Imperial College London

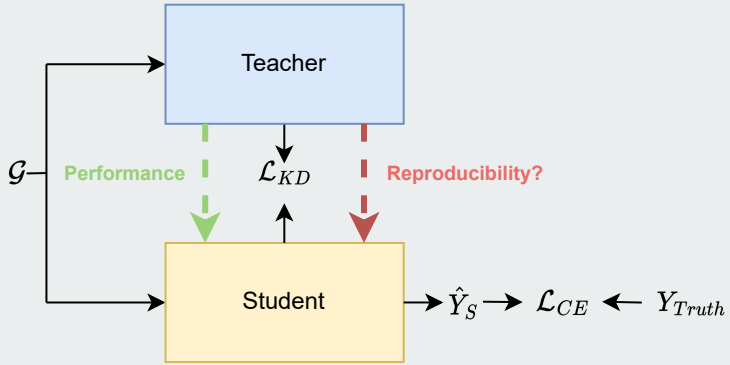
Reproducible Knowledge Distillation for Graph Neural Networks

Lorenzo G. Stigliano, September 2023

Imperial College London

Introduction

- Scalability of GNNs becomes a challenge when dealing with large graphs.
- Solution: Knowledge Distillation



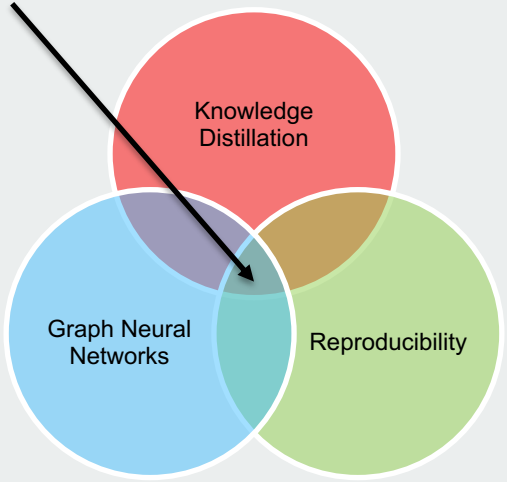
- However, there are no knowledge distillation method concerned with the reproducibility of the distilled models!

[1] Nebli A, Gharsallaoui MA, Gürler Z, Rekik I, Alzheimer's Disease Neuroimaging Initiative. Quantifying the reproducibility of graph neural networks using multigraph data representation. Neural networks. 2022 Apr 1;148:254-65.

Imperial College London

Introduction

- This project aims to bridge the three areas together, into a novel domain: **Reproducible offline Knowledge Distillation for GNNs**



Imperial College London

Contributions

1. How does one quantify the reproducibility?
2. What happens to the reproducibility of distilled models?
3. We propose a novel KD method called **Reproducibility aware Knowledge Distillation on Graphs** (RepKD) tacking the problems encountered in (2.)

Imperial College London

Quantifying reproducibility

- **Why?**

Other works [1, 2] investigate the reproducibility **between different GNN architectures** that find the same discriminate biomarkers or features; we, on the other hand, look at the **same GNN architecture**.

- **Self-reproducibility**

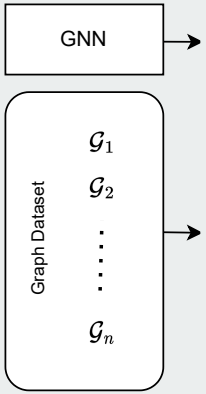
Given a GNN model trained with different data perturbation strategies (e.g., cross-validation), self-reproducibility quantifies the ability of the model to find consistent biomarkers or features across these perturbations.

[1] Nebli A, Gharsallaoui MA, Gürler Z, Rekik I, Alzheimer's Disease Neuroimaging Initiative. Quantifying the reproducibility of graph neural networks using multigraph data representation. Neural networks. 2022 Apr 1;148:254-65.
[2] Balık MY, Rekik A, Rekik I. Investigating the Predictive Reproducibility of Federated Graph Neural Networks Using Medical Datasets. InInternational Workshop on PRedictive Intelligence In MEdicine 2022 Sep 16 (pp. 160-171). Cham: Springer Nature Switzerland.

Imperial College London

Quantifying reproducibility

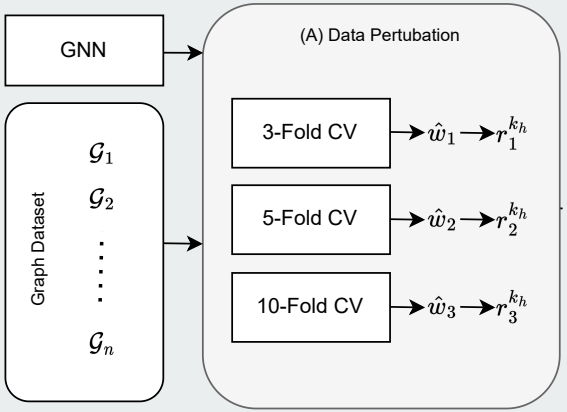
- **How?**



Imperial College London

Quantifying reproducibility

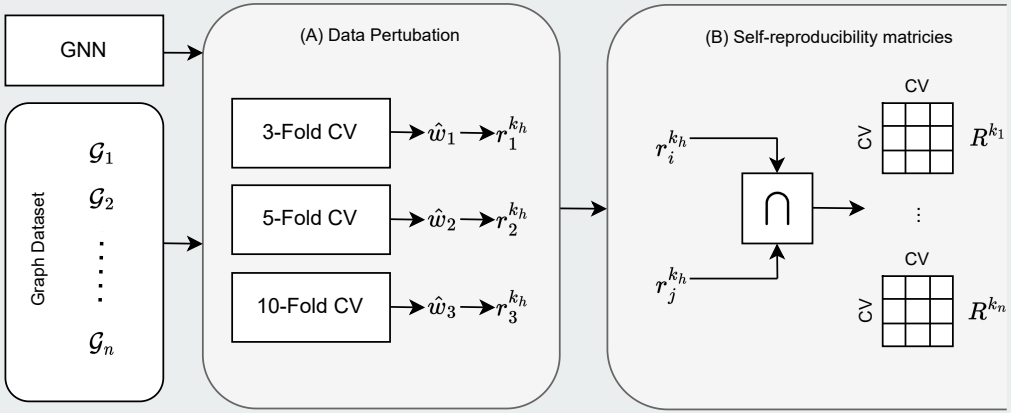
- **How?**



Imperial College London

Quantifying reproducibility

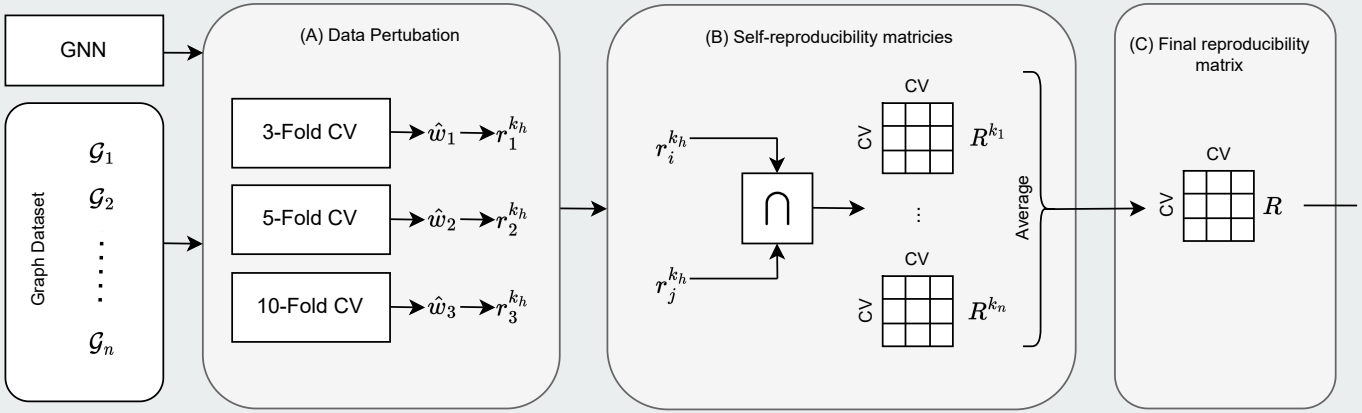
- **How?**



Imperial College London

Quantifying reproducibility

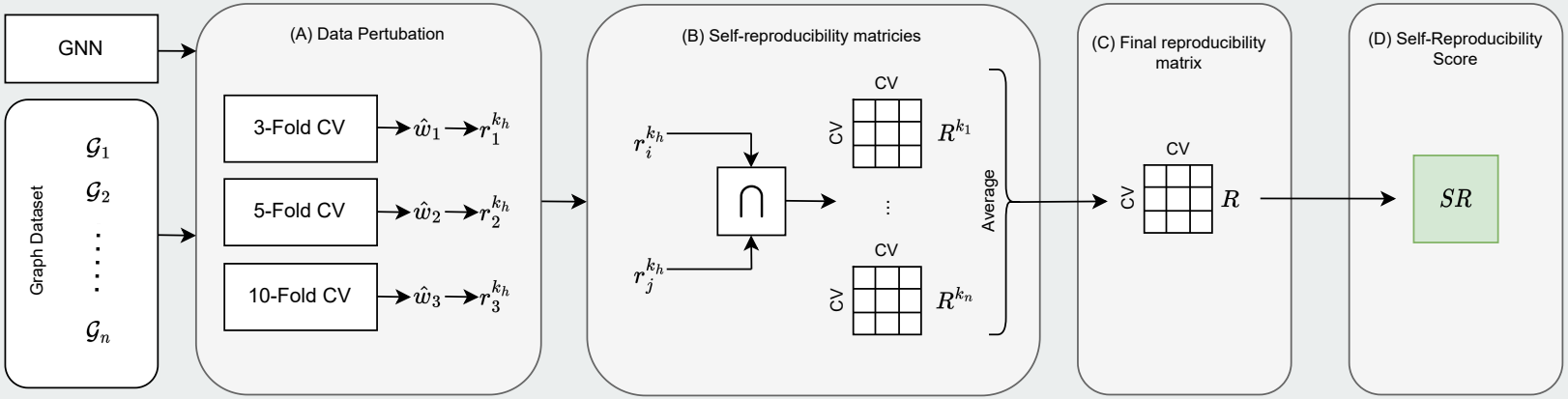
- **How?**



Imperial College London

Quantifying reproducibility

- **How?**



Imperial College London

Experimental Setup

- **Task**: Graph Classification
- **Datasets**:
  - **Brain Genomics Superstruct Project** (GSP) dataset [3]
    - Derived from structural and functional MRI scans, which are used to create cortical morphological networks (CMNs):
      - maximum principal curvature C_1, cortical thickness network C_2, sulcal depth network C_3 and average curvature network C_4.
    - 698 graphs, nodes per graph is 35, two classes: male and female.
  - **BreastMNIST** dataset [4]
    - 780 ultrasound images, nodes per graph is 28, for breast cancer tumour classification (two classes).
    - We simply use the images as graph [2]

[2] Balık MY, Rekik A, Rekik I. Investigating the Predictive Reproducibility of Federated Graph Neural Networks Using Medical Datasets. InInternational Workshop on PRedictive Intelligence In MEdicine 2022 Sep 16 (pp. 160-171). Cham: Springer Nature Switzerland.
[3] Holmes AJ, Hollinshead MO, O'keefe TM, Petrov VI, Fariello GR, Wald LL, Fischl B, Rosen BR, Mair RW, Roffman JL, Smoller JW. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. Scientific data. 2015 Jul 7;2(1):1-6.
[4] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, Pfister H, Ni B. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data. 2023 Jan 19;10(1):41.

Imperial College London

Experimental Setup

- **Knowledge Distillation Methods**: $L = \alpha \cdot L_{CE} + \beta \cdot L_{KD}$
  1. **Vanilla KD** [5] – knowledge through logits
  2. **FitNet** [6] – knowledge through logits and feature maps
  3. **Local Structure Preserving** (LSP) [7] – knowledge through local graph structure
  4. **Multi-scale knowledge distillation** (MSKD) [8] – knowledge using multiple teachers

[5] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. 2015 Mar 9.
[6] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550. 2014 Dec 19.
[7] Yang Y, Qiu J, Song M, Tao D, Wang X. Distilling knowledge from graph convolutional networks. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 7074-7083).
[8] Zhang C, Liu J, Dang K, Zhang W. Multi-scale distillation from multiple graph neural networks. InProceedings of the AAAI Conference on Artificial Intelligence 2022 Jun 28 (Vol. 36, No. 4, pp. 4337-4344).

Imperial College London

Experimental Setup

- **Teacher and Student Backbones**:
  - Graph Convolution Networks (GCN) [9]
  - Graph Attention Networks (GAT) [10]
  - Teacher: GCN and GAT both have 2-layers
  - Student: GCN and GAT both have 1-layer
- **Training & Parameter Settings**:
  - We ran all experiments across 10 different seeds
  - Self-reproducibility score evaluated for 4 different threshold values K={5,10,15,20}

Hyperparameter	Value
Training Epochs	50
Optmiser	Adam
Optmiser Weight Decay	5x10-4
Batch Size	1
Graph Thresholding	Median
Soft Targets Temperature τ (if used)	3
KD α* (if used)	1

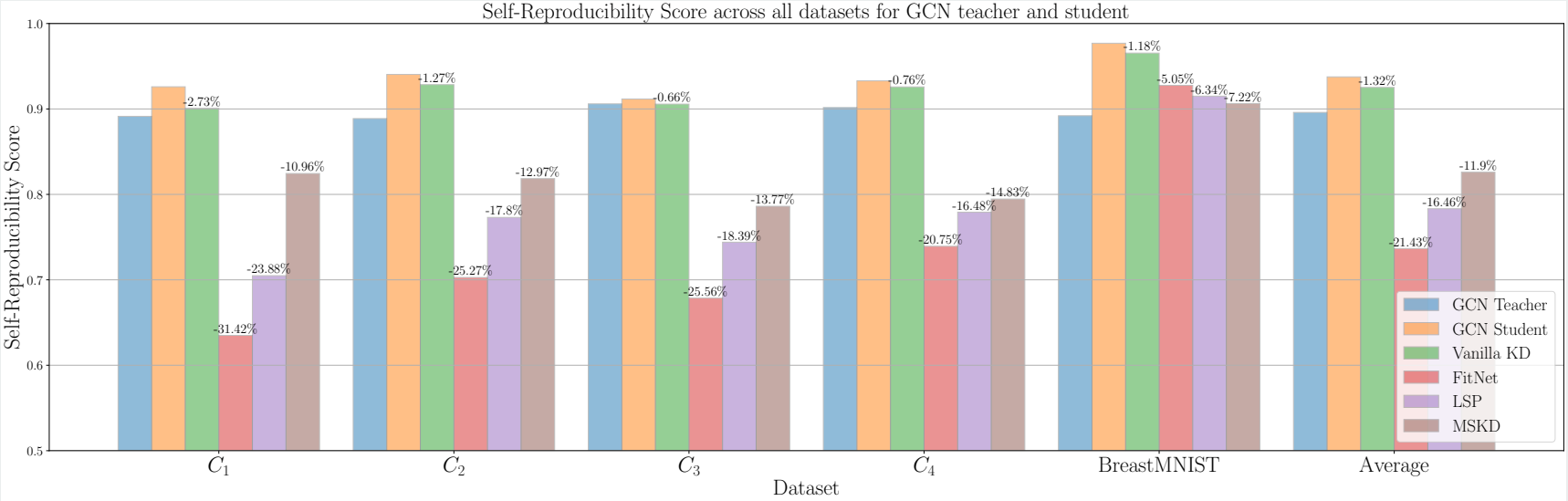
Table 6.1: Table of shared hyperparameters.

[9] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016 Sep 9.
[10] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:1710.10903. 2017 Oct 30.

Imperial College London

Motivation

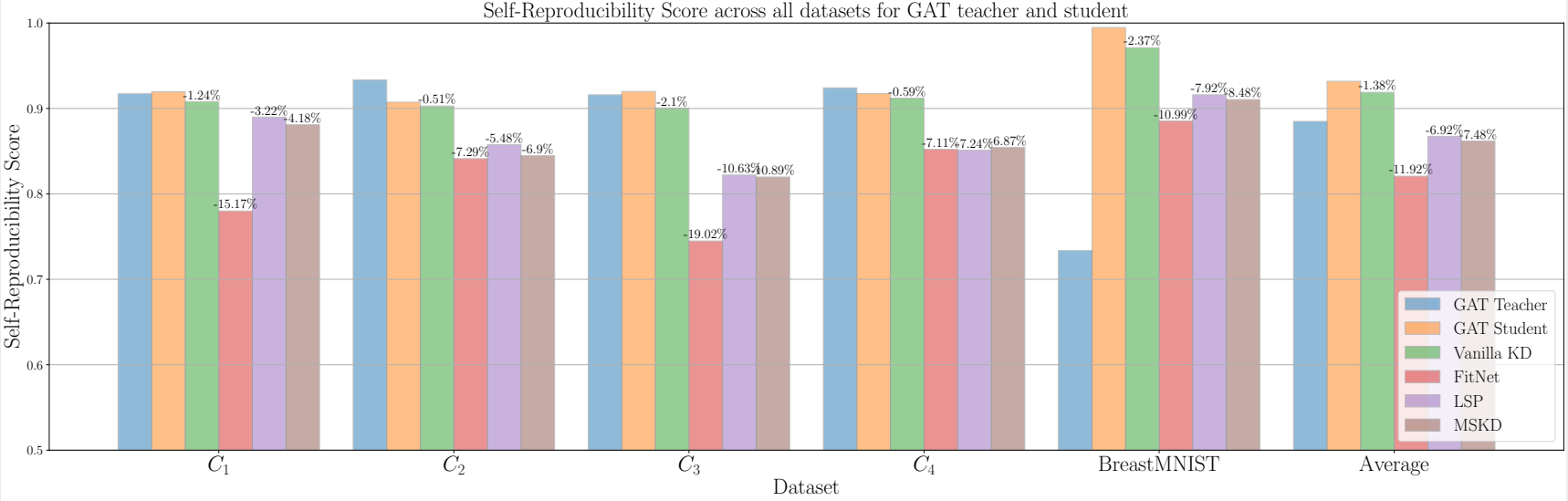
- *What happens to the reproducibility of distilled models?*



Imperial College London

Motivation

- *What happens to the reproducibility of distilled models?*



Imperial College London

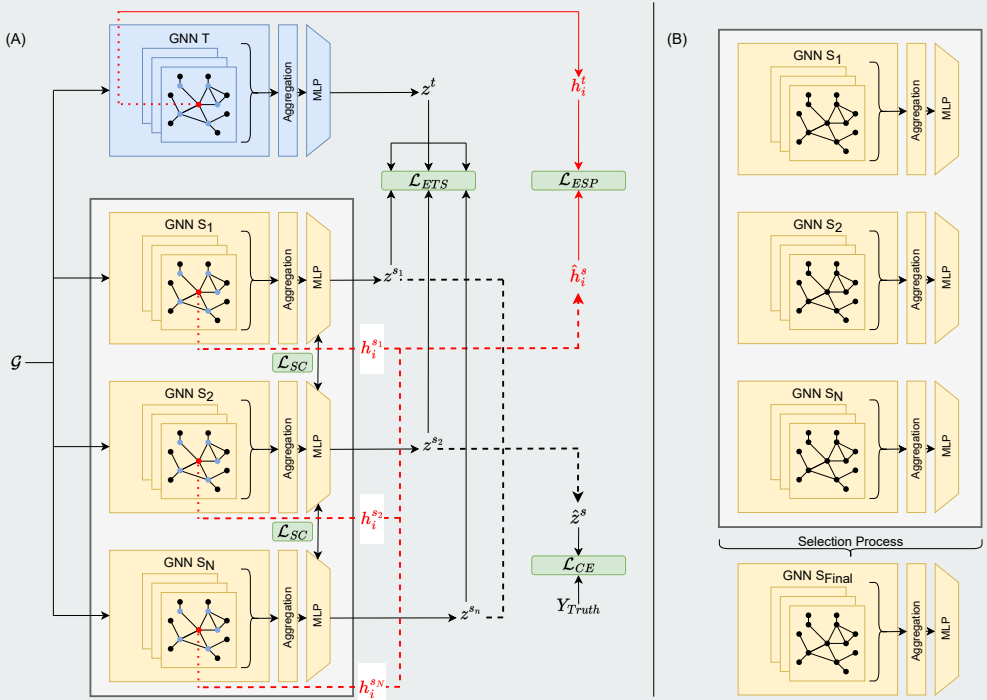
Method

- Goal: **retain and increase reproducibility** while **conserving** the performance of distilled models
- Reproducibility aware Knowledge Distillation on Graphs (RepKD)
- Method is a two-step process which consists of
  - (A) reproducibility aware knowledge distillation process – make use of a one-to-many teacher-student framework
  - (B) student selection process, which decides the final GNN student used during inference

Imperial College London

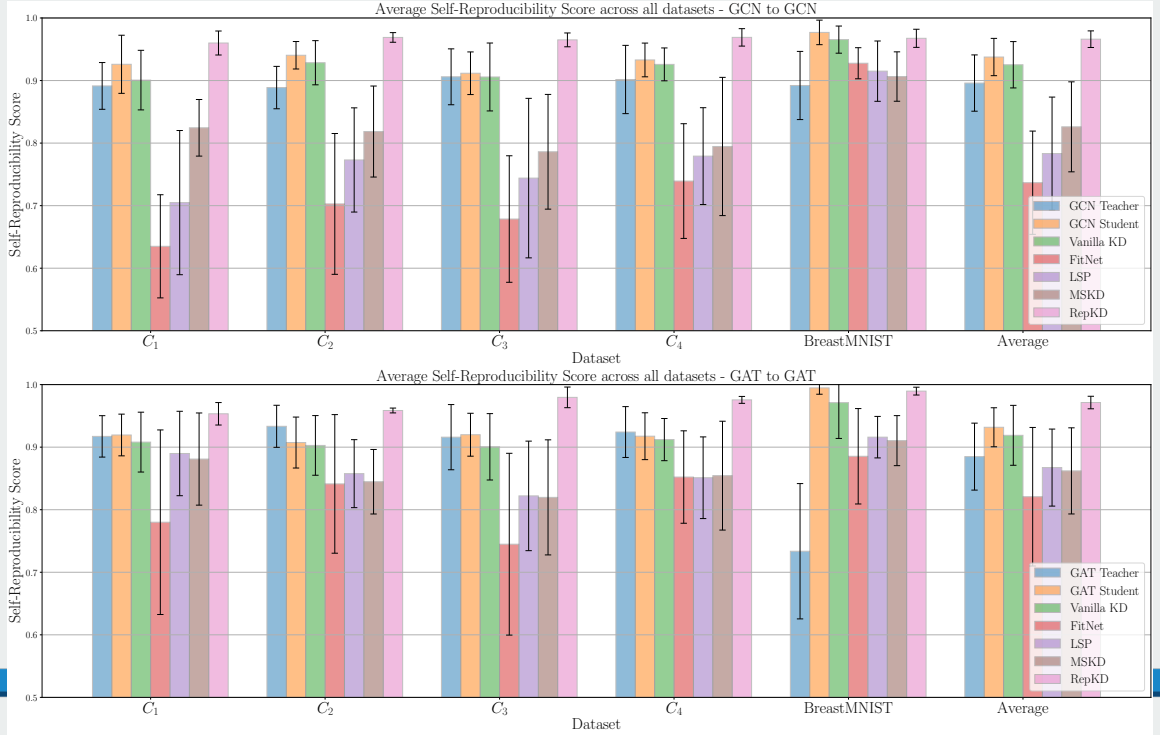
Method

- L_{CE} - Cross entropy loss
- L_{ESP} - Ensemble structure preserving loss
  - $L_{ESP} = \frac{1}{n}\sum_{i=1}^{n} D_{KL}(LS_i^t, \widehat{LS}_i^s)$
- L_{ETS} - Ensemble teacher-student loss
  - $L_{ETS} = \tau^2 \sum_{i=1}^{N} D_{KL}(\sigma_\tau(z^t, \tau), \sigma_\tau(z^{s_i}, \tau))$
- L_{IS} - Intra-student loss
  - $L_{IS} = \sum_{k=1}^{|B|} L_{SC}(B_k)$
  - B is the unique set of all pairs of student weights $B = \{(w^{s_i}, w^{s_j}) \in \binom{N}{2}\}$
- $L_{Final} = \alpha \cdot L_{CE} + \beta \cdot L_{ESP} + \gamma \cdot L_{ETS} + \lambda \cdot L_{IS}$



Imperial College London

Results – Intra-Model Performance



Imperial College London

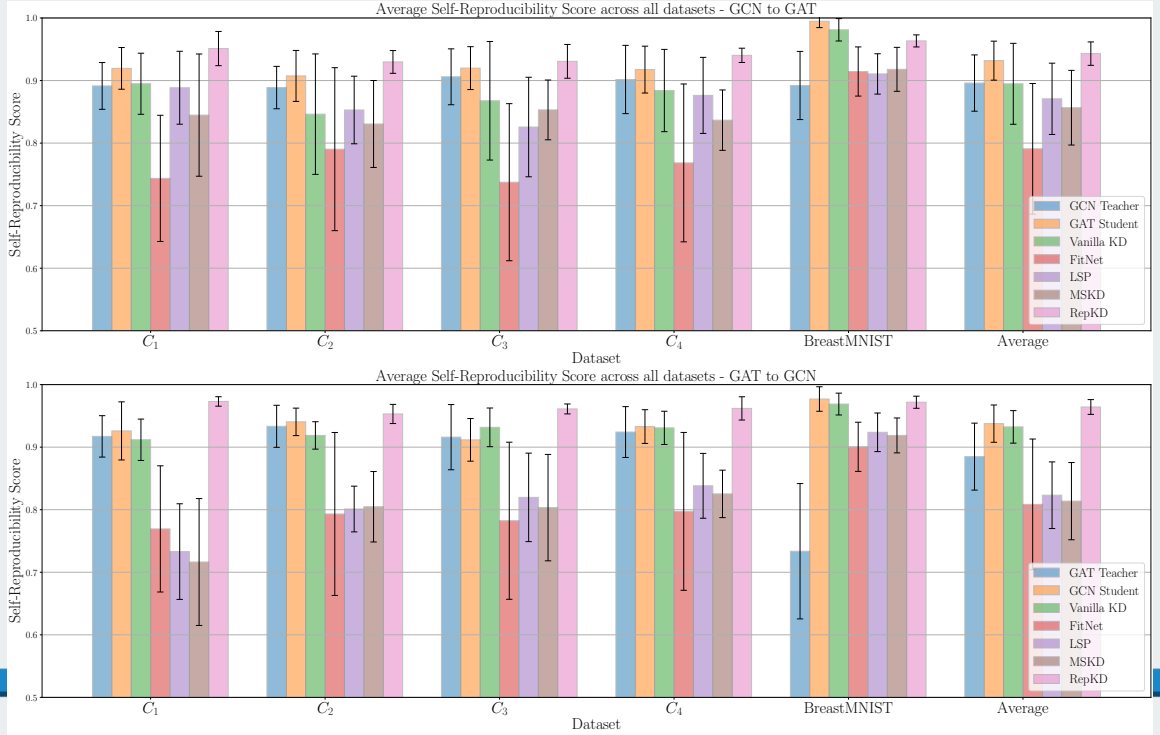
Results – Intra-Model Performance

Model	C_1	C_2	C_3	C_4	BreastMNIST	Average
GCN T.	61.01±0.30	66.16±0.09	66.91±0.09	65.44±0.12	72.48±0.31	66.40±0.18
GCN S.	57.78±0.28	63.75±0.20	63.25±0.11	63.62±0.09	69.21±0.66	63.52±0.27
Vanilla	58.16±0.24	63.98±0.14	63.75±0.15	63.77±0.23	70.34±0.60	64.00±0.27
FitNet	60.73±0.57	64.21±0.39	65.20±0.31	64.45±0.20	73.32±0.08	65.58±0.31
LSP	61.21±0.27	65.58±0.31	66.43±0.08	65.78±0.13	73.41±0.17	66.48±0.19
MSKD	**62.18±0.14**	**65.72±0.38**	**66.51±0.26**	**65.79±0.09**	**73.99±0.04**	**66.84±0.18**
RepKD	60.13±0.18	61.95±0.05	64.02±0.06	63.19±0.20	72.68±0.13	64.39±0.13
GAT T.	62.03±0.31	67.77±0.09	65.29±0.23	66.40±0.55	72.92±0.09	66.88±0.26
GAT S.	57.52±0.49	61.17±0.07	60.29±0.19	59.96±0.08	55.59±0.08	58.91±0.18
Vanilla	58.05±0.38	63.12±0.19	61.13±0.03	60.65±0.15	56.91±0.08	59.97±0.16
FitNet	62.52±0.16	66.37±0.26	65.85±0.26	66.46±0.23	73.54±0.16	66.95±0.21
LSP	63.28±0.20	66.40±0.35	66.67±0.22	66.83±0.05	73.42±0.07	67.32±0.18
MSKD	**63.45±0.35**	**66.92±0.51**	**66.93±0.11**	**66.48±0.13**	73.60±0.18	**67.47±0.25**
RepKD	57.69±0.30	64.69±0.15	56.03±0.05	56.07±0.05	**74.10±0.04**	61.72±0.22

Table 6.3: Intra-model average test-set accuracy across 3, 5 and 10-fold cross validation. Top GCN to GCN and bottom GAT to GAT. **Bold** and Underline indicate the best and second best performance among the KD methods.

Imperial College London

Results – Cross-Model Performance



Imperial College London

Results – Cross-Model Performance

Model	C_1	C_2	C_3	C_4	BreastMNIST	Average
GCN T.	61.01±0.30	66.16±0.09	66.91±0.09	65.44±0.12	72.48±0.31	66.40±0.18
GAT S.	57.52±0.48	61.17±0.07	60.29±0.19	59.96±0.08	55.59±0.08	58.91±0.18
Vanilla	57.77±0.41	62.50±0.08	61.32±0.19	61.28±0.20	56.79±0.10	59.93±0.20
FitNet	63.47±0.20	66.87±0.04	66.64±0.24	66.68±0.02	73.50±0.06	67.43±0.11
LSP	63.43±0.28	66.27±0.19	66.21±0.23	66.68±0.08	73.44±0.11	67.21±0.18
MSKD	**63.80±0.22**	**67.46±0.46**	**67.41±0.13**	**66.88±0.10**	73.51±0.05	**67.81±0.19**
RepKD	56.01±0.08	64.32±0.37	56.06±0.13	56.06±0.08	**73.98±0.03**	61.28±0.14
GAT T.	62.03±0.31	67.77±0.09	65.29±0.23	66.40±0.55	72.92±0.09	66.88±0.26
GCN S.	57.78±0.28	63.75±0.20	63.25±0.11	63.62±0.09	69.21±0.66	63.52±0.27
Vanilla	58.60±0.31	64.19±0.15	63.70±0.24	63.68±0.30	70.36±0.31	64.12±0.26
FitNet	**62.97±0.12**	64.15±0.27	66.40±0.23	65.15±0.36	73.37±0.10	66.41±0.22
LSP	60.40±0.31	64.02±0.05	64.93±0.27	65.06±0.05	**73.65±0.10**	65.81±0.16
MSKD	60.79±0.24	**65.06±0.07**	**65.19±0.41**	**65.08±0.28**	73.43±0.16	**65.91±0.23**
RepKD	60.23±0.17	61.31±0.03	64.34±0.11	62.81±0.34	70.77±0.04	63.89±0.14

Table 6.5: Cross-model average test-set accuracy across 3, 5 and 10-fold cross validation. Top GCN to GAT and bottom GAT to GCN. **Bold** and Underline indicate the best and second best performance among the KD methods.

Imperial College London

Results - Parameter Efficiency

Dataset	GNN	# Parameters			Inference Time (sec)		
		Teacher	Student	Δ	Teacher	Student	Δ
GSP	GCN	2470	108	-95.63%	3.34x10-4	2.83x10-4	-15.27%
	GAT	2536	110	-95.66%	5.14x10-3	7.36x10-4	-91.52%
BreastMNIST	GCN	2470	108	-95.63%	3.46x10-3	2.76x10-4	-20.23%
	GAT	2536	110	-95.66%	5.08x10-3	7.00x10-4	-86.22%

Table 6.9: Inference performance for Teacher and Student Models. Δ indicates the percentage change from the Teacher to Student. For number of parameters and inference time, a smaller value is better. Average inference time over 100 samples.

Imperial College London

Future work

- Exploring more complex datasets
- Exploring more GNN architectures
- Beyond classification
- Limitations of the self-reproducibility score
- RepKD - Further enhance performance (e.g, accuracy)
- RepKD - Decrease training times

Imperial College London

Demo

- https://github.com/LorenzoStigliano/thesis-imperial

Imperial College London

References

[1] Nebli A, Gharsallaoui MA, Gürler Z, Rekik I, Alzheimer's Disease Neuroimaging Initiative. Quantifying the reproducibility of graph neural networks using multigraph data representation. Neural networks. 2022 Apr 1;148:254-65.

[2] Balık MY, Rekik A, Rekik I. Investigating the Predictive Reproducibility of Federated Graph Neural Networks Using Medical Datasets. InInternational Workshop on PRedictive Intelligence In MEdicine 2022 Sep 16 (pp. 160-171). Cham: Springer Nature Switzerland.

[3] Holmes AJ, Hollinshead MO, O'keefe TM, Petrov VI, Fariello GR, Wald LL, Fischl B, Rosen BR, Mair RW, Roffman JL, Smoller JW. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. Scientific data. 2015 Jul 7;2(1):1-6.

[4] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, Pfister H, Ni B. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data. 2023 Jan 19;10(1):41.

[5] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. 2015 Mar 9.

[6] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550. 2014 Dec 19.

[7] Yang Y, Qiu J, Song M, Tao D, Wang X. Distilling knowledge from graph convolutional networks. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 7074-7083).

[8] Zhang C, Liu J, Dang K, Zhang W. Multi-scale distillation from multiple graph neural networks. InProceedings of the AAAI Conference on Artificial Intelligence 2022 Jun 28 (Vol. 36, No. 4, pp. 4337-4344).

[9] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016 Sep 9.

[10] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:1710.10903. 2017 Oct 30.