PD-KG: A Multi-Omic Knowledge Graph for Parkinson's Disease Discovery

Brook Santangelo[1*], Kishore Anekalla[1*], Anatoly Buchin[1*], Srihari Radhakrishnan[1*], Harish Ramadas[1*], Kevin Ryczko[1*], Andreas Keller[2,3,4], Jordan Crivelli-Decker[1*]

[1]SandboxAQ, Palo Alto, California, USA

[2]Saarland University (USAAR), Chair for Clinical Bioinformatics, Center for Bioinformatics, Saarbrücken, Germany

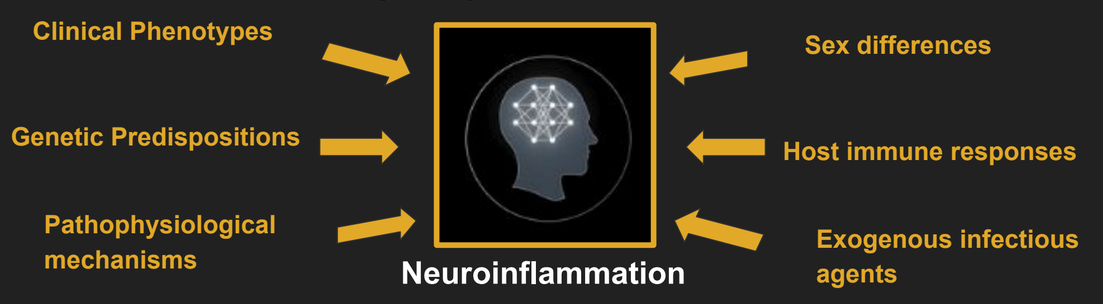
[3]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Saarbrücken, Germany

[4]Saarland University (USAAR), PharmaScienceHub (PSH), Saarbrücken,Germany

[*]These authors contributed equally

Introduction

Heterogeneity of Parkinson's disease[1]

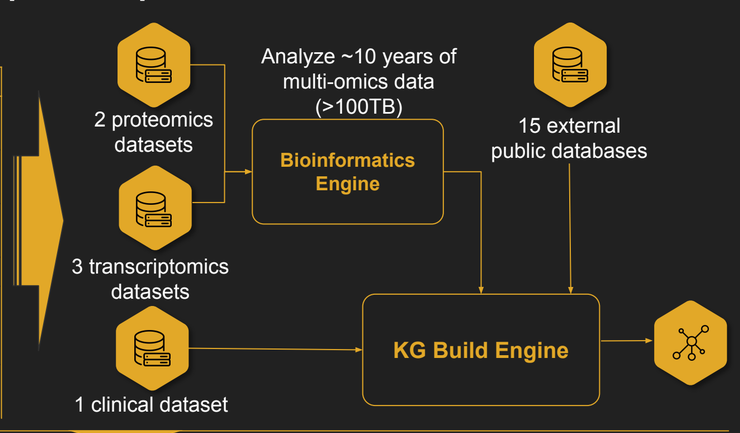


1. Varied patient progression and pathology
2. Limited number of specific biomarkers to monitor disease progression
3. No resource that properly integrates disparate, multi-omic data sources

Our Solution: A knowledge graph to empower Parkinson's disease researchers

Integration of PD Experimental Datasets

Dataset	Description
PPMI Clinical Data	Longitudinal participant-level behavioral and PD subtype data
AMP-PD Targeted Proteomics	Differential abundance analysis for CSF and Plasma
PDBMS Somascan5k	Differential abundance analysis for CSF
AMP-PD Bulk RNAseq	Differential expression analysis
Allen Brain Atlas snRNAseq (ASAP-CRN)	Cell-type specific differential expression analysis
Lange et al., 2025[2] GWAS	GWAS results



PD-KG: Comprehensive Parkinson's Disease KG
Composed of: 6.7m Edges, 222k Nodes

PD-KG Content

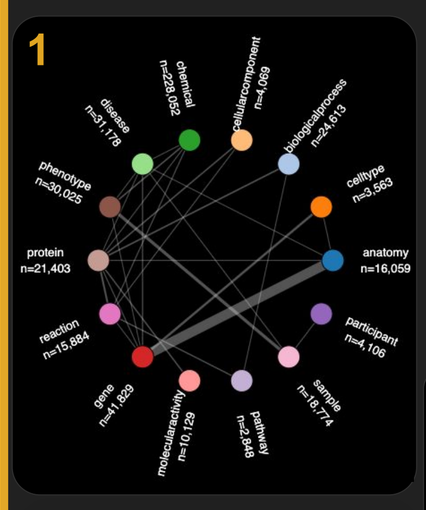
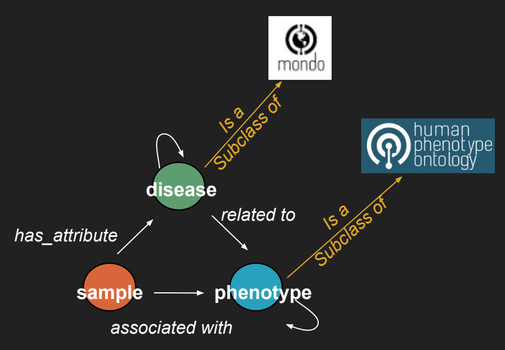


Figure 1. Schematic representation of PD-KG, according to node type and the relationships between all types. Thickness of the lines represents how many unique edges exist between nodes of those types.

Heterogeneity of PD-KG enables **genomic and phenotype-based** analyses.



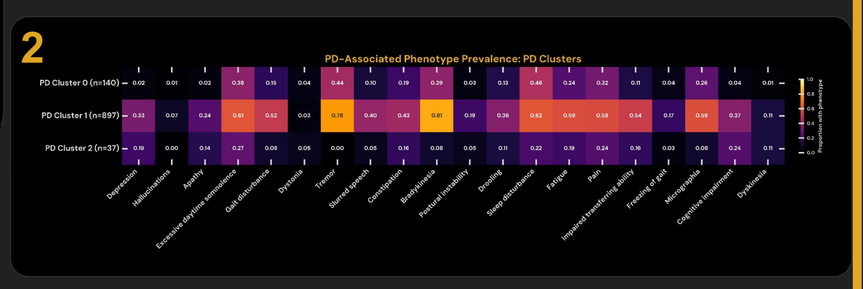
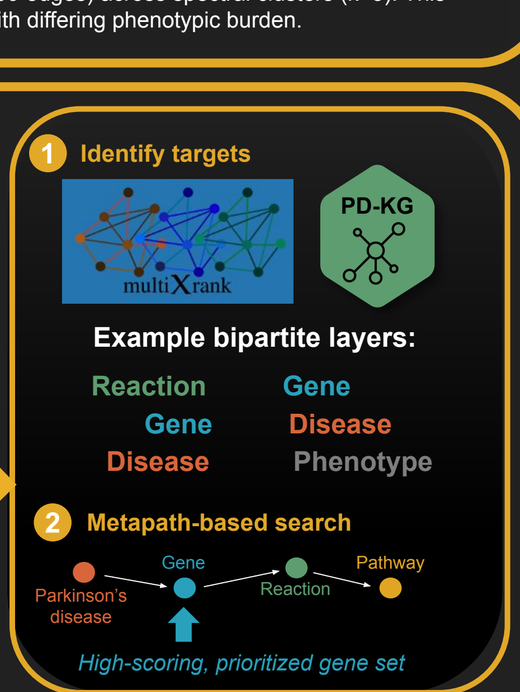


Figure 2. Heatmap showing jaccard similarity of patient clinical representation at first visit (*sample-phenotype* edges) across spectral clusters (k=3). This reveals distinct subgroups with differing phenotypic burden.

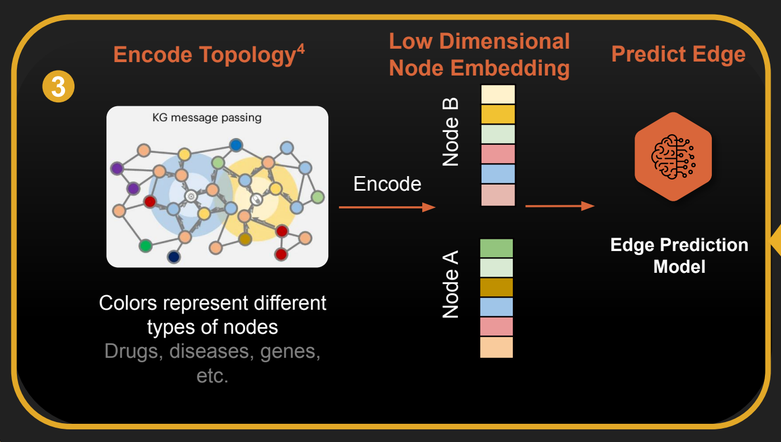
Methods

Page rank-based target prioritization using Random Walk with Restart (RWR)[3].

- Random walks across **heterogeneous bipartite layers** of PD-KG
- **Prioritize targets** (genes) based on probability scores between genes and Parkinson's disease
- Contextualize results using the broader KG heterogeneity in a **metapath-based search**



Embedding based approach to edge prediction.

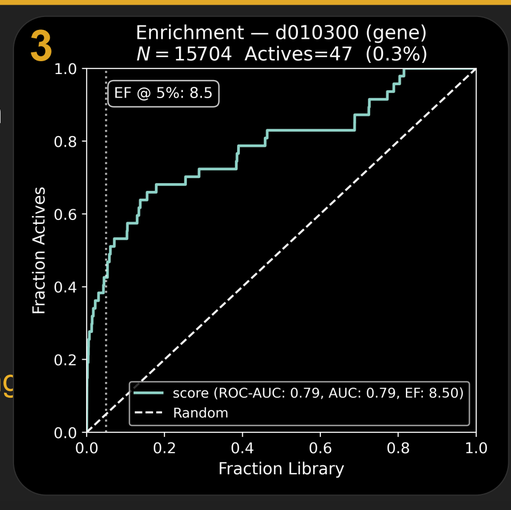


- Heterogeneous Graph Transformer (HGT) trained on edge prediction task
- Initial benchmarking utilized random splits stratified by edge type

Results

RWR-Based Target Prioritization

- Curated a list of 49 **clinically validated targets** in PD from Minikel et al.[4]
- Evaluated ranking performance of multiXrank[3]
- Evaluated pathway context using a metapath based approach



Target	Rank
UCHL1	1
SNCA	4
LRRK2	7
PRKN	10
PINK1	11
PARK7	12

Figure 3. Enrichment of known PD targets from MultiXrank gene prioritizations of 15,703 genes from PD-KG. The curve demonstrates the recovery of 49 high-confidence targets (Minikel et al.)[5] from the PD-KG genes. Well studied genes involved in PD are provided in an accompanying table, showing how they ranked.[6]

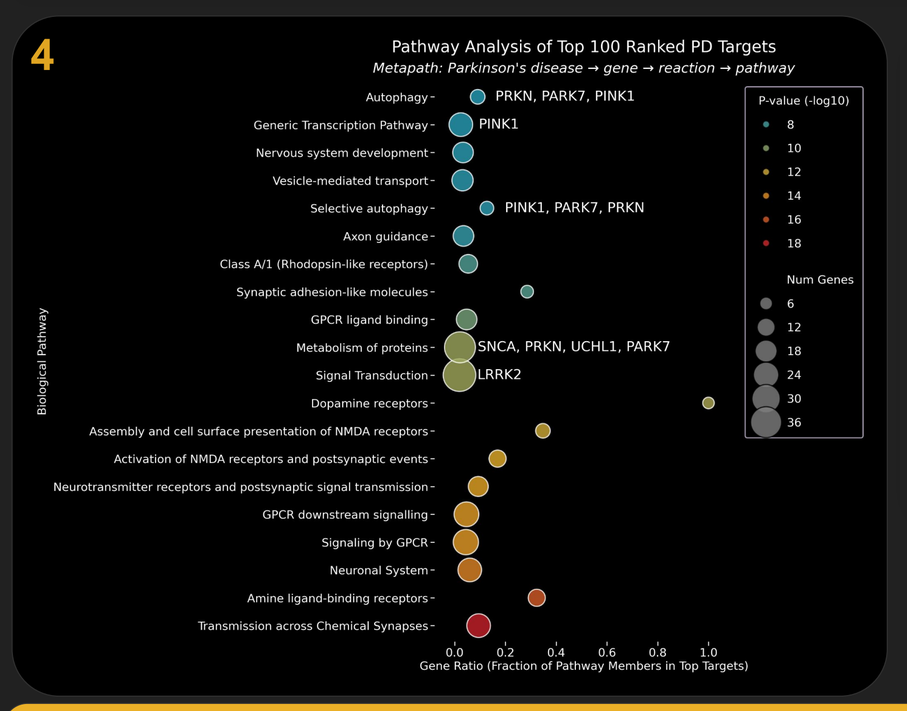
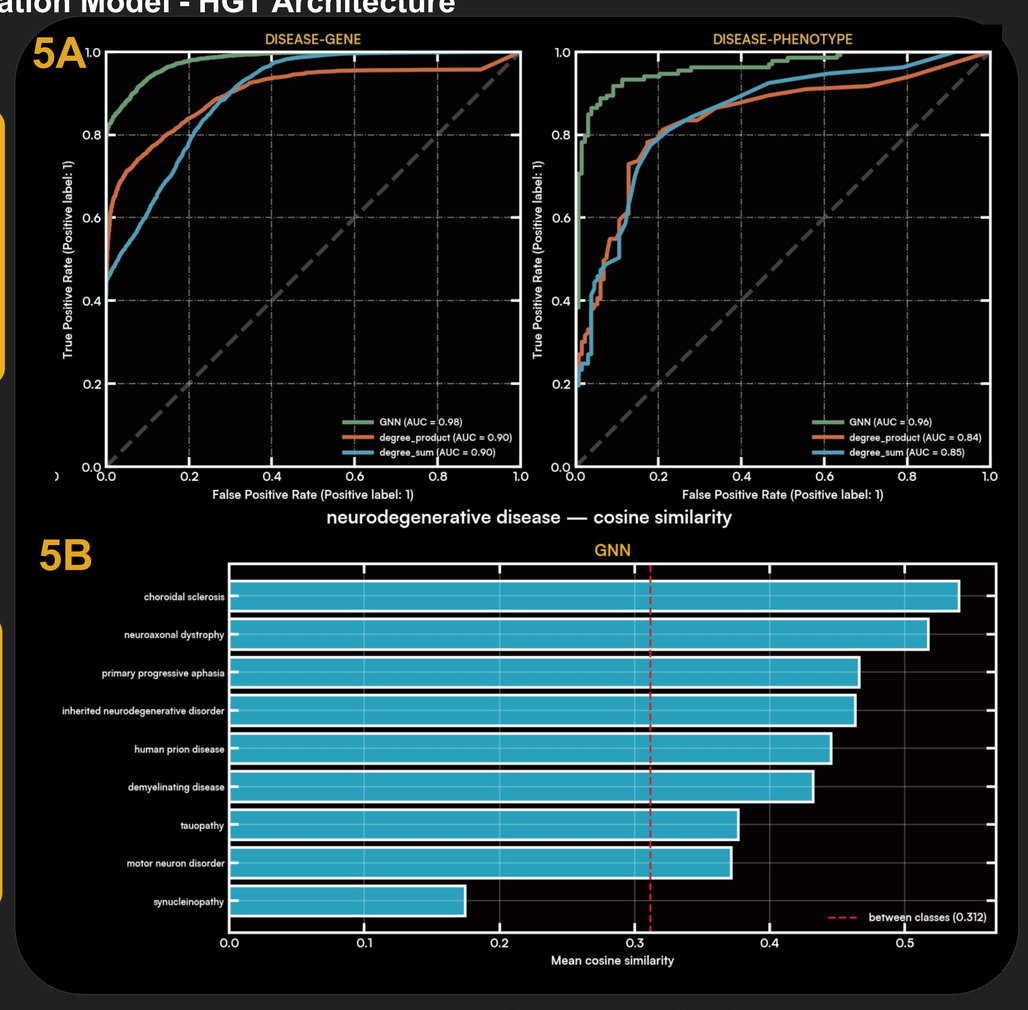


Figure 4. Summarized disease-linked pathway context through the defined metapath-based approach using data from Reactome. For each pathway, we aggregated the number of distinct candidate genes linked through the metapath (bubble size). Statistical overlap with the top 100 gene set from the multiXrank ranking was summarized per pathway using a hypergeometric test (P-value by color). Annotations of the well studied genes involved in PD are also included.

- Ranked known PD targets highly showing **PD-KG is suitable for identifying novel targets**
- Graph heterogeneity enables accurate **mechanistic contextualization via pathway enrichment analysis**

Graph Foundation Model - HGT Architecture

Edges can be predicted between pairs of nodes of a broad range of types.



Node embeddings encode **biological and ontological context.**

Figure 5. (A) A decoding edge prediction model was applied to the HGT-based embeddings for all edge types, and ROC-AUC curves are provided for disease-gene and disease-phenotype node pairs. (B) Cosine similarity of embeddings for disease nodes of each neurological disease class within (blue bars) and across (red dashed line) classes.

Conclusions and Future Directions

- Large, multi-omic PD-KG applicable to clinical and genomics focused analyses.
- Network-Based prioritization approaches perform well on target prediction tasks relevant to PD.
- GNN-based embeddings accurately predict both PD-relevant and other held-out relationships and demonstrates neurological-disease relevant context, will be applied to patient subtyping and target prediction tasks.

References

1. Wüllner, Ullrich et al. "The heterogeneity of Parkinson's disease." Journal of neural transmission (Vienna, Austria : 1996) vol. 130,6 (2023): 827-838.
2. Lange, L.M., Cerquera-Cleves, C., Schipper, M. et al. Prioritizing Parkinson's disease risk genes in genome-wide association loci. npj Parkinsons Dis. 11, 77 (2025).
3. Baptista, A., Gonzalez, A. & Baudot, A. Universal multilayer network exploration by random walk with restart. Commun Phys 5, 170 (2022).
4. Huang, K., Chandak, P., Wang, Q. et al. A foundation model for clinician-centered drug repurposing. Nat Med 30, 3601–3613 (2024).
5. Minikel, E.V., Painter, J.L., Dong, C.C. et al. Refining the impact of genetic evidence on clinical success. Nature 629, 624–629 (2024).
6. Huang, K., Chandak, P., Wang, Q. et al. A foundation model for clinician-centered drug repurposing. Nat Med 30, 3601–3613 (2024).