Diversity Is Not Enough: Why Representational Divergence Fails to Predict Fusion Utility in Pathology Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

A common heuristic for choosing which pretrained models to combine is representational diversity: prefer models whose internal representations differ most, on the assumption that divergent representations carry complementary signal. Using Centered Kernel Alignment (CKA) across four foundation-model encoders spanning three architecture families, evaluated on three histopathology classification benchmarks, we show that this heuristic can fail, and characterise when. We first establish, under controlled conditions, that architectural inductive bias—not training objective—governs representational similarity: two ViTs differing only in objective reach CKA 0.83, whereas a ViT and a Swin differing only in architecture reach 0.25. We then show that this measured divergence does not, on its own, predict fusion utility. In our pool the most representationally divergent teacher is also the weakest individually, and fusing it yields a worse result, on every dataset, than fusing a representationally redundant but individually strong teacher; across the teacher additions we examine, individual competence tracks fusion utility while representational distinctness does not. Because divergence and competence are entangled in our encoders, we present this as a concrete demonstration that diversity is not sufficient for utility rather than a general law, and argue that competence must be weighed alongside distinctness in any similarity-based selection. Building on this, a lightweight (11–31 parameter) class-conditional logit fusion improves accuracy over the strongest single model by +2.99% pooled ($p < 10^{-4}$, Cohen's $d = 3.70$) and reduces expected calibration error by 85–96%, with both effects persisting under cross-institutional distribution shift, while fusion architectures with up to 5×10^5 parameters fail to exceed the dominant teacher. All results are reproducible from released features, splits, and code.

1 Introduction

When several strong pretrained encoders are available, a practitioner must decide which of them to combine. A common answer appeals to representational diversity. The reasoning is that encoders which represent inputs differently will tend to make different predictions, so a useful combination should favour models whose representations diverge and set aside those that are redundant. Centered Kernel Alignment (CKA) (Kornblith et al., 2019) makes this reasoning operational: one measures how similarly two models represent the data, keeps the dissimilar ones, and prunes the similar ones. Recent work has begun to use representational similarity in exactly this way, as a criterion for comparing and selecting models (Klabunde et al., 2025; Mishra & Lotter, 2025).

This paper tests that criterion and finds that, at least in the pool we study, it does not hold. What CKA measures and what fusion rewards come apart sharply enough to reverse the selection a practitioner would make. We studied four foundation-model encoders drawn from three architecture families and evaluated them on three histopathology classification benchmarks. On every benchmark, the most representationally divergent teacher (CKA 0.25 with the reference model) produced a worse fusion than the most representationally redundant teacher (CKA 0.83). A practitioner following the diversity criterion would, in each case,

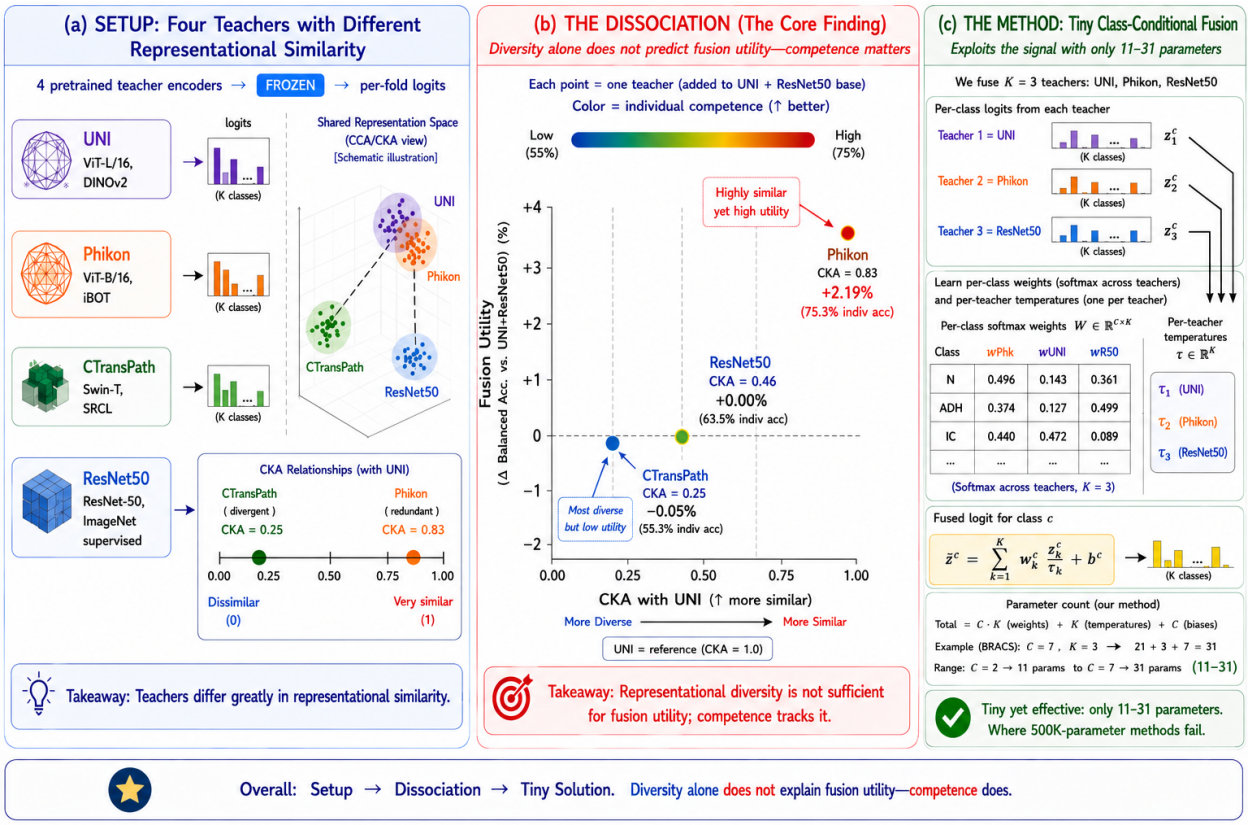

Figure 1: Overview of our study. (a) Four frozen teacher encoders that differ in representational similarity to UNI, measured by CKA: Phikon-v2 is redundant (CKA 0.83), CTransPath is divergent (CKA 0.25). (b) The dissociation. Each point is a teacher added to a UNI+ResNet50 base, positioned by its CKA with UNI and its marginal fusion utility, and coloured by individual competence. The most divergent teacher (CTransPath) gives near-zero utility, while the most redundant (Phikon-v2) gives the highest; utility tracks competence, not diversity. (c) The class-conditional fusion, with 11 to 31 parameters, that exploits this signal where far larger methods fail.

have chosen the worse of the two. We are explicit about the scope of this claim: in our pool the divergent teacher is also the weakest, so divergence and competence are entangled, and we therefore establish that diversity is not *sufficient* for utility rather than that it is uninformative in general (Section 4.2, Section 6).

The cause is that CKA reports representational divergence, whereas fusion utility depends on divergence together with individual competence. The most divergent teacher in our pool was also the weakest, so its distance from the others reflected the inputs it misclassified rather than exploitable signal. Across every teacher we added, individual competence tracked fusion utility while representational distinctness did not (Section 4.2): diversity is necessary for complementarity but not sufficient, and on its own it misleads selection.

Two methodological points were needed to establish this cleanly. Because every quantity in the argument derives from per-teacher logits, and a probe fit and evaluated on the same data inflates all of them, we generate all logits under a strict per-fold protocol and release the features, splits, and code so that each number is reproducible from public data (Section 3.2, Appendix B). And to show that the finding concerns the fusion problem rather than any one fusion design, we evaluate a range of learned methods: none exceeds the strongest single teacher, some carrying up to half a million parameters, whereas a class-conditional scheme of 11 to 31 parameters does. The limiting factor is the structure of the complementary signal, not model capacity.

The contributions of this work are as follows.

- We report a dissociation between representational diversity and fusion utility. Across three datasets and every teacher addition we examined, the most representationally divergent teacher gave the worst fusion while individual competence tracked usefulness. Because divergence and competence are entangled in our four encoders, we establish this as a demonstration that diversity is not sufficient for utility, not as a general law, and argue against selecting models on representational similarity alone without accounting for competence.

- We give controlled evidence on what drives representational similarity. With architecture separated from training objective and from training domain, architectural inductive bias rather than objective governed CKA among these encoders. This is a prerequisite for interpreting the dissociation, and it runs counter to the view that representations converge in this setting.

- We show that fusion capacity is not the obstacle. Fusion architectures with up to 5×10^5 parameters failed to exceed the dominant teacher, while an 11 to 31 parameter class-conditional method improved accuracy (pooled +2.99%, $p < 10^{-4}$) and calibration (85 to 96% ECE reduction), with both effects holding under cross-institutional distribution shift.

- We document the leakage-sensitive per-fold logit protocol on which all results depend, and we release the features, splits, and code that regenerate every reported number.

2 Related Work

2.1 Representational similarity and the diversity assumption

CKA (Kornblith et al., 2019) has become the standard way to compare representations across networks of different width and architecture, where CCA and SVCCA (Raghu et al., 2017) are less dependable. Klabunde et al. (2025) surveyed measures of network similarity and separated representational similarity, computed from activations, from functional similarity, computed from outputs; they observed that the two can diverge when architectures differ, a point our results bear on directly. Later work refined the metric, with manifold-aware alignment (Islam et al., 2025) addressing its sensitivity to outliers and an analysis of the large-bandwidth limit (Alvarez, 2022) establishing the soundness of linear CKA. Building on this reliability, several authors have used representational similarity to compare and select models (Nakai et al., 2025; Mishra & Lotter, 2025).

That use rests on an assumption: that representational divergence signals functional complementarity, so that models which represent inputs differently make usefully different predictions and are good candidates to combine. We tested the assumption and found that it fails. A representationally divergent model can be functionally redundant when it is individually weak, and a redundant model can be functionally valuable when it is strong.

2.2 Whether representations converge

There is disagreement over whether models with different architectures arrive at similar representations. Huh et al. (2024) advanced the Platonic Representation Hypothesis, under which representations move toward a shared structure as models grow, largely irrespective of architecture or objective. Wu et al. (2025a) reached a related conclusion from a data-driven comparison across metrics, reporting that self-supervised training could override architectural bias and group self-supervised CNNs with self-supervised ViTs. Our controlled comparison points the other way within a single domain. Holding the training domain fixed and changing only the architecture lowered CKA to 0.25, whereas holding the architecture fixed and changing only the objective left it at 0.83. Architecture governed representational similarity in our setting, though, as we go on to show, that divergence did not carry through to fusion utility.

2.3 Ensemble diversity and model fusion

The idea that diversity helps combination is long established. Kuncheva & Whitaker (2003) showed that ensemble accuracy gains are bounded by classifier diversity, and Wood et al. (2023) developed an information-theoretic account of the same relationship; oracle analysis (Brown et al., 2005) supplies the headroom benchmark we adopt. These results concern diversity among outputs. Our contribution is to show that representational diversity, as read from CKA, is a poor proxy for the output-level complementarity those analyses assume, since two models can be far apart representationally yet redundant in their predictions. For the fusion of pretrained encoders, earlier methods combined models through mixtures of experts, multi-teacher distillation, and attention-based fusion, typically with sizeable trainable components (Wu et al., 2025b; Zheng et al., 2023; Ativi et al., 2026; Huang et al., 2026). Our results take the opposite view: parameter-rich per-input fusion did not exceed the dominant teacher, whereas a class-conditional scheme of 11 to 31 parameters did, which locates the difficulty in the structure of the complementary signal rather than in model capacity.

2.4 Calibration

Deep networks are often miscalibrated (Guo et al., 2017), and ECE (Naeini et al., 2015) remains the standard measure despite documented shortcomings (Futami & Fujisawa, 2024). Class imbalance skews predicted probabilities toward the majority class (Nunes et al., 2025), and self-supervised models grow overconfident under distribution shift (Xu & Rajpurkar, 2024), with post-hoc and multi-branch methods reducing the error (Alkurdi & Sallow, 2026). Our method obtains calibration from the per-teacher temperatures it already learns for fusion, without a separate stage, and the improvement held under cross-institutional shift.

3 Method

We study a pool of M frozen foundation-model encoders as teachers and ask which subset, if any, should be fused. Our analysis proceeds in three parts: a set of representational and functional diagnostics on the teacher pool (Section 3.1), a lightweight class-conditional logit-fusion method (Section 3.3), and a representational-diversity selection heuristic that we state precisely and then test (Section 3.4). All logits are produced under a strict per-fold protocol (Section 3.2) that we describe in detail, as its correctness is a prerequisite for every downstream quantity.

3.1 Diagnostics on the Teacher Pool

Representational similarity. For teacher feature matrices $\boldsymbol{F}_i \in \mathbb{R}^{n \times d_i}$ and $\boldsymbol{F}_j \in \mathbb{R}^{n \times d_j}$ on a shared set of n inputs, we measure representational similarity with linear Centered Kernel Alignment (Kornblith et al., 2019):

$$\mathrm{CKA}(\boldsymbol{F}_i, \boldsymbol{F}_j) = \frac{\|\boldsymbol{F}_j^\top \boldsymbol{F}_i\|_F^2}{\|\boldsymbol{F}_i^\top \boldsymbol{F}_i\|_F \, \|\boldsymbol{F}_j^\top \boldsymbol{F}_j\|_F}, \tag{1}$$

with column-centred features. Values near 1 indicate near-identical representations; low values indicate representational divergence. CKA is a function of the features alone and is independent of the classification logits, so it is unaffected by how those logits are produced. Unlike CCA or SVCCA (Raghu et al., 2017), CKA compares representations reliably across networks of differing width and architecture.

Oracle upper bound. For a teacher set, the oracle accuracy is the fraction of inputs that at least one teacher classifies correctly,

$$\mathrm{acc}_{\mathrm{oracle}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[\, \exists\, m : \hat{y}_i^{(m)} = y_i \,\Big], \tag{2}$$

and bounds the accuracy any fusion of that set can achieve. The oracle headroom $\Delta = \mathrm{acc}_{\mathrm{oracle}} - \max_m \mathrm{acc}_m$ is the maximum attainable gain over the best single teacher.

Per-teacher and per-class structure. For each teacher we compute its *unique contribution*: the fraction of inputs that only it classifies correctly. For each class we compute the *all-wrong rate*: the fraction of inputs no teacher classifies correctly, separating classes where fusion can help from those where no teacher succeeds. These diagnostics describe where complementary signal exists; a central finding of this work (Section 4.2) is that representational divergence among teachers does not, by itself, imply such signal is exploitable.

3.2 Per-Fold Logit Protocol

All fusion and oracle quantities depend on per-teacher class logits, which we obtain from linear probes over frozen features under strict K-fold cross-validation. Within each fold, a probe is fit on that fold's training partition only and produces logits for the held-out partition only; feature standardisation statistics are likewise estimated on the training partition alone. Concatenating held-out predictions across folds yields one logit vector per input, with the guarantee that no input contributes to the probe that generates its own logit. This protocol is not incidental: fitting a probe on all data and scoring the same data inflates apparent teacher accuracy and, in turn, every oracle bound, unique-contribution count, and fusion gain derived from it. We therefore treat per-fold logit generation as part of the method and release the exact procedure (Appendix B) so that all reported numbers are reproducible from public features.

3.3 Class-Conditional Logit Fusion

Our fusion method is deliberately minimal, because our experiments (Section 4.3) show that the obstacle to fusing strong teachers is not model capacity but the structure of the complementary signal. Learned fusion architectures that route *per input*—deciding case by case which teacher to trust—fail to exceed the dominant teacher, since the features that would indicate when to override it are the same features used to classify. We therefore route *per class*. Because inputs vastly outnumber classes ($N \gg C$), class-level teacher preferences are stably estimable from data whereas per-input preferences are not.

For K selected teachers and C classes, we learn class-conditional weights $\boldsymbol{W} \in \mathbb{R}^{C \times K}$ (softmax-normalised per class), per-teacher temperatures $\in \mathbb{R}^K$, and per-class biases $\boldsymbol{b} \in \mathbb{R}^C$. The fused logit for class c is

$$z_{\text{fused}}(c) = \sum_{k=1}^{K} \alpha_{c,k} \frac{z_k(c)}{\tau_k} + b_c, \qquad \alpha_{c,:} = \text{softmax}(\boldsymbol{W}_{c,:}), \quad \tau_k = \exp(\hat{\tau}_k). \tag{3}$$

The total parameter count is $CK + K + C$, between 11 and 31 parameters in our experiments. Training minimises cross-entropy with an entropy regulariser $\mathcal{H}() = -\frac{1}{C}\sum_c \sum_k \alpha_{c,k} \log(\alpha_{c,k} + \epsilon)$, weighted by $\lambda = 0.01$, which discourages weight from collapsing onto a single dominant teacher while still permitting single-teacher reliance where it helps. Two properties matter downstream: the per-class softmax makes $\alpha_{c,:}$ directly interpretable as the trust placed in each teacher for class c, and the learned temperatures τ_k rescale logit magnitudes, yielding improved calibration as a byproduct of fusion (Section 4.4). Optimisation uses AdamW with early stopping.

3.4 A Diversity Heuristic for Teacher Selection

Given the diagnostics above, a natural strategy for choosing which teachers to fuse is to prefer representational diversity: retain teachers whose representations diverge (low CKA) and prune those that are redundant (high CKA). Algorithm 1 states this heuristic. It prunes a teacher only when it is both representationally redundant—CKA above θ_{CKA} with a retained teacher—and functionally minor—unique contribution below θ_{unique}—and otherwise retains representationally distinct teachers.

We present this heuristic not as a contribution but as a hypothesis to be tested. It formalises the widely-held intuition that representational divergence implies complementarity, and it makes a concrete, falsifiable prediction about which subset to fuse. Section 4.2 shows the prediction is wrong: the subset this heuristic selects is outperformed, on every dataset, by a subset it prunes. We therefore use Algorithm 1 as the object of study rather than a recommended procedure.

Algorithm 1 CKA diversity heuristic for teacher selection (tested, not endorsed).

Require: CKA matrix $\boldsymbol{C}$, unique contributions $\{u_m\}$, teacher set $\mathcal{T}$, thresholds $\theta_{\text{CKA}}, \theta_{\text{unique}}$
Ensure: Diversity-selected subset $\mathcal{T}^{\text{div}}$

```
T^div ← T
for each pair (T_i, T_j) with C_ij > θ_CKA do
    T_weak ← arg min(u_i, u_j)
    if u_weak < θ_unique then
        T^div ← T^div \ {T_weak}
    end if
end for
return T^div
```

Table 1: Pairwise CKA (BRACS). Varying architecture drops CKA far more than varying the training objective.

Pair	Differs in	CKA
UNI – Phikon-v2	objective	0.834
UNI – CTransPath	architecture	0.246
UNI – ResNet50	arch. + domain	0.450

4 Experiments

Setup. We evaluate four frozen encoders—UNI (ViT-L), Phikon-v2 (ViT-B), CTransPath (Swin-T), and a supervised ResNet-50—on BRACS (7 classes), SICAPv2 (4 classes), and MHIST (2 classes). All per-teacher logits are generated by the per-fold protocol of Section 3.2; all reported numbers are reproducible from released artifacts (Appendix B). We report accuracy and expected calibration error (ECE, 15 bins) under 5-fold cross-validation, and use the BRACS held-out set (570 samples from different institutions) for distribution-shift evaluation.

4.1 Architecture Governs Representational Similarity

We first establish what drives representational divergence among the teachers, as this is needed to interpret the dissociation that follows. Holding training domain fixed and varying only architecture (UNI, a ViT, vs. CTransPath, a Swin, both pathology-pretrained) drives CKA to 0.246 on BRACS; holding architecture fixed and varying only the self-supervised objective (UNI vs. Phikon-v2, both ViTs) leaves CKA at 0.834 (Table 1). The gap is attributable to architecture alone and is consistent across all three datasets. Architecture, not training objective or domain, is the dominant driver of representational similarity here—the natural reading of which is that the Swin teacher, being most representationally distinct, should be the most valuable to fuse. The next subsection shows this reading is wrong.

4.2 Representational Diversity Does Not Predict Fusion Utility

We now test whether representational divergence predicts fusion utility. Starting from a fixed UNI+ResNet50 base, we add each candidate third teacher and measure the resulting fusion accuracy (Table 2). The result is consistent across all three datasets: adding the redundant ViT (Phikon-v2, CKA 0.69–0.83 with UNI) yields a better fusion than adding the divergent Swin (CTransPath, CKA 0.006–0.32), by +2.24%, +1.48%, and +1.49% on BRACS, SICAPv2, and MHIST respectively. The teacher that representational diversity would select is, in every case, the worse choice.

The mechanism is visible when individual competence is placed alongside diversity (Table 3). The divergent teacher is also the weakest (55–61% individual accuracy), and its marginal contribution to fusion is near zero or negative. The redundant teacher is individually strong (82–89%) and contributes the largest marginal gain

Table 2: Adding a third teacher to a UNI+ResNet50 base. The high-CKA (redundant) teacher yields a better fusion than the low-CKA (divergent) teacher on every dataset.

Added teacher	CKA w/ UNI	Indiv. acc.	Fusion acc.
BRACS			
Phikon-v2 (redundant)	0.834	75.3%	**78.81%**
CTransPath (divergent)	0.246	55.3%	76.62%
SICAPv2			
Phikon-v2 (redundant)	0.696	89.2%	**93.10%**
CTransPath (divergent)	0.317	73.4%	91.62%
MHIST			
Phikon-v2 (redundant)	0.688	82.0%	**86.55%**
CTransPath (divergent)	0.006	60.5%	85.06%

Table 3: Competence versus diversity (BRACS). Marginal fusion gain tracks individual accuracy, not CKA-measured divergence.

Teacher	Indiv. acc.	CKA w/ UNI	Marginal Δ
Phikon-v2	75.3%	0.834	+2.19%
ResNet50	63.5%	0.450	+0.00%
CTransPath	55.3%	0.246	−0.05%

of any teacher. What predicts a teacher's fusion utility is its competence, not its representational distinctness from the reference; CKA, which captures only the latter, is therefore a misleading selection signal.

To separate competence from diversity across the whole pool rather than in a single comparison, Table 4 lists both variables against marginal fusion utility for every teacher and dataset. The orderings come apart. Pooling across all teacher and dataset pairs, the Spearman rank correlation between competence and utility is strongly positive ($\rho = 0.89$), whereas the correlation between diversity and utility is weak and negative ($\rho = -0.24$): representational divergence is, if anything, slightly associated with lower rather than higher utility. The least confounded teacher, ResNet50, sits at intermediate diversity and intermediate competence, and its utility matches what competence predicts. With four teachers these coefficients are underpowered, and we report them only as a compact summary of an ordering visible directly in the table; the qualitative pattern, that utility follows competence and not diversity, holds independently on all three datasets.

4.3 Fusion Capacity Is Not the Obstacle

If representational diversity does not identify useful teachers, perhaps a sufficiently expressive fusion model can extract complementary signal regardless. On BRACS, it cannot. Table 5 evaluates fusion methods spanning five orders of magnitude in parameter count, all consuming the same clean logits; we report this comparison on BRACS, the largest-class-count dataset, and note that verifying it across all three datasets is a natural extension. None exceeds the strongest single teacher by a meaningful margin: they cluster near UNI's 75.2% regardless of capacity. Our class-conditional method, with 31 parameters, reaches 78.8%. The obstacle is not model capacity but the information available: per-input routing requires anticipating which teacher is correct for each input, and the features that would carry that signal are the same features used to classify, so the routing problem is close to unlearnable (Section 5.2).

4.4 Accuracy, Calibration, and Robustness to Shift

The class-conditional fusion of the competence-selected combination (Phikon-v2+UNI+ResNet50) improves accuracy over the strongest single teacher on all three datasets: +3.60% (BRACS), +3.15% (SICAPv2),

Table 4: Diversity, competence, and marginal fusion utility per teacher. Marginal utility follows the competence ordering, not the diversity ordering, on every dataset. Pooled Spearman ρ: competence–utility = 0.89; diversity–utility = -0.24.

Dataset	Teacher	CKA w/ UNI	Indiv. acc.	Marg. util.
BRACS	Phikon-v2	0.834	75.3%	+2.19%
	ResNet50	0.450	63.5%	+0.00%
	CTransPath	0.246	55.3%	−0.05%
SICAPv2	Phikon-v2	0.696	89.2%	+1.48%
	ResNet50	0.477	76.2%	+0.00%
	CTransPath	0.317	73.4%	+0.00%
MHIST	Phikon-v2	0.688	82.0%	+1.49%
	ResNet50	0.496	77.5%	+0.00%
	CTransPath	0.006	60.5%	+0.00%

Table 5: Fusion methods on BRACS. Parameter-rich methods fail to exceed the dominant teacher; a 31-parameter class-conditional method does.

Method	Params	Acc.
UNI (best single)	0	75.21%
Equal ensemble	0	75.38%
MoE fusion	∼327K	74.68%
Cross-attention	∼559K	75.13%
Concat + MLP	∼5K	76.34%
Class-conditional (ours)	31	**78.81%**

+2.22% (MHIST), winning on every one of the 5 folds in each case (Table 6). Pooled across all 15 folds, the mean gain is +2.99% with $p < 10^{-4}$ and Cohen's $d = 3.70$ (Wilcoxon $p = 6 \times 10^{-5}$). The learned per-teacher temperatures also sharply improve calibration, reducing ECE by 85–96%. On the BRACS held-out set, the accuracy gain grows to +4.74% (57.7% → 62.5%) and calibration also improves under shift (ECE 0.316 → 0.089), indicating both effects are robust to cross-institutional distribution change rather than confined to the training distribution.

5 Discussion

5.1 Why Representational Diversity Fails to Predict Fusion Utility

The central finding of this work is a dissociation between two properties that are commonly conflated. Representational similarity, measured by CKA, captures how differently two models encode their inputs. Fusion utility captures how much combining them improves predictions. The intuition underlying diversity-based selection is that the former predicts the latter. Our results show it does not: the most representationally divergent teacher yields a worse fusion than the most representationally redundant one, on every dataset we study.

This has a direct methodological consequence. Representational-similarity metrics are increasingly proposed as a basis for model selection, on the assumption that they identify complementary models. Our results caution against this use: a low-CKA candidate is not, on that basis alone, a useful one to add. A sound selection criterion must weigh individual competence alongside representational distinctness. We do not offer a validated positive criterion—establishing and testing one is future work—but the negative result is

Table 6: Accuracy and ECE (15 bins) under 5-fold CV. Best in bold.

Method	BRACS Acc.	BRACS ECE	SICAPv2 Acc.	SICAPv2 ECE	MHIST Acc.	MHIST ECE
UNI (best single)	75.21	.174	89.97	.069	84.30	.119
Equal ensemble	75.38	.132	89.53	.061	85.06	.089
Phk+UNI+R50 (ours)	**78.81**	**.027**	**93.11**	**.003**	**86.52**	**.010**

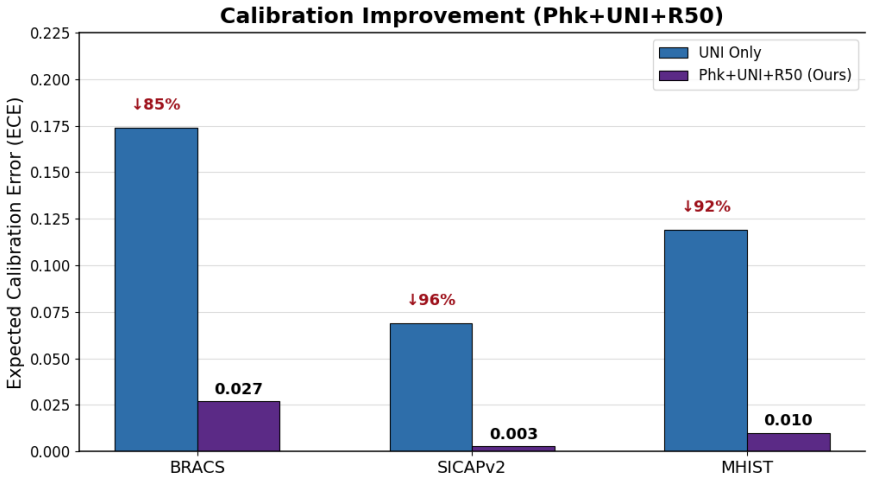

Figure 2: Expected calibration error (ECE) for the best single teacher (UNI) and the fusion, on each dataset. Fusion reduces ECE by 85 to 96% in distribution.

clear and, we expect, relevant wherever practitioners select among a growing pool of pretrained models to combine.

5.2 Capacity Is Not the Obstacle to Fusion

A second finding, established on BRACS, concerns why elaborate fusion architectures underperform: the obstacle is a mismatch between the learning problem and the available information, not a shortage of capacity. Routing at the class level rather than the input level changes the granularity of the decision. The claim "teacher A resolves class c better than teacher B" is a stable, estimable statistic; the claim "teacher A is correct on input i" is not learnable from these representations. Reducing N per-input decisions to C per-class decisions converts an ill-posed problem into a well-posed one, which is why so few parameters suffice. This also delimits the method's scope: it is expected to help when $C \ll N$ and per-class teacher rankings are non-uniform, and to offer little when they are uniform.

5.3 Calibration as a Byproduct of Fusion

The accuracy gains are meaningful but modest; the calibration gains are large. Because the method learns a temperature per teacher, it rescales logit magnitudes as part of fusion, and the resulting predictions are substantially better calibrated without any separate post-hoc stage. This matters wherever predicted confidence, and not only the top-1 label, is consumed downstream—for triage, deferral, or selective prediction. Notably, for the selected combination the calibration improvement persists under cross-institutional distribution shift, the regime in which the constituent models are individually most overconfident. That reliability, rather than the accuracy margin, may be the more consequential property for deployment.

6 Limitations and Future Directions

Scope of the dissociation. Our central claim is a demonstrated dissociation between representational diversity and fusion utility, established across three datasets and four teachers from three architecture families. We do not claim it is universal. In particular, the most divergent teacher in our pool is also the weakest, which is precisely the configuration that produces the dissociation; whether the same pattern arises in pools where divergence and competence are differently distributed is an open question. Enlarging the teacher pool—more architecture families, and encoders whose individual strength and representational distinctness vary independently—would test the generality of the finding and is the most important direction for future work.

Measurement of similarity. We use linear CKA, the standard choice for cross-architecture comparison. Kernel or manifold-aware variants might capture structure that linear CKA misses. However, our claim concerns the relationship between a widely-used similarity measure and realised utility, not the choice of measure; establishing that a different similarity metric *does* predict utility would be a complementary positive result and a valuable extension.

Frozen features. Logits are produced by linear probes over frozen features. End-to-end fine-tuning could alter both individual teacher strength and the resulting complementarity, potentially changing the diversity–utility relationship, at higher computational cost. Characterising the dissociation under fine-tuning is a natural next step.

A positive selection criterion. We establish a negative result—that representational diversity alone does not predict fusion utility—but do not provide a validated criterion that jointly accounts for distinctness and competence. Constructing such a criterion, for instance by weighting CKA-based diversity by per-teacher reliability, and testing whether it predicts fusion utility out of sample, would convert our negative finding into a positive method.

Beyond the present domain. Our experiments use histopathology as a testbed because it provides encoders matched on some factors and differing on others, enabling the controlled comparison. The finding itself concerns representations and fusion rather than the domain, and testing whether "divergent but weak loses to redundant but strong" holds in other multi-source settings would establish how far it generalises.

7 Conclusion

We asked whether representational diversity predicts which pretrained models are worth combining, and found that it does not. Across four encoders from three architecture families and three classification benchmarks, the teacher that was most distinct from the reference in representation was also the weakest in isolation, and adding it to the pool produced a worse fusion than adding a representationally redundant but individually strong teacher, on every dataset. What tracked fusion utility was individual competence, not representational distance. Representational similarity metrics such as CKA therefore describe a real property of a model pool, but that property is not, on its own, a sound basis for deciding which models to fuse. Two further findings supported this conclusion. Separating architecture from training objective and domain showed that architecture governed representational similarity among these encoders, which established that the divergent teacher really was divergent for structural reasons rather than incidental ones. Evaluating a range of learned fusion methods showed that the failure to exploit complementary signal was not a matter of insufficient capacity: methods with up to half a million parameters did not exceed the strongest single teacher, while a class-conditional scheme with between 11 and 31 parameters improved accuracy by 2.99% pooled and reduced calibration error by 85 to 96%, with both gains holding under cross-institutional distribution shift, where the accuracy gain grew to 4.74%. The limiting factor was the structure of the complementary signal, which is accessible at the level of classes but not of individual inputs. We do not offer a validated criterion for teacher selection that accounts for competence alongside distinctness, and constructing one is the natural next step. So too is testing whether the dissociation persists in pools where representational divergence and individual competence are less closely tied than they were here, and in domains beyond the one we used as

a testbed. The finding itself concerns representations and fusion rather than any particular application, and we have released the features, splits, and code so that it can be checked and extended directly.

References

Ahmed AH Alkurdi and Amira Bibo Sallow. Calibrated transformer fusion for dual-view low-energy cesm classification. *Journal of Imaging*, 12(1):41, 2026.

Sergio A Alvarez. Gaussian rbf centered kernel alignment (cka) in the large-bandwidth limit. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):6587–6593, 2022.

Wisdom Xornam Ativi, Wenyu Chen, Lazarus Kwao, Williams Ayivi, Francis Sam, Ali Alqahtani, Yeong Hyeon Gu, and Mugahed A Al-Antari. Cmaf-net: cross-modal attention fusion with information-theoretic regularization for imbalanced breast cancer histopathology. *Scientific Reports*, 2026.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20, 2005.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

Futoshi Futami and Masahiro Fujisawa. Information-theoretic generalization analysis for expected calibration error. *Advances in neural information processing systems*, 37:84246–84297, 2024.

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 CVPR*, pp. 770–778, 2016.

Gexin Huang, Anqi Li, Yusheng Tan, Beidi Zhao, Gang Wang, Zu-Hua Gao, and Xiaoxiao Li. Plug-and-play logit fusion for heterogeneous pathology foundation models. *arXiv preprint arXiv:2604.07779*, 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Mohammad Tariqul Islam, Du Liu, and Deblina Sarkar. Manifold approximation leads to robust kernel alignment. *arXiv preprint arXiv:2510.22953*, 2025.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9): 1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

Vaibhav Mishra and William Lotter. Comparing computational pathology foundation models using representational similarity analysis. *arXiv preprint arXiv:2509.15482*, 2025.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Toshiki Nakai, Ravikiran Chikkala, Lena Oberkircher, Nicholas Jennings, Natalia Skachkova, Tatiana Anikina, and Jesujoba Alabi. Treplina: Layer-wise cka+ repina alignment improves low-resource machine translation in aya-23 8b. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pp. 25–34, 2025.

João D Nunes, Diana Montezuma, Domingos Oliveira, Tania Pereira, Inti Zlobec, and Jaime S Cardoso. Beyond accuracy: The role of calibration in computational pathology. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2025.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Julio Silva-Rodríguez. Sicapv2-prostate whole slide images with gleason grades annotations. *Mendeley Data*, 1:2020, 2020.

Xiyue Wang, Sen Yang, Jun Zhang, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *International Conference on Artificial Intelligence in Medicine*, pp. 11–24, 2021.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359):1–49, 2023.

Jialin Wu, Shreya Saha, Yiqing Bo, and Meenakshi Khosla. A data-driven typology of vision models from integrated representational metrics. *arXiv preprint arXiv:2509.21628*, 2025a.

Junxian Wu, Minheng Chen, Xinyi Ke, Tianwang Xun, Xiaoming Jiang, Hongyu Zhou, Lizhi Shao, and Youyong Kong. Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5144–5153, 2025b.

Jenny Xu and Pranav Rajpurkar. Calibration and uncertainty estimation challenges in self-supervised chest x-ray pathology classification models. In *Medical Imaging with Deep Learning*, 2024.

Yuchao Zheng, Chen Li, Xiaomin Zhou, Haoyuan Chen, Hao Xu, Yixin Li, Haiqing Zhang, Xiaoyan Li, Hongzan Sun, Xinyu Huang, et al. Application of transfer learning and ensemble learning in image-level classification for breast histopathology. *Intelligent Medicine*, 3(02):115–128, 2023.

A Appendix

B Reproducibility

Every result in this paper can be reproduced from publicly available features and the released code. This appendix documents the data, the per-fold logit protocol on which all downstream quantities depend, and the artifacts released for verification.

B.1 Data

We evaluate on three public histopathology classification benchmarks. BRACS (Brancati et al., 2022) contains 3,969 regions of interest across seven breast lesion subtypes, together with a separate test set of 570 regions drawn from different institutions, which we use for the distribution-shift evaluation. SICAPv2 (Silva-Rodríguez, 2020) contains 12,081 prostate biopsy patches labelled with four Gleason grades. MHIST (Wei

et al., 2021) contains 3,152 colorectal polyp images in a binary classification of hyperplastic polyps against sessile serrated adenomas. All three datasets are available from their original sources. Teacher features are extracted once by frozen inference and cached; no encoder is fine-tuned.

B.2 Teacher encoders

Four encoders span three architecture families. UNI (Chen et al., 2024) is a ViT-L/16 trained with DINOv2 on histopathology. Phikon-v2 (Filiot et al., 2024) is a ViT-B/16 trained with iBOT on histopathology. CTransPath (Wang et al., 2022) is a Swin-T trained with semantically relevant contrastive learning on histopathology. ResNet50 (He et al., 2016) is a supervised ImageNet model, included as a maximally distinct reference. UNI and Phikon-v2 share an architecture but differ in objective, and UNI and CTransPath share a training domain but differ in architecture, which allows architecture and objective to be separated in the CKA comparison of Section 4.1.

B.3 Per-fold logit generation

Per-teacher class logits are produced by linear probes over the frozen features under 5-fold stratified cross-validation with a fixed seed. The procedure enforces train and validation separation at every step, so that no input contributes to the probe that produces its own logit.

1. The N inputs are partitioned into five stratified folds using fixed, released indices. The fold indices have no train and validation overlap, which we verify programmatically.

2. For each fold, a `StandardScaler` is fit on the training partition of that fold only, and both partitions are transformed with these statistics. No validation data influences the scaling.

3. A multinomial logistic regression probe (`scikit-learn` `LogisticRegression`, $C = 1.0$, `max_iter`$=5000$, `solver=lbfgs`) is fit on the standardised training partition and produces logits for the held-out partition only.

4. The held-out logits from all five folds are concatenated into one logit matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times C}$ per teacher, aligned to the original input order.

A probe that is fit on all inputs and then scored on the same inputs, rather than following this procedure, inflates apparent teacher accuracy, and that inflation propagates into every oracle bound, unique-contribution count, and fusion gain. We therefore treat per-fold generation as part of the method and release the generation script so that the logits can be regenerated exactly.

B.4 Released artifacts

The release contains the cached teacher features and fixed fold indices for all three datasets, the per-fold logit-generation script, the class-conditional fusion implementation of Equation 3, and a single driver script that regenerates every number reported in the main text: per-teacher accuracies, the CKA matrices, oracle bounds and unique contributions, the diversity-against-utility comparison, the fusion results with their pooled statistics, and the held-out test-set results. Running the driver on the released features reproduces all reported tables.

B.5 Compute

Feature extraction aside, the full pipeline runs on a single commodity GPU in minutes. Because the encoders are frozen and fusion operates on cached logits with at most 31 trainable parameters, the analysis and fusion stages run on CPU.

Table 7: Pairwise CKA for all teacher pairs, computed from frozen features.

Pair	Differs in	BRACS	SICAPv2	MHIST
UNI – Phikon-v2	objective	0.834	0.696	0.688
UNI – CTransPath	architecture	0.246	0.317	0.006
UNI – ResNet50	arch. + domain	0.450	0.477	0.496
Phikon-v2 – CTransPath	arch. + obj.	0.322	0.422	0.004
Phikon-v2 – ResNet50	arch. + domain	0.425	0.490	0.440
CTransPath – ResNet50	arch. + domain	0.228	0.307	0.005

Table 8: Linear versus RBF kernel CKA (BRACS).

Pair	Differs in	Linear CKA	RBF CKA
UNI – Phikon-v2	objective	0.834	0.836
UNI – CTransPath	architecture	0.246	0.248
UNI – ResNet50	arch. + dom.	0.450	0.472

C Full CKA Matrices

Table 7 reports the pairwise CKA for all teacher pairs on the three datasets, computed from the frozen features. The ordering is consistent across datasets: the two ViTs (UNI and Phikon-v2) are the most similar pair, and CTransPath, the only Swin encoder, is the most distinct from every other teacher. This consistency supports the claim in Section 4.1 that architecture, rather than objective or domain, governs representational similarity here.

C.1 Robustness to CKA variant

The main text uses linear CKA, the standard choice for cross-architecture comparison. As a robustness check we also recompute pairwise similarity with an RBF (Gaussian) kernel CKA, using the median-distance heuristic for the bandwidth, on a 1,500-sample subsample of the frozen features (Table 8). The RBF values track the linear ones closely and preserve the ordering: the two ViTs (UNI and Phikon-v2) remain the most similar pair and the ViT-to-Swin pair (UNI and CTransPath) the least similar. The conclusion that architecture governs representational similarity therefore holds under both linear and kernel CKA.

D Per-Class Fusion Weights

Table 9 reports the learned per-class teacher weights for the fused combination (Phikon-v2, UNI, ResNet50) on BRACS. Each row is a softmax over the three teachers and sums to one, so the entries can be read as the trust placed in each teacher for that class. The weights vary by class, and the pattern is interpretable: for the well-separated IC class, weight concentrates on the strongest single teacher (UNI, 0.472), whereas the harder atypical classes such as ADH and UDH distribute weight more evenly toward Phikon-v2 and ResNet50, drawing on more than one teacher to resolve the ambiguity.

Table 9: Learned per-class teacher weights for the fused combination on BRACS. Each row sums to one.

Class	$w_{\textbf{Phikon}}$	$w_{\textbf{UNI}}$	$w_{\textbf{ResNet50}}$
N	0.496	0.143	0.361
PB	0.454	0.111	0.434
UDH	0.419	0.123	0.457
FEA	0.431	0.170	0.399
ADH	0.374	0.127	0.499
DCIS	0.365	0.204	0.431
IC	0.440	0.472	0.089