Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

No gain from attention or graph modules in matched ResNet-50, ResNet-18, EfficientNet-B0 and MobileNetV2 baselines for nine-class chest radiographs

Kui Wang
Shihezi University
Rui Li
Shihezi People's Hospital
Yinzhen Li
First Affiliated Hospital of Shihezi University Medical College
Yazhou Zhang
Manas County People's Hospital
Licong Wang
Baiyin City Center for Disease Control and Prevention
Xiaolei Tong
Shihezi University
Qianwen Wang
Shihezi University
Yong Feng
Shihezi University
Dongsheng Rui
Shihezi University
Shangzhi Xu
Shihezi University
Xianghui Zhang
zxh818@shzu.edu.cn

Shihezi University

Article

Keywords:

Posted Date: October 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11065495/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Lightweight attention and graph modules are reported to improve chest radiograph classification, yet the reported gains are small, module-dependent and rarely tested with paired statistics. We compared five such modules (SE, CBAM, BAM, GCN and GAT) with an otherwise identical ResNet-50 baseline on a non-saturated nine-class chest radiograph task (NIH ChestX-ray14, 9,532 training and 2,383 validation images), using ten seeds per model, bootstrap confidence intervals, paired Wilcoxon tests with false-discovery-rate control and a pre-specified equivalence margin of 0.01 AUC. No module outperformed the baseline: all five lay within the equivalence margin, although CBAM and GAT were marginally negative. Three insertion positions of SE and CBAM, including the in-branch placement of the original papers, were likewise equivalent, while gating a block's complete output degraded performance (− 0.0274 and − 0.0703 against the baseline, negative in 10 of 10 seeds). Raising the training set 2.4-fold and repeating the comparison on ResNet-18, EfficientNet-B0 and MobileNetV2 reproduced the absence of gain (19 of 21 comparisons equivalent, two worse and none better), and a patient-level split left it unchanged, although the per-block penalty was specific to ResNet-50. Such gains should be substantiated by multi-seed, paired comparisons against an identical baseline.

1. Introduction

Deep convolutional networks pretrained on ImageNet dominate chest radiograph classification. In practice the standard recipe is to take a ResNet-family backbone and adapt it to the target disease-labelling task [1, 2]. A large body of later work then proposes to improve that backbone by inserting lightweight attention or graph modules: channel attention (squeeze-and-excitation, SE), spatial-channel attention (CBAM), bottleneck attention (BAM), and graph convolutional or graph attention layers (GCN, GAT) [3, 4, 5, 6, 7]. The premise is that such modules sharpen the feature representation and thereby raise accuracy. Their appeal is easy to see, since they add few parameters and slot into an existing backbone with little effort.

Whether that premise holds is less clear. Many studies report that attention-augmented networks improve medical image applications such as classification [8] and segmentation [9], yet the gains are inconsistent across modules. A systematic comparison of self-attention in medical imaging found module-dependent effects, occasionally negative, with one global-context attention module performing below its ResNet-18 baseline [10]. Reported gains are also rarely subjected to paired statistical testing, and among the few studies that applied formal tests, one found no significant difference between attention-augmented architectures [11]. A recent systematic integration study embedded SE and CBAM into five CNN backbones, reported consistent gains, and also varied module placement, preferring selective, deeper placement over inserting a block after every layer; it compared architectures, however, without a paired, multi-seed position contrast [12]. There is reason to expect the margin to be thin. When the baseline is already strong, as it typically is after ImageNet pretraining and fine-tuning, the room for measurable improvement narrows, and a difference of a few thousandths in area under the curve (AUC) becomes hard

to separate from the run-to-run variability introduced by stochastic optimization and CUDA non-determinism.

This study tests one specific and falsifiable hypothesis: that the five modules named above produce a measurable improvement over an identical ResNet-50 baseline on multi-class chest radiograph disease classification. The task is deliberately nine-class and non-saturated, so that a genuine gain, if it exists, has room to appear. Our earlier benchmarks had made that point concrete, since a tuberculosis-versus-normal binary task saturated at an AUC of about 0.999, which leaves no headroom for any enhancement to act in. Everything else is held constant, including the backbone, the data split, the augmentation, the optimizer and the schedule, and each model is trained across multiple random seeds. Differences between models are evaluated with bootstrap confidence intervals and paired statistical tests that correct for multiple comparisons.

2. Methods

2.1 Dataset

The data come from the NIH ChestX-ray14 collection [2], the 14-label release of the database introduced in that work. It is a public set of frontal-view chest radiographs whose labels were extracted by natural-language processing of the radiology reports, and it therefore carries the label noise documented for this collection. From the 112,120 images we retained the 107,010 whose label set contains at least one of the nine target classes (atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax and no-finding). Within that group 15,497 images carry two or more target labels and 91,513 carry exactly one. From the latter we excluded the 4,982 images that paired a single target label with one of the six non-target labels, which left 86,531 images carrying exactly one label drawn entirely from the nine-class space. All images are 1024×1024 grayscale PNG files.

The class distribution is heavily imbalanced (Table 1). To hold training cost down while keeping headroom for a measurable effect, we subsampled each class to a per-class cap of 1,200 images for training and 300 for validation, using stratified 80/20 splitting with a fixed random seed. The two classes with fewer images than the cap, cardiomegaly and pneumonia, were used in full. The resulting training set holds 9,532 images and the validation set 2,383. To mitigate the imbalance, the loss applied class weights proportional to the inverse class frequency.

Table 1
Class distribution of the nine-class subset and the subsampled benchmark splits.

Class	Full subset (n)	Train (n)	Validation (n)
No finding	60,361	1,200	300
Infiltration	9,547	1,200	300
Atelectasis	4,215	1,200	300
Effusion	3,955	1,200	300
Nodule	2,705	1,200	300
Pneumothorax	2,194	1,200	300
Mass	2,139	1,200	300
Cardiomegaly	1,093	874	219
Pneumonia	322	258	64
Total	**86,531**	**9,532**	**2,383**

2.2 Models

The baseline is ResNet-50 with ImageNet-pretrained weights [1]. All variants share an identical backbone and an identical classification head, and differ only in the inserted module. We follow the "backbone → transition → prediction" paradigm of Wang et al. [2] and insert each module once at the transition between the backbone and the classification head, that is, after the last residual stage, rather than inside every stage. SE applies squeeze-and-excitation to the stage-4 feature map [3]. CBAM applies channel attention followed by spatial attention [4]. BAM is implemented in its standard form, with batch-normalised branches, channel and spatial attention summed and passed through a single sigmoid, and output F + F·M [5]. GCN and GAT first project the stage-4 feature map to 512 channels with a 1×1 convolution and adaptively pool it to an 8×8 grid of nodes; they then apply two graph-convolutional layers [6] or one four-head graph-attention layer [7] over a 4-neighbour grid adjacency, and average the node features. After the module the feature map is globally average-pooled and passed through a shared head, a linear projection, ReLU, dropout (0.5) and a nine-way linear classifier. The projection maps 2,048→512 for the attention modules and 512→512 for the graph modules, which pool inside the module.

To reproduce a "freeze the front portion" fine-tuning regime, we froze the first two stages, conv1, bn1, relu, max-pool and residual layers 1–2, and trained only the last two residual stages, the inserted module and the classification head. Figure 1 shows the overall design.

In a separate position ablation (Section 3.2) we compared three placements of the same SE and CBAM implementations, holding the data, the frozen layers, the schedule and the seed set identical. The first is a single application at the transition after the backbone, y = M(backbone(x)), which is the placement used in the main benchmark. The second is the placement of the original papers: the module acts on the residual

function of every trainable block, before the identity addition, so that y = relu(x + M(F(x))). The third applies the module to each block's complete output, after the residual addition, so that y = M(block(x)); we refer to it as per-block output gating.

We note explicitly that the original SE and CBAM papers place the module inside the residual branch, on the block's residual function and before the identity addition. The in-branch arm reproduces that placement for the two modules we study, whereas the per-block arm does not: it multiplies the block's complete output, including the identity path, and therefore isolates the gating form rather than the position alone. The comparison is also restricted to lightweight attention and graph modules. Dual-path networks and pyramid attention networks are sometimes grouped with attention mechanisms, but the first is a standalone backbone architecture and the second a semantic-segmentation network, so both fall outside the scope of an enhancement-module comparison on a fixed ResNet-50 backbone.

2.3 Experimental design

The comparison is a controlled benchmark in which the inserted module is the only variable. All six models share the same backbone, the same fixed data split and subsampling (data seed 42), the same image preprocessing, the same batch size of 16, the same optimizer (Adam, learning rate 1e-4, weight decay 1e-4), the same learning-rate schedule (step decay by 0.1 every five epochs) and the same evaluation procedure. Each model was trained across ten random initialisation seeds (42–51), which is what allows run-to-run variability to be separated from any module effect.

2.4 Training configuration

Images were resized to 256×256 and cached in memory as 8-bit arrays. Training used random crops of 224×224 with random horizontal flipping, and validation used a centre crop of 224×224; both used ImageNet normalisation. Models were trained for up to 15 epochs with early stopping (patience 5) on the validation macro-AUC, and the loss was cross-entropy with inverse-frequency class weights. Every model was trained in a single default configuration. We did not tune a module-specific learning rate, so the inserted modules are trained from random initialisation at the same learning rate as the partially pretrained unfrozen backbone stages. It is an asymmetry that a per-module learning-rate search could in principle attenuate. Training ran on a single NVIDIA GeForce RTX 4070 Ti SUPER.

2.5 Evaluation metrics

The primary metric is macro-averaged area under the receiver operating characteristic curve (macro-AUC), computed one-versus-rest, so that the nine classes carry equal weight. Because the per-class cap leaves the validation set nearly balanced (Table 1), macro-AUC here does not correct for severe imbalance; its role is to keep the two small classes, cardiomegaly and pneumonia, from being down-weighted. Alongside it we report accuracy and a sample-weighted AUC, weighted by the class frequencies of the capped validation set, which are close to uniform. That weighting differs from the natural prevalence of the full subset, and it should not be read as a prevalence-weighted performance estimate.

2.6 Statistical analysis

We report each model's macro-AUC as the mean ± standard deviation across the ten seeds. Uncertainty at the instance level was quantified by averaging the per-seed predicted probabilities into an ensemble prediction per model and computing 2,000-iteration bootstrap 95% confidence intervals for the macro-AUC. To test whether each enhancement differs from the baseline, we applied a paired Wilcoxon signed-rank test to the per-seed macro-AUC differences between each variant and the baseline. Those five comparisons against the shared baseline make up a single pre-specified family, so we control the false-discovery rate across the family with the Benjamini–Hochberg procedure and report the FDR-adjusted significance alongside the uncorrected two-sided p-values. We also give the Bonferroni threshold (0.0100) as a conservative reference, so that the multiplicity of the family is addressed explicitly rather than left implicit. The position experiment of Section 3.2 was designed after the main benchmark had been analysed, and its four placement contrasts form a second, secondary family that we control separately under the same procedure; the two in-branch modules against the baseline are reported with uncorrected p-values inside that family. Because a reader may prefer to treat every module-versus-baseline comparison as one family, we also report the pooled analysis of all nine such comparisons, which leaves the main conclusions unchanged: CBAM and GAT remain the only two of the five main-benchmark modules that differ from the baseline, and the two per-block arms join them.

As a complementary instance-level test, we computed McNemar's exact test on the paired predictions of each variant versus the baseline within each seed and report the median and range of the per-seed p-values.

We additionally assessed equivalence directly. For each module we report the 95% confidence interval of the paired per-seed Δmacro-AUC from a 5,000-resample bootstrap, and a two one-sided tests (TOST) equivalence test against a pre-specified margin of 0.01 AUC. The test uses the 95% confidence interval rather than the conventional 90%, a deliberately more conservative choice, and declares equivalence when the whole interval lies inside ± 0.01. We also computed the design's minimum detectable effect and its power to declare equivalence. The two bootstrap procedures resample different quantities, a single ensemble prediction in one case (2,000 resamples) and the ten paired per-seed differences in the other (5,000 resamples), so their counts are not interchangeable; both are large relative to the quantities involved.

The minimum detectable effect at 80% power follows $(z_{0.975}+z_{0.80})\cdot\sigma_d/\sqrt{n}$, where σ_d is the standard deviation of the ten paired per-seed differences and $n = 10$. Across the five modules σ_d ranged from 0.0029 to 0.0059 and the minimum detectable effect from 0.0026 to 0.0052 AUC (median 0.0028). The power to declare equivalence within ± 0.01 when the true difference is zero was approximately 1.0 (≥ 0.998) for every module.

2.7 Ethics, data provenance and tool use

This study used a previously collected, publicly available and de-identified dataset (NIH ChestX-ray14); no new data were collected and no ethical approval was required. During the preparation of this work the authors used DeepSeek to revise language and to assist with code documentation. After using this tool,

the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

3. Results

3.1 Overall performance and statistical significance

All six models completed training across all ten seeds (60 runs, 15 epochs maximum with early stopping on the validation set). Table 2 reports overall performance. The baseline reached a per-seed macro-AUC of 0.7990 ± 0.0023 (mean ± SD over the ten seeds) and an ensemble macro-AUC of 0.8139 (95% CI 0.8028–0.8252). No variant improved on it: SE 0.7991 ± 0.0035, BAM 0.7999 ± 0.0039, GCN 0.7977 ± 0.0034, CBAM 0.7962 ± 0.0027 and GAT 0.7936 ± 0.0022, with ensemble values between 0.8073 and 0.8142 and all six ensemble confidence intervals overlapping. The ensemble macro-AUC exceeds the per-seed mean by about 0.015 for every model, which is expected when ten seeds are averaged, and it is reported for comparison rather than as the primary quantity. Weighted-AUC follows the same ordering. Accuracy is low in absolute terms (0.42 to 0.46) because the nine classes are imbalanced and inversely weighted cross-entropy rewards recall of the rarest class, so it is not a sensitive measure here.

Table 4 reports the paired comparisons against the baseline. On the seed-level test only two modules differed: CBAM (− 0.0028, 8/10 seeds negative, $p = 0.0195$) and GAT (− 0.0053, 10/10 seeds negative, $p = 0.0020$), both in the negative direction. Both effect sizes nonetheless lay inside the pre-specified equivalence margin of 0.01 AUC, and the equivalence analysis placed all five modules within it, with the 95% confidence interval of each paired Δ macro-AUC inside ± 0.01. Under Benjamini–Hochberg control of the false-discovery rate across the five comparisons (threshold $2/5 \times 0.05 = 0.02$ for the two smallest p-values), CBAM and GAT remain significant and no module is significant in the positive direction; under Bonferroni control only GAT remains significant.

The instance-level McNemar test agreed in direction with the seed-level test throughout. It was less sensitive for the post-backbone modules, reaching $p \leq 0.0305$ for GAT only, which is the behaviour expected of a test on 2,383 validation images rather than a disagreement between tests; for the two per-block variants it reached $p < 0.001$, matching the seed-level result described in Section 3.2.

Per-class values are given in Supplementary Table S1. Task difficulty is strongly class-dependent, with cardiomegaly easiest (baseline ensemble AUC 0.9480) and pneumonia hardest (0.7223; it is the smallest class, with 258 training and 64 validation images), and the ranking of class difficulty is identical in all six models.

Table 2
Overall performance of the six models.

Model	Per-seed macro-AUC (mean ± SD)	Ensemble macro-AUC (95% CI)	Weighted-AUC (mean ± SD)	Accuracy (mean ± SD)
baseline	0.7990 ± 0.0023	0.8139 (0.8028–0.8252)	0.8039 ± 0.0024	0.4324 ± 0.0046
SE	0.7991 ± 0.0035	0.8134 (0.8026–0.8245)	0.8042 ± 0.0030	0.4337 ± 0.0065
CBAM	0.7962 ± 0.0027	0.8113 (0.8003–0.8226)	0.8015 ± 0.0029	0.4299 ± 0.0067
BAM	0.7999 ± 0.0039	0.8142 (0.8036–0.8255)	0.8054 ± 0.0030	0.4336 ± 0.0040
GCN	0.7977 ± 0.0034	0.8129 (0.8015–0.8242)	0.8028 ± 0.0029	0.4338 ± 0.0060
GAT	0.7936 ± 0.0022	0.8073 (0.7955–0.8187)	0.7994 ± 0.0024	0.4223 ± 0.0048

Per-seed macro-AUC is the mean ± SD over the ten seeds and is the primary quantity used in the paired tests. The ensemble macro-AUC is a different estimand: the ten per-seed probability matrices are averaged and a single macro-AUC is computed from the averaged probabilities, with a 2,000-resample bootstrap 95% CI. It is reported only as a lower-variance secondary estimate, and its CI is not a confidence interval of the per-seed mean, the two need not agree (for the baseline, the ensemble point estimate 0.8139 lies above the per-seed mean 0.7990). Weighted AUC and accuracy are per-seed mean ± SD.

Table 4
Statistical comparison of each enhancement versus the baseline.

Comparison	Δ macro-AUC (mean ± SD)	95% CI of Δ	Equivalent (± 0.01 AUC)	Seeds negative	Wilcoxon p	BH-FDR significant?	McNemar p (median [min–max])
SE vs baseline	+ 0.0001 ± 0.0031	[− 0.0019, + 0.0018]	Yes	3/10	0.6953	No	0.5846 [0.3081–0.9600]
CBAM vs baseline	−0.0028 ± 0.0032	[− 0.0047, − 0.0009]	Yes	8/10	0.0195	Yes (worse)	0.3069 [0.0609–1.0000]
BAM vs baseline	+ 0.0009 ± 0.0059	[− 0.0026, + 0.0042]	Yes	4/10	0.4922	No	0.4770 [0.2224–0.8405]
GCN vs baseline	−0.0013 ± 0.0048	[− 0.0040, + 0.0015]	Yes	5/10	0.4922	No	0.3336 [0.1626–0.8751]
GAT vs baseline	−0.0053 ± 0.0029	[− 0.0071, − 0.0035]	Yes	10/10	0.0020	Yes (worse)	0.2999 [0.0319–0.8337]

Paired tests on the ten seeds (Wilcoxon signed-rank on per-seed macro-AUC; McNemar's exact test on paired per-image predictions within each seed, summarised across seeds). p = 0.0020 corresponds to the smallest attainable two-sided Wilcoxon p-value at n = 10 (0.001953). The five comparisons against the shared baseline form one pre-specified family; "BH-FDR significant?" refers to Benjamini–Hochberg control across this family (CBAM clears the second-order BH threshold 2/5 × 0.05 = 0.02 by 0.0005) and the more conservative Bonferroni threshold is 0.0100. The 95% CI of Δ is a 5,000-resample bootstrap of the paired per-seed differences; "Equivalent (± 0.01 AUC)" is the two one-sided tests (TOST) conclusion at a margin of 0.01 AUC, pre-specified before the analysis. The minimum detectable effect at 80% power was 0.0026–0.0052 AUC across the five modules (median 0.0028), and the power to declare equivalence within ± 0.01 was ≈ 1.0 (≥ 0.998). Paired differences are computed from the unrounded per-seed macro-AUC values; subtracting the four-decimal values printed in the tables can therefore produce ~ 0.0001-level discrepancies (e.g. the SE per-block SD is 0.0082 from the unrounded values, 0.0081 from the rounded Table S2 entries).

3.2 Insertion position: single, in-branch and per-block placement

The main benchmark inserts each module once after the last residual stage. Because SE and CBAM were originally proposed for insertion inside every residual block, we repeated the comparison at two further placements, keeping the data, the schedule and the ten seeds identical: inside the residual branch of every trainable block, before the identity addition (the placement of the original papers), and at the output of every trainable block, where the module acts on the block's complete output.

Both in-branch variants stayed at the baseline (Table 5, Fig. 3). SE inserted inside the residual branch reached 0.8003 ± 0.0025 and CBAM 0.8006 ± 0.0026, that is + 0.0013 and + 0.0016 against the baseline, with 3/10 and 2/10 seeds negative. Both differences are not only statistically indistinguishable from zero ($p = 0.3223$ and $p = 0.1934$) but smaller than the minimum detectable effect of their own comparison (0.0034 and 0.0033), so this experiment could not have resolved a difference of the size the module was claimed to produce. Inserting a module in the canonical form therefore neither rescued performance nor harmed it.

Both per-block variants performed worse than either the baseline or their post-backbone counterparts. SE at the output of every block reached 0.7716 ± 0.0082 and CBAM 0.7286 ± 0.0184, which is − 0.0274 ± 0.0087 and − 0.0703 ± 0.0186 against the baseline, and − 0.0275 ± 0.0089 and − 0.0675 ± 0.0182 against the same module inserted once after the backbone. All four comparisons were negative in 10 of 10 seeds (paired Wilcoxon $p = 0.0020$, the smallest attainable two-sided value at $n = 10$), and the instance-level test agreed: McNemar's exact test gave $p < 0.001$ in every seed for CBAM per-block against the baseline and $p \leq 0.0312$ for the remaining three comparisons.

Because every seed is negative, that p-value sits at the floor imposed by the sample size, so we also report the intervals, which are far from zero: −0.0274 with 95% CI [− 0.0331, − 0.0231] for SE per-block and − 0.0703 with 95% CI [− 0.0810, − 0.0592] for CBAM per-block against the baseline. The degradation is therefore not an artefact of statistical power. Its direction also contradicts the expectation that per-block placement should be the more favourable one, and it revises the interpretation we would otherwise give to the position ablation.

Table S6 compares the three placements directly. The two non-degenerate placements are indistinguishable from each other: SE in-branch versus SE post-backbone is + 0.0012 ($p = 0.1055$) and CBAM in-branch versus CBAM post-backbone is + 0.0044 ($p = 0.0137$), the latter significant under the position family but small, with no counterpart for SE. Both placements are clearly better than per-block output gating: +0.0287 for SE and + 0.0720 for CBAM ($p = 0.0020$ for both). Because the three arms have the same capacity, depth and number of inserted modules and differ only in the gating form, the harm is attributable to gating a block's complete output, which also gates the identity path, rather than to inserting a module inside a block. All combinations of data, schedule and seed were held identical across the three placements, so the contrast is paired seed by seed.

Table 5
Insertion position: SE and CBAM inserted once after the backbone, inside the residual branch of every trainable block, and at the output of every trainable block (ten seeds, identical data and schedule).

Model	macro-AUC (mean ± SD)	Ensemble 95% CI	Δ vs baseline	Δ vs same module post-backbone	Wilcoxon p (vs baseline / vs post-backbone)
baseline	0.7990 ± 0.0023	0.8028–0.8252	n/a	n/a	n/a
SE, post-backbone	0.7991 ± 0.0035	0.8026–0.8245	+ 0.0001 (3/10 negative)	n/a	0.6953
SE, in-branch	0.8003 ± 0.0025	0.8016–0.8240	+ 0.0013 (3/10 negative)	+ 0.0012 (2/10 negative)	0.3223 / 0.1055
SE, per-block	0.7716 ± 0.0082	0.7772–0.8003	−0.0274 (10/10 negative)	−0.0275 (10/10 negative)	0.0020 / 0.0020
CBAM, post-backbone	0.7962 ± 0.0027	0.8003–0.8226	−0.0028 (8/10 negative)	n/a	0.0195
CBAM, in-branch	0.8006 ± 0.0026	0.7999–0.8229	+ 0.0016 (2/10 negative)	+ 0.0044 (1/10 negative)	0.1934 / 0.0137
CBAM, per-block	0.7286 ± 0.0184	0.7478–0.7725	−0.0703 (10/10 negative)	−0.0675 (10/10 negative)	0.0020 / 0.0020

Ensemble 95% CI from the 2,000-resample bootstrap of the averaged per-seed probabilities, computed over the same fixed validation set for every arm. Δ values are computed from the unrounded per-seed macro-AUC values; subtracting the four-decimal entries printed here can therefore produce differences at the level of 0.0001.

3.3 Robustness to training-set size, backbone choice and data leakage

The main benchmark subsamples 9,532 training images and uses a single backbone, so the absence of gain could in principle be a property of a small training set, of ResNet-50 in particular, or of leakage between the training and validation splits. We tested all three.

In the first experiment the per-class training cap was raised from 1,200 to the limit of per-class availability (Normal 4,800, Pneumonia 258, Cardiomegaly 874), which increased the training set to 22,898 images while the validation set remained fixed at 2,383 images, so every configuration stays paired with the main benchmark seed by seed. The baseline rose from 0.7990 ± 0.0023 to 0.8112 ± 0.0019, and no module recovered a gain. SE inserted after the backbone remained equivalent (Δ + 0.0012, 95% CI [− 0.0002, + 0.0025], 3/10 seeds negative, p = 0.1309) and CBAM remained marginally negative (Δ − 0.0018, 95% CI [−

0.0035, − 0.0005], 8/10 seeds negative, p = 0.0195), again inside the margin but negative in direction; both per-block variants remained clearly worse (SE − 0.0317, 95% CI [− 0.0388, − 0.0250]; CBAM − 0.0636, 95% CI [− 0.0713, − 0.0559]; 10/10 seeds negative, p = 0.0020), and the position contrasts were preserved (− 0.0330 and − 0.0619, 10/10 seeds negative, p = 0.0020).

In the second experiment the full grid of eight configurations by ten seeds was repeated on ResNet-18, EfficientNet-B0 and MobileNetV2 with the same split, schedule and seeds, freezing the first half of the feature hierarchy of each, which mirrors the ResNet-50 freeze proportionally rather than matching it on parameter count. Of the 21 module-versus-baseline comparisons, 19 fell inside the 0.01 equivalence margin, 2 were worse and none was better; both exceptions are CBAM applied to the output of every block (− 0.0168 on EfficientNet-B0 and − 0.0169 on MobileNetV2, 10/10 seeds negative, p = 0.0020). The three baselines were 0.7866 ± 0.0026, 0.7890 ± 0.0022 and 0.7872 ± 0.0028. The magnitude of the per-block penalty did not transfer: the position contrast averaged − 0.0033 for SE and − 0.0126 for CBAM across the three backbones, against − 0.0330 and − 0.0619 on ResNet-50, and only CBAM was negative on all three. Significance must be read together with direction here: under Benjamini–Hochberg control across the seven comparisons of each backbone, the rejections were SE on ResNet-18 in the positive direction (Δ + 0.0021, p = 0.0039), CBAM per-block on EfficientNet-B0, and SE per-block together with CBAM per-block on MobileNetV2, where SE per-block was equivalent within the margin yet negative by the paired test (Δ − 0.0058, p = 0.0098). EfficientNet-B0 already contains squeeze-and-excitation inside its MBConv blocks, so its baseline includes channel attention and adding SE to it is partly redundant, whereas MobileNetV2 contains no such block and provides a second clean control.

The third experiment replaced the image-level split with a patient-level split, so that no patient contributes images to both sets (9,548 training and 2,367 validation images, no patient overlap; Supplementary Table S5). The baseline was 0.7893 ± 0.0033, which is 0.0097 lower than under the image-level split and consistent with the optimistic bias expected from the patient overlap described below. The conclusions were unchanged: SE remained equivalent (+ 0.0012, 3/10 seeds negative, p = 0.3223) and so did CBAM (− 0.0007, 7/10 seeds negative, p = 0.5566), while both per-block variants remained worse (SE − 0.0240 and CBAM − 0.0933, 10/10 seeds negative, p = 0.0020), with the CBAM penalty larger than under the image-level split. One effect did change: the small negative effect of post-backbone CBAM, which reached Benjamini–Hochberg significance under the image-level split (p = 0.0195), did not under the patient-level split (p = 0.5566), which matches our expectation that patient overlap attenuates differences between models. Because a patient-level split necessarily changes the validation set, its comparison with the main benchmark is between designs rather than paired, while the comparisons within each split remain paired seed by seed.

Taken together, the absence of gain over a matched, identical baseline holds for four backbones and is not an artefact of the training-set size or of leakage between splits, while the magnitude of the per-block penalty does not transfer beyond ResNet-50. Tables S3 and S4 report the full results.

4. Discussion

In a benchmark in which six models differed only in the module inserted into an otherwise identical ResNet-50 pipeline, none of the five attention or graph modules improved multi-class chest radiograph classification. The two whose difference from the baseline reached significance after Benjamini–Hochberg correction, CBAM and GAT, did so in the negative direction, although both effect sizes fell within the 0.01-AUC equivalence margin (Table 4). That result is at odds with the prevailing positive framing of lightweight attention modules in medical imaging [8, 9], which recent studies continue to report: a CBAM-enhanced CNN improved pneumonia-detection accuracy to 98.6% from a 92.08% baseline, with 96.25% for the SE-enhanced variant [13], and a hybrid ViT–DenseNet201 chest-radiograph classifier reached 97.9% overall accuracy [14]. It is, however, consistent with the smaller body of work that applied formal paired testing. A paired McNemar test between two SE-augmented architectures for four-class chest-radiograph classification found no significant difference between them ($\chi^2 = 0.125$, $p > 0.05$) [11], and a systematic comparison of attention modules found module-dependent and occasionally negative effects, with one global-context attention variant falling below its ResNet-18 baseline [10].

Our contribution is not the idea of varying module placement, which has been studied before. Ullah et al. [12], for example, embedded SE and CBAM into five backbones and found selective, deeper placement preferable to inserting a block after every layer. Our contribution is to evaluate such choices under a fully controlled, paired, multi-seed protocol with a pre-specified equivalence analysis, and to show that in this regime neither a single post-transition insertion, nor the in-branch placement of the original papers, nor per-block output gating produces a gain. Two methodological differences from that study are relevant: we held the inserted modules' learning rate equal to the backbone's rather than tuning it separately, and we selected checkpoints on the validation set rather than on the test set. Our position contrast covers three placements, including the in-branch placement of the original papers, and it is the strongest result we report. Adding insertions in the canonical form neither rescued performance nor harmed it: SE and CBAM inserted inside the residual branch of every trainable block stayed equivalent to the baseline (+ 0.0013 and + 0.0016, both below the minimum detectable effect of their own comparison). The degradation appeared only where the module is applied to each block's complete output: in that form SE and CBAM fell by − 0.0274 and − 0.0703 against the baseline, with 10 of 10 seeds negative and $p = 0.0020$. The absence of a gain therefore cannot be explained by a single, late insertion point, and the harm the per-block arms show cannot be attributed to placing the module inside a block, because the in-branch arm does exactly that and remains at the baseline. That harm is specific to ResNet-50 in our data: repeated on ResNet-18, EfficientNet-B0 and MobileNetV2, the same contrast was several times smaller (− 0.0033 for SE and − 0.0126 for CBAM on average, against − 0.0330 and − 0.0619), and only CBAM was negative on all three backbones.

Two non-exclusive explanations account for these results, and our design cannot fully separate them. The first is information redundancy. An ImageNet-pretrained ResNet-50 fine-tuned on 9,532 images already encodes the discriminative structure available at 224×224, and channel or spatial reweighting of its final-stage feature map cannot recover information that the frozen convolutional layers never represented. The pattern of results, rather than any single comparison, supports this view. The three modules whose mean difference from the baseline was indistinguishable from zero (SE, BAM, GCN) are exactly those whose

modification is closest to a reweighting of features that the classifier head can partly absorb, and the in-branch arms, which add the largest parameter count of any arm and still sit at the baseline, show that added capacity alone changes nothing here. The only arms that fell below the baseline are the two that gate a block's complete output, which is a change in how information flows through the block rather than in how much capacity it holds.

The second explanation is optimisation cost. In a small-data, partially frozen-backbone regime, extra trainable parameters inside the trainable stages increase the variance of the learned function without adding useful capacity, which is what the per-seed spread shows: the standard deviation of macro-AUC across seeds grows from 0.0023 for the baseline to 0.0082 for SE per-block and 0.0184 for CBAM per-block. The class-imbalance handling may contribute to the same effect, since inversely weighted cross-entropy emphasises the smallest class, pneumonia with 258 training images, and so rewards solutions that fit rare examples at the expense of overall accuracy, which stayed at 0.42–0.46 in all models.

A third possibility, that the choice of a nine-class single-label task with a weak baseline is itself unfavourable, is unlikely, because the baseline is far from saturation at a macro-AUC of 0.799 and the task therefore does leave headroom in which a genuine gain could appear. We cannot, however, exclude that the effective ceiling of this task is set by label noise rather than by model capacity, since ChestX-ray14 labels are extracted by natural-language processing and are known to be noisy [2], and every model plateaus near a macro-AUC of 0.80. Distinguishing a noise-limited ceiling from a capacity-limited one would require re-annotated labels, which this collection does not provide. We therefore state our result as "no module produced a gain beyond the resolution of this benchmark" rather than as evidence that attention cannot help.

Several limitations qualify these conclusions. The benchmark uses one dataset (NIH ChestX-ray14) at a resolution of 224×224, so the results speak to a computationally constrained fine-tuning regime rather than to full-resolution training on the complete collection; Section 3.3 shows that neither the small training subset of the main benchmark nor the choice of backbone accounts for the absence of gain, but that check has two bounds of its own, since the non-ResNet backbones had the first half of their feature hierarchy frozen proportionally rather than matched on parameter count, and EfficientNet-B0 already contains squeeze-and-excitation inside its MBConv blocks, so its baseline is not a module-free control. The validation set is used both for early stopping and for reporting, which introduces a mild optimistic bias common to all six models; it does not affect the comparisons, but the absolute macro-AUC values should not be read as held-out performance estimates.

The split is also image-level rather than patient-level in the main benchmark. We verified that no content-identical images are shared between the training and validation sets, but 1,009 patients contribute images to both, so 1,350 validation images (56.7%) come from patients who also appear in training. This overlap is a further optimistic bias. It applies to all six models and therefore does not by itself create a module-specific advantage, but because it attenuates between-model differences it acts in the same direction as our equivalence conclusion. The patient-level re-split reported in Section 3.3 confirms both consequences: the baseline fell by 0.0097, and the small negative effect of post-backbone CBAM, which was significant

under the image-level split, was not significant once the overlap was removed. We therefore report the absolute macro-AUC values as development-set figures rather than held-out estimates, and state the equivalence result as holding under this image-level split.

Two classes are small, pneumonia at 258 training and 64 validation images and cardiomegaly at 874 and 219, so class-specific estimates for them are unstable across seeds (± 0.015 for pneumonia). Each module was implemented in one reasonable way, BAM in its standard combined channel-plus-spatial form and GCN and GAT over an 8×8 grid graph built from the stage-4 feature map, and we did not tune the inserted modules per module, whether in learning-rate multipliers, the placement of the graph projection or the number of graph layers. A deliberately tuned variant could conceivably behave differently. Finally, we froze the first two residual stages to reproduce the widely used "freeze the early layers, fine-tune the rest" deployment regime. Whether end-to-end fine-tuning, larger training sets, higher input resolution or a different task would change the conclusion is untested here; the clearest example of such a task is lesion segmentation or detection, where spatial attention has a more direct target.

The practical implication is procedural rather than a blanket rejection of attention mechanisms. When a pretrained backbone is fine-tuned on a medical imaging dataset of the size used here, the baseline is strong enough that the burden of proof falls on the enhancement: it should be evaluated across multiple seeds, compared with a paired test against an identical baseline, and reported with confidence intervals, as any claim about a small effect should be. Our results also indicate that the insertion point and the number of insertions are not free parameters to be copied from an ImageNet setting, since the canonical placement was neutral rather than helpful and gating a block's complete output was actively harmful. Reporting decisions of this kind, including negative ones, is what allows the field to distinguish robust architectural gains from run-to-run variation.

5. Conclusion

In a controlled, multi-seed benchmark of five lightweight attention and graph modules on a nine-class chest radiograph task, with 9,532 training and 2,383 validation images and a ResNet-50 backbone whose early stages were frozen, no enhancement module produced a measurable gain over the identical ResNet-50 baseline, and the absence of gain survived a 2.4-fold enlargement of the training set, repetition on three further backbones and a patient-level re-split.

Two modules were significantly worse than the baseline after Benjamini–Hochberg false-discovery-rate correction across the five comparisons. CBAM fell by − 0.0028 ($p = 0.0195$), significant under FDR but threshold-marginal, since it does not meet the more conservative Bonferroni threshold, and GAT fell by − 0.0053 ($p = 0.0020$), negative in 10 of 10 seeds. No module was significantly better. The placement of the original papers did not change this: SE and CBAM inserted inside the residual branch of every trainable block reached 0.8003 ± 0.0025 and 0.8006 ± 0.0026, that is + 0.0013 and + 0.0016 against the baseline, both below the minimum detectable effect and both inside the equivalence margin. Only output gating degraded performance: SE and CBAM applied to the complete output of every trainable residual block reached 0.7716 ± 0.0082 and 0.7286 ± 0.0184, that is − 0.0274 and − 0.0703 against the baseline and −

0.0275 and − 0.0675 against the same modules inserted once after the backbone, with all four comparisons negative in 10 of 10 seeds (p = 0.0020) and convergent McNemar results within seeds. That penalty, unlike the absence of gain, is a ResNet-50 finding that we do not generalize to the other backbones we tested.

The per-class analysis showed no uniform advantage for any module, and the ranking of class difficulty was identical in all six models, with cardiomegaly easiest and pneumonia hardest.

We conclude that, for this task and regime, lightweight attention and graph modules should not be assumed to improve a pretrained convolutional baseline, and that evaluating them requires multi-seed paired comparisons against an identical baseline together with explicit testing of the insertion position.

Additional Information

Declarations

Competing interests

The authors declare no competing interests.

Data availability

This study re-analyses a public dataset. The NIH ChestX-ray14 collection is available from the NIH Clinical Center (https://nihcc.app.box.com/v/ChestXray-NIHCC). All results reported here were computed from the per-seed prediction files produced by the training runs. The training, evaluation and analysis scripts (train.py, stats.py, compare_A.py, models.py, analyze_inbranch.py, analyze.py, analyze_cross_backbone.py, selfcheck.py, verify_patientsplit.py, export_stage_csv.py), the per-seed prediction files for all 440 runs (60 main + 20 per-block + 20 in-branch + 50 training-set-size + 240 multi-backbone + 50 patient-level), and the split assignment files for the patient-level experiment, and an environment specification (Python, PyTorch and CUDA versions; Python/NumPy/PyTorch seeds fixed per run, with cuDNN non-determinism acknowledged) are openly available (see Code availability). The patient-level split is released as the assignment file for the images actually used (training set, validation set, class and patient identifier), together with the export script, which reproduces the split from the public metadata; the full per-image patient mapping for the complete collection is available from the corresponding author on request. Randomness controls: Python, NumPy and PyTorch (CPU and CUDA) seeds are all set from the per-run initialisation seed; DataLoader workers use num_workers = 0; the cuDNN backend retains its default non-deterministic settings, so bit-exact reproducibility of individual GPU kernels is not guaranteed. The data split and per-class subsampling use a fixed data seed (42) that is independent of the initialisation seed.

Code availability

The analysis code, the per-seed predicted probabilities and the derived result tables are openly available at https://github.com/kwang311/chestxray14-nine-class-module-benchmark (MIT License).

Author contributions

Kui Wang: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration. Rui Li: Writing – review & editing, Validation. Yinzhen Li: Writing – review & editing, Validation. Yazhou Zhang: Writing – review & editing, Validation. Licong Wang: Writing – review & editing, Validation. Xiaolei Tong: Visualization. Qianwen Wang: Formal analysis. Yong Feng: Data curation. Dongsheng Rui: Conceptualization, Writing – review & editing. Shangzhi Xu: Project administration, Funding acquisition, Writing – review & editing. Xianghui Zhang: Conceptualization, Project administration, Supervision, Funding acquisition, Writing – review & editing.

Acknowledgements

This work was supported by the Science and Technology Program of XPCC [2025ZD014], the Xinjiang Talent Development Fund [XJRC-2025-BTJY-YJ-GX-QNQZ-003], and the Shihezi Municipal Finance Science and Technology Program [2026RC06]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); pp. 770–778. (2016). 10.1109/CVPR.2016.90
2. Wang, X. et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); pp. 3462–3471. (2017). 10.1109/CVPR.2017.369
3. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); pp. 7132–7141. (2018). 10.1109/CVPR.2018.00745
4. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); pp. 3–19. (2018). 10.1007/978-3-030-01234-2_1
5. Park, J., Woo, S., Lee, J. Y. & Kweon, I. S. BAM: bottleneck attention module. In: Proceedings of the British Machine Vision Conference (BMVC); Preprint at arXiv:1807.06514 (2018).
6. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR); Preprint at arXiv:1609.02907 (2017).
7. Veličković, P. et al. Graph attention networks. In: International Conference on Learning Representations (ICLR); Preprint at arXiv:1710.10903 (2018).
8. Ferdous, M., Mahmud, S. & Shimul, M. E. Z. MedNet: a lightweight attention-augmented CNN for medical image classification. *Sci. Rep.* **15**, 41936. 10.1038/s41598-025-25857-w (2025).

9. Gu, R. et al. CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging*. **40** (2), 699–711. 10.1109/TMI.2020.3035253 (2021).
10. Rao, A., Park, J., Woo, S., Lee, J. Y. & Aalami, O. Studying the effects of self-attention for medical image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); pp. 3409–3418. (2021). 10.1109/ICCVW54120.2021.00381
11. Huang, K. A., Choudhary, H. K., Santiago, A. & Prakash, N. S. Squeeze-and-excitation enhanced convolutional neural networks for multi-class pneumonia classification on chest radiographs. *Cureus* **17** (12), e99649. 10.7759/cureus.99649 (2025).
12. Ullah, Z., Hong, M., Mahmood, T. & Kim, J. Systematic integration of attention modules into CNNs for accurate and generalizable medical image classification. *Mathematics* **13** (22), 3728. 10.3390/math13223728 (2025).
13. Potharaju, S. et al. Enhanced X-ray image classification for pneumonia detection using deep learning based CBAM and SE mechanisms. *Intell. Based Med.* **12**, 100299. 10.1016/j.ibmed.2025.100299 (2025).
14. Oltu, B. et al. Automated classification of chest X-rays: a deep learning approach with attention mechanisms. *BMC Med. Imaging*. **25**, 71. 10.1186/s12880-025-01604-5 (2025).

Figures

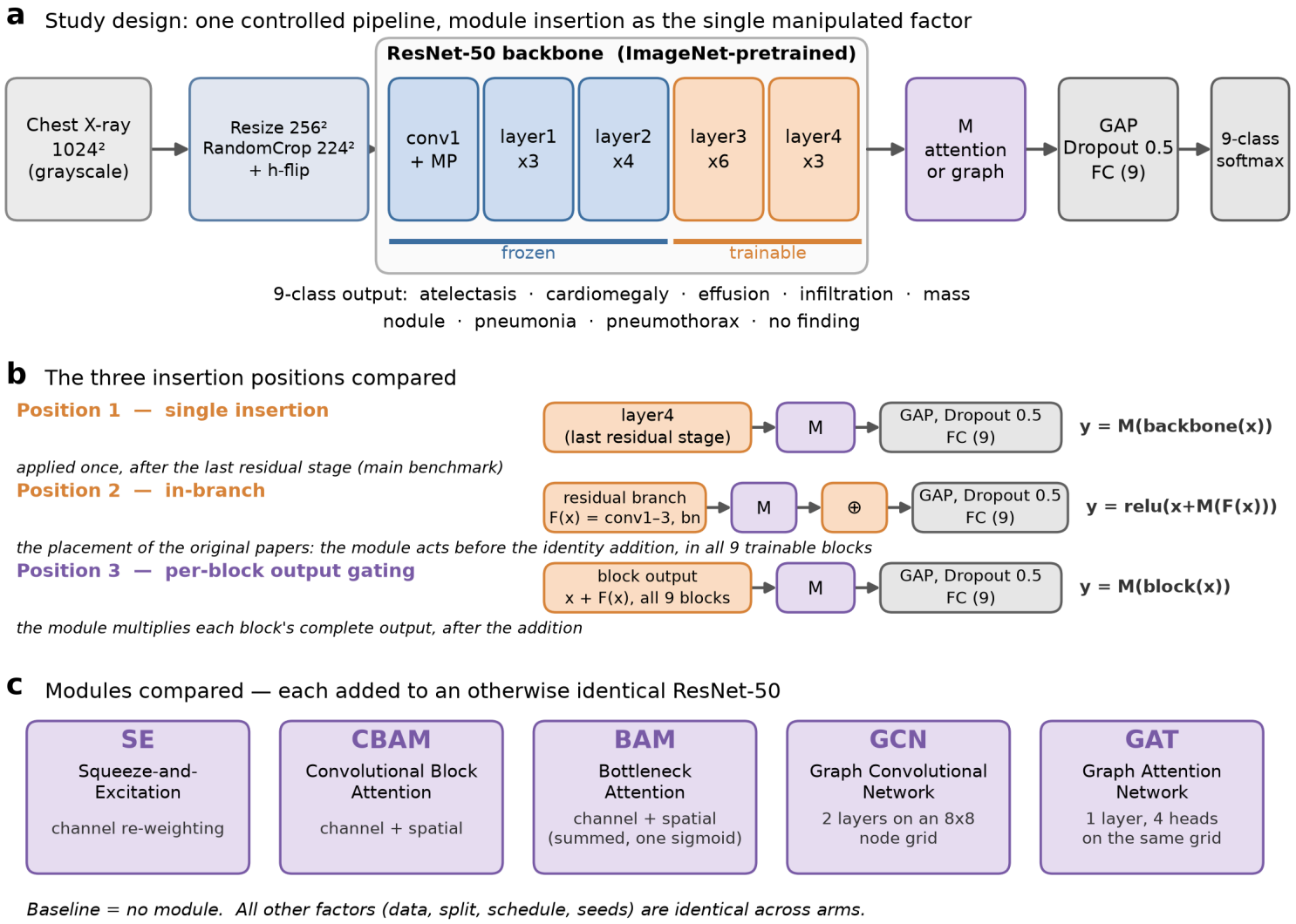


Figure 1

Study design: one controlled pipeline, with module insertion as the only manipulated factor. (a) A single ResNet-50 backbone (ImageNet-pretrained) is fine-tuned on the nine-class subset of NIH ChestX-ray14. The first two stages (conv1, max-pool and residual layers 1–2) are frozen, and only the last two residual stages, the inserted module M and the classification head are trained. Inputs are 224 × 224 crops of the 1024 × 1024 grayscale radiographs, and the head (global average pooling → dropout 0.5 → linear) produces a softmax over atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax and no finding. (b) The three insertion positions compared. In *Position 1* (main benchmark) the module is applied once, at the transition after the last residual stage, and the baseline is the identical network with M removed. In *Position 2* the module acts inside the residual branch of each of the nine trainable residual blocks of layers 3–4, before the identity addition, i.e. y = relu(x + M(F(x))), which is the placement of the original papers. In *Position 3* the module multiplies the complete output of each of those blocks, i.e. y = M(block(x)), which we call per-block output gating. (c) The five modules compared: squeeze-and-excitation (SE), convolutional block attention (CBAM), bottleneck attention

(BAM), graph convolutional network (GCN) and graph attention network (GAT). Dataset, split, augmentation, optimizer, schedule and seed set are identical across all arms.

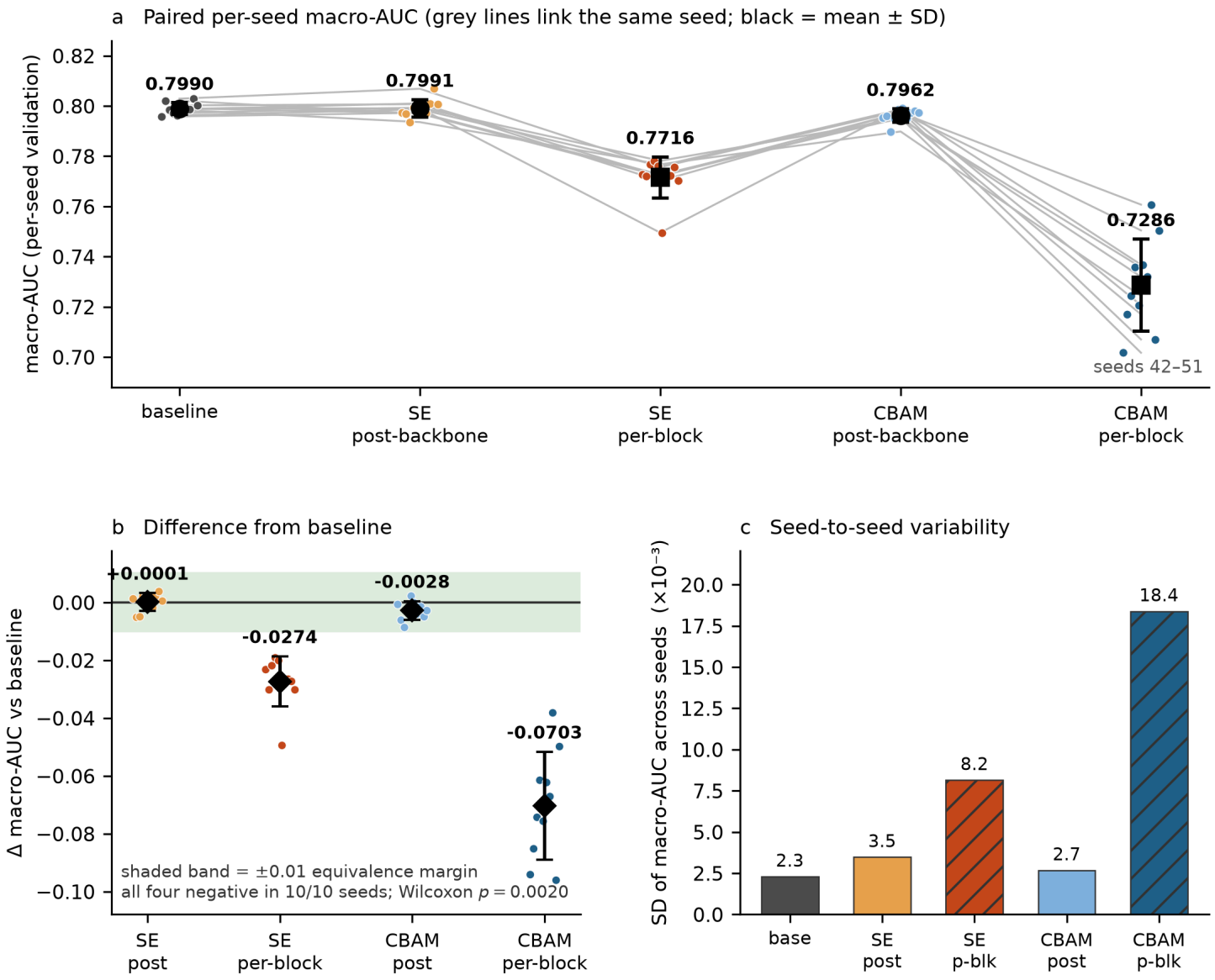


Figure 2

Position ablation: SE and CBAM inserted once after the backbone versus at the output of every trainable residual block. All panels use the same ten initialisation seeds (42–51) and the same fixed data split, so the two experiments are paired seed by seed. (a) Per-seed validation macro-AUC for the baseline, the two post-backbone variants and the two per-block variants; grey lines connect the same seed across conditions and black markers give the mean ± SD across seeds (the y-axis is truncated to 0.688–0.826 so that the per-block shift is visible). (b) Difference from the baseline for each variant: points are individual seeds, diamonds the mean ± SD, and the shaded band the pre-specified ±0.01 AUC equivalence margin. All four comparisons were negative in 10/10 seeds (paired Wilcoxon p = 0.0020, the smallest attainable two-sided value at n = 10). (c) Seed-to-seed SD of macro-AUC (×10⁻³), showing that the added capacity of

per-block insertion inflates run-to-run variability from 2.3 (baseline) to 8.2 (SE per-block) and 18.4 (CBAM per-block); hatched bars mark the per-block variants.

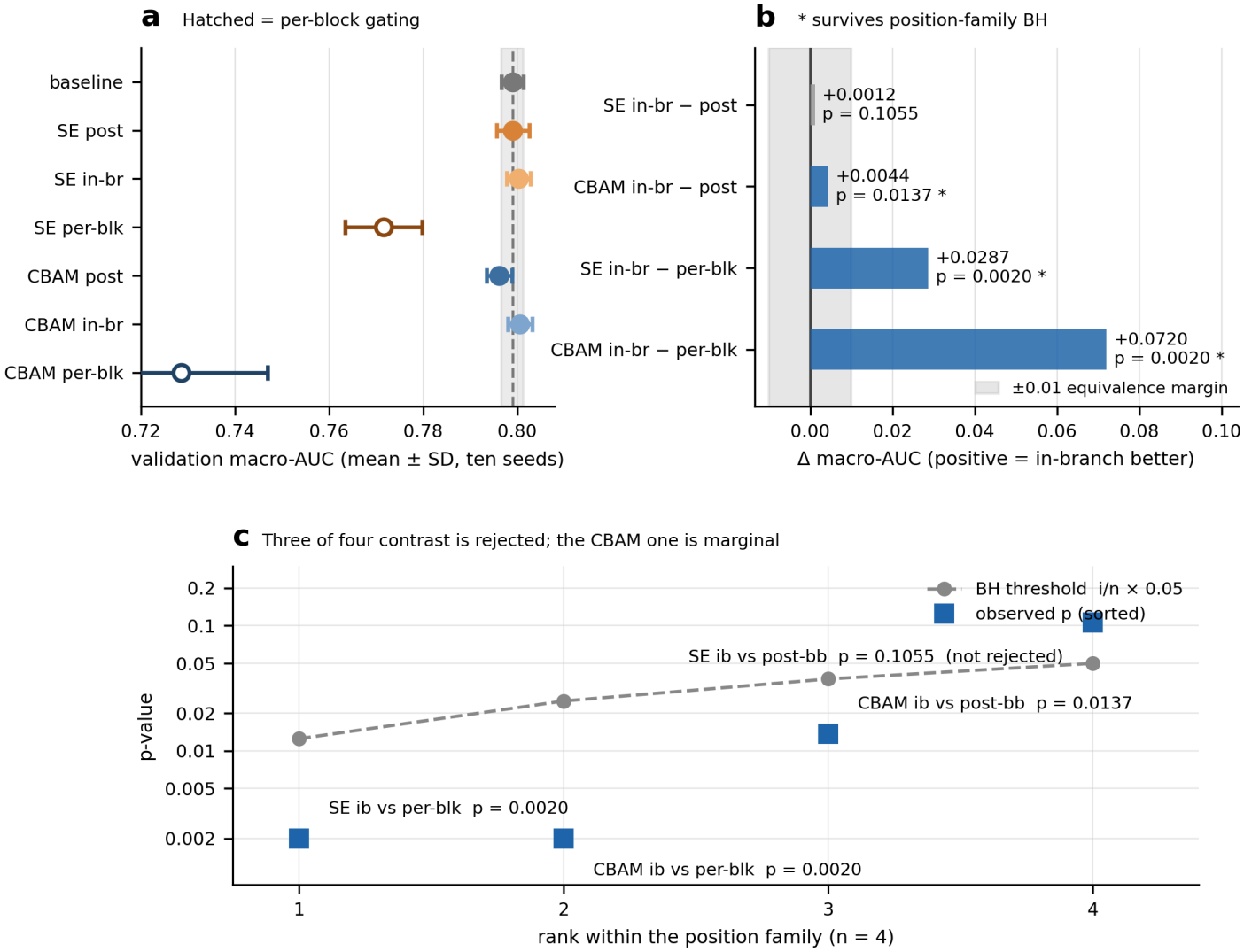


Figure 3

Insertion position compared across three placements for SE and CBAM, on the same data, schedule and seed set. (a) Validation macro-AUC for the baseline (grey band and dashed line, mean ± SD across the ten seeds) and for SE and CBAM at each of the three placements; symbols give the mean and error bars one standard deviation across seeds, and the per-block arms are drawn with open symbols and marked by hatching. (b) Difference from the in-branch arm for the four position contrasts; positive values mean the in-branch placement scores higher, and the grey band marks the pre-specified ±0.01 AUC equivalence margin; asterisks mark the contrasts that remain significant under Benjamini–Hochberg control of the position family (Table S6). (c) Benjamini–Hochberg diagnostic for the position family, plotting the sorted observed p-values against the i/n × 0.05 threshold; three of the four contrasts are rejected, the SE in-branch versus post-backbone contrast being the one that is not.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- paper1supplementary20260916.pdf