Revisiting the Calibration of Modern Neural Networks

Matthias Minderer **Josip Djolonga** **Rob Romijnders** **Frances Hubis**
Xiaohua Zhai **Neil Houlsby** **Dustin Tran** **Mario Lucic**
Google Research, Brain Team
{mjlm, lucic}@google.com

Abstract

Accurate estimation of predictive uncertainty (model calibration) is essential for the safe application of neural networks. Many instances of miscalibration in modern neural networks have been reported, suggesting a trend that newer, more accurate models produce poorly calibrated predictions. Here, we revisit this question for recent state-of-the-art image classification models. We systematically relate model calibration and accuracy, and find that the most recent models, notably those not using convolutions, are among the best calibrated. Trends observed in prior model generations, such as decay of calibration with distribution shift or model size, are less pronounced in recent architectures. We also show that model size and amount of pretraining do not fully explain these differences, suggesting that architecture is a major determinant of calibration properties.

1 Introduction

Neural networks, especially vision models, are increasingly used in safety-critical applications such as autonomous driving (Bojarski et al., 2016), medical diagnosis (Esteva et al., 2017; Jiang et al., 2012), and meteorological forecasting (Sønderby et al., 2020). For such applications, it is essential that model predictions are not just accurate, but also well calibrated. Model calibration refers to the accuracy with which the scores provided by the model reflect its predictive uncertainty. For example, in a medical application, we would like to defer images for which the model makes low-confidence predictions to a physician for review (Kompa et al., 2021). Skipping human review due to confident, but incorrect, predictions, could have disastrous consequences.

While intense research and engineering effort has focused on improving the predictive accuracy of models, less attention has been given to model calibration. In fact, over the last few years, there have been many reports that calibration of modern neural networks can be surprisingly poor, despite the advances in accuracy (e.g. Guo et al. 2017; Lakshminarayanan et al. 2017; Malinin & Gales 2018; Thulasidasan et al. 2019; Hendrycks et al. 2020b; Ovadia et al. 2019; Wenzel et al. 2020; Havasi et al. 2021; Rahaman & Thiery 2020; Leathart & Polaczuk 2020). Some works suggest a trend for larger, more accurate models to be worse calibrated (Guo et al., 2017).

These concerns are more relevant than ever, since the architecture size, amount of training data, and computing power used by state-of-the-art models continue to increase. At the same time, rapid advances in model architecture (Tolstikhin et al., 2021; Dosovitskiy et al., 2021) and training approaches (Chen et al., 2020; Mahajan et al., 2018; Radford et al., 2021) raise the question whether past results on calibration, largely obtained on standard convolutional architectures, extend to current state-of-the-art models. Since model advances are quickly translated to real-world, safety-critical applications (e.g. Mustafa et al. 2021), there is an urgent need to re-assess the calibration properties of current state-of-the-art models.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Contributions. To address this need, we provide a systematic comparison of recent image classification models, relating their accuracy, calibration, and design features. We find that:

1. The best current models, including the non-convolutional MLP-Mixer (Tolstikhin et al., 2021) and Vision Transformers (Dosovitskiy et al., 2021), are well calibrated compared to past models and their performance is more robust to distribution shift.
2. In-distribution calibration slightly deteriorates with increasing model size, but this is outweighed by a simultaneous improvement in accuracy.
3. Under distribution shift, calibration *improves* with model size, reversing the trend seen in-distribution.
4. Accuracy and calibration are correlated under distribution shift, such that optimizing for accuracy may also benefit calibration.
5. Model size, pretraining duration, and pretraining dataset size cannot fully explain differences in calibration properties between model families.

Our results suggest that further improvements in model accuracy will continue to benefit calibration. They also hint at architecture as an important determinant of model calibration. We provide code and a large dataset of calibration measurements, comprising 180 distinct models from 16 families, each evaluated on 79 ImageNet-scale datasets and 28 metric variants.[1]

2 Related Work

Measures of model calibration. The losses that are commonly used to train classification models, such as cross-entropy and squared error, are proper scoring rules (Gneiting et al., 2007) and are therefore guaranteed to yield perfectly calibrated models at their minimum—in the infinite-data limit. However, in practice, due to model mismatch and overfitting, even losses based on proper scoring rules may result in poor model calibration. Miscalibration is commonly quantified in terms of Expected Calibration Error (ECE; Naeini et al. 2015), which measures the absolute difference between predictive confidence and accuracy. We focus on ECE because it is a widely used and accepted calibration metric. Nevertheless, it is well understood that estimating ECE accurately is difficult because estimators can be strongly biased and many estimator variants exist (Nixon et al., 2019; Roelofs et al., 2020; Vaicenavicius et al., 2019; Gupta et al., 2021). Section 5 discusses these issues and our approaches to mitigate them.

Alternatives to ECE include likelihood measures, Brier score (Brier, 1950), Bayesian methods (Gelman et al., 2013), and conformal prediction (Shafer & Vovk, 2008). Further, model calibration can be represented visually with reliability diagrams (DeGroot & Fienberg, 1983). Figure 8 and Appendix F provide likelihoods, Brier scores, and reliability diagrams for our main analyses.

Empirical studies of model calibration. There have been many recent empirical studies on the robustness (accuracy under distribution shift) of image classifiers (Geirhos et al., 2019; Taori et al., 2020; Djolonga et al., 2020; Hendrycks et al., 2020a). Several works have also studied calibration. Most notable is Guo et al. (2017), who found that "modern neural networks, unlike those from a decade ago, are poorly calibrated", that larger networks tend to be calibrated worse, and that "miscalibration worsen[s] even as classification error is reduced." Other works have corroborated some of these findings (e.g., Thulasidasan et al. 2019; Wen et al. 2021). This line of work suggests a trend that larger models are worse calibrated, which would have major implications for research toward bigger models and datasets. We show that for more recent models, this trend is negligible in-distribution and in fact reverses under distribution shift.

Ovadia et al. (2019) empirically study calibration under distribution shift and provide a large comparison of methods for improving calibration. They report that both accuracy and calibration deteriorate with distribution shift. While we observe the same trend, we find that the calibration of some recent model families decays so slowly under distribution shift that the decay in accuracy is likely more relevant in practice (Section 4.3).

Ovadia et al. also find that, *across methods for improving calibration*, improvements on in-distribution data do not necessarily translate to out-of-distribution data. This finding may suggest that there is little correlation between in-distribution and out-of-distribution calibration in general. However, our

[1] Available at `https://github.com/google-research/robustness_metrics/tree/master/robustness_metrics/projects/revisiting_calibration`.

results show that, *across model architectures*, the models with the best in-distribution calibration are also the best-calibrated on a range of out-of-distribution benchmarks. The important implication of this result is that designing models based on in-distribution performance likely also benefits their out-of-distribution performance.

Improving calibration. Many strategies have been proposed to improve model calibration such as post-hoc rescaling of predictions (Guo et al., 2017), averaging multiple predictions (Lakshminarayanan et al., 2017; Wen et al., 2020), and data augmentation (Thulasidasan et al., 2019; Wen et al., 2021). Here, we focus on the intrinsic calibration properties of state-of-the-art model families, rather than methods to further improve calibration.

As a baseline on top of a model's intrinsic calibration properties, we study temperature scaling (Guo et al., 2017). It is effective in improving calibration and so simple that it can be applied in many cases at minimal additional cost, in contrast to many more sophisticated methods. Temperature scaling re-scales a model's logits by a single parameter, chosen to optimize the model's likelihood on a held-out portion of the training data. This temperature factor changes the model's *confidence*, i.e., whether the model predictions are on average too certain (overconfident), optimally confident, or too uncertain (underconfident). The classification accuracy of the model is not affected by temperature scaling. A large fraction of model miscalibration is typically due to average over- or underconfidence, e.g. due to suboptimal training duration (Guo et al., 2017). By normalizing a model's confidence, temperature scaling not only improves calibration, but also removes a primary confounder that can hide trends in calibration between models (see Section 4.2 and Appendix D). Therefore, we study both unscaled and temperature-scaled predictions in the paper.

3 Definitions and Notation

We consider the multi-class classification problem, as analyzed by Bröcker (2009), where we observe a variable X and predict a categorical variable $Y \in \{1, 2, \ldots, k\}$. We model our predictor f as a function that maps every input instance X to a categorical distribution over k labels, represented using a vector $f(X)$ belonging to the $(k-1)$-dimensional simplex $\Delta = \{p \in [0,1]^k \mid \sum_{y=1}^{k} p_y = 1\}$.

Intuitively, a model f is well-calibrated if its output truthfully quantifies the predictive uncertainty. For example, if we take all data points x for which the model predicts $[f(x)]_y = 0.3$, we expect 30% of them to indeed take on the label y. Formally, the model f is said to be calibrated if (Bröcker, 2009)

$$\forall p \in \Delta \colon P(Y = y \mid f(X) = p) = p_y. \tag{1}$$

We will focus on a slightly weaker, but more practical condition, called top-label or argmax calibration (Kumar et al., 2019; Guo et al., 2017). This requires that the above holds only for the most likely label, i.e., $\forall p^* \in [0,1]$

$$P(Y \in \arg\max p \mid \max f(X) = p^*) = p^*, \tag{2}$$

where the $\max$ and $\arg\max$ act coordinate-wise.

The most common measure of the degree of miscalibration is the *Expected Calibration Error (ECE)*, which computes the expected disagreement between the two sides of eq. (2)

$$\mathbb{E}\big[|p^* - E[Y \in \arg\max f(X) \mid \max f(X) = p^*|\big]. \tag{3}$$

Unfortunately, eq. (3) cannot be estimated without quantization as it conditions on a null event. Hence, one typically first buckets the predictions into m bins $B_1, \ldots, B_m$ based on their top predicted probability, and then takes the expectation over these buckets. Namely, if we are given a set of n i.i.d. samples $(x_1, y_1), \ldots, (x_n, y_n)$ distributed as $P(X, Y)$, then we assign each $j \in \{1, \ldots, n\}$ to a bucket B_i based on $\max f(x_j)$. Then, we compute in each bucket B_i the $\text{confidence}(B_i) = \frac{1}{|B_i|} \sum_{j \in B_i} \max f(x_j)$ and the $\text{accuracy}(B_i) = \frac{1}{|B_i|} \sum_{j \in B_i} [\![y_j \in \arg\max f(x_j)]\!]$, where $[\![\cdot]\!]$ is the Iverson bracket. Finally, we construct an estimator by taking the expectation over the bins

$$\widehat{\text{ECE}} = \sum_{i=1}^{m} \frac{|B_i|}{n} \left|\text{accuracy}(B_i) - \text{confidence}(B_i)\right|. \tag{4}$$

In Section 5 we discuss the statistical properties of this estimator, possible pitfalls, and several mitigation strategies.

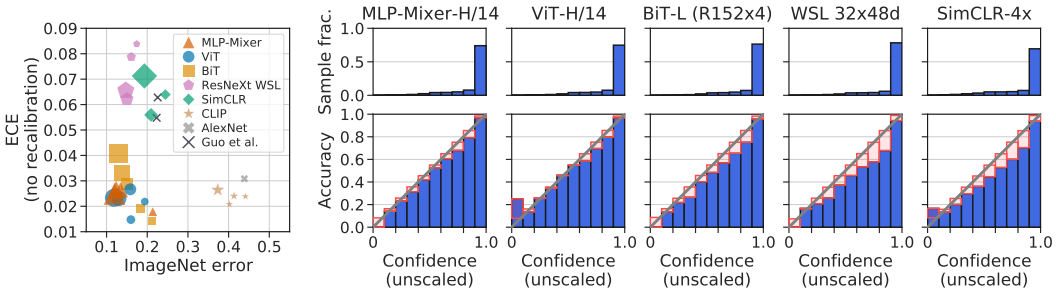

Figure 1: Some modern neural network families are both highly accurate and well-calibrated. Left: Expected calibration error (ECE) vs. classification error on IMAGENET for state-of-the-art image classification models. Marker size indicates relative model size within its family. Points labeled "Guo et al." are the values reported for DenseNet-161 and ResNet-152 in Guo et al. (2017). Right: Confidence distribution (top row) and reliability diagrams (bottom row) for some of the models.

4 Empirical Evaluation

4.1 Experimental Setup

Model families. In this study, we consider a range of recent and some historic state-of-the-art image classification models. Our selection of models covers convolutional and non-convolutional architectures, as well as supervised, weakly supervised, unsupervised and zero-shot training. We follow the original publications in naming the model variants within each family (e.g. different model sizes). See Appendix A.1 for a detailed description of all used models.

1. MLP-Mixer (Tolstikhin et al., 2021) is based exclusively on multi-layer perceptrons (MLPs) and is pre-trained on large supervised datasets.
2. ViT (Dosovitskiy et al., 2021) processes images with a transformer architecture originally designed for language (Vaswani et al., 2017) and is also pre-trained on large supervised datasets.
3. BiT (Kolesnikov et al., 2020) is a ResNet-based architecture (He et al., 2016). It is also pre-trained on large supervised datasets.
4. ResNext-WSL (Mahajan et al., 2018) is based on the ResNeXt architecture and trained with weak supervision from billions of hashtags on social media images.
5. SimCLR (Chen et al., 2020) is a ResNet, pretrained with an unsupervised contrastive loss.
6. CLIP (Radford et al., 2021) is pretrained on raw text and imagery using a contrastive loss.
7. AlexNet (Krizhevsky et al., 2012; Krizhevsky, 2014) was the first convolutional neural network to win the ImageNet challenge.

All models are either trained or fine-tuned on the IMAGENET training set, except for CLIP, which makes zero-shot predictions using IMAGENET class names as queries.

Datasets. We evaluate accuracy and calibration on the IMAGENET validation set and the following out-of-distribution benchmarks using the Robustness Metrics library (Djolonga et al., 2020):

1. IMAGENETV2 (Recht et al., 2019) is a new IMAGENET test set collected by closely following the original IMAGENET labeling protocol.
2. IMAGENET-C (Hendrycks & Dietterich, 2019) consists of the images from IMAGENET, modified with synthetic perturbations such as blur, pixelation, and compression artifacts at a range of severities.
3. IMAGENET-R (Hendrycks et al., 2020a) contains artificial renditions of IMAGENET classes such as art, cartoons, drawings, sculptures, and others.
4. IMAGENET-A (Hendrycks et al., 2021) contains images that are classified as belonging to IMAGENET classes by humans, but adversarially selected to be hard to classify for a ResNet50 trained on IMAGENET.

For the post-hoc recalibration of models, we reserve 20% of the IMAGENET validation set (randomly sampled) for fitting the temperature scaling parameter. All reported metrics are computed on the remaining 80% of the data. For evaluations on IMAGENET-C, we also exclude the 20% of images that are based on the IMAGENET images used for temperature scaling.

Calibration metric. Throughout the paper, we estimate ECE using equal-mass binning and 100 bins. Appendix E shows that our results hold for other ECE variants and are consistent with the Brier score and model likelihood.

4.2 In-Distribution Calibration

We begin by considering ECE on clean IMAGENET images (referred to as in-distribution). Figure 1 shows in-distribution ECE and reliability diagrams before any recalibration of the predicted probabilities. We find that several recent model families (MLP-Mixer, ViT, and BiT) are both highly accurate *and* well-calibrated compared to prior models, such as AlexNet or the models studied by Guo et al. (2017). This suggests that there may be no continuing trend for highly accurate modern neural networks to be poorly calibrated, as suggested previously (Guo et al., 2017; Lakshminarayanan et al., 2017; Malinin & Gales, 2018; Thulasidasan et al., 2019; Hendrycks et al., 2020b; Ovadia et al., 2019; Wenzel et al., 2020; Havasi et al., 2021; Rahaman & Thiery, 2020; Leathart & Polaczuk, 2020). In addition, we find that a recent zero-shot model, CLIP, is well-calibrated given its accuracy.

Temperature scaling reveals consistent properties of model families. The poor calibration of past models can often be remedied by post-hoc recalibration such as temperature scaling (Guo et al., 2017), which raises the question whether a difference between models remains after recalibration. We find that the most recent architectures are better calibrated than past models *even after temperature scaling* (Figure 2, right).

More generally, temperature scaling reveals consistent trends in the calibration properties between families that are obscured in the unscaled data by simple over- or underconfidence miscalibration. Before temperature scaling (Figure 2, left), several families overlap in their accuracy/calibration properties (MLP-Mixer, ViT, BiT). After temperature scaling (Figure 2, right), a clearer separation of families and consistent trends between accuracy and calibration within each family become apparent. Notably, temperature scaling reconciles our results for BiT (a ResNet architecture) with the results reported by Guo et al. for ResNets trained on IMAGENET. Furthermore, models pretrained without additional labels (SimCLR) or with noisy labels (ResNeXt-WSL) tend to be calibrated worse for a given accuracy than ResNets trained with supervision (BiT and the models studied by Guo et al.). Finally, non-convolutional model families like MLP-Mixer and ViT can perform just as well, if not better, than convolutional ones.

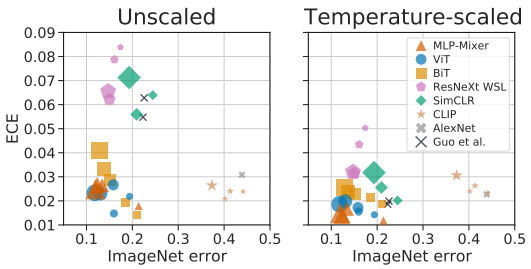

Figure 2: Temperature scaling reveals consistent properties of model families. Left: ECE vs. classification error as in Figure 1. Right: ECE after applying temperature scaling.

Differences between families are not explained by model size or pretraining amount. We next attempt to disentangle how the differences between model families affect their calibration properties. We focus on model size and amount of pretraining, both important trends in state of-the-art models.

We first consider model size. Prior work has suggested that larger neural networks are worse calibrated (Guo et al., 2017). We also find that within most families, larger members tend to have higher calibration error (Figure 2, right). However, at the same time, larger models have consistently lower classification error. This means that each model family occupies a different Pareto set in the tradeoff between accuracy and calibration. For example, our results suggest that, at any given accuracy, ViT models are better calibrated than BiT models. Changing the size of a BiT model cannot move it into the Pareto set of ViT models. Model size can therefore not fully explain the intrinsic calibration differences between these model families.[2]

We next consider model pretraining. Many current state-of-the-art image models use transfer learning, in which a model is pre-trained on a large dataset and then fine-tuned to the task of interest (Kolesnikov et al., 2020; Chen et al., 2020; Xie et al., 2020). With transfer learning, large data sources can be

[2]This relationship between model size, accuracy and calibration holds for all families we study except ResNeXt-WSL, for which increasing model sizes improves *both* accuracy *and* calibration. While investigating this difference was out of the scope of this work, it may be a promising direction for future research.

exploited to train the model, even if little data are available for the final task. To test how the amount of pretraining affects calibration, we compare BiT models pretrained on IMAGENET (1.3M images), IMAGENET-21K (12.8M images), or JFT-300 (300M images; Sun et al. 2017).

More pretraining data consistently increases accuracy, especially for larger models. It has no consistent effect on calibration (Figure 3). In particular, after temperature scaling, ECE is essentially unchanged across this 300-fold increase in pretraining dataset size (e.g. for BiT-R50x1 pretrained on IMAGENET, IMAGENET-21K and JFT-300, the ECEs are 0.0185, 0.0182, 0.0185, respectively; for BiT-R101x3, they are 0.0272, 0.0311, 0.0236; Figure 3, bottom). Therefore, regardless of the pretraining dataset, BiT always remains Pareto-dominant over SimCLR and Pareto-dominated by ViT and MLP-Mixer in our experiments.

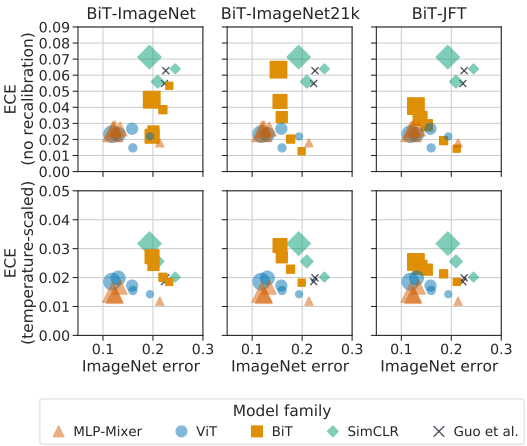

Figure 3: Family differences are not fully explained by the amount of pretraining. Each column shows ECE vs. classification error on ImageNet for BiT models pre-trained with a different dataset: IMAGENET (1.3M images), IMAGENET-21K (12.8M images), or JFT-300 (300M images). The values for other models are provided for reference in light shading (same values as in Figure 2). Note how all BiT models remain in the same relative location between ViT and SimCLR across a 300-fold difference in pretraining data size.

The BiT models compared in Figure 3 differ in both the amount of pretraining data and the duration of pretraining (see Kolesnikov et al. (2020) for details). To further disentangle these variables, we trained BiT models on varying numbers of pretraining examples while holding the number of training steps constant, and vice versa. We find that pretraining dataset size has no significant effect on calibration, while pretraining duration only shifts the model within its accuracy/calibration Pareto set (longer-trained models are more accurate and worse calibrated; Figure 10). These results suggest that pretraining alone cannot explain the differences between model families that we observe.

In summary, our results show that some modern neural network families combine high accuracy and state-of-the-art calibration on in-distribution data, both before and after post-hoc recalibration by temperature scaling. In Figure 8 and Appendices E and F, we show that these results generally hold for other measures of model calibration (other ECE variants, Brier score, and model likelihood). Our experiments further suggest that model size and pretraining amount do not fully explain the intrinsic calibration differences between model families. Given that the best-calibrated families (MLP-Mixer and ViT) are non-convolutional, we speculate that model architecture, and in particular its spatial inductive bias, play an important role.

4.3 Accuracy and Calibration Under Distribution Shift

For safety-critical applications, the model should produce reasonable uncertainty estimates not just in-distribution, but also under distribution shifts that were not anticipated at training time. We first assess out-of-distribution calibration on the IMAGENET-C dataset, which consists of images that have been synthetically corrupted at five different severities. As expected, both classification and calibration error generally increase with distribution shift (Figure 4; Ovadia et al. 2019; Hendrycks & Dietterich 2019). Interestingly, this decay in calibration performance is slower for MLP-Mixer and ViT than for the other model families, both before and after temperature scaling.

Regarding the effect of model size on calibration, we observed some trend towards worse calibration of larger models on in-distribution data. However, the trend is reversed for most model families as we move out of distribution, especially after accounting for confidence bias by temperature scaling (note positive slope of the gray lines at high corruption severities in Figure 4, bottom row). In other words, the calibration of larger models is more robust to distribution shift (Figure 5).

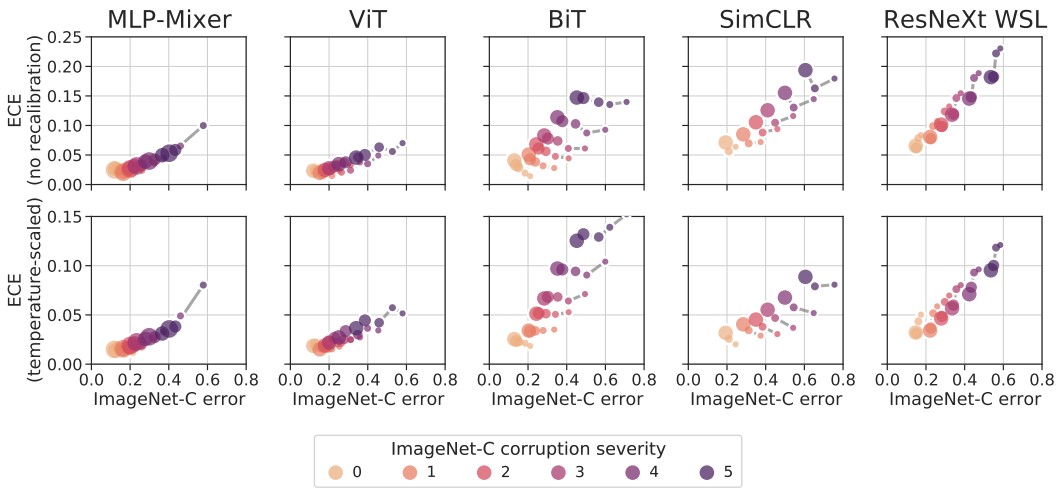

Figure 4: Calibration and accuracy on IMAGENET-C before (top) and after (bottom) temperature scaling on IMAGENET. Severity 0 refers to the clean IMAGENET test set; marker size indicates relative model size within its family (see Table 1 for model details). The calibration of some recent model families, e.g. MLP-Mixer and ViT, is more robust to distribution shift than past models.

We next consider to what degree the insights from in-distribution and IMAGENET-C calibration transfer to natural out-of-distribution data. Previous work on IMAGENET-C suggests that, when comparing recalibration methods, better in-distribution calibration and accuracy do not usually predict better calibration under distribution shift (Ovadia et al., 2019). Here, comparing model families, we find that the performance on several natural out-of-distribution datasets is largely consistent with that on IMAGENET (Figure 6). In particular, models that are Pareto-optimal (i.e. no other model is both more accurate and better calibrated) on IMAGENET remain Pareto-optimal on the OOD datasets. Further, we observe a strong correlation between accuracy and calibration on the OOD datasets. This relationship is consistent across models within a family and across datasets, over a wide range of accuracies (Figure 11).

These results suggest that larger and more accurate models, and in particular MLP-Mixer and ViT, can maintain their good in-distribution calibration even under severe distribution shifts. Based on the observed relationship between calibration and accuracy, we can reasonably hope that good calibration on in-distribution data (and anticipated distribution shifts) generally translates into good calibration on unanticipated out-of-distribution data, similar to what has been observed for accuracy (Djolonga et al., 2020).

4.4 Relating Accuracy and Calibration Within Model Families

Our data suggest that most model families lie on different Pareto sets in the accuracy/calibration space, which establishes a clear preference *between* families. We next consider how to compare individual models *within* a family (or more specifically, within a Pareto set), where one model is more accurate but worse calibrated, and the other is less accurate but better calibrated. Which model should a practitioner choose for a safety-critical application?

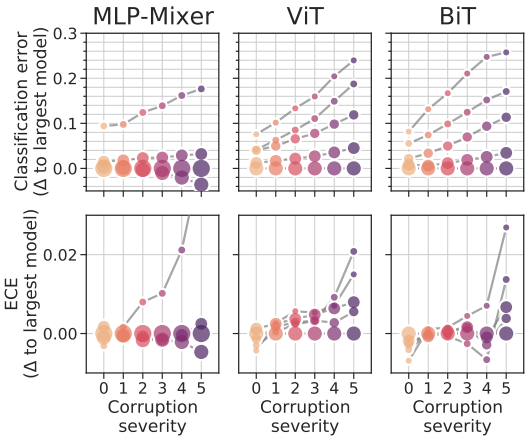

Figure 5: Classification error and ECE for the top three families on IMAGENET-C, relative to the largest model variant in each family. As distribution shift increases, both errors tend to increase more slowly for larger models. Also note that changes in ECE are much smaller than changes in classification error.

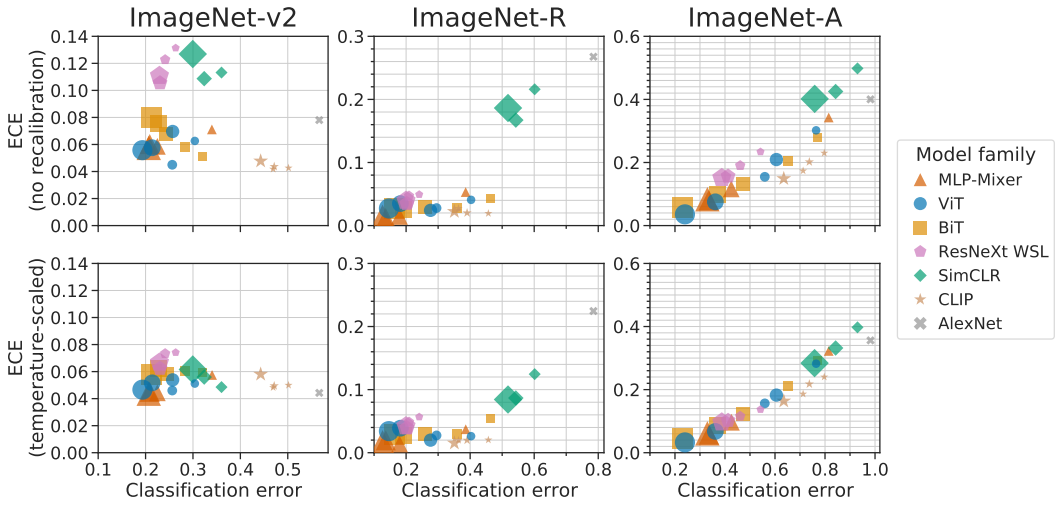

Figure 6: Calibration and accuracy before (top row) and after (bottom row) temperature scaling on out-of-distribution benchmarks. Marker size indicates relative model size within its family. IMAGENET-R and IMAGENET-A use a reduced subset of 200 classes; we follow the literature and select the subset of the model logits for these classes before evaluation. Out-of-distribution calibration tends to correlate with in-distribution calibration (Figure 1) and out-of-distribution accuracy.

The answer depends on the cost structure of the specific application (Hernández-Orallo et al., 2012). As an example, consider the scenario of *selective prediction*, which is common in medical diagnosis. In this task, one can choose to ignore the model prediction ("abstain") at a fixed cost if the prediction confidence is low, rather than risking a (more costly) prediction error.

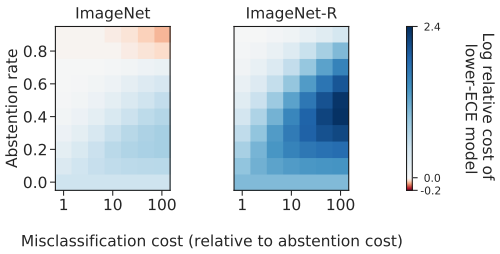

Figure 7: Relative cost of BiT-R152x4 and BiT-R50x1 models in a selective prediction scenario, computed as a combination of the misclassification and abstention costs at a given cost ratio (x-axis) and abstention rate (y-axis). Blue indicates regions where the higher-accuracy model (R152x4) achieves a lower cost than the better-calibrated model (R50x1). The accuracy advantage outweighs the calibration advantage for practical rejection rates, across all tested abstention costs.

Figure 7 compares the expected cost for two BiT variants, one with better with better classification error (R152x4, by 0.08), and one with better ECE (R50x1, by 0.009). For abstention rates up to 70% (which covers most practical scenarios with abstention rates low enough for the model to be useful), the model with better accuracy has a lower overall cost than the model with better ECE. The same is true for all other model families we study (Appendix B.3). For these families and this cost scenario, a practitioner should therefore always choose the most accurate available model regardless of differences in calibration. Ultimately, real-world cost structures are complex and may yield different results; Figure 7 presents one common scenario with downstream ramifications for the importance of the calibration differences compared to accuracy.

5 Pitfalls and Limitations

For this study, we approached calibration with a simple, practical question: *Given two models, one more accurate and the other better calibrated, which should a practitioner choose?* While working towards answering this question, we encountered several pitfalls that complicate the interpretation of calibration results.

Measuring calibration is challenging, and while the quantity we want to estimate is well specified, the estimator itself can be biased. There are two sources of bias: (i) from estimating ECE by binning, and (ii) from the finite sample size used to estimate the per-bin statistics.

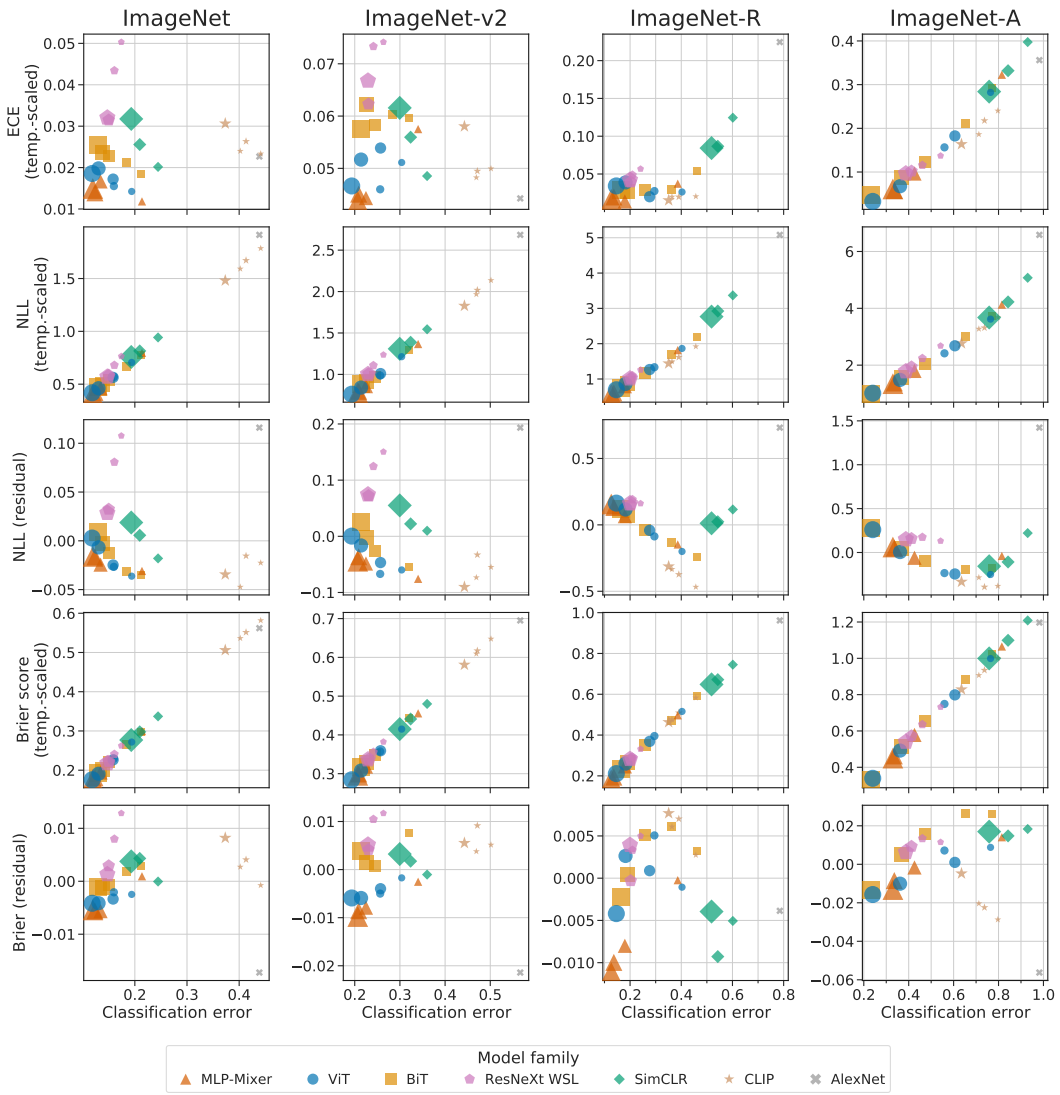

Figure 8: Alternative calibration metrics: negative log-likelihood (NLL) and Brier score. For comparison, the first row shows ECE as in Figure 6. Since NLL, Brier score, and classification error are all highly correlated, we also provide the residuals of NLL and Brier score after regressing out classification error (third and fifth row). Specifically, we first fit a linear regression $y_i = \beta_0 + \beta_1 x_i$, where x_i is the classification error and y_i is the calibration measure of model i. We then report the residual $y_i - (\beta_0 + \beta_1 x_i)$ on the y-axis of the plots in the third and fifth row. The residuals show which models have better (or worse) NLL and Brier score than what can be expected from their accuracy alone. The relationships between model families are largely similar across all calibration metrics.

The first of these biases is always negative (Kumar et al., 2019), while the second one is always positive. Thus, the estimator can both under- and over-estimate the true value, and the magnitude of the bias can depend on multiple factors. In practice, this means that the ranking of models depends on which ECE variant is chosen to estimate calibration (Nixon et al., 2019). As we show below, this is especially problematic for the positive bias, because this bias *depends on the accuracy of the model*. It is therefore possible to arrive at opposite conclusions about the relationship between accuracy and calibration, depending on the chosen bin size (Figure 9), especially when comparing models with widely varying accuracies.

Intuitively, a larger number of bins implies fewer points per bin and thus higher variance of the estimate of the model accuracy in each bin, which adds positive bias to the estimate of ECE. More formally, to estimate accuracy(B_i) precisely, we need a number of samples inversely proportional to the standard deviation $\sqrt{p_i(1-p_i)}/|B_i|$, where p_i is the expected accuracy in B_i. This indicates that the bias would be smaller for models with extreme average accuracies (i.e. close to 0 or 1) and larger for models with an accuracy close to 0.5. A detailed analysis reveals additional effects that further reduce the bias for higher-accuracy models (Appendix C). In particular, if we estimate $(\mathbb{E}[\text{accuracy}(B_i)]-\mathbb{E}[\text{confidence}(B_i)])^2$ for any bin i with n_i samples (using the sample means of the confidences and the accuracies), the bias can be shown to be equal to (conditioning on $X \in B_i$ omitted for brevity)

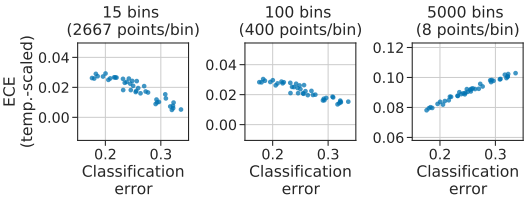

Figure 9: The effect of binning-induced bias in ECE depends on accuracy. Each dot represents a BiT ResNet model. Plotted models differ in model size, pretraining dataset size, and pretraining duration. All models are fine-tuned and evaluated on IMAGENET. After temperature scaling, there is a near-linear relationship between ECE and classification error. However, whether this relationship is positive or negative depends on the number of bins used for estimating ECE. This effect is explained by an accuracy-dependent bias that increases with the number of bins.

$$\frac{1}{n_i}\big(\mathbb{V}[A] + \mathbb{V}[C] - 2\text{Cov}[C, A]\big), \tag{5}$$

where $A = [\![Y \in \arg\max f(X)]\!]$ and $C = \max f(X)$. Hence, from Equation 5 we can conclude that higher accuracy models have a lower bias not only due to higher accuracy (lower $\mathbb{V}[A]$), but also because their outputs correlate more with the correct label (higher covariance).

In addition to a careful choice of bin size, considering accuracy and calibration jointly mitigates this issue, because the Pareto-optimal models rarely change, even if the ranking based on ECE alone does (Appendix E). In Appendices E and F, we provide the main figures of the paper for other ECE variants (number of bins, binning scheme, normalization metric, top-label, all-label, class-wise).

Finally, metrics such as Brier score (Brier, 1950) and likelihood provide alternative assessments of model calibration that do not require estimating expected calibration error. We find that the relationships between model families are consistent across ECE, NLL and Brier score (Figure 8). In particular, the same models (specifically the largest MLP-Mixer and ViT variants) remain Pareto-optimal with respect to the calibration metric and classification error in most cases. The relationship between models is visualized especially clearly after regressing out from the calibration metrics their correlation with classification error (Figure 8, third and fifth row).

6 Conclusion

We performed a large study of the calibration of recent state-of-the-art image models and its relationship with accuracy. We find that modern image models are well calibrated across distribution shifts despite being designed with a focus on accuracy. Our results suggests that there is no general trend for recent or highly accurate neural networks to be poorly calibrated compared to older or less accurate models.

Our experiments suggest that simple dimensions such as model size and pretraining amount do not fully account for the performance differences between families, pointing towards architecture as a major determinant of calibration. Of particular note is the finding that MLP-Mixer and Vision Transformers—two recent architectures that are not based on convolutions—are among the best-calibrated models both in-distribution and out-of-distribution. Self-attention (which Vision Transformers employ heavily) has been shown previously to be beneficial for certain kinds of out-of-distribution robustness (Hendrycks et al., 2020a). Our work now hints at calibration benefits of non-convolutional architectures more broadly, for certain kinds of distribution shift. Further work on the influence of architectural inductive biases on calibration and out-of-distribution robustness will be necessary to tell whether these results generalize. If so, they may further hasten the end of the convolutional era in computer vision.

Acknowledgments and Disclosure of Funding

We thank Carlos Riquelme and Balaji Lakshminarayanan for valuable comments on the manuscript. The authors declare no competing interests.

References

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv: 1604.07316*, 2016.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 1950.

Bröcker, J. Reliability, sufficiency, and the decomposition of proper scores. *Journal of the Royal Meteorological Society*, 2009.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society*, 1983.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

Djolonga, J., Matthias, M., Nado, Z., Nixon, J., Romijnders, R., Tran, D., and Lucic, M. Robustness Metrics, 2020. URL https://github.com/google-research/robustness_metrics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. CRC press, 2013.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, 2007.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. Calibration of neural networks using splines. *International Conference on Learning Representations*, 2021.

Havasi, M., Jenatton, R., Fort, S., Liu, J. Z., Snoek, J., Lakshminarayanan, B., Dai, A. M., and Tran, D. Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016.

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

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv: 2006.16241*, 2020a.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020b.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Conference on Computer Vision and Pattern Recognition*, 2021.

Hernández-Orallo, J., Flach, P. A., and Ferri, C. A unified view of performance metrics: translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 2012.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 2012.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision*, 2020.

Kompa, B., Snoek, J., and Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 2021.

Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. *arXiv: 1404.5997*, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Leathart, T. and Polaczuk, M. Temporal probability calibration. *arXiv: 2002.02644*, 2020.

Mahajan, D., Girshick, R. B., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision*, 2018.

Malinin, A. and Gales, M. J. F. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 2019.

Mustafa, B., Loh, A., Freyberg, J., MacWilliams, P., Wilson, M., McKinney, S. M., Sieniek, M., Winkens, J., Liu, Y., Bui, P., Prabhakara, S., Telang, U., Karthikesalingam, A., Houlsby, N., and Natarajan, V. Supervised transfer learning at scale for medical imaging. *arXiv: 2101.05913*, 2021.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J. V., Ren, J., Nado, Z., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. Technical report, OpenAI, 2021.

Rahaman, R. and Thiery, A. H. Uncertainty quantification and deep ensembles. *arXiv: 2007.08792*, 2020.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation. *arXiv: 2012.08668*, 2020.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N. Metnet: A neural weather model for precipitation forecasting. *arXiv: 2003.12140*, 2020.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *International Conference on Computer Vision*, 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 2020.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. MLP-Mixer: An all-MLP architecture for vision. *arXiv: 2105.01601*, 2021.

Vaicenavicius, J., Widmann, D., Andersson, C. R., Lindsten, F., Roll, J., and Schön, T. B. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Wen, Y., Tran, D., and Ba, J. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *Advances in Neural Information Processing Systems*, 2020.

Wen, Y., Jerfel, G., Muller, R., Dusenberry, M. W., Snoek, J., Lakshminarayanan, B., and Tran, D. Combining ensembles and data augmentation can harm your calibration. In *International Conference on Learning Representations*, 2021.

Wenzel, F., Snoek, J., Tran, D., and Jenatton, R. Hyperparameter ensembles for robustness and uncertainty quantification. In *Advances in Neural Information Processing Systems*, 2020.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Conference on Computer Vision and Pattern Recognition*, 2020.

Appendix

The Appendix is structured as follows:

A Models and Datasets **15**

Details and references for the models and datasets used in this work.

A.1 Models 15

A.2 Datasets 15

B Supplementary Analyses **16**

B.1 Fine-grained Analysis of Pretraining 16

Analyzes how the number of pretraining steps and the pretraining dataset size affect calibration and accuracy.

B.2 Correlation Between Calibration and Accuracy 17

Shows the relationship between calibration and accuracy, separated by model family and without averaging ImageNet-C datasets and severities.

B.3 Contribution of Accuracy and Calibration to Decision Cost 17

Compares selective prediction cost of additional model pairs.

C Sampling Bias for ℓ_2-ECE **18**

Provides further analysis of the bias of ECE and a proof that the bias depends on accuracy.

D Model Confidence **20**

Discusses factors influencing model over- and underconfidence, and how temperature scaling removes confidence miscalibration as a confounding variable when comparing models.

E ECE Variants **22**

Provides the main plots for a range of alternative ECE estimator variants (e.g. different binning scheme and number of bins).

F Alternative Calibration Metrics **22**

Provides negative log-likelihood, Brier score, and reliability diagrams for all models and datasets.

A Models and Datasets

A.1 Models

Table 1 provides an overview of the models used in this study. Model names link to the used checkpoints, where available.

Model name	Reference	Variant	Parameters
AlexNet	Krizhevsky et al. (2012)	8 layers	62.4M
BiT-L (R50-x1)	Kolesnikov et al. (2020)	ResNet50, 1× width	25.5M
BiT-L (R101-x1)	Kolesnikov et al. (2020)	ResNet101, 1× width	44.5M
BiT-L (R50-x3)	Kolesnikov et al. (2020)	ResNet50, 3× width	217.3M
BiT-L (R101-x3)	Kolesnikov et al. (2020)	ResNet101, 3× width	387.9M
BiT-L (R152-x4)	Kolesnikov et al. (2020)	ResNet154, 4× width	936.5M
CLIP	Radford et al. (2021)	ResNet50-based	25.5M
CLIP	Radford et al. (2021)	ViT-B32-based	88.3M
EfficientNet-NS (B1)	Xie et al. (2020)	18 layers, 1× width	7.9M
EfficientNet-NS (B3)	Xie et al. (2020)	31 layers, 1× width	12.3M
EfficientNet-NS (B5)	Xie et al. (2020)	45 layers, 2× width	30.6M
EfficientNet-NS (B7)	Xie et al. (2020)	64 layers, 2× width	66.7M
Mixer (B)	Tolstikhin et al. (2021)	B/16, JFT-300m	59.9M
Mixer (L)	Tolstikhin et al. (2021)	L/16, JFT-300m	280.5M
Mixer (H)	Tolstikhin et al. (2021)	H/14, JFT-300m	589.7M
Mixer (H)	Tolstikhin et al. (2021)	H/14, JFT-2.5b	589.7M
ResNeXt-WSL	Mahajan et al. (2018)	ResNeXt 101, 32x8d	88M
ResNeXt-WSL	Mahajan et al. (2018)	ResNeXt 101, 32x16d	193M
ResNeXt-WSL	Mahajan et al. (2018)	ResNeXt 101, 32x32d	466M
ResNeXt-WSL	Mahajan et al. (2018)	ResNeXt 101, 32x48d	829M
SimCLR (1x)	Chen et al. (2020)	ResNet50, 1× width	25.6M
SimCLR (2x)	Chen et al. (2020)	ResNet50, 2× width	98.1M
SimCLR (4x)	Chen et al. (2020)	ResNet50, 4× width	383.8M
ViT (B)	Dosovitskiy et al. (2021)	B/32	88.3M
ViT (B)	Dosovitskiy et al. (2021)	B/16	86.9M
ViT (L)	Dosovitskiy et al. (2021)	L/32	306.6M
ViT (L)	Dosovitskiy et al. (2021)	L/16	304.7M
ViT (H)	Dosovitskiy et al. (2021)	H/14	633.2M

Table 1: Overview of models used in this study. Per model family, the rows are sorted by increasing marker size in Figure 1 (i.e. approximate relative model size in terms of pretraining compute). We chose a qualitative scale to indicate model size because quantitative measures such as the number of parameters do not always reflect the representational power of a model. For example, ViT-B/16 has slightly fewer parameters than ViT-B/32 but requires more compute and is a more powerful model.

A.2 Datasets

We evaluate accuracy and calibration the following benchmark datasets:

1. IMAGENET (Deng et al., 2009) refers to the ILSVRC-2012 variant of the ImageNet database, a dataset of images of 1 000 diverse object classes. For evaluation, we use 40 000 images randomly sampled from the public validation set. We reserve the remaining 10 000 images for fitting the temperature scaling parameter.
2. IMAGENETV2 (Recht et al., 2019) is a new IMAGENET test set collected by closely following the original IMAGENET labeling protocol. The dataset contains 10 000 images.
3. IMAGENET-C (Hendrycks & Dietterich, 2019) consists of the images from IMAGENET, modified with synthetic perturbations such as blur, pixelation, and compression artifacts at a range of severities. The dataset includes 15 perturbations at 5 severities each, for a total of 75 datasets. For evaluation, we use the 40 000 images that were not derived from the IMAGENET images we used for temperature scaling.

4. IMAGENET-R (Hendrycks et al., 2020a) contains artificial renditions of IMAGENET classes such as art, cartoons, drawings, sculptures, and others. The dataset has 30 000 images of 200 classes. Following Hendrycks et al., we sub-select the model logits for the 200 classes before computing accuracy and calibration metrics.
5. IMAGENET-A (Hendrycks et al., 2021) contains images that are classified as belonging to IMAGENET classes by humans, but adversarially selected to be hard to classify by a ResNet50 trained on IMAGENET. The dataset has 7 500 samples of 200 classes. As for IMAGENET-R, we sub-select the logits for the 200 classes before computing accuracy and calibration metrics.

In addition, the following datasets are used for pretraining as described in the text:

1. IMAGENET-21K (Deng et al., 2009) refers to the full variant of the ImageNet database. It contains 14.2 million images of 21 000 object classes, organized by the WordNet hierarchy. Each image may have several labels.
2. JFT-300 (Sun et al., 2017) consists of approximately 300 million images, with 1.26 labels per image on average. The labels are organized into a hierarchy of 18 291 classes.

B Supplementary Analyses

B.1 Fine-grained Analysis of Pretraining

Section 4.2 and Figure 3 discuss the effect of the amount of pretraining on accuracy and calibration by comparing models pretrained on three different datasets. Figure 10 provides a more fine-grained analysis. We pretrained BiT models with varying dataset sizes or number of pretraining steps, while holding the other constant. Learning rate schedules were appropriately adapted to the number of steps, i.e. a separate model was trained with a full schedule for each condition, rather than comparing different checkpoints from the same training run. After pretraining, all models were finetuned on IMAGENET as in Kolesnikov et al. (2020).

We find that pretraining dataset size has little consistent effect on calibration error (Figure 10, left). Longer pretraining causes a slight increase in calibration error, but also decreases classification error (Figure 10, right).

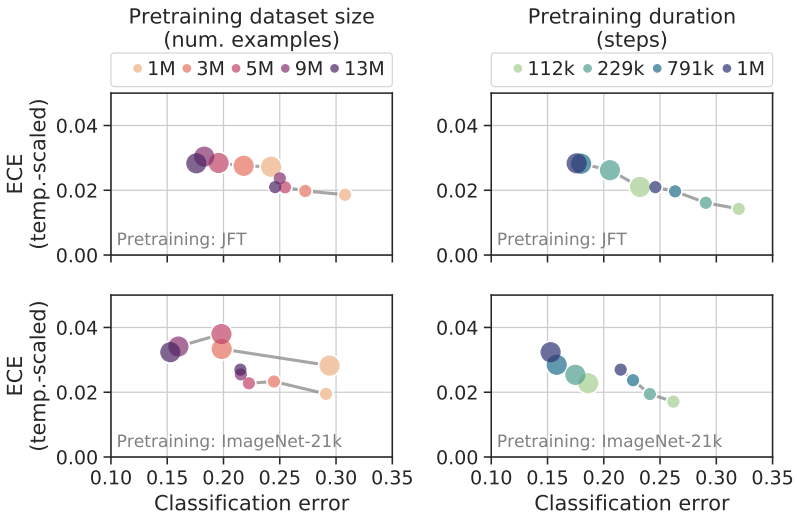

Figure 10: Effect of pretraining dataset size and duration on calibration. Larger dots indicate BiT-R101x3, smaller dots indicate BiT-R50x1. The pretraining datasets are subsampled from JFT-300 (top) or IMAGENET-21K (a larger variant of IMAGENET; bottom). Classification error is on IMAGENET after fine-tuning.

B.2 Correlation Between Calibration and Accuracy

Figures 4 and 6 show that, across a sufficiently large range of distribution shift, calibration error and classification error are correlated. Figure 11 illustrates this correlation for each model family across model variants and datasets.

In general, it is expected that calibration error and classification error are correlated to some degree due to noise in the model predictions, since adding random noise to the model confidence score would increase both calibration and classification error. Indeed, all model families show a strong positive correlation between calibration and classification error. However, there are consistent differences between model families, reflecting their intrinsic calibration properties. The relationship can be remarkably strong and lawful. For example, a simple power law of the form $y = ax^k$ (where x is classification error and y is ECE) provides a good fit for some model families (e.g. ResNeXt WSL; Figure 11). The parameters of the fit provide a quantitative description of the intrinsic calibration properties of a model family that goes beyond ECE on a specific dataset.

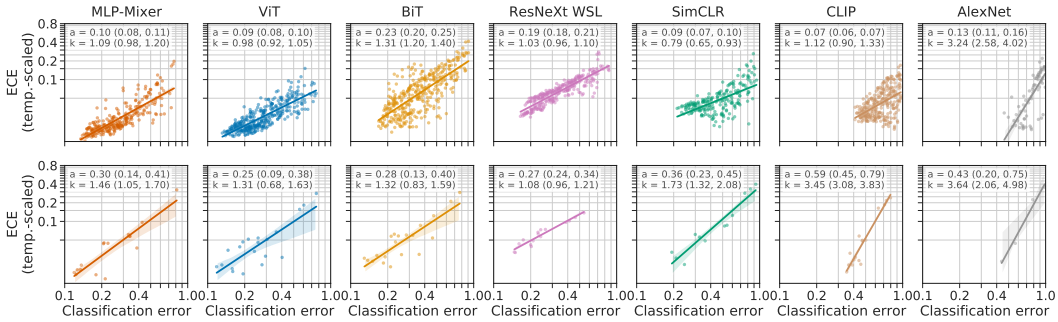

Figure 11: Correlation between ECE and classification error. Each dot represents a different combination of model variant and evaluation dataset (top: IMAGENET-C variants *without* averaging corruptions and severities; bottom: allo other datasets). Lines show power laws of the form $y = ax^k$ where x is classification error and y is ECE. Range in parentheses indicates the 95% confidence interval by bootstrap.

B.3 Contribution of Accuracy and Calibration to Decision Cost

In Section 4.4, we use a selective prediction task as a practical scenario in which we can quantify the relative impact of accuracy and calibration on the ultimate decision cost incurred by a model user. In this task, the user can either accept a model prediction and incur a misclassification cost if the prediction is wrong, or reject (abstain from) the prediction and incur an abstention cost (which is independent of whether the model prediction would have been correct). This decision is made based on the model's confidence. The total cost therefore depends on both the accuracy and the calibration of the model. A concrete example is a medical diagnosis task in which we can choose to use the model's diagnosis as-is, or refer the case to a human for review. Figure 12 shows cost planes for eight model pairs.

First, we compare models **from the same family** (Figure 12, a–e). In the blue regions, the relative cost for the model with higher accuracy (always the larger model) is lower (better); in red regions, the relative cost of the model with lower accuracy (always the smaller model) is lower (better). For most models and for most practical cost settings, the higher accuracy model is preferred over the better calibrated model. In the ViT family, for example, the bigger model has 0.076 lower classification error (0.194 vs. 0.118) and 0.007 higher ECE (0.017 vs. 0.01). For these models, the cost analysis shows that the difference in classification error outweighs the difference in ECE across all tested misclassification costs and abstention rates.

Next, we compare models **pretrained for a different number of steps** (same models as used in Figure 10) and provide the results in Figure 12, f–g. Again, the models with lower classification error (e.g. for R101x3, 0.176 vs 0.232 in favor of the longer-trained model) reach a lower total cost than the models with lower ECE (e.g. for R101x3, 0.019 vs 0.028 in favor of the shorter-trained model).

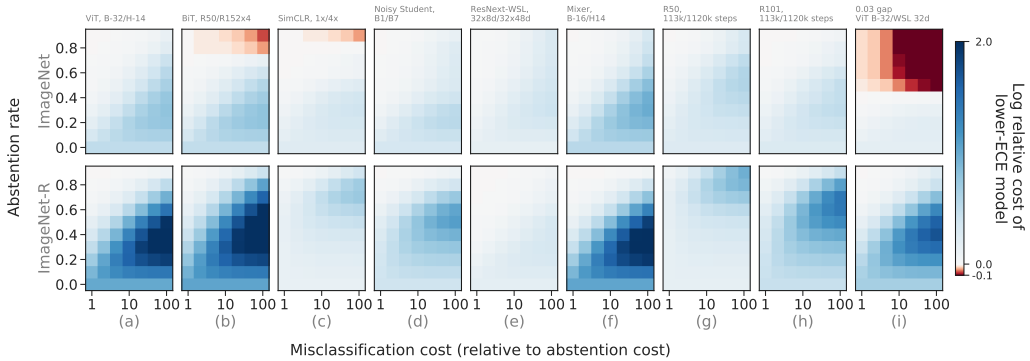

Figure 12: Relative impact of accuracy and calibration in a selective prediction scenario. Each heatmap compares two models and shows the relative cost of the better-calibrated (lower-ECE) model with respect to the other model. Total cost is computed as a linear combination of the misclassification cost and abstention cost at a given cost ratio (x-axis) and abstention rate (y-axis). Compared model pairs are indicated above each column. The top row shows IMAGENET, the bottom row IMAGENET-R. In most scenarios, the higher-accuracy model is preferred over the better-calibrated model (blue regions). Only in a few cases and at very high abstention rates does the difference in calibration outweigh the difference in accuracy (red regions). In other words, for practical abstention rates and across a wide range of abstention costs, the accuracy advantage outweighs the calibration advantage.

Finally, we compare models which attain **similar classification error and ECE difference** (Figure 12, h). In particular, we compare ViT-B/32 and ResNeXt-WSL 32d models. The latter model has 0.03 lower (better) classification error while the ViT model has 0.03 lower (better) ECE. Again, for most practical cost settings, the model with better accuracy has lower cost than (is preferred over) the model with better ECE.

C Sampling Bias for ℓ_2-ECE

In Section 5, we hinted at the fact that the bias of the ECE estimator depends on the model accuracy. Here, we expand on Equation 5 and fully derive the bias for a variant of the ECE score, when we take the squared instead of the absolute differences in each bucket for tractability.

Lemma 1 *Define the random variables $A = Y \in \arg\max f(X)$ and $C = \max f(X)$, consider the squared ECE metric*

$$ECE_2 = \sum_{i=1}^{m} P(X \in B_i)(accuracy(B_i) - confidence(B_i))^2,$$

where the B_i represent the m disjoint buckets. If we estimate the the per-bin statistics using their sample means, the statistical bias is equal to

$$\sum_i \frac{1}{n}\mathbb{V}[C - A \mid X \in B_i] = \frac{1}{n}\sum_i (\alpha_i(1-\alpha_i)(1-\delta_i) + \mathbb{V}[C \mid X \in B_i]),$$

where α_i is the accuracy in bucket B_i and δ_i is the expected difference in the confidences of the correct and incorrect predictions.

We assume that the buckets are fixed, s.t., there are n_i points in bucket B_i, and a total of $n = \sum_i n_i$ points (we will take the expectation over n_i). We introduce two random variables — the model confidence by $C = \max f(X)$ and the corresponding true/false indicator by $A = [\![Y \in \arg\max f(X)]\!]$. For each realization (x_j, y_j) we denote by c_j and a_j the corresponding values. We further define for each bucket B_i

- $\alpha_i = \mathbb{E}[A \mid X \in B_i]$, the accuracy in bucket B_i.
- $\gamma_i = \mathbb{E}[C \mid X \in B_i]$, the expected confidence in bucket B_i.

- $\delta_i = \mathbb{E}[C \mid A = 1, X \in B_i] - \mathbb{E}[C \mid A = 0, X \in B_i]$, the confidence difference for the correct and wrong predictions.
- $\bar{c}_i = \sum_{c \in B_i} c/n_i$, the sample average confidence in bucket B_i.
- $\bar{a}_i = \sum_{a \in B_i} a/n_i$, the sample average accuracy in bucket B_i.

We consider the squared ECE ℓ_2 loss, which after bucketing is equal to

$S^2 = \sum_i P(X \in B_i)(\underbrace{\mathbb{E}[C \mid X \in B_i]}_{\gamma_i} - \underbrace{\mathbb{E}[A \mid X \in B_i]}_{\alpha_i})^2$, and the corresponding sample estimate is

$\hat{S}^2 = \sum_i \frac{n_i}{n}(\frac{1}{n_i}\sum_{i \in B_i} c_i - \frac{1}{n_i}\sum_{j \in B_i} a_j)^2.$

The goal is to understand the bias $\hat{S}^2 - S^2$. Note that

$$\mathbb{E}[(\bar{c}_i - \bar{a}_i)^2 \mid n_i] = (\gamma_i - \alpha_i)^2 + \mathbb{V}[\bar{c}_i - \bar{a}_i \mid n_i] = (\gamma_i - \alpha_i)^2 + \frac{1}{n_i}\mathbb{V}[C - A \mid X \in B_i].$$

We further have

$$\mathbb{V}[C - A \mid X \in B_i] = \mathbb{V}[C \mid X \in B_i] + \mathbb{V}[A \mid X \in B_i] - 2\text{Cov}[C, A \mid X \in B_i].$$

Hence, we have that

$$\begin{aligned}
\mathbb{E}[\hat{S}^2] &= \mathbb{E}[\sum_i \frac{n_i}{n}(\frac{1}{n_i}\sum_{i \in B_i} c_i - \frac{1}{n_i}\sum_{i \in B_i} a_i)^2] \\
&= \mathbb{E}[\mathbb{E}[\sum_i \frac{n_i}{n}(\frac{1}{n_i}\sum_{i \in B_i} c_i - \frac{1}{n_i}\sum_{i \in B_i} a_i)^2 \mid n_i]] \\
&= \mathbb{E}[\sum_i \frac{n_i}{n}\big((\gamma_i - \alpha_i)^2 + \frac{1}{n_i}\mathbb{V}[C - A \mid X \in B_i]\big)] \\
&= \sum_i (\gamma_i - \alpha_i)^2 \underbrace{\mathbb{E}[\frac{n_i}{n}]}_{P(X \in B_i)} + \sum_i \frac{1}{n}\mathbb{V}[C - A \mid X \in B_i] \\
&= S^2 + \underbrace{\frac{1}{n}\mathbb{V}[C - A \mid X \in B_i]}_{\text{bias}}.
\end{aligned}$$

We can decompose the covariance as follows (see this MathOverflow answer), using the fact that A is binary:

$$\text{Cov}[C, A \mid X \in G_i] = \mathbb{V}[\alpha_i]\delta_i,$$

Here δ_i is defined as $\mathbb{E}[C \mid A = 1, X \in B_i] - \mathbb{E}[C \mid A = 0, X \in B_i]$. Now the total bias can be written as

$$\text{bias} = \frac{1}{n}\sum_i \alpha_i(1 - \alpha_i)(1 - 2\delta_i) + \mathbb{V}[C \mid X \in B_i].$$

Note that $\partial_{\alpha_i}\text{bias} = (1 - 2\alpha_i)(1 - 2\delta_i)$, which is negative when $\alpha_j > 1/2$, if we assume we have enough bins so that $\delta_j < 1/2$. Hence, for models that have at least 50% top-1 accuracy, increasing the accuracy reduces the bias.

D Model Confidence

An important aspect of the calibration of a model is its average *confidence*, i.e. the systematic bias of the model's scores to be too high (overconfident) or too low (underconfident) compared to the true accuracy. A large fraction of the miscalibration of modern neural networks is typically due to over- or underconfidence (Guo et al., 2017). In this section, we argue that over- and underconfidence are not just a source of miscalibration, but also a confounder that obscures the intrinsic calibration properties of models and makes it harder to compare across model families.

Quantifying confidence. Predictions of overconfident models tend to be overly "peaky" (low entropy), such that an increase in temperature (positive temperature factor) would be necessary to make them optimally confident, and vice versa for underconfident models. We can therefore quantify confidence in terms of the temperature scaling factor by which the logits of the unscaled model would have to be multiplied to provide optimal confidence.

The optimal confidence depends on the model *and* the dataset. Ideally, a model would be optimally confident across all distribution shifts, indicating that its confidence is well calibrated to the difficulty of the data. In practice, most models are slightly overconfident in-distribution, and tend to become more overconfident as data moves further from the training distribution (Figure 13, bottom row).

Models can show the opposite trend if they are *under*confident in-distribution. As an example, we include the EfficientNet-NoisyStudent family in Figure 13. These models tend to be underconfident (optimal temperature factor < 1; Figure 13, bottom right). Underconfident models may paradoxically show *improved* calibration under distribution shift (lower ECE for higher corruption severities), because their underconfidence balances out the general tendency towards overconfidence on OOD data. However, such underconfident models are not better calibrated in general—they are simply biased towards a high level of distribution shift, and are calibrated worse at weak or no distribution shift. A well-calibrated model should have optimal confidence both in- and out-of-distribution.

Normalizing confidence. The example of EfficientNet-NoisyStudent illustrates how confidence bias can confound trends in model calibration. This counfounder can be removed by *temperature scaling* (Guo et al., 2017), i.e. by rescaling model logits to optimize the likelihood on a held-out

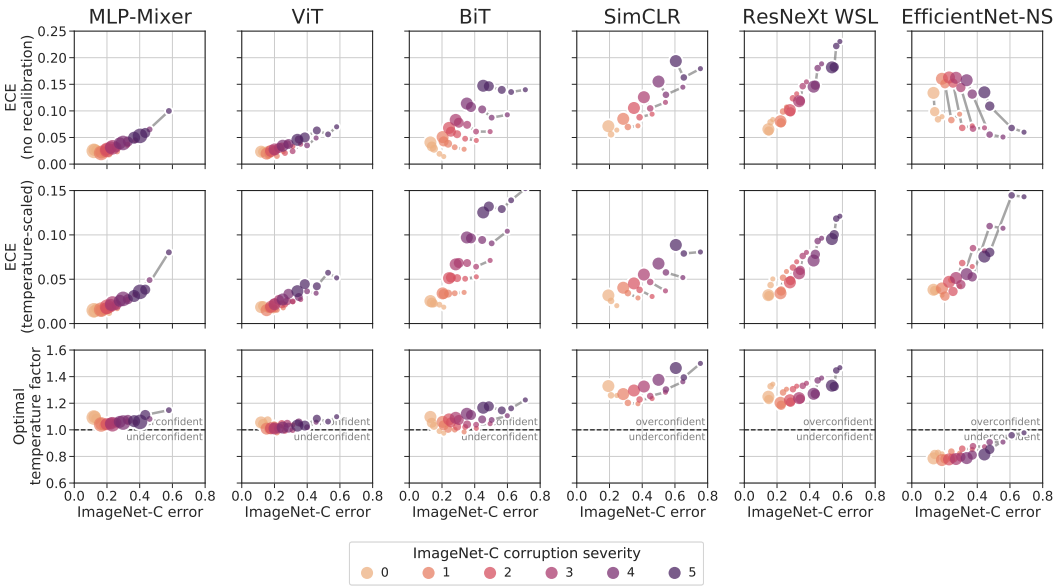

Figure 13: Related to Figure 4. Calibration and accuracy on IMAGENET-C. Here, the model confidence is shown in the third row (top two rows are identical to Figure 4). Model confidence is quantified in terms of the temperature scaling factor by which the logits of the unscaled model would have to be multiplied to provide optimal confidence for a given dataset. Values above 1 mean that the unscaled model is overconfident on the given dataset, and below 1, that the unscaled model is underconfident.

part of the in-distribution dataset (IMAGENET in our case). By removing differences in confidence bias between models, temperature scaling reveals a consistent trend for higher calibration error under distribution shift for all models, including EfficientNet-NoisyStudent (Figure 13, second row). Temperature scaling also reveals consistent differences between model families and trends within families for in-distribution calibration (Figures 2 and 3). We therefore study calibration after temperature scaling, in addition to unscaled calibration error and other calibration metrics (Appendix F), throughout this work. The benefit of temperature scaling for understanding model calibration is separate from its well-established benefit in reducing calibration error (Guo et al., 2017).

Label smoothing. One method to directly influence the confidence of a model during training is *label smoothing* (Szegedy et al., 2016). In label smoothing, uniformly distributed probability mass is added to the training targets. This decreases the implied confidence of the targets and thus of the model trained on these targets, which can reduce overfitting and improve accuracy.

Label smoothing has been reported to improve calibration (Müller et al., 2019). Here, we argue that label smoothing creates artificially underconfident models and may therefore improve calibration for a specific amount of distribution shift, but does *not* generally improve the intrinsic calibration properties of a model (i.e. its overall calibration across distribution shifts and datasets).

Figure 14 shows the ECE before and after temperature scaling of models trained with different amounts of label smoothing on IMAGENET and evaluated on IMAGENET-C. Before temperature scaling (Figure 14, left), the best-calibrated models (lowest ECE) are those trained with label smoothing. Depending on the amount of distribution shift (corruption severity), a different amount of label smoothing is necessary to optimize calibration. After temperature scaling on a held-out part of the IMAGENET validation set (Figure 14, center), it becomes clear that training without label smoothing actually results in the lowest ECE across all IMAGENET-C severities. The optimal temperature factor (Figure 14, right) reveals that label smoothing simply biases the model confidence, like temperature scaling, but without targeted optimization. These data suggest that, from a calibration perspective, models should be trained without label smoothing and then recalibrated by temperature scaling *post hoc*.

Label smoothing may explain the anomaly observed for EfficientNet-NoisyStudent under distribution shift (Figure 13, far right). In contrast to all other model families we consider, EfficientNet-NS shows strong *under*confidence before temperature scaling (Appendix D); it is also the only model family trained with label smoothing.

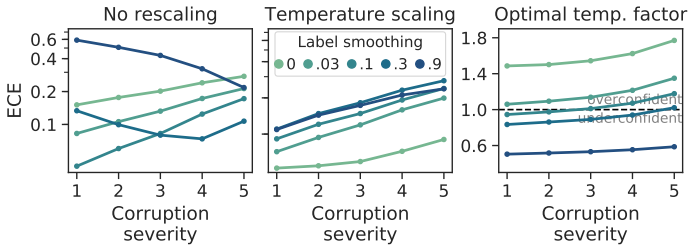

Figure 14: Effect of label smoothing on calibration. EfficientNet-B4 models were trained with the indicated label smoothing on IMAGENET and evaluated on IMAGENET-C. Before rescaling, different amounts of non-zero label smoothing appear to yield the best calibration, depending on distribution shift (left). After temperature scaling, it becomes clear that training without label smoothing is best (center). Label smoothing reduces confidence (right). ResNet architectures show similar behavior.

E ECE Variants

As discussed in Sections 2 and 5, while ECE is a well-defined quantity, estimating it requires binning and thus a choice of binning scheme and bin size. In addition, variants of ECE such as root mean squared calibration error (RMSCE, Nixon et al. 2019) exist. In RMSCE, the difference between accuracy and confidence in each bin is ℓ_2-normalized, in contrast to the ℓ_1-normalization of standard ECE. This causes larger errors to be upweighted in RMSCE. Further ECE variants consider all classes, instead of just the class with the highest predicted probability (top-label), or consider classes independently and report an average of class-wise ECEs. Different ECE variants may rank models differently (Nixon et al., 2019), which could lead to the conclusion that ECE estimators are fundamentally inconsistent. However, we find that such inconsistencies in model rank are resolved by considering ECE and classification error jointly (Figures 15 to 17). While ranks between models may change across ECE variants, these models differ in classification error, such that it is always clear which model is Pareto-optimal in terms of ECE and classification error. For example, for IMAGENETV2 in Figure 16, the ranking of BiT models (orange squares) changes slightly between some of the ECE variants. However, the models differ so much in classification error that the differences in ECE between metric variants are likely irrelevant (also see Appendix B.3, which shows that differences in classification error typically have a larger influence on decision cost than differences in ECE).

F Alternative Calibration Metrics

To confirm that our findings are not dependent on our choice of Expected Calibration Error as our main calibration metric, we provide results for two alternative calibration metrics: negative log-likelihood (NLL) and Brier score (Brier, 1950). Figure 8 in the main text covers IMAGENET, IMAGENETV2, IMAGENET-R, and IMAGENET-A. Results for IMAGENET-C are provided in Figure 18

Furthermore, we provide reliability diagrams (DeGroot & Fienberg, 1983) on IMAGENET for all models, both before (Figure 19) and after (Figure 20) temperature scaling. These diagrams visualize model calibration across the whole confidence range, rather than summarizing calibration into a scalar value.

Figures for Appendix E and Appendix F are on the following pages.

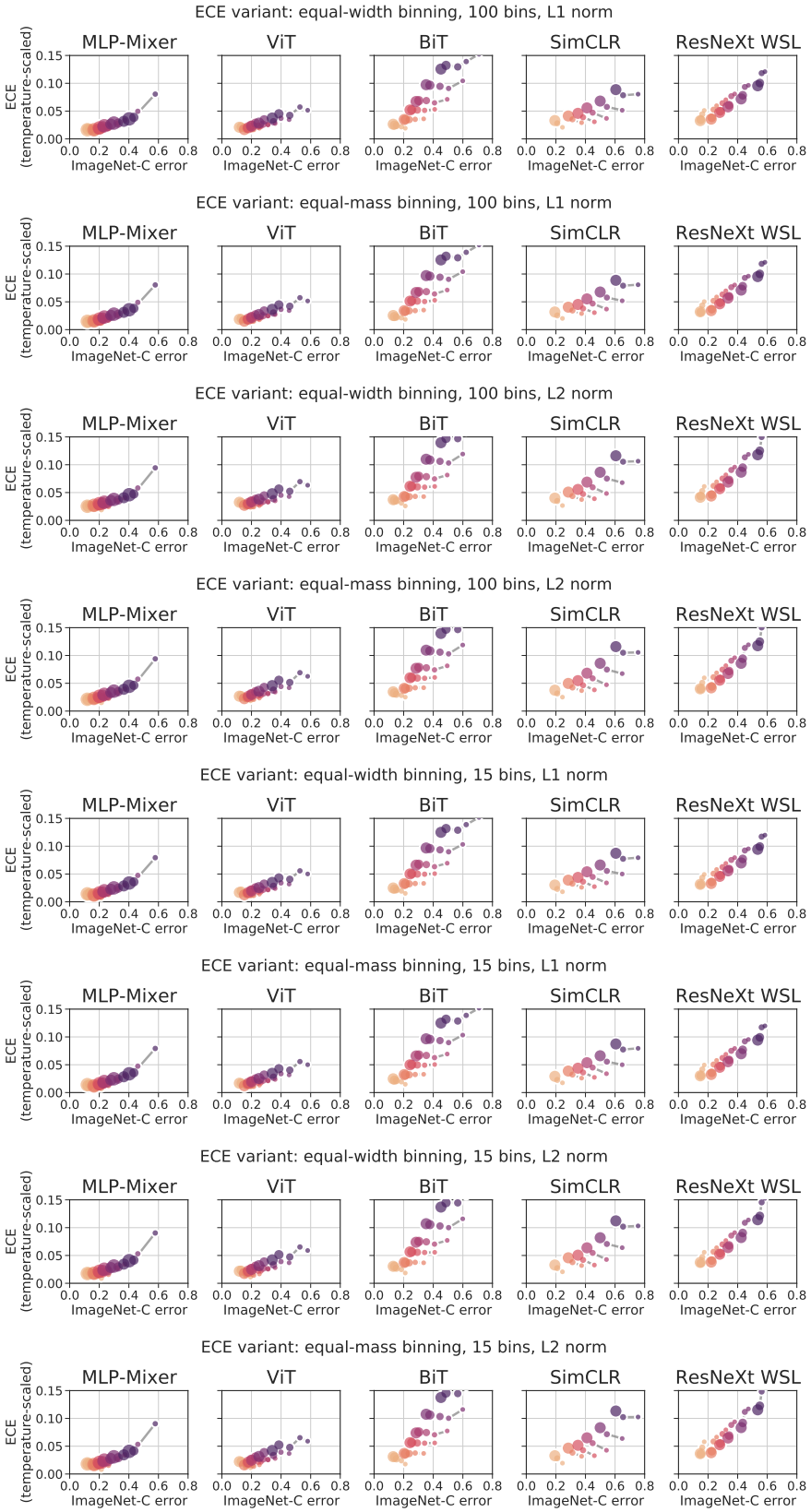

Figure 15: Related to Figure 4. Each row shows the calibration and accuracy on IMAGENET-C as in Figure 4, bottom row (i.e. after temperature scaling), but for different ECE variants. The variant is indicated in the title of each row. While absolute values can differ between variants, relative relationships between models are robust to the metric variant.

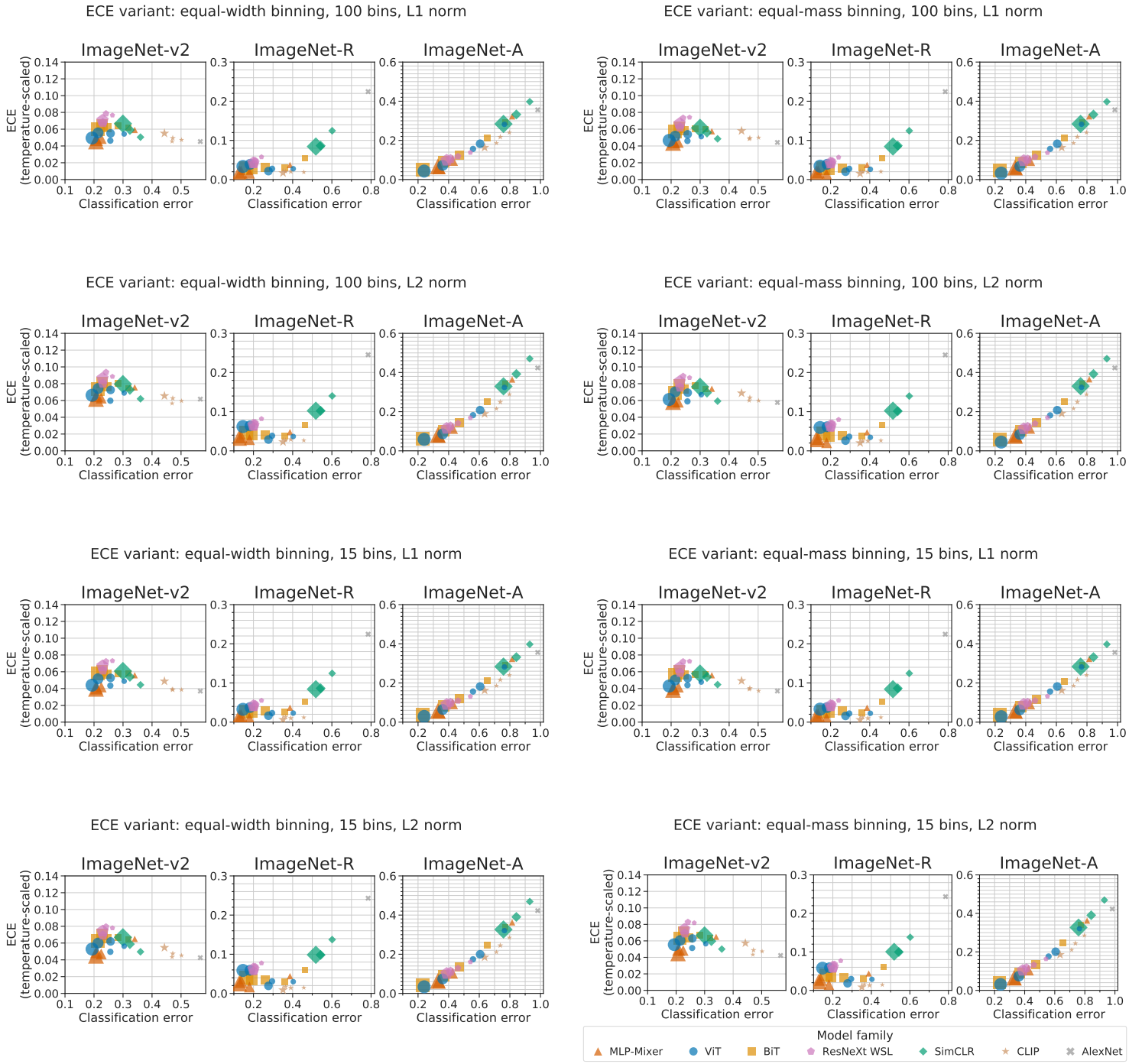

Figure 16: Related to Figure 6. Calibration and accuracy on OOD datasets as in Figure 6, bottom row (i.e. after temperature scaling), but for different ECE variants. The variant is indicated in the title of each set of plots. While absolute values can differ between variants, relative relationships between models are robust to the metric variant.

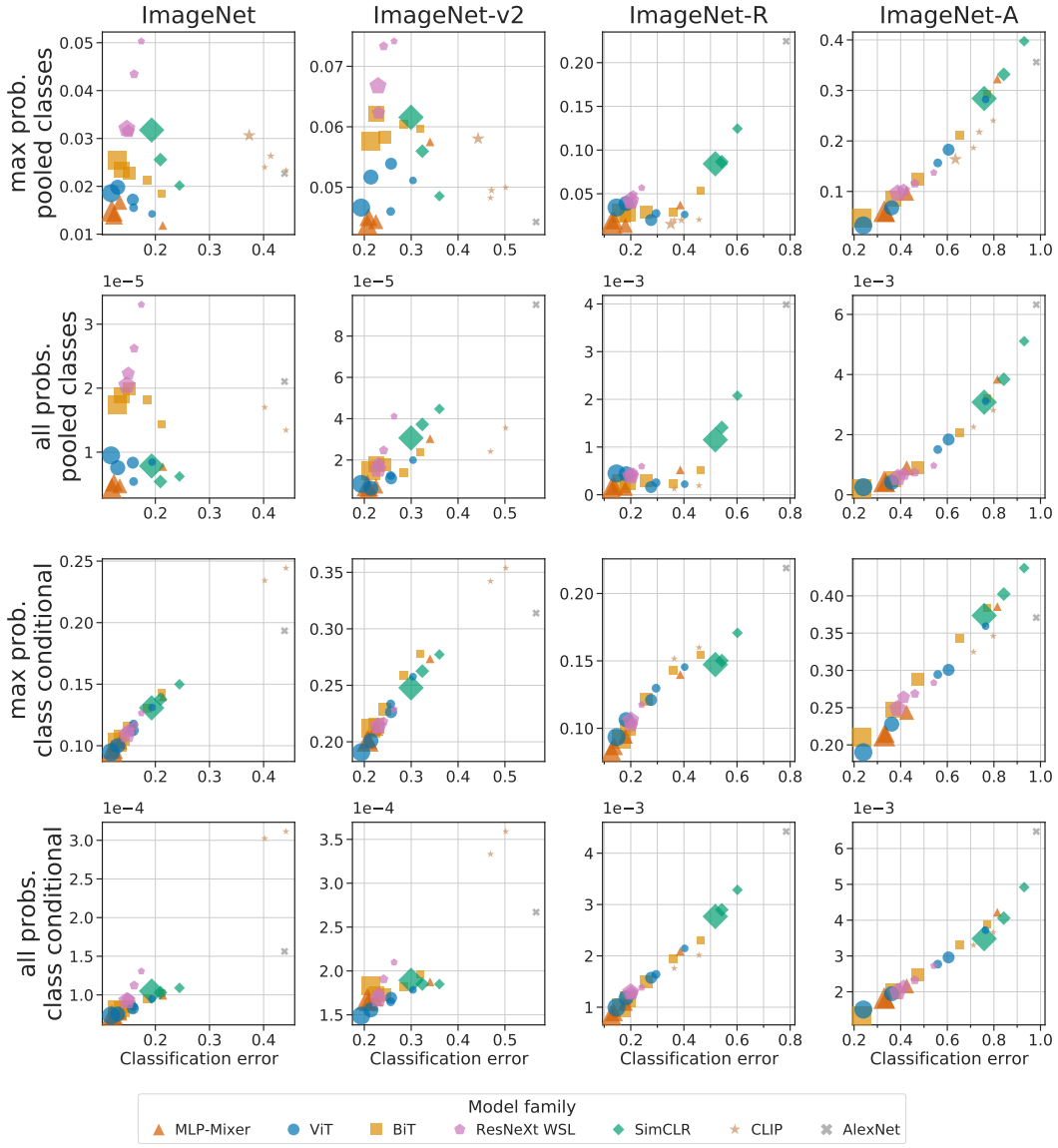

Figure 17: Further ECE variants (after temperature scaling). The top row shows the variant used in the main paper, which considers only the maximum predicted probability ("top-label calibration") and pools across classes. The remaining rows show other variants as discussed in Nixon et al. (2019). L1-normalization and adaptive binning was used in all cases (100 bins for pooled-class metrics; 15 bins for class-conditional metrics). Although the specific rankings between models depend on the ECE variant (Nixon et al., 2019), our main conclusions hold for all variants. Specifically, the same model families tend to be Pareto-optimal across all ECE variants. Also, the relationship between ECE and accuracy is largely consistent across ECE variants.

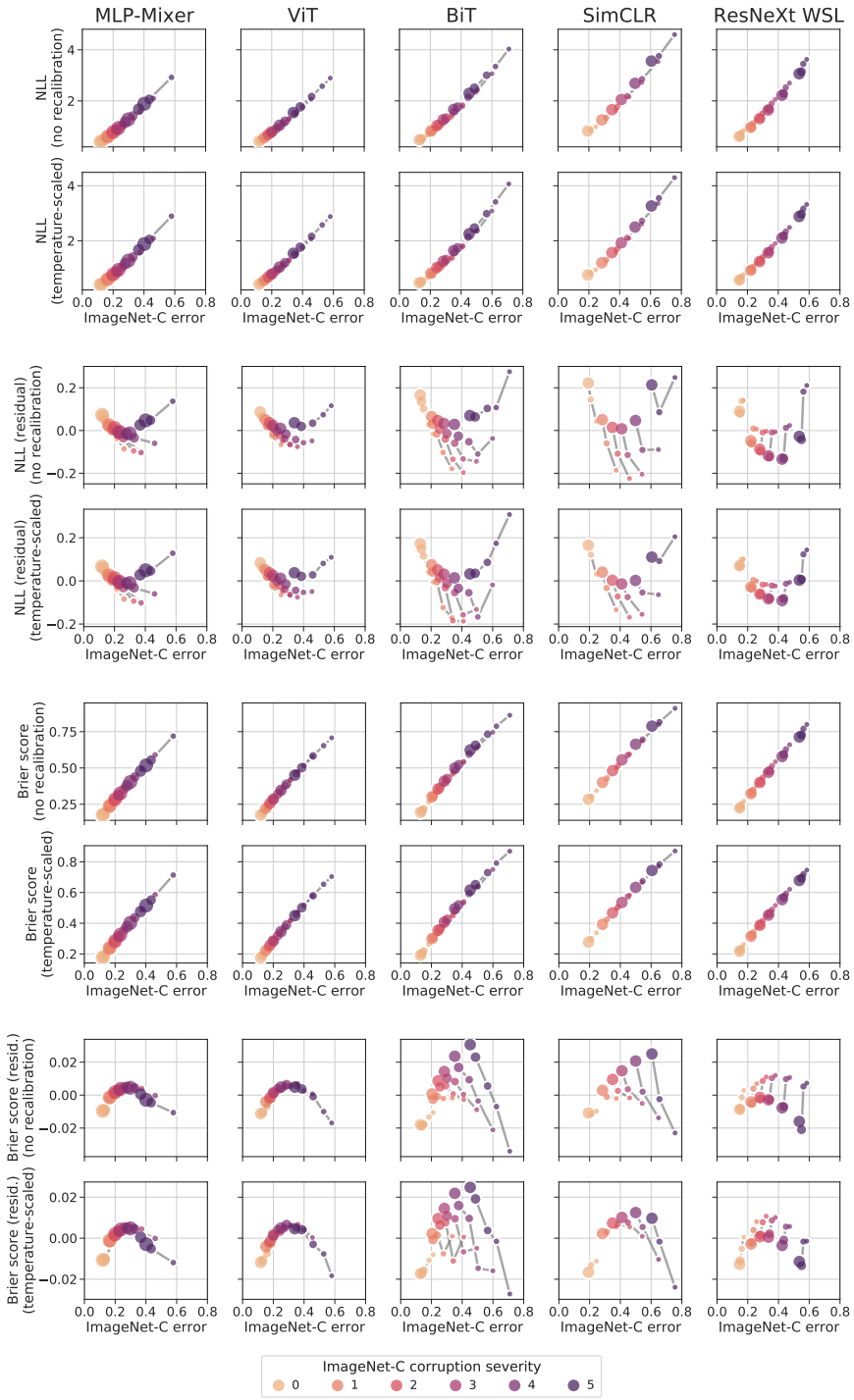

Figure 18: Alternative calibration metrics for IMAGENET-C: negative log-likelihood (NLL) and Brier score. Plotted as in Figure 4. Second and fourth rows show residuals as described in Figure 8.

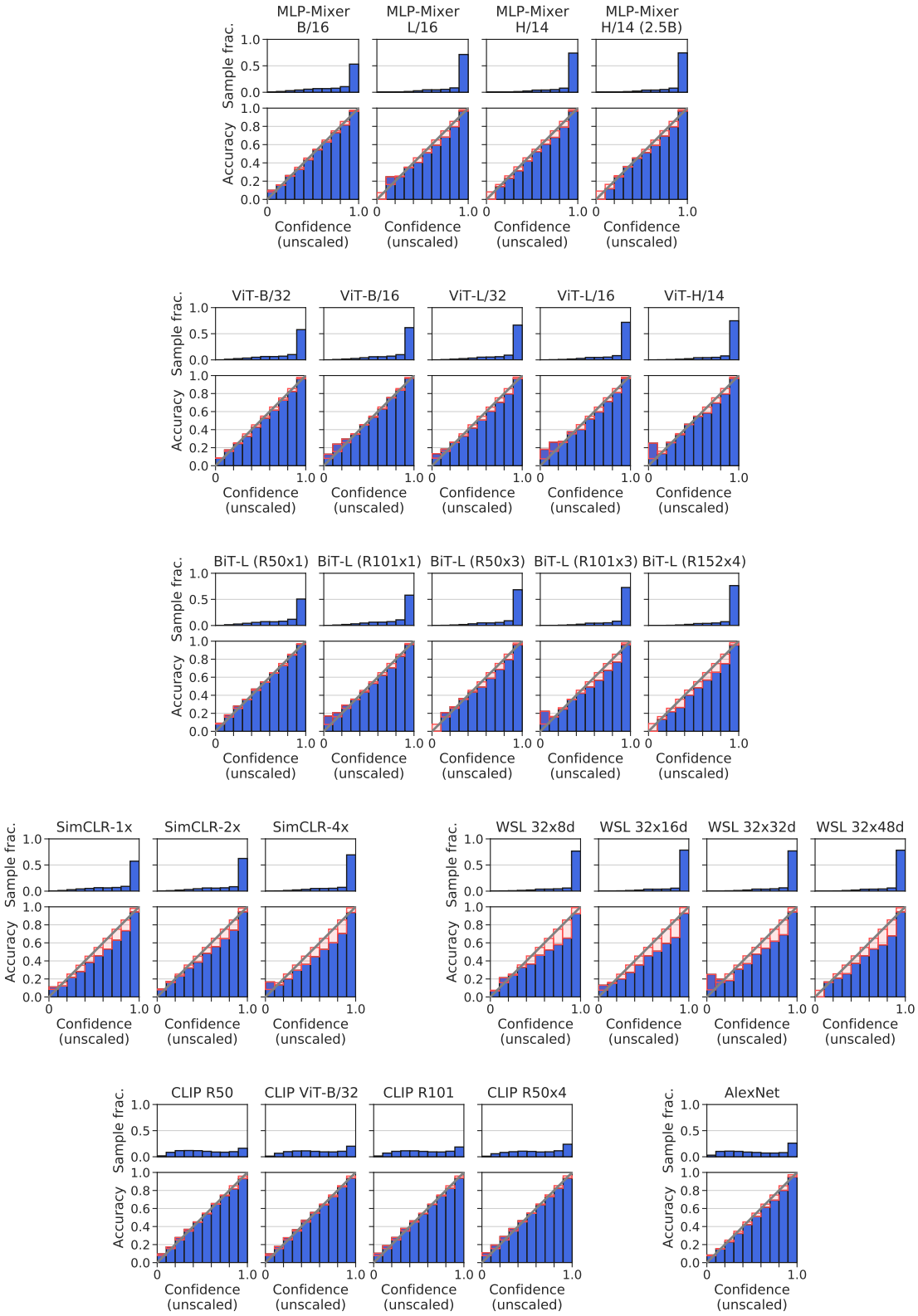

Figure 19: Reliability diagrams on IMAGENET for all models, before temperature scaling. Red boxes indicate the error compared to perfect calibration. The histogram at the top shows the distribution of confidence values for the dataset.

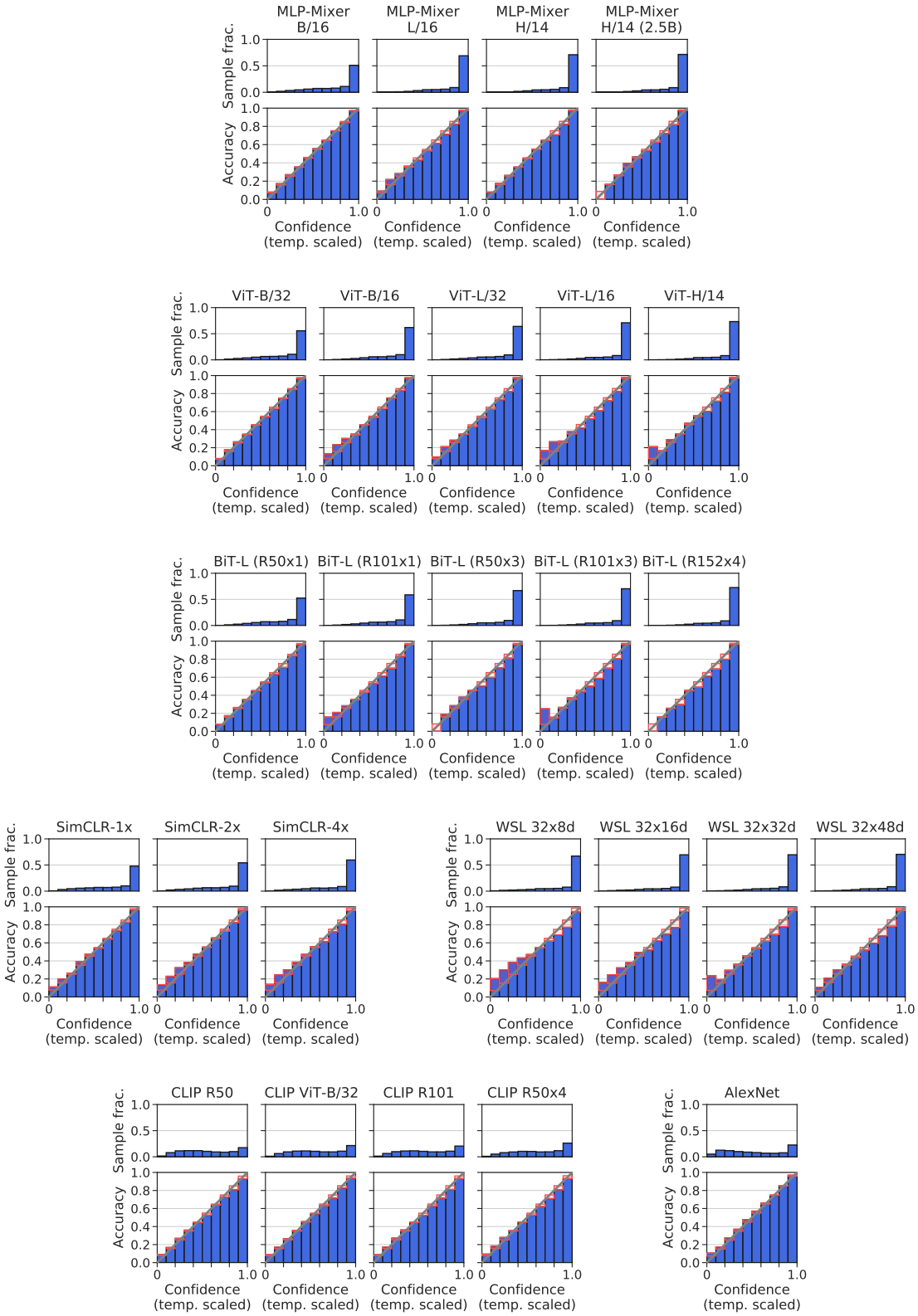

Figure 20: Reliability diagrams on IMAGENET for all models, after temperature scaling. Red boxes indicate the error compared to perfect calibration. The histogram at the top shows the distribution of confidence values for the dataset.