# BEYOND OVERCONFIDENCE: REVISITING CALIBRATION IN MODERN VISION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Modern training recipes for large-scale vision models — combining aggressive augmentation, strong regularization, and web-scale pretraining — have driven substantial gains in image classification accuracy across diverse architectures and benchmarks. Their joint effect on calibration, however, remains poorly characterized. We present a large-scale calibration benchmark comprising 1,218 ImageNet classifiers, a controlled ablation across 2,015 ViT configurations, and a transfer evaluation on four medical imaging tasks, spanning training paradigms from classical from-scratch baselines to web-scale pretrained models with aggressive augmentation and regularization. Four findings emerge: (i) in the high-accuracy regime, modern classifiers are systematically underconfident: 98.4% of the checkpoints in the top 5% of accuracy have negative calibration bias; (ii) the training recipe, not the architecture, is the primary driver of the calibration regime, evidenced by a recipe-shaped Pareto front, consistent shifts across two architecture families, and a controlled ViT ablation; (iii) under severe distribution shift, in-distribution underconfidence reverses to overconfidence; while post-hoc calibration mitigates under-confidence in the in-distribution setting, this exacerbates overconfidence under severe distribution shifts; and (iv) the recipe-driven regime holds for medical imaging: modern-recipe models remain underconfident across four imaging modalities, while supervised ResNet baselines are overconfident on the same tasks. Together, these findings suggest that modern training recipes induce a structural calibration regime that can reverse under distribution shift.

## 1 INTRODUCTION

Reliable confidence estimates are a prerequisite for deploying deep neural networks in safety-critical pipelines that abstain, defer, or escalate based on predicted probabilities. Guo et al. (2017) presented the first systematic benchmark of calibration in deep neural networks and reported that modern deep networks are poorly calibrated, with predicted probabilities systematically exceeding empirical accuracies. This finding has since been replicated across architectures and datasets (Hendrycks et al., 2021b; Lakshminarayanan et al., 2017; Rahaman & Thiery, 2021; Cheng & Vasconcelos, 2024; Wang et al., 2021). Post-hoc calibration methods such as temperature scaling (TS) have become the standard remedy, although their effectiveness has been shown to degrade under distribution shift (Ovadia et al., 2019; Hendrycks & Dietterich, 2019).

Minderer et al. (2021) revisited this question with a calibration benchmark on a more recent generation of vision models and reported that modern vision transformers and MLP-Mixers are well-calibrated out of the box and better calibrated than convolutional baselines. Attributing this gap to the inductive biases of non-convolutional architectures, they concluded that architecture is a major determinant of calibration.

Training pipelines for image classification have since changed substantially. Modern augmentation and regularization techniques (e.g., MixUp (Zhang et al., 2018), RandAugment (Cubuk et al., 2020), label smoothing (Szegedy et al., 2016), stochastic depth (Huang et al., 2016)) bring classical ResNets to accuracy competitive with vision transformers (Wightman et al., 2021) and push ViTs to new state-of-the-art accuracy on ImageNet (Steiner et al., 2022). Web-scale pre-training (e.g., CLIP (Radford et al., 2021) and DINO (Oquab et al., 2024; Siméoni et al., 2025)) has produced further accuracy gains. The calibration effects of individual recipe components have been documented



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.

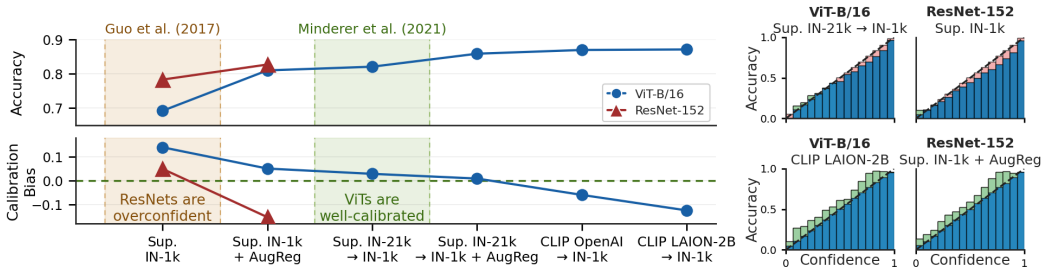


Figure 1: Left: Classification accuracy and calibration bias (mean confidence minus accuracy) on ImageNet (IN)-1k for ViT-B/16 (blue) and ResNet-152 (red) across six training paradigms, from supervised training on IN-1k, supervised training on IN-1k with aggressive augmentation and regularization (AugReg), pretraining on IN-21k followed by IN-1k fine-tuning, to CLIP pre-training. Orange and green bands mark the regimes of Guo et al. (2017) and Minderer et al. (2021), respectively. Right: reliability diagrams for four representative checkpoints; top row: overconfident models studied by Minderer et al. (2021) and Guo et al. (2017); bottom row: corresponding underconfident models trained with modern recipes.

in isolation (e.g., Müller et al., 2019; Thulasidasan et al., 2019). Their joint contribution relative to architectural inductive biases, however, has not been quantified at scale. Existing calibration benchmarks predate the recent generation of vision models: modern architectures such as ConvNeXt and EVA, and self-supervised or web-scale pre-trained models such as DINO and CLIP, are not systematically represented.

These developments — in both training recipes and architectures — motivate a calibration benchmark that jointly covers recent recipe and architecture innovations and characterizes their interaction. Figure 1 previews the magnitude of the recipe effect: on a fixed ViT-B/16, varying the training recipe across six paradigms simultaneously raises classification accuracy and moves calibration bias from the overconfident regime of Guo et al. (2017) through the well-calibrated regime of Minderer et al. (2021) to clearly underconfident predictions with recent web-scale pretraining. Here, we present a comprehensive benchmark that assesses how this effect generalizes across architectures, training recipes, and distribution shifts.

**Contributions.** We systematically benchmark the calibration properties of models trained with modern training recipes. Our main findings are:

1. We present a large-scale empirical calibration study across 1,218 ImageNet classifiers, including state-of-the-art architectures and training recipes. The majority of these models are underconfident, particularly in the high-accuracy regime.

2. We provide three complementary lines of evidence that the training recipe is a dominant driver of the calibration regime: a recipe-shaped Pareto front across 1,218 public checkpoints, consistent recipe-aligned calibration shifts within the ViT and ConvNeXt families considered in isolation, and a controlled ablation across 2,015 ViT configurations in which architecture and optimizer are held fixed.

3. We show that under input distribution shift, modern-recipe models often transition from underconfidence to overconfidence. As a consequence, temperature scaling fitted on an in-distribution validation set can increase calibration error under shift relative to the unscaled model.

4. We find that the recipe-driven regime extends beyond natural images: fine-tuned on four medical imaging tasks, modern-recipe models remain underconfident, while from-scratch ResNet baselines are overconfident.

## 2 Related Work

**Empirical Studies of Model Calibration.** Guo et al. (2017) first established that deep neural networks produce overconfident predictions, identifying depth, width, weight decay, and batch normalization as key factors. This finding has been widely replicated across architectures, datasets, and application domains (Thulasidasan et al., 2019; Hendrycks et al., 2021b; Lakshminarayanan et al., 2017; Rahaman & Thiery, 2021).

Distribution shift exacerbates overconfidence. Ovadia et al. (2019) demonstrated that calibration degrades consistently with increasing dataset shift. Hendrycks & Dietterich (2019) corroborated this finding using synthetic corruptions, and Recht et al. (2019) observed similar effects under temporal distribution drift.

Recent architectural advancements have challenged this overconfidence regime. Minderer et al. (2021) documented improved calibration in Vision Transformers and MLP-Mixers compared to previous generations of models, and emphasized the importance of model architecture in determining calibration quality. Tao et al. (2024) substantiated this architectural dependency with a large-scale benchmark over NAS-searched architectures. The training pipelines underlying their comparison, however, were not held fixed across model families. The accuracy effects of modern training recipes are by now well documented: Wightman et al. (2021) showed that updated training recipes alone bring classical ResNets to accuracy competitive with modern architectures, and Steiner et al. (2022) systematically traced how augmentation, regularization, and pre-training data jointly determine ViT performance. Their calibration effects, however, have not been characterized at the same scale. Existing calibration benchmarks remain largely confined to classical convolutional architectures and the first generation of vision transformers, with conventional supervised training as the dominant regime. Modern architectures such as ConvNeXt and EVA, and self-supervised or web-scale pre-trained models such as DINO and CLIP — which have become standard in many real-world applications — have not been systematically evaluated for calibration at scale.

**Post-hoc Calibration.** A complementary line of work addresses miscalibration after training, by transforming model outputs without modifying the underlying network. Temperature scaling (Guo et al., 2017) rescales the pre-softmax logits by a single scalar $T > 0$ fitted by minimizing negative log-likelihood on a held-out validation set; for $T > 1$ the resulting predictions are softened, for $T < 1$ sharpened. Since rescaling preserves the argmax, temperature scaling leaves classification accuracy unchanged. Despite its simplicity, temperature scaling has been reported to yield competitive results on in-distribution data and has become the standard post-hoc baseline in the calibration literature; extensions include vector and matrix scaling (Guo et al., 2017), Dirichlet calibration (Kull et al., 2019), histogram binning (Zadrozny & Elkan, 2001), and isotonic regression (Zadrozny & Elkan, 2002). Its effectiveness, however, degrades under distribution shift (Ovadia et al., 2019; Tomani et al., 2021). We revisit this question for models trained with modern recipes in Section 4.3.

## 3 Experimental Setup

We evaluate three model sets — a large-scale population of ImageNet-1k classifiers, a controlled ViT ablation, and foundation models fine-tuned on medical imaging tasks — across three evaluation settings: in-distribution ImageNet, distribution-shifted ImageNet, and medical imaging. The remainder of this section details the calibration metrics, datasets, and model sets.

**Calibration metrics.** Throughout, let $\hat{p} = \max_k f_k(x)$ denote the top-label confidence and $\hat{y} = \arg\max_k f_k(x)$ the predicted class for input $x$ with softmax outputs $f_k(x)$, and let $y$ denote the ground-truth label. We report the Expected Calibration Error as our primary metric, and the calibration bias as a complementary signed measure.

*Expected Calibration Error.* The Expected Calibration Error (ECE) partitions the confidence range $[0, 1]$ into $M$ bins $B_1, \ldots, B_M$ and aggregates the absolute gap between bin-wise accuracy and confidence:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\big|, \tag{1}$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the mean accuracy and mean confidence within bin $B_m$, and $N$ is the total number of samples. ECE is the primary metric in prior calibration benchmarks and we adopt it accordingly. Following Guo et al. (2017), we use $M = 15$ equal-width bins throughout the paper.

*Calibration bias.* To resolve the sign of miscalibration, we additionally report the calibration bias

$$b = \mathbb{E}[\hat{p}] - \mathbb{E}[\mathbf{1}\{\hat{y} = y\}], \tag{2}$$

i.e. mean confidence minus accuracy. As a signed quantity, $b$ disambiguates overconfidence ($b > 0$) from underconfidence ($b < 0$), which ECE does not. While $b$ can in principle understate miscalibration when per-bin deviations of opposite sign cancel, empirically $|b|$ tracks ECE closely across the investigated model population (Pearson $r = 0.995$; Appendix B.4), so the sign of $b$ reliably reflects the dominant miscalibration regime.

*Complementary Metrics* Since ECE is known to be sensitive to the binning scheme (Nixon et al., 2019), we additionally compute negative log-likelihood (NLL) and the top-label Brier score, and report ECE under alternative bin counts (Appendices C and D).

**Datasets.** We evaluate on four data settings: in-distribution ImageNet-1k, synthetic ImageNet-1k corruptions, natural ImageNet distribution shifts, and medical image tasks.

*ImageNet-1k (in-distribution).* As our primary in-distribution evaluation, we use the ImageNet-1k validation set (Russakovsky et al., 2015), partitioned 80/20 into a 40,000-image evaluation set used for all reported metrics and a held-out 10,000-image fitting set for post-hoc temperature scaling.

*ImageNet-C (synthetic distribution shift).* ImageNet-C (Hendrycks & Dietterich, 2019) applies 19 synthetic corruptions at 5 severity levels to the ImageNet-1k validation images; we reuse the identical 40k/10k partition across all corruption-severity combinations.

*Natural distribution shifts.* We evaluate on four natural distribution shifts: ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-Sketch (Wang et al., 2019), each on its full released test split. Temperature scaling is fitted on the held-out 10k ImageNet-1k images.

*Medical image tasks.* We evaluate on four tasks, each covering a different imaging modality: dermoscopy (Derma, 7-class) (Tschandl et al., 2018), histopathology (Path, 9-class) (Kather et al., 2019), retinal OCT (OCT, 4-class) (Kermany et al., 2018), and blood-cell microscopy (Blood, 8-class) (Acevedo et al., 2020). We follow the standardized train/val/test splits of Yang et al. (2023): models are trained on the training split, hyperparameters are optimized on the validation split, and all metrics are computed on the test split.

Further details on all datasets are provided in Appendix A.2.

**Models.** We evaluate three model sets: a large-scale population of ImageNet-1k classifiers, covering 87 distinct architectures; a controlled ViT ablation; and a set of large-scale models fine-tuned on medical imaging tasks.

*Large-scale population.* We benchmark 1,218 publicly available ImageNet-1k classifiers from the `timm` library (Wightman, 2019). The population covers convolutional, transformer, hybrid, and MLP-based architectures, and spans training paradigms from supervised ImageNet-1k training to web-scale pre-training, with and without modern augmentation and regularization. We only include public pretrained weights whose final training stage is ImageNet-1k fine-tuning; this curated sample is consequently weighted toward recent, well-performing checkpoints. Details about the model population and the subsets used in subsequent analyses are provided in Appendix A.1.

*Controlled ViT ablation.* To isolate recipe effects under a fixed architecture, we use 2,015 ViT configurations released by Steiner et al. (2022), which jointly vary augmentation strength, regularization, model size, and pre-training corpus while holding architecture and optimizer fixed.

*Medical imaging models.* On each medical task, we fine-tune four ViT-B/16 foundation models with modern training recipes: DINOv2, DINOv3, CLIP-WIT, and CLIP-LAION. We use a single fine-tuning protocol across all models to ensure that observed differences between models do not reflect fine-tuning hyperparameters (Appendix A.1). From-scratch supervised ResNet-18 and ResNet-50 baselines from Yang et al. (2023) serve as a reference.

# 4 RESULTS

## 4.1 HIGH-ACCURACY CLASSIFIERS ARE SYSTEMATICALLY UNDERCONFIDENT

We begin with a large-scale empirical view of calibration in modern ImageNet classifiers. For each checkpoint of the 1,218 ImageNet classifiers, we report accuracy, top-label ECE and calibration bias.

Figure 2 summarizes the result. The dominant calibration regime in this population is underconfidence rather than overconfidence: 75.5 % of all checkpoints have negative calibration bias, in contrast to the overconfidence regime characterized by Guo et al. (2017). The effect concentrates at the high-accuracy end of the distribution: among models in the top 5% of top-1 accuracy, 98.4% are underconfident, with a median calibration bias of -0.053. The most accurate model in the population, EVA-Giant (Fang et al., 2023) fine-tuned from CLIP pre-training (top-1 accuracy 0.894), reaches a calibration bias of $-0.229$, almost five times larger in magnitude than the bias of the ResNet-152 reference of Guo et al. (2017) ($+0.050$), and of opposite sign. While Minderer et al. (2021) reported that recent vision transformers and MLP-Mixers are well-calibrated at high accuracy, our broader population extends this picture: as accuracy increases further with modern training recipes, calibration errors shifts beyond near-zero into clear underconfidence.

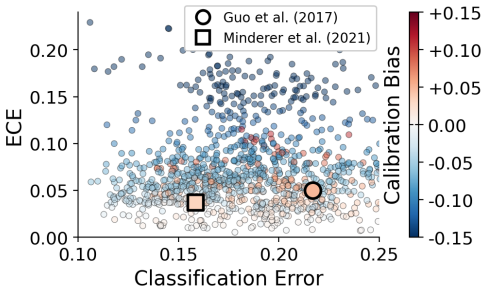


Figure 2: **Most modern ImageNet classifiers are underconfident, particularly at high accuracy.** Expected calibration error (ECE) against classification error for 1,218 ImageNet-1k classifiers (each point one model). Fill colour encodes the signed calibration bias (red: overconfident; blue: underconfident); the reference checkpoints from Guo et al. (2017) (ResNet-152, circle) and Minderer et al. (2021) (ViT-B/16, square) are highlighted. An interactive version of this scatter, with per-checkpoint metadata and filtering by architecture and recipe, is available online.

## 4.2 TRAINING RECIPE AS A DOMINANT DRIVER OF CALIBRATION ACROSS ARCHITECTURES

We next ask whether the calibration regime in this population is better explained by the architecture or by the training recipe. Three complementary views converge on the same picture: calibration aligns closely with training-recipe modernization and varies substantially less across architectures at matched recipes. The evidence comprises (i) the structure of the Pareto front across our large-scale population, (ii) calibration trajectories within two prominent architecture families (ViT and ConvNeXt) considered in isolation, and (iii) a controlled ablation on 2,015 ViT configurations in which architecture and optimizer are held fixed.

The Pareto front of classification error and ECE (Figure 3a) – the boundary of models that are not dominated on either axis – is populated by a wide range of architectures (ViT, ConvNeXt, EVA, RegNetY, and MLP-Mixer) but a narrow band of training recipes: web-scale pre-training dominates the high-accuracy region of the front, while supervised pre-training on ImageNet-21k with aggressive augmentation and regularization populates the lower-accuracy region. Notably, neither from-scratch training nor supervised pre-training without modern augmentation and regularization yields a checkpoint on the front; this includes the references from Guo et al. (2017) and Minderer et al. (2021).

To examine the recipe effect within a fixed architecture, we isolate two representative architecture families spanning transformer and modern convolutional designs that each contribute models to the Pareto front: ViT and ConvNeXt (Figure 3). In both families, classification error decreases monotonically with recipe modernization, consistent with the recipe-driven accuracy gains documented by Wightman et al. (2021) and Steiner et al. (2022). Calibration bias shifts in the same direction: from clear overconfidence under from-scratch training, through near-zero values under supervised IN-21k pre-training, to clear underconfidence under web-scale pre-training (per-scale reliability diagrams in

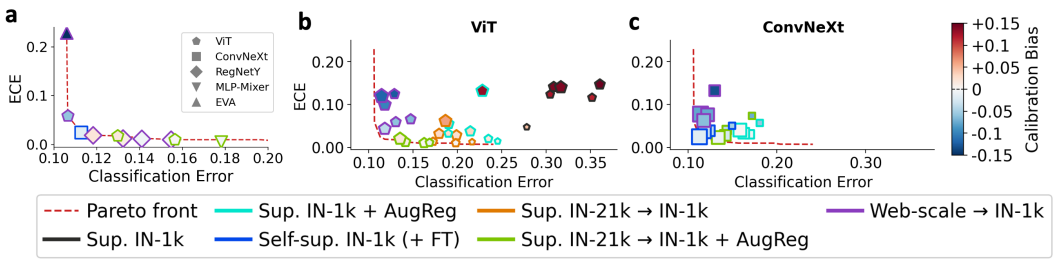


Figure 3: **Calibration shifts along the training-recipe spectrum across architecture families.** Expected calibration error (ECE) against classification error on ImageNet-1k validation. **a:** Pareto-optimal checkpoints from the large-scale study (Figure 2); marker shape encodes the architecture family. **b, c:** ViT and ConvNeXt checkpoints across the training-recipe spectrum (model details in Appendix A.1). In all panels, outline colour encodes the training recipe, fill colour encodes signed calibration bias (blue: underconfident; red: overconfident). The dashed red line marks the Pareto front from Figure 2. Modern recipes (IN-21k pre-training with AugReg and web-scale pre-training) populate the Pareto front and shift calibration into the underconfident regime in both families.

Appendix B.1). AugReg alone brings ViT pretrained on ImageNet-1k into the calibration regime of supervised ImageNet-21k pre-training, indicating that aggressive augmentation and regularization can replicate parts of the calibration shift that pre-training induces. Web-scale pre-training reaches the lowest classification error in both families and produces clearly underconfident predictions in both.

The first two analyses confound recipe components: augmentation/regularization, pre-training corpus, and pre-training scale typically co-vary across publicly available checkpoints rather than being independently controlled. To disentangle their individual contributions, we turn to 2,015 ViT configurations, in which recipe components are varied systematically while architecture and optimizer are held fixed Steiner et al. (2022). Both pre-training corpus and augmentation/regularization strength shift calibration bias monotonically from overconfidence toward underconfidence (Figure 4). The pre-training corpus is the dominant driver: switching from ImageNet-1k to ImageNet-21k shifts mean bias by 0.116 without AugReg and by 0.065 at the strongest AugReg setting. AugReg has a smaller and corpus-dependent effect: bias shifts by 0.07 under ImageNet-1k pre-training but only by 0.019 under ImageNet-21k. Bin-wise averaged reliability diagrams across the same ablation grid (Appendix B.2) show the same monotonic shift on the diagram level. In this ablation, mean calibration bias does not fall below $-0.011$. The substantially stronger underconfidence in Figures 2 and 3 is therefore driven by recipes outside this ablation, in particular web-scale pre-training.

### 4.3 Calibration Under Distribution Shift Strongly Depends on the Training Recipe

Section 4.2 established that the training recipe, not the architecture, drives in-distribution calibration. We now ask whether the same conclusion holds under distribution shift, both synthetic corruptions and natural distribution shifts, and how post-hoc temperature scaling (TS) interacts with the recipe-driven regime.

Figure 5 reports ECE against ImageNet-C error for ViT models across five training recipes and corruption severities 0–5, with and without temperature scaling. We first focus on unscaled predictions. The ECE trajectory under increasing severity differs markedly across recipes. Classical supervised ImageNet-1k training shows the steep growth in ECE with severity documented by Ovadia et al. (2019) and Hendrycks & Dietterich (2019); modern recipes progressively flatten this trajectory. Adding augmentation and regularization (AugReg) to ImageNet-1k training lowers both the in-distribution ECE and its growth rate with severity, matching the calibration profile of supervised ImageNet-21k pre-training. Combining ImageNet-21k pre-training with AugReg flattens the trajectory further. Web-scale pre-training departs qualitatively from this pattern: ECE decreases with severity, as the in-distribution underconfidence is offset by shift-induced confidence inflation. The flattened ECE trajectory previously associated with architectural property of vision transform-




Figure 4: **Augmentation/regularization and pre-training shift calibration in the same direction on ViT architectures.** Mean signed calibration bias on ImageNet-1k for the 2,015 ViT configurations of Steiner et al. (2022), averaged over model sizes, learning rate, and weight decay (Appendix A.1 for selection and aggregation details). Rows: pre-training corpus (IN-1k from scratch vs. IN-21k pre-training followed by IN-1k fine-tuning). Columns: augmentation and regularization strength, from *none* to *strong-2*. Both axes shift calibration bias monotonically in the same direction.

Table 1: **Calibration under natural distribution shift is shaped by the training recipe across architectures.** Top-1 accuracy and ECE on ImageNet-1k and four natural distribution shifts. Values in parentheses report ECE after temperature scaling with temperature $T$ fit on a held-out ImageNet-1k validation set.

| | | IN-1k | | IN-V2 | | IN-A | | IN-R | | IN-Sketch | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Recipe | $T$ | Acc | ECE (TS) | Acc | ECE (TS) | Acc | ECE (TS) | Acc | ECE (TS) | Acc | ECE (TS) |
| **ViT-B/16** | | | | | | | | | | | |
| IN-1k | 1.20 | .704 | .130 (.082) | .545 | .200 (.129) | .064 | .355 (.265) | .199 | .168 (.091) | .100 | .254 (.146) |
| IN-1k + AugReg | 1.05 | .815 | .052 (.042) | .692 | .101 (.086) | .272 | .210 (.188) | .404 | .069 (.050) | .298 | .207 (.179) |
| IN-21k | 1.03 | .824 | .032 (.027) | .709 | .068 (.059) | .370 | .173 (.158) | .444 | .074 (.060) | .314 | .133 (.116) |
| IN-21k + AugReg | 1.00 | .865 | .012 (.012) | .765 | .041 (.041) | .629 | .053 (.052) | .602 | .023 (.023) | .459 | .094 (.093) |
| Web-scale (L2B) | 0.88 | .876 | .127 (.030) | .779 | .087 (.032) | .576 | .082 (.075) | .659 | .080 (.033) | .535 | .025 (.097) |
| **ResNet-152** | | | | | | | | | | | |
| IN-1k | 1.14 | .784 | .051 (.026) | .669 | .096 (.061) | .060 | .509 (.453) | .413 | .180 (.130) | .286 | .246 (.186) |
| IN-1k + AugReg | 0.88 | .832 | .155 (.043) | .718 | .096 (.056) | .234 | .174 (.241) | .468 | .034 (.049) | .361 | .081 (.190) |

ers thus appears within a single architecture once the recipe is varied, suggesting that it reflects the training recipe rather than the architecture itself.

Table 1 corroborates this on four natural distribution shifts. The recipe-not-architecture pattern extends to the ResNet-152 family: replacing the classical supervised recipe with AugReg reduces ECE from 0.509 to 0.174 on ImageNet-A and from 0.246 to 0.081 on ImageNet-Sketch, while raising accuracy from 0.060 to 0.234 and from 0.286 to 0.361, respectively. A modern training recipe applied to a classical architecture thus reaches lower ECE under shift than recent architectures trained with the classical recipe.

The bottom row of Figure 5 shows the same models after temperature scaling fitted on a held-out ImageNet-1k validation set. TS effectively recalibrates classical supervised ImageNet-1k training, reducing ECE across all severities. Under ImageNet-1k training with AugReg and supervised ImageNet-21k pre-training, the TS benefit significantly attenuates at higher severity, and for smaller ViT configurations TS yields higher ECE than the unscaled model. Under web-scale pre-training, this divergence is most pronounced: TS reduces ECE in-distribution but yields higher ECE than the unscaled model once severity increases. This pattern follows directly from the recipe-driven calibration regime. Web-scale models are underconfident in-distribution, so TS fits $T < 1$ to sharpen predictions and reduce in-distribution ECE. Under shift, however, the unscaled predictions move toward the overconfident regime, and the same $T < 1$ that recalibrates the in-distribution underconfidence no longer matches the shifted miscalibration regime. The same pattern – TS yielding higher ECE than the unscaled model at higher severity – can also be seen on the natural distribution shifts (Table 1).

Together, these observations suggest that in-distribution underconfidence is not only an ID phenomenon but a structural property of modern training recipes that determines how models behave under shift.

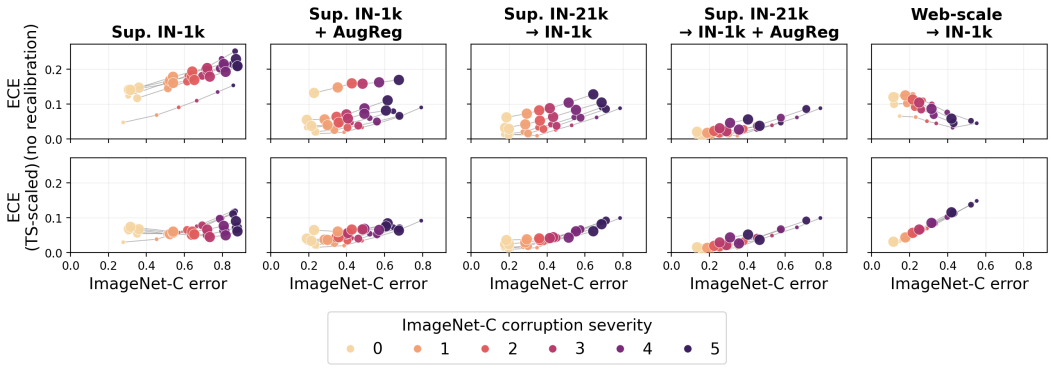


Figure 5: **Calibration under distribution shift depends on the training recipe.** Expected calibration error (ECE) against ImageNet-C error for ViT models with five training recipes, evaluated under increasing corruption severity 0-5 (colour). Marker size encodes model size. The first four columns show, for each recipe, the ViT configuration with the highest validation accuracy from the ViT ablation; the last column shows web-scale CLIP models pre-trained on LAION-2B (model details for both in Appendix A.1). **Top row:** unscaled predictions. **Bottom row:** predictions after temperature scaling (TS) fitted on a held-out ImageNet-1k validation set. The unscaled trajectory of ECE under increasing severity differs markedly across recipes. TS effectively recalibrates classical IN-1k training across all severities, while its benefit attenuates with severity under modern recipes, most strongly under web-scale pre-training.

### 4.4 Calibration Behaviour Transfers to Medical Imaging Tasks

Sections 4.1 and 4.2 characterized calibration on ImageNet, where modern recipes produce systematic underconfidence. We now ask whether this pattern is specific to natural images or carries over to a different visual domain. Figure 6 reports ECE against classification error on four medical imaging datasets (Path, Derma, OCT, Blood), comparing fine-tuned modern-recipe foundation models (DINOv2, DINOv3, CLIP-WIT, CLIP-LAION; all ViT-B/16) with from-scratch supervised ResNet-18 and ResNet-50 baselines.

Across all four datasets, the foundation models reach substantially lower classification error than the ResNet baselines and remain underconfident, mirroring their behavior on ImageNet. The four foundation models cluster tightly in both error and calibration bias. The ResNet baselines display the opposite regime: they are overconfident on all four medical datasets, reproducing the classical calibration profile of Guo et al. (2017) in this new visual domain.

The calibration regime characterized on ImageNet thus carries over to medical imaging: foundation models trained with modern recipes remain underconfident in a substantially different visual domain, while supervised ResNet baselines remain overconfident. This suggests that the recipe-driven calibration patterns of Section 4.2 reflect a structural property of how the model was trained, rather than of the dataset on which it was originally evaluated.

## 5 Implications and Limitations

Our benchmark characterizes calibration in modern vision models across architectures, training recipes, and distribution shifts. Two implications follow for calibration research:

**Modern training recipes should be the default in calibration research.** Post-hoc calibration methods have been developed and evaluated predominantly on baselines trained from scratch with light augmentation and regularization. These models occupy an overconfident regime that lies far from the Pareto front of accuracy and ECE and constitutes a minority among publicly available checkpoints (Section 4.1). Models obtained via large-scale pre-training and modern augmentation and regularization attain substantially higher accuracy, exhibit qualitatively different calibration behavior under distribution shift (Section 4.3), and are the de facto starting point for downstream

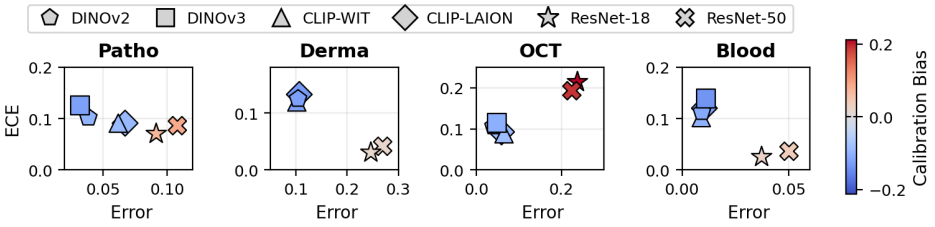


Figure 6: **Calibration on medical imaging tasks under domain shift.** Expected calibration error (ECE) against classification error for four medical imaging datasets. Marker shape encodes the pre-training model: modern-recipe models (DINOv2, DINOv3, CLIP-WIT, CLIP-LAION), all using a ViT-B/16 backbone and fine-tuned on the target task, and from-scratch supervised ResNet-18 and ResNet-50 baselines from Yang et al. (2023). Marker fill colour encodes the calibration bias (blue: underconfident; red: overconfident). Across all four domains, the fine-tuned foundation models achieve substantially lower classification error than the ResNet baselines and remain underconfident, mirroring the in-distribution behaviour observed on ImageNet, while the ResNet baselines are clearly overconfident.

deployment. To make evaluation on this regime tractable for follow-up work, we release cached logits for all 1,218 checkpoints across the in-distribution and shifted evaluation sets (Appendix F).

**Post-hoc calibration under distribution shift warrants re-consideration.** Temperature scaling and related post-hoc methods rely on the assumption that the miscalibration regime at training time is representative of the regime at application time. Our results suggest that this assumption may be challenged under modern training recipes. Recalibration methods that target the modern-recipe regime may therefore benefit from accounting for distribution shift explicitly, for example by fitting the temperature on shift-aware validation data or by adapting the temperature at inference time.

**Limitations.** Our benchmark covers single-label image classification on ImageNet and four medical imaging tasks and does not directly extend to other tasks such as detection or segmentation, multi-label settings, or non-vision modalities. The medical imaging evaluation contrasts modern-recipe foundation models (all ViT-B/16) with from-scratch supervised ResNet baselines, so architecture and recipe co-vary in this particular comparison; the recipe-not-architecture conclusion rests on the controlled comparisons in Section 4.2. Finally, our controlled ViT ablation decomposes the recipe-driven regime along two axes — pre-training corpus and overall augmentation/regularization strength — but does not further isolate the individual contributions of specific recipe components such as label smoothing, mixup, or cutmix. Identifying which components individually drive the underconfidence we observe is a natural direction for future work.

## 6 Conclusions

We presented a large-scale calibration benchmark of modern vision models, comprising 1,218 ImageNet classifiers, a controlled ViT ablation, and a transfer evaluation on four medical imaging tasks. Across these analyses, the dominant calibration regime of modern high-accuracy classifiers is underconfidence rather than overconfidence, and the calibration differences attributable to training recipe in our benchmark are consistently larger than those attributable to architecture at matched recipes. Under distribution shift, the in-distribution underconfidence reverses to overconfidence, potentially limiting the effect of post-hoc calibration. Under domain shift to medical imaging, modern-recipe foundation models remain underconfident while from-scratch ResNet baselines reproduce the classical overconfidence profile. These observations suggest that calibration in modern vision models is best understood as a property set by the training pipeline, persisting across architectures and visual domains. We hope this benchmark supports calibration research focused on the modern training recipes used in current deployments, and on recalibration methods that remain robust under distribution shift.

## REFERENCES


Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020. doi: 10.1016/j.dib.2020.105474.

Jiacheng Cheng and Nuno Vasconcelos. Towards calibrated multi-label deep neural networks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision (ECCV)*, 2016.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Frank Gerrit Zöllner, Niels Halama, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Hermann Brenner. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):e1002730, 2019. doi: 10.1371/journal.pmed.1002730.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A. N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. doi: 10.1016/j.cell.2018.02.010.

Meelis Kull, Miquel Perelló-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 (NeurIPS)*, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, pp. 13991–14002, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Rahul Rahaman and Alexandre Thiery. Uncertainty Quantification and Deep Ensembles. *Advances in Neural Information Processing Systems*, 34:20063–20075, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-3.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research (TMLR)*, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Linwei Tao, Younan Zhu, Haolan Guo, Minjing Dong, and Chang Xu. A benchmark study on calibration. In *The Twelfth International Conference on Learning Representations*, 2024.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10132, 2021.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5 (180161), 2018. doi: 10.1038/sdata.2018.161.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Neural Information Processing Systems*, January 2021.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ross Wightman. PyTorch Image Models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm. In *NeurIPS Workshop on ImageNet: Past, Present, and Future*, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pp. 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 694–699, 2002.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

# APPENDIX

The Appendix is structured as follows:

**A Models and Datasets** . . . . . 14
*Details and references for the models and datasets used in this work.*

A.1 Models . . . . . 14
A.2 Datasets . . . . . 19

**B Supplementary Analyses** . . . . . 21
*Additional results supporting the main claims.*

B.1 Per-scale reliability diagrams for ConvNeXt across pre-training regimes . . . . . 21
B.2 Averaged reliability diagrams from the controlled ViT ablation . . . . . 21
B.3 BEiT and EVA: in-distribution underconfidence and behavior under input shift . . . . . 22
B.4 Empirical relationship between calibration bias and ECE . . . . . 23

**C ECE under Alternative Binning Schemes** . . . . . 24
*Sensitivity of ECE to bin count and bin-construction strategy.*

**D Beyond ECE: NLL and Brier Score** . . . . . 25
*Complementary calibration metrics.*

**E AI Usage Disclosure** . . . . . 27
*Statement for AI Usage.*

**F Reproducibility and Benchmark Artifacts** . . . . . 28
*Released code, cached predictions, checkpoints, and standardized splits.*

**G Compute Resources** . . . . . 28
*GPU-hours and wall-clock times for all reported experiments.*

# A Models and Datasets

## A.1 Models

### A.1.1 Large-scale ImageNet Benchmark

This section details the 1,218 timm (Wightman, 2019) models that constitute our large-scale ImageNet benchmark. All models share two properties: (i) public pretrained weights with an ImageNet-1k fine-tuning step as their final training stage, and (ii) successful evaluation on ImageNet validation under our standard preprocessing pipeline. Of the 1,278 timm checkpoints meeting condition (i) under timm version 1.0.26, 1,218 passed our evaluation pipeline (95.3%); the remaining 60 failed due to memory constraints or incompatible preprocessing requirements Note that where multiple checkpoints exist for the same architecture at different input resolutions, timm exposes them under distinct model identifiers (e.g., `beit_large_patch16_224` vs. `beit_large_patch16_512`); each such identifier is counted as a separate model.

**Architecture coverage.** Table 3 lists the 87 architecture families represented in the benchmark. The benchmark spans classical convolutional networks (ResNet, VGG, DenseNet), modern convolutional networks (ConvNeXt, RegNet, EfficientNet), hybrid architectures (CoAtNet, MaxViT, MetaFormer), pure transformers (ViT, DeiT, BEiT, EVA, Swin), MLP-based architectures (MLP-Mixer, ResMLP), and recent mobile-efficient designs (MobileNetV4, EfficientViT, RepViT, FastViT).

**Resolutions.** Evaluation input resolutions range from 106×106 to 800×800 pixels, with 756 of 1,218 models evaluated at the canonical $224 \times 224$ resolution. Each model is evaluated at the resolution shipped with its public checkpoint.

### A.1.2 Pareto-front checkpoints

Figure 2 highlights the Pareto-front of classification error and ECE within the 1,218-model population, together with the reference checkpoints of Guo et al. (2017) and Minderer et al. (2021). Table 2 lists these checkpoints with their full timm identifiers, parameter counts, input resolutions, and training recipes for reproducibility. The Pareto front is dominated by web-scale and AugReg-based recipes; classical from-scratch training and supervised pre-training without AugReg do not yield checkpoints on the front.

Table 2: **Pareto-front checkpoints from Figure 2 and reference models.** Models highlighted in the Pareto-front analysis, sorted by top-1 accuracy on ImageNet-1k. Pareto-front checkpoints span web-scale, AugReg-based, and self-supervised recipes; the two reference models follow the recipes of Guo et al. (2017) and Minderer et al. (2021), respectively.

| timm Identifier | Family | Params (M) | Input | Recipe |
|---|---|---|---|---|
| eva_giant_patch14_336.clip_ft_in1k | EVA | 1013 | 336 | Web-scale pretrained → IN-1k |
| vit_so400m_patch14_siglip_378.webli_ft_in1k | ViT | 429 | 378 | Web-scale pretrained → IN-1k |
| convnextv2_huge.fcmae_ft_in22k_in1k_384 | ConvNeXt | 660 | 384 | Self-sup. IN-1k (+ FT) |
| regnety_1280.swag_ft_in1k | RegNetY | 645 | 384 | Web-scale pretrained → IN-1k |
| vit_large_patch16_384.augreg_in21k_ft_in1k | ViT | 305 | 384 | Sup. IN-21k → IN-1k + AugReg |
| regnety_320.swag_ft_in1k | RegNetY | 145 | 384 | Web-scale pretrained → IN-1k |
| regnety_1280.swag_lc_in1k | RegNetY | 645 | 224 | Web-scale pretrained → IN-1k |
| regnety_320.swag_lc_in1k | RegNetY | 145 | 224 | Web-scale pretrained → IN-1k |
| vit_base_patch16_224.augreg_in21k_ft_in1k | ViT | 87 | 224 | Sup. IN-21k → IN-1k + AugReg |
| vit_base_patch16_384.orig_in21k_ft_in1k | ViT | 87 | 384 | Reference (Minderer et al.) |
| mixer_b16_224.miil_in21k_ft_in1k | MLP-Mixer | 60 | 224 | Sup. IN-21k → IN-1k + AugReg |
| resnet152.tv_in1k | ResNet | 60 | 224 | Reference (Guo et al.) |

### A.1.3 Architecture-family case studies

Section 4.2 compares two representative architecture families: ConvNeXt and ViT (Figure 3). The ConvNeXt panel uses checkpoints from the timm population throughout. For ViT, the from-scratch

Table 3: Architecture families covered in the large-scale benchmark. Families with fewer than six models are aggregated into the last row.

| Family | # Models | Parameter range | Input resolution |
|---|---|---|---|
| ResNet/SE/ResNeXt (variants) | 192 | 5M–469M | 160–448 |
| ViT | 104 | 6M–632M | 224–448 |
| EfficientNet/V2 | 81 | 0.4M–480M | 106–600 |
| RegNet (X/Y/Z) | 67 | 3M–1.3B | 224–384 |
| ConvNeXt/V2 | 57 | 2M–846M | 224–384 |
| XCiT | 42 | 3M–189M | 224–384 |
| MetaFormer (Conv-/CA-Former) | 32 | 26M–100M | 224–384 |
| MaxViT/MaxxViT | 31 | 8M–476M | 224–512 |
| Swin/V2 | 26 | 28M–197M | 224–384 |
| EfficientViT | 25 | 2M–246M | 224–384 |
| HGNet/V2 | 19 | 6M–75M | 224 |
| MobileNetV4 | 18 | 2M–38M | 224–384 |
| DeiT III | 16 | 22M–632M | 224–384 |
| MobileViT/V2 | 16 | 1M–18M | 256–384 |
| FastViT | 14 | 4M–44M | 256 |
| HRNet | 14 | 13M–128M | 224 |
| MobileNetV3 | 14 | 2M–15M | 224–256 |
| DarkNet/CS3 | 13 | 3M–51M | 256 |
| GhostNet/V2 | 13 | 2M–12M | 224 |
| DLA | 12 | 1M–53M | 224 |
| FlexiViT/NaFlexViT | 12 | 22M–304M | 240–384 |
| RepViT | 12 | 5M–24M | 224 |
| CoAtNet | 11 | 15M–74M | 224–384 |
| CrossViT | 11 | 7M–105M | 240–408 |
| MambaOut | 11 | 7M–102M | 224–384 |
| NFNet | 11 | 10M–316M | 192–384 |
| Next-ViT | 11 | 32M–58M | 224–384 |
| RepVGG | 11 | 9M–133M | 224–320 |
| BEiT/V2 | 10 | 87M–326M | 224–384 |
| LeViT | 10 | 8M–39M | 224 |
| PoolFormer | 10 | 12M–73M | 224 |
| CaiT | 9 | 12M–271M | 224–384 |
| Res2Net | 9 | 25M–48M | 224 |
| ResMLP | 9 | 15M–129M | 224 |
| ViT (relpos) | 9 | 22M–119M | 224–256 |
| CoaT | 8 | 5M–45M | 224–384 |
| DeiT | 8 | 6M–88M | 224–384 |
| EVA/EVA-02 | 8 | 6M–1.0B | 196–336 |
| Hiera | 8 | 28M–673M | 224–256 |
| PiT | 8 | 5M–75M | 224 |
| DPN | 7 | 13M–87M | 224 |
| EfficientFormer/V2 | 7 | 4M–82M | 224 |
| PVT/V2 | 7 | 4M–82M | 224 |
| ReXNet | 7 | 5M–35M | 224 |
| ResNeSt | 7 | 11M–70M | 224–320 |
| TinyViT | 7 | 5M–21M | 224–384 |
| VGG | 7 | 133M–144M | 224 |
| VOLO | 7 | 27M–295M | 224–384 |
| Xception | 7 | 23M–42M | 299 |
| DenseNet | 6 | 8M–29M | 224 |
| EdgeNeXt | 6 | 1M–19M | 256 |
| FasterNet | 6 | 4M–93M | 224 |
| FocalNet | 6 | 28M–89M | 224 |
| HardCoreNAS | 6 | 5M–8M | 224 |
| MixNet | 6 | 4M–7M | 224 |
| Twins | 6 | 24M–99M | 224 |
| Other (small families, $n < 6$) | 111 | 0.3M–305M | 160–384 |
| **Total** | **1218** | | |

and supervised pre-training recipes are drawn from the controlled ablation of Steiner et al. (2022) (see Appendix A.1.4); web-scale ViT checkpoints are drawn from the timm population. Tables 4 and 5 list these checkpoints with their full identifiers.

Table 4: **ConvNeXt family checkpoints used in Figure 3.** Models grouped by training recipe; within each recipe sorted by parameter count.

| timm Identifier | Family | Params (M) | Input | Recipe |
|---|---|---|---|---|
| *Scratch + AugReg* | | | | |
| convnext_tiny.fb_in1k | ConvNeXt | 29 | 224 | Sup. IN-1k + AugReg |
| convnext_small.fb_in1k | ConvNeXt | 50 | 224 | Sup. IN-1k + AugReg |
| convnext_base.fb_in1k | ConvNeXt | 89 | 224 | Sup. IN-1k + AugReg |
| convnext_large.fb_in1k | ConvNeXt | 198 | 224 | Sup. IN-1k + AugReg |
| *Sup. pretrained + AugReg* | | | | |
| convnext_tiny.fb_in22k_ft_in1k | ConvNeXt | 29 | 224 | Sup. IN-22k → IN-1k + AugReg |
| convnext_small.fb_in22k_ft_in1k_384 | ConvNeXt | 50 | 384 | Sup. IN-22k → IN-1k + AugReg |
| convnext_base.fb_in22k_ft_in1k | ConvNeXt | 89 | 224 | Sup. IN-22k → IN-1k + AugReg |
| convnext_large.fb_in22k_ft_in1k | ConvNeXt | 198 | 224 | Sup. IN-22k → IN-1k + AugReg |
| *Self-sup. IN-1k (+ FT)* | | | | |
| convnextv2_tiny.fcmae_ft_in22k_in1k_384 | ConvNeXt | 29 | 384 | Self-sup. IN-1k (+ FT) |
| convnextv2_base.fcmae_ft_in22k_in1k_384 | ConvNeXt | 89 | 384 | Self-sup. IN-1k (+ FT) |
| convnextv2_large.fcmae_ft_in22k_in1k_384 | ConvNeXt | 198 | 384 | Self-sup. IN-1k (+ FT) |
| convnextv2_huge.fcmae_ft_in22k_in1k_384 | ConvNeXt | 660 | 384 | Self-sup. IN-1k (+ FT) |
| *Web-scale pretrained* | | | | |
| convnext_base.clip_laion2b_augreg_ft_in12k_in1k_384 | ConvNeXt | 89 | 384 | Web-scale (LAION-2B) → IN-1k |
| convnext_large_mlp.clip_laion2b_augreg_ft_in1k_384 | ConvNeXt | 200 | 384 | Web-scale (LAION-2B) → IN-1k |
| convnext_large_mlp.clip_laion2b_soup_ft_in12k_in1k_384 | ConvNeXt | 200 | 384 | Web-scale (LAION-2B) → IN-1k |
| convnext_xxlarge.clip_laion2b_soup_ft_in1k | ConvNeXt | 846 | 256 | Web-scale (LAION-2B) → IN-1k |

Table 5: **ViT family checkpoints (web-scale recipe) used in Figure 3.** Web-scale ViT checkpoints are drawn from the timm population; checkpoints for the other recipes are drawn from the controlled ablation of Steiner et al. (2022) and listed in Appendix A.1.4.

| timm Identifier | Family | Params (M) | Input | Recipe |
|---|---|---|---|---|
| vit_base_patch32_clip_384.laion2b_ft_in12k_in1k | ViT | 88 | 384 | Web-scale (LAION-2B) → IN-1k |
| vit_base_patch16_clip_384.openai_ft_in12k_in1k | ViT | 87 | 384 | Web-scale (OpenAI) → IN-1k |
| vit_base_patch16_clip_384.laion2b_ft_in12k_in1k | ViT | 87 | 384 | Web-scale (LAION-2B) → IN-1k |
| vit_large_patch14_clip_336.openai_ft_in12k_in1k | ViT | 305 | 336 | Web-scale (OpenAI) → IN-1k |
| vit_large_patch14_clip_336.laion2b_ft_in12k_in1k | ViT | 305 | 336 | Web-scale (LAION-2B) → IN-1k |
| vit_huge_patch14_clip_336.laion2b_ft_in12k_in1k | ViT | 632 | 336 | Web-scale (LAION-2B) → IN-1k |

### A.1.4 ViT ablation

To isolate recipe effects under a fixed architecture (Section 4.2, Figure 4), we evaluate 2,015 publicly released Vision Transformer checkpoints from Steiner et al. (2022). We use the subset of their release that targets ImageNet-1k (`adapt_ds = imagenet2012`) and covers the six non-hybrid architecture variants ViT-Ti/16, ViT-S/16, ViT-S/32, ViT-B/16, ViT-B/32, and ViT-L/16. Across these checkpoints, six training-recipe axes are varied jointly while architecture and optimizer are held fixed (Table 6); augmentation configurations are detailed in Table 7. The two ImageNet-21k pre-training variants released by Steiner et al. (2022) (`i21k` and `i21k_30`) are pooled into a single "IN-21k → IN-1k" category for our analysis.

**Augmentation configurations.** We evaluate seven combinations of RandAugment (Cubuk et al., 2020) and Mixup (Zhang et al., 2018). Each configuration is described by a triple $(l, m, \alpha)$, where $l$ specifies the number of RandAugment operations, $m$ controls augmentation magnitude, and $\alpha$ is the Mixup interpolation parameter.

Table 6: **Search space of the ViT ablation.** Hyperparameter combinations are evaluated independently for each of the six ViT variants.

| Hyperparameter | Values |
|---|---|
| Pretraining | {None (IN-1k from scratch), ImageNet-21k} |
| Augmentation (Table 7) | 7 configurations |
| Weight decay | {0.03, 0.1} |
| Stochastic depth & dropout | {(0, 0), (0.1, 0.1)} |
| Adapt learning rate | {0.01, 0.03} |
| Adapt resolution | {224, 384} |

Table 7: **Augmentation configurations used in the ViT ablation.**

| Setup | $l$ | $m$ | $\alpha$ |
|---|---|---|---|
| none | 0 | 0 | 0 |
| light0 | 2 | 0 | 0 |
| light1 | 2 | 10 | 0.2 |
| medium1 | 2 | 15 | 0.2 |
| medium2 | 2 | 15 | 0.5 |
| strong1 | 2 | 20 | 0.5 |
| strong2 | 2 | 20 | 0.8 |

**Configurations used in Section 4.2 and Section 4.3.** For the case studies of Section 4.2 (Figure 3) and the distribution-shift evaluation of Section 4.3 (Figure 5, Table 1), we select within each architecture and each (pretraining, augmentation-status) regime the ablation configuration with the highest top-1 accuracy on the ImageNet-1k validation set, restricted to checkpoints with adapt learning rate $0.01$ and resolution $384$. The resulting checkpoints are listed in Table 8.

Table 8: **Steiner et al. configurations used in Section 4.2 (Figure 3) and Section 4.3 (Figure 5, Table 1).** Each row identifies one ablation configuration by its filename in the public release of Steiner et al. (2022).

| Architecture | Configuration filename |
|---|---|
| *Sup. IN-1k (scratch)* | |
| ViT-Ti/16 | `Ti_16-i1k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/16 | `S_16-i1k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/32 | `S_32-i1k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/16 | `B_16-i1k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/32 | `B_32-i1k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-L/16 | `L_16-i1k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| *Sup. IN-1k + AugReg* | |
| ViT-Ti/16 | `Ti_16-i1k-300ep-lr_0.001-aug_light0-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/16 | `S_16-i1k-300ep-lr_0.001-aug_medium2-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/32 | `S_32-i1k-300ep-lr_0.001-aug_medium2-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/16 | `B_16-i1k-300ep-lr_0.001-aug_strong2-wd_0.1-do_0.1-sd_0.1--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/32 | `B_32-i1k-300ep-lr_0.001-aug_medium2-wd_0.1-do_0.1-sd_0.1--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-L/16 | `L_16-i1k-300ep-lr_0.001-aug_medium2-wd_0.1-do_0.1-sd_0.1--imagenet2012-steps_20k-lr_0.01-res_384` |
| *Sup. IN-21k → IN-1k* | |
| ViT-Ti/16 | `Ti_16-i21k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/16 | `S_16-i21k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/32 | `S_32-i21k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/16 | `B_16-i21k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/32 | `B_32-i21k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-L/16 | `L_16-i21k-300ep-lr_0.001-aug_none-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| *Sup. IN-21k → IN-1k + AugReg* | |
| ViT-Ti/16 | `Ti_16-i21k-300ep-lr_0.001-aug_light0-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/16 | `S_16-i21k-300ep-lr_0.001-aug_light1-wd_0.03-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-S/32 | `S_32-i21k-300ep-lr_0.001-aug_light0-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/16 | `B_16-i21k-300ep-lr_0.001-aug_medium1-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-B/32 | `B_32-i21k-300ep-lr_0.001-aug_light1-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384` |
| ViT-L/16 | `L_16-i21k-300ep-lr_0.001-aug_strong1-wd_0.1-do_0.1-sd_0.1--imagenet2012-steps_20k-lr_0.01-res_384` |

### A.1.5 DISTRIBUTION-SHIFT EVALUATION

This section lists the models used in the distribution-shift evaluation of Section 4.3.

**ViT configurations for Figure 5.** The first four columns of Figure 5 show, for each recipe (Sup. IN-1k, Sup. IN-1k + AugReg, Sup. IN-21k → IN-1k, Sup. IN-21k → IN-1k + AugReg), the ViT configuration with the highest ImageNet-Val accuracy from Steiner et al. (2022); the corresponding identifiers and full configurations are listed in Appendix A.1.4. The fifth column shows web-scale CLIP models pre-trained on LAION-2B and fine-tuned on ImageNet-1k (Table 9).

Table 9: **Web-scale CLIP checkpoints used in Figure 5.** All checkpoints are pre-trained on LAION-2B and fine-tuned on ImageNet-1k.

| timm Identifier | Family | Params (M) | Input | Recipe |
|---|---|---|---|---|
| vit_base_patch32_clip_384.laion2b_ft_in12k_in1k | ViT | 88 | 384 | Web-scale (LAION-2B) → IN-1k |
| vit_base_patch16_clip_384.laion2b_ft_in12k_in1k | ViT | 87 | 384 | Web-scale (LAION-2B) → IN-1k |
| vit_large_patch14_clip_336.laion2b_ft_in12k_in1k | ViT | 305 | 336 | Web-scale (LAION-2B) → IN-1k |
| vit_huge_patch14_clip_336.laion2b_ft_in12k_in1k | ViT | 632 | 336 | Web-scale (LAION-2B) → IN-1k |

**Models for Table 1.** Table 1 reports accuracy and ECE on natural distribution shifts for two architectures and seven recipes. The ResNet-152 entries use timm checkpoints (Table 10); the ViT-B/16 entries are drawn from Steiner et al. (2022) and listed in Appendix A.1.4.

Table 10: **ResNet-152 checkpoints used in Table 1.**

| timm Identifier | Family | Params (M) | Input | Recipe |
|---|---|---|---|---|
| resnet152.tv_in1k | ResNet | 60 | 224 | Sup. IN-1k |
| resnet152.a1h_in1k | ResNet | 60 | 224 | Sup. IN-1k + AugReg |

### A.1.6 MEDICAL IMAGING MODELS

Section 4.4 (Figure 6) compares four modern-recipe foundation models, each fine-tuned on four medical imaging tasks, against from-scratch supervised ResNet-18 and ResNet-50 baselines from Yang et al. (2023).

**Foundation models.** We fine-tune four ViT-B/16 backbones obtained from different pre-training paradigms (Table 11). All backbones are loaded via the timm library (Wightman, 2019).

Table 11: **ViT-B/16 backbones used in the medical imaging evaluation.**

| Pre-training paradigm | Pre-training corpus | timm Identifier |
|---|---|---|
| DINOv2 | LVD-142M | `vit_base_patch14_dinov2.lvd142m` |
| DINOv3 | LVD-1689M | `vit_base_patch14_dinov3.lvd1689m` |
| CLIP-WIT | WIT-400M (OpenAI) | `vit_base_patch16_clip_224.openai` |
| CLIP-LAION | LAION-2B | `vit_base_patch16_clip_224.laion2b` |

**Fine-tuning protocol.** Each backbone is fine-tuned end-to-end on each medical task with a single, near-identical protocol. Hyperparameters constant across all model–task combinations are summarized in Table 12. Training length and vertical-flip augmentation are adapted per task to dataset size and anatomical orientation; the corresponding values are listed in Table 13.

**ResNet baselines.** ResNet-18 and ResNet-50 baselines are taken from (Yang et al., 2023). We use the trained weights provided by the authors without additional fine-tuning and evaluate them under the same protocol as the foundation models.

Table 12: **Fine-tuning hyperparameters held constant across all foundation models and medical tasks.**

| Hyperparameter | Value |
|---|---|
| Optimizer | AdamW |
| Base learning rate | $5 \times 10^{-5}$ |
| Minimum learning rate | $1 \times 10^{-7}$ |
| Warmup learning rate | $1 \times 10^{-6}$ |
| Weight decay | 0.05 |
| Layer-wise LR decay | 0.75 |
| Schedule | Cosine |
| Batch size | 128 |
| Input resolution | $224 \times 224$ |
| Mixed precision | bfloat16 |
| Gradient clipping | 3.0 (norm) |
| Label smoothing | 0.1 |
| Drop path | 0.1 |
| Mixup | 0.8 (batch mode, prob. 1.0) |
| CutMix | 1.0 |
| RandAugment | rand-m9-mstd0.5-inc1 |
| Random erase | 0.25 (pixel) |
| Color jitter | 0.4 |
| Horizontal flip | 0.5 |
| Crop scale / aspect ratio | $[0.08, 1.0]$ / $[0.75, 1.33]$ |
| Crop pct (eval) | 0.95 |
| Model EMA | 0.9998 |
| Random seed | 42 |

Table 13: **Per-dataset fine-tuning hyperparameters.** Training length is shorter for the larger Path and OCT datasets.

| Hyperparameter | Path | Derma | OCT | Blood |
|---|---|---|---|---|
| Epochs | 20 | 50 | 20 | 50 |
| Warmup epochs | 2 | 5 | 2 | 5 |
| Cooldown epochs | 2 | 5 | 2 | 5 |

### A.2 DATASETS

This section details the datasets used in our evaluation: in-distribution ImageNet-1k, synthetic and natural ImageNet distribution shifts, and four medical imaging tasks.

**ImageNet-1k.** ImageNet-1k (Russakovsky et al., 2015) is a 1,000-class object recognition benchmark assembled from web-scraped photographs and is the standard in-distribution evaluation set for image classification on ImageNet. We use the public validation set with 50,000 images and randomly split it into a 40,000-image evaluation set used for all reported metrics and a held-out 10,000-image fitting set used for post-hoc temperature scaling. The split is fixed with seed 42 and reused across all experiments.

**ImageNet-C.** ImageNet-C (Hendrycks & Dietterich, 2019) probes robustness to common corruptions by applying 19 synthetic perturbations (e.g., Gaussian noise, motion blur, JPEG compression, brightness, fog, pixelation) to the ImageNet-1k validation images at 5 severity levels each. The corrupted images preserve the underlying 1,000-class label structure of ImageNet-1k, which makes ImageNet-C a controlled testbed for graceful degradation under input distribution shift. We evaluate on all 19 corruptions and reuse the same 40,000/10,000 split as for in-distribution evaluation across all corruption-severity combinations.

**Natural distribution shifts.** We evaluate on four natural ImageNet distribution shifts using their full released test splits. *ImageNet-V2* (Recht et al., 2019) contains 10,000 images across the same 1,000 classes as ImageNet-1k, collected from the same source pipeline a decade later, and quantifies

sampling shift on otherwise unchanged label semantics. *ImageNet-R* (Hendrycks et al., 2021a) comprises 30,000 artistic renditions (sketches, paintings, sculptures, embroidery, etc.) restricted to a 200-class subset of ImageNet-1k and probes robustness to texture and style shift. *ImageNet-A* (Hendrycks et al., 2021b) contains 7,500 unmodified natural images from the same 200-class subset that are systematically misclassified by a held-out ResNet-50 and thus represent naturally hard examples. *ImageNet-Sketch* (Wang et al., 2019) provides 50,889 black-and-white sketches across all 1,000 ImageNet-1k classes and isolates shift from photographic to non-photographic input. For these benchmarks, post-hoc temperature scaling is fitted on the held-out 10,000-image fitting set of ImageNet-1k.

**Medical imaging tasks.** We use four medical classification tasks from (Yang et al., 2023) at $224 \times 224$ resolution, covering four imaging modalities and a wide range of dataset sizes (Table 14). *Path* (Kather et al., 2019) contains hematoxylin-and-eosin-stained colorectal cancer histology patches across nine tissue classes. *Derma* (Tschandl et al., 2018) comprises dermoscopic images of pigmented skin lesions across seven diagnostic categories. *OCT* (Kermany et al., 2018) consists of retinal optical coherence tomography scans across four diagnostic categories. *Blood* (Acevedo et al., 2020) provides peripheral blood-cell microscopy images across eight cell types. For each task we follow the standardized train/validation/test splits of Yang et al. (2023): foundation models are fine-tuned on the training split (Appendix A.1.6), training-time hyperparameters are selected on the validation split, and all reported metrics are computed on the test split.

Table 14: **Medical datasets used in the transfer evaluation.**

| Dataset | Modality | Classes | Train | Val | Test |
|---|---|---|---|---|---|
| Path | Histopathology | 9 | 89,996 | 10,004 | 7,180 |
| Derma | Dermoscopy | 7 | 7,007 | 1,003 | 2,005 |
| OCTM | Retinal OCT | 4 | 97,477 | 10,832 | 1,000 |
| Blood | Blood-cell microscopy | 8 | 11,959 | 1,712 | 3,421 |

# B SUPPLEMENTARY ANALYSES

This section collects additional observations that support the claims but that did not fit the main paper's page budget. Each subsection reports a self-contained experiment and is referenced from the main text where appropriate.

## B.1 PER-SCALE RELIABILITY DIAGRAMS FOR CONVNEXT ACROSS PRE-TRAINING REGIMES

To complement the aggregated case study in Figure 3 of the main paper, we report per-model reliability diagrams for four ConvNeXt scales (T/S/B/L). The top row shows checkpoints trained from scratch on ImageNet-1k; the bottom row shows checkpoints obtained by ImageNet-21k pre-training followed by ImageNet-1k fine-tuning. For the ImageNet-1k–only checkpoints, the reliability curves stay close to the diagonal, with a small but consistent overconfidence component that grows mildly with model scale. After ImageNet-22k pre-training, the same architectures shift uniformly into the underconfidence region: the empirical accuracy in each confidence bin lies above the diagonal, and the gap is largest in the low- and mid-confidence bins. The effect is monotonic in scale only to a limited extent — ConvNeXt-B and ConvNeXt-L exhibit the most pronounced underconfidence, while ConvNeXt-T and ConvNeXt-S show a smaller but qualitatively identical shift. This is consistent with the recipe-level reading of Figure 1: holding the architecture family fixed and changing only the pre-training stage moves the calibration regime in a controlled direction.

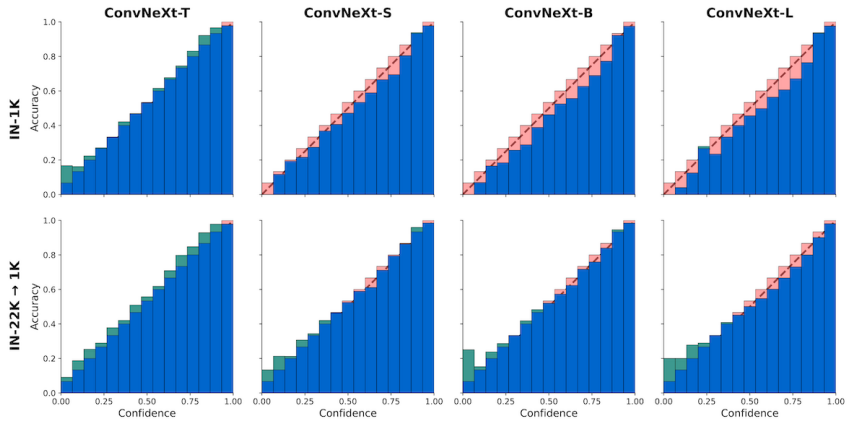


Figure 7: **Reliability diagrams for ConvNeXt-T, ConvNeXt-S, ConvNeXt-B, and ConvNeXt-L (15 equal-width bins).** Top row: supervised training on ImageNet-1k. Bottom row: ImageNet-21k pre-training followed by ImageNet-1k fine-tuning. Red shading indicates bins for which mean confidence exceeds bin accuracy (overconfidence), green shading indicates the opposite (underconfidence). Across all four scales, switching from ImageNet-1k–only training to ImageNet-21k pre-training shifts the diagrams from well-calibrated/overconfident to clearly underconfident, with no change to the architecture or its capacity.

## B.2 AVERAGED RELIABILITY DIAGRAMS FROM THE CONTROLLED VIT ABLATION

We additionally report bin-wise averages over the full 2,015-run ViT ablation grid, grouped by the two recipe factors with the largest effect: (a) presence vs. absence of upstream pre-training, and (b) augmentation strength. For each group, we average bin accuracies across all runs and report bin-wise standard deviations as error bars. Without pre-training, the averaged reliability diagram lies systematically below the diagonal across the full confidence range, indicating that runs in this group are, on average, overconfident. Adding pre-training removes this gap almost entirely: the averaged curve tracks the diagonal closely, with bin-wise variability that is comparable to or smaller than in the no-pre-training group. The augmentation factor exhibits a similar but more graded pattern. With no augmentation, the averaged curve shows a clear overconfidence gap. Light augmentation (Light1) reduces this gap substantially, and strong augmentation (Strong2) further closes it, with the bin-wise mean tracking the diagonal across most of the confidence range. The two factors are not

redundant: their effect adds up across the ablation grid, but each, in isolation, moves the average reliability diagram in the same direction.

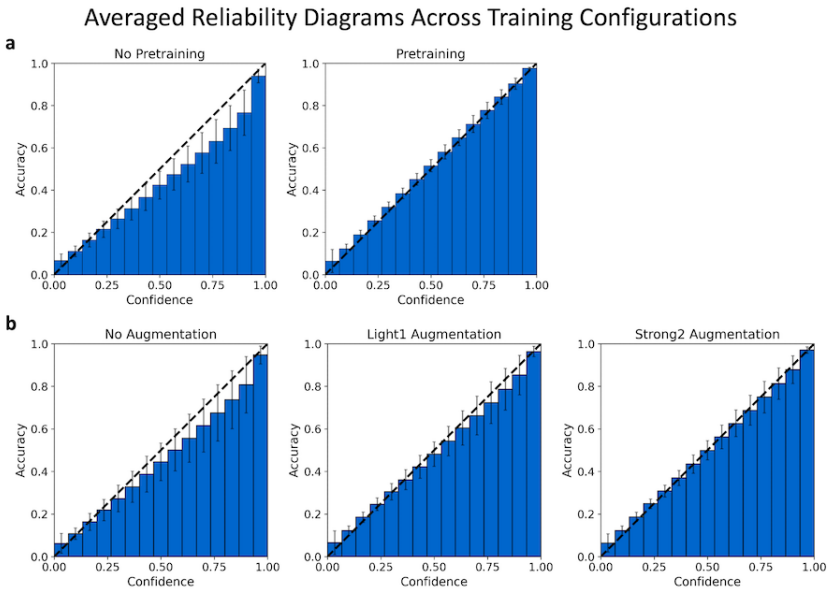


Figure 8: **Averaged reliability diagrams across the controlled ViT ablation (2,015 runs, 15 equal-width bins).** Bars show bin-wise mean accuracy averaged across runs in each group; error bars show bin-wise standard deviation. (a) Runs grouped by absence (left) vs. presence (right) of upstream pre-training. (b) Runs grouped by augmentation strength: no augmentation, Light1, and Strong2. Both factors shift the averaged reliability diagram from a regime of mild overconfidence toward the diagonal; the effects are visible in isolation and accumulate across the grid.

### B.3 BEiT AND EVA: IN-DISTRIBUTION UNDERCONFIDENCE AND BEHAVIOR UNDER INPUT SHIFT

To check that the recipe effect is not specific to thefamilies in the main text, we report the calibration behavior of two further Transformer-based architectures, BEiT and EVA, both released as checkpoints trained with modern recipes (large-scale (pre-)training combined with modern augmentation and regularization). We evaluate both models in two settings: the in-distribution ImageNet-1k validation set, and ImageNet-C across all five corruption severities. In-distribution (left), both checkpoints are significantly underconfident.

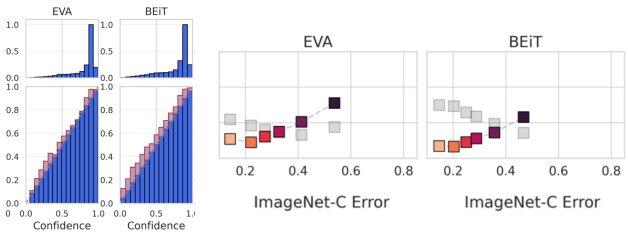


Figure 9: **Calibration of BEiT and EVA. Left: in-distribution evaluation on ImageNet-1k.** Top: sample-fraction histograms over confidence bins. Bottom: reliability diagrams (15 equal-width bins); red shading marks overconfidence, transparent shading marks the gap to the diagonal. Both checkpoints are underconfident, with predictions concentrated in the highest-confidence bins. Right: ECE versus ImageNet-C classification error for severities 0 (clean) through 5. Coloured markers report TS-scaled ECE, with the temperature fitted once on the in-distribution ImageNet-1k validation set and applied unchanged across severities; grey markers report the corresponding unscaled ECE.

### B.4 Empirical relationship between calibration bias and ECE

For an input $x$ with ground-truth label $y$, let $\hat{y} = \arg\max_k f_k(x)$ denote the model's predicted class and $\hat{p} = \max_k f_k(x)$ its associated top-label confidence, where $f_k(x)$ is the softmax probability assigned to class $k$. The calibration bias is then defined as

$$b = \mathbb{E}[\hat{p}] - \mathbb{E}[\mathbf{1}\{\hat{y} = y\}], \tag{3}$$

i.e. mean confidence minus accuracy. As a global mean, $b$ can in principle understate miscalibration when per-bin deviations of opposite sign cancel. We therefore verify empirically that $|b|$ tracks ECE across the 1,218 ImageNet-1k models considered in Section 4.1. Figure 10 plots $|b|$ against ECE for all checkpoints.

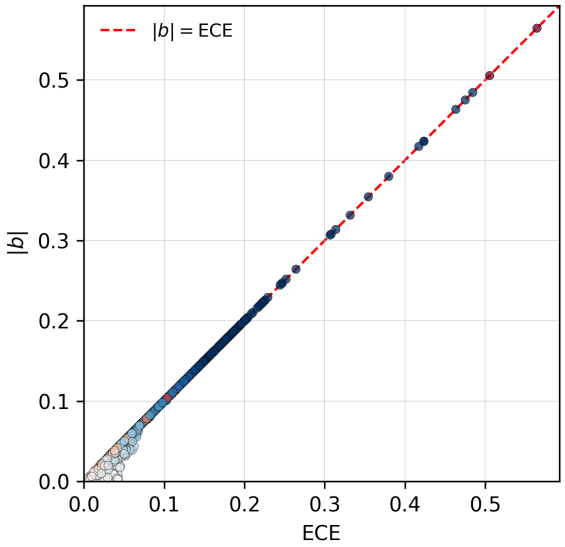


Figure 10: **Calibration bias magnitude $|b|$ tracks ECE across the 1,218-model population.** Each point is one ImageNet-1k checkpoint. The dashed line marks $|b| = \mathrm{ECE}$, the upper bound on $|b|$ implied by the binned ECE definition; all points lie at or below it by construction. Most points lie close to the bound (Pearson $r = 0.995$, mean gap $0.003$), indicating that cancellation across confidence bins is empirically limited.

Two observations characterize the empirical relationship between $b$ and ECE in this population:

**(i) $|b|$ is a tight lower bound on ECE.** By construction $|b| \leq \mathrm{ECE}$, since $|b|$ collapses all confidence bins into a single signed average. Empirically, $|b|$ remains close to ECE across the population, with Pearson $r(|b|, \mathrm{ECE}) = 0.995$ and Spearman $\rho = 0.989$.

**(ii) Per-bin cancellation is empirically negligible.** The gap $\mathrm{ECE} - |b|$ measures the contribution of opposite-sign bin deviations: if all per-bin deviations had the same sign as $b$, the two quantities would coincide. Across the population, this gap is on average $0.003$ and at most $0.043$.

Together, these observations support the interpretation of $b$ as a directional summary of miscalibration alongside ECE.

## C ECE UNDER ALTERNATIVE BINNING SCHEMES

The Expected Calibration Error depends on the binning scheme used to estimate the per-bin accuracy and confidence. This appendix reports two complementary checks: (i) sensitivity to the number of bins under equal-width binning (Figure 11), and (ii) sensitivity to the bin-construction strategy.

**Sensitivity to the number of bins.** Following Guo et al. (2017), we use $M = 15$ equal-width bins as the primary metric throughout the paper. Minderer et al. (2021) use $M = 100$ for higher resolution at the expense of bin-size variability. Across the 1,218-model population, ECE values at $M = 15$ and $M = 100$ are essentially identical: Pearson $r = 0.9997$, Spearman $\rho = 0.9988$, mean absolute difference $0.0017$, and maximum absolute difference $0.0094$. This indicates that the conclusions of Section 4.1 and Section 4.2 are not driven by the choice of $M$. Figure 11 confirms this visually.

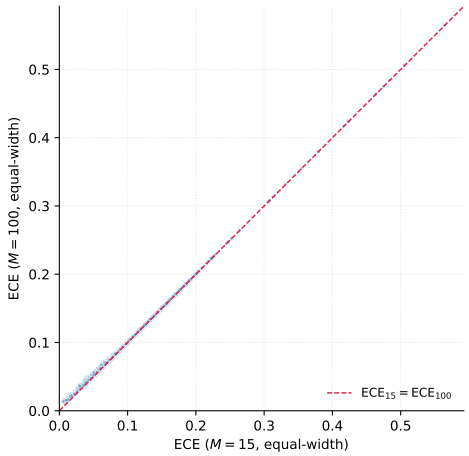


Figure 11: **ECE is robust to the choice of M.** Each point is one of the 1,218 ImageNet-1k checkpoints from our large-scale benchmark; the dashed line marks $\mathrm{ECE}_{15} = \mathrm{ECE}_{100}$. Across the population, the two binning resolutions agree closely (Pearson $r = 0.9997$, Spearman $\rho = 0.9988$, max absolute difference $0.0094$), confirming that the conclusions of Section 4.1 and Section 4.2 do not depend on the bin count.

The Pareto-front analysis of Section 4.2 (Figure 3) is similarly unchanged when ECE is recomputed with $M = 100$ bins (Figure 12). The same checkpoints populate the Pareto front, and the recipe-level structure — web-scale pre-training and AugReg-based recipes dominating, classical from-scratch training absent — is preserved.

**Sensitivity to the binning strategy.** Equal-width binning, the default in our analysis, partitions the confidence range $[0, 1]$ into $M$ bins of equal width. Equal-mass binning, used by Minderer et al. (2021), partitions it into $M$ bins each containing the same number of samples. The two schemes can disagree when confidence is heavily concentrated in narrow ranges — a regime that may apply to high-accuracy underconfident models. We re-compute ECE under equal-mass binning on the architecture-family case-study checkpoints (Tables 2–5).

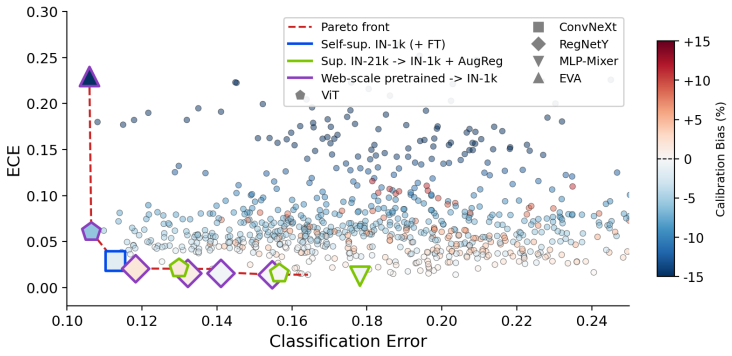


Figure 12: **Pareto front under M=100 equal-width binning.** Reproduction of Figure 3a using the higher-resolution binning scheme of Minderer et al. (2021). The Pareto-front checkpoints (Table 2) and the recipe-level structure of the front are unchanged.

## D BEYOND ECE: NLL AND BRIER SCORE

We complement the ECE-based analysis with two binning-free alternatives: the negative log-likelihood (NLL) and the top-label Brier score. Tables 15 and 16 report all three calibration metrics together with classification accuracy for the architecture-family case studies of Figure 3.

Unlike ECE, both NLL and Brier decompose into a calibration component and a refinement (sharpness) component: a model that improves classification accuracy while keeping its calibration regime constant will mechanically achieve lower NLL and lower Brier score. Consequently, NLL and Brier values across the recipe progression decrease primarily because accuracy increases (NLL drops from $\approx 1.7$ at scratch ViT to $\approx 0.5$ under web-scale pre-training; Brier from $0.16$ to $0.09$), and not because calibration improves. The signed calibration bias — mean confidence minus accuracy — is independent of accuracy and is therefore the cleanest single-number indicator of the regime shift documented in Section 4.2.

Table 15: **ConvNeXt family checkpoints (Figure 3) under multiple calibration metrics on ImageNet-1k.**NLL and Brier values decrease in part because accuracy increases along the recipe progression; only ECE and calibration bias isolate the calibration regime independently of refinement.

| Architecture | Recipe | Top-1 | ECE | NLL | Brier | Bias |
|---|---|---|---|---|---|---|
| *Sup. IN-1k + AugReg* | | | | | | |
| ConvNeXt-T | Sup. IN-1k + AugReg | 0.825 | 0.060 | 0.78 | 0.106 | −0.060 |
| ConvNeXt-S | Sup. IN-1k + AugReg | 0.835 | 0.028 | 0.70 | 0.098 | −0.008 |
| ConvNeXt-B | Sup. IN-1k + AugReg | 0.843 | 0.035 | 0.68 | 0.095 | +0.002 |
| ConvNeXt-L | Sup. IN-1k + AugReg | 0.848 | 0.036 | 0.67 | 0.094 | −0.002 |
| *Sup. IN-22k → IN-1k + AugReg* | | | | | | |
| ConvNeXt-T | Sup. IN-22k → IN-1k + AugReg | 0.832 | 0.072 | 0.72 | 0.103 | −0.072 |
| ConvNeXt-S | Sup. IN-22k → IN-1k + AugReg | 0.861 | 0.044 | 0.56 | 0.088 | −0.042 |
| ConvNeXt-B | Sup. IN-22k → IN-1k + AugReg | 0.859 | 0.026 | 0.56 | 0.086 | −0.023 |
| ConvNeXt-L | Sup. IN-22k → IN-1k + AugReg | 0.865 | 0.022 | 0.53 | 0.083 | −0.004 |
| *Self-sup. IN-1k (+ FT)* | | | | | | |
| ConvNeXtV2-T | Self-sup. IN-1k (+ FT) | 0.853 | 0.049 | 0.61 | 0.091 | −0.048 |
| ConvNeXtV2-B | Self-sup. IN-1k (+ FT) | 0.877 | 0.039 | 0.51 | 0.079 | −0.034 |
| ConvNeXtV2-L | Self-sup. IN-1k (+ FT) | 0.882 | 0.037 | 0.50 | 0.078 | −0.030 |
| ConvNeXtV2-H | Self-sup. IN-1k (+ FT) | 0.885 | 0.022 | 0.46 | 0.075 | −0.014 |
| *Web-scale pretrained (LAION-2B)* | | | | | | |
| ConvNeXt-B | Web-scale (LAION-2B) → IN-1k | 0.877 | 0.136 | 0.62 | 0.102 | −0.136 |
| ConvNeXt-L | Web-scale (LAION-2B) → IN-1k | 0.881 | 0.078 | 0.55 | 0.085 | −0.076 |
| ConvNeXt-L | Web-scale (LAION-2B) → IN-1k | 0.885 | 0.064 | 0.50 | 0.080 | −0.063 |
| ConvNeXt-XXL | Web-scale (LAION-2B) → IN-1k | 0.887 | 0.081 | 0.52 | 0.082 | −0.079 |

Table 16: **ViT family checkpoints (Figure 3) under multiple calibration metrics on ImageNet validation.** NLL and Brier values decrease in part because accuracy increases along the recipe progression; only ECE and calibration bias isolate the calibration regime independently of refinement.

| Architecture | Recipe | Top-1 | ECE | NLL | Brier | Bias |
|---|---|---|---|---|---|---|
| *Sup. IN-1k (scratch)* | | | | | | |
| ViT-Ti/16 | Sup. IN-1k (scratch) | 0.725 | 0.049 | 1.17 | 0.131 | +0.045 |
| ViT-S/16 | Sup. IN-1k (scratch) | 0.704 | 0.120 | 1.51 | 0.155 | +0.120 |
| ViT-S/32 | Sup. IN-1k (scratch) | 0.655 | 0.116 | 1.72 | 0.160 | +0.116 |
| ViT-B/16 | Sup. IN-1k (scratch) | 0.704 | 0.130 | 1.63 | 0.155 | +0.129 |
| ViT-B/32 | Sup. IN-1k (scratch) | 0.644 | 0.144 | 1.92 | 0.173 | +0.144 |
| ViT-L/16 | Sup. IN-1k (scratch) | 0.688 | 0.140 | 1.79 | 0.159 | +0.140 |
| *Sup. IN-1k + AugReg* | | | | | | |
| ViT-Ti/16 | Sup. IN-1k + AugReg | 0.759 | 0.018 | 0.97 | 0.119 | +0.007 |
| ViT-S/16 | Sup. IN-1k + AugReg | 0.814 | 0.031 | 0.73 | 0.104 | +0.026 |
| ViT-S/32 | Sup. IN-1k + AugReg | 0.768 | 0.023 | 0.92 | 0.117 | +0.017 |
| ViT-B/16 | Sup. IN-1k + AugReg | 0.815 | 0.052 | 0.76 | 0.103 | +0.049 |
| ViT-B/32 | Sup. IN-1k + AugReg | 0.794 | 0.035 | 0.83 | 0.111 | +0.026 |
| ViT-L/16 | Sup. IN-1k + AugReg | 0.779 | 0.127 | 1.25 | 0.141 | +0.126 |
| *Sup. IN-21k → IN-1k* | | | | | | |
| ViT-Ti/16 | Sup. IN-21k → IN-1k | 0.784 | 0.012 | 0.81 | 0.115 | −0.007 |
| ViT-S/16 | Sup. IN-21k → IN-1k | 0.831 | 0.015 | 0.62 | 0.097 | +0.009 |
| ViT-S/32 | Sup. IN-21k → IN-1k | 0.802 | 0.015 | 0.73 | 0.107 | +0.004 |
| ViT-B/16 | Sup. IN-21k → IN-1k | 0.824 | 0.032 | 0.67 | 0.100 | +0.029 |
| ViT-B/32 | Sup. IN-21k → IN-1k | 0.805 | 0.028 | 0.73 | 0.106 | +0.021 |
| ViT-L/16 | Sup. IN-21k → IN-1k | 0.821 | 0.059 | 0.73 | 0.107 | +0.059 |
| *Sup. IN-21k → IN-1k + AugReg* | | | | | | |
| ViT-S/16 | Sup. IN-21k → IN-1k + AugReg | 0.834 | 0.011 | 0.60 | 0.096 | −0.008 |
| ViT-B/16 | Sup. IN-21k → IN-1k + AugReg | 0.865 | 0.012 | 0.48 | 0.082 | +0.008 |
| ViT-B/32 | Sup. IN-21k → IN-1k + AugReg | 0.843 | 0.011 | 0.57 | 0.093 | +0.002 |
| ViT-L/16 | Sup. IN-21k → IN-1k + AugReg | 0.869 | 0.021 | 0.47 | 0.082 | +0.017 |
| *Web-scale pretrained* | | | | | | |
| ViT-B/32 | Web-scale (LAION-2B) → IN-1k | 0.857 | 0.065 | 0.61 | 0.091 | −0.065 |
| ViT-B/16 | Web-scale (OpenAI) → IN-1k | 0.873 | 0.061 | 0.55 | 0.085 | −0.061 |
| ViT-B/16 | Web-scale (LAION-2B) → IN-1k | 0.876 | 0.127 | 0.63 | 0.099 | −0.127 |
| ViT-L/14 | Web-scale (OpenAI) → IN-1k | 0.884 | 0.047 | 0.49 | 0.078 | −0.040 |
| ViT-L/14 | Web-scale (LAION-2B) → IN-1k | 0.882 | 0.097 | 0.55 | 0.087 | −0.097 |
| ViT-H/14 | Web-scale (LAION-2B) → IN-1k | 0.888 | 0.121 | 0.56 | 0.090 | −0.121 |

## E AI USAGE DISCLOSURE

Generative AI tools were used during the preparation of this work to assist with language editing, improving readability, and refining the presentation and structure of the manuscript. They were also used for limited support in writing and editing software code.

Generative AI tools were not used to generate synthetic datasets, formulate mathematical claims or proofs, develop the core methodology, design the experiments, or independently interpret the experimental results. All AI-assisted text, code, and suggestions were reviewed, verified, and, where necessary, modified by the authors.

The authors take full responsibility for the final content of this work, including all claims, results, code, and text produced with the assistance of generative AI tools.

# F Reproducibility and Benchmark Artifacts

We release the benchmark as a structured collection of evaluation code, model predictions, fine-tuned checkpoints, and standardized splits, with the goal of enabling future calibration research to plug into our pipeline without re-running inference on the full timm model zoo. Two release stages are planned. *For the review period*, the supplementary material accompanying this submission contains the aggregated metrics table for all 1,218 ImageNet-1k checkpoints, a reference implementation of logits extraction and evaluation of all metrics. *At the camera-ready stage*, we will additionally release the full set of cached logits, the medical fine-tuning checkpoints, and a documented evaluation interface, hosted on a public repository under a permissive license.

**Evaluation code (review and camera-ready).** A reference implementation of all metrics reported in this paper — ECE under multiple binning schemes (equal-width with $M \in \{15, 100\}$), top-label and multi-class Brier scores, NLL, and signed calibration bias — is provided as a self-contained Python module. The module operates on PyTorch tensors and is independent of the model loading pipeline, so users can import the metrics directly to evaluate their own logits. A second module provides the temperature-scaling implementation used in Section 4.3, fitted via L-BFGS on the held-out validation set.

**Aggregated metrics table (review and camera-ready).** A CSV containing per-model values for top-1 accuracy, ECE under $M = 15$ and $M = 100$ equal-width binning, signed calibration bias, the timm model identifier, the recipe categorization, parameter count, and input resolution is provided for the full 1,218-model population. This table is the basis for Figure 2 and can be re-aggregated by users to filter for specific architectures or recipes.

**Standardized splits (camera-ready).** The 80/20 ImageNet-1k validation split (40,000-image evaluation set, 10,000-image fitting set) is released as a deterministic list of image indices generated with seed 42, so that future evaluations reuse the same split.

**Cached model predictions (camera-ready).** At the camera-ready stage, we release the cached logits and predicted labels for all 1,218 timm checkpoints of the large-scale benchmark on the ImageNet-1k validation set, the four natural distribution shifts (ImageNet-V2, -R, -A, -Sketch) and the 19 ImageNet-C corruptions at five severity levels each, and the medical fine-tuning models on the four medical test splits. Cached logits avoid the need to re-download model weights and re-run inference, which substantially lowers the cost of benchmarking new recalibration methods on our model population.

**Medical fine-tuning checkpoints (camera-ready).** We release the fine-tuned weights of the four foundation-model backbones (DINOv2, DINOv3, CLIP-WIT, CLIP-LAION; Appendix A.1.6) on each of the four medical tasks. These checkpoints reach state-of-the-art accuracy on Path, Derma, OCT, and Blood and may be of independent interest beyond the calibration-focused analysis of this paper.

**Evaluation interface (camera-ready).** A new recalibration method can be benchmarked on our population with three steps: load the cached logits for the 1,218 models, apply the recalibration transform to each logit tensor, and compute metrics with the released module. Results can be compared directly against the population-level statistics of this paper without re-running any model inference.

# G Compute Resources

All experiments were run on an internal cluster with access to 8 NVIDIA A100 80GB GPUs.

**Inference on the large-scale ImageNet benchmark.** The 1,218 `timm` checkpoints were evaluated on the ImageNet-1k validation set (50k images), the four natural distribution shifts (ImageNet-V2, -R, -A, -Sketch; 7.5k–50k images each), and the 19 ImageNet-C corruptions at five severity

levels each, yielding approximately 122,000 model–split evaluations in total. Inference was distributed across the 8 A100 GPUs and required approximately 20–25 GPU-days in total (≈3 days of wall-clock time). Logits are cached so that subsequent metric computation and recalibration experiments do not require re-running inference.

**Controlled ViT ablation.** The 2,015 ViT configurations of Steiner et al. (2022) were evaluated on the ImageNet-1k validation set under the same inference pipeline. This required approximately 3–5 GPU-days, corresponding to less than a day of wall-clock time on the 8 A100 GPUs.

**Medical imaging fine-tuning.** Each (backbone, task) combination was fine-tuned end-to-end on a single A100 for 20 epochs on the larger datasets (Path, OCT) and 50 epochs on the smaller datasets (Derma, Blood); see Table 14. Single-run wall-clock times ranged from approximately 2–4 hours on Derma and Blood to 6–10 hours on Path and OCT. The full medical evaluation comprises 16 runs (4 backbones $\times$ 4 tasks) and required approximately 15–20 GPU-days in total, parallelized across the 8 A100 GPUs (≈2–3 days of wall-clock time).

**Total compute.** The full set of experiments reported in this paper required approximately 40–50 A100 GPU-days. Preliminary experiments — including pipeline development, debugging of the inference loop, and exploratory fine-tuning runs that did not enter the final evaluation — added approximately 20–30% on top of this figure.