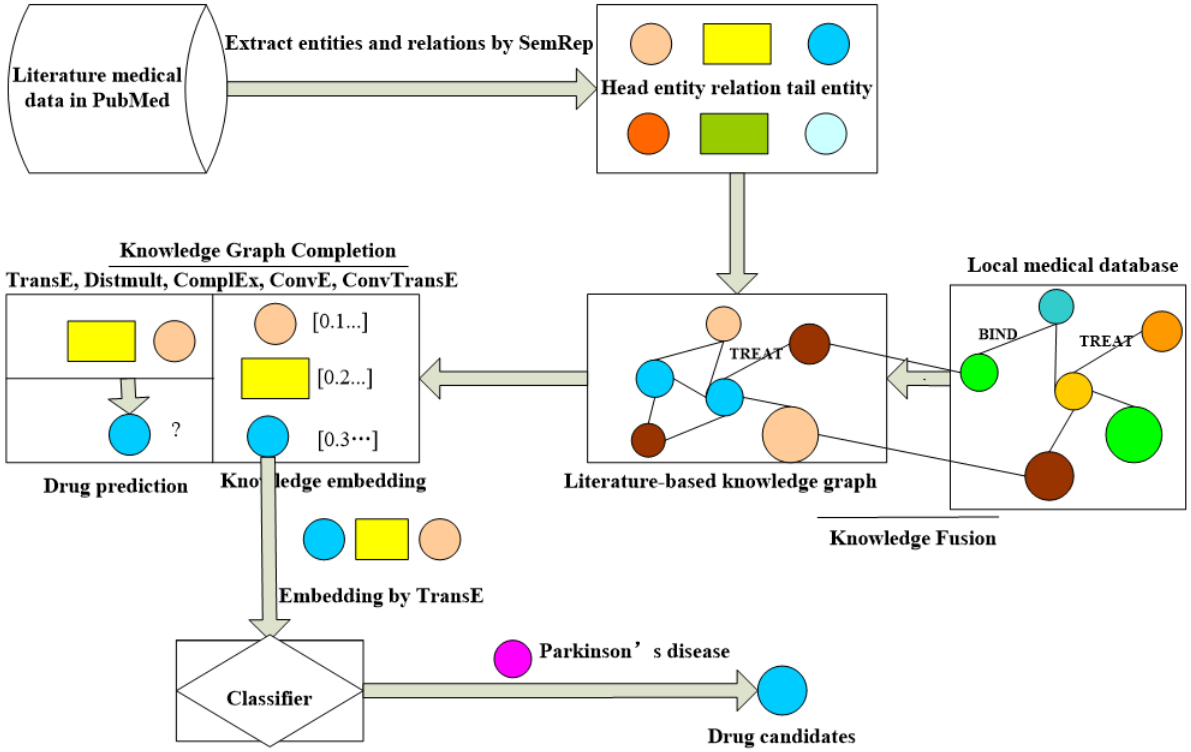


Figure 2. The framework of our drug repurposing method.

3.2.1. Preprocessing and Extraction of Medical Entities and Their Relationships

We preprocess the abstracts of the literature downloaded from the PubMed database by SemRep [24], which is a UMLS-based program that extracts information from the literature. We employ SemRep to extract entities and their relationships from the abstract to construct triples. A triple contains a subject, a predicate, and an object, which corresponds to a medical entity and their relation. In addition, there are types and unique concept identifiers information of these biomedical entities in the UMLS vocabulary. For example, in the extracting result, ***"00000000.tx.2|relation|C0178601|Dopamine Agonists|phsu|phsu|||TREATS|C0030567|Parkinson Disease|dsyn|dsyn|"***, "Dopamine Agonists" and "Parkinson Disease" represents the head entity and tail entity, respectively, and "TREATS" is the relation between them. It means that dopamine agonists have the relation of treatment to Parkinson's disease. "C0178601" and "C0030567" corresponds to the concept identifiers of "Dopamine Agonists" and "Parkinson Disease" in the UMLS vocabulary, respectively. In addition, "|phsu|" and "|dsyn|" is the abbreviation of "pharmacologic substance" and "disease or syndrome", respectively. In order to construct a more standardized and professional knowledge graph, we use the UMLS concept identifier, rather than biomedical entity to construct a knowledge graph, like the form of "C0178601 TREATS C0030567". Subsequently, we create the "IS_TYPE" relationship between the entity, like "C0178601 IS_TYPE phsu|phsu" in our knowledge graph and the reason is that

the types of entities in UMLS, like "phsu | phsu" contain hierarchical information that could more accurately describe entity information. For example, the type of sildenafil and aspirin are both drugs, while the type of cancer is disease. It is easy to obtain the information that sildenafil is more closely related to aspirin, rather than cancer. Subsequently, we construct inverse relationships, like "C0030567 TREAT_INVERSE C0178601" based on main relationships that exist multiple times in literature-based knowledge graph. Those inverse relationships can not only expand the knowledge graph, but also provide more semantic information to it. In addition, the number of some triples is very small, and these triples are made due to the errors or noises in the extraction process; therefore, we deleted these triples.

3.2.2. Construction and Fusion of the Medical Knowledge Graph

After the above processing of the medical literature data, we construct a literature-based knowledge graph that is centered on Parkinson's disease. The knowledge graph contains 115,300 triples, which is composed of 12,497 medical entities and 43 relations. It includes many types of entities, such as diseases, drugs, genes, and relevant entity types. This is the information in our literature-based knowledge graph.

The literature-based knowledge graph and a local medical base that contains accurate, but relatively old, information are integrated to construct a fused medical knowledge graph. The detailed steps are as follows:

Firstly, the data in local knowledge base are all represented in the form of "head_entity relation tail_entity". The form of triples can intuitively represent the relational network in the knowledge graph. Additionally, it is very helpful for the detailed operation process of knowledge fusion. Subsequently, the main problem of knowledge fusion is the medical entity name in local knowledge base cannot correspond to the entity name in the literature-based knowledge graph. The same entity has different form of expression, such as "Parkinson Disease" and "Parkinson's disease". That would construct a inconsistent knowledge graph if they are directly integrated into the literature-based knowledge graph. In order to address the inconsistency problem of the data. In the local medical knowledge base, we found that every entity has corresponding UMLS concept identifier. We utilize the UMLS identifier instead of the specific entity in local medical knowledge base. For example, the UMLS identifier corresponding to "Parkinson's Disease" and "Parkinson's disease" are both "C0030567", which can solve the problem and fuse them together. Finally, the literature-based knowledge graph includes a large number of medical entities and the target of drug repurposing is Parkinson's disease. Therefore, in the process of knowledge fusion, we employ the information that is related to Parkinson's disease in the local knowledge base. This information is extracted from the local knowledge base in the form of triples and integrated into the literature-based knowledge graph to construct a fused knowledge graph. The fused knowledge graph is a medical knowledge network that is composed of nodes and edges, where nodes represent medical entities and edges represent relationships between those medical entities. The fused knowledge graph contained 165,901 triples, which is composed of 12,497 medical entities and 43 relations. Table 1 shows the data sets of medical literature and knowledge fusion data. We assumed that there are undeveloped drugs used to treat Parkinson's disease in the knowledge graph. These medical entities have direct or indirect relations with Parkinson's disease. Therefore, knowledge graph completion models and machine learning methods can be employed in order to repurpose the drug candidates for Parkinson's disease by using the fused medical knowledge graph.

Table 1. Comparison of medical data sets before and after fusion.

Data Sets	Medical Literature Data	Knowledge Fusion Data
Entities	12,497	12,497
Relations	43	43
triples	115,300	165,901

3.2.3. The Prediction of the Drug Candidate for Parkinson's Disease by Knowledge Graph Completion Methods

When considering a knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{E})$, where $\mathcal{E}$ represents the set of medical entities, $\mathcal{R}$ stands for the set of relations among entities, a triple can be represented as (h, r, t). By employing these triples to knowledge graph completion methods, it could infer new triples (h', r', t'), where the entities of h', t' are both in set $\mathcal{E}$ and the relations of r' are in set $\mathcal{R}$. The knowledge graph completion task could be represented as a ranking task, in which we learn a prediction function $\psi(h, r, t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \mapsto \mathbb{R}$ that could judge the true or false of triples. We employed five knowledge graph completion methods: DistMult and ComplEx for semantic matching models, ConvE [25] and ConvTransE [26] for neural network models, and TransE for translational distance models. The ways that these methods encode entities and relations into a low dimensional vector space are different. However, those methods can also be used for knowledge reasoning.

Translational distance models (TransE). TransE model defines a triple (h, r, t) as a translation between head entity h and tail entity t through relation r in a continuous vector space. It is like the form of $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, in which h, r, t is the embedding of h, r, and t, respectively. The score function of TransE is $s(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_1/L_2}$. Either L_1 or L_2 norm can be employed.

Semantic matching models (DistMult and ComplEx). DistMult is a relatively simple semantic matching models. The score function of the DistMult model is defined as $s(h, r, t) = \langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle$. DistMult is limited to symmetric relations, which make it unable to distinguish head and tail entities. ComplEx extends DistMult to the complex domain in order to improve this problem. The embeddings of Head and tail for the same entity are complex conjugates that enable the ComplEx model could capture asymmetric relations information. Its score function is defined as $s(h, r, t) = \mathrm{Re}(\langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle)$, where Re() is a real part of a complex vector and k is a dimension of an embedding.

Neural network models (ConvE and ConvTransE). ConvE is a relatively simple method among neural network models. The Score function of ConvE is defined as $\psi_r(\mathbf{e}_s, \mathbf{e}_o) = f(\mathrm{vec}(f([\overline{\mathbf{e}}_s; \overline{\mathbf{r}}_r] * \omega))\mathbf{W})\mathbf{e}_o$, where $f()$ is the nonlinear activation function vec means to flatten the tensor into a vector and the [] operator is to deform and splice the embedding e and r. It projects embedding vectors to another spaces for characterization and, due to the powerful feature extraction capability of the convolution structure, it can obtain good link prediction results and obtain less parameter utilization. The ConvTransE model maintains the characteristics of translation, like TransE between entities and relationships on the basis of ConvE. The score function of ConvTransE is defined as $f(\mathrm{vec}(\mathbf{M}(e_s, e_r))W)e_o$. f denotes a non-linear function. The feature map matrix is reshaped into a vector vec that is projected into a F^L dimensional space while using W for linear transformation.

We employed these knowledge graph completion models to predict drug candidates that can potentially treat Parkinson's disease with the medical knowledge graph. The knowledge graph completion model can project the entities and relationships in the medical knowledge graph to the low-dimensional continuous vector space, and then predict the new treatment relations in medical entities from the medical knowledge graph. The processes and results of the experiment can be seen in the next section.

3.2.4. The Prediction of the Drug Candidate for Parkinson's Disease by Machine Learning Methods

In machine learning tasks, we regard drug repurposing as a binary classification problem. It divides the relation between a drug and a disease into two categories: treatment and no treatment. The classifier is trained by the treatment mechanism of drugs that are used to treat non-Parkinson's disease, and then the learned model is employed in order to predict potential drug candidates for Parkinson's disease. Support vector machine (SVM) [27], random forest [28], logistic regression, and decision tree are classic machine learning models. Here, we use those machine learning models to repurpose the potential drug candidates for Parkinson's disease.

4. Experiment

4.1. Experimental Setup

In the knowledge graph completion method task, the number of training sets, validation sets, and test sets from the literature-based knowledge graph is 108,348, 5703, and 1249, respectively.

In the same model, for the data sets that fused knowledge graph, the number of training sets, valid sets, and test sets is 149,567, 15,027, and 1307, respectively.

We employ different hyperparameters on the training set. We manually set the hyperparameter ranges: learning rate in $\{0.01, 0.03, 0.05\}$, embedding size in $\{50, 100\}$, number of kernels in $\{50, 100\}$, dropout rate in $\{0.2, 0.3, 0.4, 0.5\}$, and kernel size in $\{2 \times 1, 2 \times 4\}$ for those knowledge graph completion methods.

In the task of machine learning classification, we regard drug repurposing as a binary classification problem. It divides the relation between a drug and disease into two categories: treatment and no treatment. Classifier model is trained by the treatment mechanism of drugs that are used to treat non-Parkinson's disease, and then the learned model is used to predict drugs candidates for Parkinson's disease. The specific operation is as follows. Firstly, we obtain the vector representations of entities and their relations using the classic model TransE [15]. Subsequently, we extract the triples with TREATS relationships (like the form of (head entity, TREATS, tail entity)) in the knowledge graph. Afterwards, the triples are divided into two types: Parkinson's disease and other disease. For the triples of tail entity is other disease; we extract and randomly change the tail entity in order to construct a negative sample with a ratio of 1:3, and replace the entities and relationships in the triples with corresponding vector representations and the use of these triples as the training set. The triples in which the tail entity is Parkinson's disease are extracted and they perform same operations with the training set and employed it as the test set. According to statistics, the number of training sets is 76,607 and the number of test sets is 5228.

4.2. Evaluation Metrics

We employ two kinds of evaluation metrics. The experiments use the proportion of correct entities ranked in the top one, three, and ten (Hits@1, Hits@3, Hits@10)for knowledge graph completion task. hit@n represents the proportion of the results in the test set among the top-k prediction results. In addition, Precision (P), Recall (R), and F-1 scores are employed in the machine learning classification task.

4.3. Experimental Results and Analysis

In the task of knowledge graph completion methods, drug repurposing is considered to be a task of predicting missing medical entities for given triples like the form of (?, TREAT, Parkinson's disease). It predicts drug candidates that are undeveloped for treating Parkinson's disease by those knowledge completion models trained by the data sets that did not contain the triples, like (?, TREAT, Parkinson's disease). In this way, we use these trained models to repurpose the drug candidates against Parkinson's disease.

In the task of machine learning classification, we employ four machine learning methods in the data sets of fused knowledge graph and only in the medical literature-based

knowledge graph to classify whether there exists the relation of treatment between drugs and Parkinson's disease and compare the classification results with the results obtained by Zhu et al. [18]. The data set that was employed by Zhu et al. is the data containing 48,378 triples with 4653 medical entities in the PubMed database and it is a subset of our medical literature-based knowledge graph.

4.3.1. Comparison of Knowledge Graph Completion Models

Firstly, we compare the prediction results of the medical candidates using and not using knowledge graph fusion. For the ConvTransE model, the prediction results using knowledge fusion can increase Hits@1 by 0.77%, Hits@3 by 3.59%, and Hits@10 by 0.56% as compared with the result obtained not using knowledge graph fusion, as we can see from Table 2. For the ConvE model, Hits@1 increased by 7.51%, Hits@3 increased by 16.84%, and Hits@10 increased by 1.48%. For the ComplEx model, Hits@1 increased by 13.68%, Hits@3 increased by 6.03%, and Hits@10 increased by 1.86%. For the DistMult model, Hits@1 increased by 7.65%, Hits@3 increased by 5.34%, and Hits@10 increased by 2.98%. For the TransE model, Hits@1 increased by 12.83%, Hits@3 increased by 6.77%, and Hits@10 increased by 6.57%. From the experimental results, we can see that, through knowledge fusion, we not only expand the medical knowledge graph, but also integrate the novel knowledge in the literature-based knowledge graph with the accuracy knowledge in the local medical base. It significantly improves the accuracy of drug repurposing on Parkinson's disease.

Subsequently, we compared the prediction results of drug repurposing among different models. Firstly, we found that the TransE model, as a classic knowledge graph embedding model, is simple, but still achieved high prediction. The TransE model is an classic graph embedding model that treats the relationship as a certain translation vector between entities and it regards the relationship vector r as the translation between the head entity vector h and the tail entity vector t for each triple (h, r, t). However, this model is less effective in dealing with complicated relationships, such as one-to-many and many-to-many. Secondly, the ComplEx model has a good effect on capturing and predicting the semantics of asymmetric relations. The reason is that the ComplEx [17] model proposes a method that is based on the representation of complex numbers and it can better capture the semantic information of symmetric relations and make accurate predictions. Subsequently, we can see that the prediction results of the ComplEx model are close to the results of the DistMult model in drug repurposing, because the two models are analogous, and ComplEx introduced complex embedding extends DistMult. The ConvE model has a certain performance improvement when compared with the ComplEx model in the prediction of drug repurposing. The reason is that the ConvE model employs a convolution method. The ConvE model has a good improvement in the prediction results, due to the powerful feature extraction capability of the convolution structure. The ConvTransE model outperforms the ConvE model and it has obtained the best prediction results. The reason is that ConvTransE model adds features of translation on the basis of the ConvE model in order to further improve the accuracy of predictions. Throughout the results, we found that ConvTransE has achieved relatively better results. Therefore, we speculate that ConvTransE method may have better prospects and it is valuable in making further improvements for employing in drug repurposing field.

Table 2. The drug repurposing results of five knowledge graph completion models.

Models	No Knowledge Fusion			Knowledge Fusion		
	Hits@1	**Hits@3**	**Hits@10**	**Hits@1**	**Hits@3**	**Hits@10**
TransE	42.11%	56.12%	69.02%	**54.94%**	62.89%	75.59%
DistMult	28.99%	49.74%	67.80%	36.64%	55.08%	70.78%
ComplEx	23.35%	45.92%	66.58%	37.03%	51.95%	68.44%
ConvE	28.82%	56.68%	73.52%	36.33%	**73.52%**	75.00%
ConvTransE	50.95%	67.97%	86.71%	51.72%	71.56%	**87.27%**

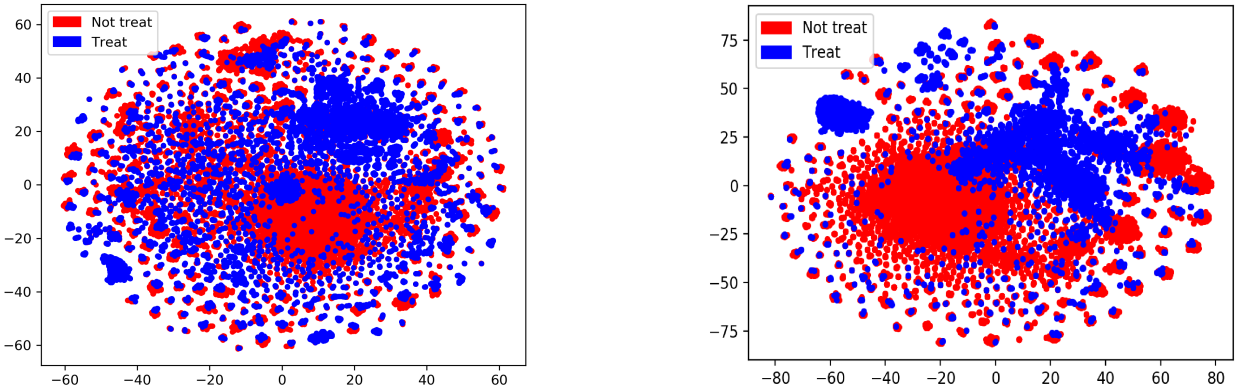
4.3.2. Comparison of Machine Learning Methods

We use classic knowledge graph completion method TransE to embed the medical entities and their relations into low-dimensional continuous entity vectors. Subsequently, we classify treatment relation by employing four machine learning methods to learn the treatment mechanism from the existing treatment relations of drug candidates which is undeveloped in treating Parkinson's disease. After that, the trained classifier is used to predict drug candidates to repurpose Parkinson's disease. Table 3 shows that the best-performing result has an F1 score of 98.42%, which is trained by SVM classifier in the fused knowledge graph. Additionally, the classification results trained on literature-based knowledge graph are better than the results presented in Zhu et al. The reason is that we employ a more rich and complete data sets of the literature. The performance of the machine learning methods that were trained on the fused knowledge graph has a more obvious improvement than other medical knowledge graphs. The most important reason is that we integrate the triples in the medical knowledge base containing a large amount of precise medical information with the literature-based knowledge graph that contains novel medical knowledge. Therefore, we employ machine learning methods in predicting drug candidates against Parkinson's disease in order to further confirm the effectiveness of knowledge fusion.

Table 3. Experimental results of machine learning methods using different knowledge graph.

Models	Literature-Based Knowledge Graph			Fused Knowledge Graph			Results in Zhu et al. [18]		
	Recall	**Precision**	**F1-Score**	**Recall**	**Precision**	**F1-Score**	**Recall**	**Precision**	**F1-Score**
SVM	98.78%	96.42%	97.58%	100.00%	96.90%	**98.42%**	98.72%	94.14%	96.38%
LogisticRegression	97.55%	93.07%	95.26%	99.92%	93.48%	**96.59%**	93.97%	91.42%	92.68%
RandomForest	96.56%	93.48%	95.00%	97.12%	94.91%	**96.00%**	83.41%	93.01%	87.95%
DecisionTree	83.51%	81.55%	82.52%	89.14%	82.27%	**85.57%**	72.16%	76.13%	74.09%

We employ the t-SNE [29] dimensionality reduction tool to perform dimensionality reduction and visualization processing on the training set in the fused knowledge graph and the literature-based knowledge graph that was used in machine learning methods. The classification effect of Figure 3b has a higher degree of aggregation than the result of Figure 3a, as shown in the Figure 3. The information expression ability of the embedding acquired by entities in fused knowledge graph is significantly better than that of the knowledge graph without fusion. The superiority of knowledge fusion is more intuitively demonstrated through visual display.



(**a**) The representation using knowledge fusion (**b**) The representation not using knowledge fusion

Figure 3. Two-dimensional representation of the vectors learned by TransE.

4.4. Results and Discussion

We made some predictions based on given drugs in knowledge graph through the trained machine learning model SVM classifier, which is a binary classifier. We analyzed the relation between Parkinson's disease and drug candidates in the knowledge graph to understand why these drugs are more likely to treat Parkinson's disease than other drugs. The candidate drug is associated with existing drugs to treat Parkinson's disease, as we can see from triples (Drug_01 ASSOCIATE_WITH Drug_02) and (Drug_02 TREAT Parkinson's disease). Subsequently, the possibility that the drug might treat Parkinson's disease is higher. Therefore, we consider that these drugs are more likely to have a therapeutic relationship with Parkinson's disease than other drugs. Therefore, we could speculate that Drug_01 has potential treatment to Parkinson's disease. According to this treatment mechanism, our classifier predicts several drug candidates for Parkinson's disease and we have verified our results through some literature, and we have taken several meaningful results to present in Table 4.

Table 4. Results of drug repurposing by classifier for Parkinson's disease.

UMLS ID	Drug Name	Source that Have Been Proved
C4754962	Terazosin	Medical literature in the Journal of clinical investigation [30]
C1367795	Ambroxol	Medical literature in JAMA neurology [31]
C1721377	Nilotinib	Medical literature in JAMA neurology [32]

It can be seen from the above experimental results that our framework has achieved relatively good results. However, there are too many medical entities in the medical knowledge graph. The information of medical entities are not related to Parkinson's disease in the medical knowledge graph. Those entities perhaps interfere with drug repurposing for Parkinson's disease. Therefore, there is still a lot of room to explore related research in the future.

5. Conclusions

In this paper, we proposed a drug repurposing framework by integrating literature-based medical knowledge graph and local medical base. Through this framework, we fused the literature-based data that contain novel knowledge and a local medical knowledge base with high accuracy information. The results of the drug repurposing have been improved by employing the fused knowledge graph. In addition, in the knowledge graph completion methods, it is found that ConvTransE has achieved better results in drug repurposing than other models, which can provide new directions for subsequent research. After that, we use machine learning models to explore treatment mechanisms,

and utilize this potential information to repurpose drug candidates against Parkinson's disease and further confirm the effectiveness of knowledge fusion. The drug repurposing framework that was proposed in this paper uses knowledge fusion, knowledge graph completion approaches, and machine learning methods in order to predict drug candidates for Parkinson's disease, and the experimental results provide researchers with valuable research ideas to further explore the drug repurposing for Parkinson's disease.

Author Contributions: Conceptualization, X.Z. and C.C.; methodology, X.Z.; validation, C.C.; formal analysis, X.Z.; investigation, X.Z.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, C.C.; visualization, X.Z.; project administration, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 62076045) and the Guidance Program of Liaoning Natural Science Foundation (No. 2019-ZD-0569).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Armstrong, M.J.; Okun, M.S. Diagnosis and treatment of Parkinson disease: A review. *JAMA* **2020**, *323*, 548–560. doi:10.1001/jama.2019.22360. [CrossRef] [PubMed]
2. Reddy, D.H.; Misra, S.; Medhi, B. Advances in drug development for Parkinson's disease: Present status. *Pharmacology* **2014**, *93*, 260–271. [CrossRef] [PubMed]
3. Shameer, K.; Readhead, B.; Dudley, J.T. Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Curr. Top. Med. Chem.* **2015**, *15*, 5–20. [CrossRef] [PubMed]
4. Hubsher, G.; Haider, M.; Okun, M.S. Amantadine: The journey from fighting flu to treating Parkinson disease. *Neurology* **2012**, *78*, 1096–1099. [CrossRef]
5. Xue, H.; Li, J.; Xie, H.; Wang, Y. Review of drug repositioning approaches and resources. *Int. J. Biol. Sci.* **2018**, *14*, 1232. [CrossRef]
6. Ashburn, T.T.; Thor, K.B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **2004**, *3*, 673–683. [CrossRef]
7. Sertkaya, A.; Birkenbach, A.; Berlind, A.; Eyraud, J. *Examination of Clinical Trial Costs and Barriers for Drug Development: Report to the Assistant Secretary of Planning and Evaluation (ASPE)*; Department of Health and Human Services: Washington, DC, USA, 2014.
8. Gottlieb, A.; Stein, G.Y.; Ruppin, E.; Sharan, R. PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **2011**, *7*, 496. [CrossRef]
9. Li, J.; Zhu, X.; Chen, J. Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts. *PLoS Comput. Biol.* **2009**, *5*, 14. [CrossRef]
10. Rastegar-Mojarad, M.; Elayavilli, R.K.; Li, D. A new method for prioritizing drug repositioning candidates extracted by literature-based discovery. In Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 9–12 November 2015.
11. Wu, C.; Gudivada, R.C.; Aronow, B.J.; Jegga, A.G. Computational drug repositioning through heterogeneous network clustering. *BMC Syst. Biol.* **2013**, *7*, S6. [CrossRef]
12. Napolitano, F.; Zhao, Y.; Moreira, V.M.; Tagliaferri, R.; Kere, J.; D'Amato, M.; Greco, D. Drug repositioning: A machine-learning approach through data integration. *J. Cheminf.* **2013**, *5*, 30. [CrossRef]
13. Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; Eliassi-Rad, T. Collective classification in network data. *AI Mag.* **2008**, *29*, 93–106. [CrossRef]
14. Zhang, J.; Yu, P.S.; Zhou, Z.H. Meta-path based multi-network collective link prediction. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14), New York, NY, USA, 24–27 November 2014. [CrossRef]
15. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 2787–2795.
16. Yang, B.; Yih, W.; He, X.; Gao, J.; Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. *arXiv* **2014**, arXiv:1412.6575.

17. Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; Bouchard, G. Complex Embeddings for Simple Link Prediction. In Proceedings of the 33rd International Conference on International Conference on Machine Learning—Volume 48 (ICML'16), New York, NY, USA, 20–22 June 2016.
18. Zhu, Y.; Jung, W.; Wang, F.; Che, C. Drug repurposing against Parkinson's disease by text mining the scientific literature. *Libr. Hi Tech* **2020**, *38*, 741–750. [CrossRef]
19. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, 668–672. [CrossRef] [PubMed]
20. Thorn, C.F.; Klein, T.E.; Altman, R.B. PharmGKB: The pharmacogenomics knowledge base. *Pharmacogenomics* **2013**, 311–320._20. [CrossRef]
21. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **2020**. [CrossRef]
22. Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; et al. Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **2020**, *48*, 1031–1041. [CrossRef]
23. Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L.J.; Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **2010**, *6*, 343. [CrossRef]
24. Rindflesch, T.C.; Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *J. Biomed. Inf.* **2003**, *36*, 462–477. [CrossRef]
25. Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2D Knowledge Graph Embeddings. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), Hilton New Orleans Riverside, New Orleans, LA, USA, 2–7 February 2018.
26. Shang, C.; Tang, Y.; Huang, J.; Bi, J.; He, X.; Zhou, B. End-to-End Structure-Aware Convolutional Networks for Knowledge Base Completion. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-2019), Hilton Hawaiian Village, Honolulu, HI, USA, 27 January–1 February 2019. [CrossRef]
27. Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Model.* **2003**, *43*, 1882–1889. [CrossRef]
28. Cao, D.-S.; Zhang, L.-X.; Tan, G.-S.; Xiang, Z.; Zeng, W.-B.; Xu, Q.-S.; Chen, A.F. Computational Prediction of Drug Target Interactions Using Chemical, Biological, and Network Features. *Mol. Inform.* **2014**, *33*, 669–681. [CrossRef] [PubMed]
29. Van Der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
30. Cai, R.; Zhang, Y.; Simmering, J.E.; Schultz, J.L.; Li, Y.; Fernandez-Carasa, I.; Consiglio, A.; Raya, A.; Polgreen, P.M.; Narayanan, N.S.; et al. Enhancing glycolysis attenuates Parkinson's disease progression in models and clinical databases. *J. Clin. Investig.* **2019**, *129*. [CrossRef] [PubMed]
31. Mullin, S.; Smith, L.; Lee, K.; D'Souza, G.; Woodgate, P.; Elflein, J.; Hällqvist, J.; Toffoli, M.; Streeter, A.; Hosking, J.; et al. Ambroxol for the treatment of patients with Parkinson disease with and without glucocerebrosidase gene mutations: A nonrandomized, noncontrolled trial. *JAMA Neurol.* **2020**, *77*, 427–434. [CrossRef] [PubMed]
32. Pagan, F.L.; Hebron, M.L.; Wilmarth, B.; Torres-Yaghi, Y.; Lawler, A.; Mundel, E.E.; Yusuf, N.; Starr, N.J.; Anjum, M.; Arellano, J.; et al. Nilotinib effects on safety, tolerability, and potential biomarkers in Parkinson disease: A phase 2 randomized clinical trial. *JAMA Neurol.* **2020**, *77*, 309–317. [CrossRef]