UNIVERSITY OF CAMBRIDGE

GATES Cambridge

Multimodal Lego: Model Merging and Fine-Tuning Across Topologies and Modalities

Konstantin Hemker, Nikola Simidjievski, Mateja Jamnik {kh701, ns779, mj201}@cam.ac.uk

Code Paper

ICLR

Problem

Multimodal models require many **paired training samples** for competitive performance.

Fusion & ensembling methods may sacrifice performance from **signal interference**

Model merging requires equivalent architectures to interpolate weights.

How can we build performant multimodal models from pre-trained unimodal encoders?

Solution

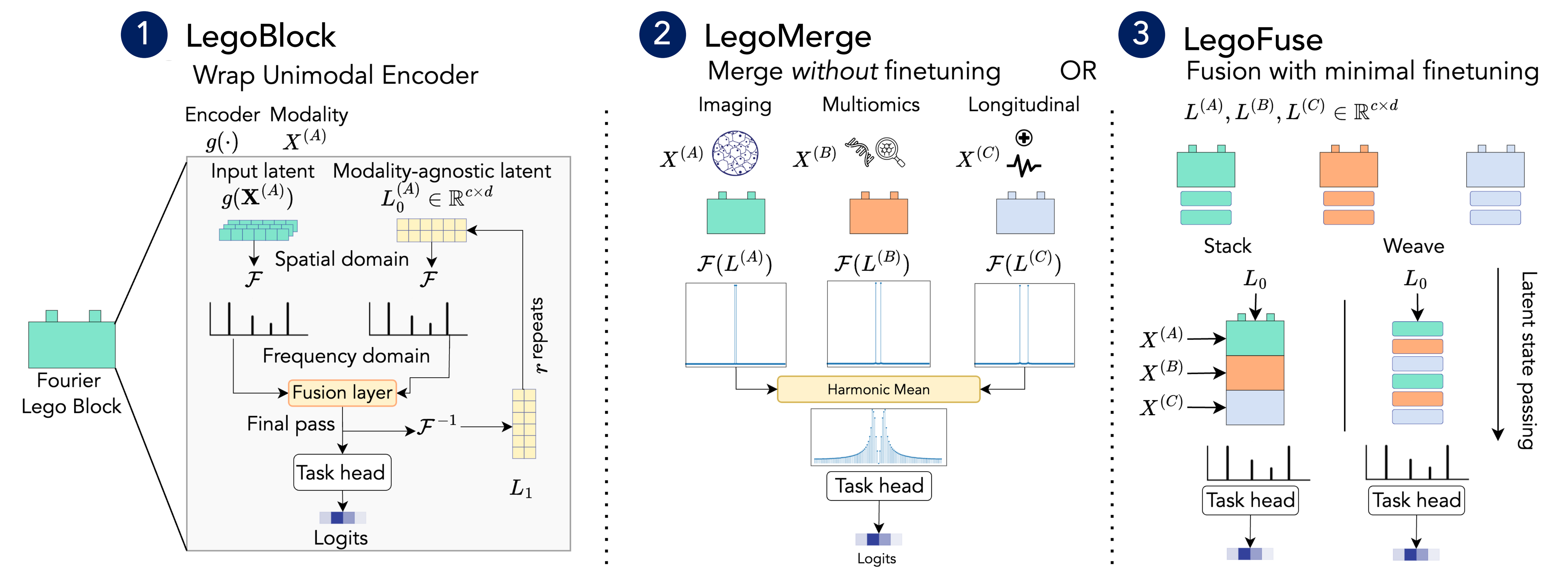
Novel multimodal model merging paradigm with three components:

1. **LegoBlocks:** fits modality-specific adapter to pre-trained model with any topology
2. **LegoMerge:** effectively merges the blocks with little signal interference
3. **LegoFuse:** allows for parameter-efficient fine-tuning

Contributions

- ✓ Performant multimodal models without costly e2e training
- ✓ Agnostic to model architecture, enabling flexible multimodal learning
- ✓ Scalable to any number of modalities
- ✓ Robust to high cross-modal imbalance and/or missing modalities

Architecture Overview



Learns modality-agnostic frequency representation

No paired data required

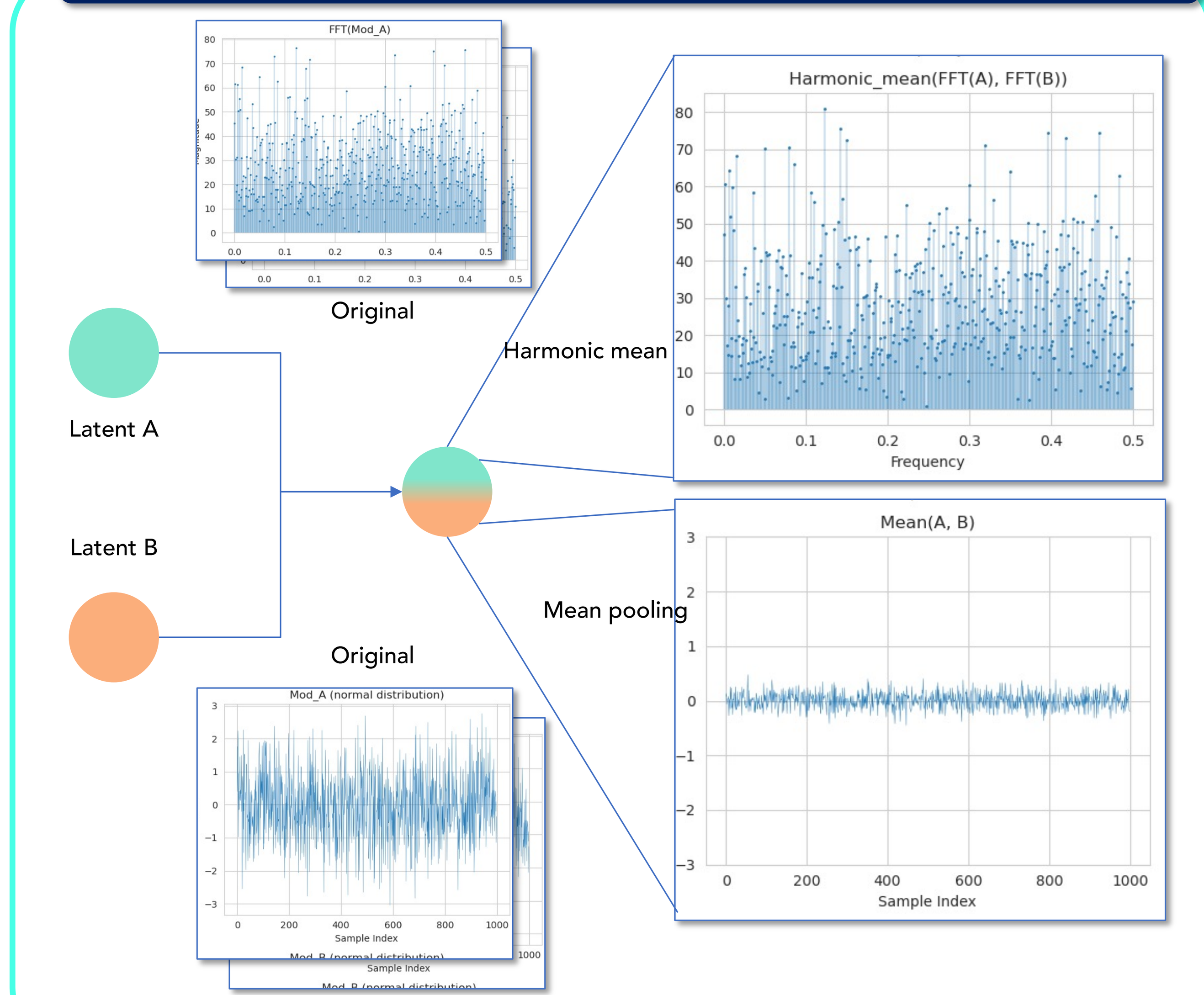
Performant merge into multimodal model "for free"

Outperforms unimodal models

Cross-modal contextualisation

State-of-the-art performance with little fine-tuning

Frequency-domain representations preserve signal

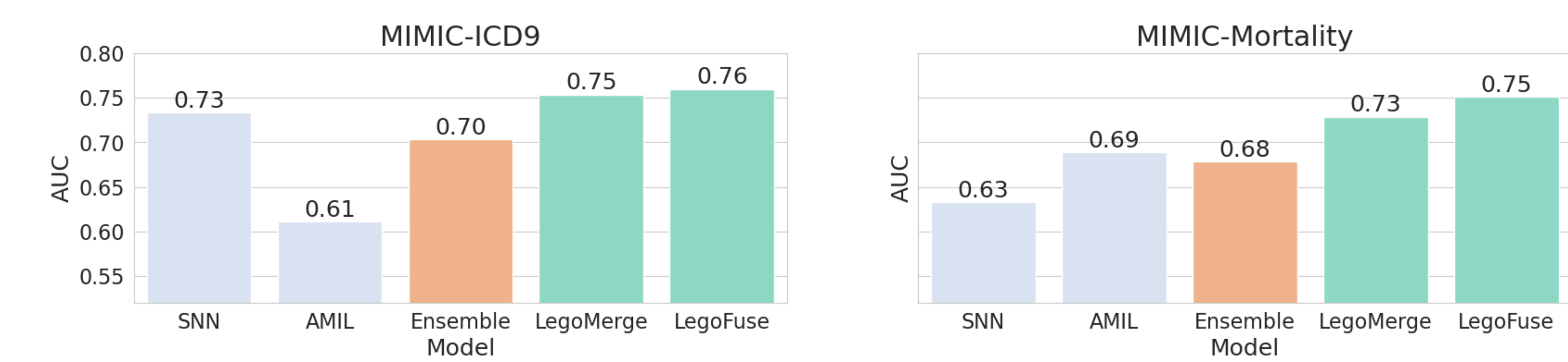


Results

	BLCA	BRCA	KIRP	UCEC	ICD9	MORT	ISIC
Samples	n=436	N=1021	n=284	n=538	n=32616	n=32616	n=2875
Modalities	img, mut, cnv, rna	img, mut, cnv, rna	img, mut, cnv, rna	img, mut, cnv, rna	tab, ts	tab, ts	tab, img
Metric	c-Index	c-Index	c-Index	c-Index	AUC	Macro AUC	AUC
Unimodal (Tabular)							
SNN	$0.689_{\pm 0.012}$	$0.544_{\pm 0.020}$	$0.798_{\pm 0.035}$	$0.589_{\pm 0.057}$	$0.731_{\pm 0.023}$	$0.634_{\pm 0.020}$	$0.507_{\pm 0.005}$
MultiModN	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.525_{\pm 0.140}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$
Perceiver	$0.686_{\pm 0.009}$	$0.557_{\pm 0.016}$	$0.836_{\pm 0.053}$	$0.615_{\pm 0.035}$	$0.629_{\pm 0.023}$	$0.658_{\pm 0.000}$	$\mathbf{0.840_{\pm 0.084}}$
LegoBlock	$0.681_{\pm 0.015}$	$0.591_{\pm 0.021}$	$0.840_{\pm 0.135}$	$0.615_{\pm 0.031}$	$0.645_{\pm 0.017}$	$0.619_{\pm 0.028}$	$0.668_{\pm 0.141}$
Unimodal (Image/Time-series)							
ABMIL	$0.591_{\pm 0.057}$	$0.610_{\pm 0.093}$	$0.741_{\pm 0.080}$	$0.558_{\pm 0.040}$	$0.614_{\pm 0.025}$	$0.691_{\pm 0.014}$	$0.500_{\pm 0.000}$
MultiModN	$0.520_{\pm 0.022}$	$0.527_{\pm 0.150}$	$0.570_{\pm 0.156}$	$0.564_{\pm 0.097}$	$0.500_{\pm 0.000}$	$0.544_{\pm 0.033}$	$0.500_{\pm 0.000}$
Perceiver	$0.532_{\pm 0.027}$	$0.604_{\pm 0.064}$	$0.716_{\pm 0.063}$	$0.534_{\pm 0.106}$	$0.700_{\pm 0.013}$	$0.715_{\pm 0.016}$	$0.719_{\pm 0.050}$
LegoBlock	$0.568_{\pm 0.029}$	$0.533_{\pm 0.000}$	$0.630_{\pm 0.182}$	$0.565_{\pm 0.069}$	$0.643_{\pm 0.013}$	$0.711_{\pm 0.008}$	$0.706_{\pm 0.147}$
Multimodal							
LegoMerge	$0.701_{\pm 0.021}$	$0.601_{\pm 0.025}$	$0.825_{\pm 0.114}$	$0.625_{\pm 0.080}$	$0.684_{\pm 0.015}$	$\mathbf{0.751_{\pm 0.027}}$	$\mathbf{0.721_{\pm 0.143}}$
Uplift (Merge vs. best Block)	2.9%	1.7%	-1.8%	1.6%	5.7%	5.3%	2.1%
SNN + ABMIL (CC, Late)	$0.561_{\pm 0.000}$	$0.541_{\pm 0.104}$	$0.841_{\pm 0.128}$	$0.601_{\pm 0.018}$	$0.628_{\pm 0.020}$	$0.617_{\pm 0.015}$	$0.661_{\pm 0.196}$
SNN + ABMIL (BL, Late)	$0.622_{\pm 0.054}$	$0.557_{\pm 0.089}$	$0.811_{\pm 0.108}$	$\mathbf{0.666_{\pm 0.031}}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.001}$	$0.501_{\pm 0.002}$
Perceiver (CC, Early)	$0.547_{\pm 0.060}$	$0.561_{\pm 0.105}$	$0.692_{\pm 0.000}$	$0.548_{\pm 0.000}$	$0.733_{\pm 0.028}$	$0.723_{\pm 0.015}$	$\mathbf{0.721_{\pm 0.198}}$
MultiModN (Inter.)	$0.524_{\pm 0.018}$	$0.500_{\pm 0.000}$	$0.602_{\pm 0.076}$	$0.512_{\pm 0.008}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$
MCAT (Inter.)	$0.702_{\pm 0.032}$	$0.564_{\pm 0.000}$	$0.823_{\pm 0.076}$	$0.633_{\pm 0.068}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.627_{\pm 0.059}$
HEALNet (Inter.)	$\mathbf{0.714_{\pm 0.025}}$	$\mathbf{0.618_{\pm 0.063}}$	$\mathbf{0.842_{\pm 0.063}}$	$0.594_{\pm 0.023}$	$\mathbf{0.767_{\pm 0.022}}$	$0.748_{\pm 0.009}$	$0.639_{\pm 0.09}$
LegoFuse, w/ 2 epochs	$\mathbf{0.734_{\pm 0.032}}$	$\mathbf{0.626_{\pm 0.046}}$	$\mathbf{0.863_{\pm 0.112}}$	$\mathbf{0.634_{\pm 0.010}}$	$\mathbf{0.771_{\pm 0.020}}$	$\mathbf{0.759_{\pm 0.041}}$	$0.701_{\pm 0.023}$

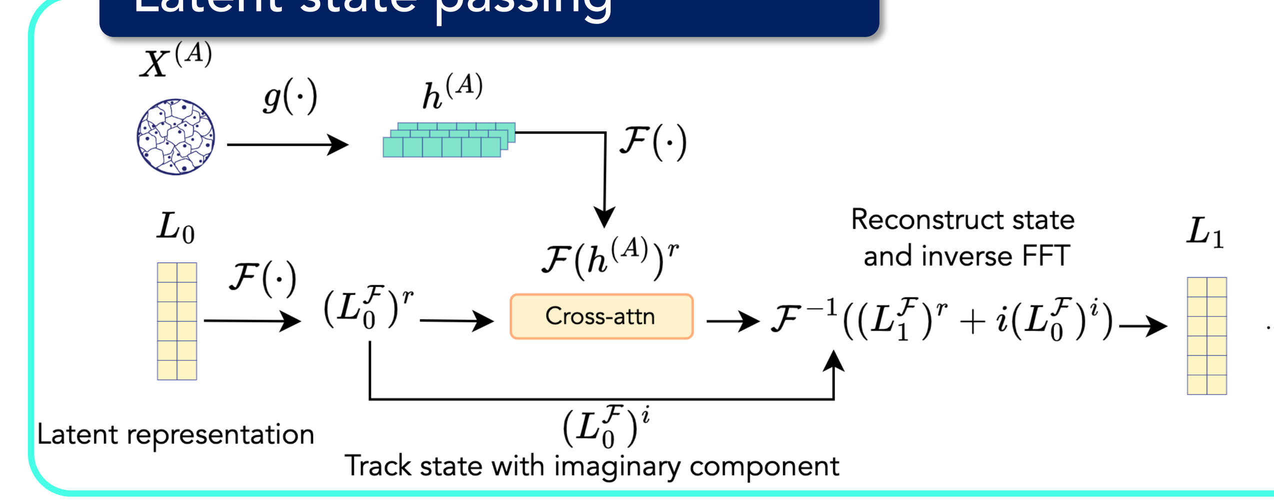
Mean and standard deviation of task performance, showing the concordance Index (survival) and AUC (classification) on 5 random sub-sampling folds with the best and second-best models highlighted.

Merge vs. Ensemble



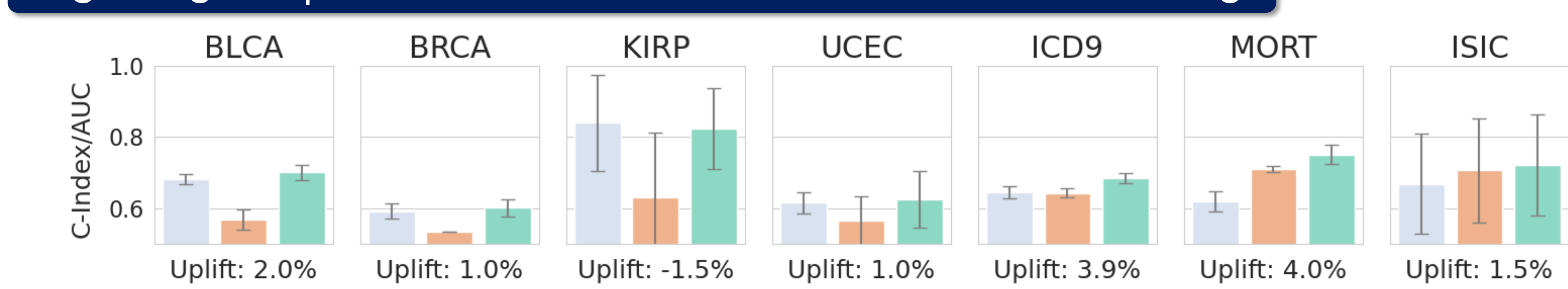
AUC performance on the MIMIC dataset when merging pre-trained using LegoMerge and LegoFuse. Our multimodal model merge shows significantly improved performance over using and ensemble, exhibiting the performance gains at no additional cost through the merge.

Latent state passing



Frequency-domain state passing in the LegoBlock. The real components of the FFT are used in the cross-attention while the imaginary component is used for reconstruction.

LegoMerge outperforms unimodal models without additional training



Mean task performance (c-Index/AUC) of LegoBlock (Tabular), LegoBlock (Image/Time Series) and LegoMerge, showing the increase in task performance by applying the multimodal model merge without any fine-tuning. Our proposed method shows improved performance on 6 out of 7 tasks.