Generating Drug Repurposing Hypotheses through the Combination of Disease-Specific Hypergraphs

First Author 1[*] ABC@SAMPLE.COM
University X, Country 1

First Author 2[*] DEF@SAMPLE.COM
University Y, Country 2

Last Author GHI@SAMPLE.COM
University Z, Country 3

Abstract

The drug development pipeline for a new compound can last 10-20 years and cost over $10 billion. Drug repurposing offers a more time- and cost-effective alternative. Computational approaches based on biomedical knowledge graph representations, comprising a mixture of nodes (e.g., disease, drug, protein, symptom, side effect) and their interactions, have recently yielded new drug repurposing hypotheses, including suitable candidates for COVID-19. In this study, we present a novel, disease-specific hypergraph representation learning technique to derive contextual embeddings of biological pathways of various lengths but that all start at any given drug and all end at the disease of interest. Further, we extend this method to multi-disease hypergraphs. Specifically, we combine the hypergraph of the disease of interest with those of its main associated risk factors to guide the search of repurposing candidates towards likely relevant drug classes. We determine the repurposing potential of 1,522 drugs based on the median cosine similarity between all biological pathways starting at that drug and ending at the disease of interest and all biological pathways starting at drugs currently prescribed against that disease and ending similarly. We illustrate our approach with Alzheimer's Disease (AD), which affects over 55 million patients worldwide but still has no cure. We compare each drug's rank across four hypergraph settings (single- or multi-disease): AD only, AD + Hypertension (HTN), AD + Type 2 Diabetes (T2D), and AD + HTN + T2D. Notably, our framework led to the identification of two promising drugs, i.e., whose repurposing potential was significantly higher in two-disease combined hypergraphs: dapagliflozin (antidiabetic; moved up, from top 32% to top 7%, across all considered drugs) and debrisoquine (antihypertensive; moved up, from top 76% to top 23%). Our approach serves as a hypothesis generation tool, to be paired with a validation pipeline relying on lab experiments and semi-automated parsing of the biomedical literature.

Keywords: Precision Medicine, Drug Repurposing, Disease Specificity, Hypergraphs

1. Introduction

The development of new drugs can take more than 15 years, from the discovery and pre-clinical phase to review by regulatory agencies Xue et al. (2018). Hence, repurposing drugs already approved by the Food and Drug Administration or European Medicines Agency serves as a convenient alternative since they have already known to be safe in human populations. From a research and development perspective, drug repurposing is a less risky enterprise. Indeed, following compound identification, repositioned drugs would generally hit the market in less than 10 years. Beyond time savings, this strategy brings significant cost savings, potentially reducing the average pharmaceutical pipeline's budget by over $5 billion compared to traditional drug development. To date, drug repurposing encompasses three main approaches: computational biomedicine Jarada et al. (2020), biological experimentation, and their combination, e.g., through systems pharmacology Zhao and Iyengar (2012).

Computational approaches are both more time-effective and cost-effective than *in vitro* or *in vivo* biological experiments, which involve high-throughput

[*] These authors contributed equally

screening or phenotypic screening based on animal and human models, respectively. Examples of available strategies include signature matching, genome-wide association studies, and the retrospective analysis of real-world clinical information Wu et al. (2022). Their use has been unlocked by the concurrent emergence of technical advances such as biological microarrays and the increase in data accessibility, as illustrated by the rapid growth of electronic health records and biobanks Tan et al. (2023).

Simultaneously, massive genomic databases and cell lines have yielded 20+ high-quality biological and biomedical knowledge graphs (KG) such as SPOKE Morris et al. (2023) and PrimeKG Chandak et al. (2023) and aggregating platforms such as the KG-Hub noa to ensure that the former can be shared and made interoperable for downstream graph machine learning tasks. Network-based methods for drug repurposing rely on the encoding of interactions between entities (i.e., drugs, diseases, proteins, biological functions) that can be heterogeneous (i.e., inhibition, binding). These representations can help address both predictive (e.g., polypharmacy side effects) and inferential (e.g., reasoning over causal pathways) questions. Prior graph representations such as the multi-scale interactome (MSI) Ruiz et al. (2021) have proved useful in identifying known drug repurposing agents and formulating potential candidates.

In a previous study, we have shown how a disease-specific hypergraph representation learning technique could identify likely repurposing targets that were missed by the multiscale interactome Jain et al. (2023). Building on the promise of disease-specific hypergraph representation learning for the identification of suitable drug repurposing candidates by biological pathway similarity search, we propose to combine knowledge graph information pertaining to a disease of interest and co-morbid diseases. The objective is to assess how this ”perturbation” may affect the ranks of drugs currently prescribe to mitigate the repercussions of comorbidities. Specifically, we hypothesize that combining disease-specific hypergraphs of co-morbid diseases (e.g., HTN and T2D) with that of the condition of interest (e.g., AD) will boost the ranks of their respective antagonist drugs (antihypertensives and antidiabetics, respectively) upwards. Our findings will hopefully support the design of disease-specific network representation models and yield new drug repurposing insights. Additionally, our framework could enable precision medicine through combined hypergraphs of increasing granularity that closely match the disease profile of a pre-specified patient population.

2. Methods

2.1. Hypergraph Construction and Combination

The Hetionet is accompanied by a graph database that allows us to query the paths starting at 1 of 1,522 drugs and ending at a disease of interest. It returns us the paths in groups of metapaths where one metapath group might be defined as ”Drug-protein-protein-disease,” for example. This functionality allows us to focus on the biological paths: the metapaths with that contain one or more protein nodes. If after we have aggregated all of the biological paths starting at a drug and ending at a disease of interest, we have more than 10,000 paths, we select only the top 10% of paths in each remaining biological metapath group. The top 10% of paths are found by their direct weighted path count, a metric that quantifies path significance to a metapath category by the number of connections each node in the path has in the graph Himmelstein et al. (2023).

After this filtering, we have selected an induced disease specific subgraph from the aggregate heterogeneous knowledge graph (See Figure 1 (a), (b), (d)). Then, we unify the biological paths under hyperedges to create a disease specific hypergraph (See Figure 1 (c) and (e)). Finally, we turn these disease specific hypergraphs into a weighted graph where the nodes represent biological paths and they are connected by an edge if they share one or more middle nodes. A middle node here is defined by a node not at the start or end of the path, since the start of the path is a drug and the end of the path is the disease that all paths of the same disease would share in common. We defined the edge connection of two paths in this way to focus on the biological relationships between the paths that is captured by the middle node similarities between paths (see Figure 1 (f)). The weight, w, on each edge connecting the paths is directly proportional to the number of similar middle nodes between two paths. Each weight is normalized to be in this interval: (0,1].

2.2. Hypergraph Representation Learning

We learned biological pathways by initiating a random walk on the transformed graph delineated in Figure 1 (d), commencing from any of the nodes with a

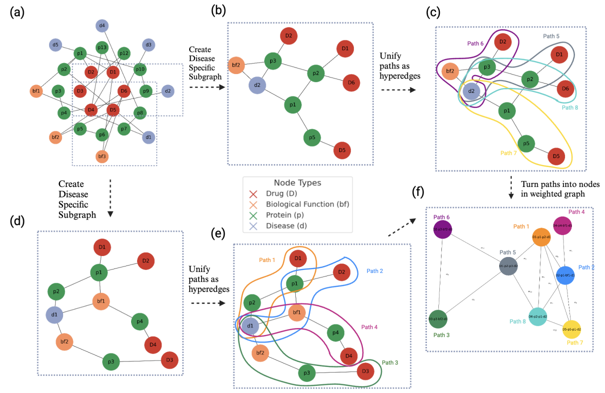

Figure 1: Creating and combining two disease-specific hypergraphs into a weighted graph

drug currently prescribed as the first element on the biological pathway. We accounted for the presence of weighted edges by sampling neighboring nodes proportionally to the strength of the connection. The random walker began at a selected node, then proceeded iteratively to an adjacent node chosen at random, and repeated this process for a predetermined number of steps. Each random walk was replicated 10 times for every start node in our projected graph, with a walk path length set at 80.

At each step of the random walk, the probability of transition from biological pathway v to biological pathway x, where the weighted edge between two pathways is w is expressed as:

$$P(v_i = x | v_{i-1} = v) = \frac{w \text{ between } v \text{ and } x}{\text{sum of all } w \text{ leaving } v} \tag{1}$$

2.2.1. Skip-Gram Model

We interpreted the resulting random walks as sentences, utilizing the Word2Vec Skip-Gram model provided by gensim to develop node embeddings for each biological pathway Grover and Leskovec (2016). This model predicts context words (nodes within the same walk) given a target word (a node). Applied to the context of our disease-specific weighted graph, the embeddings of biological pathways learned through this process encapsulate the local neighborhood structure of the nodes and are subsequently used for our pathway similarity search. The skip-gram part of the algorithm is elaborated upon in Section B.

2.3. Evaluating Path Embeddings

Given a specific disease of interest, our study aimed to identify biological pathways analogous to those associated with drugs currently used to treat it. In particular, we conducted a case study on Alzheimer's disease and considered medications prescribed to alleviate the symptoms and behavioral complications of Alzheimer's Disease (AD). We focused primarily on three compounds: donepezil, memantine, and galantamine (reference drugs) Tan et al. (2014); Howes (2014); Bond et al. (2012), approved by the FDA in 1996, 2003, and 2001, respectively. Additionally, we not only learned the embeddings of an AD only hypergraph, but we combined it with two of its main risk factors: HTN and T2D. Our approach is disease-agnostic and can be extended to other diseases, upon the supply of a list of compounds currently used in clinical practice or previously suggested as repurposing candidates.

We represented each pathway as a vector: pathways starting at a drug were denoted as vector A, and those ending at AD or a reference drug were denoted as vector B.

To quantify the similarities between these pathways, we calculated the cosine similarity for each pair of A and B vectors, which led to a distribution of cosine similarities for each drug. Subsequently, we identified the median cosine similarity for every drug and ranked each drug from 1 (highest median similarity) to 1,552 (lowest median similarity). This process can be mathematically articulated as:

$$R_d = \text{rank}\left(\text{median}\left(\left\{\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}\right\}\right)\right) \tag{2}$$

where R_d is the rank of drug d, and A and B represent the path vectors as defined earlier.

Exhaustive literature review on a few of the repurposing suggestions from each method are provided to support their potential for AD. Further, ranks were compared across hypergraphs to observe the effect of combining comorbidities on the task of learning pathway embeddings. Full lists of suggestions as well as code used to generate them are available here.

3. Results

3.1. Comparing graph construction and embedding overlap

We constructed four distinct hypergraphs of increasing size: AD only, AD + HTN, AD + T2D, and AD + T2D + HTN. We compare the overlap of their top biological pathway embeddings (ranked by median cosine similarity between all paths starting at a drug and ending at AD and all paths starting at galantamine, donepezil, or memantine and ending at AD). Notably, differences appeared in the composition of the top 6% there was less than 100% overlap when comparing in turn each of the three multi-disease hypergraphs to the AD only one. However, for all three combined hypergraphs, full overlap was reached beyond the top 6% of pathway embeddings (see Figure 2 (a)).

Additionally, while the clustering coefficient is high overall across all hypergraph settings (see Figure 2(b)), we noticed that the level of clustering in the projected, weighted graph decreases as we combine a disease and its risk factors into a fused hypergraph. The observation of a lesser extent of triadic closure suggests that overlaying disease-specific hypergraphs is effectively adding information.

Concurrently, the weighted graph density decreases as the number of diseases in the underlying hypergraph increases, from AD only to AD + HTN + T2D. The latter has the lowest density of all three combined weighted graphs (See Figure 2 (d)), a trend similar to that observed with the clustering coefficient. However, the AD + T2D weighted graph has a lower density than its AD + HTN counterpart (Figure 2(d)). This result is not consistent with the above-described clustering coefficient differences (Figure 2(b)).

Lastly, we note that there is limited variability in the composition of gene nodes present in each hypergraph (See Figure 2(c)), besides some shifting in the top 6% pathway embeddings when the T2D hypergraph is added to that of AD. This result suggests the existence of some redundancy in the distribution of underlying biological pathways, the likely source of robustness to hypergraph perturbation observed in our AD case study (i.e., removing or adding comorbidities from the baseline, single-disease hypergraph).

4. Discussion & Future Directions

We proposed a novel hypergraph representation learning method to identify drug repurposing targets for a specific disease through the combination of its hypergraph with those of one or several of its associated comorbidities.

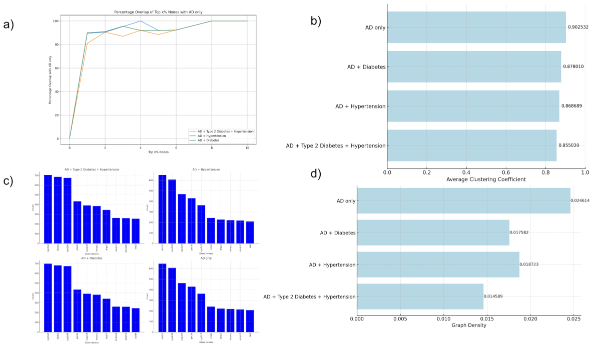

Figure 2: Comparison of Combined Hypergraphs

We showed how ranks embedding overlap varied as a disease specific hypergraphs were combined with its comorbidities. Additionally, we saw changes in the graph densities and clustering of the graphs while the gene composition stayed rather constant. Note that the existence of links among biological pathways may be specific to each tissue or organ; our current study is tissue-agnostic but future research will focus on allowing clinical experts using our tool to manually edit (e.g., adding or removing) connections that may uniquely apply to certain tissues or organs such as in brain health (e.g., hippocampus).

We also plan to do further analyses upon the ranks shifts of specific drugs to see if we can declare the drugs that improve in ranks as better repurposing candidates in situations that a patient presents both diseases. Potentially, this could lead to a future study on patients with hypertension and AD to observe the effects of the drug on this patient population.

We provide full links to all code used to generate the hypergraphs, our hypergraph representation learning algorithm, and data used to create figures here. In the future, we plan on scaling our proposed method to 800+ prevalent diseases and their known risk factors and comorbidities to formulate new therapeutic hypotheses—both single compounds and drug combinations, while also providing pathway-based explanations to help elucidate the underlying mechanisms of action.

References

KG-Hub—building and exchanging biological knowledge graphs | Bioinformatics | Oxford Academic. URL https://academic.oup.com/bioinformatics/article/39/7/btad418/7211646.

Rana Moustafa Al AdAwi, Zainab Jassim, Dina Elgaily, Hani Abdelaziz, Bhagya Sree, and Mohamed Izham Mohamed Ibrahim. Assessment of dapagliflozin effectiveness as add-on therapy for the treatment of type 2 diabetes mellitus in a qatari population. *Scientific Reports*, 9(1):6864, May 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-43052-6. URL https://doi.org/10.1038/s41598-019-43052-6.

M. Bond, G. Rogers, J. Peters, R. Anderson, M. Hoyle, A. Miners, T. Moxham, S. Davis, P. Thokala, A. Wailoo, M. Jeffreys, and C. Hyde. The effectiveness and cost-effectiveness of donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease (review of Technology Appraisal No. 111): a systematic review and economic model. *Health technology assessment (Winchester, England)*, 16(21):1–470, 2012. ISSN 1366-5278. URL https://researchonline.lshtm.ac.uk/id/eprint/210187/. Number: 21 Publisher: NIHR Journals Library.

K. Brøsen and L. F. Gram. Clinical significance of the sparteine/debrisoquine oxidation polymorphism. *European Journal of Clinical Pharmacology*, 36(6):537–547, Nov 1989. ISSN 1432-1041. doi: 10.1007/BF00637732. URL https://doi.org/10.1007/BF00637732.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, February 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-01960-3. URL https://www.nature.com/articles/s41597-023-01960-3. Number: 1 Publisher: Nature Publishing Group.

Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks, July 2016. URL http://arxiv.org/abs/1607.00653. arXiv:1607.00653 [cs, stat].

Daniel S. Himmelstein, Michael Zietz, Vincent Rubinetti, Kyle Kloster, Benjamin J. Heil, Faisal Alquaddoomi, Dongbo Hu, David N. Nicholson, Yun Hao, Blair D. Sullivan, Michael W. Nagle, and Casey S. Greene. Hetnet connectivity search provides rapid insights into how two biomedical entities are related, January 2023. URL https://www.biorxiv.org/content/10.1101/2023.01.05.522941v1. Pages: 2023.01.05.522941 Section: New Results.

Laurence Guy Howes. Cardiovascular Effects of Drugs Used to Treat Alzheimer's Disease. *Drug Safety*, 37(6):391–395, June 2014. ISSN 1179-1942. doi: 10.1007/s40264-014-0161-z. URL https://doi.org/10.1007/s40264-014-0161-z.

Ayush Jain, Marie Laure-Charpignon, Irene Y. Chen, and Ahmed Alaa. Generating new drug repurposing hypotheses using disease-specific hypergraphs. September 2023. URL https://github.com/ayujain04/hypergraph_psb_paper/blob/main/PSB_paper.pdf.

Tamer N. Jarada, Jon G. Rokne, and Reda Alhajj. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of Cheminformatics*, 12(1):46, July 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00450-7. URL https://doi.org/10.1186/s13321-020-00450-7.

John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, Adil Harroud, Lauren Sanders, Sylvain V Costes, Krish Bharat, Arjun Chakraborty, Alexander R Pico, Taline Mardirossian, Michael Keiser, Alice Tang, Josef Hardi, Yongmei Shi, Mark Musen, Sharat Israni, Sui Huang, Peter W Rose, Charlotte A Nelson, and Sergio E Baranzini. The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics*, 39(2): btad080, 02 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad080. URL https://doi.org/10.1093/bioinformatics/btad080.

Camilo Ruiz, Marinka Zitnik, and Jure Leskovec. Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications*, 12(1):1796, March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21770-8. URL https://www.nature.com/articles/

s41467-021-21770-8. Number: 1 Publisher: Nature Publishing Group.

Chen-Chen Tan, Jin-Tai Yu, Hui-Fu Wang, Meng-Shan Tan, Xiang-Fei Meng, Chong Wang, Teng Jiang, Xi-Chen Zhu, and Lan Tan. Efficacy and Safety of Donepezil, Galantamine, Rivastigmine, and Memantine for the Treatment of Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Journal of Alzheimer's Disease*, 41(2):615–631, January 2014. ISSN 1387-2877. doi: 10.3233/JAD-132690. URL https://content.iospress.com/articles/journal-of-alzheimers-disease/jad132690. Publisher: IOS Press.

George S. Q. Tan, Erica K. Sloan, Pete Lambert, Carl M. J. Kirkpatrick, and Jenni Ilomäki. Drug repurposing using real-world data. *Drug Discovery Today*, 28(1):103422, January 2023. ISSN 1359-6446. doi: 10.1016/j.drudis.2022.103422. URL https://www.sciencedirect.com/science/article/pii/S1359644622004159.

Patrick Wu, QiPing Feng, Vern Eric Kerchberger, Scott D. Nelson, Qingxia Chen, Bingshan Li, Todd L. Edwards, Nancy J. Cox, Elizabeth J. Phillips, C. Michael Stein, Dan M. Roden, Joshua C. Denny, and Wei-Qi Wei. Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension. *Nature Communications*, 13(1):46, January 2022. ISSN 2041-1723. doi: 10.1038/s41467-021-27751-1. URL https://www.nature.com/articles/s41467-021-27751-1. Number: 1 Publisher: Nature Publishing Group.

Hanqing Xue, Jie Li, Haozhe Xie, and Yadong Wang. Review of Drug Repositioning Approaches and Resources. *International Journal of Biological Sciences*, 14(10):1232–1244, July 2018. ISSN 1449-2288. doi: 10.7150/ijbs.24612. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6097480/.

Shan Zhao and Ravi Iyengar. Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. *Annual Review of Pharmacology and Toxicology*, 52:505–521, 2012. ISSN 1545-4304. doi: 10.1146/annurev-pharmtox-010611-134520.

Appendix A. Drug Rank Shifts

When we combine the AD only hypergraph with its co-morbidity hypergraphs, Hypertension and Type 2 Diabetes, we see some significant changes in some antihypertensive and antidiabetic drug ranks. Notably, our framework led to the identification of two promising drugs, i.e., whose repurposing potential was significantly higher in two-disease combined hypergraphs: dapagliflozin Al AdAwi et al. (2019) (antidiabetic; moved up, from top 32% to top 7%, across all considered drugs) and debrisoquine Brøsen and Gram (1989) (antihypertensive; moved up, from top 76% to top 23%).

Figure 3 shows the difference between AD only rank - X, where X = AD + Hypertension, AD + Type 2 Diabetes, or AD + Hypertension + Type 2 Diabetes. These results could provide insight into another use case of this method: building drug repurposing portfolios for specific patient populations based on the set of disease(s) they are presenting at the time of intervention deployment (e.g., initiation of a pharmaceutical treatment strategy).

In the future, we plan to simulate clinical trials among patient populations who have AD only, AD + Hypertension, AD + Type 2 Diabetes, and AD + Hypertension + Type 2 Diabetes to validate our hypothesis that the combining of comorbidity hypergraphs correctly selects for drugs that better work in populations with both diseases.

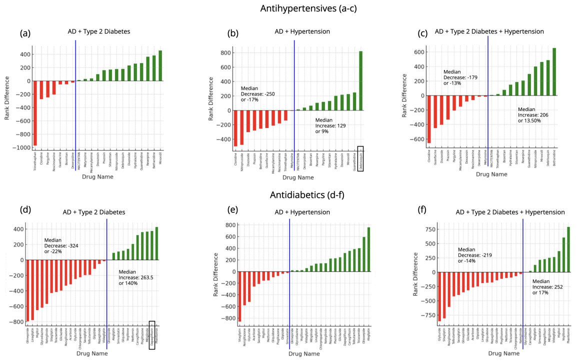

Figure 3: Rank Differences compared to AD only hypergraph

We will also further explore the literature on each drug that had a significant shift in rank once the additional biological pathways from the antagonist disease were added.

Appendix B. Skip-Gram Method Continued

The Skip-Gram model's objective is to devise word representations that effectively predict surrounding words in a sentence or document. Formally stated, given a sequence of training words $w_1, w_2, ..., w_T$, the model aims to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-k\leq j\leq k, j\neq 0}\log P(w_{t+j}|w_t) \tag{3}$$

where k denotes the size of the training context and T denotes the total number of training words. We guided the model to learn embeddings of 64 dimensions and subsequently used cosine similarity as the metric to quantify the similarity between any two biological pathways.

Our decision to utilize the Skip-Gram algorithm for learning embeddings was driven by our intent to infer semantic contextual relationships among biological pathways, given a specific disease.