FROZEN NETWORKS, BETTER DECISIONS: SNR-CONDITIONED READOUT FOR MODULATION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic modulation classification identifies how information is encoded in a received wireless signal. Noise can hide the patterns that distinguish modulation formats, making accurate recognition difficult. We ask how much accuracy can be gained after a classifier has been trained. We study affine readouts that adjust its output probabilities while keeping the network fixed. The maps are fitted on validation data and can use signal-to-noise ratio (SNR) to adapt their decisions to signal quality. This conditional form requires SNR at prediction time. In a post-hoc study across five public benchmarks, we compare these readouts on matched hard-label models and models trained with SNR-dependent soft targets. On the two related RadioML2016 benchmarks, independently fitted bin maps improve mean accuracy by about ten percentage points without retraining the network. This gain occurs in a validation-defined transition between low and high baseline accuracy. Sharing parameters across bins adds little on average. Gains are smaller on other benchmarks, and some conditional maps perform worse than a global map that does not use SNR. Equal readout treatment also leaves some training advantages intact outside the transition band. Simulated SNR errors can erase the readout gains. These results identify substantial room to improve trained modulation classifiers, while showing why training improvements should be evaluated against fitted output rules with their information requirements made explicit.

1 INTRODUCTION

Wireless transmitters carry information by changing a signal's amplitude, frequency, or phase. The rule governing these changes is its modulation format. Automatic modulation classification (AMC) identifies an unknown format from received signals, helping a receiver interpret the transmission (Dobre et al., 2007). Neural networks can learn this task directly from signal samples (O'Shea et al., 2016). Their task becomes harder when noise hides the patterns that distinguish one format from another. Signal-to-noise ratio (SNR) measures signal power relative to noise power.

A dataset can give a precise class label even when the received signal provides weak evidence. A simulator chooses a modulation format, generates a waveform, and adds noise (O'Shea et al., 2016). The label still records the chosen format, but the receiver may no longer be able to identify it with confidence. Ordinary hard-label training assigns all target probability to the recorded class, regardless of SNR. An alternative is to vary the weight of that class and plausible alternatives with signal quality. We ask whether such an adjustment can improve a network's predictions from noisy signals.

Existing work provides several ways to change supervision. Label smoothing mixes class labels with a uniform distribution and has improved generalization and calibration in studied settings (Müller et al., 2019). Distillation uses a teacher's probabilities to convey similarities between classes that a single label does not express (Hinton et al., 2015). In AMC, FR-AMC addresses prediction ambiguity through fuzzy regularization with adaptive loss scaling (Liang et al., 2025). These approaches motivate changing how a network learns from ambiguous signals. They act in different ways, however, and neither softer targets nor sharper predictions alone establish better recognition.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We use a fixed SNR-dependent soft-target scheme, DPC, as a training comparison. It mixes the recorded label with teacher probabilities and a uniform distribution, giving the latter components greater weight at lower SNR. For an unrestricted predictor, expected hard-label cross-entropy is minimized by the true class probabilities given its inputs (Gneiting & Raftery, 2007). Any benefit from changing these targets must therefore be tested with finite data and the chosen networks. We assess both classification accuracy and probability quality. The former evaluates the selected class; the latter assesses the full predicted distribution.

The usual decision selects the class with the largest predicted probability. These probabilities can also be adjusted after training. A transformation fitted on validation data can change their values and the predicted class, while keeping the original network fixed (Kull et al., 2019). We call this output rule a *readout*. A readout may additionally use SNR to adapt its decisions to signal quality, which requires SNR at prediction time. This provides another route to improving predictions without retraining the network. We first ask how much accuracy a frozen network can gain through readout, and where these gains occur. We then ask which training differences remain when both models receive the same readout procedure.

We measure readout gains and remaining training differences on five public modulation benchmarks. Matched hard-label and DPC models receive the same readout families and validation splits, with readout parameters fitted separately for each model. We evaluate the full SNR range and two regions defined by baseline validation accuracy: a middle transition band and an upper plateau. We report classification accuracy, probability scores, and sensitivity to simulated SNR errors. This is a post-hoc study of existing models and predictions.

On the two RadioML2016 benchmarks, SNR-conditioned affine readout improves mean transition-band accuracy by about ten percentage points with the hard-label networks fixed. Independently fitted bin maps already give this gain; sharing their parameters adds little on average. Other benchmarks show smaller gains and some losses. DPC remains useful in a different part of the comparison: its average all-SNR and plateau accuracy advantages persist after equal readout treatment, while most transition-band differences remain uncertain. Global readout, without SNR, also reduces DPC's average transition-band probability-score disadvantage. Simulated SNR errors can erase the conditional readout gains over the original models. These results show substantial scope to improve predictions after network training, while identifying the benchmark and SNR conditions on which that benefit depends.

2 Comparing Training and Readout

We first measure frozen-readout gains, then the training differences that remain. Each backbone has a hard-label model (H) and a matched DPC model (D). Both receive the same readout procedure, fitted separately.

Three comparisons. Let $A(a, f)$ denote accuracy for training method a and readout f on common samples. Relative to raw outputs, the readout gain is

$$G_a(f) = A(a, f) - A(a, \mathrm{raw}). \tag{1}$$

The training difference and readout interaction are

$$\begin{aligned} D(f) &= A(D, f) - A(H, f), \\ I(f) &= G_D(f) - G_H(f) = D(f) - D(\mathrm{raw}). \end{aligned} \tag{2}$$

Positive $D(f)$ means DPC is more accurate after equal readout treatment; positive $I(f)$ means it gains more from readout. An existing advantage can give $D(f) > 0$ without $I(f) > 0$. Accuracy differences use percentage points. For probability scores, the same contrasts are favourable when negative.

Equal treatment and information. Both models share evaluation samples, validation folds, readout families and selection rules. Global readouts need only the signal at prediction time; conditional readouts also need SNR. Their comparison changes both information and parameter structure, not SNR alone. All models use the hard model's validation-defined regions within each seed (Section 4).

3 Training Targets and Readout Models

SNR-dependent training comparison. DPC is our fixed training comparison. It mixes the class label with teacher probabilities and a uniform distribution. The combined soft weight is $0.7(1 - r)$ for $r < 0.8$, otherwise zero; $r \in [0, 1]$ normalizes SNR by the dataset range. The paired hard model predicts its own training samples, giving in-sample self-distillation. Appendix G specifies all fixed constants. The trained student needs only the signal at prediction time.

Global and separate readouts. Let $p_a(x)$ be a model's exported probabilities over K classes. Our affine readouts use $u_a(x) = \log \mathrm{clip}(p_a(x), 10^{-12}, 1)$. These are log probabilities, not native logits or intermediate features. The common raw reference is $\mathrm{softmax}(u_a)$, which normalizes the clipped probabilities. An affine readout returns $\mathrm{softmax}(Wu_a + b)$ (Kull et al., 2019). A global readout fits one W, b for all SNRs. An independent readout fits W_s, b_s separately for each SNR bin s, allowing different output adjustments at different noise levels.

Sharing across SNR. Independent readouts estimate $K^2 + K$ parameters per bin. We test whether encouraging similar maps improves estimation. Let $q_s = N_s/N$ be the sample fraction in bin s of the current fitting subset. The weighted centres are $W_0 = \sum_s q_s W_s$ and $b_0 = \sum_s q_s b_s$. These centres are computed from the bin maps, not fitted separately. Define $R(A, b) = \|A\|_F^2/K^2 + \|b\|_2^2/K$ and the score loss $\ell(z, y) = \mathrm{logsumexp}(z) - z_y$. The shared readout minimizes

$$\frac{1}{N}\sum_i \ell(W_{s_i}u_i + b_{s_i}, y_i) + 0.01R(W_0 - I_K, b_0) + \lambda \sum_s q_s R(W_s - W_0, b_s - b_0). \tag{3}$$

The penalties keep the centre near the identity map I_K and limit differences between bins.

Global and independent readouts instead use penalties $\lambda R(W - I_K, b)$ and $\lambda \sum_s q_s R(W_s - I_K, b_s)$. Each custom family selects $\lambda \in \{0.001, 0.01, 0.1, 1\}$. At $\lambda = 0.01$, the independent and shared objectives are identical by the weighted variance identity. Each bin retains a full affine map. A fourth family adds adjacent-bin smoothing, scaled by actual dB spacing, with strength fixed at 0.01 and the same sharing-penalty grid (Appendix A).

Structured Matrix Scaling. We also use the official Structured Matrix Scaling (SMS) method (Berta et al., 2026), globally and separately within each SNR bin. Both use `probmetrics==1.3.0`, with original exported probabilities and the official preprocessing and fixed default fitting rule. SMS does not use our four-candidate search, so the comparison does not equalize selection budgets. These two SMS variants and the four affine families form the complete readout comparison.

4 Experimental Design

Models and data. We use five public benchmarks and 57 benchmark–backbone combinations, called cells (Table 1). Each cell has matched hard-label and DPC models from three seeds, giving 342 checkpoints. These are all complete, usable pairs in the 69-cell inventory. This inventory excludes the denoising autoencoder (DAE) backbone post hoc (Appendix H). Each pair shares its backbone, data split, optimizer, batch size and student-training budget. DPC requires additional teacher preparation and prediction export, so total training cost is not matched. Dataset sources and the AMR-Benchmark implementation are documented in Appendix F.

Validation and selection. Nested validation assesses each custom readout after selecting its penalty. We use three outer folds. Within each outer fitting subset, we create a fresh three-fold inner split for penalty selection. The criterion is negative log-likelihood (NLL), averaged over all inner holdout samples across all SNRs. Among candidates within 10^{-6} of the best NLL, we choose the strongest penalty. Splits are stratified by class and SNR; sample-identifier hashes keep the folds identical across training methods. Each selected readout is refitted on its outer fitting subset and evaluated on the outer holdout.

Table 1: Complete study coverage. K is the number of classes; sample counts are per checkpoint. Every cell has three seeds and both training methods. All checkpoints receive all six readout families. Models within a paired seed share samples and evaluation regions.

Benchmark	Cells	K	N_{val}	N_{test}
2016.10B	13	10	240,000	240,000
2016.10A	13	11	44,000	44,000
2018.01A	12	24	511,056	511,056
RML22	8	10	84,000	84,000
HisarMod	11	26	104,000	260,000

For test evaluation, we repeat selection using the predefined three folds of the complete validation set, then fit on all validation samples. We freeze configurations and parameters before evaluating test predictions. SMS uses the same outer folds with its fixed default rule, followed by a full-validation fit. Optimizer settings, failure handling and diagnostics are retained in Appendix A.

Regions and metrics. Each hard model's validation curve defines the evaluation regions for its paired seed. The primary transition band contains bins within 20–80% of the curve's observed accuracy range. The plateau contains bins above that relative 80% threshold. These are neither absolute accuracy cutoffs nor common dB thresholds. All-SNR evaluation includes every test bin. The selected band varies across seeds in 29 of 57 cells. Exact rules and fallbacks are given in Appendix F.

We report accuracy, NLL and Brier score in all three regions. Brier sums squared probability errors over classes, without dividing by K. Each metric averages over samples within the region. Effects are then averaged equally over seeds within a cell, and over cells within a benchmark or the full grid. The transition band remains the original primary region; all-SNR and plateau evaluations provide additional comparisons. SNR-error experiments cover all RadioML2016.10B cells under the specified Gaussian error model (Section 5.5).

Uncertainty and study status. Cell-level 95% confidence intervals use paired effects from three seeds and Student's t with two degrees of freedom. The intervals are not adjusted for multiple comparisons. Benchmark and grid means are descriptive. For shared-minus-independent transition-band accuracy, we report Holm-adjusted p values separately within the 57 cells of each training arm. Three post-hoc comparisons between regions are defined for each cell. Their p values are adjusted together by Holm across all 171 tests (Appendix A.1). Other comparisons remain unadjusted. The intervals include variation in validation-selected regions across seeds. Differences between regions are not an additive decomposition of total accuracy.

This study was informed by earlier test results, and validation data had already been used to select the checkpoints. Nested fitting controls readout selection for these fixed models; it does not provide a new, independent confirmation. The earlier preregistered comparisons are reported separately in Appendix C.

5 Results

5.1 Large readout gains without retraining the network

The largest readout gains occur on the two RadioML2016 benchmarks. Independently fitting an affine map in each SNR bin raises hard-model transition-band accuracy by 11.30 pp on 10B and 9.39 pp on 10A, averaged over their backbones. All 26 cells have positive nominal paired intervals. The networks and their exported probabilities are fixed throughout this comparison. The improvement comes from fitting the output maps on validation data and supplying SNR at prediction time.

Shared readout gives a similar result: 11.26 and 9.46 pp for hard models on 10B and 10A. We retain this fixed family for the main figures and report all six families in Table 2; no family is selected separately for each test cell. Figure 1a shows shared-minus-raw gains for all 57 cells and both training methods. The positive nominal intervals extend to all 26 RadioML2016 cells in both arms.

The other benchmarks give smaller hard-model means: 2.31 pp on 2018, 1.04 on RML22 and 0.80 on HisarMod. A large gain on RadioML2016 therefore does not describe every benchmark in the study.

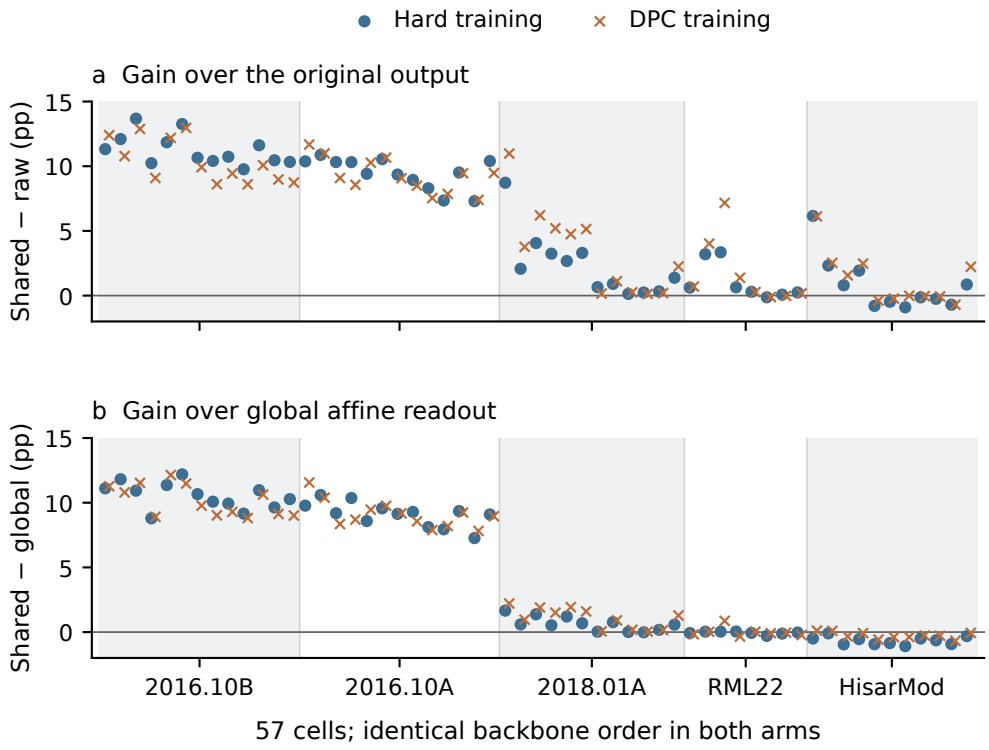


Figure 1: Accuracy gains from a frozen network's readout. (a) Shared-minus-raw transition-band accuracy. (b) Shared-minus-global affine accuracy on the same samples. Both arms use the same benchmark–backbone order for all 57 cells. Points are three-seed means; complete nominal 95% intervals accompany the data. The spread across cells is not population uncertainty. Conditional maps require inference-time SNR, whereas global maps do not. Figure 4 compares shared and independent maps.

Global affine also adjusts probabilities but uses one map without SNR. Relative to this reference, the 26 RadioML2016 cells gain an average of 9.81 pp for hard models and 9.61 pp for DPC (Figure 1b). All their nominal intervals are positive. Across the remaining 31 cells, the means are -0.01 and $+0.32$ pp. The comparison changes both the available information and the map's parameter structure, so it does not isolate the contribution of SNR alone.

HisarMod makes the boundary concrete. Shared readout is worse than global affine by 0.67 pp for hard models and 0.26 pp for DPC on average. Ten of its eleven hard-model intervals for this comparison are negative. Five hard-model cells also have negative nominal gains relative to raw predictions. More flexible, SNR-dependent maps can therefore make decisions worse. The RadioML2016 pattern does not transfer uniformly to the other benchmarks.

5.2 Additional sharing contributes little to the mean gain

Independent bin maps already produce the large RadioML2016 gains. We test whether coupling those maps adds further value. Over all 57 cells, the mean difference is only $+0.0196$ pp for hard models and $+0.0220$ pp for DPC. Nominal intervals include both improvements and losses. Holm correction retains one hard-model improvement and one DPC loss (Table 3). The experiment does not establish a consistent additional accuracy benefit from sharing. Adding adjacent-bin smoothing changes the mean by a further -0.0113 and -0.0169 pp, respectively.

Table 2: All readout families, averaged equally over the 57 cells. Panel A reports transition-band readout gains G for both training methods and DPC-minus-hard accuracy D in all three regions, in pp. Panel B reports probability-score differences and readout interactions I. Negative NLL and Brier differences favour DPC. Raw predictions have zero readout gain and no readout interaction. These are descriptive means; the complete cell intervals are retained in the supplement.

A. Accuracy differences (pp)

Readout	G_H band	G_D band	D band	D all	D plateau
Raw	+0.00	+0.00	+0.14	+0.71	+1.71
Global	+1.04	+1.10	+0.20	+0.85	+1.69
Independent	+5.49	+5.63	+0.28	+0.95	+1.84
Shared	+5.51	+5.66	+0.28	+0.93	+1.81
Shared + smooth	+5.50	+5.64	+0.28	+0.93	+1.81
SMS global	+0.92	+0.98	+0.21	+0.84	+1.69
SMS per SNR	+5.13	+5.02	+0.03	+0.75	+1.71

B. Probability-score differences (lower is better)

Readout	$D^{\text{band}}_{\text{NLL}}$	$D^{\text{all}}_{\text{NLL}}$	$D^{\text{band}}_{\text{Brier}}$	$D^{\text{all}}_{\text{Brier}}$	$I^{\text{band}}_{\text{NLL}}$	$I^{\text{band}}_{\text{Brier}}$
Raw	+0.0379	+0.0046	+0.0032	-0.0063	—	—
Global	-0.0019	-0.0126	-0.0029	-0.0097	-0.0398	-0.0061
Independent	-0.0032	-0.0157	-0.0023	-0.0105	-0.0412	-0.0055
Shared	-0.0036	-0.0159	-0.0023	-0.0104	-0.0415	-0.0055
Shared + smooth	-0.0038	-0.0159	-0.0024	-0.0104	-0.0417	-0.0056
SMS global	-0.0006	-0.0116	-0.0028	-0.0095	-0.0385	-0.0060
SMS per SNR	+0.0078	-0.0082	+0.0012	-0.0081	-0.0302	-0.0020

Official per-SNR SMS also improves mean transition-band accuracy over raw outputs: 5.13 pp for hard models and 5.02 pp for DPC across the grid. Shared readout exceeds these means by 0.38 and 0.64 pp. These are comparisons between complete pipelines with different selection budgets. Since independent affine gives nearly the same mean gain as shared affine, the SMS comparison cannot assign an advantage to parameter sharing.

The search range also matters. On HisarMod, shared and smooth readouts choose the largest allowed penalty, $\lambda = 1$, at every checkpoint in both arms. The tested penalties do not repair the unfavourable comparison with global affine there. Selection at the grid boundary leaves open what a different range would do. We retain the complete tested comparison, without extending the search after seeing the test outcomes. These results support the readout finding without establishing a new best estimator.

5.3 Equal readout treatment leaves some training benefits intact

A stronger frozen-model baseline changes how training gains should be judged. It does not by itself show that another training method has no value. We therefore give hard and DPC models the same readout procedure and compare their predictions in each evaluation region.

Before readout, DPC-minus-hard averages $+0.14$ pp in the transition band, $+0.71$ pp over all SNRs and $+1.71$ pp on the plateau. After shared readout, the corresponding means are $+0.28$, $+0.93$ and $+1.81$ pp (Table 2). In the primary transition band, 47 of 57 after-readout intervals include zero; six favour DPC and four favour hard training. Most cells therefore remain uncertain. Over all SNRs and on the plateau, twenty cells have positive nominal intervals in each region, with one and zero negative intervals, respectively.

Most of the plateau advantage is already present before output adjustment. Its mean rises from 1.71 to 1.81 pp, giving a direct interaction of only $+0.10$ pp. The plateau interaction has one positive and five negative nominal intervals. Training advantages can remain even when both models benefit similarly from readout.

Across six readouts, grid-mean DPC advantages are 0.75–0.95 pp over all SNRs and 1.69–1.84 pp on the plateau (Figure 2a). These are correlated transforms of the same models, not independent replications.

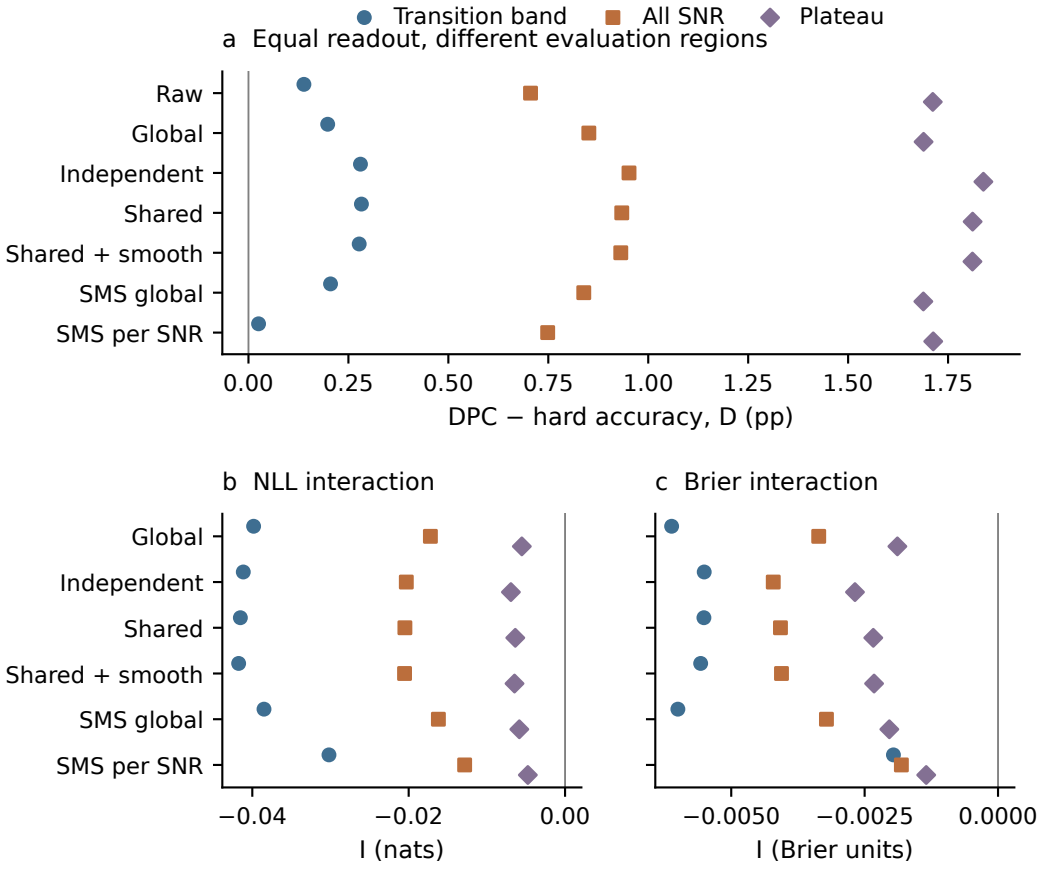


Figure 2: Training differences after equal readout treatment. (a) DPC-minus-hard accuracy for raw predictions and all six readout families in the transition band, all-SNR set and validation-defined plateau. (b,c) Direct readout interactions in NLL and Brier for the fitted families; negative values mean that DPC's score improves more than hard's. Global maps require no inference-time SNR. All displayed means describe the same 57 cells. Probability scores and accuracy have different units and directions of benefit.

We also compare the two regions within each seed. Under shared readout, DPC's mean advantage is 1.53 pp larger on the plateau than in the transition band. No cell in that training-region contrast survives the joint 171-test Holm correction. Its regional difference remains a descriptive and nominal-interval result, not evidence for separate learning mechanisms. Appendix A.1 retains every regional contrast, including the band-minus-plateau readout gains in both arms.

Benchmark-level changes also need this distinction. On 10B, the transition-band DPC difference moves from $+0.62$ pp before readout to -0.29 pp after shared readout. On 2018 it moves from -0.79 to $+0.25$ pp, yet every after-readout cell interval includes zero. A change in the mean's sign is not an established reversal.

5.4 Global readout can reduce a probability-score disadvantage

A global map changes the probability-score comparison even without SNR. In the transition band, raw DPC predictions have worse mean NLL than hard predictions by 0.0379 nats. After global affine, the difference is -0.0019 nats; after shared affine, it is -0.0036. Their direct NLL interactions are -0.0398 and -0.0415 nats (Figure 2b). Twenty-nine nominal cell intervals for the global NLL interaction are negative; none is positive. A map that receives no SNR can therefore remove much of DPC's average relative NLL disadvantage in this region.

Brier supports the need to examine more than one score. Its mean DPC-minus-hard difference changes from $+0.0032$ for raw predictions to -0.0029 under global affine and -0.0023 under shared affine. The global map gives the larger average correction in Brier (Figure 2c). Extra SNR conditioning does not improve every probability comparison, despite its large accuracy benefit on RadioML2016. Nor does every readout remove the average disadvantage: per-SNR SMS leaves transition-band differences of $+0.0078$ in NLL and $+0.0012$ in Brier.

These statements concern the transition band. Over all SNRs, raw DPC already has better mean Brier, and shared-readout DPC has better means in both scores (Table 2). DPC cannot be described as having poorer probabilities everywhere. The practical distinction is that global output adjustment can change probability-score comparisons substantially, whereas the large extra conditional accuracy gains are concentrated on RadioML2016. The scores do not identify whether calibration, class-specific errors or feature changes explain either result.

5.5 SNR ERRORS CAN ERASE THE READOUT GAINS

The large RadioML2016 gains require usable SNR information at prediction time. We test this condition on all thirteen 10B cells at the thirteen fixed Gaussian error levels. Hard routing selects the nearest SNR bin; soft routing mixes bin probabilities under the Gaussian error model and a uniform bin prior. Both training methods receive the same perturbations. Figure 3 separates each method's gain over its own raw predictions from the DPC-minus-hard difference.

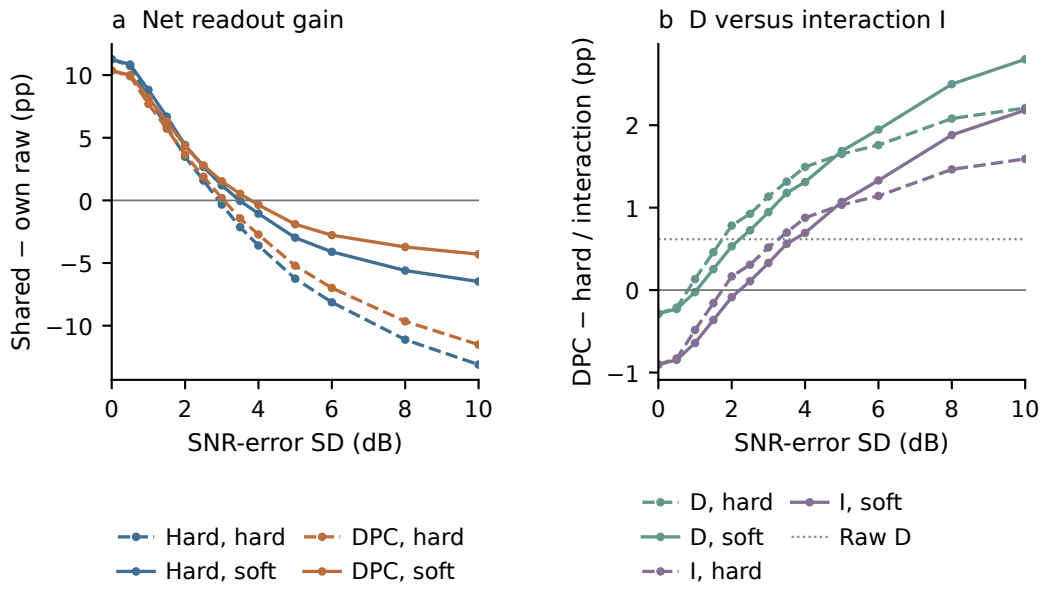


Figure 3: SNR errors affect net gains and relative advantages differently. Shared readout is evaluated at all thirteen specified Gaussian error levels. Net gains use each training method's own raw reference; D and I compare the same paired models and perturbations. Means cover all 39 model–seed pairs in their hard-validation transition bands. Lines connect sampled levels; they do not estimate continuous thresholds. Dashed and solid lines denote hard and soft routing, and the grey dotted line is raw D. A positive relative advantage can remain when both net gains are negative.

At error SD 2 dB, hard routing retains mean shared-readout gains of 3.48 pp for hard models and 3.65 pp for DPC. Soft routing retains 4.41 and 4.33 pp. These are useful net gains within the specified simulation. The hard-route DPC advantage is 0.78 pp, but its raw advantage was already 0.62 pp. The direct interaction is only 0.17 pp, with no positive nominal cell interval. The after-readout training difference alone would give an incomplete account of complementarity.

At larger errors, relative advantage can hide a loss for both methods. At 6 dB with hard routing, DPC exceeds hard by 1.76 pp, while hard and DPC lose 8.13 and 6.99 pp relative to their own raw

outputs. All thirteen nominal net-gain intervals are negative in both arms. Here the larger DPC advantage describes less damage; neither method benefits from adding this readout under the tested conditions.

For hard models, the first sampled nonpositive mean net gain is at 3.0 dB with hard routing and 3.5 dB with soft routing. For DPC, these sampled levels are 3.5 and 4.0 dB. All four mean net gains are negative at every measured level from 4 dB onward. Soft routing reduces sensitivity but does not remove the dependence on SNR quality. These sampled values apply to this benchmark and error model; they do not validate a physical SNR estimator or give a universal tolerance.

6 Related Work

Probability calibration provides several ways to adjust a trained classifier's outputs. Temperature and vector scaling (Guo et al., 2017), affine maps of log probabilities (Kull et al., 2019), and Structured Matrix Scaling (Berta et al., 2026) supply relevant readout families. Class-frequency logit adjustment instead addresses long-tail recognition (Menon et al., 2021). Our contribution is the empirical finding about conditional readout in AMC, its limits, and its effect on training comparisons; it is not the affine transformation itself. Proper scores assess probability quality alongside the resulting class decisions (Gneiting & Raftery, 2007).

Label smoothing and distillation motivate the fixed training comparison (Müller et al., 2019; Hinton et al., 2015). AMC methods also adapt training or network processing to signal quality (Liang et al., 2025; Wang et al., 2026; Wu et al., 2026). We examine what remains after both training arms receive equal readout treatment. A separate matched FR-AMC study provides further context under its own protocol (Appendix J); it does not reproduce published numbers or establish the same conclusions for all training methods. Appendix K gives extended context.

7 Discussion and Limitations

The RadioML2016 results make fitted output rules a useful reference when evaluating an additional training method. A global map needs no SNR; a conditional map can offer larger gains when reliable SNR is available. Validation data should guide that choice. Comparing both trained models under equal readout treatment then reveals which training advantages remain. Reporting each model's net gain also prevents a relative advantage from concealing losses for both models. These comparisons concern the complete prediction rule, not changes in internal representations.

The evidence remains conditional on the studied grid. The analysis is post hoc, the two RadioML2016 benchmarks are related, and three seeds give limited precision. DAE exclusion and cell highlighting occurred after results were visible; the exclusion favours the original training comparison. The old ladder's eleven nominal gains include seven reproduced by readout and four that remain, but none passes Bonferroni or Holm correction. Remaining gains do not establish feature changes. Full sensitivity results appear in Appendix H. The Gaussian SNR-error study does not validate a real SNR estimator. Fresh test data and measured SNR estimates are needed to assess transfer beyond these conditions.

8 Conclusion

Frozen SNR-conditioned readout yields large accuracy gains on RadioML2016 without retraining the network. Independent affine maps already show this effect; additional sharing adds little on average. Equal readout treatment still leaves some broader-region training advantages intact. The gains vary across benchmarks and depend on SNR quality. Evaluating fitted output rules alongside training methods therefore reveals both an opportunity to improve existing classifiers and the conditions under which that opportunity is useful.

Ethics Statement

This study uses public simulated-radio benchmarks and collects no human-subject data or personally identifiable information. It studies evidence standards for recognition under degraded observations. Complete results, including failures and negative outcomes, are retained to reduce selective reporting.

AI Use Statement

Generative AI tools assisted in refining follow-up hypotheses and experimental comparisons, checking mathematical claims, implementing methods and experiment adapters, monitoring execution, recomputing statistics, interpreting results, preparing scientific figures, drafting and editing the manuscript, translating the Chinese reading version, verifying references, and checking artifacts. Numerical checks used tests, separately implemented statistical calculations and source hashes; AI-assisted cross-checking does not substitute for author review. The authors remain responsible for verifying the code, analyses, claims and artifacts and for all content submitted for review.

Reproducibility Statement

The main study is a post-hoc comparison of existing checkpoints and predictions. The supplement contains authenticated unit metrics, complete comparison reports, analysis and figure code, and instructions for rebuilding the tables and figures. Appendix A specifies fitting and evaluation; Appendix F gives the data and region rules. The earlier preregistered ladder is reported separately in Appendix C, with its original decisions and deviations. Reaggregating packaged metrics does not rerun training or authenticate omitted prediction arrays; the full-run requirements are documented separately.

References

Eugène Berta, David Holzmüller, Michael I. Jordan, and Francis Bach. Structured matrix scaling for multi-class calibration. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1342–1350. PMLR, 2026. URL https://proceedings.mlr.press/v300/berta26a.html.

Octavia A. Dobre, Ali Abdi, Yeheskel Bar-Ness, and Wei Su. Survey of automatic modulation classification techniques: classical approaches and new trends. *IET Communications*, 1(2):137–156, 2007. doi: 10.1049/iet-com:20050176. URL https://digitalcommons.njit.edu/fac_pubs/13459/.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894. URL https://doi.org/10.1109/TNNLS.2013.2292894.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2204.02937.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Xinyan Liang, Ruijie Sang, Yuhua Qian, Qian Guo, Feijiang Li, and Liang Du. Robust automatic modulation classification with fuzzy regularization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37396–37408. PMLR, 2025. URL https://proceedings.mlr.press/v267/liang25o.html.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324. URL https://doi.org/10.1109/ICCV.2017.324.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

Timothy J. O'Shea, Johnathan Corgan, and T. Charles Clancy. Convolutional radio modulation recognition networks. *arXiv preprint arXiv:1602.04105*, 2016. URL https://arxiv.org/abs/1602.04105.

Timothy J. O'Shea, Tamoghna Roy, and T. Charles Clancy. Over-the-air deep learning based radio signal classification. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):168–179, 2018. doi: 10.1109/JSTSP.2018.2797022. URL https://arxiv.org/abs/1712.04578.

Venkatesh Sathyanarayanan, Peter Gerstoft, and Aly El Gamal. RML22: Realistic dataset generation for wireless modulation classification. *IEEE Transactions on Wireless Communications*, 22(11): 7663–7675, 2023. doi: 10.1109/TWC.2023.3254490. URL https://ieeexplore.ieee.org/document/10070586/.

Kursat Tekbiyik, Ali Riza Ekti, Ali Gorcin, Gunes Karabulut Kurt, and Cihat Kececi. Robust and fast automatic modulation classification with CNN under multipath fading channels. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1–6, 2020. doi: 10.1109/VTC2020-Spring48590.2020.9128408. URL https://arxiv.org/abs/1911.04970.

Run Wang, Jizhe Li, Youze Yang, Shasha Wang, and Bing Zheng. KADNet: Low SNR automatic modulation classification via SNR aware deformable convolution and Kolmogorov–Arnold networks. *Digital Signal Processing*, 174:105942, 2026. doi: 10.1016/j.dsp.2026.105942. URL https://www.sciencedirect.com/science/article/pii/S105120042600062X.

Teng Wu, Quan Zhu, Runze Mao, Changzhen Hu, and Shengjun Wei. CAIC-Net: Robust radio modulation classification via unified dynamic cross-attention and cross-signal-to-noise ratio contrastive learning. *Sensors*, 26(3):756, 2026. doi: 10.3390/s26030756. URL https://www.mdpi.com/1424-8220/26/3/756.

Mert Yüksekgönül, Linjun Zhang, James Zou, and Carlos Guestrin. Beyond confidence: Reliable models should also consider atypicality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38420–38453, 2023. doi: 10.52202/075280-1669. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7900318ffaf5e9bc60250f134c6cc3c7-Abstract-Conference.html.

Fuxin Zhang, Chunbo Luo, Jialang Xu, Yang Luo, and Fu-Chun Zheng. Deep learning based automatic modulation recognition: Models, datasets, and challenges. *Digital Signal Processing*, 129:103650, 2022. doi: 10.1016/j.dsp.2022.103650. URL https://www.sciencedirect.com/science/article/abs/pii/S1051200422002676.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

A Complete Frozen-Readout Protocol and Results

This section records the post-hoc follow-up separately from the original preregistered ladder. It uses the complete 57-cell cohort and all six readout families in both arms. Numerical results come from the completed validation, test and routing reports. The supplementary package retains every cell and metric, including negative outcomes and uncertainty, without choosing a winning readout family separately for each test cell.

Penalty identities and spacing. For B fitted SNR bins, let $d_s = (\mathrm{SNR}_{s+1} - \mathrm{SNR}_s)/(2\,\mathrm{dB})$. The smooth penalty is

$$P_{\mathrm{smooth}} = P_{\mathrm{shared}} + \frac{0.01}{B-1} \sum_{s=1}^{B-1} \frac{R(W_{s+1} - W_s, b_{s+1} - b_s)}{d_s^2},$$

with no adjacency term for $B = 1$. Counts, weights and centres are recalculated within every fitting subset, with no use of a held-out fold. Since $\sum_s q_s(W_s - W_0) = 0$ and $\sum_s q_s(b_s - b_0) = 0$,

$$\sum_s q_s R(W_s - I, b_s) = R(W_0 - I, b_0) + \sum_s q_s R(W_s - W_0, b_s - b_0).$$

This proves equality of independent and shared objectives at $\lambda = 0.01$. All finite candidates retain per-bin affine capacity. The search does not include an infinite-sharing candidate, and no result is inferred for that limit.

Nested fit counts. A custom unit contains three outer folds, each with 4×3 inner candidate fits and one selected refit, followed by 4×3 full-validation selection fits and one final refit: $3(12+1) + 12 + 1 = 52$ fits. All 1,368 custom units completed, comprising 71,136 fits. The 684 SMS units use three outer calls and one final call, totalling 2,736 wrapper calls. Per-SNR wrappers contain several official pipelines, so wrapper calls are not counts of individual optimizer fits and their cost is not directly comparable to the custom search.

Custom objectives and gradients were evaluated in TensorFlow float64 on H100 GPUs and optimized with SciPy L-BFGS-B. The cumulative limit was 3,000 iterations, in stages of at most 500. Fisher coordinate changes between stages left the physical objective and current parameters

unchanged. Settings were `maxcor=200`, `ftol=1e-12`, `gtol=1e-8` and `maxls=50`. The final custom environment used TensorFlow 2.17, NumPy 1.26.4 and SciPy 1.17.1 over the NVIDIA TensorFlow 25.02 container. Independent NumPy/SciPy checks covered objectives, gradients, extreme scores, shared-centre constraints and predictions. Failures in earlier engineering attempts are archived separately from the complete final execution.

SMS ran on CPU in a separate environment with `probmetrics==1.3.0`, NumPy 2.2.6 and SciPy 1.15.3. Its official pipeline returned successfully for every final unit. This records an official return, not an exposed optimizer convergence certificate. Both environments, dependency locks, package hashes, per-fit diagnostics and timing fields are preserved. H100 GPUs accelerated custom computation; SMS is not reported as GPU-accelerated.

Probability and metric numerics. The common raw reference is obtained by normalizing probabilities after clipping at 10^{-12}. Custom readout NLL is evaluated from scores by stable log-sum-exp. SMS consumes original exported probabilities and retains its official numerical rules; reported SMS NLL clips target probabilities at 10^{-300}. Accuracy is unaffected by the custom input clipping's monotone transformation except for possible ties at the clipping floor. The old ladder retains its historical probability-loss implementation, including the possible loss/gradient mismatch at extreme clipping. Its outputs were not overwritten by the new stable objective. Differences between old and new readouts therefore cannot be assigned to sharing alone.

Within an evaluation region R, every metric averages over its samples. In particular, the multiclass Brier score is

$$\mathrm{Brier}_R = \frac{1}{|R|} \sum_{i \in R} \sum_{k=1}^{K} \big(p_{ik} - \mathbb{1}[y_i = k]\big)^2,$$

without division by K. Accuracy and NLL use the same sample weights. Paired effects are then averaged equally over the three seeds within a cell; benchmark and grid summaries weight cells equally, not by their sample counts.

Fold isolation and interpretation. Fold identities depend on dataset, seed, stage and sample identifier, excluding arm, model predictions and outcome metrics. Inner folds are constructed anew within each outer training subset; outer holdouts cannot enter their fits. Final selection uses the predefined full-validation folds and does not select on outer test scores. Candidate failure makes the family unavailable, rather than permitting selection from a reduced grid. The complete final execution has no unavailable family unit. This workflow addresses readout selection conditional on previously selected checkpoints and the existing research history; it is explicitly not a new independent confirmation.

Full primary comparison. Table 3 reports every shared-minus-independent cell interval. Holm is applied separately to the 57 hard-arm and 57 DPC-arm tests, only for this primary band-accuracy comparison. No correction is implied across arms, metrics, stages, other readout contrasts or SNR-error levels. The package provides all 17,955 cell-comparison rows for each validation/test stage, including all-SNR and plateau results and proper scoring rules. A separate implementation recalculated all test means, paired intervals and p values from the 2,052 authenticated unit results and matched the reported statistics up to floating-point rounding. For Figure 1a, the figure data export all $57 \times 2 = 114$ shared-minus-raw comparisons, with the three paired seed effects and existing nominal intervals. Panel (b) shows the 114 shared-minus-global comparisons from the supplementary regional analysis. Both panels retain the existing unadjusted p values in the figure data. These comparisons are separate from the 171-test regional Holm family and from the two 57-test shared-minus-independent families. The latter comparison is shown in Figure 4 and Table 3.

Table 3: Shared minus independent affine: all 57 cells. Values are percentage points with paired 95% intervals over three seeds. $*$ marks a Holm-adjusted $p < 0.05$ within the separate 57-cell family for that arm.

Benchmark	Backbone	Hard	DPC
2016.10B	1dcnnpf	-0.14 [-0.28, +0.00]	-0.06 [-0.10, -0.01]
2016.10B	cgdnet	-0.05 [-0.17, +0.07]	-0.20 [-0.23, -0.18]$*$
2016.10B	cldnn	-0.04 [-0.16, +0.08]	-0.18 [-0.43, +0.06]

Benchmark	Backbone	Hard	DPC
2016.10B	cldnn2	-0.04 [-0.22, +0.14]	-0.05 [-0.17, +0.07]
2016.10B	cnn1	-0.01 [-0.03, +0.02]	-0.13 [-0.27, +0.02]
2016.10B	cnn2	-0.06 [-0.19, +0.07]	-0.05 [-0.19, +0.09]
2016.10B	densenet	-0.00 [-0.22, +0.22]	-0.05 [-0.21, +0.11]
2016.10B	gru2	-0.05 [-0.15, +0.05]	-0.04 [-0.09, +0.01]
2016.10B	lstm2	-0.01 [-0.13, +0.11]	-0.03 [-0.11, +0.06]
2016.10B	mcldnn	-0.07 [-0.12, -0.02]	-0.02 [-0.07, +0.04]
2016.10B	mcnet	-0.01 [-0.18, +0.16]	-0.16 [-0.30, -0.03]
2016.10B	petcgdnn	-0.05 [-0.11, +0.01]	-0.08 [-0.13, -0.03]
2016.10B	resnet	+0.02 [-0.03, +0.06]	-0.06 [-0.08, -0.04]
2016.10A	1dcnnpf	-0.04 [-0.25, +0.17]	+0.09 [-0.44, +0.63]
2016.10A	cgdnet	-0.01 [-0.21, +0.18]	+0.06 [-0.54, +0.67]
2016.10A	cldnn	-0.10 [-0.51, +0.32]	+0.09 [-0.22, +0.39]
2016.10A	cldnn2	+0.38 [-0.30, +1.07]	+0.02 [-0.09, +0.14]
2016.10A	cnn1	+0.22 [-0.19, +0.63]	+0.03 [-0.56, +0.61]
2016.10A	densenet	+0.35 [-0.23, +0.94]	+0.11 [-0.17, +0.39]
2016.10A	gru2	-0.18 [-0.87, +0.51]	+0.04 [-0.28, +0.37]
2016.10A	icamcnet	-0.15 [-0.82, +0.52]	+0.05 [-0.22, +0.31]
2016.10A	lstm2	+0.28 [-0.58, +1.13]	-0.03 [-0.24, +0.18]
2016.10A	mcldnn	-0.13 [-1.53, +1.27]	+0.01 [-0.26, +0.28]
2016.10A	mcnet	+0.17 [-0.33, +0.66]	+0.15 [-0.35, +0.65]
2016.10A	petcgdnn	+0.03 [-0.49, +0.54]	+0.15 [-0.01, +0.30]
2016.10A	resnet	+0.14 [-0.17, +0.45]	+0.18 [+0.05, +0.32]
2018.01A	1dcnnpf	-0.14 [-0.33, +0.06]	+0.00 [+0.00, +0.00]
2018.01A	cgdnet	+0.05 [-0.04, +0.14]	-0.13 [-0.22, -0.05]
2018.01A	cldnn	-0.09 [-0.33, +0.15]	-0.07 [-0.36, +0.22]
2018.01A	cldnn2	-0.03 [-0.07, +0.01]	+0.00 [+0.00, +0.00]
2018.01A	cnn1	-0.01 [-0.11, +0.08]	-0.04 [-0.36, +0.28]
2018.01A	cnn2	-0.08 [-0.45, +0.29]	+0.00 [+0.00, +0.00]
2018.01A	gru2	-0.06 [-0.10, -0.02]	+0.05 [-0.07, +0.18]
2018.01A	icamcnet	-0.02 [-0.14, +0.10]	-0.01 [-0.09, +0.07]
2018.01A	lstm2	+0.03 [-0.05, +0.12]	+0.05 [-0.03, +0.13]
2018.01A	mcldnn	-0.03 [-0.12, +0.05]	+0.01 [-0.07, +0.08]
2018.01A	petcgdnn	-0.03 [-0.20, +0.13]	+0.05 [-0.12, +0.22]
2018.01A	resnet	-0.01 [-0.15, +0.12]	+0.02 [-0.02, +0.06]
RML22	1dcnnpf	+0.12 [-0.18, +0.43]	+0.30 [+0.12, +0.49]
RML22	cgdnet	+0.31 [+0.15, +0.47]	+0.09 [+0.06, +0.11]
RML22	cldnn	+0.05 [-0.36, +0.46]	-0.06 [-0.40, +0.28]
RML22	cnn1	-0.11 [-0.64, +0.42]	-0.04 [-0.30, +0.22]
RML22	lstm2	+0.13 [-0.18, +0.44]	+0.14 [+0.10, +0.19]
RML22	mcnet	+0.01 [-0.10, +0.12]	+0.06 [-0.10, +0.23]
RML22	petcgdnn	+0.07 [-0.01, +0.15]	+0.03 [-0.01, +0.07]
RML22	resnet	+0.12 [-0.03, +0.26]	+0.15 [-0.22, +0.52]
HisarMod	cgdnet	+0.00 [-0.06, +0.06]	+0.07 [+0.05, +0.09]
HisarMod	cldnn2	+0.12 [-0.10, +0.35]	+0.35 [+0.06, +0.63]
HisarMod	cnn1	+0.04 [-0.03, +0.11]	+0.14 [+0.04, +0.25]
HisarMod	cnn2	+0.05 [-0.10, +0.21]	+0.11 [-0.11, +0.32]
HisarMod	gru2	-0.09 [-0.15, -0.02]	-0.03 [-0.14, +0.08]
HisarMod	icamcnet	-0.05 [-0.14, +0.05]	-0.01 [-0.07, +0.04]
HisarMod	lstm2	-0.04 [-0.12, +0.03]	-0.01 [-0.31, +0.29]
HisarMod	mcldnn	-0.02 [-0.08, +0.04]	-0.01 [-0.09, +0.07]
HisarMod	mcnet	+0.02 [-0.03, +0.07]	+0.01 [-0.08, +0.11]
HisarMod	petcgdnn	-0.09 [-0.20, +0.02]	-0.12 [-0.19, -0.05]
HisarMod	resnet	+0.45 [+0.41, +0.50]$*$	+0.31 [+0.02, +0.61]

Routing protocol and complete scope. The 10B study contains all 39 model–seed pairs, two arms and six families, giving 468 available family units. Gaussian error SDs are $\{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 8, 10\}$ dB. Perturbation coordinates are paired across arms and readout families using the fixed seed sequence $[0, \mathrm{round}(10\sigma)]$. Hard routing selects the nearest SNR bin. Soft routing uses Gaussian bin weights under a uniform prior and mixes probabilities, rather than scores. Global maps are route-invariant; the zero-error hard and soft predictions coin-

cide. The report retains all 11,830 paired cell comparisons, as well as the original frozen hard/soft routing curves for sensitivity comparison. The historical curves use a different fitting protocol and rounded percentages; their difference from the new curves does not isolate parameter sharing.

Reproduction layers. Default supplementary reproduction authenticates stored unit metrics and reconstructs paired statistics, tables and figures in a separate directory. It does not rerun training, selection or prediction, and external parameter and source-prediction hashes remain provenance rather than a fresh verification of unbundled files. A full readout rerun requires the exported predictions and the recorded GPU/CPU environments. Full backbone reproduction additionally requires the public data, pinned upstream sources, splits and three-stage hard/teacher/DPC commands. These levels are documented separately.

A.1 Regional contrasts and cross-readout consistency

The additional regional analysis uses the frozen per-seed test metrics and does not change region selection, readout fitting or the original primary endpoint. For each of the 57 cells, three contrasts are formed within each canonical seed: shared-readout $D_{\mathrm{plateau}} - D_{\mathrm{band}}$, $G_{H,\mathrm{band}} - G_{H,\mathrm{plateau}}$ and $G_{D,\mathrm{band}} - G_{D,\mathrm{plateau}}$. A paired Student-$t$ interval with two degrees of freedom is then computed from the three regional differences. Holm adjustment uses all $3 \times 57 = 171$ two-sided tests together; it is separate from the two original 57-cell sharing families. Every mean, seed effect, nominal interval, unadjusted and adjusted p value is provided in the machine-readable regional analysis, including negative outcomes.

The mean plateau-minus-band training contrast is $+1.5289$ pp, with 15 positive and one negative nominal interval and no Holm rejections. Band-minus-plateau readout gains average $+4.2152$ pp for hard and $+4.2593$ pp for DPC; their nominal positive/negative counts are 31/3 and 32/3. The joint correction retains five and four positive contrasts, respectively, and no negative contrasts. This asymmetry in corrected evidence is retained: regional training patterns remain descriptive rather than a corrected discovery of mechanism separation.

These regional contrasts were selected after the existing results had been examined and are exploratory. A contrast between plateau and transition band describes the already selected operating regions of each pipeline. The regions need not share an SNR threshold across models or seeds, and all-SNR metrics contain rather than independently replicate their constituent regions. The counts of cells with a positive mean under every readout family use intersections of identical cell identifiers; positive-interval intersections are recorded separately. Shared checkpoints and test samples mean that these family intersections are consistency checks, not independent replications.

Table 4: Benchmark detail for the transition-band training contrast before and after shared readout. Effects are descriptive means in pp; positive and negative counts use nominal three-seed intervals for the after-readout difference. The remaining cell intervals include zero. This table complements the all-family, all-region comparison in the main text.

Benchmark	Raw	Shared	Interaction	Positive	Negative
2016.10B	+0.62	-0.29	-0.90	0/13	2/13
2016.10A	+0.87	+0.68	-0.19	2/13	0/13
2018.01A	-0.79	+0.25	+1.04	0/12	0/12
RML22	+0.06	+0.72	+0.66	1/8	1/8
HisarMod	-0.22	+0.20	+0.43	3/11	1/11

B PARAMETER SHARING: COMPLETE CELL COMPARISON

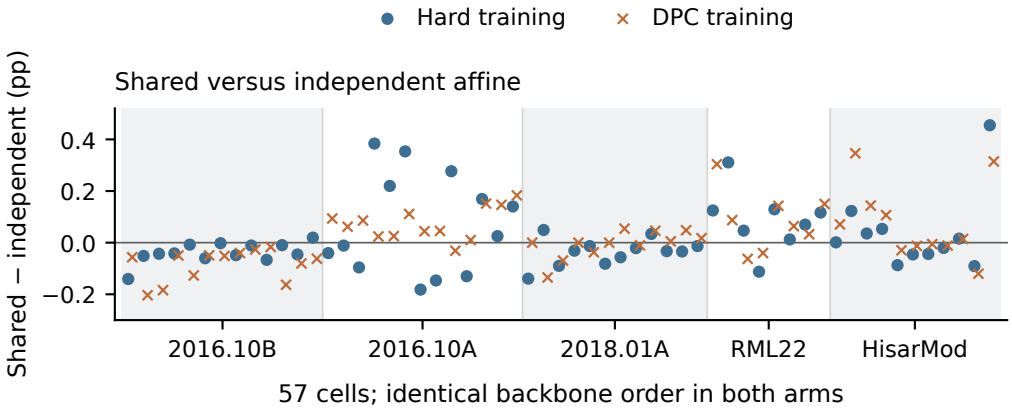


Figure 4: Shared-minus-independent transition-band accuracy in all 57 cells. Both arms use the same benchmark and backbone order as Figure 1. Points are means of three paired seed effects. Complete nominal 95% intervals and the separately defined 57-cell Holm adjustments for each arm remain in Table 3. This comparison uses the same fixed checkpoints and validation procedure as the main-figure contrasts.

C ORIGINAL FROZEN-READOUT LADDER: RETAINED RESULTS

The results in this appendix retain the original fitting protocol and preregistered adjudication. They are distinct from the nested-selection follow-up in the main text. Both studies use probability-derived frozen scores; the historical term logits in the ladder denotes these log-probability scores, not observed native pre-softmax activations. No old verdict has been recomputed with the new objective.

Table 5: Original-ladder headroom by rung: median across benchmark cells of three-seed mean accuracy differences from hard, in percentage points. These medians differ from the benchmark means in Table 8.

benchmark	cells	TS	VS	Global	F_{shift}	F_{VS}	F_{aff}
RadioML2016.10B	13	+0.00	−0.38	−0.12	+1.33	+4.03	+10.60
RadioML2016.10A	13	+0.00	−0.18	−0.20	+1.74	+3.80	+10.01
RadioML2018.01A	12	+0.00	+0.23	+0.20	+1.00	+1.19	+1.60
RML22	8	+0.00	+0.33	+0.24	+0.32	+0.32	+0.30
HisarMod2019.1	11	+0.00	+0.16	+0.17	+0.24	+0.19	−0.31

C.1 THE GRID AND ITS NOMINAL TRAINING GAINS

The roster contains fourteen backbones per benchmark except HisarMod2019.1, which ships thirteen. The DAE scope change leaves 69 cells; 57 have all three canonical seeds usable on both matched arms. The remaining twelve remain in the inventory (Table 12).

The preregistered cell criterion uses the two-sided paired Student-t interval across three per-seed differences. We call a gain *nominally established* when this unadjusted interval lies above zero. Eleven cells meet the criterion, forty-three are inconclusive and three are negative or show no gain (Table 7). These outcomes do not survive multiplicity correction: Bonferroni and Holm retain zero cells. Three seeds yield wide intervals, so an inconclusive cell does not establish absence of a useful effect.

Table 6: The frozen-readout ladder. Only the lower block is a nested capacity chain; it also uses the test-time reliability bin. In our protocol, a training-time method must beat F_{aff} to trigger a feature-level test. Survival shows only that the specified affine map of frozen logits did not reproduce the gain.

rung	map on frozen logits z	needs test-time r?	re-ranks?
Temperature scaling (TS)	z/T	no	no
Vector scaling (VS)	$a_k z_k + b_k$	no	yes
Global class-bias	$z + \boldsymbol{\beta}$	no	yes
F_{shift} per-bin shift	$z + \boldsymbol{\beta}_b$	yes	yes
F_{VS} per-bin VS	$a_{k,b} z_k + c_{k,b}$	yes	yes
F_{aff} per-bin affine	$W_b\, z + \boldsymbol{\beta}_b$	yes	yes (full)

Table 7: Evidence census over the rebuilt grid. Cell verdicts use the preregistered, unadjusted three-seed paired Student-t interval; no cell passes Bonferroni or Holm. The pooled interval treats the 57 cell effects as independent observations. Fixed-seed intervals use 5,000 paired resamples.

evidence state	cells
cells in the grid	69
analysable (three usable seeds on both arms)	57
not analysable (Table 12)	12
establishes a gain over hard training	11
reproduced by a frozen rung	7
exceeds the affine readout F_{aff}	4
comparison with affine inconclusive	0
plateau not retained	0
gain inconclusive	43
negative or no gain	3
pooled effect over the 57 analysable cells (pp)	$+0.14$ $[-0.39, +0.66]$
plateau retained (cell–seeds)	159/171

The equal-weight mean over 57 cell effects is $+0.14$ pp, with a descriptive across-cell Student-t interval of $[-0.39, +0.66]$. That interval treats cell effects as independent observations; shared benchmarks and backbones challenge this assumption. It is neither a dependence-robust interval nor a forecast for a new benchmark. Plateau accuracy is retained in 159 of 171 cell–seeds and in all three seeds of every nominally established cell. Appendix H reports dependence-sensitive global-count calculations; their conclusion changes when the related RadioML2016 datasets are treated as one cluster.

C.2 Frozen-logit headroom varies across benchmarks

Measurement on the full analysable grid. Define *headroom* as $\Delta_{\text{aff}} = F_{\text{aff}} - \text{hard}$ in each cell's validation-defined transition band. This difference is defined on all 57 analysable cells regardless of their training-gain verdict. Benchmark means are $+11.0$, $+9.8$, $+2.2$, $+0.9$ and $+0.5$ pp on RadioML2016.10B, 10A, RadioML2018.01A, RML22 and HisarMod2019.1, respectively (Table 8, Figure 5). All 26 RadioML2016 cell means exceed $+7.7$ pp and hence the largest nominally established training gain anywhere in the grid ($+3.93$ pp).

Exploratory association, conditional on the analysis gates. An additive benchmark–backbone model gives $F(4, 39) = 171$ and partial $\eta^2 = 0.95$ for benchmark given backbone, versus $F(13, 39) = 4.3$ and $\eta^2 = 0.59$ for backbone given benchmark. Within-backbone permutations give $p < 10^{-4}$ for benchmark; within-benchmark permutations give $p = 6 \times 10^{-4}$ for backbone, using 10^5 draws. These post hoc analyses treat the observed 57-cell set as fixed and assume exchangeability within the permutation strata; they do not account for outcome-related gate selection or identify a causal dataset property. The same decomposition of training gains gives $F(4, 39) = 2.2$, $p = 0.08$.

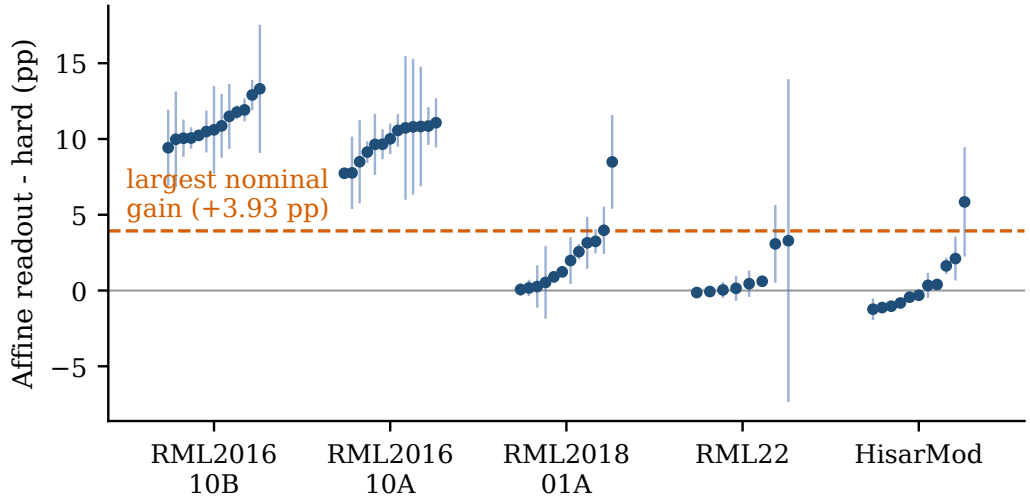


Figure 5: Frozen-logit headroom in each of the 57 analysable cells. Points are three-seed means and bars are unadjusted paired 95% intervals. The dashed line is the largest nominally established training gain in the grid (+3.93 pp). Every RadioML2016 cell mean exceeds it with a frozen readout alone.

Table 8: Frozen-affine headroom in validation-defined transition bands over all 57 analysable cells: mean, range and nominal interval counts. The exploratory decomposition conditions on the fixed analysable set.

benchmark	cells	mean	range	interval > 0
RadioML2016.10B	13	+11.01	[+9.42, +13.31]	13/13
RadioML2016.10A	13	+9.79	[+7.74, +11.07]	13/13
RadioML2018.01A	12	+2.21	[+0.07, +8.49]	8/12
RML22	8	+0.93	[−0.13, +3.29]	2/8
HisarMod2019.1	11	+0.49	[−1.23, +5.85]	4/11

This does not establish absence of a benchmark effect, nor does significance in one analysis and not the other establish that the two effects differ.

Matched contrasts and estimation limits. RadioML2016.10B and RML22 both have ten classes. Across their eight shared backbones, paired headroom differences all favour 10B by +8.9–+11.6 pp (Appendix H). Class count alone cannot explain that contrast. RadioML2016.10A and HisarMod each provide 200 validation frames per class and SNR bin, yet have mean headroom +9.8 versus +0.5 pp. This does not equate their estimation difficulty: HisarMod's larger affine map has only 7.4 frames per parameter and performs worse than F_{shift} in eight of eleven cells, compared with 36 frames per parameter on RML22. Poor top-rung estimation can suppress measured headroom and make absorption harder, biasing HisarMod toward producing survivors. Benchmark labels therefore summarize observed differences, including estimation conditions, rather than explain their cause.

The ladder among nominal gains. For each of the eleven nominally established gains we record the first rung whose three-seed method-minus-rung interval is entirely at or below zero. Seven are reproduced: two by F_{shift}, three by F_{VS} and two by F_{aff} (Table 9). All seven lie on RadioML2016.10A or 10B. Four survive the fitted ladder: IC-AMCNet and GRU2 on HisarMod2019.1, and ResNet and PET-CGDNN on RML22. These are candidate locations for feature probes; reproduction elsewhere does not establish unchanged representations. Four reproduced cells have per-seed ladder outcomes that disagree with their three-seed outcome. Under an alternative, non-preregistered per-seed crite-

rion, the reproduced side remains seven RadioML2016 cells but the survivor side grows to seven and includes RadioML2018.01A. This restriction on survivors is therefore criterion-dependent (§7).

Table 9: The eleven nominally established gains under the preregistered three-seed criterion. Columns 3–5 give per-seed transition-band differences; column 6 gives the unadjusted paired Student-t interval. The absorbing rung is adjudicated from three-seed intervals, not a vote over per-seed verdicts.

benchmark	backbone	method − hard per seed (pp) 2026	2027	2028	three-seed mean [95%]	lowest sufficient rung
Exceeds the affine readout						
HisarMod2019.1	IC-AMCNet	+2.92	+4.68	+4.20	+3.93 [+1.67, +6.20]	—
HisarMod2019.1	GRU2	+2.22	+1.37	+2.73	+2.11 [+0.40, +3.82]	—
RML22	ResNet	+1.37	+2.27	+1.48	+1.71 [+0.49, +2.93]	—
RML22	PET-CGDNN	+1.03	+1.55	+2.04	+1.54 [+0.29, +2.79]	—
Reproduced by a frozen readout						
RadioML2016.10A	LSTM2	+2.64	+4.38	+2.64	+3.22 [+0.72, +5.72]	F_{aff}
RadioML2016.10A	GRU2	+2.75	+1.53	+3.52	+2.60 [+0.12, +5.08]	F_{aff}
RadioML2016.10A	IC-AMCNet	+1.11	+0.88	+0.56	+0.85 [+0.17, +1.53]	F_{shift}
RadioML2016.10B	PET-CGDNN	+1.70	+2.33	+2.29	+2.10 [+1.23, +2.98]	F_{shift}
RadioML2016.10B	MCNET	+1.59	+2.41	+1.61	+1.87 [+0.70, +3.04]	F_{VS}
RadioML2016.10B	GRU2	+1.37	+2.59	+1.60	+1.85 [+0.26, +3.45]	F_{VS}
RadioML2016.10B	LSTM2	+1.15	+1.14	+1.75	+1.35 [+0.47, +2.22]	F_{VS}

Implication for reporting. Within this grid, frozen logits support larger improvements on RadioML2016 than any nominal training gain we observed. Thus an accuracy improvement alone leaves a substantial readout explanation unresolved. A comparable frozen-logit control should precede a representation argument, with its metadata requirements and estimation quality reported. This evidence does not establish that every previous or future method's gain is recoverable. Headroom avoids selecting cells by their gain verdict, but still depends on the chosen readout, training protocol and analysis gates.

C.3 The readout's ten points are not free: it needs the SNR

Headroom is measured with the true SNR in hand. A per-bin readout needs an SNR estimate at inference; the training-time method, once trained, runs on the frame alone. The accuracy differences alone do not establish whether the ten points are useful when SNR must be estimated.

We route each test sample by $\widehat{\text{SNR}} = \text{SNR} + \mathcal{N}(0, \sigma^2)$ and sweep σ over all thirteen analysable RadioML2016.10B cells at three seeds each, every cell keeping its own validation-defined band. With perfect metadata the readout is 10.4 pp above the method, whereas Table 8 compares the readout with hard training. The advantage decays to $+3.1$ pp at $\sigma = 2$ dB; its first nonpositive value on the sampled error grid is at $\sigma = 3.0$ dB. Gaussian soft routing, which spreads a sample over bins in proportion to its estimate, reaches its first nonpositive sampled value at 3.5 dB. These are grid-based crossings, not estimated continuous thresholds; the metadata-free method remains a comparator when metadata are unreliable.

So the practical rule is conditional, not a threshold to quote: on this benchmark the observed average readout advantage requires sufficiently accurate SNR estimation; a deployment decision must also check plateau retention and local validation performance. These numbers are specific to this benchmark, this readout, these validation-defined bands and an additive Gaussian error model.

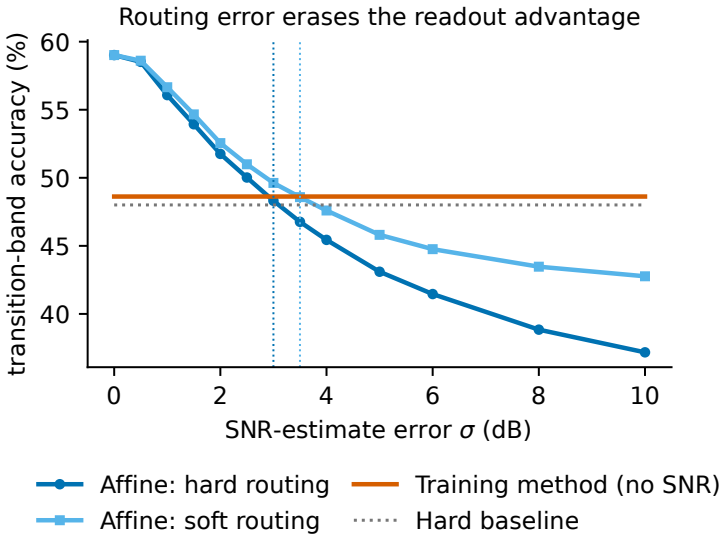


Figure 6: RadioML2016.10B under additive Gaussian SNR-estimation error, averaged over all thirteen analysable cells and three seeds each. The frozen affine readout first loses its average advantage on the sampled error grid at $\sigma = 3.0$ dB with hard routing and 3.5 dB with soft routing (dotted lines). These are sampled points, not fitted continuous thresholds. Accuracy is measured in each cell's own validation-defined transition band.

D Theory and Proofs

This section states the readout's information limits and the population identities motivating the training targets. Throughout, $\eta_r(x) = \Pr(Y \mid X = x, R = r)$ is the reliability-conditioned Bayes posterior, q a surrogate target on the simplex, π the class prior, and $H(\cdot)$, $D_{\mathrm{KL}}(\cdot\|\cdot)$, $I(\cdot;\cdot)$ the Shannon entropy, KL divergence, and mutual information. These are population quantities; public data do not reveal the true posterior. The audit instead evaluates predictive distributions under fixed rules.

The ladder and its operational interpretation. Fix the exported probability vector $p(x)$ and score vector $u(x) = \log \mathrm{clip}(p(x), 10^{-12}, 1)$. With reliability bin $b = b(r)$, the three conditional score-map families are $F_{\mathrm{shift}} = \{u + \beta_b\}$, $F_{\mathrm{VS}} = \{a_b \odot u + c_b\}$ and $F_{\mathrm{aff}} = \{W_b u + \beta_b\}$, each followed by softmax. They satisfy $F_{\mathrm{shift}} \subset F_{\mathrm{VS}} \subset F_{\mathrm{aff}}$. The experiments compare validation-NLL-fitted members of these families using test accuracy intervals. NLL fitting does not guarantee maximal test accuracy, and finite validation data do not identify the best member of a family.

Proposition 1 (nested readout families). A shift is a vector-scaling map with $a_b = \mathbf{1}$, and vector scaling is an affine map with $W_b = \mathrm{diag}(a_b)$. Hence any performance achievable by a member of a lower family is achievable by some member of a higher family. This inclusion is a capacity statement; it does not force separately estimated members to have ordered test performance.

Proposition 2 (information available to the control). For fixed fitted parameters, a readout depends on a frame only through (u, b). Two frames with the same (u, b) therefore receive identical predictions; the readout cannot distinguish information absent from that interface. This is an information restriction, not a linear-rank statement about internal features. Even without clipping, $\log \mathrm{softmax}(z)$ subtracts a sample-dependent log-normalizer from native logits z, so an affine map on log probabilities need not be affine in the features producing z. We make no capacity ordering between this control and a linear feature probe. If the fitted readout reproduces a trained method's accuracy, that performance is recoverable from the exported outputs and specified metadata. It does not imply unchanged representations. Failure to reproduce a gain can reflect estimation error, a more flexible output map, or information in features that the exported outputs omit. A feature-level test must assess these alternatives under matched information and fitting conditions.

Figure 7 fixes the vocabulary the body uses without restating: the transition band and the low-information floor on the left, and on the right why a one-hot training label and a non-one-hot posterior are consistent for the same realized frame.

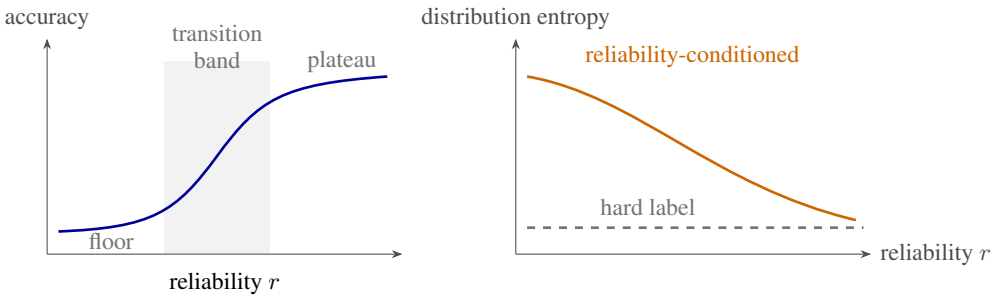


Figure 7: Schematic relationship between reliability, accuracy and uncertainty; the curves are not measured or numerically scaled. **Left**: the accuracy waterfall. **Right**: a one-hot training label always has zero entropy (the offset dashed line is a schematic baseline), whereas a true posterior in a specified generator, or a posterior-motivated surrogate target on public data, can become more diffuse when observations are less informative. We test whether training with such targets beats a readout of the frozen hard-label model.

Posterior risk and the target gap. For fixed (x, r) the log-loss risk of a predictor $p(x, r)$ is $-\sum_k \eta_{r,k}(x) \log p_k(x, r) = H(\eta_r(x)) + D_{\mathrm{KL}}(\eta_r(x) \| p(x, r))$, so the unique minimizer is $p = \eta_r$. Replacing η_r with a surrogate q adds exactly $D_{\mathrm{KL}}(\eta_r \| q)$ to the pointwise risk. This posterior belongs to the inverse problem faced by a classifier. The forward generator has already fixed Y, but the classifier does not observe the latent clean signal, the realized noise, or the generator's class record. Marginalizing these hidden variables can leave overlapping class-conditional likelihoods. Every realized label therefore remains one-hot, while its population conditional mean satisfies $\mathbb{E}[\delta_Y \mid X{=}x, R{=}r] = \eta_r(x)$. This expectation is over draws from the generator, not over alternative labels for one realized frame.

More generally, let V denote the inputs supplied to a predictor. For a fixed V, the population cross-entropy risk is minimized by $\Pr(Y{=}\cdot \mid V)$. With $V = (X, R)$ this is $\eta_R(x)$; with only the observed frame, $V = X$, it is $\Pr(Y{=}\cdot \mid X{=}x)$ after reliability has been marginalized. This population statement does not guarantee that finite data, a restricted model class, or an optimizer reaches the minimizer. At high reliability, either posterior can concentrate near the realized class. We call $D_{\mathrm{KL}}(\eta_r \| q)$ the *target gap*. Absolute NLL also contains the unobserved conditional-entropy term. On the same data distribution, NLL differences can reflect differences in target gap; an absolute validation NLL is not itself the target gap.

Group posterior moment. For any validation-admissible group g and constant target q_g,

$$\mathbb{E}[D_{\mathrm{KL}}(\eta_R(X) \| q_g) \mid g] = \mathbb{E}\Big[\textstyle\sum_k \eta_{R,k}(X) \log \eta_{R,k}(X) \,\Big|\, g \Big] - \sum_k \bar{\eta}_{g,k} \log q_{g,k},$$

where $\bar{\boldsymbol{\eta}}_g = \mathbb{E}[\eta_R(X) \mid g]$. The first term does not depend on q_g; the second is the cross-entropy from $\bar{\boldsymbol{\eta}}_g$ to q_g, minimized at $q_g = \bar{\boldsymbol{\eta}}_g$. Adding and subtracting $\sum_k \bar{\eta}_{g,k} \log \bar{\eta}_{g,k}$ gives

$$\mathbb{E}[D_{\mathrm{KL}}(\eta_r \| q_g) \mid g] = \mathbb{E}[D_{\mathrm{KL}}(\eta_r \| \bar{\boldsymbol{\eta}}_g) \mid g] + D_{\mathrm{KL}}(\bar{\boldsymbol{\eta}}_g \| q_g), \tag{4}$$

whose first term is the cost of using one distribution for every pointwise posterior in the group. This motivates a *group posterior moment* as a possible target, but does not establish that the in-sample hard teacher used here estimates it accurately. The class-level case $g = (Y{=}y, R{=}r)$ averages pointwise posterior vectors over frames generated with that class and reliability. It is not the conditional distribution $\Pr(Y \mid Y{=}y, r)$, which is trivially one-hot.

Endpoint limits. Assume positive class priors π_k that do not vary with R. Bayes' rule gives $\eta_{r,k}(x) = \pi_k p_r(x \mid k) / \sum_j \pi_j p_r(x \mid j)$. At high reliability, suppose that for X_r drawn conditional on $Y = y$, $p_r(X_r \mid k)/p_r(X_r \mid y) \to 0$ in probability for every $k \neq y$. Dividing by

$\pi_y p_r(X_r \mid y)$ then gives $\eta_r(X_r) \to \delta_y$. At a non-informative floor, require the defined likelihood ratios $p_r(X_r \mid k)/p_r(X_r \mid j) \to 1$ in probability under the mixture distribution for all classes. Then $\eta_{r,k}(X_r) \to \pi_k$, which is uniform for balanced priors. Pointwise convergence of all densities to zero alone does not imply this result. If class priors depend on SNR, Bayes' rule instead uses $\pi_k(r) = \Pr(Y = k \mid R = r)$, and the floor follows those priors. These are conditional population limits, not claims that the benchmark curves reach either endpoint. The observed training label remains one-hot throughout.

Degradation-posterior consistency. For a Markov degradation chain $Y \to X_{r_h} \to X_{r_l}$ with $r_h > r_l$ and each class k,

$$\Pr(Y{=}k \mid X_{r_l}) = \mathbb{E}[\Pr(Y{=}k \mid X_{r_h}, X_{r_l}) \mid X_{r_l}] = \mathbb{E}[\Pr(Y{=}k \mid X_{r_h}) \mid X_{r_l}],$$

the posterior-flow identity. The data-processing inequality gives $I(Y; X_{r_h}) \geq I(Y; X_{r_l})$, and since $H(Y \mid X_r) = H(Y) - I(Y; X_r)$, average posterior entropy cannot decrease as degradation increases. Under this coupling the Bayes posterior predicts a monotone reliability waterfall; the measured transition curves are trained-model accuracies on data where the coupling need not hold, so this motivates the shape rather than guaranteeing it.

E Validation-Only AMC Coordinate Probe

The coordinate probe is computed entirely within validation. We assign each validation sample to one of five out-of-fold partitions, stratified by class and reliability. Samples are ordered by a SHA256 hash of dataset, model seed and sample index, then assigned round-robin within each stratum; a fixed row index is the fallback when an artifact has no sample identifier. SNR is treated as a categorical coordinate. For each held-out sample, the prediction is the mean outcome among samples at its SNR in the other four folds, with their overall mean as fallback for an unseen SNR. Outcomes are the frozen hard model's predictive entropy and correctness. We pool the held-out predictions to calculate one R^2 per outcome and run.

The primary census contains 190 canonical hard-arm runs. Eight additional redraws are retained only in a separately identified 198-run sensitivity. These sets must not be confused with the 171 hard runs in the fully paired, three-seed headroom grid. The per-run values and source receipts are released as `figures_data/coordinate_probe.csv`; the expanded sensitivity is `coordinate_probe_sensitivity_198.csv` in the same directory.

Table 10: Validation-only five-fold OOF coordinate probe over every usable hard-arm run in the canonical census (190 runs). Entries are the median across runs with the observed range in brackets. The probe measures how outcomes vary with SNR; it is not used to select a coordinate or predict readout slack, and never reads test.

benchmark	runs	entropy R^2	correctness R^2
HisarMod2019.1	36	0.17 [0.06, 0.50]	0.15 [0.03, 0.35]
RadioML2016.10A	40	0.64 [0.62, 0.68]	0.41 [0.35, 0.51]
RadioML2016.10B	42	0.84 [0.78, 0.87]	0.48 [0.32, 0.51]
RadioML2018.01A	42	0.84 [0.58, 0.88]	0.54 [0.16, 0.65]
RML22	30	0.69 [0.35, 0.75]	0.33 [0.07, 0.43]

Canonical census and redraw sensitivity. Table 10 summarizes the canonical runs. The expanded census adds eight redraws, retaining their identities rather than silently treating them as canonical seeds. Both views support the descriptive observation that predictive entropy is more closely organized by physical SNR on RadioML2016.10B and RadioML2018.01A than on HisarMod2019.1, while RML22 supplies a counterexample to using coordinate predictiveness as a forecast of headroom. Neither census is a new gate on the frozen-logit results.

The historical coordinate-replacement rule required a learned coordinate to improve OOF R^2 by at least 0.10 on predictive entropy or correctness. No learned-coordinate branch was implemented or selected in this study; the rule did not gate the reported readout results.

F EXPERIMENTAL SETUP AND PROTOCOL

Datasets and models. The public audit uses RadioML2016.10A/B (O'Shea et al., 2016) on their AMR-Benchmark splits (Zhang et al., 2022), RML22 (Sathyanarayanan et al., 2023), HisarMod2019.1 (Tekbiyik et al., 2020), and RadioML2018.01A (O'Shea et al., 2018). SNR is the physical reliability coordinate. The backbone roster is the AMR-Benchmark family: 1DCNN-PF, CGDNet, CLDNN, CLDNN2, CNN1, CNN2, DenseNet, GRU2, IC-AMCNet, LSTM2, MCLDNN, MCNET, PET-CGDNN and ResNet, on every benchmark except HisarMod2019.1, which ships no 1DCNN-PF. The denoising autoencoder of the upstream roster is outside the analysis scope (Section 7).

Matched comparison and intervals. Each pair uses the same backbone, split, optimizer, batch size, training budget, and sample indices. Every prediction artifact contains logits or probabilities, labels, SNR, sample indices, class order when available, and provenance. The audit rejects mismatched class counts, available class orders, split hashes, invalid probabilities, or misaligned labels. We target seeds $\{2026, 2027, 2028\}$ and report the actual n. Fixed-seed intervals use 5,000 stratified paired bootstrap resamples. Across-seed intervals are paired Student-t intervals with $\mathrm{df} = n - 1$.

How the runs were actually produced. The corpus was trained on a shared, heterogeneous cluster, and each job carried a per-job wall clock whose stated purpose was to catch a hung process rather than to bound run time. That distinction turned out to matter. One cell, RadioML2018.01A with MCLDNN, required roughly 64 hours in a completed hard-arm run (155 epochs at ≈ 25 minutes each). Seed 2026 was truncated twice, at 20 and then 40 hours. Those two attempts consumed 60 GPU-hours and left checkpoints and training logs but no predictions or provenance, because the harness wrote those artifacts only after a normal return. Seeds 2027 and 2028 outlived their supervising queue processes, so the queue timeout no longer applied to those attempts (deviation record, D8).[1]

Selection and verdicts. Let a_s be hard-validation accuracy in percentage points at SNR bin s, $a_{\min} = \min_s a_s$ and $h = \max_s a_s - a_{\min}$. The transition band contains bins with $a_{\min} + 0.2h \leq a_s \leq a_{\min} + 0.8h$; the plateau contains bins with $a_s > a_{\min} + 0.8h$. If $h < 10^{-9}$ pp, both regions contain all bins. An empty transition band uses the bin at zero-based index $\lfloor B/2 \rfloor$ in the accuracy-sorted order of the B bins; an empty plateau uses all maximum-accuracy bins. Each seed's regions are shared by both arms and all readouts. Plateau retention means the test plateau method-minus-hard point estimate is at least -0.5 pp; a cell retains the plateau only if all three seeds meet this tolerance. This is separate from the interval test of band gain. Reconstructing the 171 hard-validation curves in the matched follow-up gives the stored regions exactly; none invokes a region fallback.

The original readouts minimize validation NLL with fixed regularization, except temperature scaling, which has no ridge penalty. A per-bin fit uses the global fit when its validation bin has fewer than 50 samples; a test bin unseen in validation also uses the global fit. These are the original ladder's fitting rules, separate from the nested selection in the new study. The fixed audit order is TS $\to$ VS $\to$ Global $\to F_{\text{shift}} \to F_{\text{VS}} \to F_{\text{aff}}$; only the final three form a strict capacity chain. A cell is classified from intervals, not point estimates:

[1] We report this because it shapes what a wall clock may be used for. A cap short enough to truncate a healthy run does not merely shorten it; with artifacts written only on normal exit it deletes the run outright, and it does so selectively, on exactly the slowest cells. The harness now converts the termination signal into a stop-at-next-batch and then follows its ordinary dump path, so a capped run yields the same artifacts as a converged one and is labelled `stopped_by=wall_clock`; the cap is unified at 96 hours, about $1.5\times$ the slowest measured cell. Two aggregation gates refuse to pool a truncated run with converged ones, within a cell and across the two arms of a pair, since a converged method arm minus a truncated baseline would measure training length rather than method. No run in the reported corpus is wall-clock stopped. Of the 446 canonical runs, 440 terminated by early stopping and 6 at their epoch budget; of the 409 that additionally pass the aggregation gates, the split is 403 and 6. All six epoch-budget runs are the three seed pairs of the RML22/DAE cell, which is outside the analysis scope (deviation record, D16), so every run entering the reported ladder stopped early. The safeguard is therefore preventive and changes none of the numbers reported here.

Algorithm 1 Interval-based frozen-readout audit

```
Fit the hard model and every readout on train/validation only.
Fix transition and plateau bins from the hard validation curve.
Evaluate the hard model, method, and frozen readouts once on test.
if upper(method−hard)≤ 0 then
    return negative_or_no_gain
else if the method−hard interval crosses 0 then
    return gain_inconclusive
else if plateau retention fails then
    return plateau_tradeoff
else if the first rung has upper(method−rung)≤ 0 then
    return absorbed_by_<rung>
else if method−F_aff crosses 0 then
    return inconclusive_vs_affine
else
    return survives_affine_requires_probe
end if
```

G The Fixed Audited Training Recipe

The experiment uses one fixed soft-target recipe. Let $\boldsymbol{p}_i^{\mathrm{hard}}$ be the matched hard model's probabilities on training sample i, evaluated at temperature $T = 1$. Teacher exports use the same dataset, backbone, split and canonical seed as the student; sample indices and labels are checked on loading. The teacher predicts samples from its own training split, so this is in-sample self-distillation, not a cross-fitted teacher.

For dataset SNR limits $s_{\min}, s_{\max}$ define $r_i = \mathrm{clip}((s_i - s_{\min})/(s_{\max} - s_{\min}), 0, 1)$. The ranges are $[-20, 18]$ dB for RadioML2016.10A/B and HisarMod2019.1, $[-20, 30]$ dB for RadioML2018.01A and $[-20, 20]$ dB for RML22. Write $\mathrm{norm}(v) = v/\sum_k v_k$. The implementation computes

$$\widetilde{\boldsymbol{p}}_i = \mathrm{norm}(\boldsymbol{p}_i^{\mathrm{hard}} + 10^{-4}\mathbf{1}), \quad \alpha_i = 0.25(1 - r_i), \quad \boldsymbol{b}_i = \mathrm{norm}((1 - \alpha_i)\widetilde{\boldsymbol{p}}_i + \alpha_i \mathbf{1}/K).$$

With $\epsilon_i = 0.7(1 - r_i)$ for $r_i < 0.8$ and $\epsilon_i = 0$ otherwise,

$$q_i^\star = \mathrm{norm}\big((1 - \epsilon_i)\delta_{y_i} + \epsilon_i \boldsymbol{b}_i\big), \tag{5}$$

and training minimizes

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}\big(f_\theta(x_i), q_i^\star\big). \tag{6}$$

There is no separate reliability-dependent sample weight. Validation uses true one-hot labels, and its loss selects the checkpoint and controls early stopping. The constants and base above are fixed, rather than chosen by comparing candidate targets on NLL and Brier. This recipe combines self-distillation and SNR-dependent smoothing; the grid does not test a menu of uniform, prior, confusion, cross-fitted-teacher and group-moment bases. At inference the student uses the frame alone, whereas the conditional readout also uses an SNR estimate. The DPC pipeline first trains the matched hard teacher and exports its training predictions. Matching student-training budgets therefore does not match total computation including teacher preparation; no equal-total-cost claim is made.

H Supporting Numbers for the Limitations

Multiplicity. The cell statistic is $t = \bar{d}/(s_d/\sqrt{3})$ on the three per-seed transition-band differences, with two degrees of freedom, for which the two-sided p is $1 - |t|/\sqrt{t^2 + 2}$. A Bonferroni level of $0.05/57 = 8.8 \times 10^{-4}$ therefore requires $|t| \geq 33.7$. The five smallest p values in the grid are RadioML2018.01A/CNN1 ($t = -20.7$, $p = 0.0023$), RadioML2016.10B/PET-CGDNN ($t = +10.4$, $p = 0.0092$), RadioML2018.01A/CLDNN ($t = -9.3$, $p = 0.0113$), HisarMod2019.1/IC-AMCNet ($t = +7.5$, $p = 0.0174$) and HisarMod2019.1/CNN2 ($t = -7.1$, $p = 0.0192$). Holm and Bonferroni both retain zero cells at $m = 57$. Three of these five cells have negative effects. Zero corrected rejections is a property of these observations, not an impossibility result for three seeds.

Illustrative power calculation. Across-seed standard deviations have median 0.92 pp and range $[0.19, 14.9]$ pp. For an assumed true difference δ, we plug each observed s_d into a noncentral t distribution with two degrees of freedom and noncentrality $\delta\sqrt{3}/s_d$, using the unadjusted two-sided 95% threshold. Mean plug-in power is 0.308689 at $+1$ pp, 0.571925 at $+2$ pp and 0.735194 at $+3$ pp. Each variance estimate comes from only three seeds, so these are illustrative calculations conditional on noisy observed variances, not measured power or guarantees for future runs.

Dependence-sensitive global-count calculations. For the equal-tailed paired Student-t intervals, suppose each upper-tail test has size at most 0.025 under its nonpositive-effect null. This requires valid one-sided inference, as obtained under independent, identically distributed normal seed differences; two-sided 95% coverage alone is insufficient. Under the global nonpositive null, each interval then lies wholly above zero with probability at most 0.025. Under independence of all 57 cells the binomial tail for at least eleven is 1.5×10^{-7}. This independence assumption is implausible when cells share benchmarks and backbones, so it is only a sensitivity calculation.

For a cluster containing m_j cells, the union bound gives a probability of any positive interval of at most $q_j = \min(1, 0.025m_j)$, allowing arbitrary dependence inside the cluster. With benchmark cluster sizes $(11, 13, 13, 8, 12)$ and four of five clusters positive, a Poisson-binomial calculation using these upper bounds gives 0.0246146875, *provided the five cluster events are independent*. Combining the related RadioML2016.10A/B datasets yields cluster sizes $(11, 26, 8, 12)$ and three of four positive clusters; the analogous bound is 0.1127. Shared backbones may also challenge independence between clusters. These calculations also condition on the gated set; nominal interval coverage after those gates has not been established. They are assumption-dependent upper bounds, not dependence-robust evidence that the method helps somewhere.

Matched-backbone headroom contrast. The eight shared backbones between RadioML2016.10B and RML22 all favour 10B, with paired headroom differences $+8.9$–$+11.6$ pp. Conditional on independent paired differences with equiprobable signs under the null, an eight-pair two-sided sign test gives $2/2^8 = 0.0078125$. Sharing datasets can violate that independence assumption. The paired differences are the primary descriptive evidence; the sign calculation is exploratory and does not establish a causal class-count explanation.

The DAE counterfactual. Of the DAE cells present in the archive, one (RML22) has all three seeds on both arms and two more have at least one usable pair. Rerunning the pooled estimate with the identical audit at $5{,}000$ resamples gives $+0.139$ pp $[-0.386, +0.664]$ over the 57 reported cells, $+0.082$ pp $[-0.446, +0.609]$ adding the three-seed DAE cell, and $+0.025$ pp $[-0.503, +0.554]$ adding all three. Per-seed DAE effects are predominantly negative, which is why removing them raises the pooled estimate. All three intervals cross zero.

Non-analysable cells. The aggregation gates that removed seeds, with counts over the whole campaign, are: paired baseline collapsed (15), hard-arm collapsed (15), method-arm collapsed (11), redraw seed excluded from the canonical set (16), out of scope (7), bad checkpoint (2), and paired feed mismatch (2). These counts cover the campaign, including recorded scope deviations. The DAE scope change occurred after other results were visible; its direction and effect are reported separately above. Mechanical gates do not guarantee that the analysable subset is representative of the planned grid.

The alternative per-cell rule. Requiring all three per-seed bootstrap intervals to exclude zero, rather than the preregistered three-seed interval, admits four cells the preregistered rule does not (HisarMod2019.1/MCLDNN and MCNET, RadioML2016.10B/MCLDNN, RadioML2018.01A/IC-AMCNet) and drops one it does admit (RadioML2016.10A/IC-AMCNet, whose gain is individually inconclusive on two seeds yet clears the paired interval), for 14 against 11.

The swap does not leave the ladder result unchanged, and the two halves of Appendix C's benchmark split behave differently under it. The reproduced side is robust: under either rule exactly seven cells are reproduced by a frozen rung and *all* of them lie on RadioML2016.10A or 10B. The surviving side is not: it grows from four cells to seven, and the three added cells include RadioML2018.01A/IC-AMCNet, so the clean statement that survivors occur only on HisarMod2019.1 and RML22 holds

under the preregistered rule and not under the alternative. We report the preregistered result and flag this explicitly, because the restriction on survivors depends on the chosen interval rule.

I THE COMPLETE ANALYSABLE GRID

Table 11 lists all 57 analysable cells. It gives each seed's transition-band difference, the preregistered paired three-seed Student-t interval and the corresponding unadjusted cell outcome. The eleven nominally established gains are those of Table 9; none is significant after Bonferroni or Holm. Per-seed differences are displayed rather than replaced by a majority vote over seed-level decisions.

The table shows why one seed can be misleading: the spread within a cell can exceed the cell's mean effect and even reverse its sign. HisarMod2019.1/LSTM2 spans $+2.50$ to -18.40 pp across seeds. Such variation remains visible in the complete grid even when the three-seed interval is inconclusive.

Table 11: Every analysable cell of the 69-cell grid. Columns 3–5 are the per-seed transition-band differences; column 6 is their three-seed paired Student-t interval, which is the preregistered cell-level criterion; column 7 is the cell verdict recomputed at that level, not a vote over per-seed verdicts.

		method − hard per seed (pp)			three-seed	
benchmark	backbone	2026	2027	2028	mean [95%]	verdict
HisarMod	CGDNet	$+3.72$	-0.39	$+4.39$	$+2.57$ $[-3.86, +9.00]$	inconclusive
HisarMod	CLDNN2	-0.91	-0.07	-0.51	-0.49 $[-1.54, +0.55]$	inconclusive
HisarMod	CNN1	-2.55	-0.86	-3.14	-2.18 $[-5.12, +0.75]$	inconclusive
HisarMod	CNN2	-2.90	-3.35	-4.60	-3.62 $[-5.80, -1.43]$	no gain
HisarMod	GRU2	$+2.22$	$+1.37$	$+2.73$	$+2.11$ $[+0.40, +3.82]$	survives
HisarMod	IC-AMCNet	$+2.92$	$+4.68$	$+4.20$	$+3.93$ $[+1.67, +6.20]$	survives
HisarMod	LSTM2	-0.90	$+2.50$	-18.40	-5.60 $[-33.46, +22.26]$	inconclusive
HisarMod	MCLDNN	$+2.48$	$+5.51$	$+2.85$	$+3.62$ $[-0.49, +7.73]$	inconclusive
HisarMod	MCNET	$+0.94$	$+1.12$	$+0.39$	$+0.82$ $[-0.13, +1.76]$	inconclusive
HisarMod	PET-CGDNN	$+0.13$	$+1.73$	$+1.76$	$+1.20$ $[-1.11, +3.52]$	inconclusive
HisarMod	ResNet	-17.28	$+11.74$	-8.92	-4.82 $[-41.92, +32.28]$	inconclusive
RML2016.10A	1dcnnpf	-1.70	$+0.44$	-1.02	-0.76 $[-3.47, +1.95]$	inconclusive
RML2016.10A	CGDNet	$+0.05$	$+0.77$	-0.97	-0.05 $[-2.22, +2.12]$	inconclusive
RML2016.10A	CLDNN	$+0.21$	-1.86	$+2.36$	$+0.24$ $[-5.01, +5.49]$	inconclusive
RML2016.10A	CLDNN2	$+3.11$	$+1.72$	$+0.94$	$+1.92$ $[-0.81, +4.65]$	inconclusive
RML2016.10A	CNN1	$+0.94$	-0.61	-0.73	-0.13 $[-2.45, +2.18]$	inconclusive
RML2016.10A	DenseNet	-1.71	$+0.73$	-0.98	-0.66 $[-3.77, +2.45]$	inconclusive
RML2016.10A	GRU2	$+2.75$	$+1.53$	$+3.52$	$+2.60$ $[+0.12, +5.08]$	abs. F_{aff}
RML2016.10A	IC-AMCNet	$+1.11$	$+0.88$	$+0.56$	$+0.85$ $[+0.17, +1.53]$	abs. F_{shift}
RML2016.10A	LSTM2	$+2.64$	$+4.38$	$+2.64$	$+3.22$ $[+0.72, +5.72]$	abs. F_{aff}
RML2016.10A	MCLDNN	$+0.45$	$+0.38$	$+1.61$	$+0.82$ $[-0.90, +2.53]$	inconclusive
RML2016.10A	MCNET	$+0.99$	-1.61	$+0.09$	-0.18 $[-3.46, +3.11]$	inconclusive
RML2016.10A	PET-CGDNN	$+1.03$	$+3.17$	$+1.29$	$+1.83$ $[-1.07, +4.73]$	inconclusive
RML2016.10A	ResNet	$+2.60$	$+0.76$	$+1.61$	$+1.66$ $[-0.63, +3.95]$	inconclusive
RML2016.10B	1dcnnpf	-1.50	-0.66	-2.84	-1.67 $[-4.40, +1.07]$	inconclusive
RML2016.10B	CGDNet	$+1.04$	$+0.54$	-0.49	$+0.36$ $[-1.57, +2.30]$	inconclusive
RML2016.10B	CLDNN	-1.15	$+0.92$	-0.59	-0.27 $[-2.93, +2.39]$	inconclusive
RML2016.10B	CLDNN2	-0.16	-0.22	$+3.50$	$+1.04$ $[-4.25, +6.33]$	inconclusive
RML2016.10B	CNN1	-0.26	-0.91	-1.55	-0.91 $[-2.50, +0.69]$	inconclusive
RML2016.10B	CNN2	-0.47	$+0.00$	-0.39	-0.29 $[-0.92, +0.35]$	inconclusive
RML2016.10B	DenseNet	-0.71	$+0.46$	$+0.36$	$+0.04$ $[-1.57, +1.64]$	inconclusive
RML2016.10B	GRU2	$+1.37$	$+2.59$	$+1.60$	$+1.85$ $[+0.26, +3.45]$	abs. F_{VS}
RML2016.10B	LSTM2	$+1.15$	$+1.14$	$+1.75$	$+1.35$ $[+0.47, +2.22]$	abs. F_{VS}
RML2016.10B	MCLDNN	$+1.25$	$+0.38$	$+1.75$	$+1.13$ $[-0.60, +2.85]$	inconclusive
RML2016.10B	MCNET	$+1.59$	$+2.41$	$+1.61$	$+1.87$ $[+0.70, +3.04]$	abs. F_{VS}
RML2016.10B	PET-CGDNN	$+1.70$	$+2.33$	$+2.29$	$+2.10$ $[+1.23, +2.98]$	abs. F_{shift}
RML2016.10B	ResNet	$+2.43$	$+1.45$	$+0.36$	$+1.41$ $[-1.16, +3.99]$	inconclusive
RML2018.01A	1dcnnpf	-3.05	-1.73	-0.06	-1.61 $[-5.34, +2.11]$	inconclusive
RML2018.01A	CGDNet	-4.83	-0.61	-0.19	-1.87 $[-8.25, +4.50]$	inconclusive
RML2018.01A	CLDNN	-3.13	-2.32	-2.28	-2.58 $[-3.76, -1.39]$	no gain

benchmark	backbone	2026	2027	2028	mean [95%]	verdict
RML2018.01A	CLDNN2	-3.95	-0.42	-3.76	-2.71 $[-7.65, +2.23]$	inconclusive
RML2018.01A	CNN1	-2.47	-2.10	-2.35	-2.31 $[-2.79, -1.83]$	no gain
RML2018.01A	CNN2	-2.44	-0.87	-1.89	-1.73 $[-3.71, +0.24]$	inconclusive
RML2018.01A	GRU2	$+1.94$	$+2.52$	$+0.14$	$+1.53$ $[-1.56, +4.62]$	inconclusive
RML2018.01A	IC-AMCNet	$+0.35$	$+0.86$	$+0.30$	$+0.50$ $[-0.26, +1.27]$	inconclusive
RML2018.01A	LSTM2	$+1.09$	-0.05	$+0.21$	$+0.42$ $[-1.08, +1.91]$	inconclusive
RML2018.01A	MCLDNN	$+0.72$	$+1.12$	-0.17	$+0.55$ $[-1.09, +2.20]$	inconclusive
RML2018.01A	PET-CGDNN	-0.17	$+1.89$	-0.02	$+0.57$ $[-2.30, +3.43]$	inconclusive
RML2018.01A	ResNet	-0.51	-0.43	$+0.31$	-0.21 $[-1.33, +0.91]$	inconclusive
RML22	1dcnnpf	-0.38	-0.05	-0.41	-0.28 $[-0.78, +0.21]$	inconclusive
RML22	CGDNet	-2.21	$+0.71$	-0.75	-0.75 $[-4.38, +2.88]$	inconclusive
RML22	CLDNN	$+0.10$	-10.47	-1.34	-3.90 $[-18.14, +10.33]$	inconclusive
RML22	CNN1	$+1.40$	-0.32	-0.39	$+0.23$ $[-2.29, +2.75]$	inconclusive
RML22	LSTM2	$+1.96$	-0.59	$+2.37$	$+1.25$ $[-2.73, +5.23]$	inconclusive
RML22	MCNET	$+1.07$	$+0.11$	$+0.82$	$+0.67$ $[-0.58, +1.91]$	inconclusive
RML22	PET-CGDNN	$+1.03$	$+1.55$	$+2.04$	$+1.54$ $[+0.29, +2.79]$	survives
RML22	ResNet	$+1.37$	$+2.27$	$+1.48$	$+1.71$ $[+0.49, +2.93]$	survives

What is not in this table. Twelve of the 69 grid cells are absent because they are not analysable (Table 12); the aggregation gates that removed their seeds are listed with counts in the deviation record. Denoising autoencoders were moved out of scope on 2026-09-06, after results for other cells were visible; that decision, its effect on the denominator, and a counterfactual recomputation that puts the excluded cells back are recorded in D16 of the deviation record and summarised in Section 7. HisarMod2019.1 ships no 1DCNN-PF backbone, which is why its roster is thirteen rather than fourteen (D17).

Table 12: The twelve cells of the 69-cell grid that are not analysable. Nine lost one or two seeds to the aggregation gates — collapse, bad checkpoint, or feed mismatch — and three RML22 cells yielded no analysable pair on any seed. This inventory concerns the fixed 69-cell grid; the earlier DAE scope change is disclosed separately.

benchmark	backbone	usable seeds	status
HisarMod2019.1	CLDNN	1	seeds lost to the gates
HisarMod2019.1	DenseNet	2	seeds lost to the gates
RadioML2016.10A	CNN2	1	seeds lost to the gates
RadioML2016.10B	IC-AMCNet	2	seeds lost to the gates
RadioML2018.01A	DenseNet	2	seeds lost to the gates
RadioML2018.01A	MCNET	1	seeds lost to the gates
RML22	GRU2	2	seeds lost to the gates
RML22	IC-AMCNet	2	seeds lost to the gates
RML22	MCLDNN	2	seeds lost to the gates
RML22	CLDNN2	0	no analysable pair
RML22	CNN2	0	no analysable pair
RML22	DenseNet	0	no analysable pair

J FR-AMC: An External Method Under the Same Audit

FR-AMC is a published SNR-aware training method with a public implementation. It is a useful external test of the audit because it was not designed by us and its loss is not a variant of the method audited in the main grid.

Source pinning. We pin FR-AMC to commit `c5eb898f06633ef2fb478e71f775f3553d6923d6` and load it through an external adapter. The repository provides neither a license nor model weights, so we redistribute no upstream source. The adapter verifies the commit and the SHA256 of every loss, network, and preprocessing file it loads, recording them

in each run's `run_manifest.json`; it changes only paths, loading, diagnostic execution, and prediction export. The selected released model and loss are unchanged, byte for byte.

Scope, and the constraint that sets it. The four released networks hard-code a frame length of 128 — MCLDNN and DAELSTM through `nn.Linear(128*2, 80)`, FEAT through a `torch.arange(0, 128)` phase term, ThreeStream through `self.signallen=128`. HisarMod2019.1 and RadioML2018.01A have length 1024, so running them would require editing upstream source, which the byte-identity constraint forbids. This restricts the eligible benchmarks to frame length 128; it does not imply that every eligible benchmark was evaluated. The matched experiment reported here uses two of them, RadioML2016.10B and RML22.

Matched full-SNR track. Every setting is held fixed across CE, official FR, and FR-eq for seeds $2026/2027/2028$ over the full SNR grid. The official loss keeps its mini-batch-wide sum and positive exponential. The published equation prints a negative exponent, so FR-eq implements that equation with a per-sample top-k sum and is reported as a separately named sensitivity, never as "the" method. CE validation logits define the transition band and fit the ladder; the test split is evaluated once. Because official FR sums its regularizer over the minibatch, the released batch size is part of each architecture's method semantics, so this tests transfer of the released architecture–recipe pair, not an architecture-only effect.

Table 13: FR-AMC matched-track audit in the CE-validation-defined transition band, three seeds, 5,000 resamples. Differences are percentage points with paired three-seed intervals. A negative FR$-F_{\text{aff}}$ interval favours the frozen affine readout over the trained method.

architecture	variant	FR − CE (pp)	FR $-F_{\text{aff}}$ (pp)	plateau	gain status
RadioML2016.10B					
MCLDNN	official FR	$+0.79\ [-1.47, +3.05]$	$-8.10\ [-9.64, -6.56]$	yes	inconclusive
MCLDNN	FR-eq	$+0.75\ [-4.64, +6.15]$	$-8.14\ [-12.15, -4.13]$	yes	inconclusive
DAELSTM	official FR	$+0.50\ [-2.12, +3.13]$	$-2.83\ [-4.69, -0.97]$	yes	inconclusive
DAELSTM	FR-eq	$-0.80\ [-2.66, +1.07]$	$-4.13\ [-7.77, -0.49]$	no	inconclusive
FEAT	official FR	$-0.25\ [-3.52, +3.02]$	$-11.00\ [-14.34, -7.66]$	no	inconclusive
FEAT	FR-eq	$-0.05\ [-3.68, +3.59]$	$-10.80\ [-14.23, -7.36]$	yes	inconclusive
ThreeStream	official FR	$-2.70\ [-5.11, -0.30]$	$-12.22\ [-13.63, -10.81]$	no	negative / no gain
ThreeStream	FR-eq	$-0.39\ [-2.17, +1.38]$	$-9.91\ [-11.59, -8.23]$	yes	inconclusive
RML22					
MCLDNN	official FR	$+0.55\ [-2.95, +4.04]$	$+1.17\ [-2.49, +4.82]$	no	inconclusive
MCLDNN	FR-eq	$+0.02\ [-3.50, +3.54]$	$+0.64\ [-3.03, +4.31]$	no	inconclusive
DAELSTM	official FR	$+6.78\ [-23.37, +36.93]$	$+6.38\ [-23.70, +36.45]$	no	inconclusive
DAELSTM	FR-eq	$+8.46\ [-14.89, +31.81]$	$+8.06\ [-15.05, +31.16]$	yes	inconclusive
FEAT	official FR	$-2.63\ [-4.13, -1.14]$	$-2.44\ [-4.06, -0.82]$	no	negative / no gain
FEAT	FR-eq	$+0.65\ [-1.30, +2.59]$	$+0.84\ [-0.92, +2.60]$	yes	inconclusive
ThreeStream	official FR	$-3.47\ [-10.05, +3.12]$	$-3.02\ [-9.58, +3.54]$	no	inconclusive
ThreeStream	FR-eq	$+0.32\ [-1.02, +1.67]$	$+0.77\ [-0.56, +2.11]$	yes	inconclusive

Outcome. Table 13 gives all sixteen comparisons: two benchmarks, four architectures, two FR variants, three seeds each, at 5,000 resamples. **None establishes a gain over its CE baseline.** Fourteen intervals cross zero and two are negative — official FR on ThreeStream/RadioML2016.10B (-2.70 pp $[-5.11, -0.30]$) and on FEAT/RML22 (-2.63 pp $[-4.13, -1.14]$). This is an absence of established gains, not a demonstration of equivalence; several intervals are far too wide to exclude a useful effect, and DAELSTM on RML22 is extreme, spanning $[-23.37, +36.93]$ pp across three seeds.

The frozen affine readout is at least as good as FR in nine of the sixteen comparisons, and the nine are not spread evenly: F_{aff} beats FR in all eight RadioML2016.10B comparisons and in one of eight on RML22. This descriptive contrast is consistent with the headroom pattern in Section 5.1, under a separate externally authored loss. It is not a confirmed replication of a causal benchmark effect.

What this track does not contain. It contains no release-faithful reproduction track. The released protocol fixes one benchmark, one split ratio and one seed, and reproducing it is a different experi-

ment from the matched comparison above; the extension protocol frozen on 2026-09-07 deliberately ran the matched track only. A reproduction claim about the published numbers therefore cannot be made from this evidence, and we make none.

K Extended Related Work

Soft targets and label smoothing. Label smoothing (Müller et al., 2019) and distillation (Hinton et al., 2015) replace one-hot targets with softened distributions. Our fixed in-sample teacher and SNR schedule combine these established ideas. Fitting a readout after training remains distinct from selecting training targets.

Calibration and class correction. Temperature and vector scaling are post-hoc calibration methods (Guo et al., 2017). Logit adjustment uses class-frequency information for long-tail recognition (Menon et al., 2021); its frequency-based offsets differ from the validation-fitted biases used here. NLL and Brier are proper scoring rules (Gneiting & Raftery, 2007); their improvement does not isolate calibration from other changes in predictive distributions.

Output adaptation and feature reweighting. Dirichlet calibration applies an affine map to log probabilities followed by softmax (Kull et al., 2019); our per-SNR control uses that output-map form with its stated clipping, regularization and fitting protocol. Deep feature reweighting instead retrains a final layer on frozen features using data that break spurious correlations (Kirichenko et al., 2023). It has a different input interface from our exported-probability readout. Atypicality-aware methods use sample or class rarity to improve uncertainty estimates (Yüksekgönül et al., 2023); that conditioning variable is distinct from the physical SNR supplied here.

SNR-aware AMC methods. FR-AMC regularizes ambiguous predictions (Liang et al., 2025); KADNet combines SNR-aware deformable convolution with Kolmogorov–Arnold networks (Wang et al., 2026); CAIC-Net couples dynamic cross-attention and cross-SNR contrastive learning (Wu et al., 2026). Our direct external audit is of FR-AMC (Appendix J); the other citations provide context, not additional experimental comparisons.

Uncertainty and degradation. Uncertainty decompositions (Kendall & Gal, 2017) and ensembles (Lakshminarayanan et al., 2017) address probability quality. Corruption benchmarks (Hendrycks & Dietterich, 2019) and AMC under SNR (O'Shea et al., 2016; 2018) provide degradation settings. Focal loss (Lin et al., 2017), mixup (Zhang et al., 2018), label-noise learning (Frénay & Verleysen, 2014) and selective prediction (Geifman & El-Yaniv, 2017) provide further comparison directions; we do not evaluate them here.