# AI-based Identification of Diagnostic Variants from Genotype and Phenotype

Azza Althagafi[1,2], Robert Hoehndorf[1]

[1]Computer, Electrical and Mathematical Sciences & Engineering Division (CEMSE), Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia. [2]Taif University, Faculty of Computing and Information Technology, Taif, Saudi Arabia

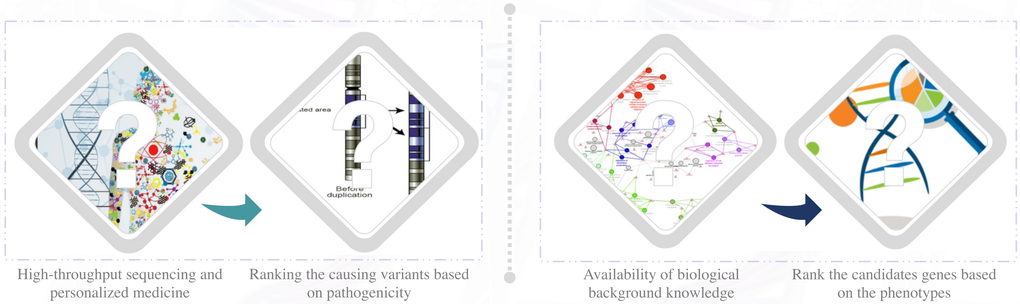
## BACKGROUND

- Understanding the molecular mechanisms underlying the set of abnormal phenotypes is important for the diagnosis, prevention, and development of therapies.
- We developed an AI model that uses background knowledge from human and model organisms to rank genes that are likely involved in a set of abnormal phenotypes.

## RESEARCH QUESTIONS

- Can we design a deep learning model that learns to link gene functions and anatomical site of expression to the phenotypes resulting from a loss of gene function?
- Can we combine this AI-based prediction of candidate genes with pathogenicity predictions of variants to improve diagnostic yield when applied to whole exome or whole genome sequencing data?

## MATERIALS AND METHODS



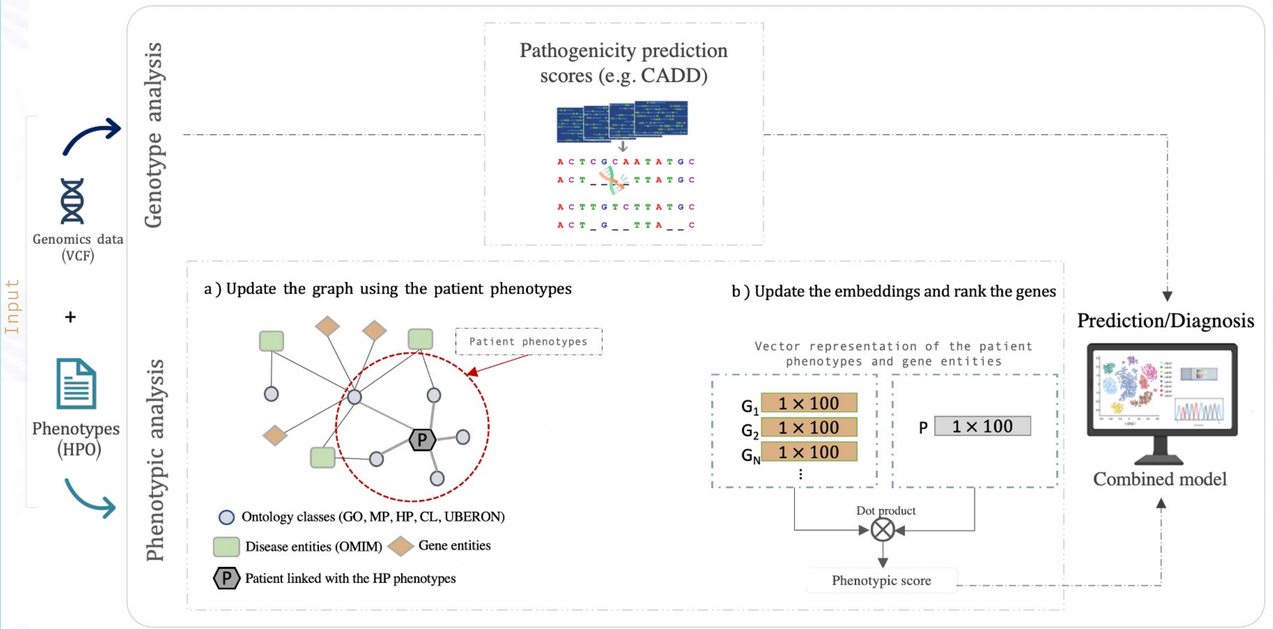




Figure 1: Model workflow

## RESULTS

We combine our deep learning model for ranking candidate genes with pathogenicity predictions of variants and benchmark our method on a set of 259 pathogenic variants from Saudi Arabia. We find that our method significantly improves over pathogenicity prediction methods as well as other phenotype-based approaches that use less background knowledge.

| | | Top hit | Top@10 | Top@30 | Top@50 |
|---|---|---|---|---|---|
| Genotype-based prediction tools | CADD | 80 (0.3089) | 202 (0.7799) | 213 (0.8224) | 216 (0.834) |
| | MCAP | 24 (0.0927) | 174 (0.6718) | 174 (0.6718) | 174 (0.6718) |
| | SIFT | 92 (0.3552) | 179 (0.6911) | 179 (0.6911) | 179 (0.6911) |
| | PolyPhen2 | 152 (0.5869) | 180 (0.695) | 180 (0.695) | 180 (0.695) |
| Phenotype-based prediction tools | Exomiser_hiPHIVE | 29 (0.112) | 78 (0.3012) | 132 (0.5097) | 155 (0.5985) |
| | Exomiser_Phenix | 31 (0.1197) | 43 (0.166) | 71 (0.2741) | 126 (0.4865) |
| | Exomiser_PHIVE | 15 (0.0579) | 30 (0.1158) | 40 (0.1544) | 44 (0.1699) |
| | DeepPVP | 8 (0.0309) | 19 (0.0734) | 25 (0.0965) | 30 (0.1158) |
| DL2vec models | DL2vec_GO | 104 ( 0.4015 ) | **218 ( 0.8417 )** | **219 ( 0.8456 )** | **219 ( 0.8456 )** |
| | DL2vec_MP | **162 ( 0.6255 )** | 172 ( 0.6641 ) | 177 ( 0.6834 ) | 177 ( 0.6834 ) |
| | DL2vec_HP | 195 ( 0.7529 ) | 215 ( 0.8301 ) | 216 ( 0.834 ) | 216 ( 0.834 ) |
| | DL2vec_CL | 85 ( 0.3282 ) | 105 ( 0.4054 ) | 105 ( 0.4054 ) | 106 ( 0.4093 ) |
| | DL2vec_UBERON | 118 ( 0.4556 ) | 206 ( 0.7954 ) | 210 ( 0.8108 ) | 210 ( 0.8108 ) |
| | DL2vec_intersection | 79 ( 0.305 ) | 93 ( 0.3591 ) | 93 ( 0.3591 ) | 93 ( 0.3591 ) |
| | DL2vec_union | 95 ( 0.3668 ) | 206 ( 0.7954 ) | 210 ( 0.8108 ) | 210 ( 0.8108 ) |

Table 1: Benchmark results on genomes matched with clinical information

## CONCLUSIONS

1) Deep learning methods over biomedical knowledge graphs improve variant interpretation.
2) We are able to relate gene functions and site of expression to phenotypes resulting from a loss of function.

**REFERENCES:**

- Hoehndorf, Robert, et al. "Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology." Bioinformatics 28.13 (2012): 1783-1789.
- Smedley, Damian, et al. "Next-generation diagnostics and disease-gene discovery with the Exomiser." Nature protocols 10.12 (2015): 2004-2015.
- PAVS - Phenotype associated Variants in Saudi Arabia, http://pavs.phenomebrowser.net

**Contact Info:** KAUST, Thuwal 23955, Saudi Arabia. azza.althagafi@kaust.edu.sa