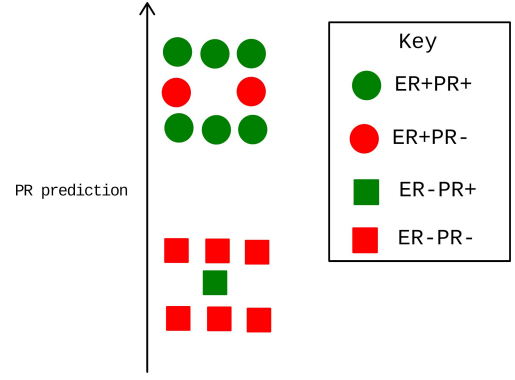
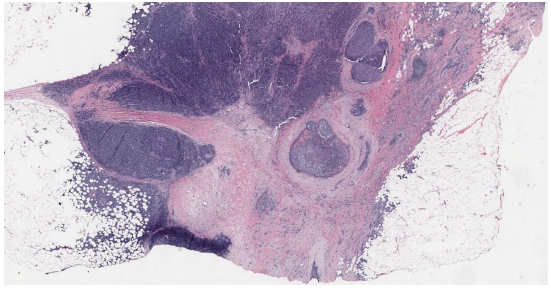
Investigating the power of deep learning for predicting breast cancer from whole slide images

Rory Sharp
CS310: Third Year Computer Science Project
University of Warwick
Supervised by:
Professor Fayyaz Minhas



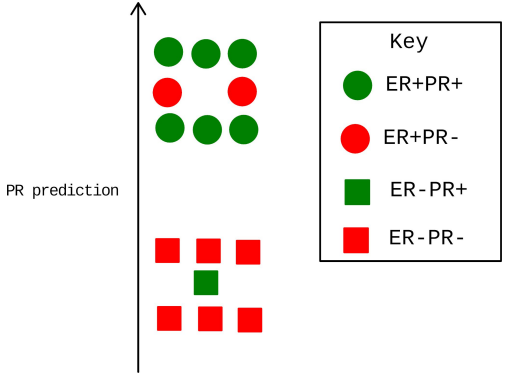


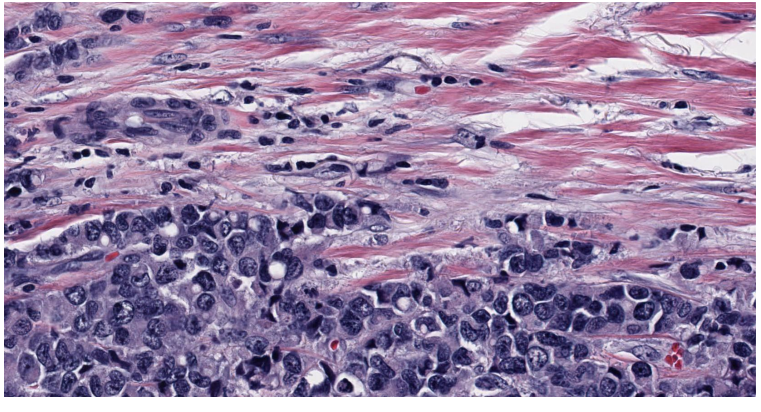
Resource acknowledgements

- “Tissues and samples were received from the Australian Breast Cancer Tissue Bank which is generously supported by the National Health and Medical Research Council of Australia, The Cancer Institute NSW and the National Breast Cancer Foundation. The tissues and samples are made available to researchers on a non-exclusive basis”
- “The results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/”
- “The authors acknowledge the use of the Batch Compute System in the Department of Computer Science at the University of Warwick, and associated support services, in the completion of this work”

Project overview

- Problem statement: Use a graph neural network to predict estrogen and progesterone hormone receptor (ER and PR) statuses of breast cancers from H&E (hematoxylin (stains nuclei blue) and eosin (stains proteins pink and stains red blood cells red)) stained whole slide images (WSIs), measuring performance using area under the ROC curve
- Questions we aim to answer:
  - How much better actually are more complex models than simpler ones for this task?
  - Do H&E stained slides even contain enough data to predict these two attributes independently?





Motivation: Biology

- Cancer is the second leading cause of death globally and the leading cause of death in the UK
- Globally, breast cancer is the most commonly diagnosed cancer in women and the most deadly cancer in women
    - Lots of data to feed the hunger of machine learning, and a high impact area to work in
- In the UK, breast cancer is the most commonly diagnosed cancer in women and the second most deadly cancer in women (after lung)
- Many breast cancers have hormone receptors, the recurrence risk of these is best managed using drugs that target the oestrogen-signalling pathway the cancer was using to grow — in some cases, this allows the patient to be completely spared the trauma of traditional chemotherapy
- H&E stain is cheap and is used by pathologists for a wide range of purposes as it highlights histological features (such as the size and shape of the nuclei, and the proportion of cells dividing)
    - In current clinical practice, receptor statuses are determined from biopsy slides stained with a single purpose IHC stain i.e. is not a task that can be done from the multi-purpose H&E slides

Motivation: Computational pathology

- Predicting ER, PR statuses from H&E is a common objective in computational pathology research — we believed it could be worthwhile to go back and see what the impact of stripping some of the complexity out of the models in the literature would be
- Although deep learning models work well on average for this task, their performance varies a lot between certain subgroups of patients — this is a major ethical barrier to the real-world deployment of such AI systems, it is very worthwhile to investigate how/whether this can be mitigated
    - Moreover, determining whether a patient is in a subgroup for which current computational approaches are effective requires determining other properties that also require specialist lab testing, thus defeating the purpose
    - More precisely, AUC_ROC is a widely used metric in computational pathology but confounding factors can skew it and this has not always been considered in historic research

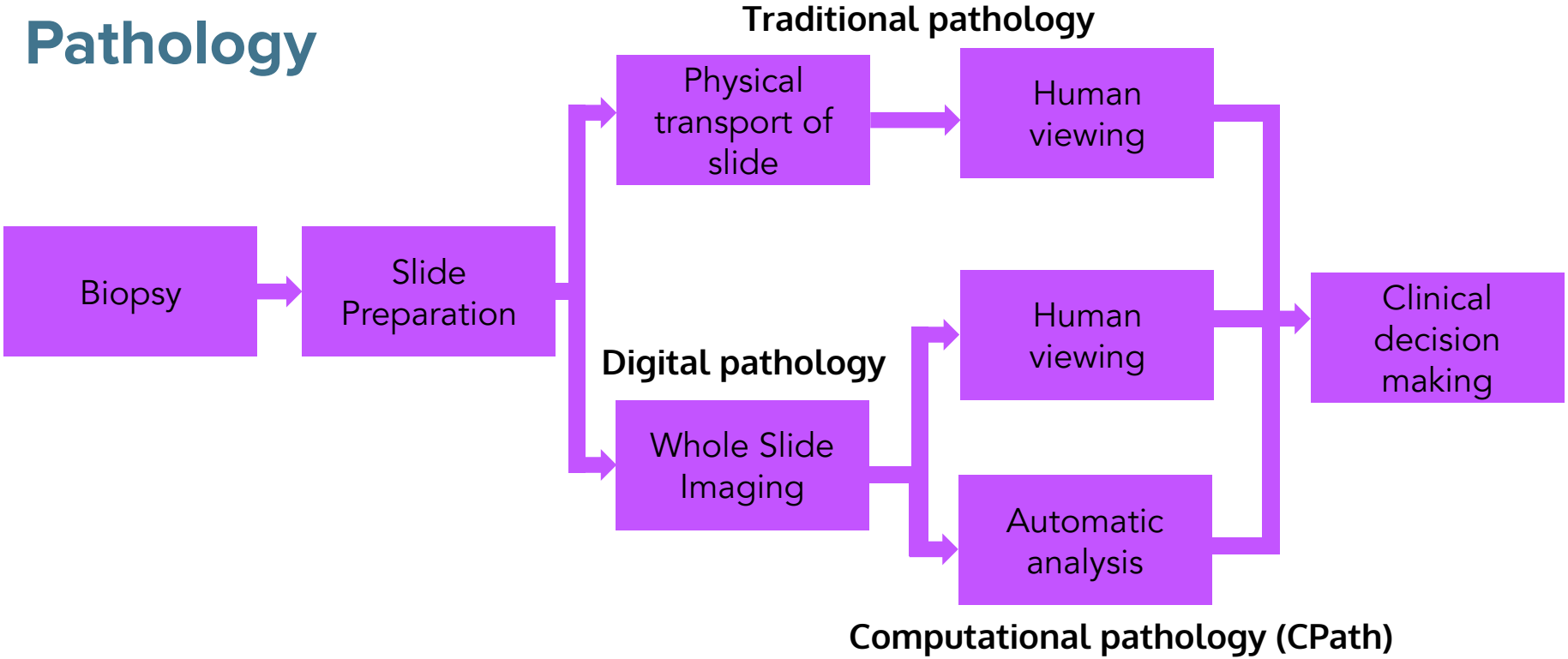
Project objectives

The initial scope of the project was to answer the following questions:
1. How important is the graph structure (i.e. message passing) for the H&E ER/PR prediction problem?
2. Would introducing attention improve performance for this problem?

At the end of term 1 the scope was expanded to add the following questions:
3. Do models that perform better overall also perform better in the hard subset? (i.e. what is the relationship between (complexity of) architecture and susceptibility to confounding?)
4. How sensitive are each of our models to their hyperparameters?

At the end of term 2 the scope was expanded to add the following questions:
5. What changes to training have reduced confounding in prior work (in other domains)?
6. How effective are these changes in this domain?

Pathology



Based on a figure in [1]

[1] F. Minhas, 'Machine Learning for Cancer Diagnostics, Prognostics and Prediction', Accessed: Mar. 08, 2025. [Online]. Available: https://warwick.ac.uk/fac/sci/dcs/people/fayyaz_minhas/bs347oncology/bs347_oncology_ml_in_cancer_fm.pdf

Hormone-sensitive breast cancer

- The presence of estrogen receptor (ER) and progesterone receptor (PR) in a breast tumour are indicators of better base prognosis and of the suitability of targeted therapies
- Immunohistochemical (IHC) testing attaches dye to antibodies which selectively bind to features of interest (ER and PR)
- Unlike H&E, IHC is sensitive to processing (e.g. formalin fixation time)
    - Tissue death of receptors (such as due to inadequate fixation) can lead to false-negatives
        - Nearly 40% of the over 2000 cases which Canadian provincial laboratory for Newfoundland and Labrador found to be ER− between 1997 and 2005 were found to be ER+ when retested by other laboratories — this lead to a public inquiry and a ≈£15M (in 2024 terms) class action lawsuit settlement
- IHC stains are significantly more expensive than H&E and are single-purpose

Hormone-sensitive breast cancer: Rates

P(ER− and PR−) = 25%	P(ER−|PR−) = 68%
P(ER− and PR+) = 01.4%	P(ER+|PR+) = 98%
P(ER+ and PR−) = 12%	P(PR−|ER−) = 95%
P(ER+ and PR+) = 62%	P(PR+|ER+) = 84%

Joint distribution of labels in TCGA dataset, and implied conditional distribution

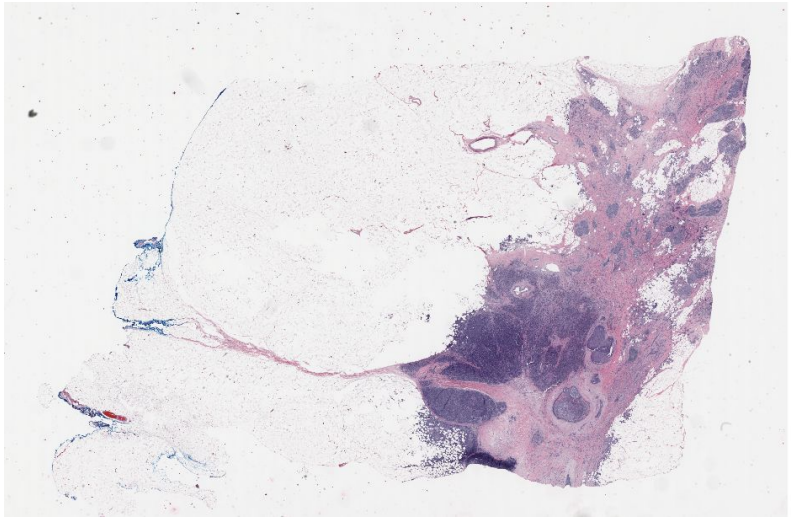
- There is class imbalance: P(ER+) = 74% and P(PR+) = 63%
- There are correlations: P(PR+) = 63% but P(PR+|ER+) = 84%
- There are near-deterministic causal relationships: P(PR−|ER−) "≈ P(ER+|PR+) "≈ 100%
  - Biologically, estrogen signalling is believed to be a necessary condition for PR expression, but in clinical practice 1–3% of patients test as ER−PR+
    - It has been proposed that all ER−PR+ cases may be test artefacts

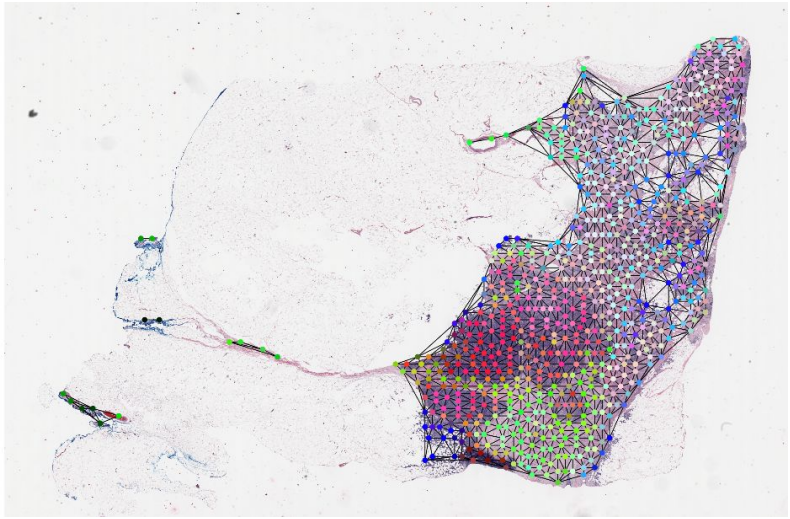
WSI Data modelling for CPath

- Due to their very high resolution, it is necessary to break a WSI up into patches for processing then aggregate the patch-level results to produce a slide-level prediction
- SlideGraph [1]: A powerful approach is to create a graph for the slide in which each node is a patch with features extracted by a patch-level model and use a graph neural network to process this graph into a graph-level prediction
- Representing the slide as a graph allows the information of the spatial distribution of the features to be used by the model
- This paradigm of producing features at the patch-level is preferable to naive aggregation of patch-level predictions as it allows the model to take a hierarchical approach: producing patch-level analyses from local context then combining these using wider context
    - This is in line with how a human analyses a slide, zooming in and remembering what they saw that was relevant then zooming back out

[1] W. Lu, S. Graham, M. Bilal, N. Rajpoot, and F. Minhas, 'Capturing Cellular Topology in Multi-Gigapixel Pathology Images', Jun. 2020, doi: 10.1109/CVPRW50498.2020.00138.

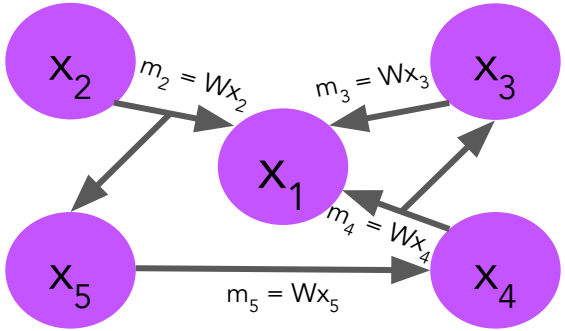
WSI Data modelling for CPath





Graph neural networks (GNNs): Message passing

- Predominant paradigm for using edges in graph neural networks is the message passing framework
- Operation of a message passing layer: Firstly, each node transforms its features into a message. Then, each node sends its message over its edges to all its neighbours. Finally, each node then updates its features by taking an aggregation over all the messages it received



- The same weights matrix W is used by all nodes
- There is not an edge directly from x5 to x1. However, if there are 2 message passing layers, then data from x5 can reach x1 via messages x5 → x4 → x1

GNNs for graph-level (i.e. slide-level) prediction

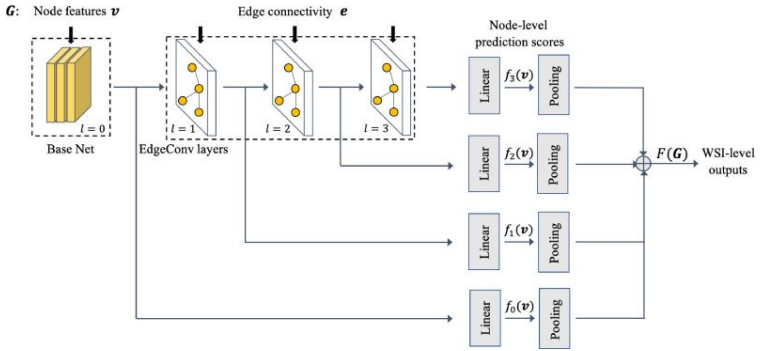
1. Graph-level layers transform the node features
2. Mean pooling produces a graph-level feature vector
3. Read-out (ordinary) layers transform the graph-level feature vector into the graph-level prediction

- Graph linear layer: $x'_i = W x_i \; \forall x_i$
- Graph convolutional layer (GCN): Message passing with $\text{agg}_i(m) = \sum_{m_j} \left(\left(\frac{1}{n} \right) m_j \right)$ where m is the messages received by node i and n is the number of messages received by node i
- Graph attention layer (GAT): Message passing with $\text{agg}_i(m) = \sum_{m_j} \left(\theta \left(m_i \# m_j \right) m_j \right)$ where θ is a small neural network and $\#$ is vector concatenation

SlideGraph+

$$\text{Recall that } \text{MessagePassing}(\phi, \text{agg})(x_i) = \text{agg}\left(\{\phi\left(x_j\right) : \ x_j \in N(x_i)\}\right).$$
$$\text{Similarly, } \text{EdgeConv}(\phi, \text{agg})(x_i) \qquad = \text{agg}\left(\{\phi\left((x_i \# \left(x_j - x_i\right)\right) : \ x_j \in N(x_i)\}\right)$$

- SlideGraph+ [1] is a state of the art model for prediction of slide-level labels from H&E WSIs
- SlideGraph+ was tested on HER2 (another IHC-based test) which is even more challenging to predict from H&E than ER/PR so a simpler model may be sufficient for ER/PR — this observation is what initially sparked this project
- The SlideGraph+ architecture is a series of EdgeConv-based blocks (using max aggregation in the EdgeConv), each of which branches off to produce a prediction which is skipped through to the end to be summed to produce the model output



Architecture of SlideGraph+ [1]

[1] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, and F. Minhas, 'SlideGraph + : Whole slide image level graphs to predict HER2 status in breast cancer', Aug. 2022, doi: 10.1016/j.media.2022.102486.

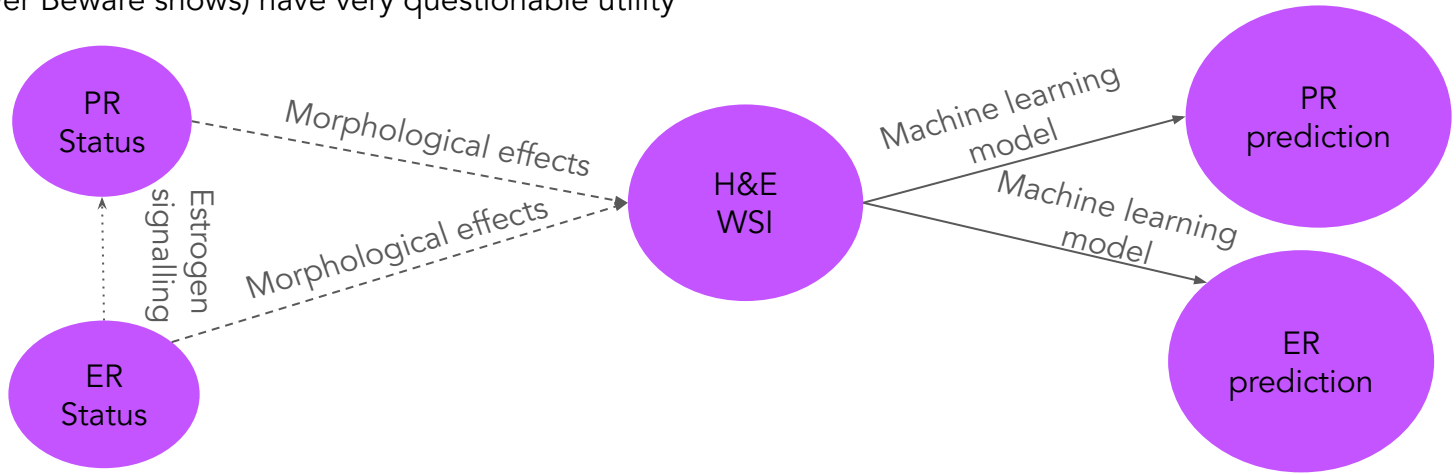
Confounding factors in CPath: Buyer Beware

- Confounding = distortion of a measure between two variables as a result of hidden variables that affect both of them
- Other than SlideGraph+, Buyer Beware [1] is the other big influence on this project: Sounded the warning bell that deep-learning-based predictors of biomarkers from WSIs heavily utilize the combined effect of associated biomarkers instead of isolating the one they are meant to be predicting
    - Our project: We know this problem that CPath models predict the combined effect of biomarkers instead of isolating the one we ask them for exists, but can it be fixed?

[1] M. Dawood, K. Branson, S. Tejpar, N. Rajpoot, and F. ul A. A. Minhas, 'Buyer Beware: confounding factors and biases abound when predicting omics-based biomarkers from histological images', Jan. 2024, doi: 10.1101/2024.06.23.600257.

Confounding factors in CPath: Our problem

- ML task: Predict the ER/PR status from a H&E stained WSI based on the visually observable effects of the not directly observable true status
  - As the PR/ER status also has relevant effects, it is challenging to separate out the effects of the individual receptors to produce independent predictions
- Effects of ER/PR in H&E may be so similar it is genuinely hard, but there is also an incentive for models to be lazy as the labels are correlated
- Medicine already has a practical method for obtaining independent ER, PR results (and thereby identifying the non-negligible proportion of ER+PR− patients), namely IHC, so confounded machine-learning based methods (like Buyer Beware shows) have very questionable utility



Simplicity bias of neural networks

- There are many combinations of weights that allow a neural network to minimize the loss function, however gradient descent has an implicit bias to learn the simplest features first, and not subsequently learn additional features that help with fewer training examples if it can already accurately classify those examples [1]
- Gradient starvation effect: The features that are most helpful on average over the training set will have the largest gradients and so will be learnt at the fastest rate. Moreover, learning these features significantly reduces the overall error and so vanishes the gradients to learn other features
- In many cases, this discouragement from learning complex features is a valuable built-in protection against overfitting
- In some cases there are simple features that we do not wish the model to learn:
    - The correlation between a simple feature and a label may be an artefact of the training data (a spurious correlation)
    - The feature may relate to a real-world property but this property may be one that external constraints require the model to be invariant towards (for example for compliance with anti-discrimination legislation)

[1] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. https://arxiv.org/abs/2006.07710, June 2020.

Feature extraction: ImageNet

- Despite the significant difference between ImageNet images and WSIs, the features from CNNs trained on ImageNet have been successfully used as patch features for slide-level classifiers in computational pathology
- This can be viewed as a form of transfer learning in which the weights of all the pre-trained blocks are frozen

Feature extraction: HoVer-Net

- HoVer-Net [1] is a CNN for computational pathology and was trained using transfer-learning fine-tuning with ImageNet weights as the starting point
- Operates over a patch to identify each nucleus (producing a segmentation mask for all nuclei) and classify its type
    - From this segmentation mask, the centre of each nucleus can be calculated alongside morphological features (such as: area, eccentricity (how much more of an oval than a circle it is)) to complement the feature of the type of the nucleus
- HoVer-Net can be used to provide patch features for slide-level classifiers by concatenating the means and standard deviations of the nuclear features of the patch
- Unlike the deep embeddings from ImageNet, HoVer-Net features are explainable and are related to pathology — this interpretability is a major advantage for the possible clinical adoption of computational pathology

[1] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis, December 2019. doi: 10.1016/j.media.2019.101563.

Foundation models

- There is experimental evidence that increasing parameter counts and training time improves generalisation if training continues for long enough [1]
  - The implicit bias towards simple features provides protection against over-fitting
  - Giving a model more parameters or more training time gives it more chances to find simpler solutions [2]
- Foundation models are very large task-agnostic models trained over very large unlabelled datasets using self-supervised learning
  - Much like traditional transfer-learning, only a small amount of labelled data is required to carry out supervised learning to fine-tune a foundation model to carry out a particular task
- The self-supervised nature of the foundation training dramatically increases the size of available datasets for the initial training (and so improves the generalisation of the base model) by removing the need for labelling
- Large Language Models (LLMs) fit into the foundation model framework: a base model is trained for the self-supervised task of predicting the next token in a very large dataset, then models for particular tasks are tuned using RLHF
  - As per simplicity bias, LLMs have very high parameter counts and have demonstrated strong generalisation
  - Also as per simplicity bias, although seemingly impressive, LLMs cannot be relied upon as they are very well known to show overconfidence when faced with situations they do not know the answers to

[1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32):15849–15854, August 2019. doi: 10.1073/pnas.1903070116.
[2] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. https://arxiv.org/abs/1803.03635, March 2018.

Vision foundation models

- Although self-supervision is most naturally suited to natural language processing, the foundation model framework has been applied to image processing to create vision foundation models
- DINO [1] (knowledge DIstillation with NO labels) and DINOv2 are self-supervision rules for image processing
- Vision foundation models are designed around vision transformers instead of CNNs
    - A vision transformer works exactly as a text transformer does except: the input is embeddings for image patches instead of word tokens, and the output is an embedding to be learnt to summarize all the patches in the sequence
- DINO applies the contrastive self-supervised learning paradigm to the knowledge distillation paradigm:
    - The vision transformer is trained as a student network which when given a cropped part of an image produces an embedding that is similar to that produced by a teacher network when given the whole image (and non-similar to those produced by the teacher network when given other images)
    - The teacher network is not directly trained but rather is an exponential moving average of the previous student networks
- DINOv2 takes the DINO term and adds on other self-supervised-learning terms

[1] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging Properties in Self-Supervised Vision Transformers. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 2021. doi:10.1109/ICCV48922.2021.00951.

Feature extraction: UNI

- UNI [1] is a recently published foundation model for computational pathology
  - A vision transformer trained over a dataset of 100,426 H&E WSIs (≈100 million patches) using the DINOv2 self-supervision
    - The number of pathologist-hours away from patients that would be required to produce patch-level annotations over a dataset of this size in order to use supervised learning would be completely unjustifiable
- High-quality task-specific models can be trained on top of UNI features using very small labelled datasets
  - However, an IHC biomarker predictor in particular from UNI had not been trained at the outset of this project, providing scope for us to make a novel contribution
    - During the project, results for this were published [2] — however, they do not critically evaluate their results in light of Buyer Beware as we do

[1] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 30(3):850–862, March 2024. doi: 10.1038/s41591-024-02857-3.

[2] Hongming Xu, Mingkang Wang, Duanbo Shi, Huamin Qin, Yunpeng Zhang, Zaiyi Liu, Anant Madabhushi, Peng Gao, Fengyu Cong, and Cheng Lu. When multiple instance learning meets foundation models: Advancing histological whole slide image analysis. Medical Image Analysis, 101, April 2025. doi: 10.1016/j.media.2025.103456.

Data: Datasets

- TCGA (from USA) and ABCTB (from Australia) are the two big datasets for CPath for breast cancer WSIs
- Due to the large computational cost of generating all the node features from enough WSIs to have enough graphs to train deep learning models, my supervisor provided pre-generated graphs:
  1. Graphs from TCGA-BRCA [1] WSIs with features from a ShuffleNet (a CNN) trained on ImageNet (1024D features) — we only had 692 of the 1098 cases of TCGA-BRCA and only the 628 that have both oestrogen receptor (ER) and progesterone receptor (PR) data were used
  2. Graphs from ABCTB [2] WSIs with features from UNI (1040D features) — 3000 graphs

[1] National Cancer Institute. GDC Projects: TCGA-BRCA. https://portal.gdc.cancer.gov/projects/TCGA-BRCA.
[2] Jane E Carpenter, Deborah Marsh, Mythily Mariasegaram, and Christine L Clarke. The Australian Breast Cancer Tissue Bank (ABCTB). Open Journal of Bioresources, July 2014. doi: 10.5334/ojb.aa.

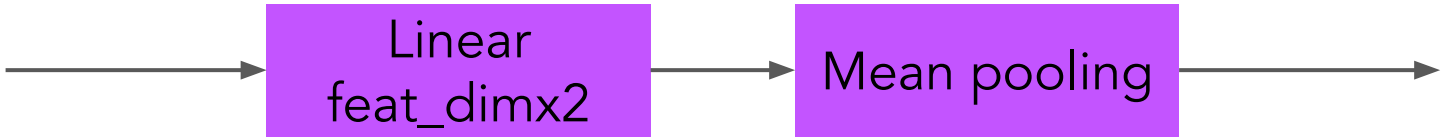
Data: Partitioning

- Data was split into a train-set for 5-fold cross-validation when experimenting and a test-set to be held out until drawing conclusions at the end of the project — as our conclusion was that no methods are effective for reducing confounding for this problem, we did not turn out to require the test set
- Prevalence of each combination of labels in each set/fold is as close as possible to the prevalence in the overall dataset
- All WSIs from the same hospital are in the same set/fold to ensure we evaluate the ability of the model to generalise to new hospitals which may pre-process slides slightly differently e.g. intensity of staining — this was only possible for TCGA as ABCTB did not have the tissue source site data

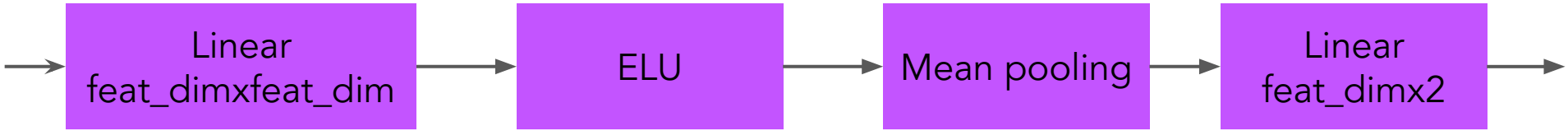
Our 5 models

- Non-deep learning baseline: Linear GNN — models logistic regression
- Non-message passing deep-learning: GNN — measures how important the graph structure is
- Message passing: GCN, and GAT — GAT is in some sense more flexible than EdgeConv (and GCN)
- SlideGraph+: EdgeConv (aggregation = max) — a current state-of-the-art model

Model architectures: Non-message passing

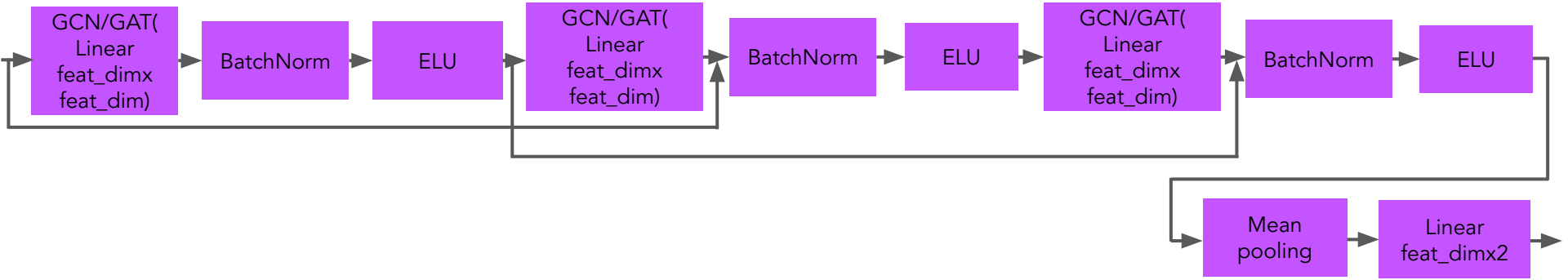


Linear GNN



GNN

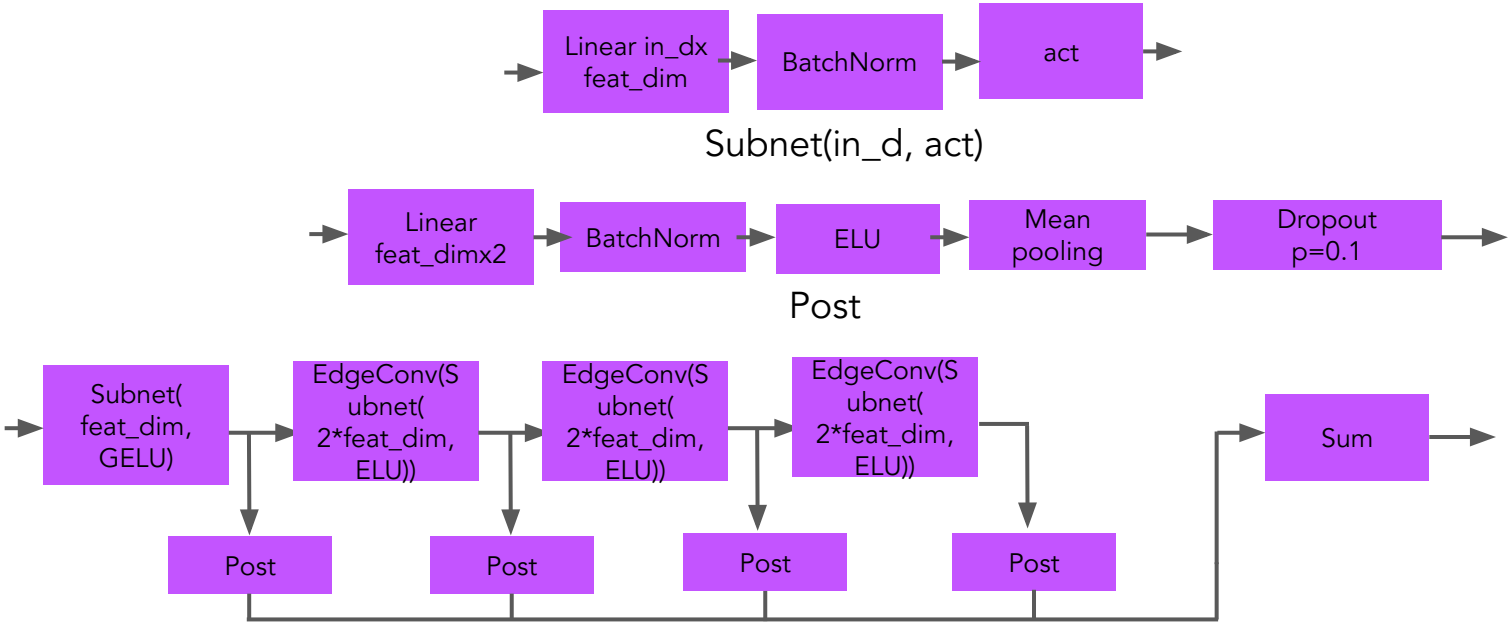
Model architectures: Node message-passing



GCN/GAT

- ResNet-like skip connections sum the input to the previous onto the output of the current message passing
- Each GAT(Linear feat_dimxfeat_dim) is the concatenation of 8 Linear feat_dimx(feat_dim/8) heads

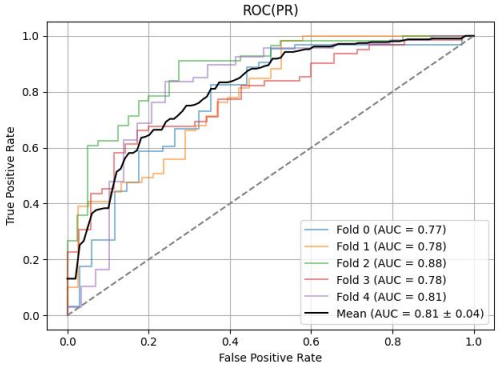
Model architectures: EdgeConv

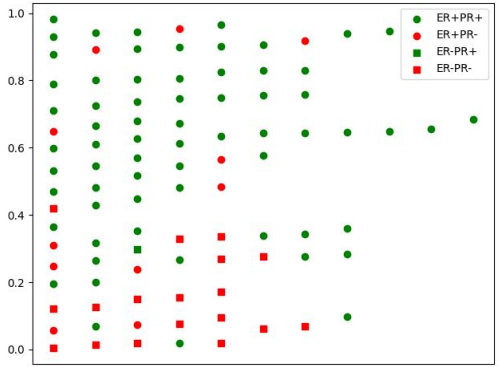


Architecture is as in Buyer Beware i.e. of SlideGraph+ form

Area under the ROC curve (AUC_ROC)

- False positive rate = (number of positive predictions)/(number of negative examples)
- True positive rate = (number of positive predictions)/(number of positive examples)
- ROC curve = false positive rate vs true positive rate for all possible decision thresholds
- Area under ROC curve = the probability that a random positive example is given a higher score by the model than a random negative example
- AUC_ROC is confoundable in multi-label settings:
  - The model on the right has decent AUC_ROC because it has good separation between red (PR−) and green (PR+)
  - However, on closer inspection the decent AUC_ROC is largely simply because most of the red data is squares (ER−) and squares tend to be ranked lower by this PR predictor than triangles (ER+)





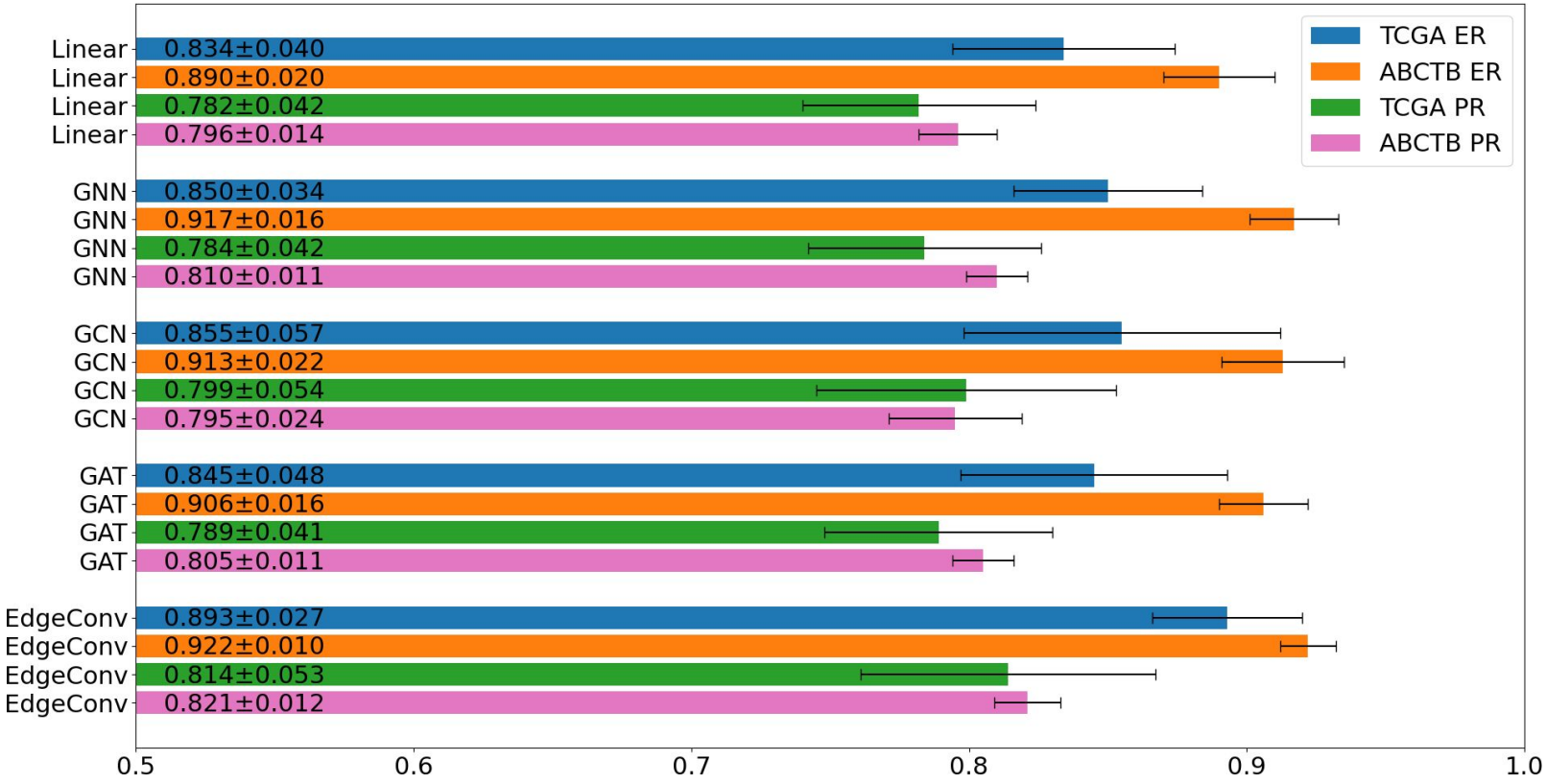
Our extensions to AUC_ROC

- The ordinary AUC_ROC implicitly values instances in proportion to the frequency of the occurrence of their other label
- By re-weighting AUC_ROC so that all instances are valued equally we obtain what we will call a balanced AUC_ROC
    - This is directly analogous to the widespread practice of re-weighting examples in a loss function to correct for a class imbalance
- We can also assign no weighting to examples that we consider to be too easy to obtain what we will conditioned AUC_ROCs
    - When classifying ER, PR+ cases are (almost) definitely ER+
    - When classifying PR, ER− cases are (almost) definitely PR−
- Conditioned AUC_ROCs have been considered by previous authors, but to the best of our knowledge the balanced AUC_ROC is a novel contribution

Training

- PyTorch Lighting was used to implement the training loop
- Loss = sum of binary cross entropy loss for ER and PR
    - Conditioned-label-based oversampling: we upweight examples of underrepresented combinations of labels to correct for class imbalance
    - ER−PR+ cases not included — it is an ongoing debate in the medical community whether there are any true ER−PR+ patients
- Optimizer = AdamW — weight decay provides regularisation
- Models were trained until 25 epochs passed with no improvement in the validation loss (early stopping), then rolled back to the epoch with the lowest validation loss (checkpointing)

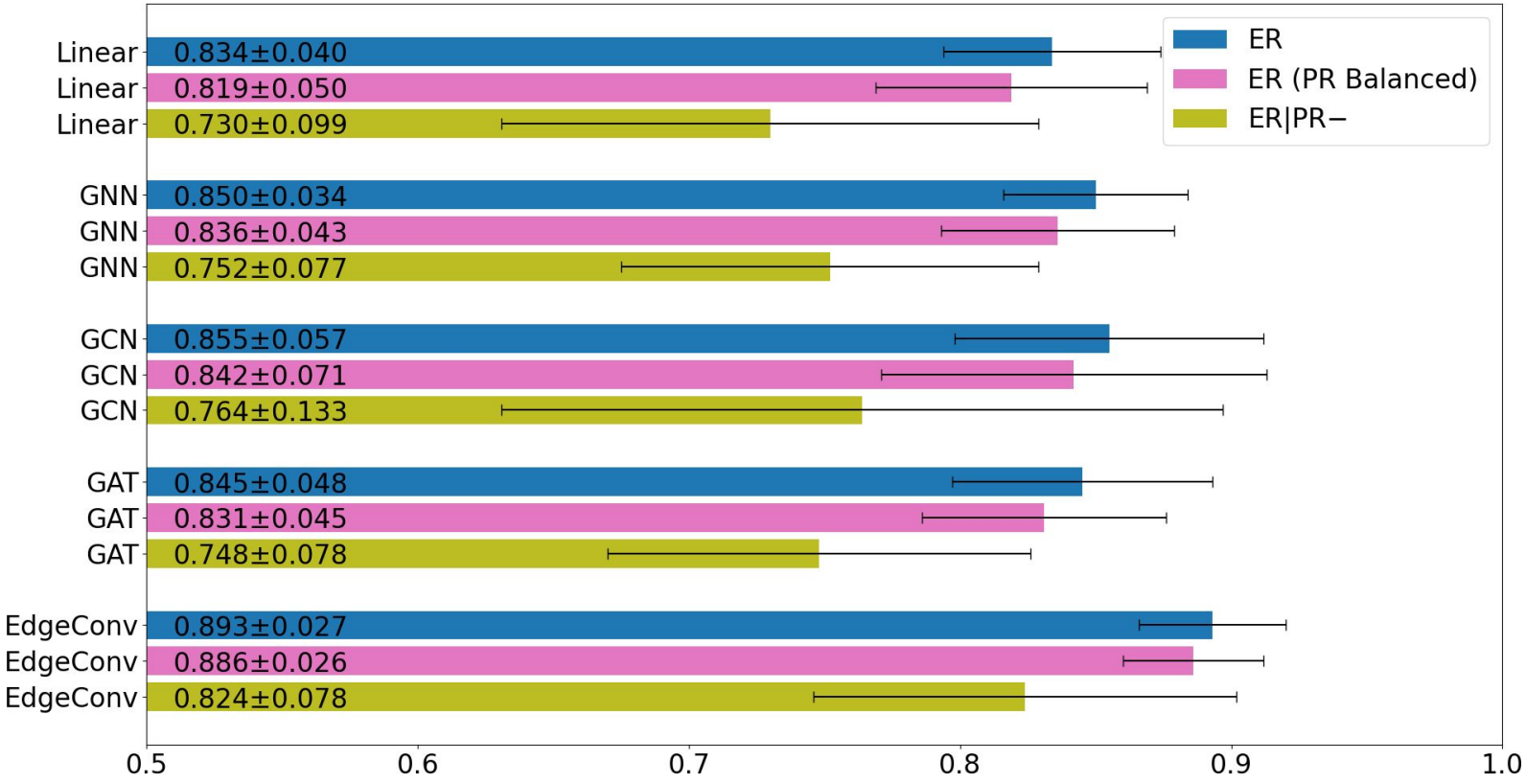
Results: Baseline effects of model type and dataset



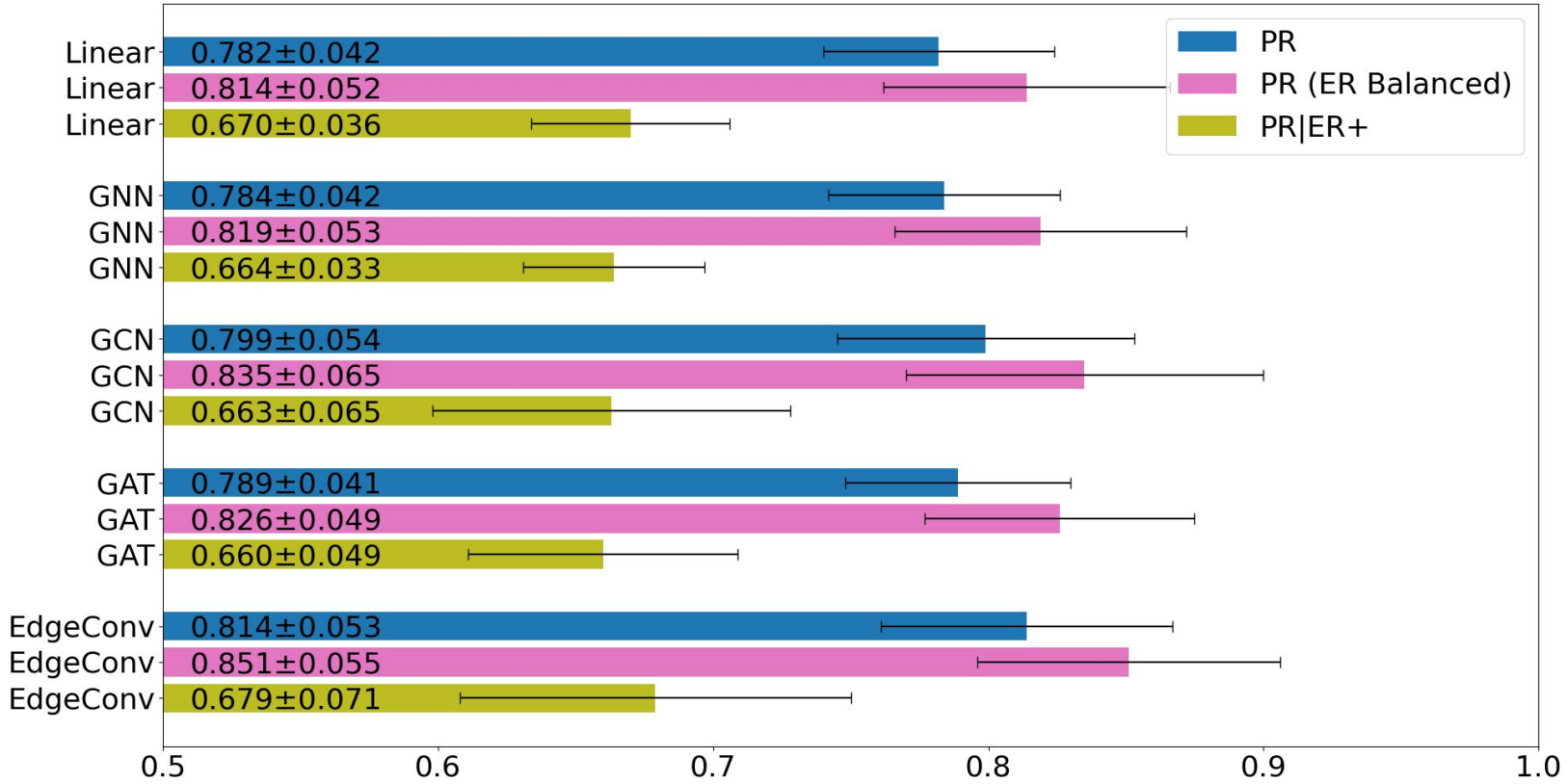
Analysis: Baseline effects of model type and dataset

- As expected, ER is easier than PR
- For TCGA there are signs of the expected hierchy: Linear is the worst performing (but simplest) model, and EdgeConv is the best performing (but most complex) model
- For ABCTB, this hierarchy collapses — this is probably due to ABCTB using UNI features and TCGA using ImageNet features rather than due to the underlying data
    - This is an interesting result that UNI features may be superior
- It is notable that the difference between the datasets is: larger for ER than for PR, and smaller for EdgeConv than for the other models

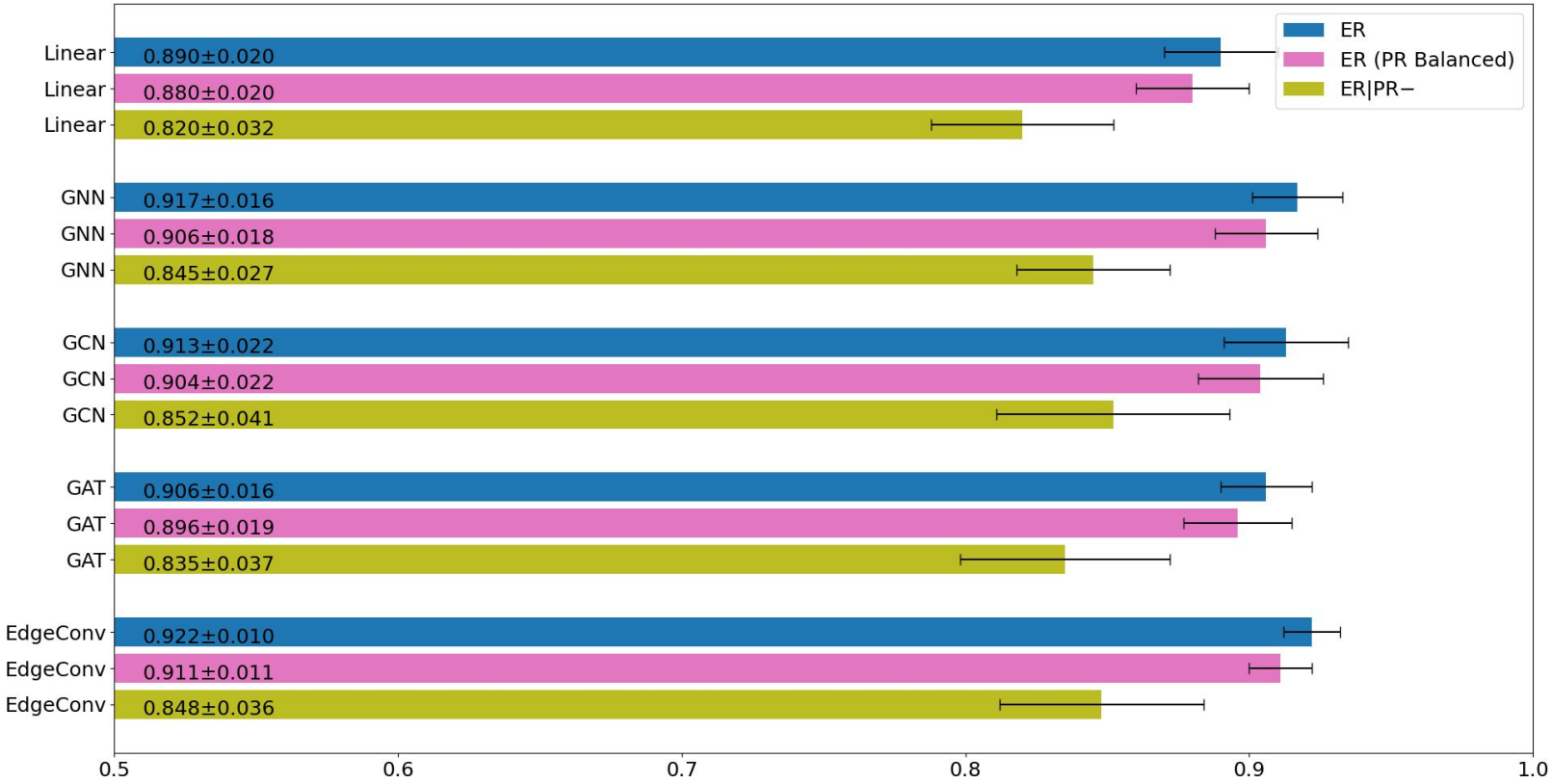
Results: Baseline levels of confounding: TCGA ER



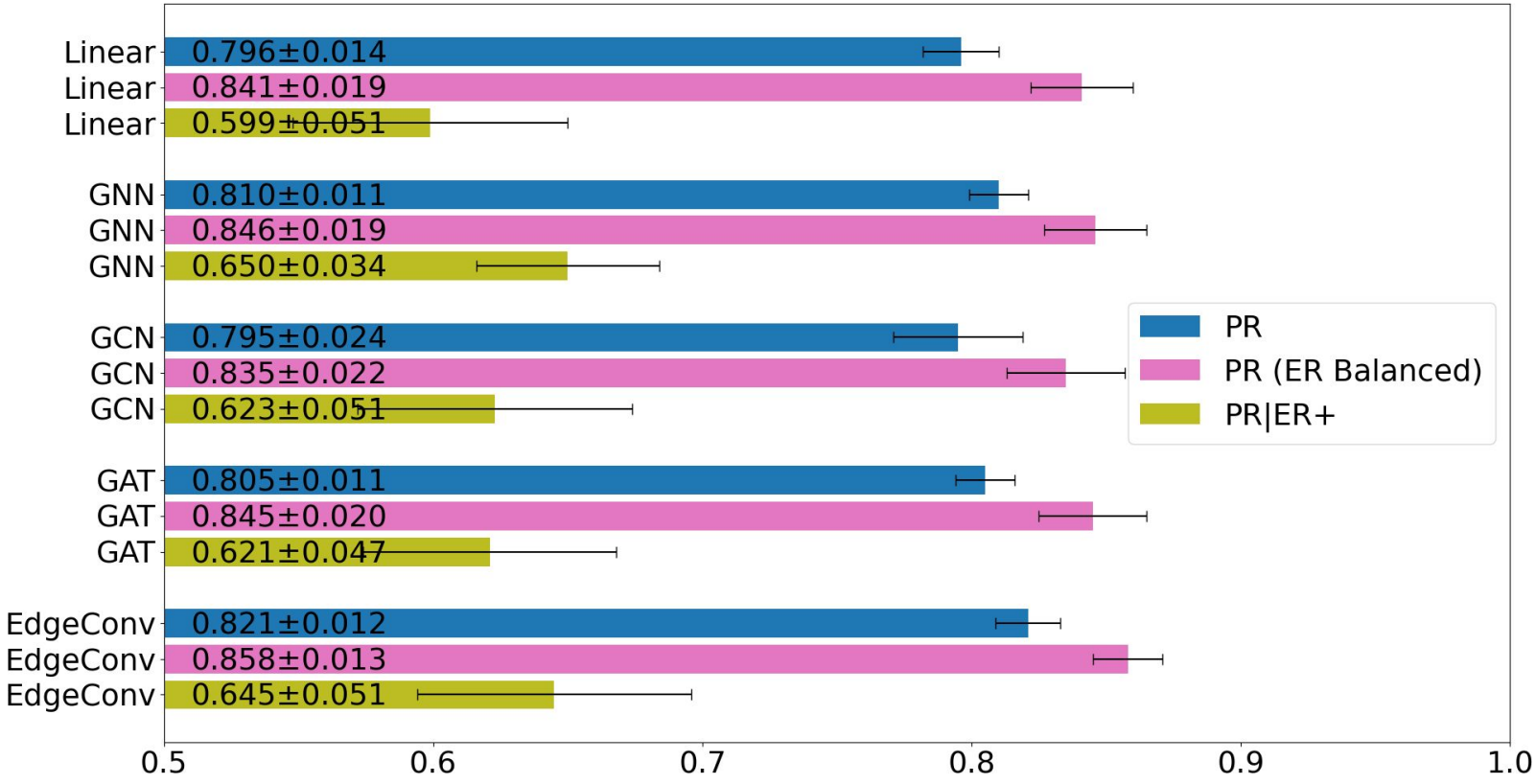
Results: Baseline levels of confounding: TCGA PR



Results: Baseline levels of confounding: ABCTB ER



Results: Baseline levels of confounding: ABCTB PR



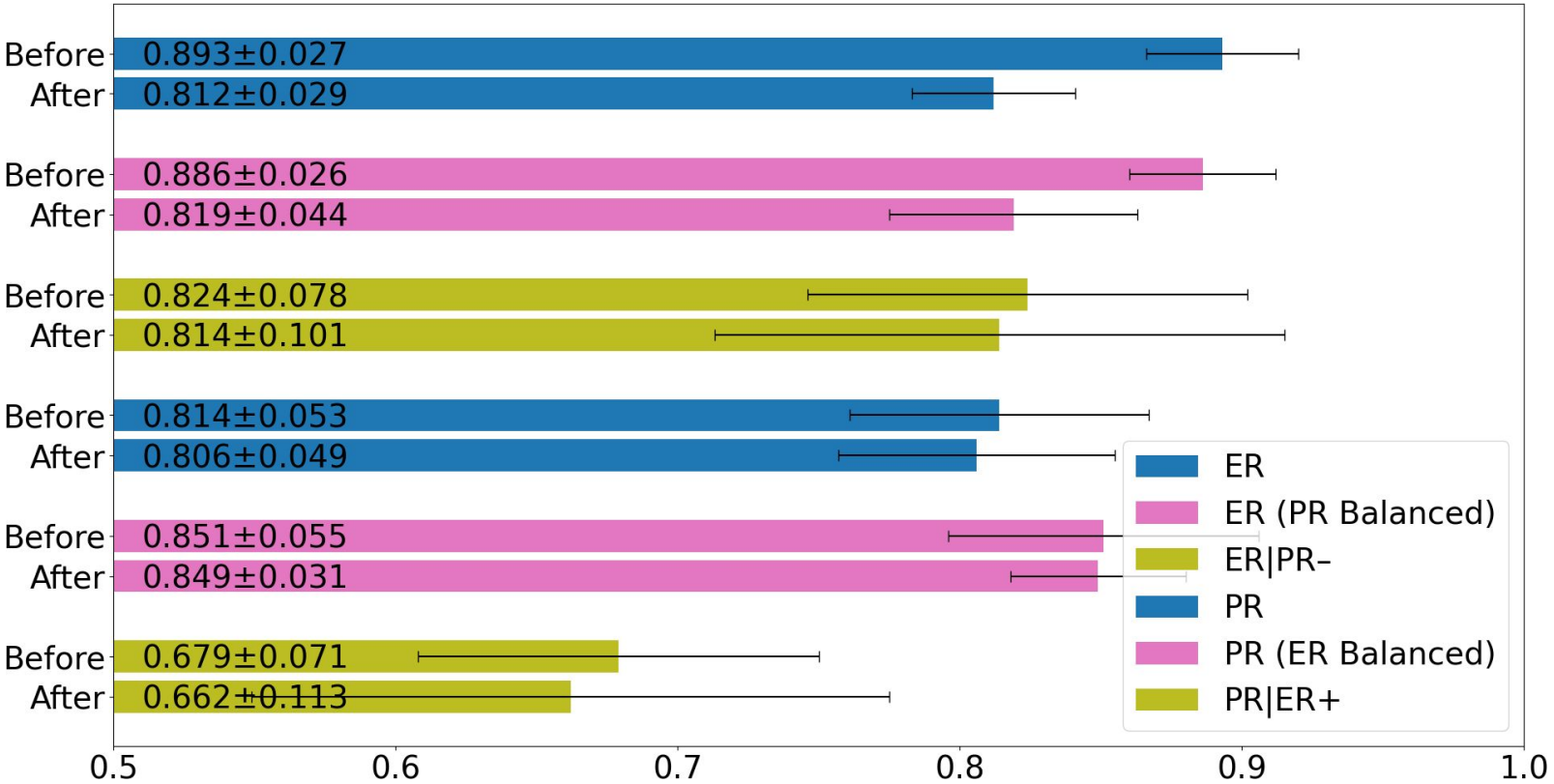
Analysis: Baseline levels of confounding

- We have reproduced the result from Buyer Beware that: although the base AUC_ROCs seem impressive, the models are actually very poor at distinguishing between ER and PR
  - Moreover, we have made the new discovery that the base AUC_ROC falls off faster with decreasing model complexity than the AUC_ROC within the difficult subset
- We see a larger drop off between overall performance and performance in the difficult subset for PR than for ER
  - This is likely because ER causes PR

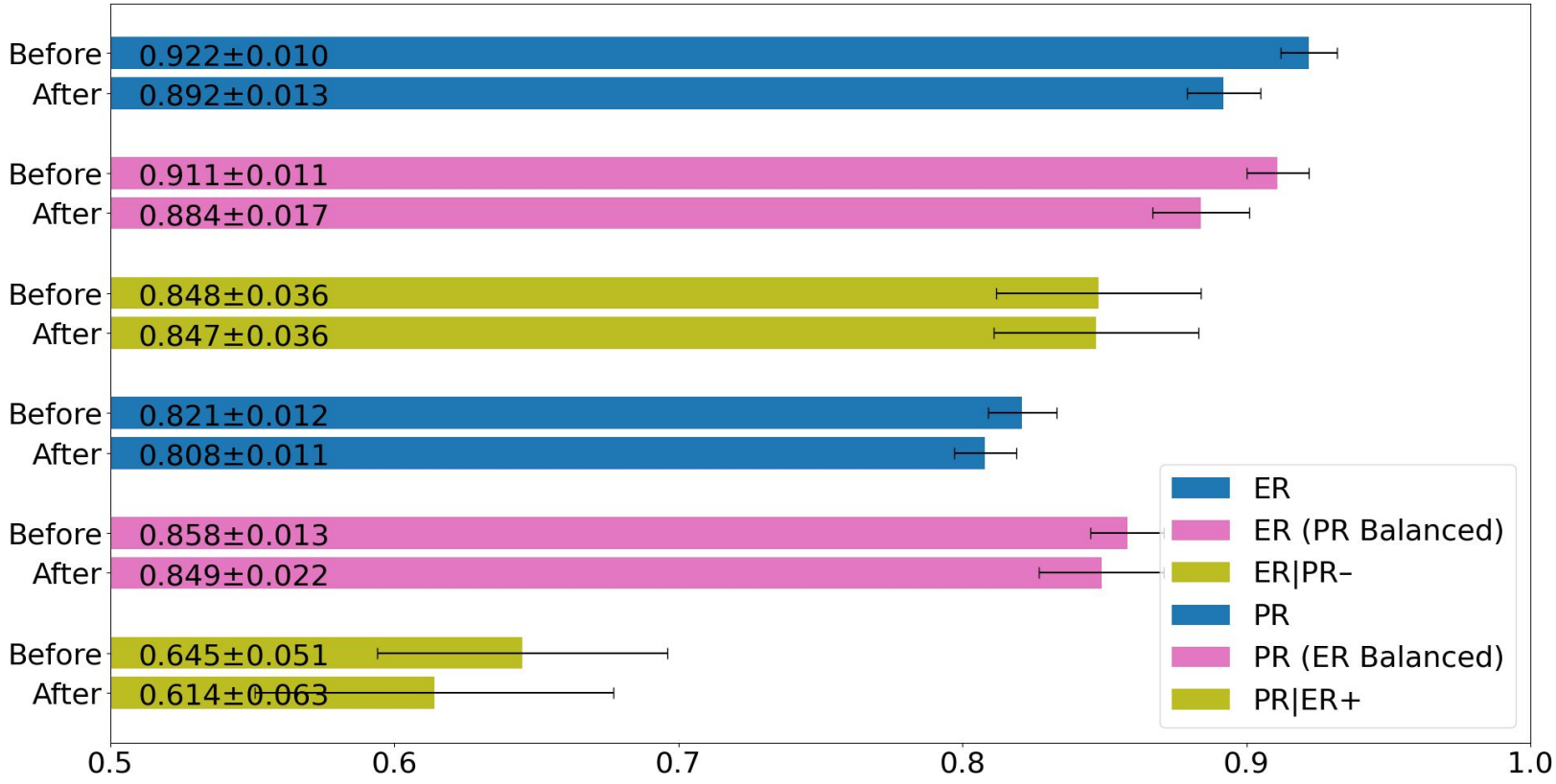
Intervention: Oversampling to remove correlations

- In line with standard machine learning practice, we have always been oversampling to remove the class imbalance between ER + and ER− (and between PR+ and PR−)
  - This stops the model from exploiting the facts that P(ER+ | ∅) > 0.5 and P(PR+ | ∅) > 0.5
- We propose that this can be generalised for our particular context of 2 correlated binary variables: remove the class imbalances between ER+|PR+ and ER −|PR+ and between ER+|PR− and ER−|PR− (mutatis mutandis PR|ER)
  - This stops the model from exploiting the facts that P(ER− | PR−) > 0.5 and so on
- It should be noted that this is not a scalable solution: We are fortunate to have been able to intuitively identify a confounder and to already have labels for it
  - Moreover, conditioning on multiple confounders at once would make the levels of oversampling to apply become very noisy as each intersection would only contain a small number of examples

Results: Oversampling: TCGA EdgeConv



Results: Oversampling: ABCTB EdgeConv



Analysis: Oversampling intervention

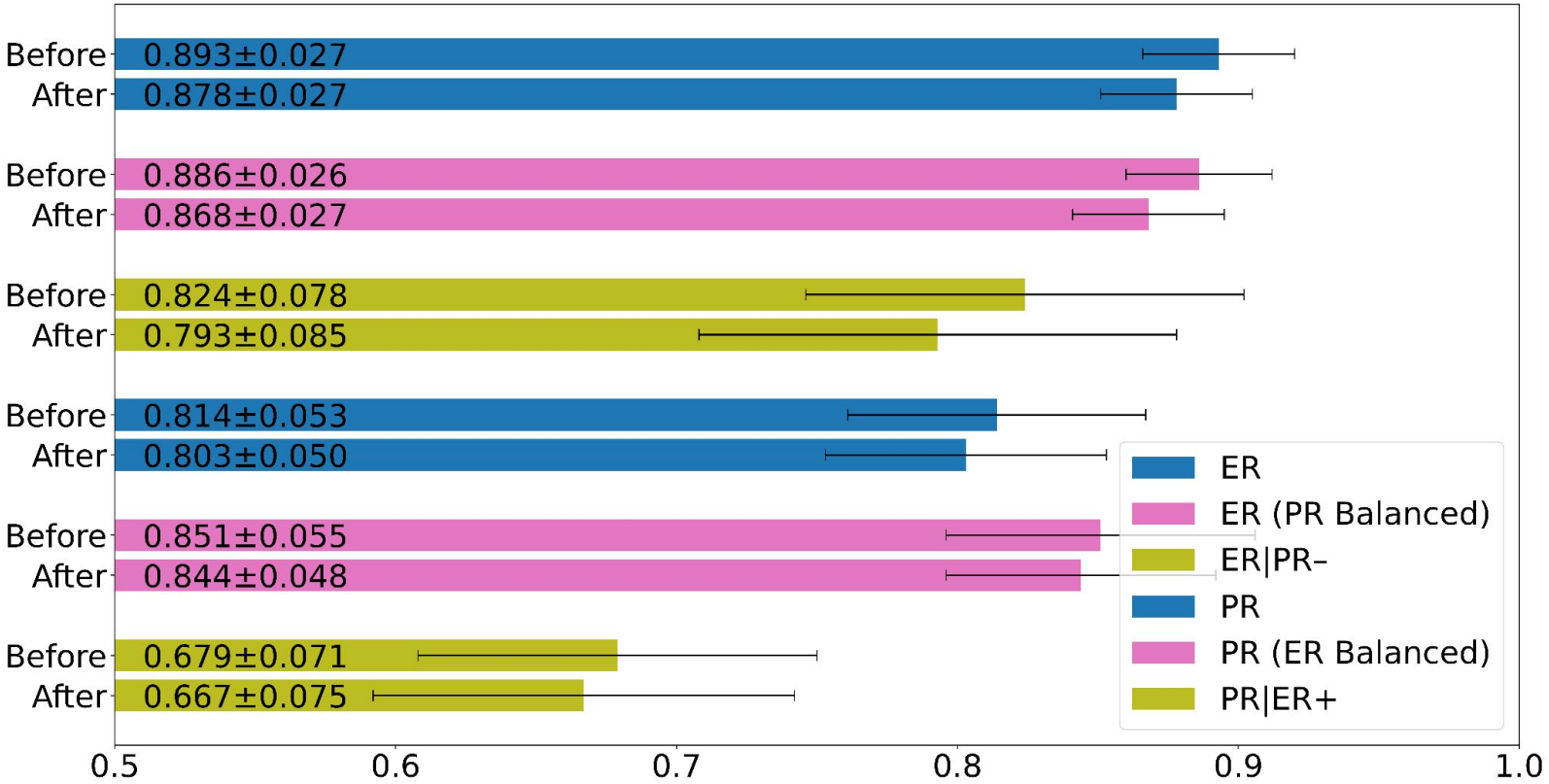
- Amplifying the gradients for the hard (ER+PR−) examples, has hurt the performance on the easy examples (ER+PR+ and ER −PR−). Critically, unfortunately, in doing so the model has not managed to improve the performance on the hard examples (at validation time)
- The most likely explanation for the complete failure of this intervention is: as there is only a small number of examples of ER+PR−, it is easy for the model to find features which work well at train time but are not present in the validation data (spurious correlations)
- We now have an indication that the model was not merely being lazy but really cannot find distinguishing features as it has not struggled just as much to do so when placed in a situation where the only way of decreasing the loss is to do so
    - It is possible that H&E stains simply do not contain enough information to independently predict ER and PR

Intervention: Hinge loss

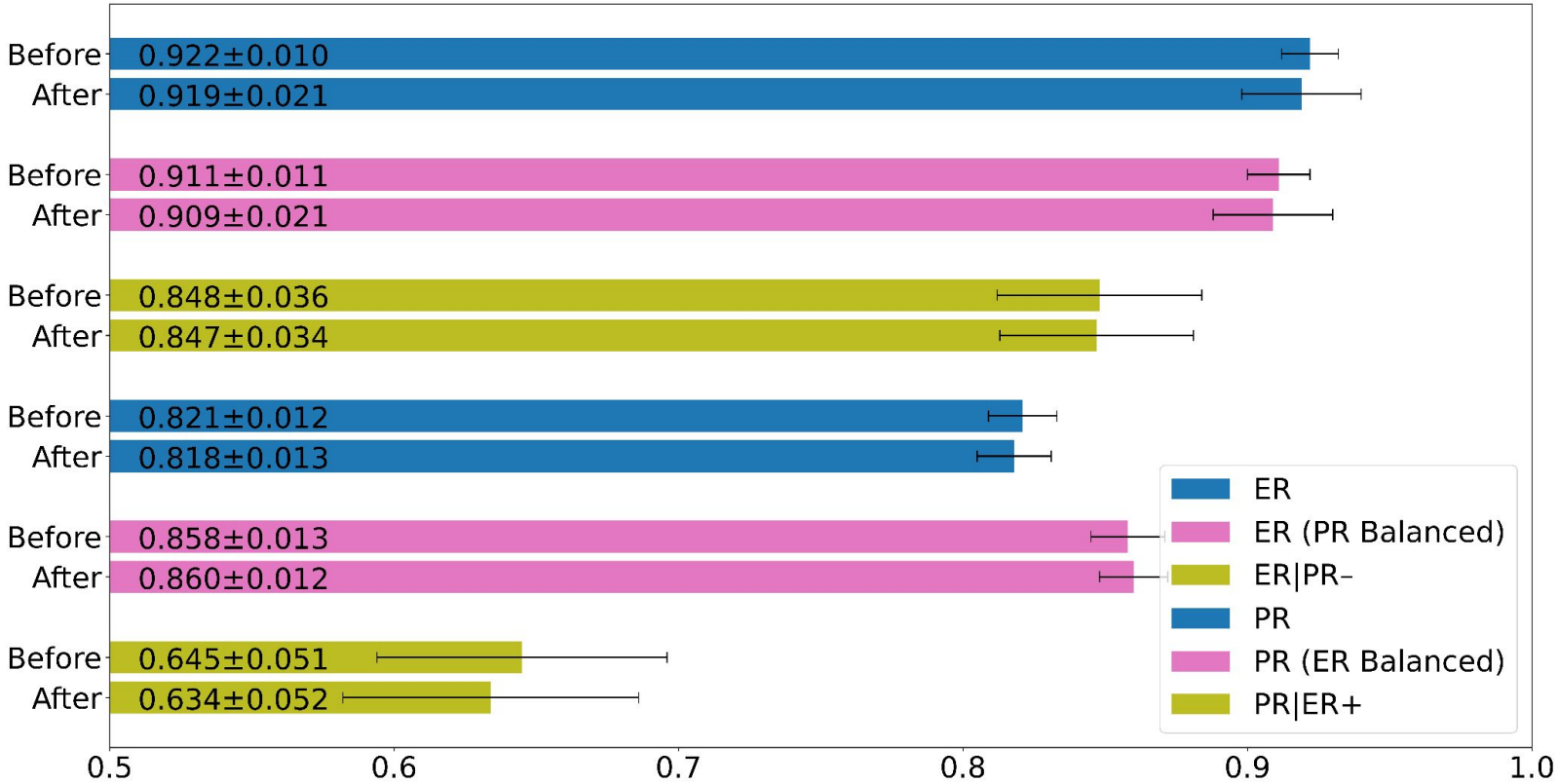
- The cross-entropy loss (and associated sigmoid on the output) is a key factor in the vanishing of the gradients (from examples that are already accurately classified) that would have enabled the learning diverse features
  - Cross-entropy loss: $-[t \cdot \log(\sigma(y)) + (1 - t) \cdot \log(1 - \sigma(y))]$
- It has been proposed that if diverse features are desired, then the hinge-loss should be used instead [1]
  - Hinge-loss: $\max(0, 1 - (2 \cdot t - 1) \cdot y)$

[1] David I. Rodenhiser, Joseph D. Andrews, Theodore A. Vandenberg, and Ann F. Chambers. Gene signatures of breast cancer progression and metastasis. Breast Cancer Research, 13(1):201, January 2011. doi: 10.1186/bcr2791.

Results: Hinge loss: TCGA EdgeConv



Results: Hinge loss: ABCTB EdgeConv



Analysis: Hinge loss intervention

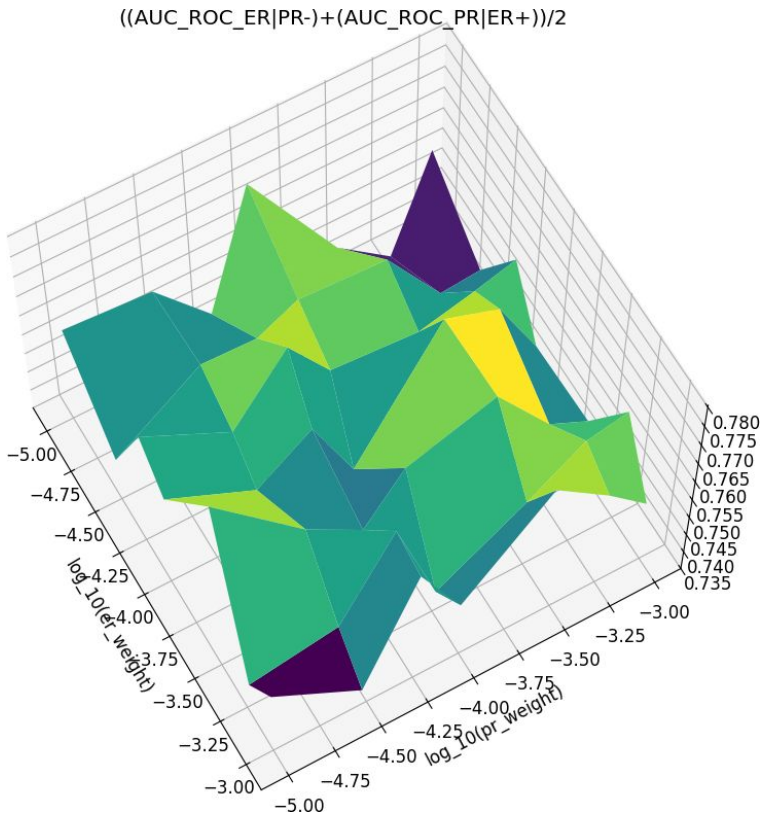
- This has had less of a negative impact on overall performance than oversampling did
- This has had the same remarkable lack of impact on confounding as oversampling did

Intervention: Spectral decoupling

- Hinge loss and spectral decoupling are the main methods that have been proposed for counteracting gradient starvation and thus encouraging the learning of diverse features
- Spectral decoupling [1]: Instead of weight decay, use the L2 norm of the predictions of the model (as raw logits (ie before the sigmoid is applied)) as a regularization term in the loss function
    - Intuition: Promote the learning of features that are only predictive in particular contexts
- The incentive from the error term to focus on features that will increase the average accuracy of predictions is counterbalanced by the incentive from this penalty term to focus on the features that decrease (or at least do not increase) the average confidence of predictions.
    - The regularisation strength must be carefully balanced to neither over- nor under-penalise the reinforcement of existing features as we do want to confidently learn some features that are shared across many examples
- Spectral decoupling has been shown to improve generalisation of prostate cancer detection [2] but they did find the regularisation strength to be a sensitive hyperparameter

[1] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. https://arxiv.org/abs/2011.09468, November 2020.
[2] Joona Pohjonen, Carolin Stürenberg, Antti Rannikko, Tuomas Mirtti, and Esa Pitkänen. Spectral decoupling allows training transferable neural networks in medical imaging. iScience, 25(2):103767, February 2022. doi: 10.1016/j.isci.2022.103767.

Results: Spectral decoupling



Analysis: Spectral decoupling intervention

- The difference in the metrics between the lowest and the highest point of our hyperparameter tuning of the regularisation strengths is very small and there is no clear pattern
- The inescapable conclusion is that we are only seeing random fluctuations

Interventions: Conclusion

- We have tried a collection of techniques that are known to encourage deep learning models to become more invariant and which take aim at differing components of the training process
- All our attempts have completely failed to have any positive impact on our problem
- We feel that the most likely explanation at this point is that there simply are not the correlations present between the H&E images and the IHC labels to be able to produce invariant predictions

Project conclusion

- ER/PR prediction from H&E has been used as a common benchmark in computational pathology, however we have confirmed and strengthened the result from Buyer Beware that this task is much harder than the headline results suggest
- It is possible that although our models are bad at detecting ER+PR− cases, they are good at detecting physical signatures of oestrogen-sensitivity and are actually "mis-classifying" cases which have oestrogen-receptors but impaired oestrogen-signalling to growth. In this case, deep-learning based approaches could be better than IHC for the purpose for which the test results are actually used
    - It would be valuable for researchers with access to data with more labels to investigate this
        - The clinical outcomes of patients following endocrine therapy could be used as a proxy for oestrogen-sensitivity (predicting benefit from endocrine therapy is the primary clinical purpose of running ER/PR tests)
- More generally, the magnitude and persistence of the performance gap we have demonstrated between overall performance and performance on difficult examples on this real-world problem highlights the importance of thorough evaluation of all machine-learning based systems before they are used to replace human judgements in any part of society

Possible future of CPath

- There is a developing area of tumour profiling tests which test for mutations to a select collection of genes of the cancer to get a broader view than the traditional biomarkers (such as ER/PR/HER2 protein expressions) which only provided shreds of evidence of how the cancer is functioning let alone how it could evolve in the face of medical interventions
- Although we focused on ER/PR in this project for the sake of alignment with previous computational pathology work, ER/PR/HER2 are actually highly imperfect proxies for the variable clinicians use them determine: can a patient be safely spared chemotherapy or not?
    - Although ER/PR/HER2 have been the early pioneers of personalising breast cancer treatment, the advent of these tumour profiling tests (although the current generation only looks at a fairly small numbers of genes) makes it possible to imagine a future in which measuring a tiny number of protein expressions to push patients into a small number of boxes rather than understanding a patient's tumour as a whole to truly personalise their treatment seems highly primitive [1]
- Computational pathology may be unduly limiting itself by focusing on trying to exactly replicate IHC rather than exploiting its ability to model any variable we choose

[1] David I. Rodenhiser, Joseph D. Andrews, Theodore A. Vandenberg, and Ann F. Chambers. Gene signatures of breast cancer progression and metastasis. Breast Cancer Research, 13(1):201, January 2011. doi: 10.1186/bcr2791.