ORTHOGONAL ENSEMBLES IN PREDICTION SPACE: GEOMETRY, CONTROLS, AND RELIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Explicit diversity objectives shape the relationships among ensemble members, but their connection to predictive reliability depends on the representation being regularized. We study prediction-space coupling through a Gram penalty on centered, normalized member logits, optimized jointly with the supervised task loss. Our analysis relates the penalty to effective dimension and derives the rank constraint imposed by ensemble centering. We also construct members that attain the centered lower bound while producing identical class probabilities, establishing a limit of logit-space diversity. Across four classification datasets and five backbones, 40 coupled and uncoupled runs separate the gains from averaging predictions from the incremental effects of coupling. Both conditions exceed their average member's accuracy in every setting. Coupling improves accuracy in 6 of 20 settings, with a mean difference of -0.0972 accuracy points and different effects on proper scores and calibration. Additional experiments examine coupling strength, predictive entropy, and aggregation rules; a brain-age case study illustrates uncertainty alongside shared prediction bias. These findings clarify what the objective controls and show why predictive reliability must be evaluated alongside representation geometry.

1 INTRODUCTION

Deep ensembles combine multiple predictors to improve predictive accuracy and estimate uncertainty (Lakshminarayanan et al., 2017). Their effectiveness depends on both the quality of individual members and the relationships among their predictions. Random initialization and stochastic training can produce distinct solutions (Fort et al., 2019), yet members may still make the same mistakes and confidently agree on an incorrect prediction. Understanding these shared errors is essential when uncertainty estimates are used to assess reliability without access to labels at inference time (Guo et al., 2017; Hüllermeier & Waegeman, 2021).

Explicit diversity objectives offer a way to shape member relationships during training. Negative-correlation learning couples ensemble errors (Buschjäger et al., 2020), DICE reduces conditional redundancy in learned features (Ramé & Cord, 2021), and orthogonality penalties regularize network parameters or ensemble representations (Bansal et al., 2018; Larrazabal et al., 2021). The choice of representation determines what each objective controls. In particular, separation between logit vectors need not imply disagreement between class probabilities, and neither property alone establishes that members make complementary errors. The central question is how the geometry encouraged by a diversity objective relates to the predictive behavior of the resulting ensemble.

We study this question through prediction-space coupling: independently parameterized networks are trained with a shared supervised objective and a Gram penalty on centered, normalized logits. The output space gives members common coordinates in which their relationships can be analyzed. Batch centering removes per-class minibatch means, ensemble centering isolates each member's deviation from the collective output, and normalization makes the penalty directional. At inference, the ensemble averages raw logits before applying softmax; predictive entropy then summarizes the concentration of the resulting distribution.

This formulation permits an explicit analysis of what the penalty can achieve. Ensemble centering introduces a linear dependence among member residuals, preventing exact pairwise orthogonality

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.

when those residuals are nonzero. Moreover, softmax-invariant shifts can attain the centered penalty's lower bound while leaving every member's class probabilities identical. These observations motivate evaluating the learned geometry together with accuracy, probabilistic scores, calibration, and error complementarity.

We evaluate coupled and uncoupled ensembles across four datasets and five backbones, holding the main training and evaluation protocol fixed. Averaging improves accuracy relative to the mean member in all 20 configurations under both conditions, whereas coupling improves ensemble accuracy in six. The comparison therefore distinguishes the benefit of combining predictions from the additional effect of regularizing their geometry. Complementary studies examine coupling strength, entropy and errors, alternative aggregation rules, and shared bias in brain-age regression.

Our contributions are threefold:

- We formulate prediction-space coupling and characterize its geometry, relating the Gram penalty to effective dimension, deriving the centering constraint, and identifying a limit through an identical-probability construction.
- We provide a 20-configuration, 40-run comparison with $\lambda = 0$ controls that separates aggregation gains from coupling effects on accuracy, proper scores, calibration, and error complementarity.
- We examine the relationship between uncertainty and error through coupling-strength experiments, error-conditioned entropy, fixed-prediction FiLM aggregation, and an exploratory brain-age case study.

2 Related work

Diversity in ensembles. Deep ensembles learn separate solutions for probabilistic prediction (Lakshminarayanan et al., 2017; Fort et al., 2019). Explicit diversity objectives include negative-correlation learning (Buschjäger et al., 2020) and conditional-redundancy reduction (Ramé & Cord, 2021). Orthogonality has been studied in network weights (Bansal et al., 2018), parallel networks (Mashhadi et al., 2021), and segmentation ensembles (Larrazabal et al., 2021). Our formulation acts on centered outputs, enabling a direct analysis of their Gram geometry and its relationship to class probabilities.

Efficient ensembles. BatchEnsemble shares weights with member-specific rank-one factors (Wen et al., 2020), while FiLM-Ensemble uses feature-wise linear modulation (Turkoglu et al., 2022). Prediction-space coupling instead trains independently parameterized members with an output-level interaction. We assess this interaction using $\lambda = 0$ controls and compare systems and aggregation rules with FiLM in Section 5.6.

Uncertainty and reliability. Calibration describes the relationship between confidence and observed frequencies (Guo et al., 2017), while Brier score and NLL evaluate the full predictive distribution against outcomes (Brier, 1950). Predictive entropy summarizes uncertainty directly from model outputs. Its interpretation depends on the predictive model and the distinction between aleatoric and epistemic uncertainty (Hüllermeier & Waegeman, 2021; Wimmer et al., 2023). The usefulness of ensemble diversity also varies with the evaluation task (Xia & Bouganis, 2022); for example, selective prediction measures the errors remaining after abstaining on uncertain inputs (Geifman & El-Yaniv, 2019). We focus on the relationship between predictive entropy and correctness, together with distribution-level scores and calibration.

3 Method: prediction-space coupling

Figure 1 presents our training and inference procedure. During training, independently parameterized members process a shared minibatch and receive the same labels (panel a). Their centered, normalized logits define a Gram penalty that is optimized jointly with mean cross-entropy (panel b). During inference, raw member logits are averaged before softmax to produce the class prediction and predictive entropy (panel c). We first define the representation and objective, then describe inference and analyze the geometry induced by the penalty.

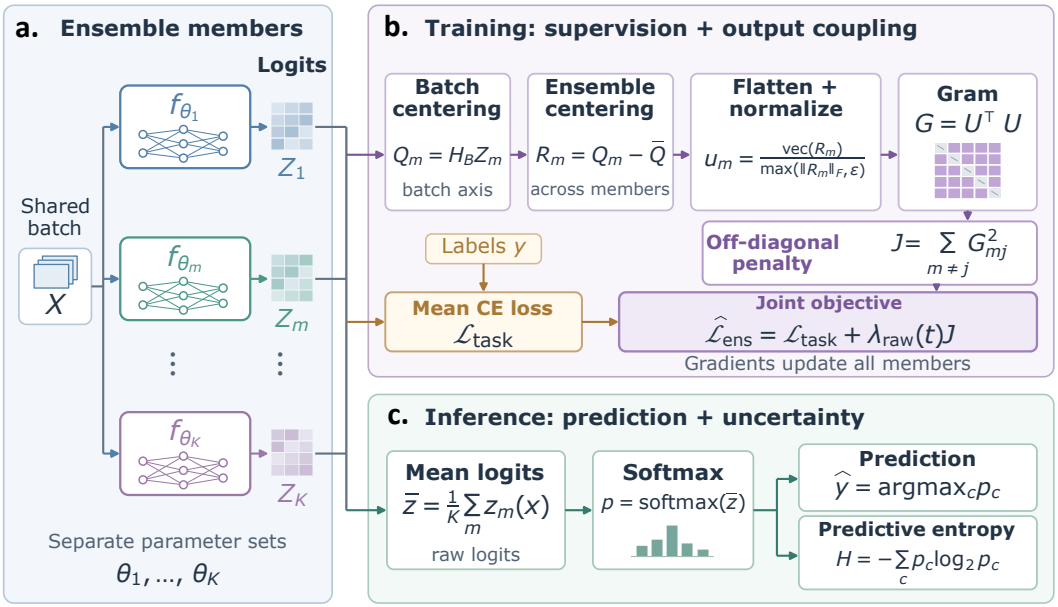


Figure 1: **Prediction-space coupling: training and inference. (a)** Independently parameterized members process a shared minibatch. **(b)** Batch and ensemble centering, followed by normalization, define the Gram penalty J. Joint training combines J with mean cross-entropy and updates all members. **(c)** At inference, raw logits are averaged before softmax to obtain the class prediction and predictive entropy. Centered residuals are used only for the training penalty.

3.1 Representations and objective

For a minibatch of B examples and C classes, each of $K \geq 2$ independently parameterized predictors produces logits $Z_m \in \mathbb{R}^{B\times C}$. With $H_B = I_B - B^{-1}\mathbf{1}_B\mathbf{1}_B^\top$, batch and optional ensemble centering give

$$Q_m = H_B Z_m, \qquad R_m = Q_m - \alpha\frac{1}{K}\sum_{j=1}^{K} Q_j, \qquad u_m = \frac{\mathrm{vec}(R_m)}{\max(\|R_m\|_F, \epsilon)}, \tag{1}$$

where $\epsilon = 10^{-8}$ and $\alpha \in \{0, 1\}$. Batch centering acts across examples for each class; ensemble centering subtracts $\bar{Q} = K^{-1}\sum_j Q_j$. Normalization removes residual magnitude when $\|R_m\|_F \geq \epsilon$. Classification uses $\alpha = 1$ (Fig. 1b); $\alpha = 0$ denotes batch centering alone.

Writing $U = [u_1, \ldots, u_K]$ and $G = U^\top U$, define

$$J = \sum_{m\neq j} G_{mj}^2, \qquad \mathcal{L}_{\mathrm{orth}} = \frac{J}{K(K-1)}. \tag{2}$$

In the main classification study, training minimizes

$$\widehat{\mathcal{L}}_{\mathrm{ens}} = \frac{1}{K}\sum_{m=1}^{K}\frac{1}{B}\sum_{i=1}^{B}\ell(Z_{m,i}, y_i) + \lambda_{\mathrm{raw}}(t)J, \tag{3}$$

where ℓ is cross-entropy and $\lambda_{\mathrm{raw}}(t)$ warms up to the reported coefficient. Since J sums ordered off-diagonal entries, its coefficient corresponds to $K(K-1)\lambda_{\mathrm{raw}}$ on $\mathcal{L}_{\mathrm{orth}}$: 0.01 becomes 0.2 for $K = 5$. The $\lambda = 0$ control optimizes the same averaged task loss, $\mathcal{L}_{\mathrm{task}}$, and omits the Gram computation.

The Gram computation requires $O(K^2BC)$ arithmetic and propagates gradients to all independently parameterized members. Centering and normalization are computed per minibatch.

3.2 Inference and uncertainty

Inference uses the original logits $z_m(x)$, without the training penalty's centering or normalization (Fig. 1c). The main classification study uses softmax of mean logits,

$$p_{\text{logit}}(y \mid x) = \text{softmax}\left(\frac{1}{K}\sum_m z_m(x)\right)_y, \tag{4}$$

with prediction $\hat{y} = \arg\max_c p_c$ and predictive entropy $H(p) = -\sum_c p_c \log_2 p_c$ in bits. Entropy summarizes distributional concentration without labels; Section 5.5 examines its relationship with errors.

3.3 Effective dimension and the centering constraint

Proposition 1 (Gram penalty and effective dimension)**.** *Let $u_1, \ldots, u_K$ be unit vectors, $G = U^\top U$, and $d_{\text{eff}}(G) = (\operatorname{tr} G)^2 / \operatorname{tr}(G^2)$. Then*

$$d_{\text{eff}}(G) = \frac{K}{1 + (K-1)\mathcal{L}_{\text{orth}}}. \tag{5}$$

If $\operatorname{rank}(U) \le r \le K$ with $r \ge 1$, then

$$\mathcal{L}_{\text{orth}} \ge \frac{K - r}{r(K-1)}. \tag{6}$$

The result follows from the classical frame-potential/Welch argument (Waldron, 2003): $\operatorname{tr} G = K$ and Cauchy–Schwarz bounds its nonzero eigenvalues (Appendix **??**).

For $\alpha = 1$, let $r_m = \text{vec}(R_m)$. Ensemble centering gives $\sum_m r_m = 0$. When normalization is exact, $\sum_m \|r_m\|_2 u_m = 0$, so the rank is at most $K - 1$ and

$$\mathcal{L}_{\text{orth}} \ge \frac{1}{(K-1)^2}. \tag{7}$$

For five members, the lower bound is 0.0625, or $J \ge 1.25$, and a regular simplex attains it in sufficient dimension. Exact pairwise orthogonality is infeasible. For two nonzero centered residuals, $u_2 = -u_1$, making the squared-alignment penalty constant away from the clipping region. Batch centering alone does not impose this ensemble-wise dependence.

Normalization. The unit-vector analysis applies when $\|R_m\|_F \ge \epsilon$ for every member. Epsilon clipping introduces a separate case: at complete collapse, $u_m = 0$ and $J = 0$. Thus the lower bound in Eq. (7) characterizes noncollapsed, exactly normalized residuals. Within this regime, Eq. (5) relates the realized Gram matrix to its effective dimension.

3.4 Logit geometry and predictive probabilities

Softmax satisfies $\text{softmax}(z + a\mathbf{1}_C) = \text{softmax}(z)$. A penalty on logits can respond to such shifts even though class probabilities, cross-entropy, and predictions are unchanged. Ensemble centering does not remove member-specific, example-dependent shifts.

Proposition 2 (Identical probabilities at the centered lower bound)**.** *For $K \ge 2$, $B \ge K$, and $C \ge 2$, fix any common logit matrix $Z_\star \in \mathbb{R}^{B\times C}$. There exist member logits $Z_1, \ldots, Z_K$ whose row-wise softmax distributions all equal those of $Z_\star$, while the ensemble-centered representation ($\alpha = 1$) has exactly normalized nonzero residuals and attains $\mathcal{L}_{\text{orth}} = 1/(K-1)^2$.*

The proof constructs regular-simplex residuals along softmax-invariant directions (Appendix **??**). These shifts attain the centered lower bound while preserving all class probabilities, separating logit geometry from predictive disagreement. Appendix **??** also gives orthogonal prediction deviations with correlated losses.

Table 1: Incremental effect of coupling across 20 configurations. Accuracy is in percent; Δ is coupled ($\lambda_{\text{raw}} = 0.01$) minus uncoupled ($\lambda = 0$). Positive ΔAcc and negative score differences favor coupling. All results use the final epoch, with one run per condition. Full metrics appear in Tables **??** and **??**.

Dataset	Backbone	Acc_0	$\text{Acc}_{.01}$	ΔAcc	ΔNLL	ΔBrier	ΔECE
CIFAR-10	R18	96.800	96.660	-0.140	+0.0010	+0.0007	+0.0003
	R50	97.870	97.890	+0.020	+0.0004	-0.0001	-0.0004
	VGG16	94.450	94.510	+0.060	-0.0005	-0.0007	-0.0003
	ConvNeXt-T	94.830	94.990	+0.160	-0.0018	-0.0008	+0.0004
	ViT-B/16	54.330	54.190	-0.140	+0.0013	+0.0003	-0.0001
CIFAR-100	R18	84.660	84.640	-0.020	-0.0006	+0.0001	-0.0001
	R50	87.210	87.200	-0.010	+0.0001	+0.0011	+0.0004
	VGG16	77.230	77.150	-0.080	-0.0012	-0.0006	-0.0018
	ConvNeXt-T	81.370	81.340	-0.030	+0.0016	+0.0005	+0.0012
	ViT-B/16	32.370	32.240	-0.130	+0.0050	+0.0017	+0.0004
Tiny ImageNet	R18	71.090	71.140	+0.050	+0.0044	+0.0007	+0.0014
	R50	76.860	76.530	-0.330	+0.0030	+0.0011	-0.0006
	VGG16	62.580	62.570	-0.010	-0.0056	-0.0004	-0.0047
	ConvNeXt-T	57.820	57.440	-0.380	+0.0174	+0.0048	+0.0018
	ViT-B/16	22.640	22.530	-0.110	+0.0046	-0.0000	-0.0015
ImageNet	R18	55.956	55.476	-0.480	+0.0220	+0.0045	-0.0013
	R50	62.922	62.820	-0.102	+0.0008	-0.0001	-0.0016
	VGG16	55.220	55.402	+0.182	-0.0044	-0.0009	-0.0024
	ConvNeXt-T	41.836	42.170	+0.334	-0.0155	-0.0028	+0.0010
	ViT-B/16	16.888	16.100	-0.788	+0.0585	+0.0041	-0.0029

4 Experimental setup

Main classification study. We evaluate five-member ensembles on CIFAR-10, CIFAR-100, Tiny ImageNet, and ImageNet with ResNet-18, ResNet-50, VGG16, ConvNeXt-Tiny, and ViT-B/16 (Krizhevsky, 2009; Russakovsky et al., 2015; He et al., 2016). Each backbone is fine-tuned from pretrained weights with a new classifier head and 224-pixel inputs. Both conditions use seed 42, ten epochs, batch size 64, and SGD with cosine learning-rate decay. Coupling warms linearly to $\lambda_{\text{raw}} = 0.01$ over the first 10% of steps. All results use the final checkpoint; Appendix **??** gives the optimizer and preprocessing details.

Table 1 compares coupling with $\lambda = 0$ under matched initialization, optimization, augmentation, and evaluation: 20 configurations and 40 runs, with one run per condition. Member-relative gains assess averaging.

Prediction and metrics. The main study follows the inference path in Fig. 1c and Eq. (4). The complementary CIFAR and STL-10 experiments instead use the probability mixture $p_{\text{mix}} = K^{-1} \sum_m \text{softmax}(z_m)$. We report accuracy, NLL in nats, class-summed Brier score, 15-bin ECE, and predictive entropy in bits (Appendix **??**).

Appendices **??** and **??** give protocols for the complementary training, coupling-strength, uncertainty, aggregation, and regression experiments.

5 Results

5.1 Classification performance and aggregation gains

Coupling increases accuracy in 6 of the 20 settings and decreases it in 14 (Table 1). The unweighted mean difference is -0.0972 accuracy points and the median is -0.055. Differences range from -0.788 for ImageNet ViT-B/16 to $+0.334$ for ImageNet ConvNeXt-T; all five CIFAR-100 controls have slightly higher accuracy than their coupled counterparts. The accuracy effect is therefore configuration-dependent under this fine-tuning protocol.

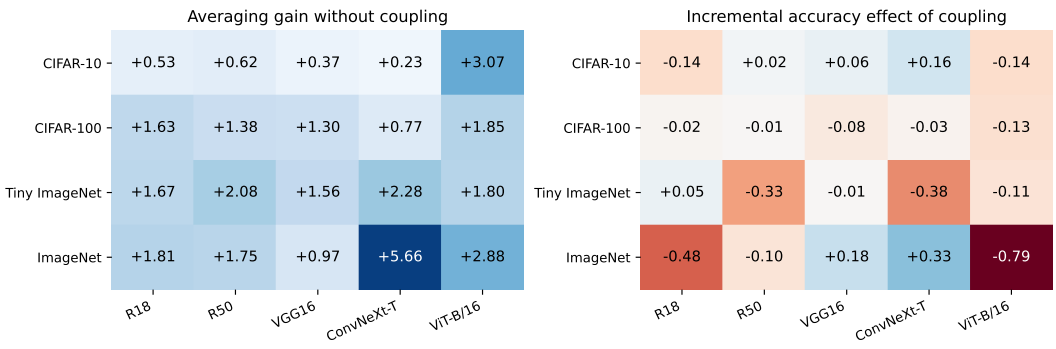


Figure 2: **Aggregation gains and coupling effects.** Left: uncoupled ensemble accuracy minus its mean member accuracy. Right: coupled minus uncoupled ensemble accuracy. Values are accuracy-point differences; panels use different color ranges. Averaging improves accuracy in every control, while the effect of coupling varies across configurations.

The effect also differs across probabilistic metrics. Coupling lowers NLL in 7 settings, Brier score in 9, and ECE in 12. The unweighted mean coupled-minus-control changes are $+0.00453$ nats, $+0.000659$, and -0.000530, respectively. The average ECE reduction accompanies small average increases in NLL and Brier score, illustrating the distinction between calibration and overall probabilistic quality. Tables **??** and **??** give the complete results.

Accuracy and calibration across backbones. The coupled CIFAR-100 ResNet-50 ensemble has higher accuracy than ConvNeXt-Tiny (87.20% versus 81.34%), but also higher ECE (0.0295 versus 0.0111). On ImageNet, ViT-B/16 has lower ECE than ResNet-50 despite worse accuracy, NLL, and Brier score. These comparisons (Table **??**) reinforce the complementary roles of accuracy, proper scores, and calibration.

All 20 ensembles in each condition exceed their average member (Fig. 2), with mean gains of 1.6764 points with coupling and 1.7107 without it. Seventeen coupled ensembles exceed their best member, compared with all 20 controls (Table **??**). Averaging thus provides a consistent benefit in both conditions, whereas the additional coupling term has a configuration-dependent effect.

5.2 Prediction geometry and error complementarity

The final-epoch, sample-weighted mean minibatch penalty J ranges from 1.2566 to 1.3983 in the coupled runs (Table **??**), numerically close to the five-member unit-vector bound of 1.25 in Proposition 1.

Pairwise correctness disagreement counts examples on which exactly one member is correct; double-fault rate counts those on which both are wrong. Coupling increases disagreement in ten settings and decreases it in ten, while reducing double faults in eight (Table **??**). Similar penalty values therefore accompany different changes in error complementarity.

5.3 Longer training and additional CIFAR comparisons

To examine behavior over a longer training horizon, Fig. 3 follows a four-member coupled CIFAR-100 ensemble trained for 200 epochs with seed 0, batch size 128, SGD learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, gradient clipping at 5, and maximum raw coupling 0.01 with 10% warmup. Accuracy rises from 57.20% at epoch ten to 81.95% at epoch 200, while NLL decreases from 1.5396 to 0.7207. Both metrics continue to improve over the longer training trajectory.

Table 2 reports four additional comparisons using probability averaging. CIFAR-100 uses five members and ten epochs; CIFAR-10 uses three members, eight epochs, batch size 48, mixed precision, learning rate 0.03, and maximum raw coupling 0.005 after 15% warmup. Coupling improves CIFAR-100 ResNet-50 but reduces accuracy in the other three settings. In particular, coupled CIFAR-10

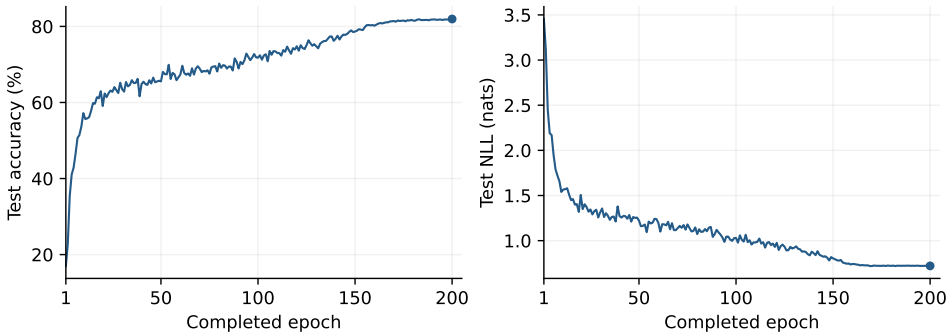


Figure 3: CIFAR-100 learning dynamics for a coupled ensemble ($K = 4$, seed 0). All 200 epochs are shown. Accuracy (percent) and NLL (nats) use mean member probabilities.

Table 2: Additional coupled and uncoupled CIFAR experiments. Accuracy is evaluated at the final epoch using probability averaging. Ensemble sizes and training budgets differ from the main matrix. The coupled CIFAR-10 VGG16 run stays at 10.00% throughout all eight epochs.

Dataset	Backbone	K	Epochs	Uncoupled (%)	Coupled (%)
CIFAR-100	ResNet-50	5	10	87.25	87.53
CIFAR-100	VGG16	5	10	74.23	73.90
CIFAR-10	ResNet-50	3	8	97.72	97.68
CIFAR-10	VGG16	3	8	94.46	10.00

VGG16 stays at 10.00%, versus 94.46% finally without coupling, highlighting the importance of backbone and coupling schedule for optimization stability.

5.4 Coupling-strength sensitivity

To examine how coupling strength affects prediction and training behavior, we compare raw coefficients $0, 10^{-3}, 10^{-2}, 10^{-1}$ on STL-10 using five pretrained ResNet-18 members, ten epochs, and probability averaging (Coates et al., 2011). Table 3 reports accuracy alongside gradient discrepancy, gradient-spectrum statistics, and absolute pairwise loss correlation. Appendix **??** defines the diagnostics and training protocol.

Table 3: STL-10 coupling-strength sweep with one run per coefficient. Gradient and spectral statistics are medians; absolute loss correlation is a mean over ten checkpoints. Accuracy is highest at $\lambda_{\text{raw}} = 10^{-2}$, while the strongest coupling produces the lowest loss correlation.

λ_{raw}	Acc. (%)	Grad. disc.	Eig. ratio	Part. ratio	Abs. corr.
0	95.775	0.7328	71.63	3.82	0.4382
10^{-3}	95.812	0.6925	77.45	4.57	0.4844
10^{-2}	95.888	0.6894	32.64	5.43	0.3818
10^{-1}	95.725	0.6036	17.91	7.15	0.3267

Accuracy peaks at $\lambda_{\text{raw}} = 10^{-2}$, whereas loss correlation is lowest at 10^{-1}. Gradient discrepancy decreases and participation ratio increases across the sweep, without a monotonic accuracy gain. The gradient participation ratio concerns a different matrix from Proposition 1. These results favor balancing diversity against task performance rather than maximizing a diversity diagnostic alone.

5.5 Predictive entropy and classification error

We next examine whether predictive entropy distinguishes correct predictions from errors in the initial CIFAR-10 and CIFAR-100 experiments. Figure 4 shows the CIFAR-100 distributions: correct predic-

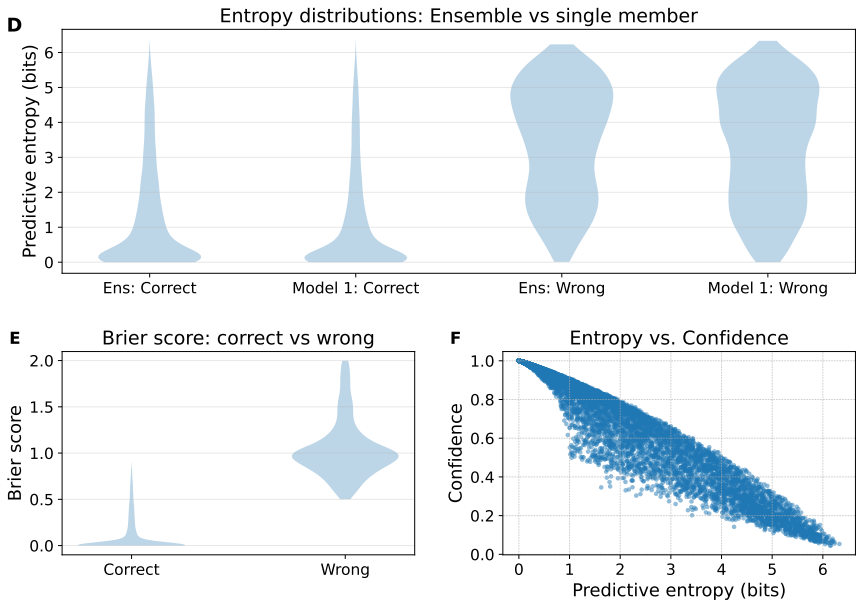


Figure 4: **Predictive uncertainty and classification errors on CIFAR-100.** Top: entropy distributions for correct and incorrect predictions of the ensemble and a representative member. Bottom left: per-example Brier scores conditional on correctness. Bottom right: confidence versus predictive entropy. Errors have higher entropy and Brier scores, with overlap between the correct and incorrect groups.

Table 4: **CIFAR-100 ensemble comparison.** All metrics use final-epoch predictions with logit averaging and 15-bin ECE. Accuracy is in percent and predictive entropy is in bits.

Method	K	Acc. (%) ↑	Brier ↓	ECE ↓	Entropy
Coupled R18	5	84.64	0.2256	0.0366	1.097
Uncoupled R18	5	84.66	0.2255	0.0367	1.0946
FiLM	4	79.39	0.2941	0.0365	1.202

R18: five pretrained ResNet-18 networks, 224-pixel inputs, ten epochs, seed 42. FiLM: shared CIFAR ResNet, 32-pixel inputs, 200 epochs from scratch, seed 0.

tions concentrate at low entropy, whereas errors extend toward higher entropy. The corresponding CIFAR-10 results appear in Fig. **??**.

Mean entropy rises from 0.390 bits on correct predictions to 1.499 on errors for VGG16, and from 0.446 to 1.999 for ResNet-50 (Table **??**). Brier scores also rise. Higher entropy on errors supports its use as an uncertainty indicator, although the distributions overlap.

5.6 FiLM COMPARISON AND AGGREGATION

Comparison of ensemble systems. The coupled ResNet-18 ensemble has higher accuracy and lower Brier score than FiLM-Ensemble (Turkoglu et al., 2022), with similar ECE (Table 4). The uncoupled ensemble reaches slightly higher accuracy. Because these systems differ in architecture and training, Table 1 provides the matched comparison needed to assess coupling.

Aggregation of fixed predictions. With FiLM predictions fixed, logit averaging increases accuracy from 78.97% to 79.39% and reduces entropy from 1.3927 to 1.2024 bits (Table 5). It also lowers Brier score, whereas probability averaging gives lower NLL and ECE. Aggregation therefore affects both accuracy and uncertainty.

Table 5: FiLM-Ensemble aggregation on CIFAR-100, evaluated from the same final-checkpoint predictions for four members and 10,000 test examples. Entropy H is in bits.

Aggregation	Acc. (%)	NLL	Brier	ECE	H
Mean probabilities	78.97	0.7907	0.2946	0.0307	1.3927
Softmax of mean logits	79.39	0.7944	0.2941	0.0365	1.2024

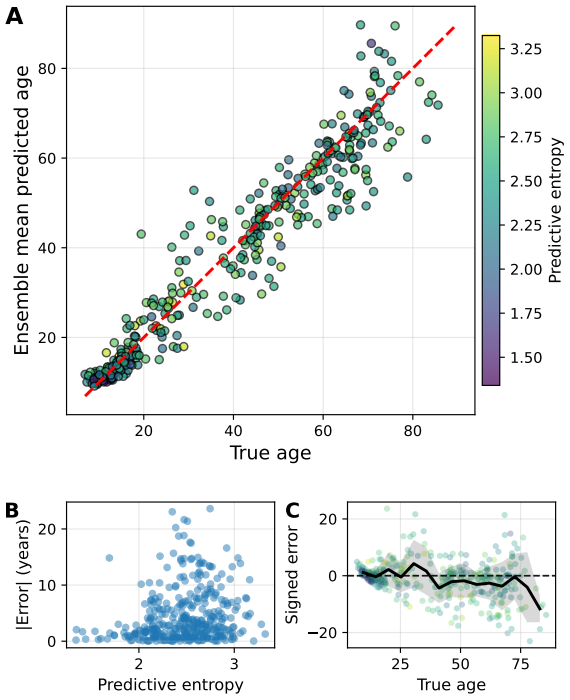


Figure 5: **Brain-age prediction, uncertainty, and age-dependent error.** (A) Ensemble-mean predicted versus chronological age, with the identity line. (B) Absolute error versus predictive entropy. (C) Signed error versus age, with a binned mean and shaded standard deviation.

5.7 Brain-age regression: uncertainty and shared error

Using NKI-Rockland FreeSurfer derivatives from Reproducible Brain Charts (Nooner et al., 2012; Shafiei et al., 2025), ten hierarchical latent networks (Xiao et al., 2024) combine regression, reconstruction, and prediction-space coupling. Features use training-set standardization; inference averages predicted ages (Appendix **??**).

Predictions follow chronological age (Fig. 5A). Higher entropy accompanies both large and small errors (panel B), while older ages are systematically underestimated (panel C). Individual uncertainty and age-conditioned signed error thus capture complementary aspects of reliability.

6 Discussion and conclusion

Prediction-space coupling links output geometry to effective dimension and centering constraints, but its lower bound permits identical class probabilities. Across 20 configurations, averaging improves mean-member accuracy, while coupling effects vary. The complementary studies support evaluating geometry alongside errors, proper scores, and calibration.

Single-run, ten-epoch results require repeated-seed validation; the exploratory brain-age study highlights shared bias that uncertainty may miss. Future work should examine longer training and representations that remove softmax-invariant directions.

AI USE STATEMENT

Generative AI was used to assist with analysis of experimental results and code, mathematical development and checking, manuscript revision, and preparation of tables and figures.

ETHICS STATEMENT

The classification results are retrospective benchmark analyses. The exploratory brain-age example concerns human neuroimaging derivatives and should not be interpreted as a diagnostic tool or as evidence of biological age acceleration. Systematic prediction bias may be shared across ensemble members and may not be signaled by disagreement. Any subsequent clinical use would require appropriate governance, population-specific evaluation, and validation beyond the analyses presented here.

REPRODUCIBILITY STATEMENT

Appendix **??** proves the geometric proposition. Appendix **??** specifies experimental protocols, metric definitions, and limitations. Appendix **??** provides full classification and member results; Appendix **??** defines the coupling diagnostics; and Appendix **??** presents the entropy–error evidence. All main-study results use the final checkpoint.

REFERENCES

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep CNNs? *arXiv preprint arXiv:1810.09102*, 2018.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Sebastian Buschjäger, Lukas Pfahler, and Katharina Morik. Generalized negative correlation learning for deep ensembling. *arXiv preprint arXiv:2011.02952*, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL `https://proceedings.mlr.press/v97/geifman19a.html`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Agostina J Larrazabal, César Martínez, Jose Dolz, and Enzo Ferrante. Orthogonal ensemble networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 594–603. Springer, 2021.

Peyman Sheikholharam Mashhadi, Sławomir Nowaczyk, and Sepideh Pashami. Parallel orthogonal deep neural network. *Neural Networks*, 140:167–183, 2021.

Kate Brody Nooner, Stanley J Colcombe, Russell H Tobe, Maarten Mennes, Melissa M Benedict, Alexis L Moreno, Laura J Panek, Shaquanna Brown, Stephen T Zavitz, Qingyang Li, et al. The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, 6:152, 2012.

Alexandre Ramé and Matthieu Cord. DICE: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=R2ZlTVPx0Gk.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Golia Shafiei, Nathalia B Esper, Mauricio S Hoffmann, Lei Ai, Andrew A Chen, Jon Cluce, Sydney Covitz, Steven Giavasis, Connor Lane, Kahini Mehta, et al. Reproducible brain charts: An open data resource for mapping brain development and its associations with mental health. *Neuron*, 113 (22):3758–3779, 2025.

Mehmet Ozgur Turkoglu, Alexander Becker, Hüseyin Anil Gündüz, Mina Rezaei, Bernd Bischl, Rodrigo Caye Daudt, Stefano D'Aronco, Jan Wegner, and Konrad Schindler. Film-ensemble: Probabilistic deep learning via feature-wise linear modulation. *Advances in neural information processing systems*, 35:22229–22242, 2022.

Shayne Waldron. Generalised Welch bound equality sequences are tight frames. Technical report, Department of Mathematics, University of Auckland, 2003. URL https://www.math.auckland.ac.nz/~waldron/Preprints/Welch-bound/CLN-02-481.pdf.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklf1yrYDr.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in artificial intelligence*, pp. 2282–2292. PMLR, 2023.

Guoxuan Xia and Christos-Savvas Bouganis. On the usefulness of deep ensemble diversity for out-of-distribution detection. *arXiv preprint arXiv:2207.07517*, 2022. URL https://arxiv.org/abs/2207.07517.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. *arXiv preprint arXiv:2404.09403*, 2024.