Multi-dataset and Transfer Learning using Gene Expression Knowledge Graphs

Rita T. Sousa, Heiko Paulheim

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Extended Semantic Web Conference (ESWC)
03.06.2025

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

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 gene fragments (sequences), and $m << n$.

	S1	S2	S3	S4	S5	S6	…	Sn
P1	$GE_{P1,S1}$	$GE_{P1,S2}$	$GE_{P1,S3}$	$GE_{P1,S4}$	$GE_{P1,S5}$	$GE_{P1,S6}$	…	$GE_{P1,Sn}$
P2	$GE_{P2,S1}$	$GE_{P2,S2}$	$GE_{P2,S3}$	$GE_{P2,S4}$	$GE_{P2,S5}$	$GE_{P2,S6}$	…	$GE_{P2,Sn}$
…	…	…	…	…	…	…	…	…
Pm	$GE_{Pm,S1}$	$GE_{Pm,S2}$	$GE_{Pm,S3}$	$GE_{Pm,S4}$	$GE_{Pm,S5}$	$GE_{Pm,S6}$	…	$GE_{Pm,Sn}$

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Gene expression integration challenge

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

	S1	S2	…
P1	0.1	0.9	…
P2	0.8	0.7	…
…	…	…	…

	S3	S4	…
P3	0.3	0.4	…
P4	0.5	0.8	…
…	…	…	…

Solutions

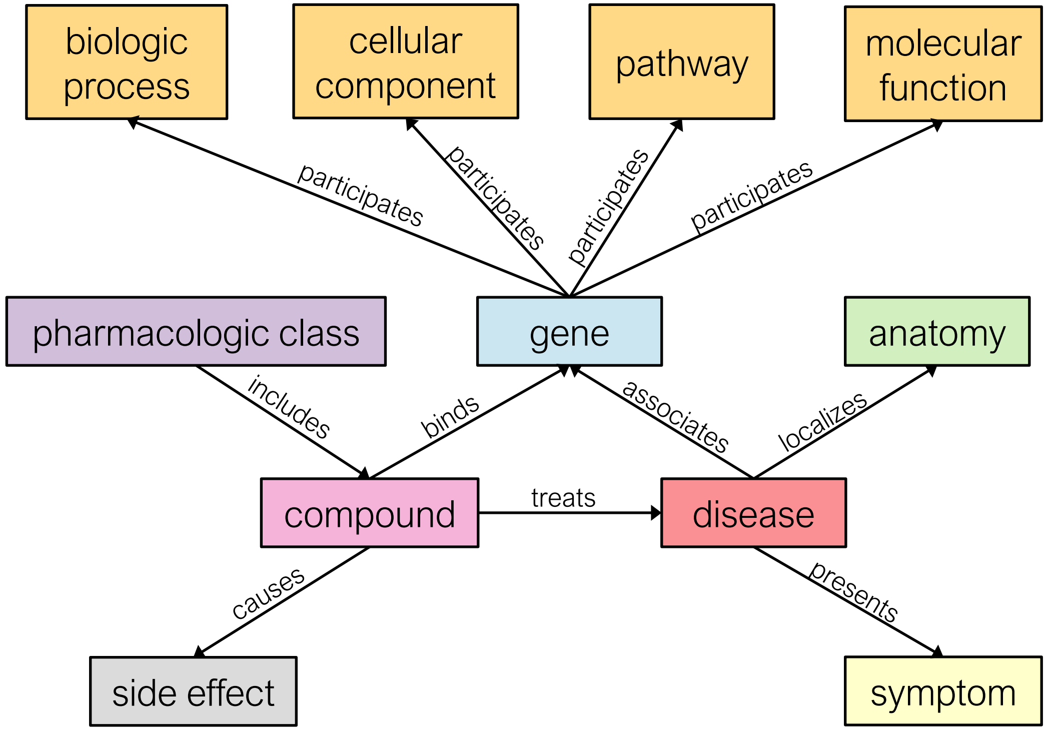
Use only one dataset, thereby having only little training data.

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

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

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.



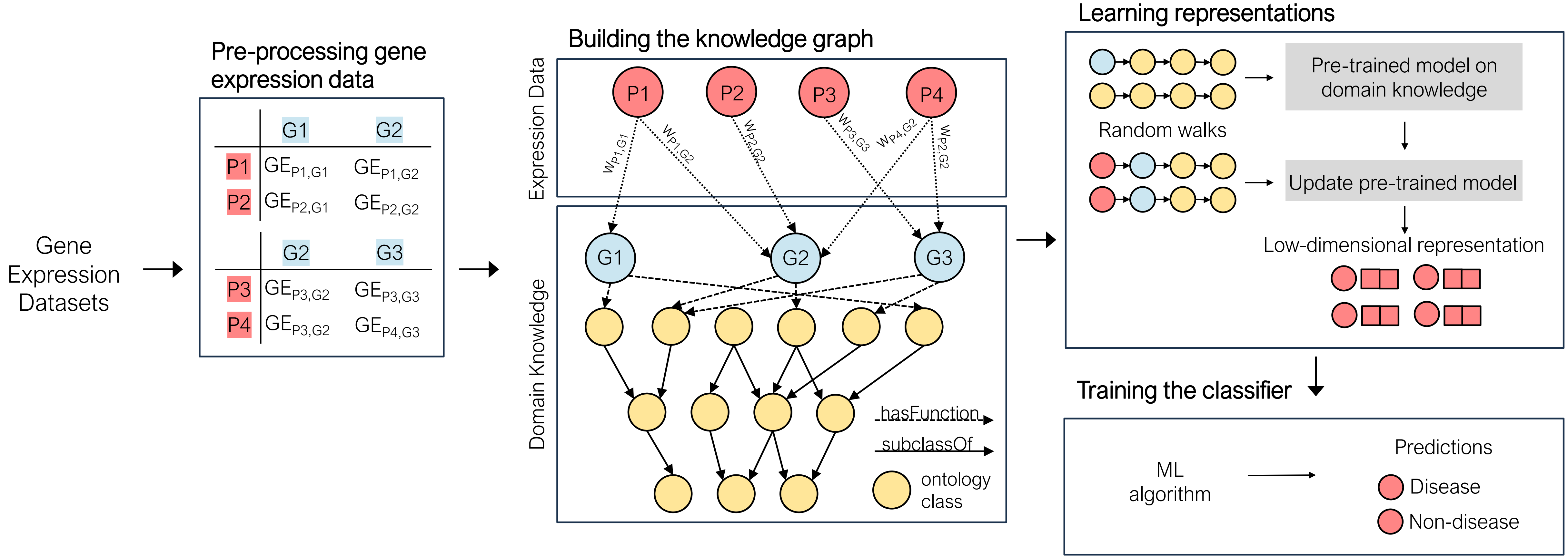
Methodology

With funding from the:

UNIVERSITY OF MANNHEIM

Federal Ministry of Research, Technology and Space

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



With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Methodology

Pre-processing of gene expression data

- Filtering out gene expression values corresponding to gene fragments without an associated gene are filtered out.

- Averaging expression values across all gene fragments corresponding to the same gene for each patient.

	S1	S2	S3	S4	S5	S6	…
P1	$GE_{P1,S1}$	$GE_{P1,S2}$	$GE_{P1,S3}$	$GE_{P1,S4}$	$GE_{P1,S5}$	$GE_{P1,S6}$	…
P2	$GE_{P2,S1}$	$GE_{P2,S2}$	$GE_{P2,S3}$	$GE_{P2,S4}$	$GE_{P2,S5}$	$GE_{P2,S6}$	…
…	…	…	…	…	…	…	…
Pm	$GE_{Pm,S1}$	$GE_{Pm,S2}$	$GE_{Pm,S3}$	$GE_{Pm,S4}$	$GE_{Pm,S5}$	$GE_{Pm,S6}$	…

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

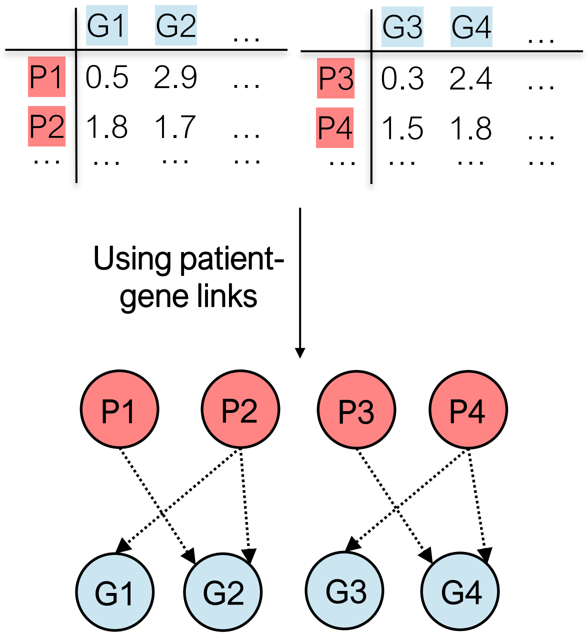
With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Methodology

Building the knowledge graph

The KG is built by integrating:

- **Gene expression data** using a strategy where a patients and genes are linked based on expression values.



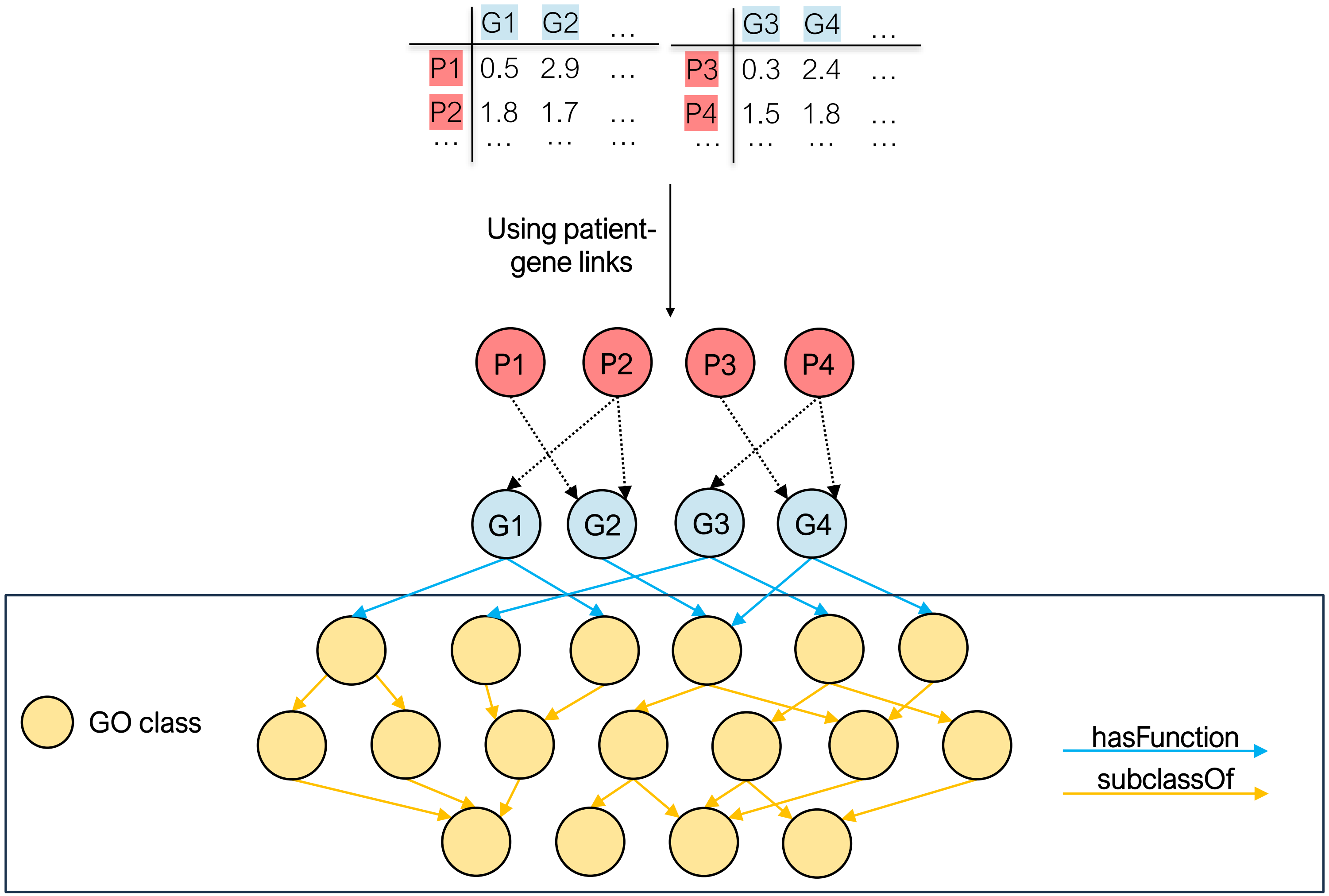
With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Methodology

Building the knowledge graph

The KG is built by integrating:

- **Gene expression data** using a strategy where a patients and genes are linked based on expression values.
- **Domain-specific knowledge** including Gene Ontology (GO) data.



With funding from the:

UNIVERSITY OF MANNHEIM

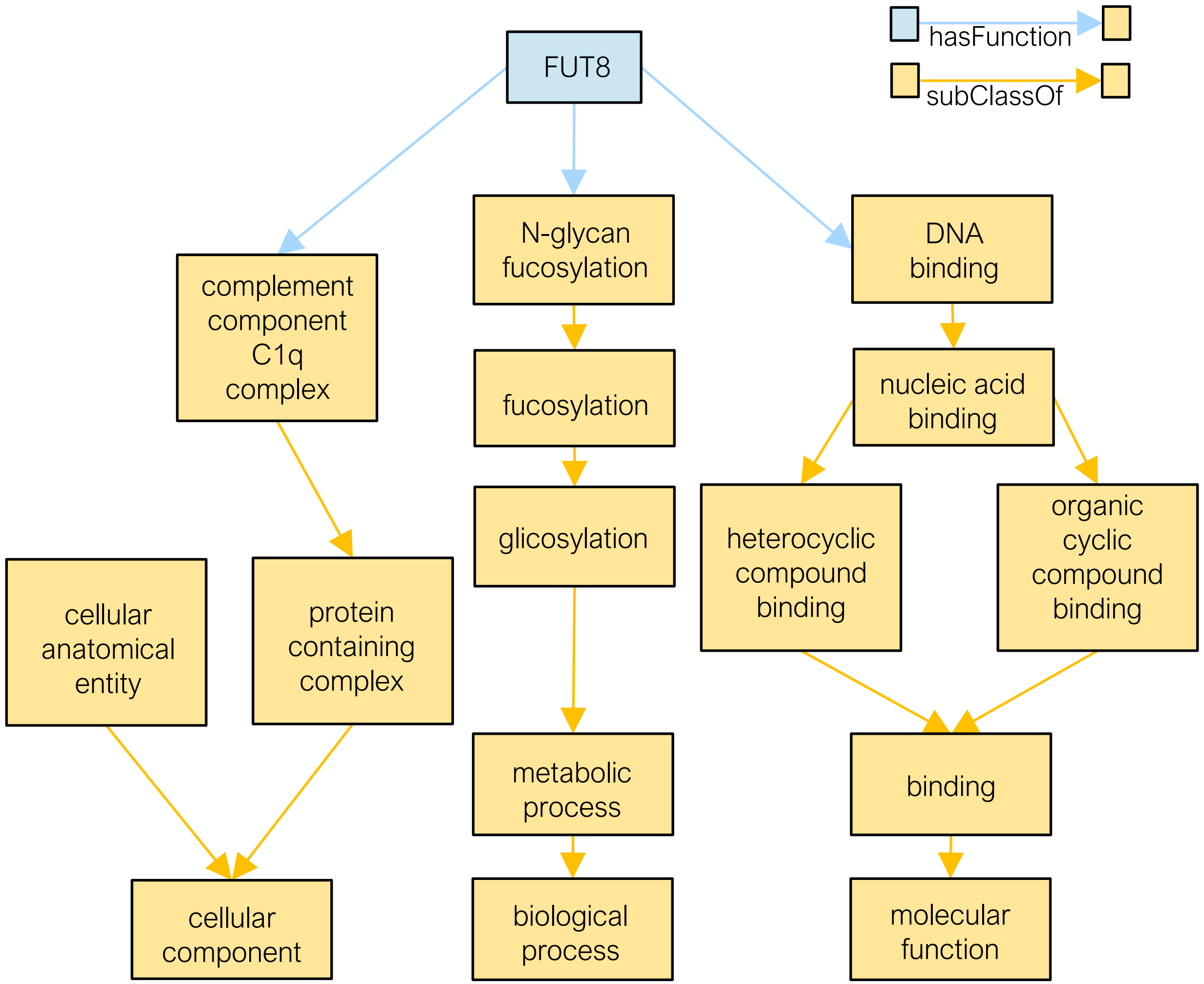
Federal Ministry of Research, Technology and Space

Methodology

Building the knowledge graph

The KG is built by integrating:

- **Gene expression data** using a strategy where a patients and genes are linked based on expression values.
- **Domain-specific knowledge** including Gene Ontology (GO) data.

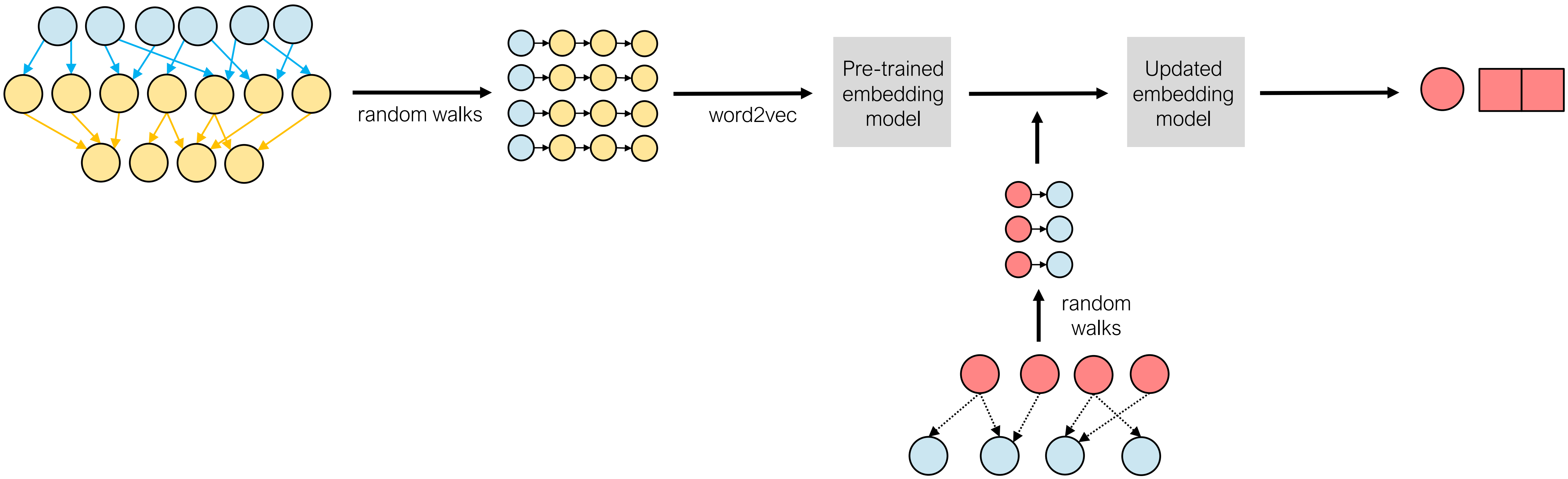


With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Methodology

Learning patient representations

- RDF2Vec is used to generate low-dimensional representations for each KG node.
- RDF2vec is capable of adapting its vectors upon updates in the knowledge graph without a full retraining.



With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

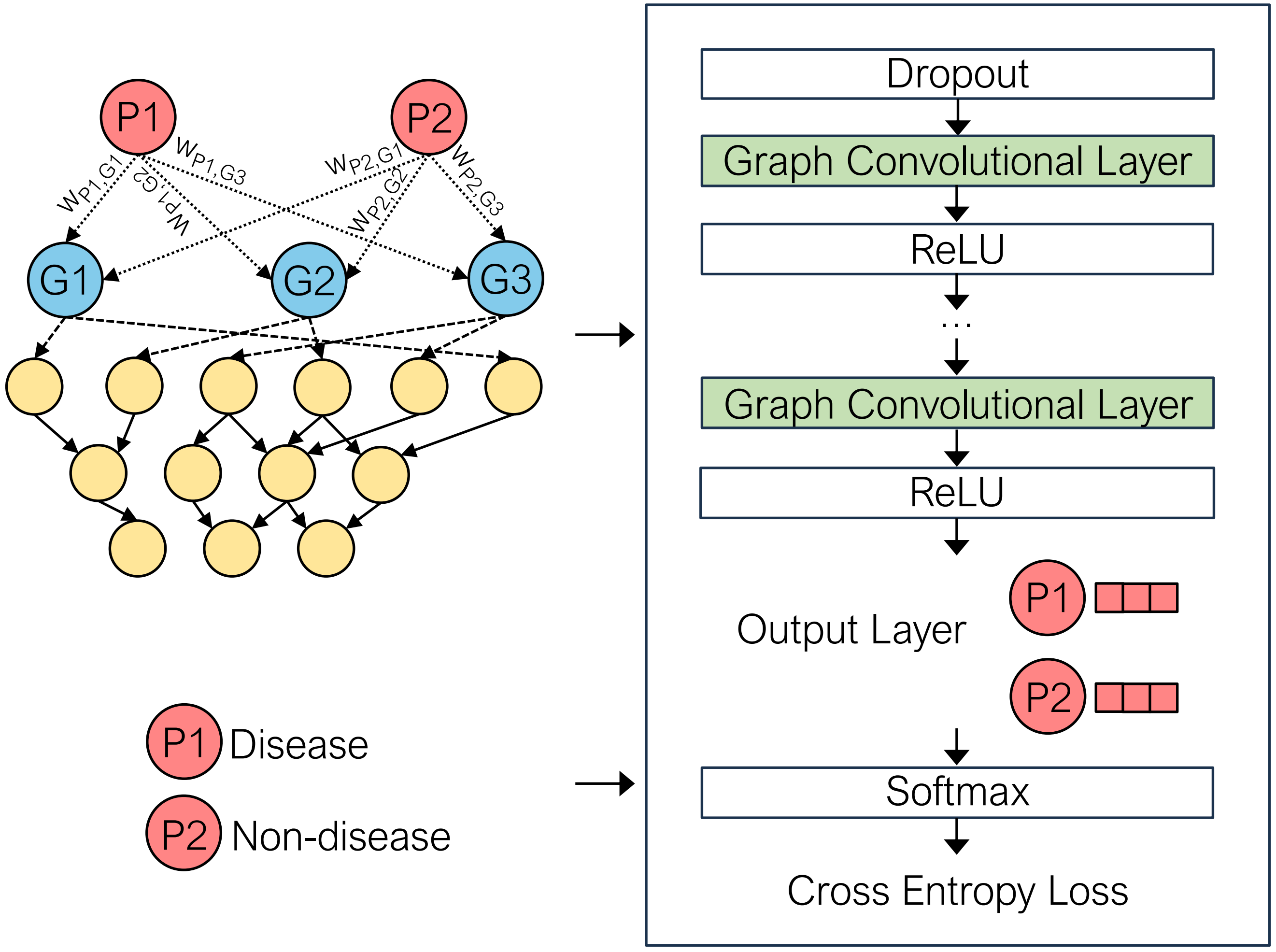
Methodology

Training a supervised learning algorithm

Patient diagnosis is formulated as a binary classification task.

Two approaches:

- MLP classifier
- GCN



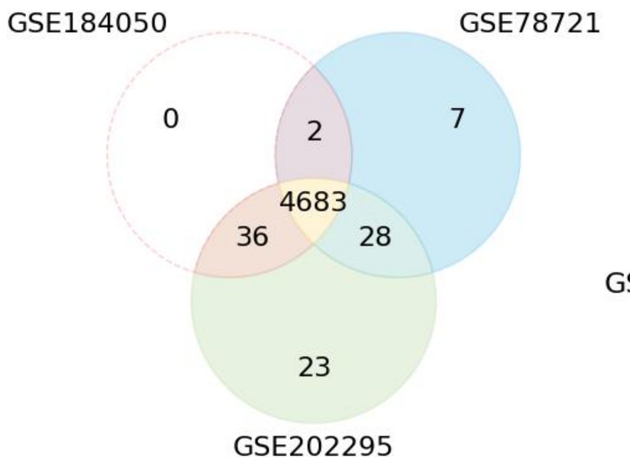
With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Experimental data

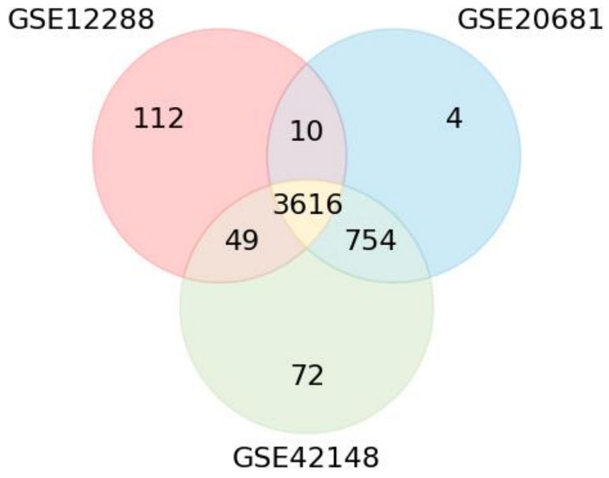
GEO datasets for three diseases are considered.

Gene Expression Omnibus

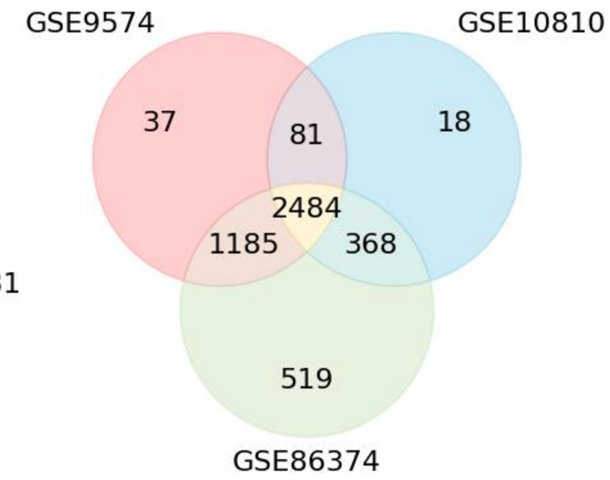
Disease	Dataset	Positive Samples	Negative Samples
Diabetes type II	GSE184050	50	66
	GSE78721	68	62
	GSE202295	61	50
Coronary artery disease	GSE12288	110	112
	GSE20681	99	99
	GSE42148	13	11
Breast cancer	GSE9574	14	15
	GSE10810	31	27
	GSE86374	124	35



Diabetes type 2



Coronary artery disease

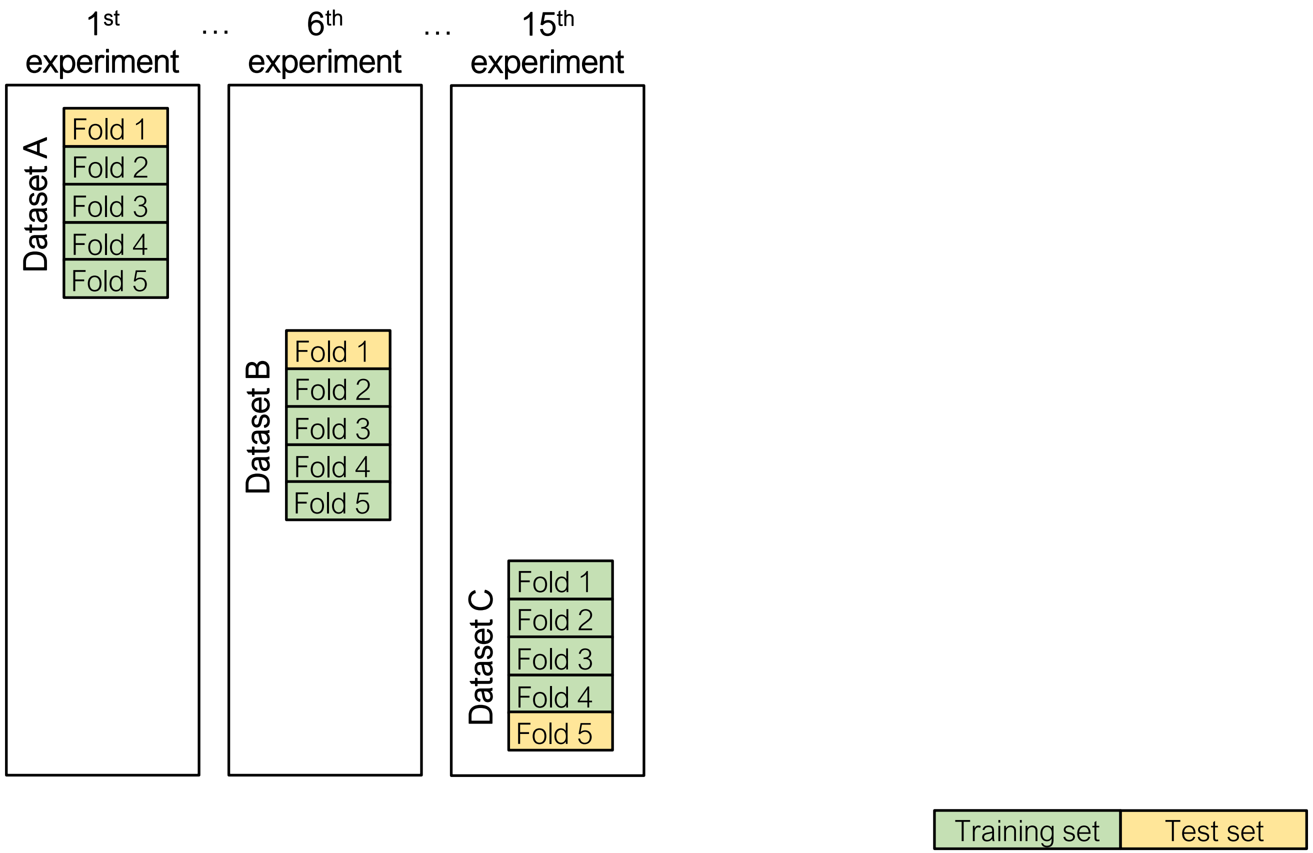


Breast cancer

Experimental setup

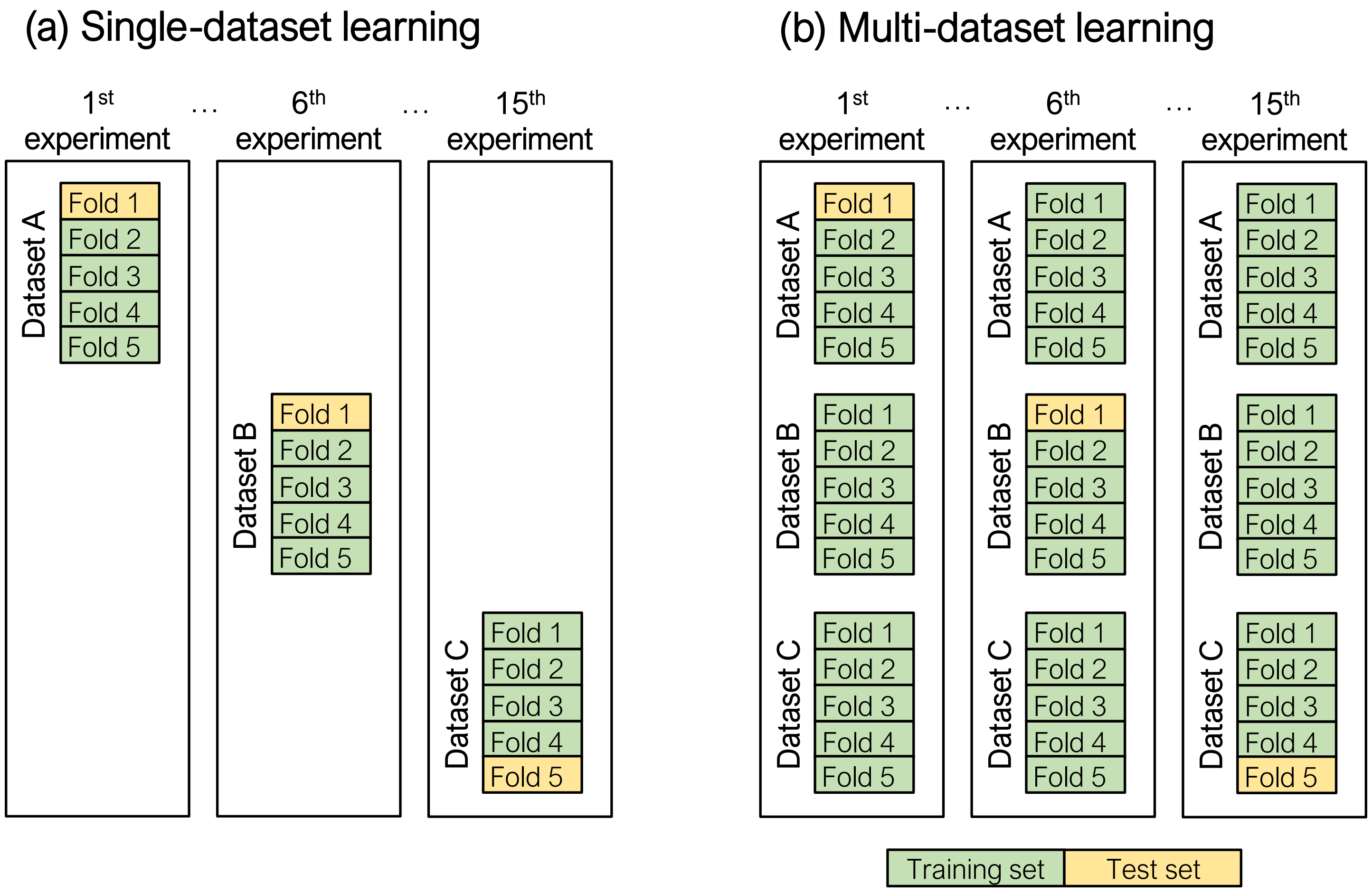
With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

(a) Single-dataset learning



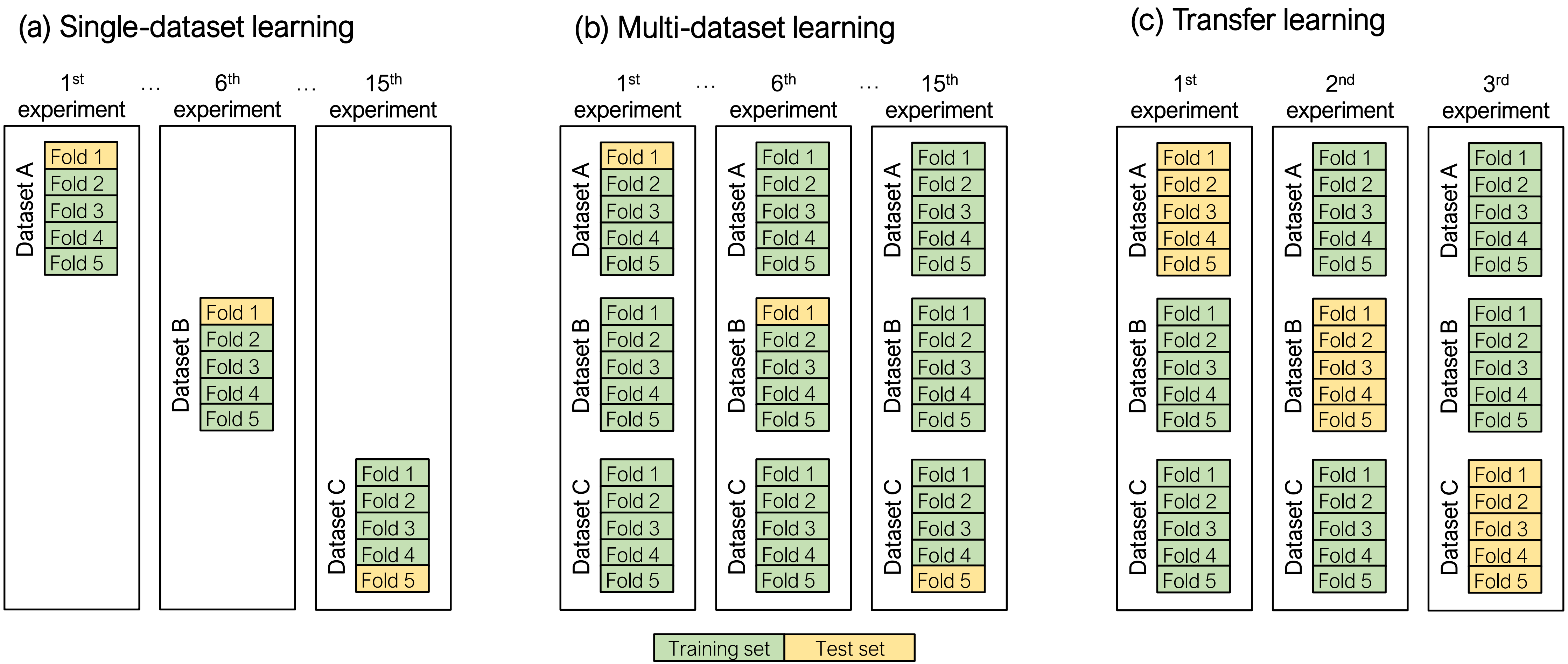
Experimental setup

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space



With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Experimental setup



With funding from the:

UNIVERSITY OF MANNHEIM

Federal Ministry of Research, Technology and Space

Baselines

- For each setting, our methodology is compared to baselines that employ the processed expression values directly as input for an MLP.

- In the multi-dataset and transfer learning settings, the baseline includes two variations: including all genes or including the overlapping genes.

	G1	G2	…
P1	0.1	0.9	…
P2	0.8	0.7	…
…	…	…	…

	G2	G3	…
P3	0.3	0.4	…
P4	0.5	0.8	…
…	…	…	…

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

	G2	…
P1	0.9	…
P2	0.7	…
P3	0.3	…
P4	0.5	…

Performance results

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Table: Mean and standard deviation for weighted average f1-score, comparing the baselines and our methodology when coupled with MLP or GCN for the 3 settings.

Disease	Dataset	Single-dataset learning			Multi-dataset learning				Transfer learning			
		Baseline	Ours		Baseline		Ours		Baseline		Ours	
			MLP	GNN	All	Overlap	MLP	GNN	All	Overlap	MLP	GNN
Diabetes type II	GSE184050	0.495 (0.095)	0.742 (0.084)	0.809 (0.086)	0.450 (0.125)	**0.559** (0.095)	0.525 (0.095)	0.525 (0.135)	**0.432**	0.256	0.373	0.363
	GSE78721	0.391 (0.114)	0.532 (0.113)	0.563 (0.128)	0.402 (0.064)	0.359* (0.021)	**0.452** (0.089)	0.401 (0.106)	*0.359	**0.513**	0.431	0.410
	GSE202295	**0.507** (0.116)	0.504 (0.107)	0.470 (0.114)	0.424 (0.123)	0.548 (0.157)	0.463 (0.048)	0.407 (0.056)	*0.390	0.318	**0.381**	0.372
Coronary artery disease	GSE12288	0.440 (0.074)	0.523 (0.049)	0.568 (0.075)	0.408 (0.079)	0.479 (0.061)	**0.496** (0.043)	0.482 (0.077)	*0.338	0.328*	**0.468**	0.466
	GSE20681	0.328 (0.007)	**0.544** (0.075)	0.542 (0.060)	0.339 (0.007)	0.333* (0.009)	0.380 (0.040)	**0.520** (0.060)	*0.333	0.333*	0.519	0.549
	GSE42148	0.338 (0.099)	0.450 (0.171)	0.564 (0.282)	**0.448** (0.288)	0.338 (0.099)	0.442 (0.207)	0.338 (0.179)	*0.288	0.391	**0.417**	0.324
Breast cancer	GSE9574	**0.405** (0.113)	0.355 (0.226)	0.394 (0.115)	0.578 (0.222)	0.479 (0.197)	0.394 (0.281)	0.537 (0.188)	*0.353	**0.425**	0.386	0.299
	GSE10810	0.558 (0.293)	0.897 (0.099)	0.879 (0.103)	0.576 (0.316)	0.700 (0.306)	0.779 (0.179)	**0.802** (0.107)	*0.372	0.372*	**0.751**	**0.751**
	GSE86374	0.441 (0.296)	0.869 (0.140)	0.865 (0.127)	0.586 (0.194)	0.562 (0.242)	**0.834** (0.054)	0.810 (0.051)	*0.079	0.683*	**0.671**	**0.671**

* The classifier predicts everything with either label 0 or label 1

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Performance results

Single-dataset learning

Table: Mean and standard deviation for weighted average f1-score, comparing the baselines and our methodology when coupled with MLP or GCN for the 3 settings.

Disease	Dataset	Single-dataset learning			Multi-dataset learning				Transfer learning			
		Baseline	Ours		Baseline		Ours		Baseline		Ours	
			MLP	GNN	All	Overlap	MLP	GNN	All	Overlap	MLP	GNN
Diabetes type II	GSE184050	0.495 (0.095)	0.742 (0.084)	**0.809** (0.086)	0.450 (0.125)	**0.559** (0.095)	0.525 (0.095)	0.525 (0.135)	**0.432**	0.256	0.373	0.363
	GSE78721	0.391 (0.114)	0.532 (0.113)	**0.563** (0.128)	0.402 (0.064)	0.359* (0.021)	**0.452** (0.089)	0.401 (0.106)	*0.359	**0.513**	0.431	0.410
	GSE202295	**0.507** (0.116)	0.504 (0.107)	0.470 (0.114)	0.424 (0.123)	[illegible]	[illegible]	[illegible] (0.056)	*0.390	0.318	**0.381**	0.372
Coronary artery disease	GSE12288	0.440 (0.074)	0.523 (0.049)	**0.568** (0.075)	0.408 (0.079)	[illegible]	[illegible]	[illegible]	*0.338	0.328*	**0.468**	0.466
	GSE20681	0.328 (0.007)	**0.544** (0.075)	0.542 (0.060)	0.339 (0.007)	[illegible]	[illegible]	[illegible] (0.060)	*0.333	0.333*	0.519	0.549
	GSE42148	0.338 (0.099)	0.450 (0.171)	**0.564** (0.282)	**0.448** (0.288)	[illegible]	0.442 (0.207)	0.338 (0.179)	*0.288	0.391	**0.417**	0.324
Breast cancer	GSE9574	**0.405** (0.113)	0.355 (0.226)	0.394 (0.115)	0.578 (0.222)	0.479 (0.197)	0.394 (0.281)	0.537 (0.188)	*0.353	**0.425**	0.386	0.299
	GSE10810	0.558 (0.293)	**0.897** (0.099)	0.879 (0.103)	0.576 (0.316)	0.700 (0.306)	0.779 (0.179)	**0.802** (0.107)	*0.372	0.372*	**0.751**	**0.751**
	GSE86374	0.441 (0.296)	**0.869** (0.140)	0.865 (0.127)	0.586 (0.194)	0.562 (0.242)	**0.834** (0.054)	0.810 (0.051)	*0.079	0.683*	**0.671**	**0.671**

Contextualizing genetic information improves patient diagnosis, with considerable improvements for some datasets.

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Performance results

Multi-dataset learning

Table: Mean and standard deviation for weighted average f1-score, comparing the baselines and our methodology when coupled with MLP or GCN for the 3 settings.

Disease	Dataset	Single-dataset learning			Multi-dataset learning				Transfer learning			
		Baseline	Ours		Baseline		Ours					
			MLP	GNN	All	Overlap	MLP	GNN	All	Overlap	MLP	GNN
Diabetes type II	GSE184050	0.495 (0.095)	0.742 (0.084)	0.809 (0.086)	0.450 (0.125)	**0.559** (0.095)	0.525 (0.095)	0.525 (0.135)	[illegible]	[illegible]	0.373	0.363
	GSE78721	0.391 (0.114)	0.532 (0.113)	0.563 (0.128)	0.402 (0.064)	0.359* (0.021)	**0.452** (0.089)	0.401 (0.106)	[illegible]	[illegible]	0.431	0.410
	GSE202295	**0.507** (0.116)	0.504 (0.107)	0.470 (0.114)	0.424 (0.123)	0.548 (0.157)	0.463 (0.048)	0.407 (0.056)	[illegible]	[illegible]	0.381	0.372
Coronary artery disease	GSE12288	0.440 (0.074)	0.523 (0.049)	0.568 (0.075)	0.408 (0.079)	0.479 (0.061)	**0.496** (0.043)	0.482 (0.077)	[illegible]	[illegible]	0.468	0.466
	GSE20681	0.328 (0.007)	**0.544** (0.075)	0.542 (0.060)	0.339 (0.007)	0.333* (0.009)	0.380 (0.040)	**0.520** (0.060)	[illegible]	[illegible]	0.519	0.549
	GSE42148	0.338 (0.099)	0.450 (0.171)	0.564 (0.282)	**0.448** (0.288)	0.338 (0.099)	0.442 (0.207)	0.338 (0.179)	[illegible]	[illegible]	0.417	0.324
Breast cancer	GSE9574	**0.405** (0.113)	0.355 (0.226)	0.394 (0.115)	0.578 (0.222)	0.479 (0.197)	0.394 (0.281)	0.537 (0.188)	[illegible]	[illegible]	0.386	0.299
	GSE10810	0.558 (0.293)	0.897 (0.099)	0.879 (0.103)	0.576 (0.316)	0.700 (0.306)	0.779 (0.179)	**0.802** (0.107)	[illegible]	[illegible]	0.751	0.751
	GSE86374	0.441 (0.296)	0.869 (0.140)	0.865 (0.127)	0.586 (0.194)	0.562 (0.242)	**0.834** (0.054)	0.810 (0.051)	[illegible]	[illegible]	0.671	0.671

* The classifier predicts everything with either label 0 or label 1

Diverse range of data sources can enhance performance in smaller datasets.

Performance results

Transfer learning

With funding from the:

UNIVERSITY OF MANNHEIM

Federal Ministry of Research, Technology and Space

Table: Mean and standard deviation for weighted average f1-score, comparing the baselines and our methodology when coupled with MLP or GCN for the 3 settings.

Disease	Dataset	Single-dataset learning: Baseline	Single-dataset learning: Ours MLP	Single-dataset learning: Ours GNN	Multi-dataset learning: Baseline All	Multi-dataset learning: Baseline Overlap	Multi-dataset learning: Ours MLP	Multi-dataset learning: Ours GNN	Transfer learning: Baseline All	Transfer learning: Baseline Overlap	Transfer learning: Ours MLP	Transfer learning: Ours GNN
Diabetes type II	GSE184050	0.495 (0.095)	0.742 (0.084)	0.809 (0.086)	0.450 (0.125)	0.559 (0.095)	0.525 (0.095)	0.525 (0.135)	0.432	0.256	0.373	0.363
	GSE78721	0.391 (0.114)	0.532 (0.113)	0.563 (0.128)	0.402 (0.064)	0.359* (0.021)	0.452 (0.089)	0.401 (0.106)	*0.359	0.513	0.431	0.410
	GSE202295	0.507 (0.116)	0.504 (0.107)	0.470 (0.114)	0.424 (0.123)	0.548 (0.157)	0.463 (0.048)	0.407 (0.056)	*0.390	0.318	0.381	0.372
Coronary artery disease	GSE12288	0.440 (0.074)	0.523 (0.049)	0.568 (0.075)	0.408 (0.079)	0.479 (0.061)	0.496 (0.043)	0.482 (0.077)	*0.338	0.328*	0.468	0.466
	GSE20681	0.328 (0.007)	0.544 (0.075)	0.542 (0.060)	0.339 (0.007)	0.333* (0.009)	0.380 (0.040)	0.520 (0.060)	*0.333	0.333*	0.519	0.549
	GSE42148	0.338 (0.099)	0.450 (0.171)	0.564 (0.282)	0.448 (0.288)	[illegible]	[illegible]	0.338 (0.179)	*0.288	0.391	0.417	0.324
Breast cancer	GSE9574	0.405 (0.113)	0.355 (0.226)	0.394 (0.115)	0.578 (0.222)	[illegible]	[illegible]	0.537 (0.188)	*0.353	0.425	0.386	0.299
	GSE10810	0.558 (0.293)	0.897 (0.099)	0.879 (0.103)	0.576 (0.316)	[illegible]	[illegible]	802 (0.107)	*0.372	0.372*	0.751	0.751
	GSE86374	0.441 (0.296)	0.869 (0.140)	0.865 (0.127)	0.586 (0.194)	[illegible]	[illegible]	0.810 (0.051)	*0.079	0.683*	0.671	0.671

* The classifier predicts everything with either label 0 or label 1

Results similar to those obtained in single and multi-dataset settings for our methodology

KG Embeddings

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

3 clusters, each from a different dataset, with no clear split between positive and negative classes

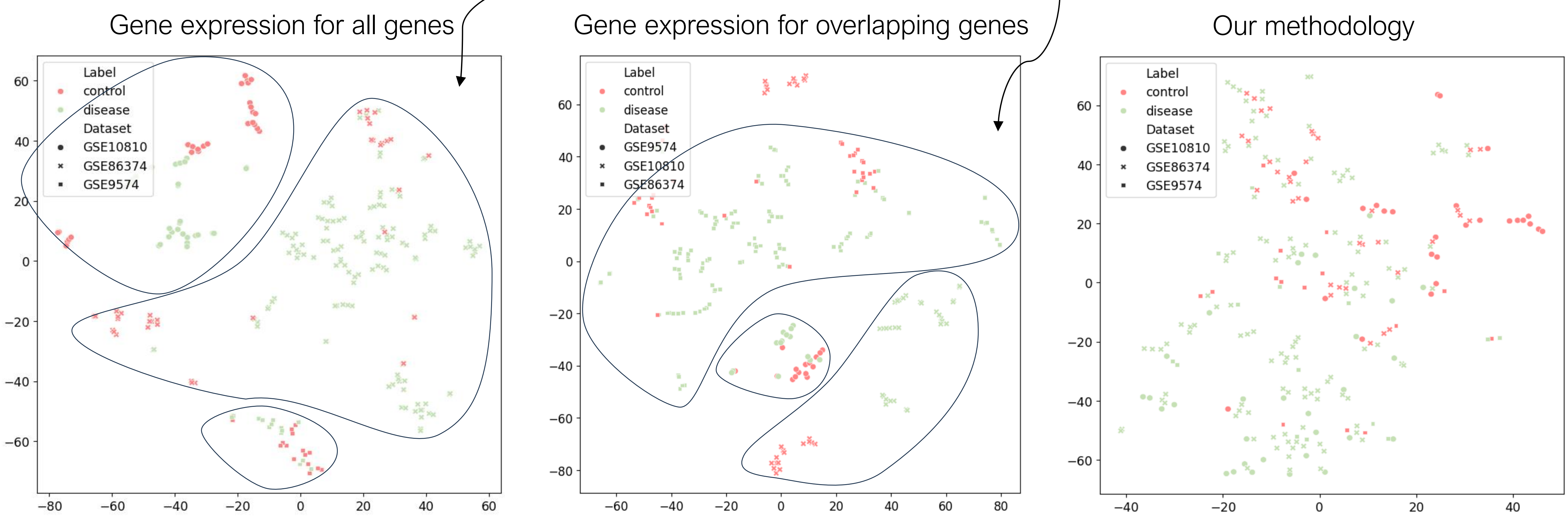


Figure: t-SNE plots comparing patient representations based on the gene expression values (using all genes or only the overlapping genes across the three datasets) to patient representations generated based on KG embeddings. Each point represents a patient, with the color indicating the label and the shape indicating the dataset they originate from.

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

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 patient diagnosis.
- The proposed approach is versatile and can be extended to combining datasets with incompatible features beyond the gene expression domain.

Thank you for your attention!

With funding from the:

UNIVERSITY OF MANNHEIM

Federal Ministry of Research, Technology and Space

rita.sousa@uni-mannheim.de

https://ritatsousa.github.io/

With funding from the:

UNIVERSITY OF MANNHEIM

Federal Ministry of Research, Technology and Space

Additional Slides

With funding from the:
UNIVERSITY OF MANNHEIM
Federal Ministry of Research, Technology and Space

Ablation studies

The performance of a GCN when the input node features are replaced with randomly initialized values and when the model receives as input unweighted graph.

KG embeddings with GCNs consistently outperforms GCNs lacking node features across all datasets.

The influence of using weighted edges seems to be less significant.

Figure: Bar plot depicting the F-score comparisons between different GCN configurations: one using weighted edges and KG embeddings as node features (pink bars), another with randomly initialized node features (blue bars), and another without weighted edges (green bars)