Research motivation

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.

Origin of our open data project

This summer, Ayush Jain was part of the Broad Institute's BSRP undergraduate internship program (https://www.broadinstitute.org/bios/ayush-jain), working under the guidance of Marie-Laure Charpignon, an MIT IDSS/LIDS PhD student, and Dr. Anthony Philippakis at the Eric and Wendy Schmidt Center. Our project – in collaboration with Dr. Ahmed Alaa and Dr.Irene Chen at UC Berkeley Computational Precision Health – aimed to advance drug repurposing strategies through hypergraph representation learning, utilizing biological pathway similarity search.

Brief description of our dataset

Our dataset, built upon the Hetionet—a comprehensive open data knowledge graph from UC Berkeley—encompasses individual hypergraphs for 137 diseases and 5,734 side effects (~30GB of data), providing a rich resource for identifying repurposing targets across varied patient demographics: https://github.com/ayujain04/HERALD. Our work, which serves as a companion to our first study accepted at the 2024 Pacific Symposium on Biocomputing, paves the way for a new approach to drug repurposing of both chronic and rare diseases.

Contribution to the literature and open data community for biomedical knowledge graphs

In this open data repository, we present a novel, disease-specific hypergraph representation learning technique to derive contextual embeddings of biological pathways of various lengths 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.

Demonstrating the value of our open dataset

To demonstrate the potential of our database of disease-specific hypergraphs and weighted graphs, 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 hypergraphs combining two diseases: 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.

Validating our approach and scaling the impact of our open dataset

We are currently partnering with three labs to validate our approach: Mark Albers (systems pharmacology, Harvard Medical School), Clotilde Lagier-Tourenne (neurology, Harvard Medical School), and Pradeep Natarajan (cardiology, Broad) offered to test the feasibility of our drug repurposing candidates using a combination of omics data and manual review of pathways. Upon Ayush's presentation at the Broad, we were also contacted by Michal Lipinski (psychiatry, Stanley Center at the Broad) to focus specifically on neuro-psychiatric disorders. Recently, we further leveraged our open dataset to evaluate the benefits of combining several disease-specific hypergraphs to identify repurposing targets, with the hypothesis that fusing the representation of a disease with those of its underlying risk factors should guide the search towards the most suitable medications. Our study is currently under review at the Machine Learning for Health conference; we are hoping other computational and experimental teams will join us in this effort.

Future vision

Going forward, with the advent of generative AI, we hope to enable other research teams to leverage the biomedical knowledge hypergraphs that we have built for downstream tasks (e.g., creating perturbed versions of disease-specific embeddings to determine the robustness and generalizability of drug repurposing candidates). In a recent preprint, Hubert et al. [1] have released work in that direction. The authors recognized that relying on a limited collection of datasets, as is often the case in health and medicine owing to the scarcity of open data, is insufficient to assess method generalizability [2, 3, 4]. To alleviate this limitation, they proposed the synthetic generation of KGs from a small set of open biomedical databases.

References

1. Hubert N, Monnin P, d'Aquin M, Brun A, Monticolo D. PyGraft: Configurable generation of schemas and knowledge graphs at your fingertips. arXiv.org. September 7, 2023. Accessed September 15, 2023. https://arxiv.org/abs/2309.03685.

2. Celi LA, Citi L, Ghassemi M, Pollard TJ. The PLOS one collection on machine learning in health and biomedicine: Towards open code and open data. PLOS ONE. Accessed September 15, 2023. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0210232.

3. Kras A, Celi LA, Miller JB. Accelerating ophthalmic artificial intelligence research: The role of an Open Access Data Repository. Current opinion in ophthalmology. September 2020. Accessed September 15, 2023. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8095451/.

4. de Kok JWTM, de la Hoz MÁA, de Jong Y, et al. A guide to sharing open healthcare data under the General Data Protection Regulation. Nature News. June 24, 2023. Accessed September 15, 2023. https://www.nature.com/articles/s41597-023-02256-2.