ESWC 2024

INTEGRATING HETEROGENEOUS GENE EXPRESSION DATA THROUGH KNOWLEDGE GRAPHS FOR IMPROVING DIABETES PREDICTION

Rita T. Sousa, Heiko Paulheim

Data and Web Science Group, Universität Mannheim, Germany

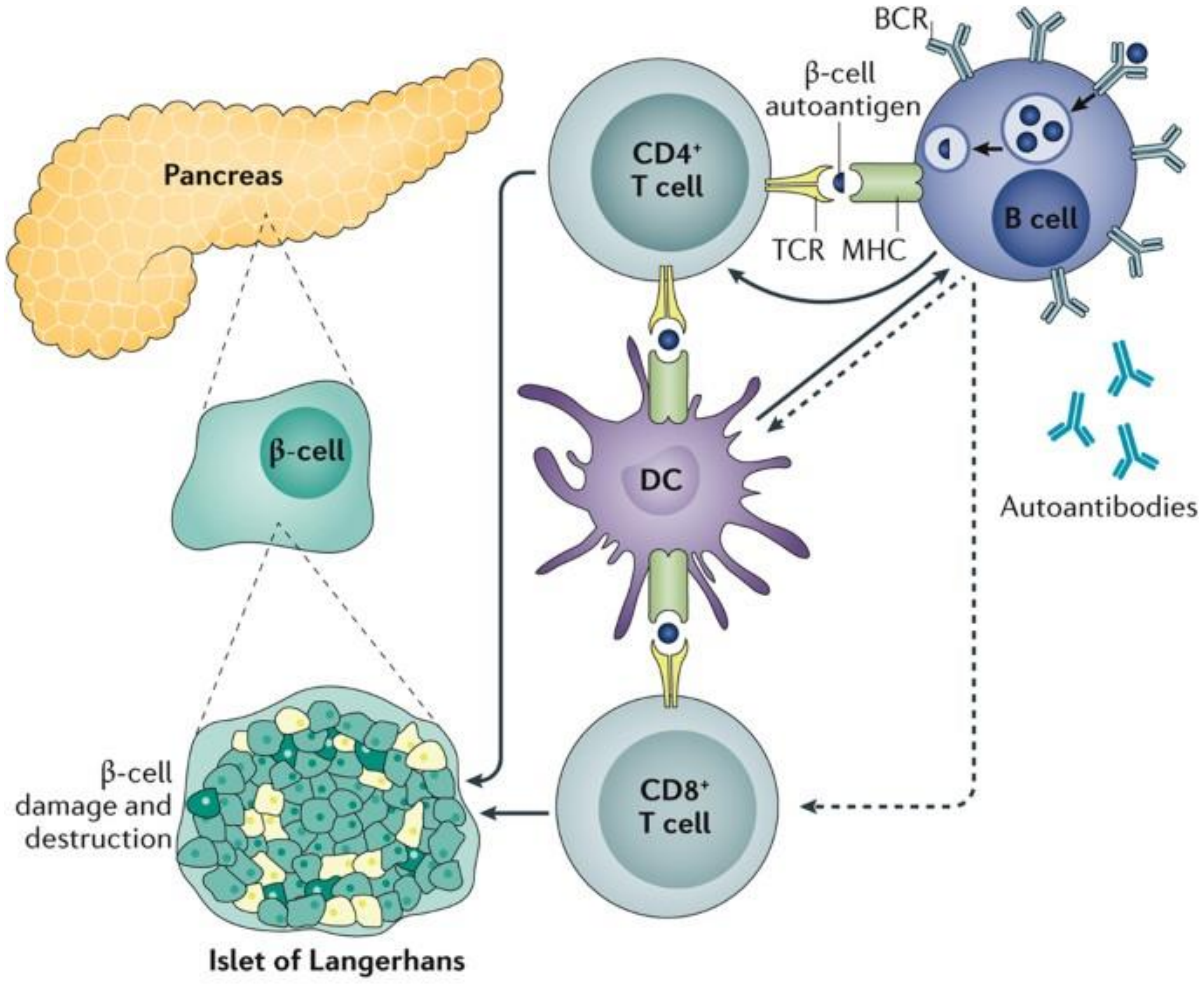
7th Workshop on SeWeBMeDA
26th of May 2023

UNIVERSITY OF MANNHEIM

DIABETES

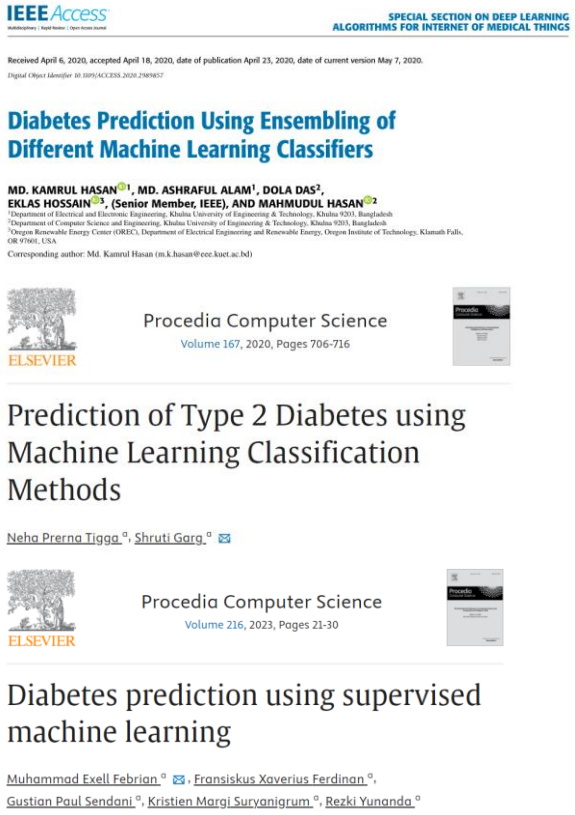
- In 2019, diabetes was the direct cause of **1.5 million deaths**.
- Diabetes is a major cause of several **comorbidities**: blindness, kidney failure, heart attacks, stroke and lower limb amputation.
- WHO launched the Global Diabetes Compact aiming for sustained improvements in **diabetes prevention**.

World Health Organization



Katsarou, A., Gudbjörnsdottir, S., Rawshani, A. et al. Type 1 diabetes mellitus. Nature Reviews Disease Primers 3, 17016 (2017). https://doi.org/10.1038/nrdp.2017.16

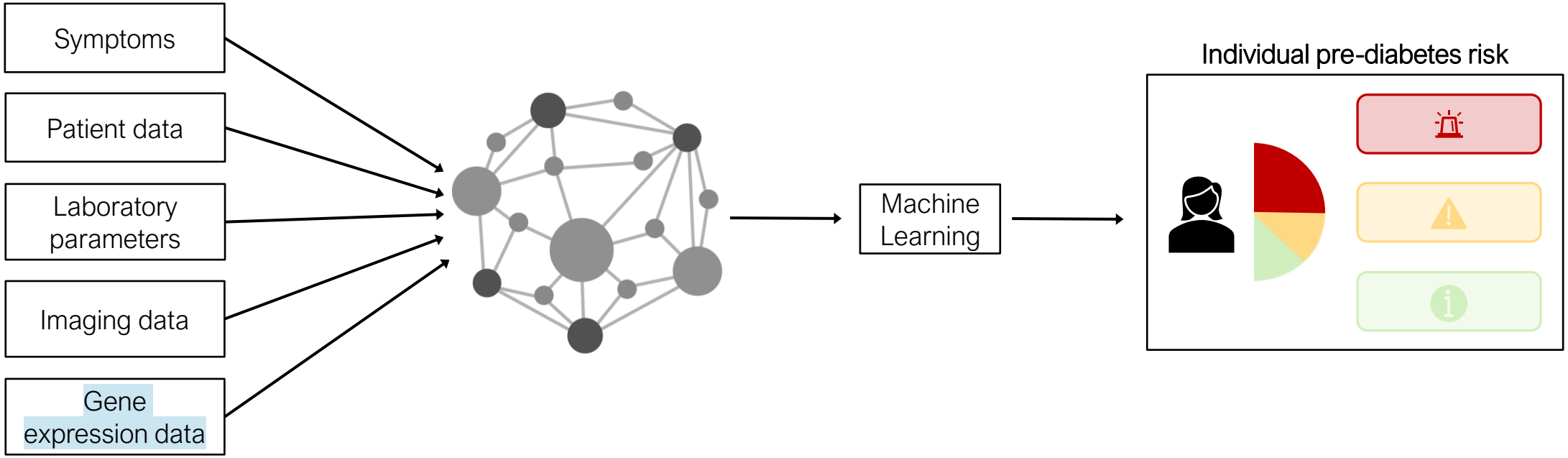
DIABETES PREDICTION USING MACHINE LEARNING



- Due to the **multidisciplinary nature** of diabetes, predicting and detecting this disease continues to pose a significant challenge.
- **Machine learning methods** have shown promise in identifying diabetes patterns and risk factors, enabling early detection and personalized interventions.

KI-DIABETES DETECTION PROJECT

The goal is to integrate data from various sources and apply machine learning methods to improve the early-stage detection of Diabetes.



GENE EXPRESSION DATA

- Gene expression values are numerical representations indicating the expression levels of genes under specific conditions.

- The expression values are organized in a matrix $m \times n$, where m is the number of samples, n is the number of genes, and $m << n$.

	G1	G2	G3	G4	G5	G6	…	Gn
P1	$GE_{P1,G1}$	$GE_{P1,G2}$	$GE_{P1,G3}$	$GE_{P1,G4}$	$GE_{P1,G5}$	$GE_{P1,G6}$	…	$GE_{P1,Gn}$
P2	$GE_{P2,G1}$	$GE_{P2,G2}$	$GE_{P2,G3}$	$GE_{P2,G4}$	$GE_{P2,G5}$	$GE_{P2,G6}$	…	$GE_{P2,Gn}$
…	…	…	…	…	…	…	…	…
Pm	$GE_{Pm,G1}$	$GE_{Pm,G2}$	$GE_{Pm,G3}$	$GE_{Pm,G4}$	$GE_{Pm,G5}$	$GE_{Pm,G6}$	…	$GE_{Pm,Gn}$

GENE EXPRESSION INTEGRATION CHALLENGE

Gene expression datasets typically only have few instances, and different datasets record different gene expressions.

	G1	G2	…
P1	0.1	0.9	…
P2	0.8	0.7	…
…	…	…	…
Avg	0.4	0.6	…

	G3	G4	…
P3	0.3	0.4	…
P4	0.5	0.8	…
…	…	…	…
Avg	0.4	0.3	…

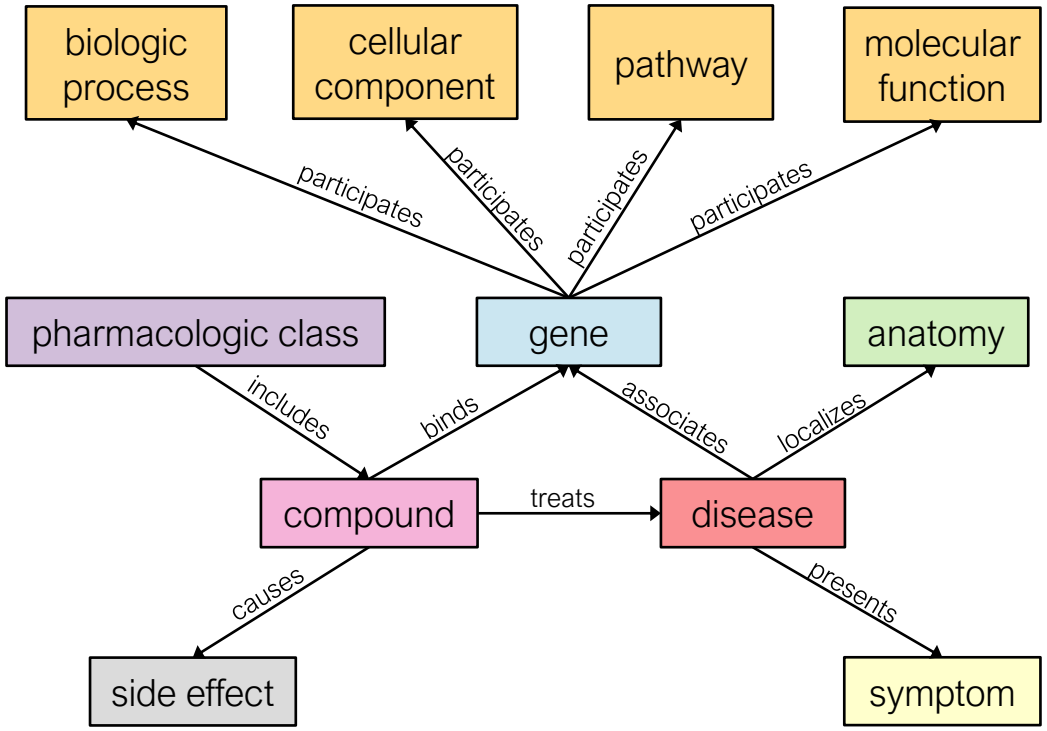
Solutions

Use only one dataset, thereby having only little training data.

Try to combine multiple datasets that are typically “incompatible“.

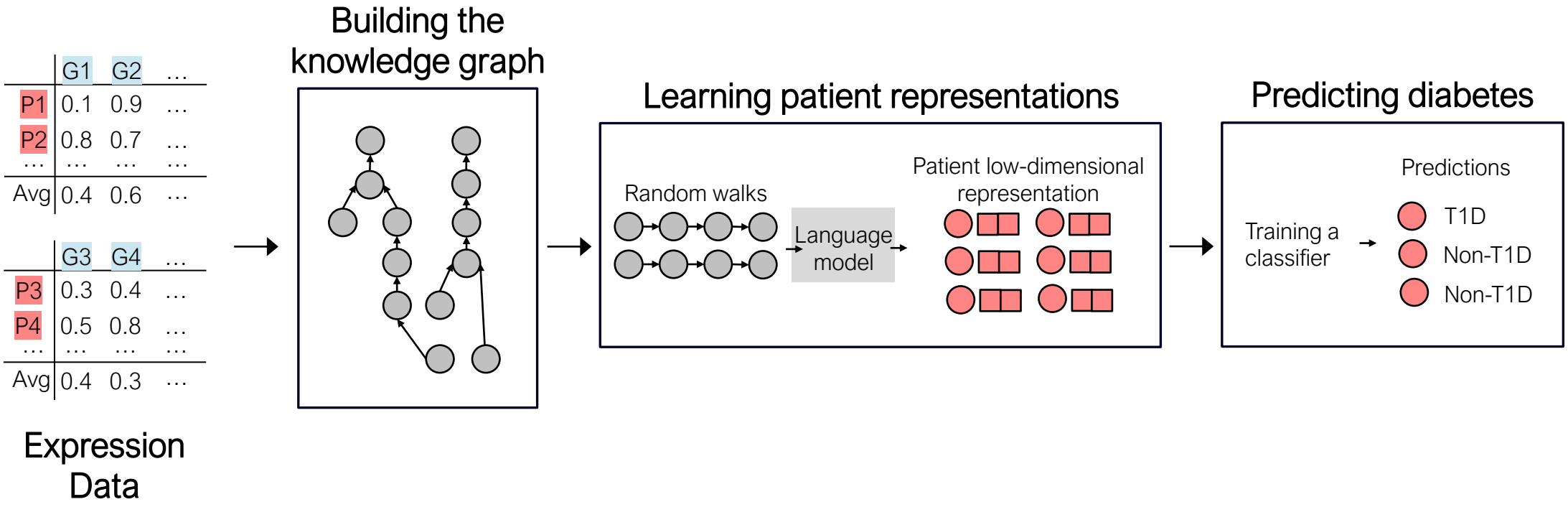
KNOWLEDGE GRAPHS AND DATA INTEGRATION

- 900+ biomedical ontologies covering many domains and fitting different applications.
- Knowledge graphs (KGs) can be explored for many biomedical applications such as finding new treatments for existing drugs, diagnosing patients, identifying associations between diseases and genes, etc.



METHODOLOGY

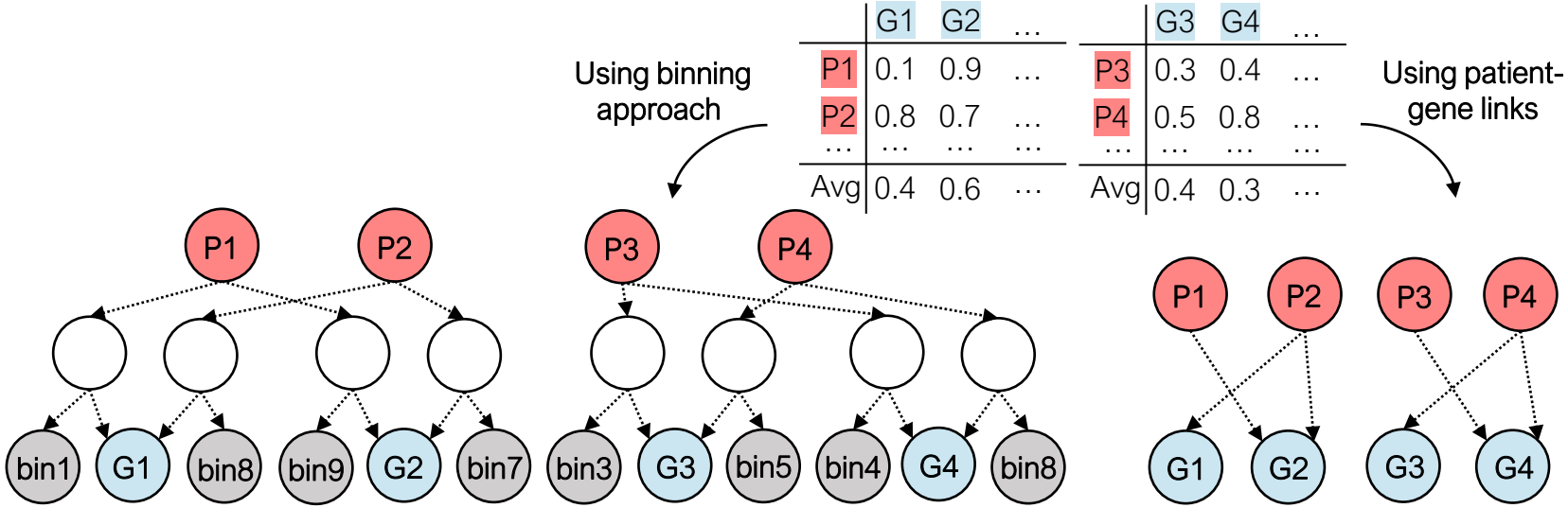
The goal is to integrate multiple expression datasets into a biomedical KG and then use it for diabetes prediction.



METHODOLOGY
STEP I: BUILDING THE KNOWLEDGE GRAPH

The KG is built by integrating:

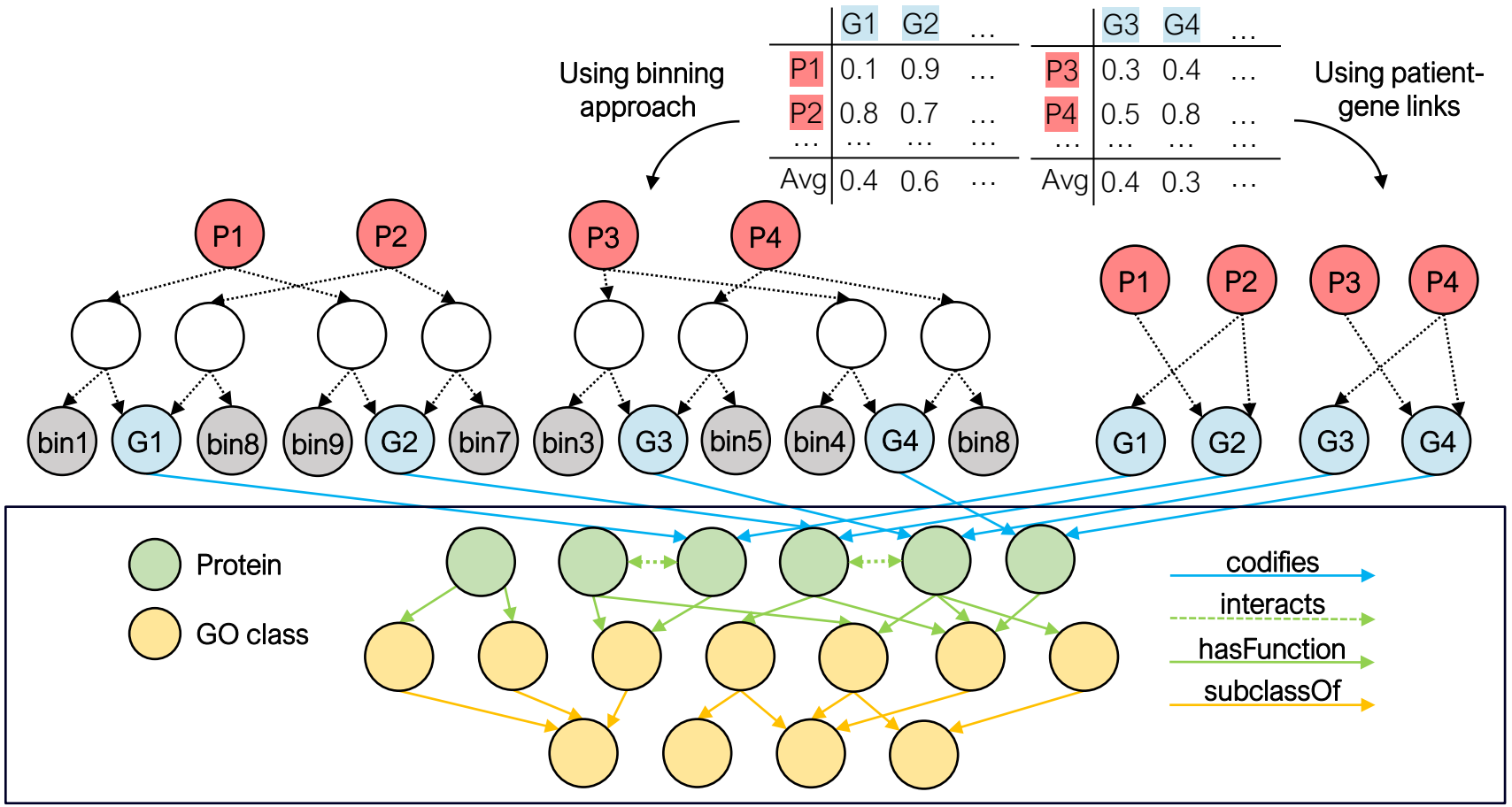
- **Gene expression data** using two strategies: representing patient gene expression values in a KG using blank nodes and binning approaches; linking patients and genes based on expression values.



METHODOLOGY
STEP I: BUILDING THE KNOWLEDGE GRAPH

The KG is built by integrating:

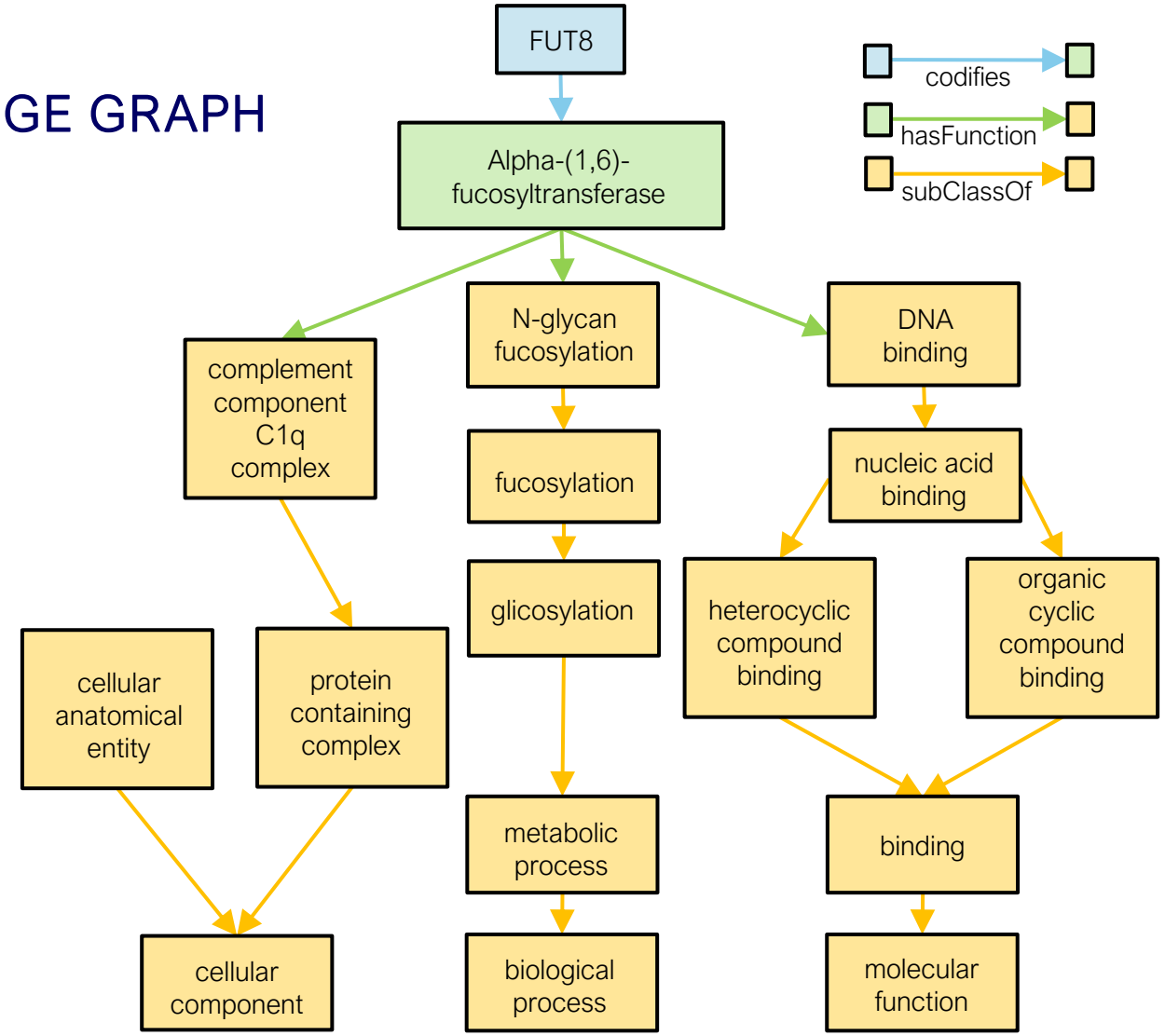
- **Gene expression data** using two strategies: representing patient gene expression values in a KG using blank nodes and binning approaches; linking patients and genes based on expression values.
- **Domain-specific knowledge** including Gene Ontology (GO) data, and protein interactions.



METHODOLOGY
STEP I: BUILDING THE KNOWLEDGE GRAPH

The KG is built by integrating:

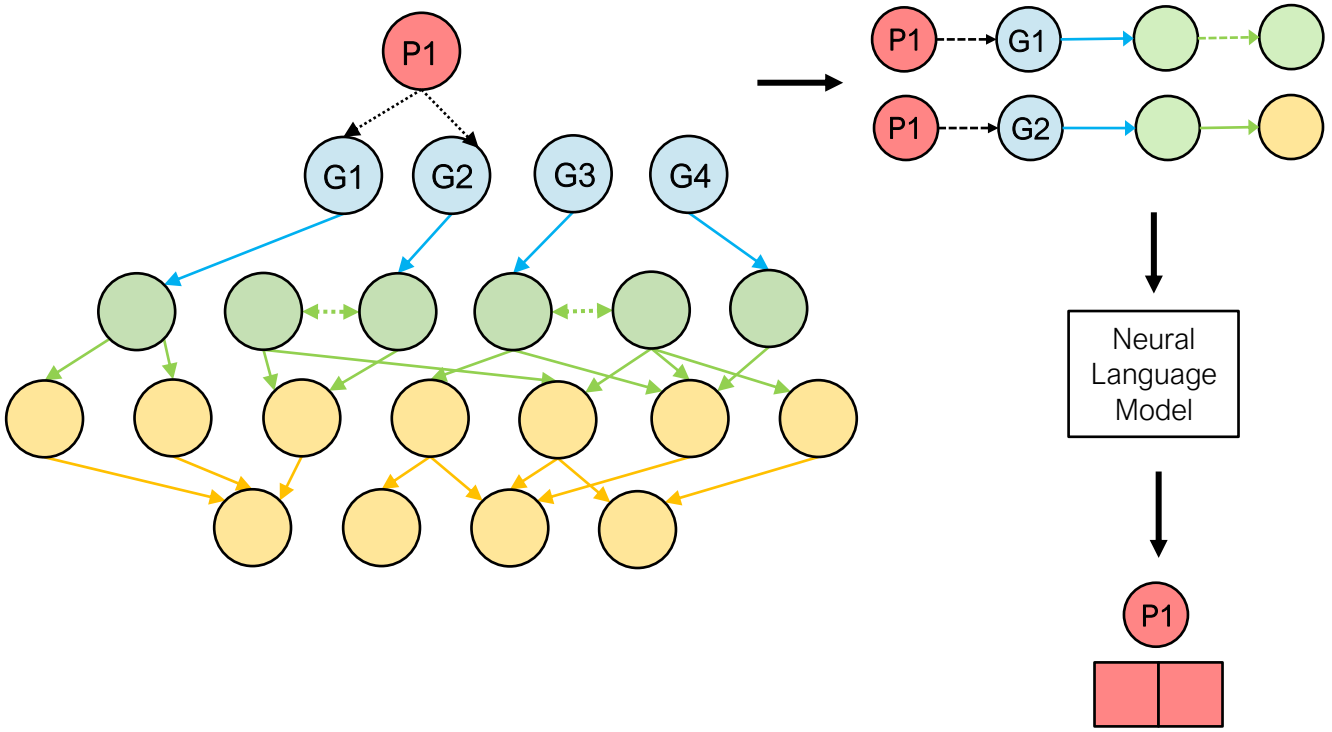
- **Gene expression data** using two strategies: representing patient gene expression values in a KG using blank nodes and binning approaches; linking patients and genes based on expression values.
- **Domain-specific knowledge** including Gene Ontology (GO) data, and protein interactions.



METHODOLOGY
STEP II: LEARNING PATIENT REPRESENTATIONS

Two distinct approaches are employed to represent patients:

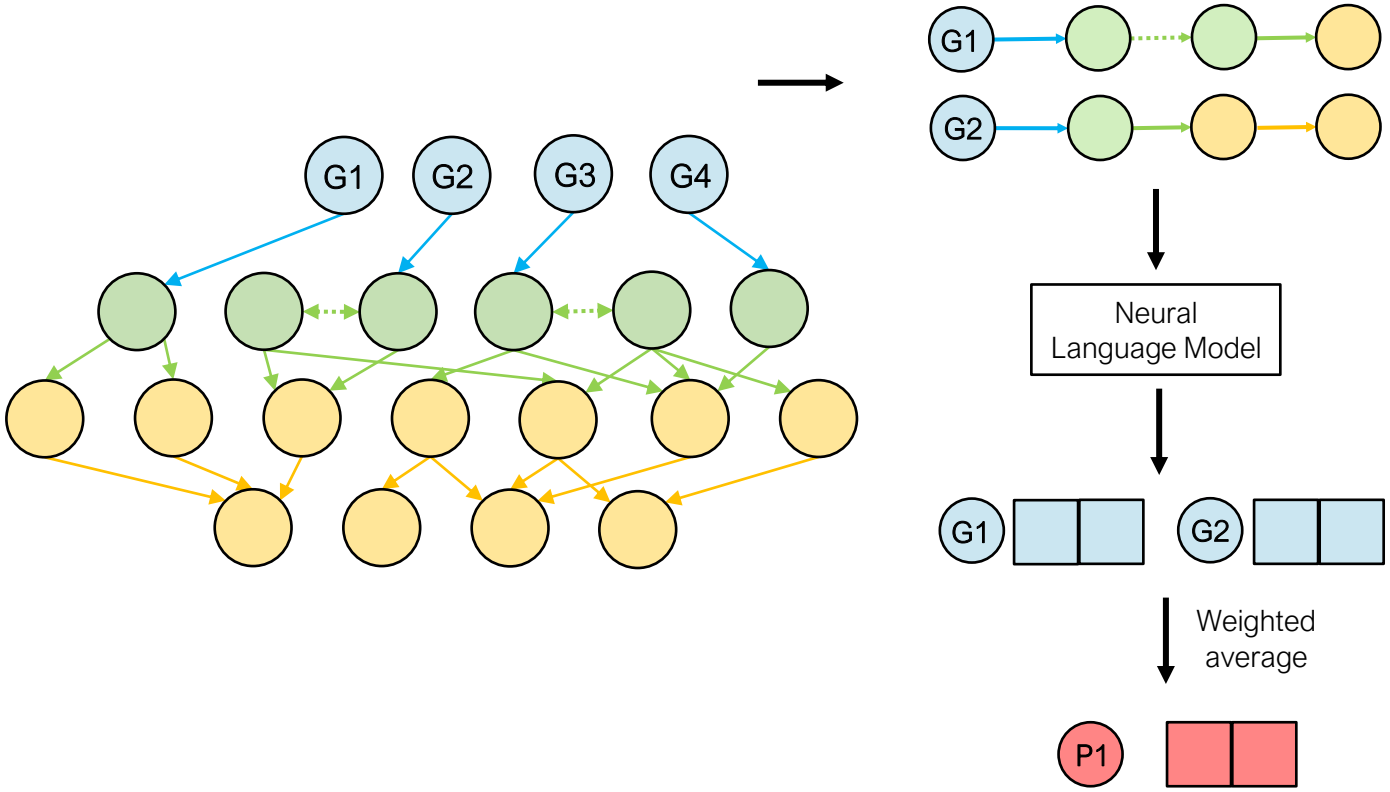
- Generating RDF2vec **embeddings directly for the patients** using the KG.



METHODOLOGY
STEP II: LEARNING PATIENT REPRESENTATIONS

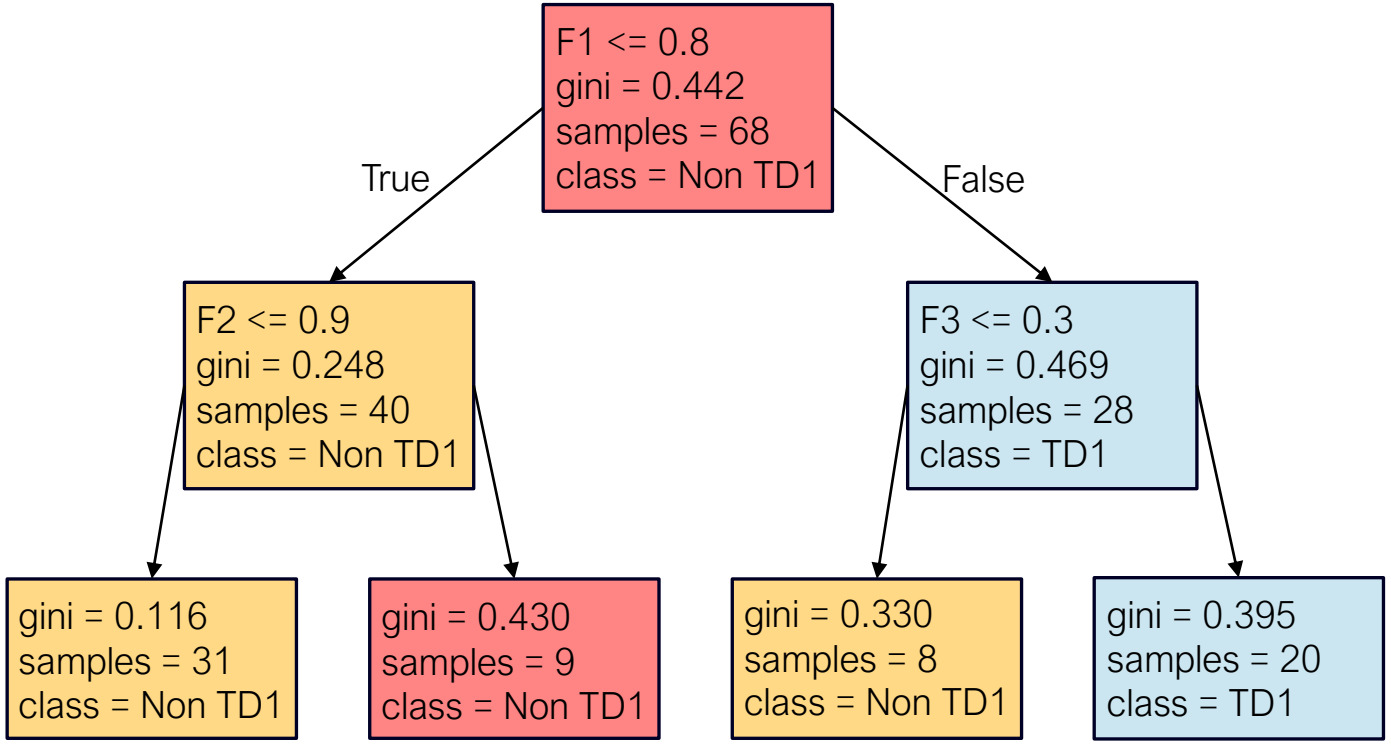
Two distinct approaches are employed to represent patients:

- Generating RDF2vec **embeddings directly for the patients** using the KG.
- Generating RDF2Vec gene embeddings and represents patients as the **weighted average of gene embeddings**, determined by the respective gene expression values.



METHODOLOGY
STEP III: PREDICTING DIABETES

- Diabetes prediction is formulated as a binary classification task.
- The patient representations are fed into a decision tree for training.



DATA

Three diabetes-related GEO datasets (GSE15932, GSE30208, and GSE55098) are considered.

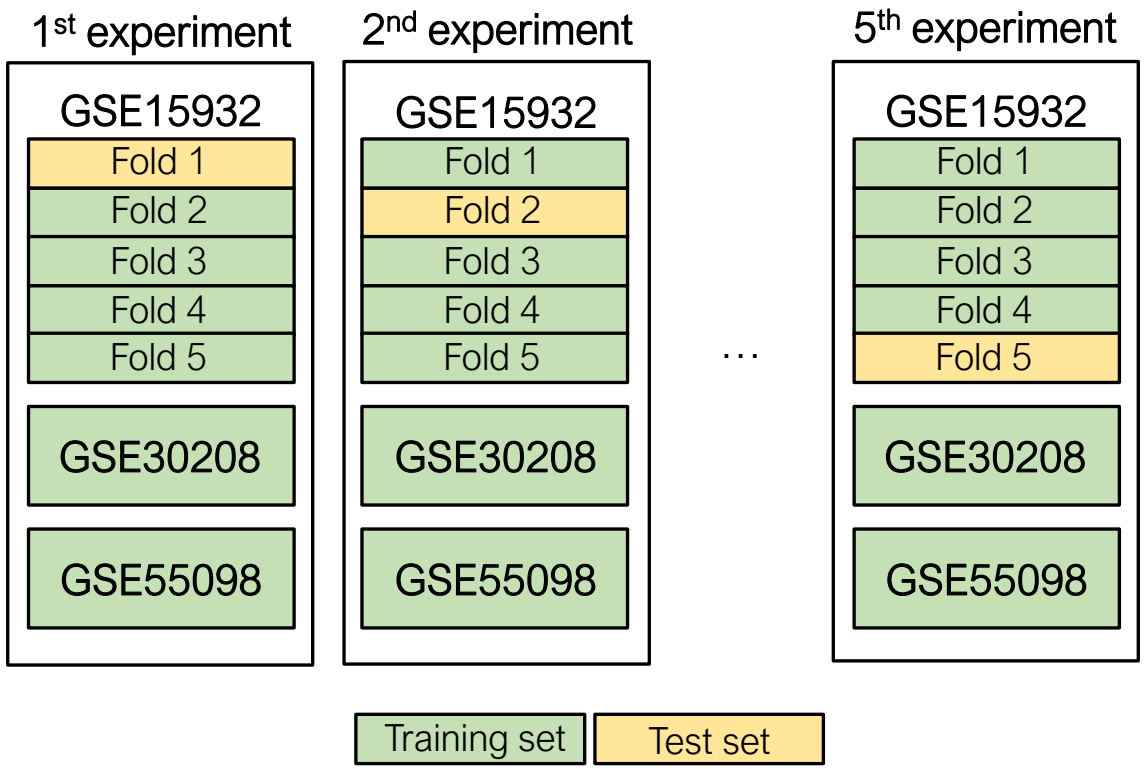
Datasets	Number of samples			Number of shared genes		
	Total	T1D	Non-T1D	GSE15932	GSE30208	GSE55098
GSE15932	63	37	26	368	0	0
GSE30208	22	12	10	0	764	337
GSE55098	16	8	8	0	337	764

	Number
Triples	2433477
Types of relations	56
GO classes	51375
Proteins	19169

EXPERIMENTAL SETUP

- To assess the efficacy of the proposed methodology, the diabetes performance on the GSE15932 dataset is analyzed by enriching the training data with information from the GSE30208 and GSE55098 datasets.

- A stratified cross-validation strategy is employed to ensure robust evaluation.



BASELINES

2 **baselines** that employ the expression values of the patient directly as input for the classifier:

Exclusively employing data from one **single dataset**.

Merging all measured genes across datasets and setting the value to 0 when the patient does not have an expression value.

	G1	G2	…
P1	0.1	0.9	…
P2	0.8	0.7	…
…	…	…	…
Avg	0.4	0.6	…

	G3	G4	…
P3	0.3	0.4	…
P4	0.5	0.8	…
…	…	…	…
Avg	0.4	0.3	…

→

	G1	G2	G3	G4
P1	0.1	0.9	0	0
P2	0.8	0.7	0	0
P3	0	0	0.3	0.4
P4	0	0	0.5	0.8

PERFORMANCE RESULTS

- The results confirm the hypothesis that **injecting other expression datasets can improve the performance of machine learning models**.
- The strategy involving the **weighted average of gene embeddings** for patient representation emerges as particularly promising.

	Acc	Pr	Re	F1	WAF	AUC
Baselines						
Only one dataset	0.554	**0.708**	0.561	0.578	0.529	0.560
Using all the datasets	0.442	0.650	0.314	0.396	0.422	0.474
Proposed Methodology						
Patient rep. using weighted avg. gene emb.	**0.619**	0.677	**0.739**	**0.683**	**0.589**	**0.606**
Patient rep. using KG with binning approach	0.481	0.565	0.579	0.551	0.460	0.466
Patient rep. using KG with patient-gene links	0.583	0.638	0.604	0.595	0.567	0.578

Table 1: Average diabetes prediction performance on the GSE30208 dataset for the baselines and our methodology.

ABLATION STUDY

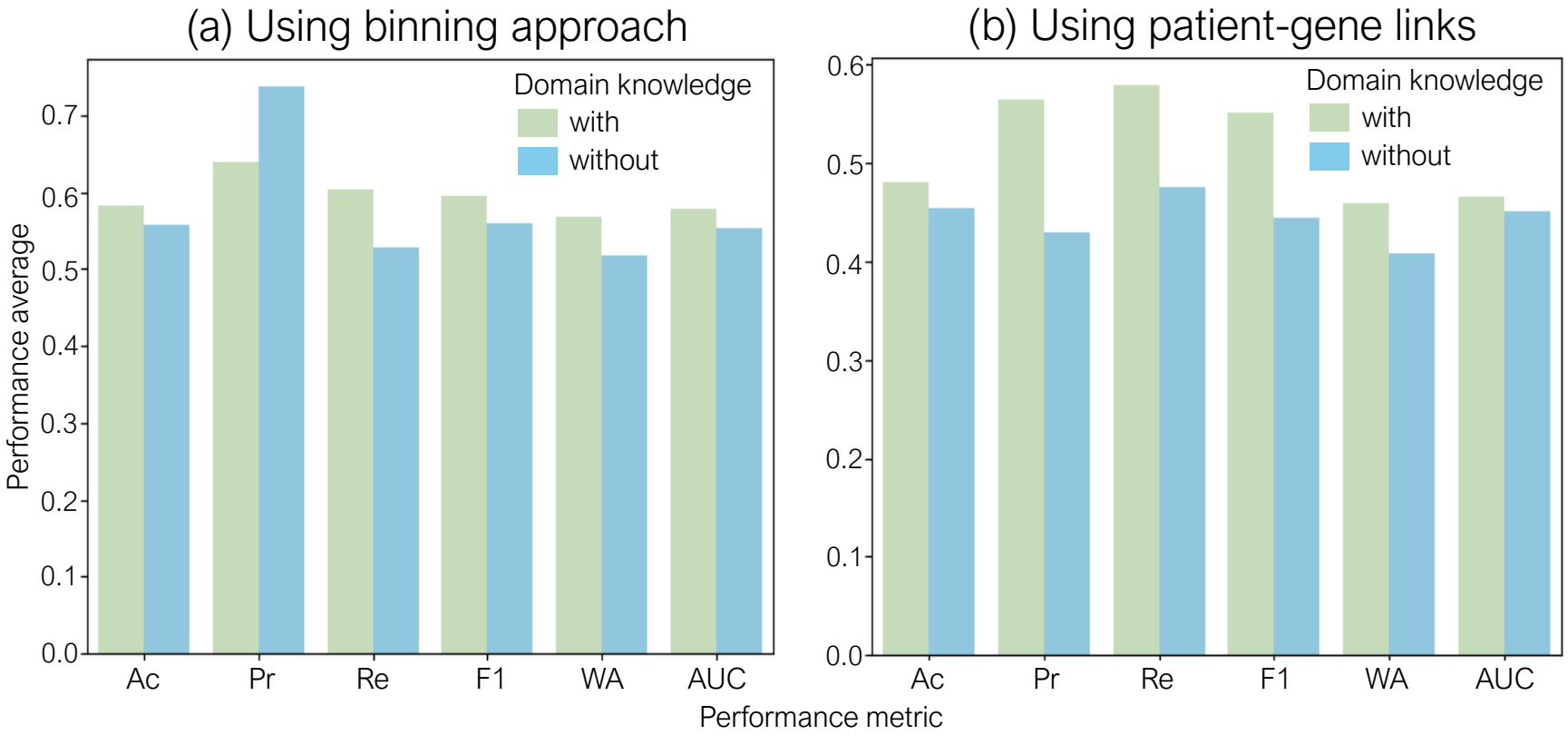


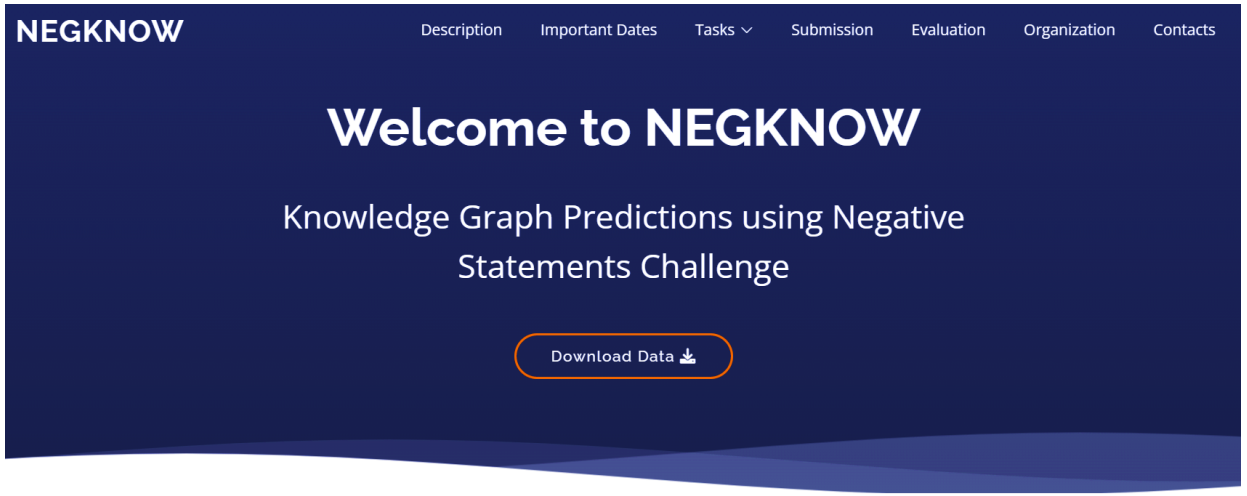
Figure 1: Performance comparison between using a KG with and without domain knowledge generated with (a) binning and (b) patient-gene links.

Knowledge about protein functions and interactions can play an important role in integrating data from datasets measuring gene expression across different genes.

CONCLUSIONS

- We present an approach that enables a comprehensive representation of gene expression data from different datasets within a KG.
- The results of our experiments showed that integrating gene expression data improves the performance of diabetes prediction.
- The proposed approach is versatile and can be extended to the prediction of other diseases.

NEGKNOW CHALLENGE @ ISWC



challengenegknow@gmail.com

https://negknow.github.io/NEGKNOW/index.html

CHALLENGE DESCRIPTION

This challenge aims to encourage participants to develop novel approaches that can effectively handle negative statements in knowledge graphs (KGs).

Since ontologies are already able to express negation and the enrichment of biomedical KGs with interesting negative statements is gaining traction, this challenge focuses on exploring ontology-rich biomedical KGs. These KGs use an ontology to provide rich descriptions of real-world entities instead of focusing on describing relations between entities themselves. Furthermore, there is an essential difference between a positive and a negative statement related to the implied inheritance in this kind of KG. A positive statement between an entity and an ontology class implies a positive statement between that entity and all the superclasses of the ontology class. Conversely, a negative statement between an entity and an ontology class implies a negative statement between the entity and all the subclasses of the ontology class.

Participants in this challenge will be evaluated on three relation prediction tasks. Relation prediction is the task of learning a relation between two KG entities (a pair) when the relation itself is not explicitly defined in the KG.

References:

- Negative statements considered useful [Arnaout et al., 2021]
- Inconsistencies, negations and changes in ontologies [Flouris et al., 2006]
- Biomedical knowledge graph embeddings with negative statements [Sousa et al., 2023]

THANK YOU FOR YOUR ATTENTION.

rita.sousa@uni-mannheim.de

@RitaTorresSousa

https://ritatsousa.github.io/

7th Workshop on SeWeBMeDA
26th of May 2023

UNIVERSITY OF MANNHEIM