

Generating Knowledge Graph Based Explanations for Drug Repurposing Predictions

Elif Ozkan¹, Remzi Celebi¹, Arif Yilmaz¹, Vincent Emonet¹ and Michel Dumontier¹

¹Institute of Data Science, Maastricht University, Maastricht, The Netherlands

Abstract

Over the past years, computer assisted drug repurposing methods have started to gain more attention as they offer a faster and a more effective way to treat many diseases. While these methods are quite promising in terms of power of prediction, the hesitation regarding the use of these methods in practice still remains due to their highly complex working mechanisms, which limits their interpretability.

Explainable Artificial Intelligence (XAI), which takes transparency, interpretability, informativeness as its main foundations, could address the limitations of the black-box models. In this context, Knowledge Graphs (KGs) could leverage the explanations provided to the user in the biomedical domain, as they are capable of represent relations between the entities in a semantically consistent way. Knowledge Graphs have the potential to generate graph-based representations, while providing the context, which make it easily interpretable by humans.

In this paper, we propose an approach, which is a KG based explainable AI framework in the field of drug repurposing as an extension of the PREDICT Method. The approach is centered on generating similarity-based explanations by extracting the relevant paths from the input, which consists of a disease and a predicted drug for the treatment of the disease. To demonstrate the utility of this approach, we demonstrate how the graphical operations used in the KG could be used to generate plausible explanations, by conducting a use case on Alzheimer Disease. Our findings suggest that the utilization of biomedical KGs and this approach has a great potential to provide transparent explanations as it is able to illustrate the relations between drug, disease entities which are quite relevant to the target input. Application of this approach to the drug repurposing and to other similar domains, could be helpful to overcome the limitations caused by the black-box nature of the computational drug repurposing models and could be a powerful tool to enhance the understanding of decision making process of models and simplify scientific communication among domain experts and computer scientists.

Keywords

Knowledge Graph, Explainable AI, XAI, drug repurposing

1. Introduction


The advancements in the field of Artificial Intelligence (AI) have been successfully utilized in the computer-assisted biomedical tasks in the past few years. AI and Machine learning methods applied in this field bears significant promise for drug discovery and repurposing as they significantly accelerate and offer new alternatives for the process of treatment [1]. Drug

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 13–16, 2023, Basel, Switzerland

✉ e.yozkan@student.maastrichtuniversity.nl (E. Ozkan); remzi.celebi@maastrichtuniversity.nl (R. Celebi); a.yilmaz@maastrichtuniversity.nl (A. Yilmaz); vincent.emonet@maastrichtuniversity.nl (V. Emonet); michel.dumontier@maastrichtuniversity.nl (M. Dumontier)



© 2023 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)

companies and researchers started to pay more attention to the computational drug repurposing methods in the recent years in order to find a faster and effective path for the treatment of COVID-19 [2]. Although the performance of many computational drug repurposing frameworks are quite promising such as the PREDICT method which utilizes similarity search based on the principle of "guilt by association" [3], their inner working mechanisms are still seen as a black-box as the way these frameworks make decisions are not entirely evident [4]. Hence, this limitation restricts the full adoption process of computational drug repurposing methods by institutions.

Explainable AI could play a critical role in addressing to the limitations which are caused by the highly complex, non-transparent nature of the computational drug repurposing models and help us understand and interpret the underlying models, in order to mitigate the lack of interpretability of certain machine learning models and to augment human reasoning and decision-making. In the biomedical context, alongside the natively interpretable models such as Random Forests or Decision Trees, Knowledge Graphs (KG), which are semantically rich, interlinking data structures that formally represents the relationship between different entities, have started to be leveraged [5]. Knowledge Graphs (KGs), graph-based representations of knowledge, are capable of encoding the complex in the form of structured statement, in the way that it is human interpretable [6].

In this paper, our approach, which is a knowledge graph based explainable AI approach in the specific context of drug repurposing, is proposed. The proposed approach involves extracting relevant subgraph, given a drug and a disease pair, in order to provide similarity-based explanations in the form of a knowledge graph as an explainable extension of the PREDICT method [3] which is based upon the principle of "Guilt by Association" and easily adaptable to the knowledge graph structure. This method is then evaluated by conducting a case study on drug candidates, which could potentially treat Alzheimer disease. It was made accessible to the user as a branch of the OpenPredict [7] model, which is the concrete implementation of a drug-repositioning framework. The outline of this paper is as follows. A more in depth information about the related work done by other researches is provided in Section 2, and it is followed by the methodology which is adopted in this research in Section 3. The results and discussion are presented in Sections 4 and 5 respectively.

2. Related Work

AI-based drug repurposing is defined as the identification, prediction and evaluation of new use cases and indications for existing and approved drugs using computational methods, such as Machine Learning and Deep Learning. One of the most effective computational approaches utilized in the context of drug repurposing is consideration similarities between entities, specifically by analyzing the drug and disease based similarities, as well as the their combined similarities.

In this sense, the PREDICT method presents a framework that provides predictions on novel associations between desired drugs and diseases [3]. The framework is mainly based on the 'Guilt by Association' (GBA) approach which was first proposed by Chiang and Butte, which involves the measurements of similarities among the known drug and disease to drug-disease pairs, given a target query drug and disease. The known associations between entities and the

later formed associations are used as features and then are fed into a classification algorithm in order to provide a final prediction. PREDICT is a quite effective framework as it enables the incorporation of additional features related to similarity between drugs and diseases. However, the usability of this effective framework itself holds some limitations as the underlying features and reasoning behind the final predictions made by the framework, since both predictive and interpretative features are isolated by the complex classification algorithms [8].

Augmentation of Knowledge Graphs into AI systems in the biomedical field, specifically for the drug discovery and repurposing tasks, allows for generating explanations of the system by providing informed and labeled visualizations by converting knowledge formalization rules and logic into a form, which is more suitable for human comprehension, to the user. For instance, Edwards et. al [9]. in their study on Explainable Biomedical recommendations via reinforcement learning, propose a neuro-symbolic approach which involves the application of multi-hop neural driven recommendation to complex biomedical knowledge graphs . They conclude that such KG based approaches has a great potential to generate explanations and improve the performance of the black-box methods.

Similarly, Liu et.al. [10] in their study regarding Neural Multi-Hop Reasoning with logical rules on Biomedical Knowledge graphs propose a novel neuro-symbolic approach PoLo (Policy-Guided Walks With Logical Rules) that leverages the interpretability and the structure of Knowledge Graphs to conduct guided policy walks. The experimental findings that they have found for this specific approach based on KGs, on the use case of drug repurposing of the novel disease COVID-19, demonstrated that path-based reasoning methods outperform existing black-box methods on the drug repurposing task as well as providing a natural transparency mechanism which makes this approach more transparent to the existing black-box methods.

Moreover, Wang et. al. [11], in their study on discovering the potential reactions of antitumor drugs adopted a Tumor-Biolink knowledge graph (TBKG) based method which is comprised of four main steps including (1) graph building, (2) reaction discovery, (3) graph verification, (4)clinical validation, and in which they explored the relations among tumors, biomarkers and drugs. It is concluded in the study that the generated knowledge graphs could have successfully been interpreted and validated by the domain experts and therefore, their approach is capable of providing explanations and transparency of their reaction discovery process.

Inherently explainable predictive models such as Decision Trees and Classification Rules as well as biomedical Knowledge Graphs are utilized for the drug interaction tasks to bring its explainability to a higher level [12]. Bresso et.al. utilize these simple classification methods to develop an explainable AI system for investigating drug interactions and they have used Decision Trees to make predictions from the generated Knowledge Graphs. Along with the quantitative performance metrics they also conduct qualitative experiments with the domain experts for explainability, similarly to the clinical validation step in Wang et. al's study. It demonstrates that the synthesis of knowledge graphs with inherently explainable prediction methods provide explainable and comprehensible models to explore activity reactions of drugs.

3. Method

In order to provide an interpretable drug repurposing framework, we developed a knowledge graph based pipeline. The primary purpose of this pipeline is to generate a knowledge graph which indicates two types of relationship; *similar_to* which is the similarity between the drug-drug and disease-disease pairs, and the *treats* relationship between a drug and a disease.

The base information regarding the similarity between drug-drug & disease-disease pairs and the relations between drug-disease pairs are curated into a dataset which includes the vector embeddings of 593 drugs obtained from *DrugBank* and 313 associated diseases from *Online Mendelian Inheritance in Man, (OMIM)* databases.

The overall strategy includes identifying a set of ranked paths through the generated Knowledge Graph that provide plausible explanations for a predicted drug indication based on their similarities to known drug-disease pairs. The generated explanation is based on a input which is composed of a desired drug-disease pair. Our approach generates a KG of ranked paths in three steps : Path Generation, Path Ranking and the Generation of Explanation Graph.

Path Generation step involves creating a set of paths based on a given input. Each path has n -length, consisting of two types of relations; *similar_to* between drug-drug and *treats* between drug-disease entities. The level of similarity between two entities, E_1 and E_2 , is obtained by taking the cosine similarity S_C of their vector embeddings which is given by :

$$S_C(E_1, E_2) = \frac{E_1 \cdot E_2}{\|E_1\| \|E_2\|} \quad (1)$$

As the paths of length $n > 3$ are less biochemically relevant and the increasing path lengths become increasingly difficult to understand, 0, 1 and 2-hop relation paths are used to connect the drug and the disease.

Five cases are taken into account during the path formation. As Figure 1 demonstrates, the *treats* relation among the known drug-disease pairs are retrieved immediately (*drug2-disease1*). For the unknown drug-disease pairs, for instance *drug1-disease1*, the structural similarity between the homotypic times should also be considered in order to form an edge which represents the relation *treats*. In this specific case, *drug1* is similar to *drug2*, which is known to be indicated for the treatment of *disease1*, and similarly, it is known that *disease1* is similar to *disease2*. Therefore, it is possible to form an edge between the entities *drug1* and *disease1*. However, the plausibility of this edge depends on the similarity scores that the other existing edges have, and in practice, the number of paths between two predicted entities is quite high due to the size of the data sets. Therefore, an additional graphical action is needed to rank the weight of the formed paths and select the most relevant paths to form the *treats* relation, based on an input drug-disease pair.

The path ranking operation is achieved by the adoption of principle of parsimony, which suggests that explanations with simpler and shorter paths are more relevant compared to the paths that are longer, and might have relatively less indirect information. Each path formed in the previous phase are ranked according to their assigned weight. The weight w_{π_k} , assigned to each path π_k is computed by :

$$w_{\pi_k} = \sum_{e_i \in \pi_k} e_i \quad (2)$$

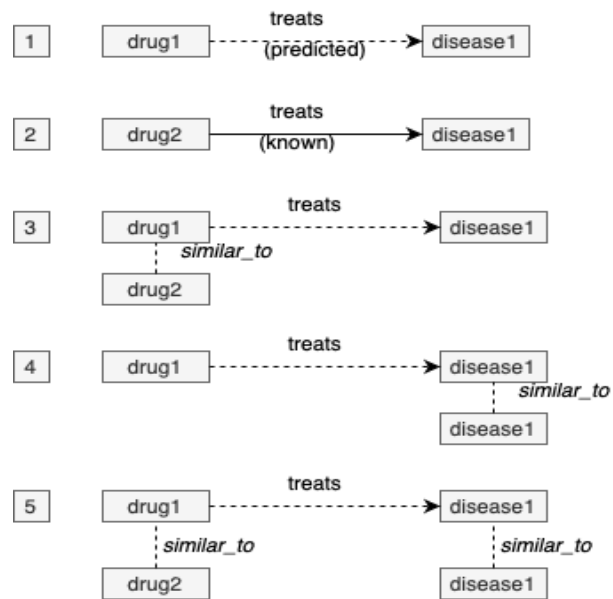


Figure 1: Knowledge graph based explanation.

where the edge weight e_i is set to 1 in case of a relation of type *treats*, and to $(1 - similarity)$ in the edge connects two entities of the same type. Once a weight is assigned to each path in the KG, they are ranked by their weights in an ascending order. As a last step, top k weighted paths with highest weights are included in the explanation graph. Moreover, to ensure the relevance and significance of the included entities in the final explanation, as well as simplifying the graphs to provide more readable explanations, we introduce an additional binary variable "*min_similarity_threshold*", which restricts the amount of included entities further, according to a desired similarity threshold. The restriction process is achieved by taking the entities, which are in the top n percentile of all entities, in terms of their similarity scores to the target. If there are no entities which satisfy a certain similarity threshold, i.e, the similarity between the found entities are too weak, an empty explanation graph is returned.

4. Use case

In order to observe the effectiveness of the generated explanation through the Knowledge Graphs, a case study for Alzheimer Disease (OMIM:104300), which is indicated to carry similar characteristics with diseases such as dementia and Parkinson's disease [13], was conducted through the OpenPredict API, using the PREDICT dataset and model.

The drug Amandatine is suggested by the PREDICT model as a potential treatment for Alzheimer's disease. In order to understand the relation between the drug Amandatine and Alzheimer Disease, we use this pair as an input to the pipeline as shown in Figure 2.

Considering the above input, the generated explanation graph, the pipeline first extracts the individual edges, then augments them into paths, as illustrated in Figure 3.



Figure 2: Amantadine is predicted for the treatment of Alzheimer Disease.

Amantadine is structurally similar to Donepezil, and Rivastigmine which are directly indicated for the treatment of Alzheimer Disease. It is also shown to be similar to drugs Carbidopa, Zonisamide and Haloperidol which treat Parkinson’s Disease, Epilepsy and Dementia & Schizophrenia, which are similar to Alzheimer, respectively. The generated paths are then merged into a single knowledge graph, displaying the relationships between the drug-disease entities as a complete semantic network as shown in Figure 4. The resulting explanation graph provides plausible explanations as the entities included in the graph are closely related to the target pair (Amantadine-Alzheimer). For instance, many studies have shown that there is strong evidence that Parkinson’s Disease and Alzheimer Disease have overlapping similarities in terms of clinical and neuropathologic features [14] and Carbidopa is indicated for treatment of early symptoms of Alzheimer [15] [16].

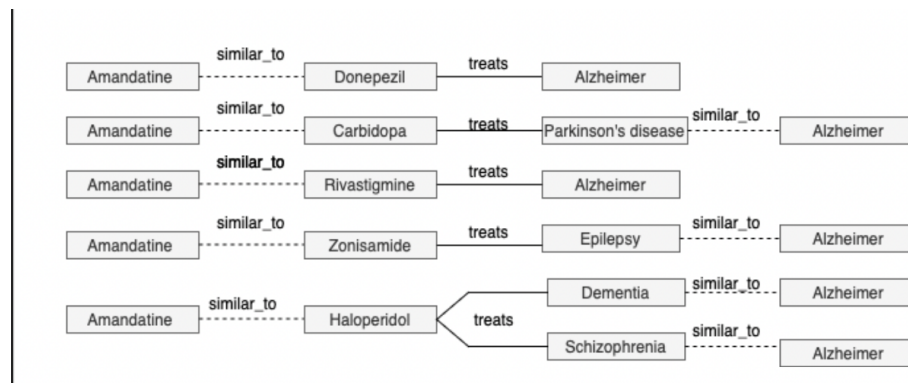


Figure 3: The paths formed by the pipeline given Amantadine-Alzheimer pair as the input

The *min_similarity_threshold* is taken as 10 in this case study, considering the availability of the instances in the data. In order to observe whether the variable *min_similarity_threshold* causes loss of information in this specific case study, an alternative explanation is generated without taking *min_similarity_threshold* into consideration. The paths formed, that are not restricted by a certain threshold, turned out to be indeed more populated with entities, as seen in Figure 5. The result is quite interesting as in this example, *min_similarity_threshold* indeed reduced the size of the explanation graph in a way that the entities included in the graph are more relevant. In Figure 5, Clonidine is shown to be similar to Amantadine. Clonidine is known to be indicated for the treatment of Gilles La Tourette Syndrome, which is a disease mostly related to neuropsychiatric movement and typically starts developing from childhood [17]. For the entity Carbidopa, in comparison with the explanation graph restricted with the similarity threshold, Multiple Sclerosis, a neuroskeletal disorder [18], is also shown to be similar to the

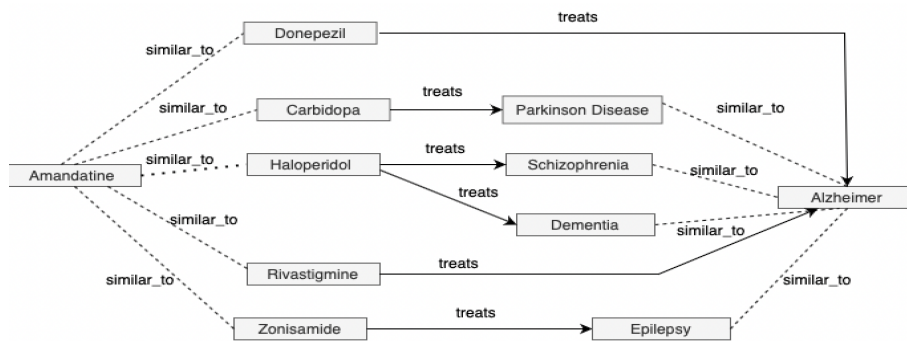


Figure 4: The final explanation graph.

Alzheimer Disease. In this context, it is possible to say that these diseases are relatively less related to Alzheimer, therefore the utilization of *min_similarity_threshold* enabled excluding less relevant entities.

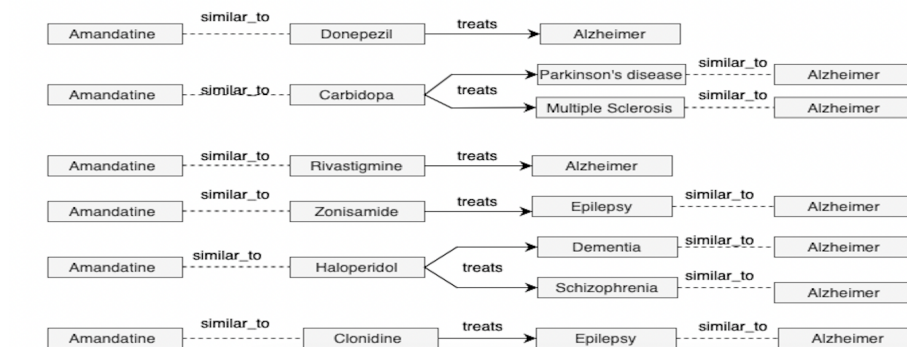


Figure 5: The paths formed without being restricted by *min_similarity_threshold*

5. Discussion

The results that have been obtained from the conducted case study demonstrated that the building semantic connections using Knowledge Graphs could provide meaningful and effective explanations in the biomedical, specifically drug repurposing, domain. Although this study is mainly focused on drug repurposing domain, it is intended to show that the proposed pipeline, which takes the Knowledge Graph structure as its main baseline, is a powerful pipeline to generate plausible explanations.

Although, the literary sources and previous studies were taken as a basis to qualitatively evaluate how effective the proposed pipeline is, in generating explanations, evaluation of conducted case studies by domain experts could be a further and a more reliable justification.

In this sense, this comes as the main limitation in the evaluation process. Furthermore, another limitation in terms of applicability in other domains and cases could be that the path ranking process might be problematic as an edge could have a dominantly large weight, especially, if the weights are not normalized. Therefore, it would be sensible to consider the alternative path ranking strategies, such as finding the shortest path in the graph, as well as the one used in this pipeline. Exclusion of entities with lower similarities through determined thresholds simplify the outputted knowledge graphs, allowing for easier interpretations by the domain experts. In order to prevent the possible hindrances, knowledge graphs generated using different thresholds could be observed.

Computational drug repurposing methods are still not fully adopted by the institutions due to the lack of explainability behind the sophisticated methods [1]. In this sense, utilization of Knowledge Graphs could help domain experts to augment the explanations provided with their expertise and reasoning to gain more insight on the studied subjects. It could also encourage considering the drug-disease relations that have not been studied yet as the Knowledge Graph explanation visualizes not only the entities related to the target but also the entities related to the intermediate entities along the paths.

This method has also drawn some challenges that are still yet to be tackled. For instance, considering more complex relations such as the interaction between the target and the intermediate drugs may foster obtaining a deeper level of understanding of the treatment potential of the target disease by the prospective drug. Another challenge might be the augmentation of new drug and disease information to the pipeline. The vector embedding conversion is easily performed as a reproducible strategy is adopted, however augmentation of large information could cause redundancy and sometimes loss of information due to the larger filtering and simplification which would be performed in parallel with the increasing search space.

Overall, the case study conducted on the proposed pipeline is an indicator of the promising potential of Knowledge Graphs, and semantic operations that come with it, in providing transparent and understandable explanations in the biomedical domain, and the challenges that it introduces are an incentive to enhance KG-Based Explainable AI methods in the domain.

6. Conclusion

In this work, a knowledge graph based explanation framework is proposed for drug repositioning task. The proposed approach could be utilized to provide explanations and improve the main principles of Explainable AI, by providing accountability, reliability and transparency regarding the decisions that were made through computational methods.

The proposed framework took the PREDICT method as a baseline in providing the explanations. This way, by enhancing the Guilt by Association strategy that PREDICT method uses by augmenting KGs, along with several graphical operations, the relations between the related entities to the given drug-disease pairs are demonstrated as transparent explanations to the users in the form of structured predicates.

References

- [1] J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence, *Nature Machine Intelligence* 2 (2020) 573–584.
- [2] S. Ekins, M. Mottin, P. R. Ramos, B. K. Sousa, B. J. Neves, D. H. Foil, K. M. Zorn, R. C. Braga, M. Coffee, C. Southan, et al., Déjà vu: stimulating open drug discovery for sars-cov-2, *Drug discovery today* 25 (2020) 928–941.
- [3] A. Gottlieb, G. Y. Stein, E. Ruppin, R. Sharan, Predict: a method for inferring novel drug indications with application to personalized medicine, *Molecular systems biology* 7 (2011) 496.
- [4] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, *Proceedings of the National Academy of Sciences* 116 (2019) 22071–22080. doi:<https://doi.org/10.1073/pnas.1900654116>.
- [5] A. Callahan, N. H. Shah, Machine learning in healthcare, in: *Key Advances in Clinical Informatics*, Elsevier, 2017, pp. 279–291.
- [6] I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey, *Artificial Intelligence* 302 (2022) 103627.
- [7] R. Celebi, J. R. Moreira, A. A. Hassan, S. Ayyar, L. Ridder, T. Kuhn, M. Dumontier, Towards fair protocols and workflows: the openpredict use case, *PeerJ computer science* 6 (2020) e281.
- [8] E. Bresso, P. Monnin, C. Bousquet, F.-E. Calvier, N.-C. Ndiaye, N. Petitpain, M. Smail-Tabbone, A. Coulet, Investigating adr mechanisms with explainable ai: a feasibility study with knowledge graph mining, *BMC medical informatics and decision making* 21 (2021) 1–14. doi:<https://doi.org/10.1186/s12911-021-01518-6>.
- [9] G. Edwards, S. Nilsson, B. Rozemberczki, E. Papa, Explainable biomedical recommendations via reinforcement learning reasoning on knowledge graphs, 2021. doi:10.48550/ARXIV.2111.10625.
- [10] Y. Liu, M. Hildebrandt, M. Joblin, M. Ringsquandl, R. Raissouni, V. Tresp, Neural multi-hop reasoning with logical rules on biomedical knowledge graphs, in: *European Semantic Web Conference*, Springer, 2021, pp. 375–391. doi:<https://doi.org/10.48550/arXiv.2103.10367>.
- [11] M. Wang, X. Ma, J. Si, H. Tang, H. Wang, T. Li, W. Ouyang, L. Gong, Y. Tang, X. He, et al., Adverse drug reaction discovery using a tumor-biomarker knowledge graph, *Frontiers in genetics* 11 (2021) 625659. doi:10.3389/fgene.2020.625659.
- [12] E. Bresso, P. Monnin, C. Bousquet, F.-E. Calvier, N.-C. Ndiaye, N. Petitpain, M. Smail-Tabbone, A. Coulet, Investigating adr mechanisms with explainable ai: a feasibility study with knowledge graph mining, *BMC medical informatics and decision making* 21 (2021) 1–14. doi:<https://doi.org/10.1186/s12911-021-01518-6>.
- [13] C. Reitz, C. Brayne, R. Mayeux, Epidemiology of alzheimer disease, *Nature Reviews Neurology* 7 (2011) 137–152.
- [14] D. P. Perl, C. O. Warren, D. Calne, Alzheimer’s disease and parkinson’s disease: distinct entities or extremes of a spectrum of neurodegeneration?, *Annals of neurology* 44 (1998) S19–S31.
- [15] C. S. Okereke, L. Kirby, D. Kumar, E. I. Cullen, R. D. Pratt, W. A. Hahne, Concurrent

administration of donepezil hcl and levodopa/carbidopa in patients with parkinson's disease: assessment of pharmacokinetic changes and safety following multiple oral doses, *British journal of clinical pharmacology* 58 (2004) 41–49.

- [16] P. Chopade, N. Chopade, Z. Zhao, S. Mitragotri, R. Liao, V. Chandran Suja, Alzheimer's and parkinson's disease therapies in the clinic, *Bioengineering & Translational Medicine* (2022) e10367.
- [17] E. Jakubovski, K. R. Müller-Vahl, Gilles de la tourette syndrome: symptoms, causes and therapy, *Psychotherapie, Psychosomatik, Medizinische Psychologie* 67 (2017) 252–268. doi:<https://doi.org/10.1186/s12911-021-01518-6>.
- [18] N. Ghasemi, S. Razavi, E. Nikzad, Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy, *Cell Journal (Yakhteh)* 19 (2017) 1.