GCondNet: A Novel Method for Improving Neural Networks on Small High-Dimensional Tabular Data

Andrei Margeloiu, Nikola Simidjievski, Pietro Lio, Mateja Jamnik
University of Cambridge, UK




Introduction

short & wide tables

D features

N samples

$D >> N$

Tabular datasets in medicine and bioinformatics are usually:

- **high-dimensional** (5,000 - 20,000 features)
- **small size** (~100s samples/datapoints) ($D >> N$)

Problem: Neural networks tend to overfit on such small datasets, partially because the networks have too many degrees of freedom.

Research question: How to reduce overfitting and improve the accuracy of neural networks on tabular datasets with $D >> N$?

We propose a general method for $D >> N$

GCondNet improves neural networks by:

1. extracting the "*implicit* relationships" between samples

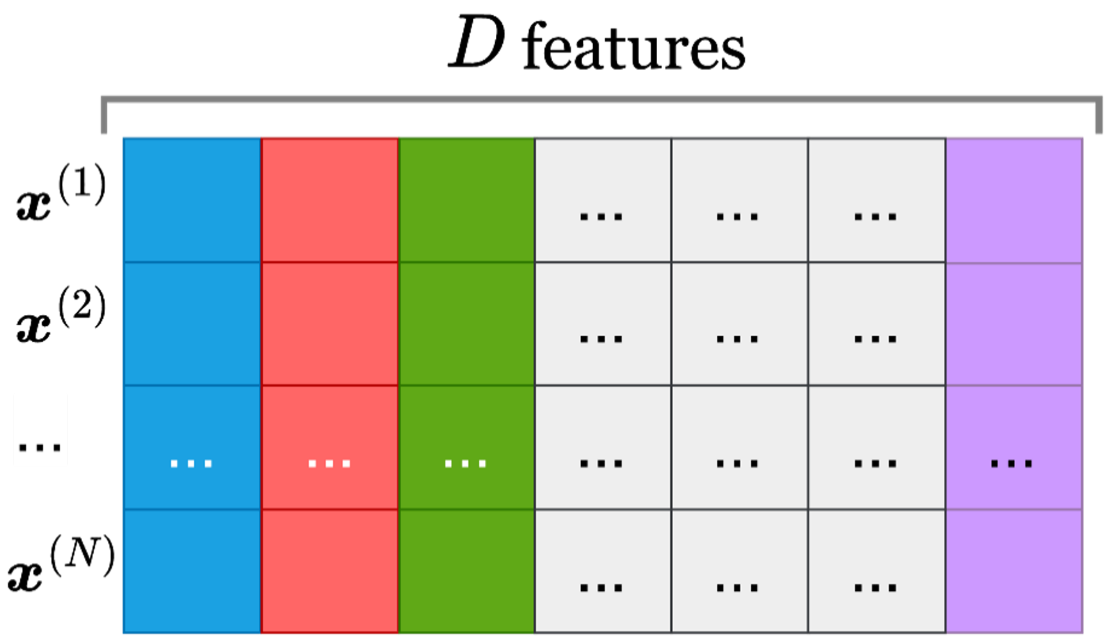
2. performing "soft parameter-sharing" to constrain the model's parameters

- ✅ simple & general
- ✅ improved accuracy
- ✅ robust to incorrect relationships

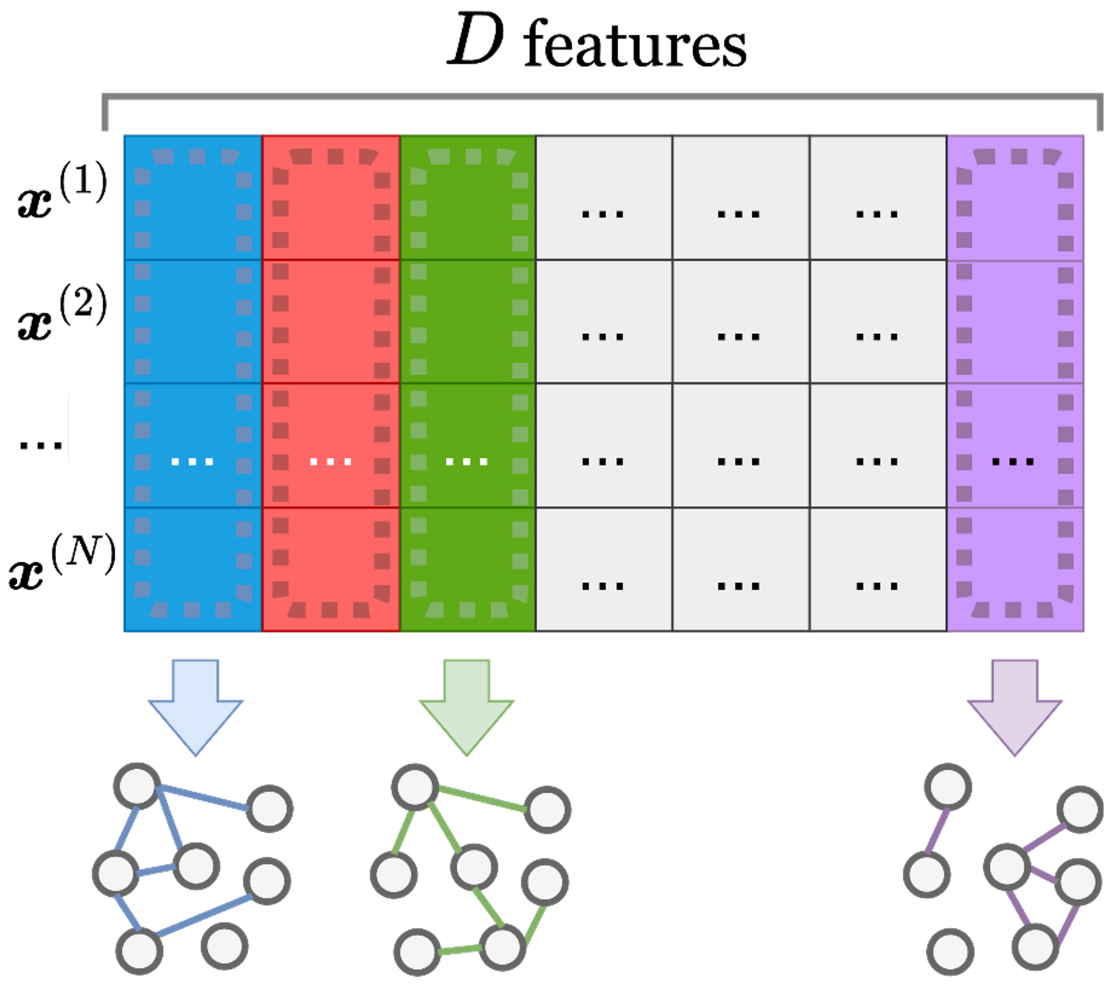
Key idea: use the implicit sample-wise relationships

Setup: Any tabular dataset, where each row $x^{(i)}$ is a sample/datapoint.

Standard methods capture the relationships between features.

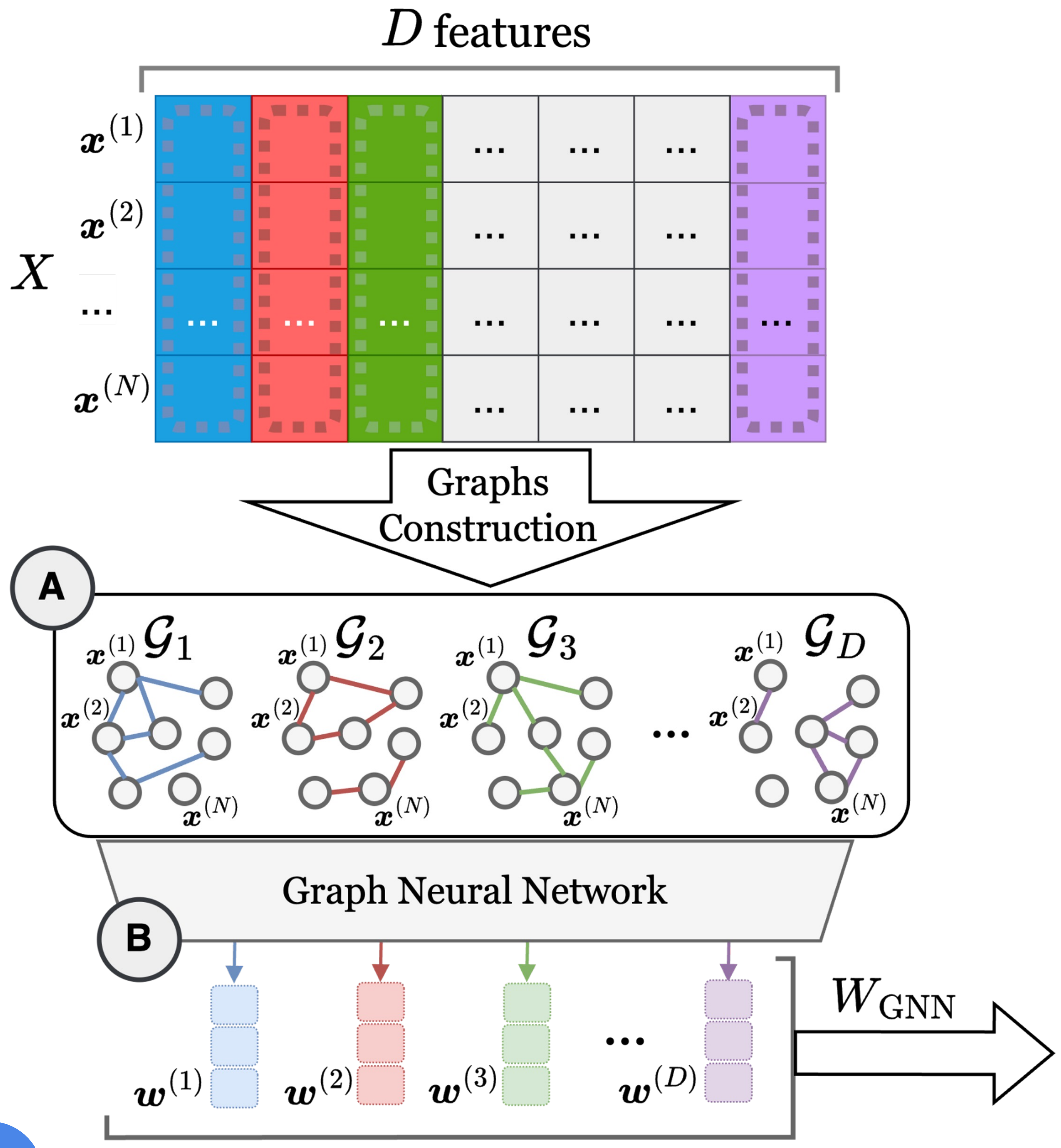


Key idea: GCondNet *additionally* leverages the implicit relationships between samples *across each feature*.



1

Represent the sample-wise relationships as graphs



A Generate a graph for each feature in the dataset (resulting in D graphs), with each node representing a sample (totalling N nodes per graph).

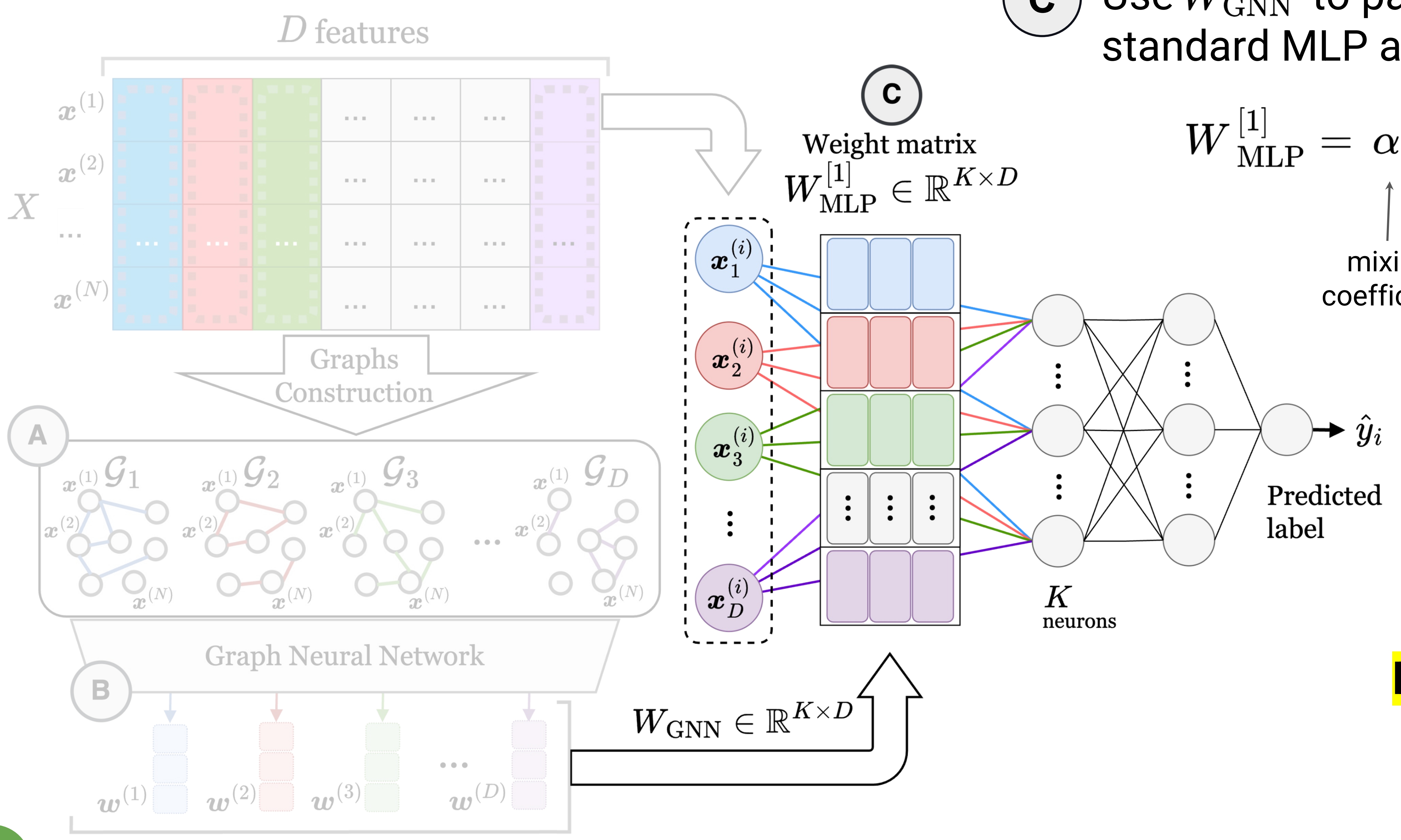
B Use a Graph Neural Network (GNN) to extract graph embeddings $w^{(j)}$ from each of the D graphs.

Concatenate all graph embeddings into matrix $W_{\text{GNN}} = \left[w^{(1)}, w^{(2)}, \ldots, w^{(D)}\right]$

1

UNIVERSITY OF CAMBRIDGE

Performing soft-parameter-sharing



C Use W_{GNN} to parameterise the first layer of a standard MLP as a convex combination

$$W_{\text{MLP}}^{[1]} = \underbrace{\alpha}_{\text{mixing coefficient}} W_{\text{GNN}} + (1-\alpha) \underbrace{W_{\text{scratch}}}_{\text{initialised to zero}}$$

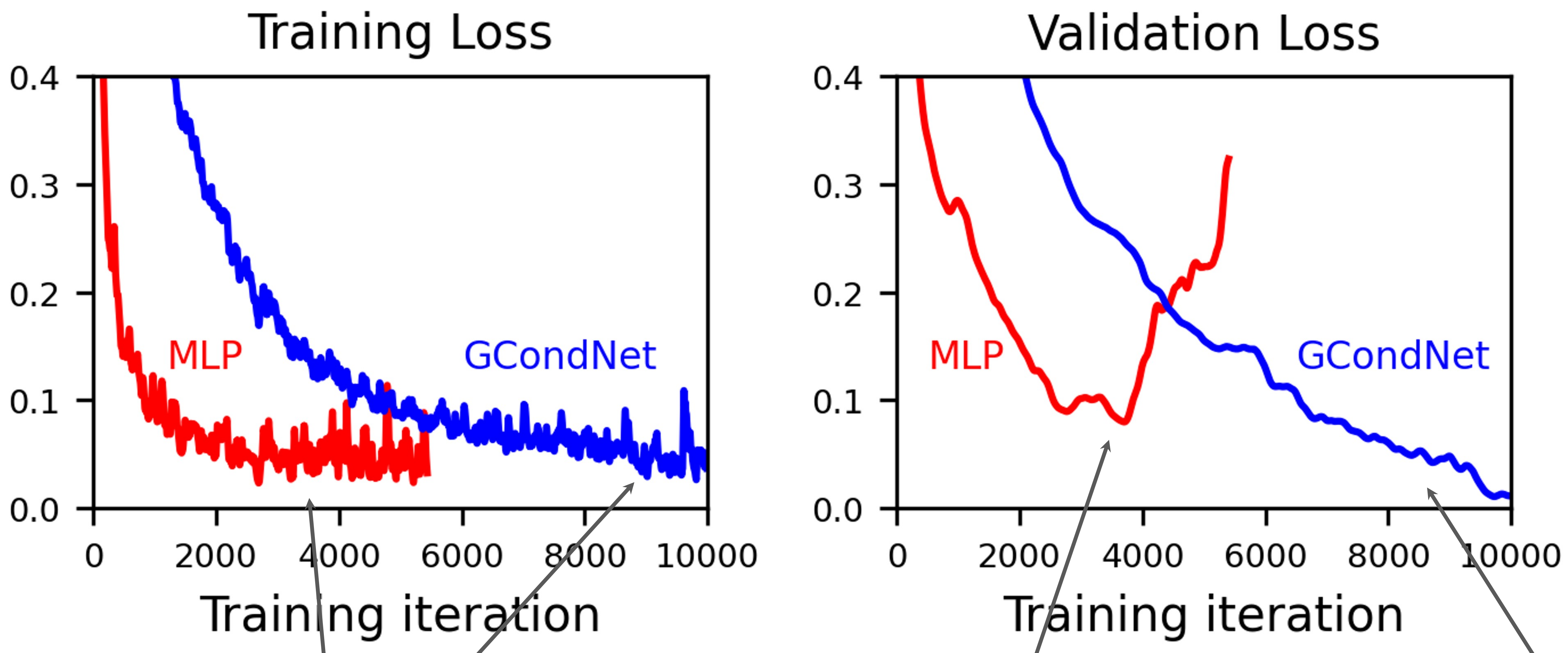
End-to-end training.

2

UNIVERSITY OF CAMBRIDGE

The impact of GCondNet's inductive bias

Experiment: Compare the loss curves on MLP and an equivalent GCondNet (*averaged over 25 runs*)



(1) Both the MLP and GCondNet fit the training data

(2) MLP overfits

(3) GCondNet achieves lower validation error

UNIVERSITY OF CAMBRIDGE

GCondNet outperforms other methods

- We evaluate on 9 real-world biomedical datasets
- GCondNet outperforms 15 standard and modern methods, including specialised methods for tabular datasets with $D >> N$

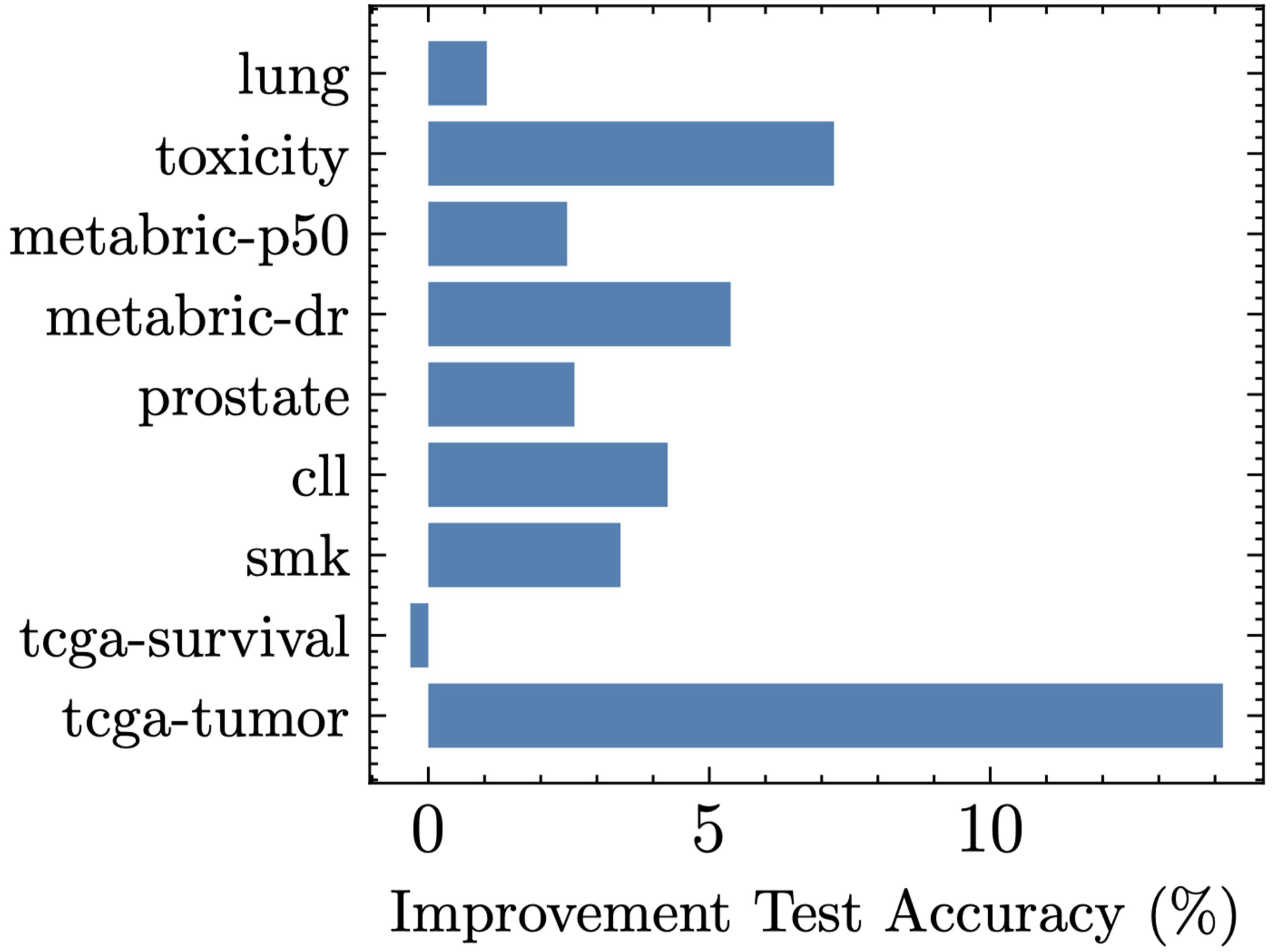
Dataset	lung	toxicity	metabric-p50	metabric-dr	prostate	cll	smk	tcga-survival	tcga-tumor	Avg. rank
MLP	94.20 ± 4.9	93.21 ± 6.1	94.31 ± 5.3	59.56 ± 5.5	88.76 ± 5.5	78.30 ± 8.9	64.42 ± 8.4	56.28 ± 6.7	48.19 ± 7.7	6.88⋆
DietNetworks	90.43 ± 6.2	82.13 ± 7.4	95.02 ± 4.7	56.98 ± 8.7	81.71 ± 11.0	68.84 ± 9.2	62.71 ± 9.3	53.62 ± 5.4	46.69 ± 7.1	10.62⋆
FsNet	91.75 ± 3.0	60.26 ± 8.1	83.86 ± 8.1	56.92 ± 10.1	84.74 ± 9.8	66.38 ± 9.2	56.27 ± 9.2	53.83 ± 7.9	45.94 ± 9.8	11.75⋆
DNP	92.83 ± 5.6	93.50 ± 6.1	93.56 ± 5.5	55.79 ± 7.0	90.25 ± 5.9	85.12 ± 5.4	66.89 ± 7.6	58.13 ± 8.2	44.71 ± 5.9	7.38
SPINN	92.26 ± 6.6	93.50 ± 4.8	93.56 ± 5.5	56.13 ± 7.2	89.27 ± 5.9	85.34 ± 5.4	68.43 ± 7.9	57.70 ± 7.0	44.28 ± 6.8	7.19
WPFS	94.83 ± 4.2	88.29 ± 5.2	95.96 ± 4.1	59.05 ± 8.6	89.15 ± 6.7	79.14 ± 4.4	66.89 ± 6.2	59.54 ± 6.9	55.91 ± 8.5	4.00
TabNet	77.65 ± 12.9	40.06 ± 11.3	83.60 ± 11.4	49.18 ± 9.6	65.66 ± 14.7	57.81 ± 9.9	54.57 ± 8.7	51.58 ± 9.9	39.34 ± 7.9	14.88⋆
TabTransformer	94.03 ± 4.7	87.67 ± 6.1	93.82 ± 4.7	52.49 ± 9.0	85.96 ± 11.5	76.81 ± 6.8	64.00 ± 9.2	56.91 ± 5.6	40.70 ± 6.9	9.62⋆
CAE	85.00 ± 5.0	60.36 ± 11.2	95.78 ± 3.6	57.35 ± 9.3	87.60 ± 7.8	71.94 ± 13.4	59.96 ± 10.9	52.79 ± 8.3	40.69 ± 7.3	9.31⋆
LassoNet	25.11 ± 9.8	26.67 ± 8.6	48.81 ± 10.8	48.88 ± 5.7	54.78 ± 10.5	30.63 ± 8.6	51.04 ± 8.5	46.08 ± 9.2	33.49 ± 7.5	16.00⋆
ElasticNet	95.19 ± 3.7	94.32 ± 4.8	95.98 ± 2.6	58.23 ± 9.6	91.36 ± 6.1	84.35 ± 7.3	70.36 ± 8.5	55.88 ± 5.7	50.73 ± 7.9	4.06
Random Forest	91.81 ± 6.9	80.75 ± 6.7	89.11 ± 6.5	51.38 ± 3.7	90.78 ± 7.1	82.06 ± 6.5	68.16 ± 7.5	61.30 ± 6.0	50.93 ± 8.4	6.62
LightGBM	93.42 ± 5.9	82.40 ± 6.4	94.97 ± 5.1	58.23 ± 8.5	91.38 ± 5.7	85.59 ± 6.5	65.70 ± 7.4	57.08 ± 7.8	49.11 ± 10.3	5.06
GCN	93.29 ± 4.6	76.13 ± 7.0	91.12 ± 8.6	58.28 ± 7.3	82.59 ± 12.4	71.99 ± 8.3	65.62 ± 8.0	58.31 ± 5.7	51.01 ± 8.1	7.75⋆
GATv2	93.33 ± 6.2	76.65 ± 11.2	86.95 ± 8.2	54.71 ± 7.1	83.23 ± 10.5	57.74 ± 14.1	66.06 ± 8.2	53.60 ± 6.8	45.45 ± 9.3	11.25⋆
GCondNet (ours)	95.34 ± 4.4	95.25 ± 4.5	96.37 ± 3.9	59.34 ± 8.9	90.37 ± 5.5	80.69 ± 5.4	68.08 ± 7.3	56.36 ± 9.4	51.69 ± 8.8	**3.62**

UNIVERSITY OF CAMBRIDGE

Applying GCondNet to TabTransformer

- GCondNet is a *general* framework for injecting graph-regularisation into various types of neural networks beyond an MLP

- Applying GCondNet to TabTransformer leads to consistent performance improvements by up to 14%

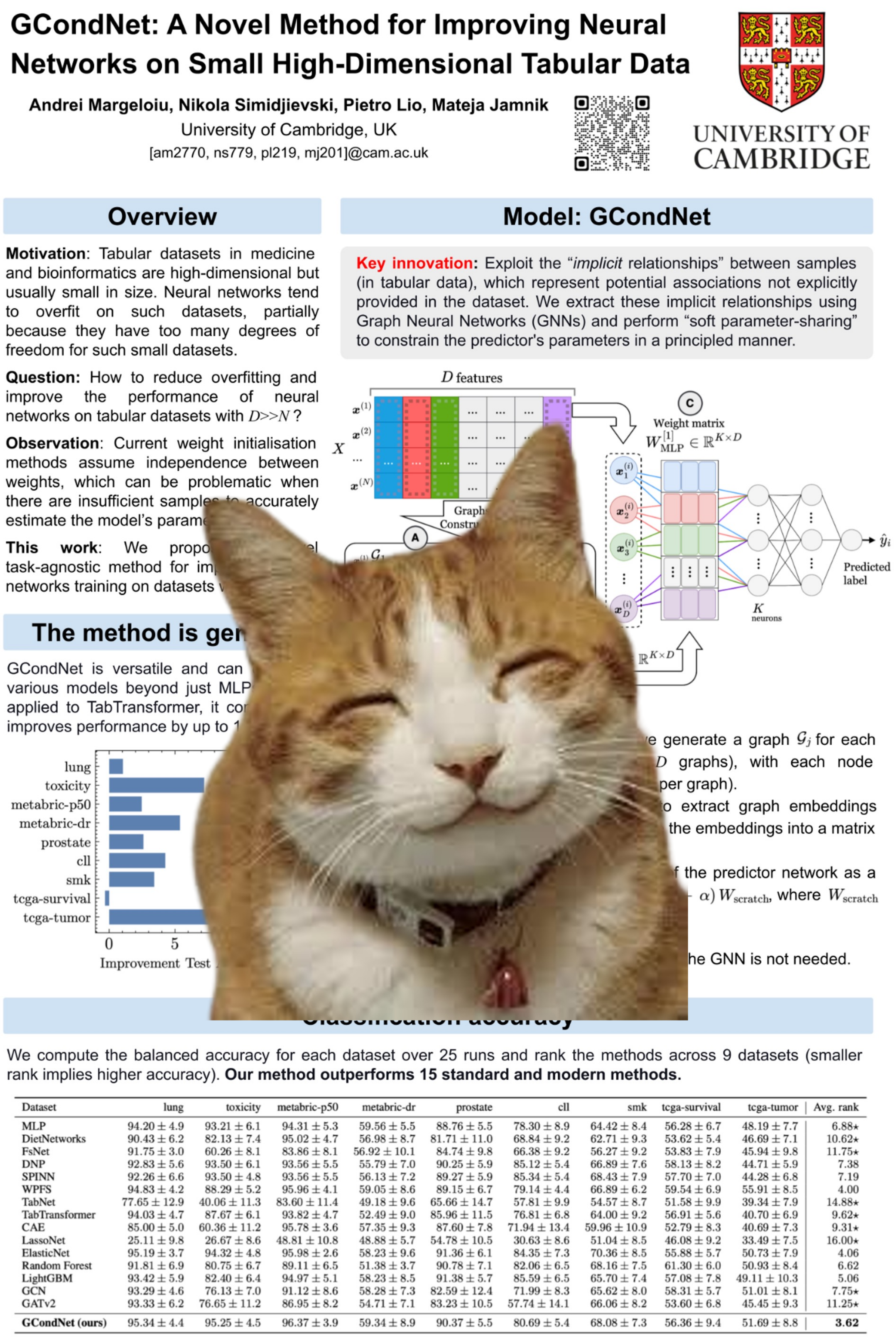


TabTransformer: Tabular data modeling using contextual embeddings. *arXiv:2012.06678* (2020)

UNIVERSITY OF CAMBRIDGE

Summary

GCondNet improves neural networks by extracting the "*implicit* relationships" between samples and performing “soft parameter-sharing” to constrain the model’s parameters.



GCondNet: A Novel Method for Improving Neural Networks on Small High-Dimensional Tabular Data

Andrei Margeloiu, Nikola Simidjievski, Pietro Lio, Mateja Jamnik

University of Cambridge, UK

[am2770, ns779, pl219, mj201]@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Overview

Motivation: Tabular datasets in medicine and bioinformatics are high-dimensional but usually small in size. Neural networks tend to overfit on such datasets, partially because they have too many degrees of freedom for such small datasets.

Question: How to reduce overfitting and improve the performance of neural networks on tabular datasets with $D>>N$?

Observation: Current weight initialisation methods assume independence between weights, which can be problematic when there are insufficient samples to accurately estimate the model’s param

This work: We propo

task-agnostic method for im

networks training on datasets

Model: GCondNet

Key innovation: Exploit the “*implicit* relationships” between samples (in tabular data), which represent potential associations not explicitly provided in the dataset. We extract these implicit relationships using Graph Neural Networks (GNNs) and perform “soft parameter-sharing” to constrain the predictor’s parameters in a principled manner.

e generate a graph $\mathcal{G}_j$ for each

D graphs), with each node

per graph).

o extract graph embeddings

the embeddings into a matrix

f the predictor network as a

$\alpha) W_{\text{scratch}}$, where W_{scratch}

he GNN is not needed.

The method is gen

GCondNet is versatile and can

various models beyond just MLP

applied to TabTransformer, it co

improves performance by up to 1

Classification accuracy

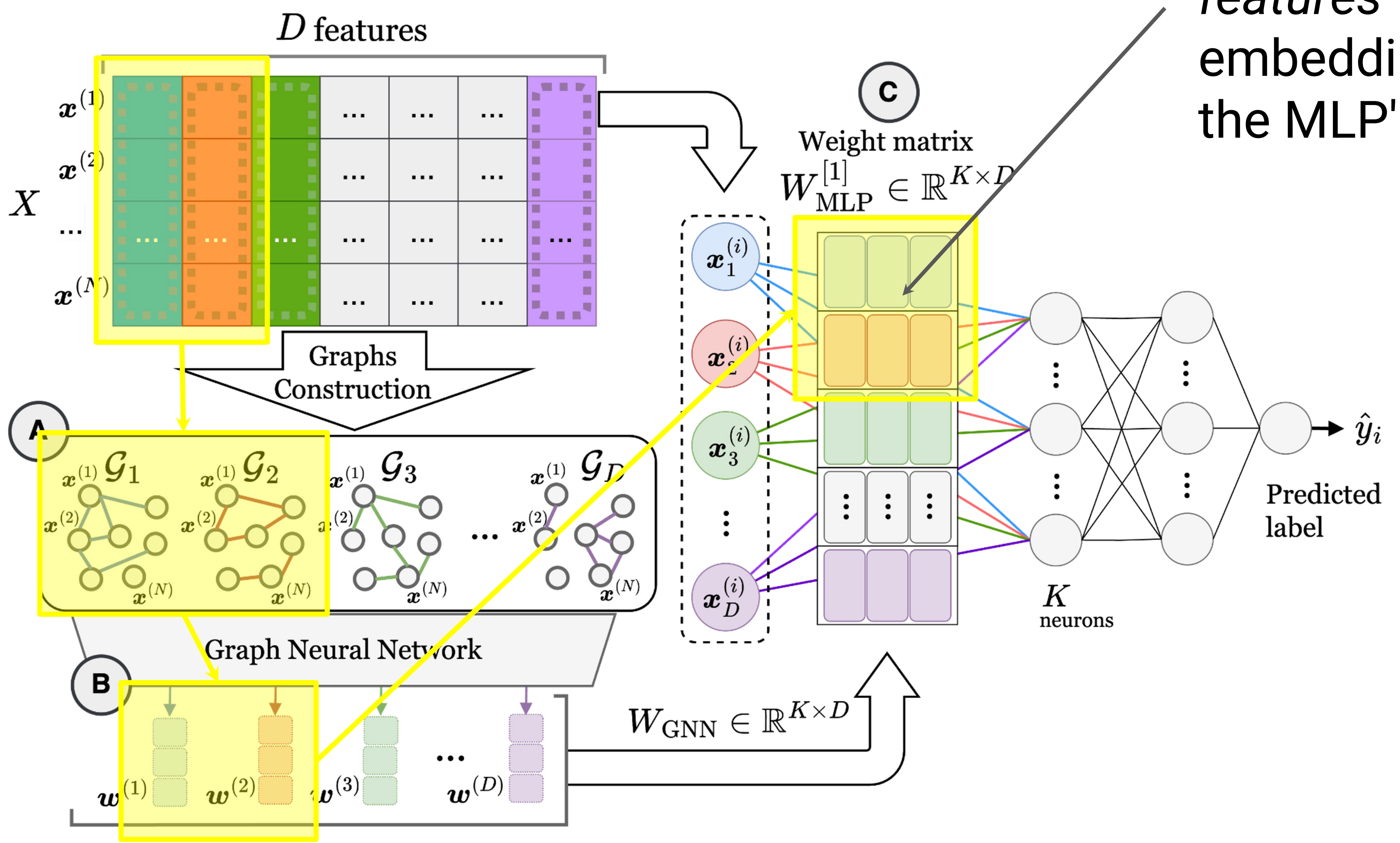
We compute the balanced accuracy for each dataset over 25 runs and rank the methods across 9 datasets (smaller rank implies higher accuracy). **Our method outperforms 15 standard and modern methods.**

Dataset	lung	toxicity	metabric-p50	metabric-dr	prostate	cll	smk	tcga-survival	tcga-tumor	Avg. rank
MLP	94.20 ± 4.9	93.21 ± 6.1	94.31 ± 5.3	59.56 ± 5.5	88.76 ± 5.5	78.30 ± 8.9	64.42 ± 8.4	56.28 ± 6.7	48.19 ± 7.7	6.88⋆
DietNetworks	90.43 ± 6.2	82.13 ± 7.4	95.02 ± 4.7	56.98 ± 8.7	81.71 ± 11.0	68.84 ± 9.2	62.71 ± 9.3	53.62 ± 5.4	46.69 ± 7.1	10.62⋆
FsNet	91.75 ± 3.0	60.26 ± 8.1	83.86 ± 8.1	56.92 ± 10.1	84.74 ± 9.8	66.38 ± 9.2	56.27 ± 9.2	53.83 ± 7.9	45.94 ± 9.8	11.75⋆
DNP	92.83 ± 5.6	93.50 ± 6.1	93.56 ± 5.5	55.79 ± 7.0	90.25 ± 5.9	85.12 ± 5.4	66.89 ± 7.6	58.13 ± 8.2	44.71 ± 5.9	7.38
SPINN	92.26 ± 6.6	93.50 ± 4.8	93.56 ± 5.5	56.13 ± 7.2	89.27 ± 5.9	85.34 ± 5.4	68.43 ± 7.9	57.70 ± 7.0	44.28 ± 6.8	7.19
WPFS	94.83 ± 4.2	88.29 ± 5.2	95.96 ± 4.1	59.05 ± 8.6	89.15 ± 6.7	79.14 ± 4.4	66.89 ± 6.2	59.54 ± 6.9	55.91 ± 8.5	4.00
TabNet	77.65 ± 12.9	40.06 ± 11.3	83.60 ± 11.4	49.18 ± 9.6	65.66 ± 14.7	57.81 ± 9.9	54.57 ± 8.7	51.58 ± 9.9	39.34 ± 7.9	14.88⋆
TabTransformer	94.03 ± 4.7	87.67 ± 6.1	93.82 ± 4.7	52.49 ± 9.0	85.96 ± 11.5	76.81 ± 6.8	64.00 ± 9.2	56.91 ± 5.6	40.70 ± 6.9	9.62⋆
CAE	85.00 ± 5.0	60.36 ± 11.2	95.78 ± 3.6	57.35 ± 9.3	87.60 ± 7.8	71.94 ± 13.4	59.96 ± 10.9	52.79 ± 8.3	40.69 ± 7.3	9.31⋆
LassoNet	25.11 ± 9.8	26.67 ± 8.6	48.81 ± 10.8	48.88 ± 5.7	54.78 ± 10.5	30.63 ± 8.6	51.04 ± 8.5	46.08 ± 9.2	33.49 ± 7.5	16.00⋆
ElasticNet	95.19 ± 3.7	94.32 ± 4.8	95.98 ± 2.6	58.23 ± 9.6	91.36 ± 6.1	84.35 ± 7.3	70.36 ± 8.5	55.88 ± 5.7	50.73 ± 7.9	4.06
Random Forest	91.81 ± 6.9	80.75 ± 6.7	89.11 ± 6.5	51.38 ± 3.7	90.78 ± 7.1	82.06 ± 6.5	68.16 ± 7.5	61.30 ± 6.0	50.93 ± 8.4	6.62
LightGBM	93.42 ± 5.9	82.40 ± 6.4	94.97 ± 5.1	58.23 ± 8.5	91.38 ± 5.7	85.59 ± 6.5	65.70 ± 7.4	57.08 ± 7.8	49.11 ± 10.3	5.06
GCN	93.29 ± 4.6	76.13 ± 7.0	91.12 ± 8.6	58.28 ± 7.3	82.59 ± 12.4	71.99 ± 8.3	65.62 ± 8.0	58.31 ± 5.7	51.01 ± 8.1	7.75⋆
GATv2	93.33 ± 6.2	76.65 ± 11.2	86.95 ± 8.2	54.71 ± 7.1	83.23 ± 10.5	57.74 ± 14.1	66.06 ± 8.2	53.60 ± 6.8	45.45 ± 9.3	11.25⋆
GCondNet (ours)	95.34 ± 4.4	95.25 ± 4.5	96.37 ± 3.9	59.34 ± 8.9	90.37 ± 5.5	80.69 ± 5.4	68.08 ± 7.3	56.36 ± 9.4	51.69 ± 8.8	**3.62**

Scan for PDF

Feel free to reach out
am2770@cam.ac.uk

What is the inductive bias?



Inductive bias: Two *highly correlated features* will have similar graphs embeddings, leading to similar weights in the MLP's first layer $W^{[1]}_{\mathrm{MLP}}$