

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Azza Althagafi

Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.

To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations

CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement

CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications

CAGI participants must acknowledge the original authors of these slides and CAGI 6. Include the acknowledgement banner across the bottom

CAGI participants must include the CAGI credits slide in their presentations

Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement

Slides with a red X over the entire slide may not be used in any circumstances

Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted

Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.



CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk

Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.

Permission request regarding CAGI 6 recordings and hybrid presentations

I understand that my presentation as associated with CAGI 6 will be recorded and broadcasted via the internet through CAGI's designated technology provider.

As the copyright holder to this presentation entitled *Prioritizing causative variants by integrating functional embedding and biological annotations for genes*, I hereby agree that the copyright for the online broadcast and recording of my presentation be released to CAGI and my recording be uploaded on to the CAGI website.

CAGI6 Conference

Clinical challenge: Rare Genomes Project

Prioritizing causative variants by integrating functional embedding and biological annotations for genes

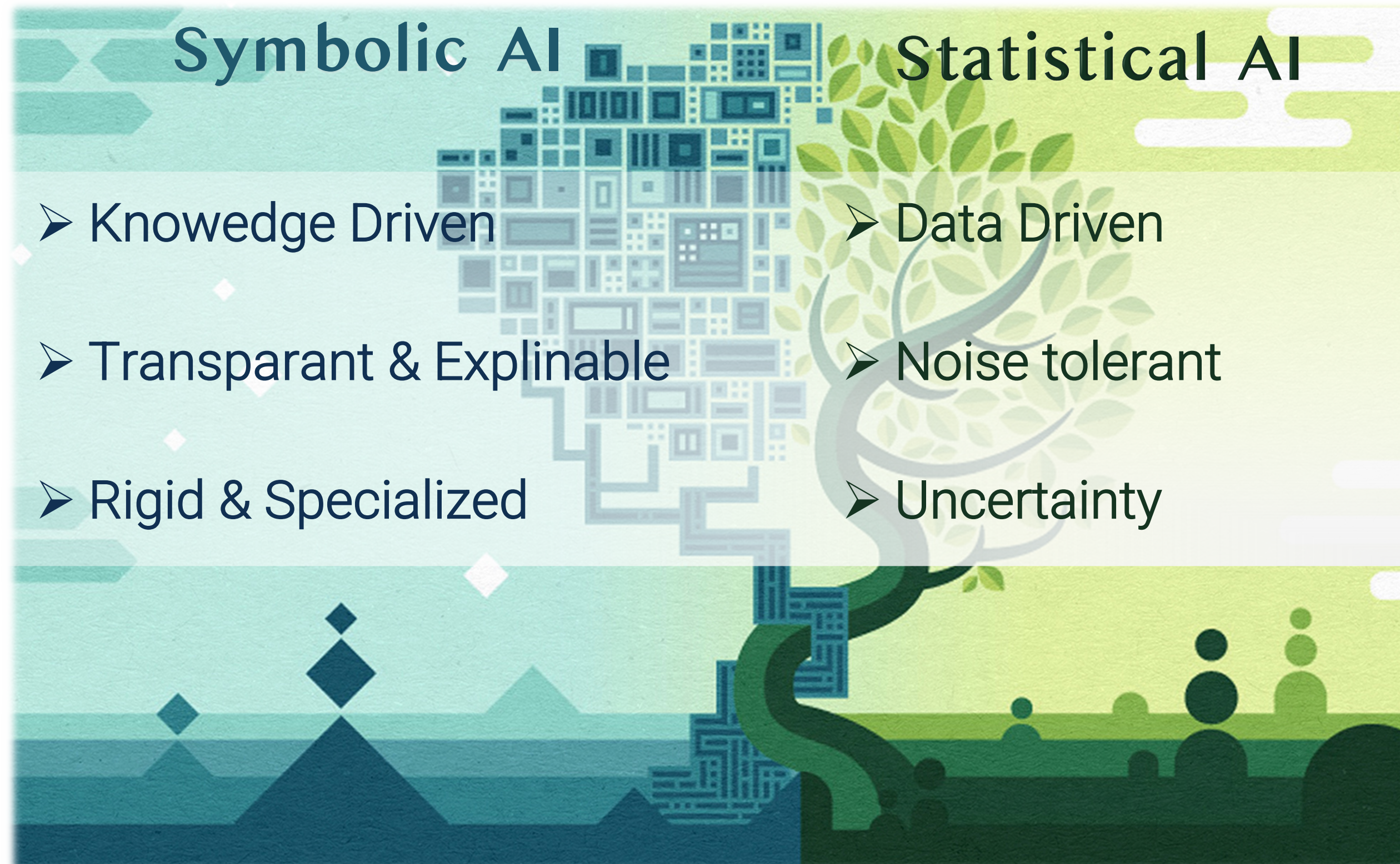
Azza Thamer Althagafi

PhD Candidate

Advisor: Prof. Robert Hoehndorf

Computational Bioscience Research Center,
Computer, Electrical & Mathematical Science and Engineering Division,
King Abdullah University of Science and Technology (KAUST),
Thuwal, Saudi Arabia

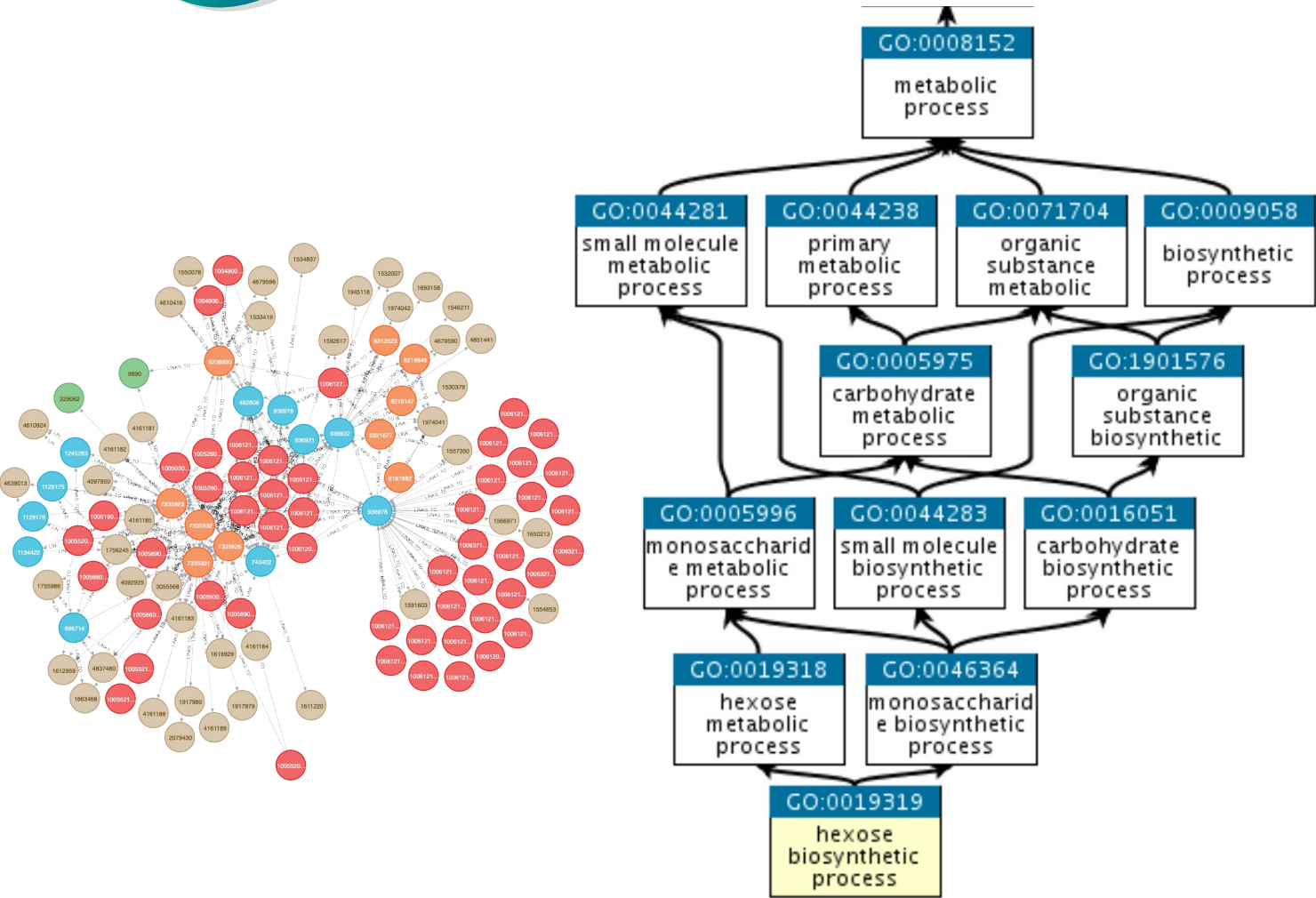
Background and Motivation



Background and Motivation

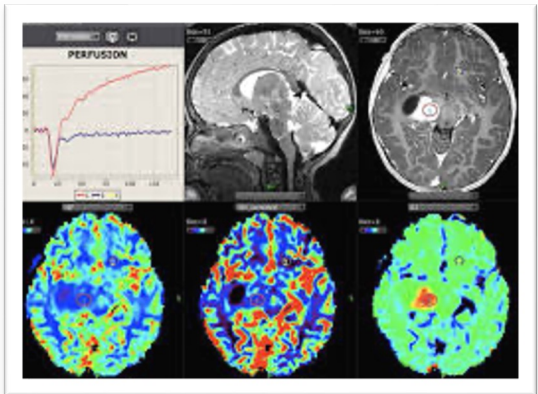


Knowledge base

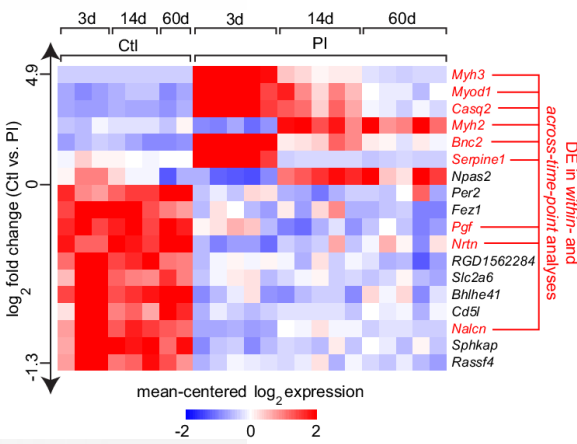


Data

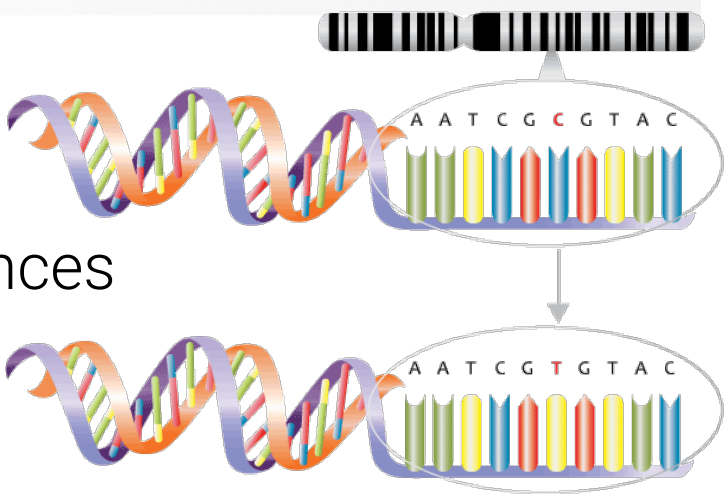
Images



Gene expressions



DNA Sequences



Problem Definition

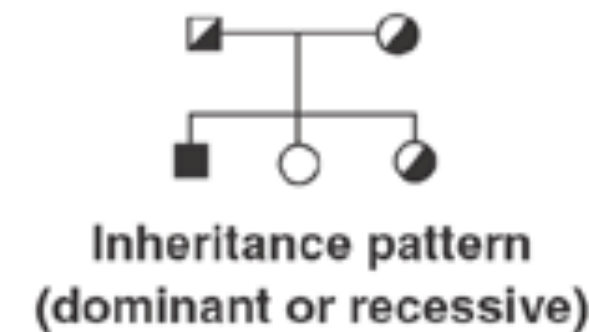
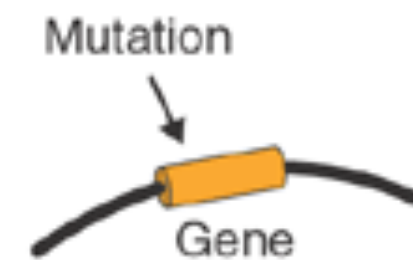


Patients Clinical Phenotypes & Genomics Data (WGS/WES)

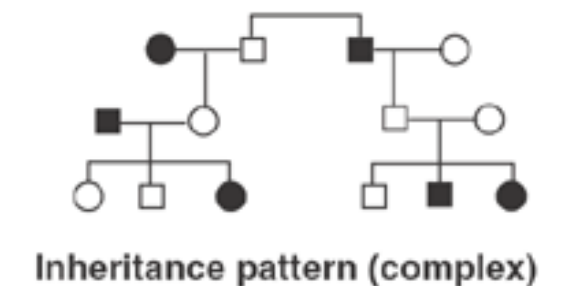
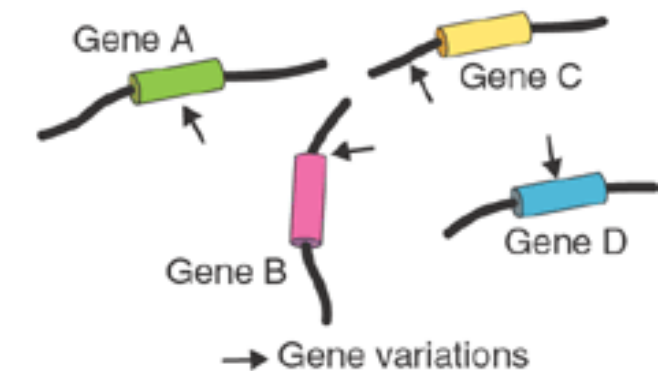


Can we find the causative structure variant associated with the phenotypes?

Monogenic Diseases



Complex Diseases



Related Research

■ Phenotype-based methods

- Using the function information for genes to predict the effect of variants in the phenotypes observed in the patients.
- Compare disease phenotypes to known genotype-phenotype associations.
- Suggest candidate genes/variants based on phenotype similarity measures.

Software | [Open Access](#) | Published: 06 February 2019

DeepPVP: phenotype-based prioritization of causative variants using deep learning

Next-generation diagnostics and disease-gene discovery with the Exomiser

Published: 20 July 2015

Phenolyzer: phenotype-based prioritization of candidate genes for human diseases

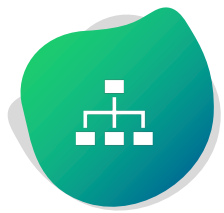
Hui Yang, Peter N Robinson & Kai Wang 

Nature Methods **12**, 841–843(2015) | [Cite this article](#)

1361 Accesses | **150** Citations | **31** Altmetric | [Metrics](#)

Related Research

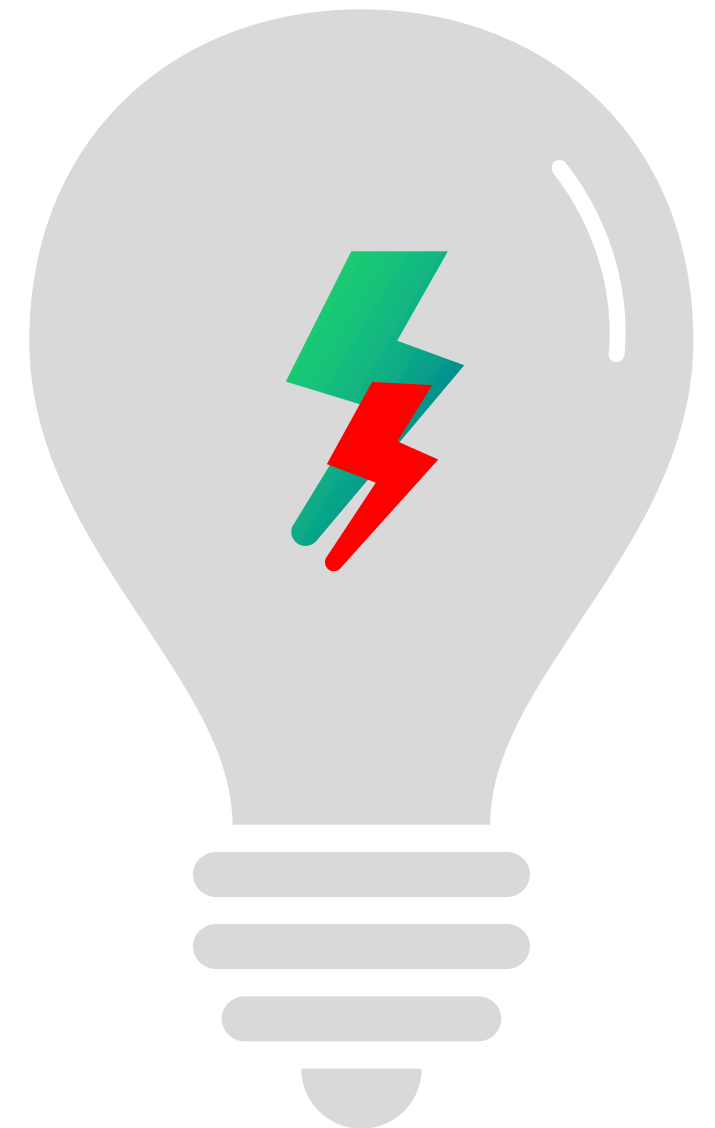
Limitations:



Incomplete knowledge of phenotypes

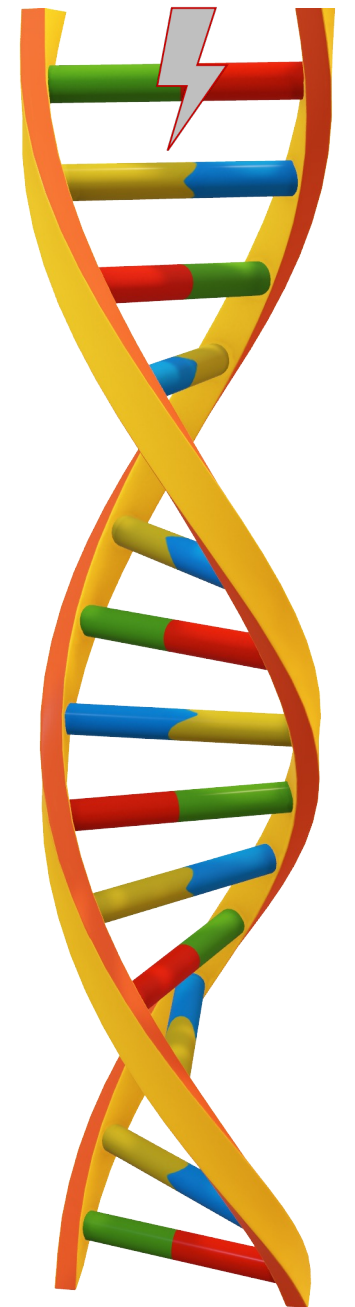


limited amount of information about phenotype associations available for genotypes in model organism



Problem Definition

- Prioritizing the causative variants.
 - Learns to link gene functions and anatomical site of expression to the phenotypes resulting from a loss of gene function
 - Determining which variants identified using Whole-exome Sequencing (WES) or Whole-genome sequencing (WGS) are most likely to **damage** gene function and underlie the **disease phenotype**.



Our Solutions



Knowledge Graph

Leverages a large amount of background knowledge from human and animal models.



Deep Learning Models

Explore the knowledge graph using computational methods to generate representations.



EmbedPVP: Different types of genomic variants

Using several genomic features to train the models and accounts for the contribution of multiple genes, SNP, indels, intergenic.

Model Overview: Phenotype-based model

DL2vec*: Embedding graph-based representations of ontologies

- A **data-driven** approach to produce **vector embeddings** which encode:
 - The logical axioms encoded in an ontology.
 - The known ontology-based annotations of biological entity:
 - The information for genes based on their associated phenotypes, functions of the gene products, and anatomical location of gene expression.

in one single embedding.

Embedding

An embedding is a mapping from one mathematical structure X into another structure Y :

$$f : X \rightarrow Y$$

Such that X is preserved in Y .

* Jun Chen, Azza Althagafi, and Robert Hoehndorf. "Predicting candidate genes from phenotypes, functions, and anatomical site of expression." (2020). DOI: 10.1093/bioinformatics/btaa879

■ Using functional and phenotypic characteristics for genes in:

Human phenome	Mouse phenome	Functions of the gene products	Gene expression in individual cell types	Anatomical site of expression from the GTEx tissue expression
Human Phenotype Ontology (HPO)	Mammalian Phenotype Ontology (MP)	Gene Ontology (GO)	Celltype Ontology (CL)	Uber-anatomy ontology (UBERON)
4,315 genes & 169,281 associations	13,529 genes & 168,550 associations	17,786 genes & 208,630 associations	6,559 genes & 17,149 associations	20,538 genes & 585,765 associations

■ All these ontologies can be applied across different species using integrated PhenomeNET ontology*

* Hoehndorf,R. et al. PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Res., 39, e119.

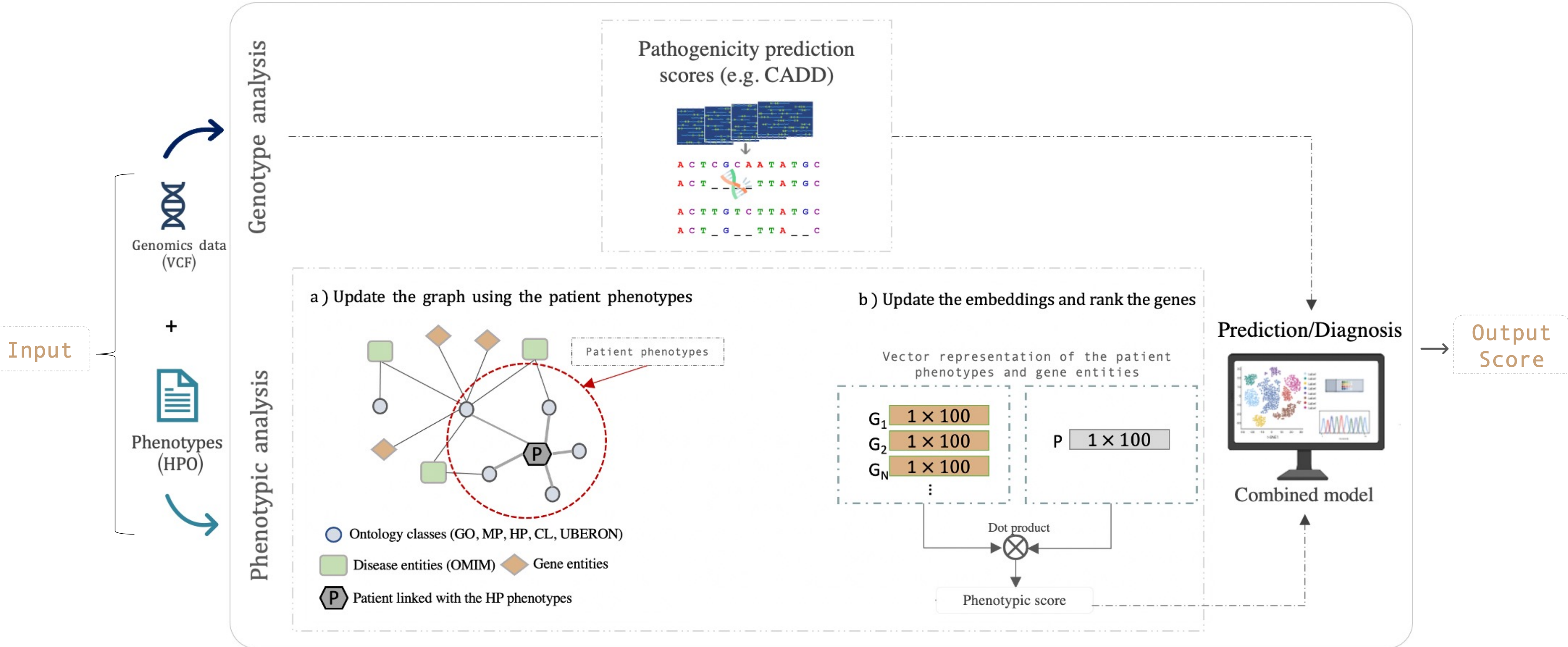
Fig1: DL2Vec workflow

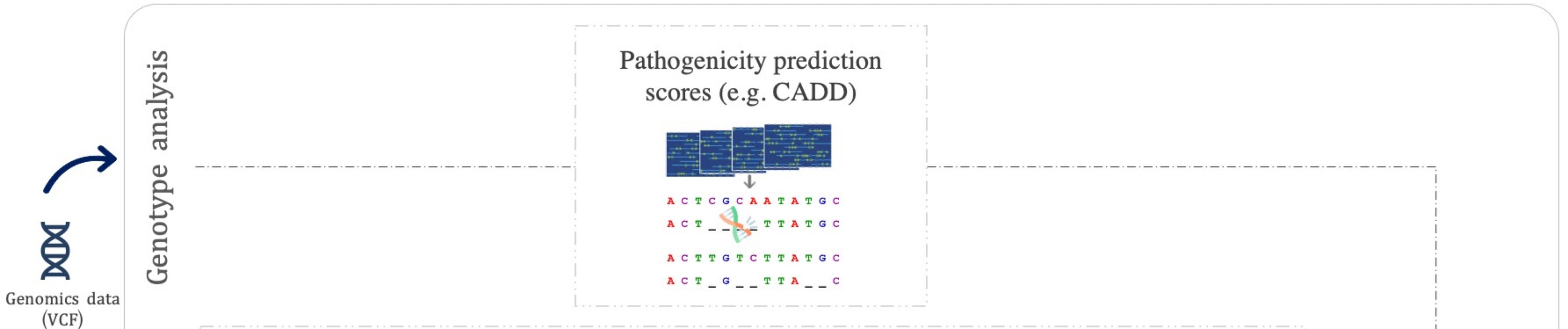
Models Overview

- Genomics Features

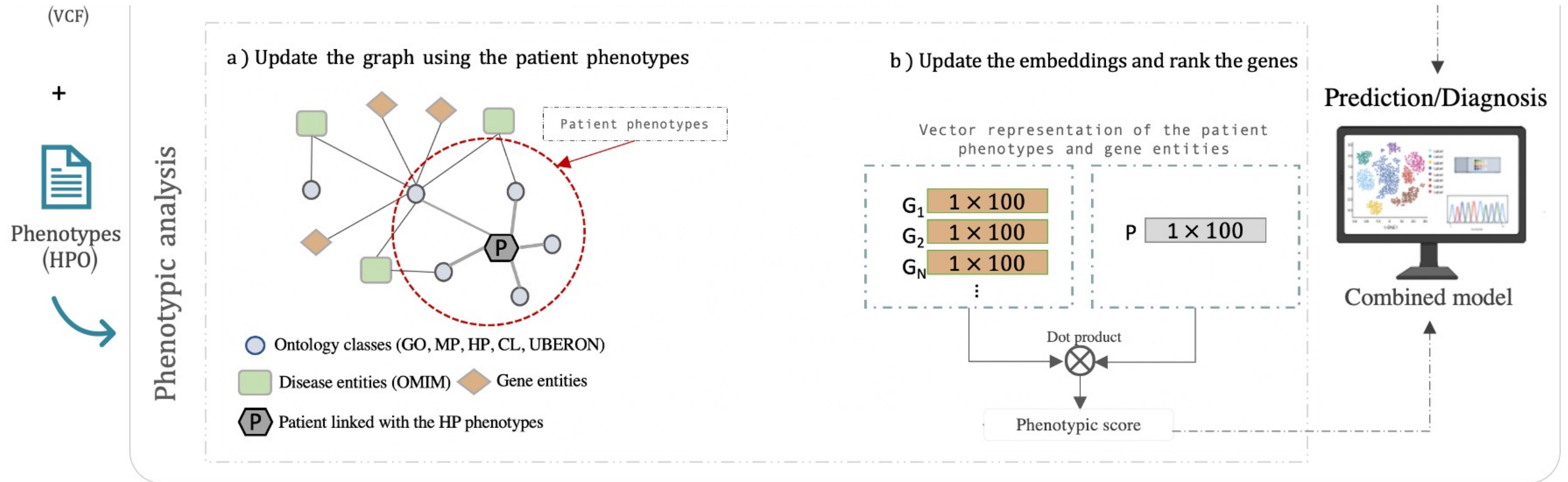
	Prioritizing Single-nucleotide variants (SNP), indels
Training datasets	Pathogenic and benign variants from the Clinically relevant variation (ClinVar)
Gene-based features	DL2Vec: Graph-based machine learning method for biomedical ontologies.
Variant-based features	Pathogenicity scores, e.g. CADD, Sift, polyphen, distance to the left and right genes in terms of intergenic variants.
Merging gene-based features into variant features	Using the gene related feature for the exonic variants, and the maximum score for the intergenic variants. $V_s = Max (G_1 \dots G_N)$

AI-based Identification of Diagnostic Variants from Genotype and Phenotype

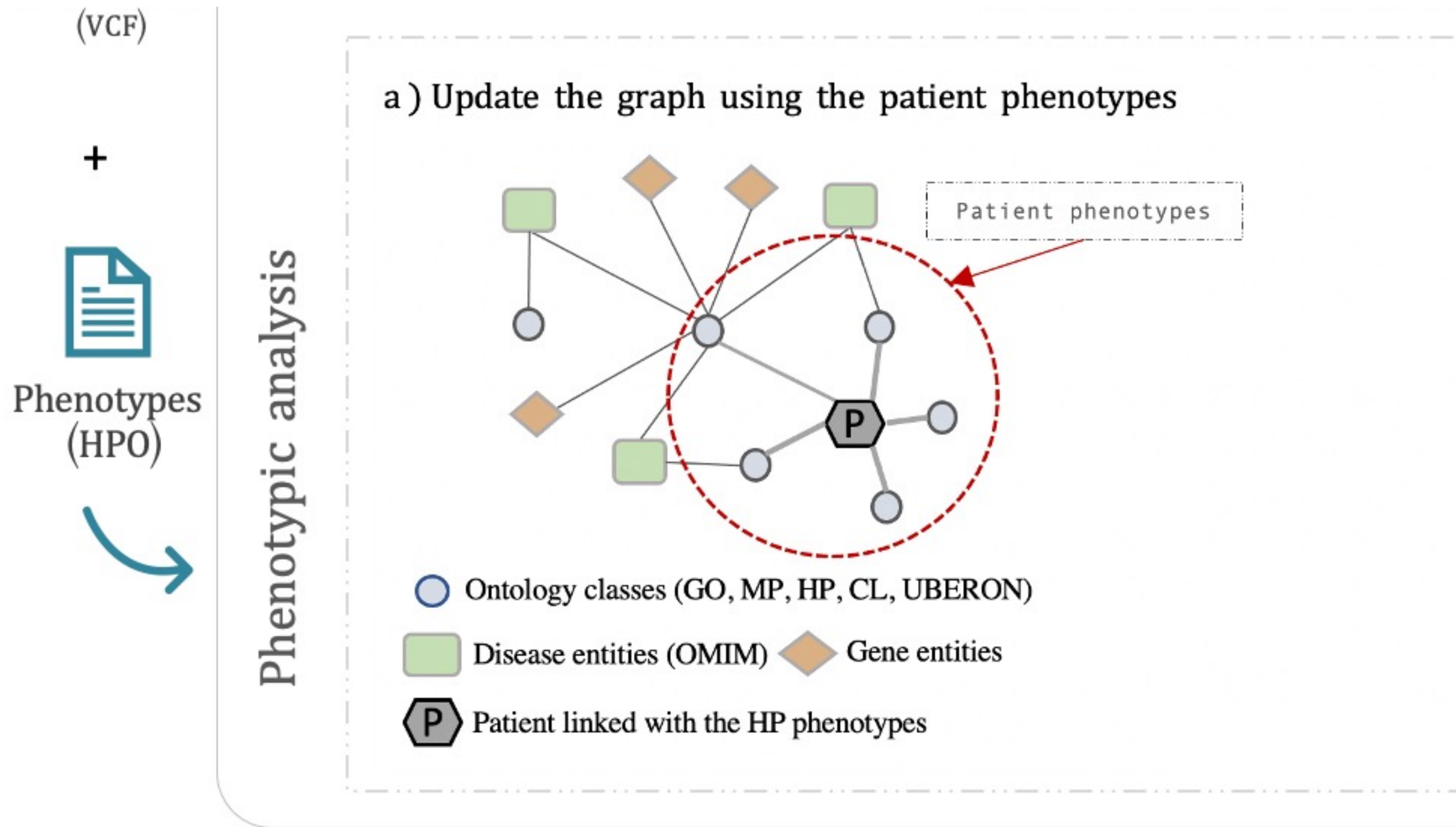




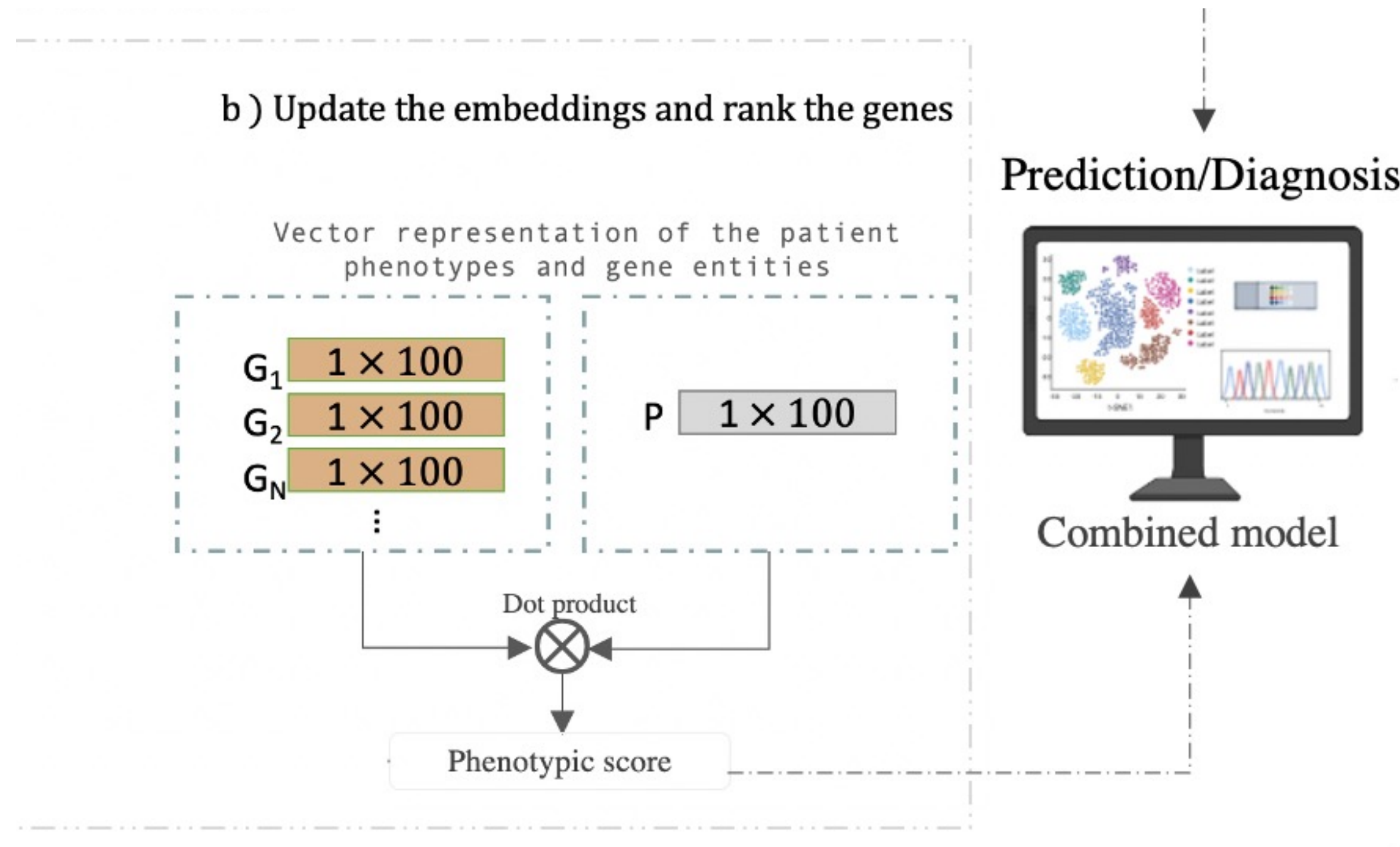
AI-based Identification of Diagnostic Variants from Genotype and Phenotype



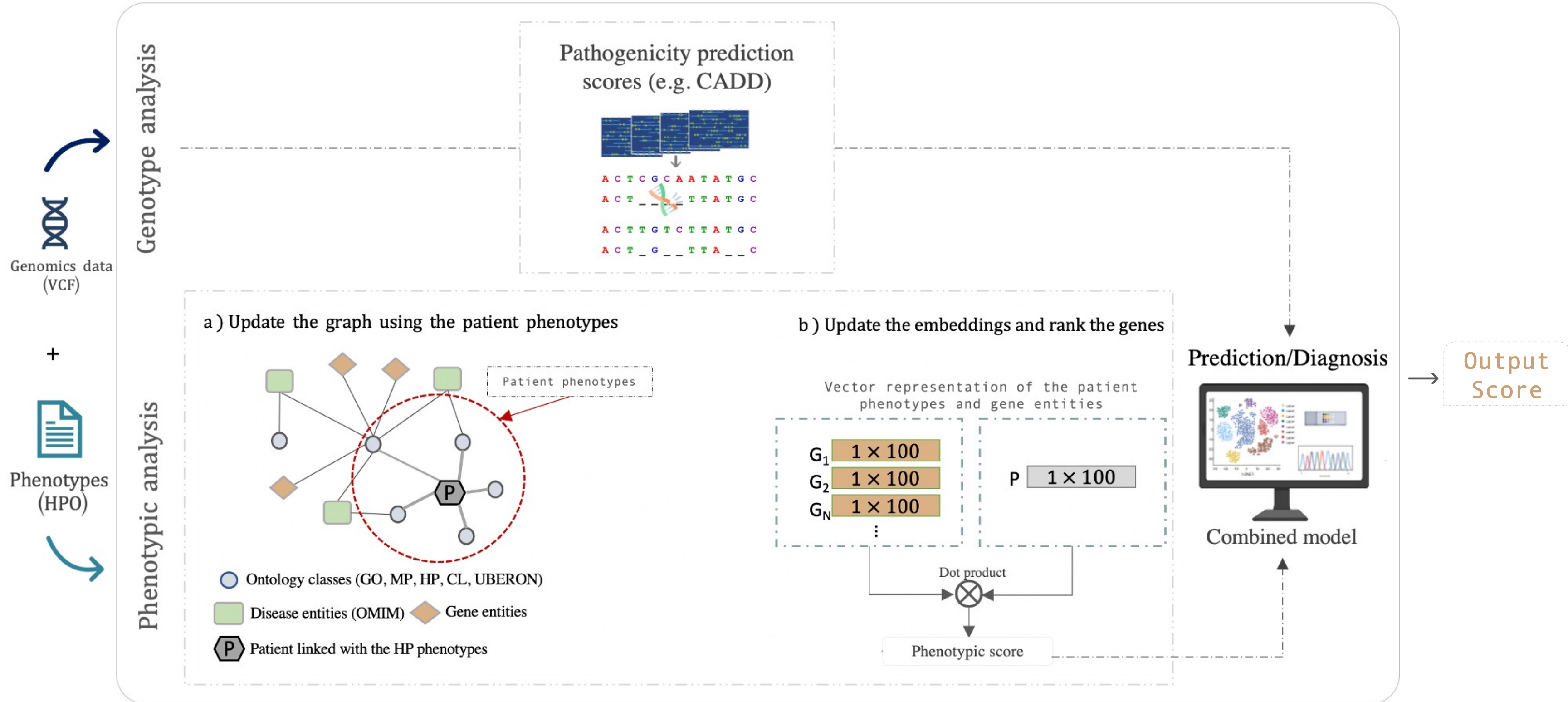
AI-based Identification of Diagnostic Variants from Genotype and Phenotype



AI-based Identification of Diagnostic Variants from Genotype and Phenotype

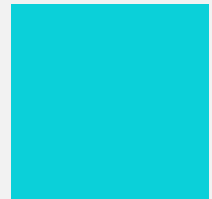


AI-based Identification of Diagnostic Variants from Genotype and Phenotype



Results & Evaluations

We use the popular evaluation measures to evaluate the models



Benchmark Synthetic/real dataset

Evaluation Metrics

- **Synthetic datasets**

- Inserting a single causative variants into a randomly selected WGS from the 1000 Genomes Project with 1% MAF.
- We link the Clinical phenotypes that are associated with the causative variant into the synthetic genome.
- Run the models and find the rank of the inserted variants.

Results & Evaluations

Model : EmbedPVP

Benchmark Synthetic dataset, with clinical phenotypes

		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

RGP Challenge : Case Example

Genotype analysis:

- MAF filter
- Family-based filtering (Slivar is available from <https://github.com/brentp/slivar>)
- Pathogenicity prediction:
 - The top 1 : 25.45%
 - The top 10 : 63.64%
 - The top 50 : 92.73%

Conclusion

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

Databases and ontologies

Predicting candidate genes from phenotypes, functions and anatomical site of expression

Jun Chen^{1,†}, Azza Althagafi^{1,2,†} and Robert Hoehndorf ^{1,*}

¹Computational Bioscience Research Center (CBRC), Computer, Electrical & Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 2455, Saudi Arabia
²Department, College of Computers and Information Technology, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia

*To whom correspondence should be addressed

[†]The authors wish it to be known that, in their work, they have used the following reagents:

Bioinformatics

Issues

Advance articles

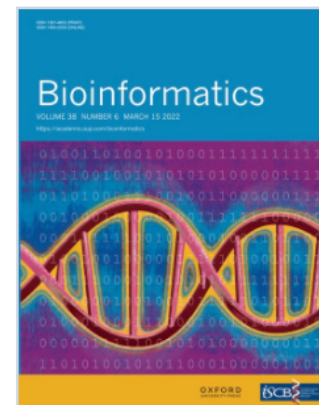
Submit ▼

Purchase

Alerts

About ▼

All Bioinformatics



Volume 38, Issue 6
15 March 2022

Article Contents

Abstract

DeepSVP: integration of genotype and phenotype for structural variant prioritization using deep learning

Azza Althagafi, Lamia Alsubaie, Nagarajan Kathiresan, Katsuhiko Mineta, Taghrid Aloraini, Fuad Al Mutairi, Majid Alfadhel, Takashi Gojobori, Ahmad Alfares, Robert Hoehndorf 

Bioinformatics, Volume 38, Issue 6, 15 March 2022, Pages 1677–1684,

<https://doi.org/10.1093/bioinformatics/btab859>

Published: 24 December 2021 **Article history** ▼



PDF

Split View

Cite

Permissions

Share ▼

Acknowledgements



Robert Hoehndorf

Principal Investigator

Bio-Ontology Research Group (BORG),
CBRC, CEMSE. Thuwal, Saudi Arabia



Marwah Abdulhakeem

Scientific Curator

Bio-Ontology Research Group (BORG),
CBRC, Thuwal , SA



Ahmad Alfares

Paediatric Consultant

Department of Pathology and Laboratory
Medicine, King Abdulaziz Medical City, Ministry
of National Guard-Health Affairs (MNG-HA),
Riyadh, SA

Acknowledgements

- *Special Thanks to:*
 - IBEX Supercomputing team at KAUST
 - Bio-Ontology Research Group (BORG)
 - CAGI6 Conference organizers.



Contact Info

Bio-Ontology Research Group (BORG), Computational Bioscience
Research Center (CBRC), PI: Professor Robert Hoehndorf



azza.althagafi@kaust.edu.sa



Office

KAUST, Building 3, Level 4,
(3315-WS30 / Conference room 4330)



Phone +966 533422088



Thank You!