COMP6211I: Trustworthy Machine Learning

Uncertainty

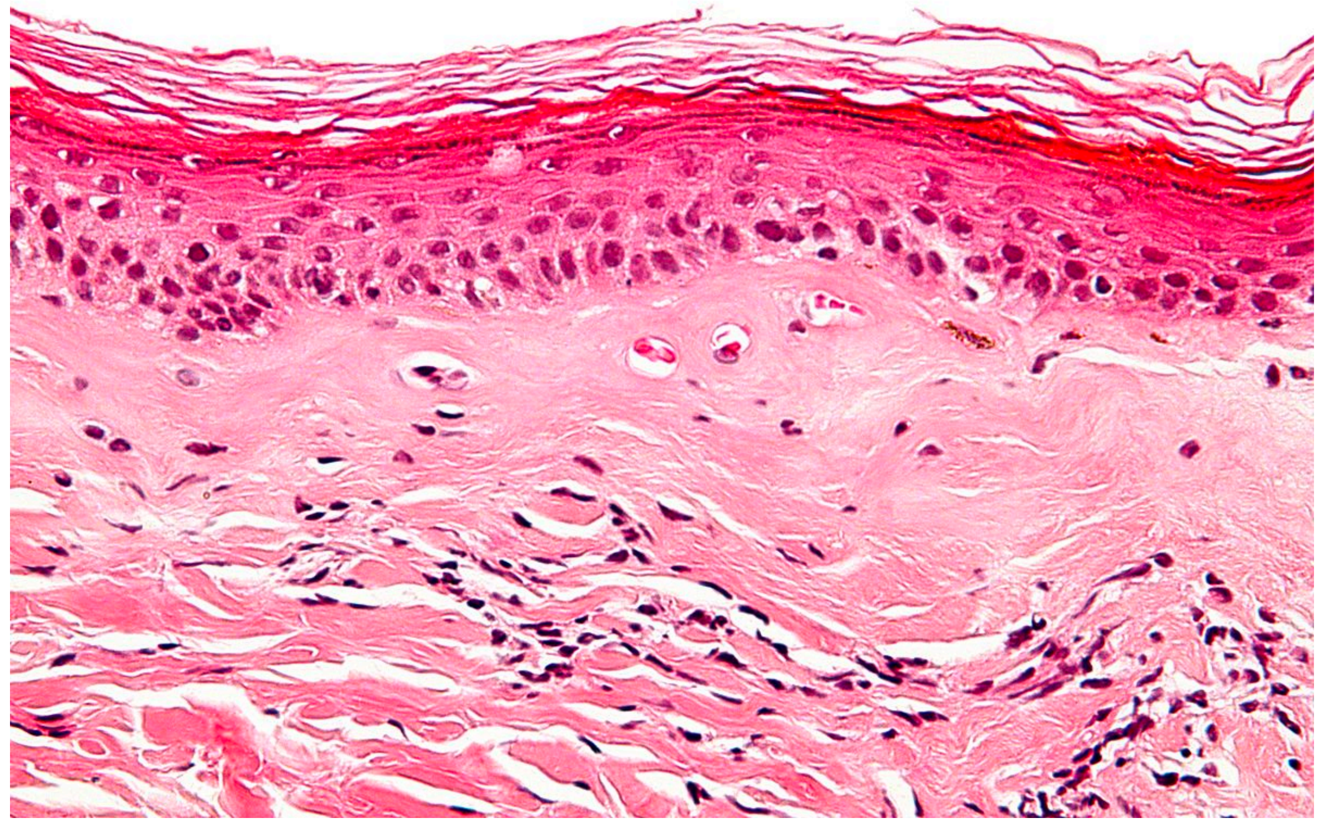
Minhao CHENG

What is uncertainty in machine learning

- We make observations using the sensors in the world
  - (e.g. camera) Based on the observations, we intend to make decisions
  - Given the same observations, the decision should be the same However,
  - The world changes, observations change, our sensors change, the output should not change!
  - We'd like to know how confident we can be about the decisions

Why calibration matters?

- Safety-critical applications.
- Example: Selective prediction in medical diagnosis

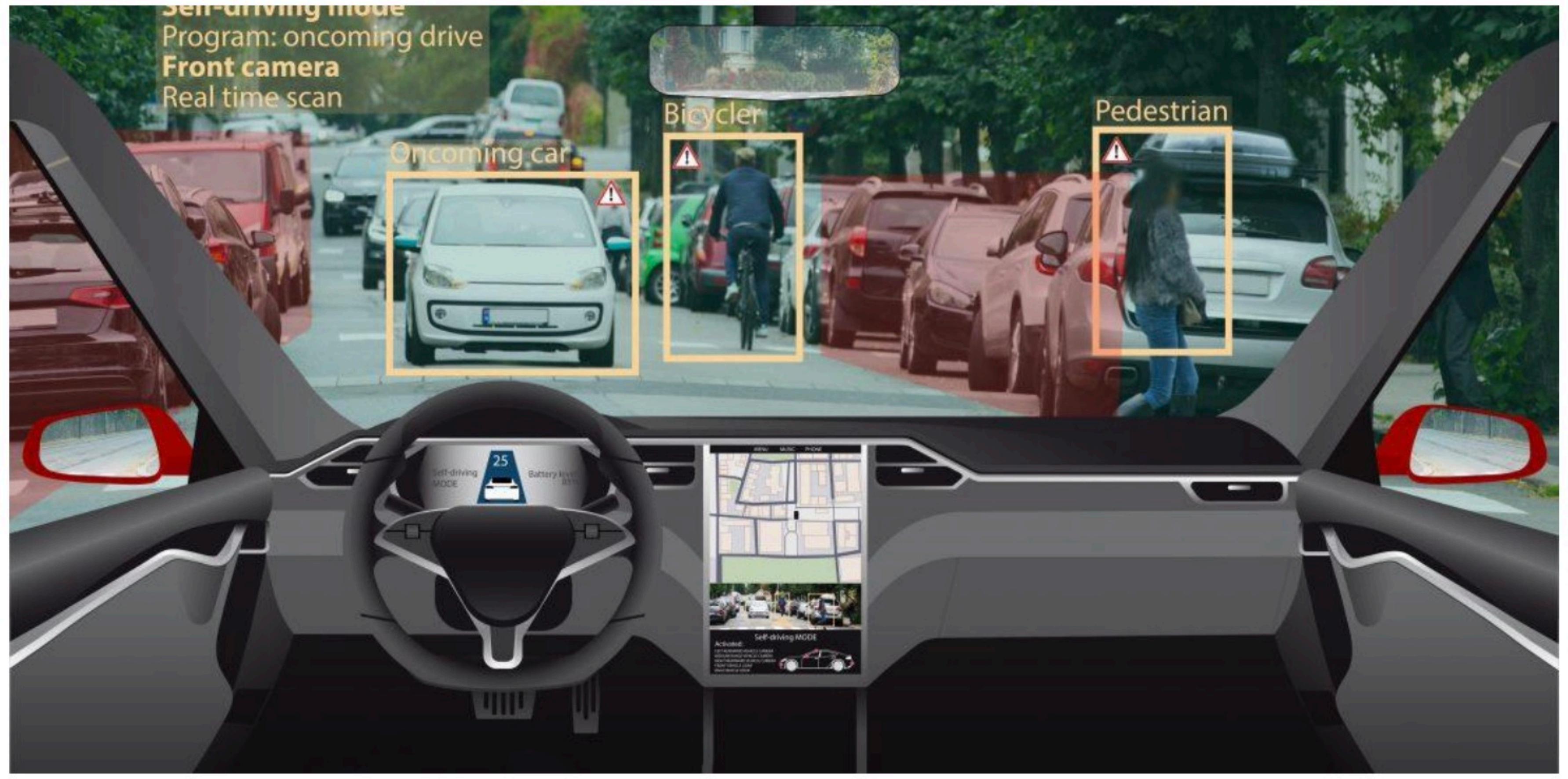


Model

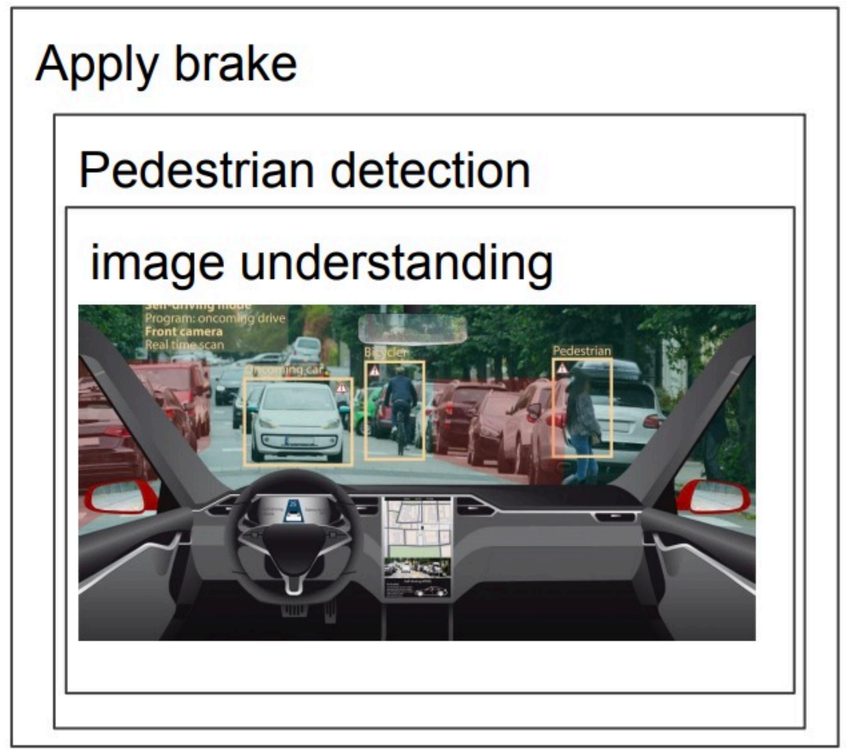
Accept model prediction? ← Not cancer: 0.99 → Refer to human specialist?

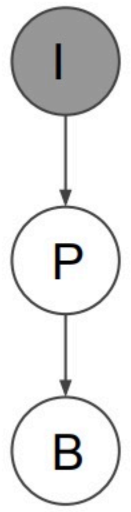
Why calibration matters?

Imagine you are designing the vision system for an autonomous vehicle



Applications that require reasoning in earlier stages





What is uncertainty in machine learning

- We build models for predictions, can we trust them? Are they certain?

Where uncertain comes from?

Remember the machine learning's objective: minimize the **expected loss**

$$\min_\theta \ \mathbb{E}_{\mathbf{x},y}[\ell(h(\mathbf{x};\theta),y)] = \int \ell(h(\mathbf{x};\theta),y)dp^*(\mathbf{x},y)$$
$$\approx \frac{1}{n}\sum_{i=1}^n \ell(h(\mathbf{x}_i;\theta),y_i) \quad (\mathbf{x}_i,y_i)\sim p^*(\mathbf{x},y)$$

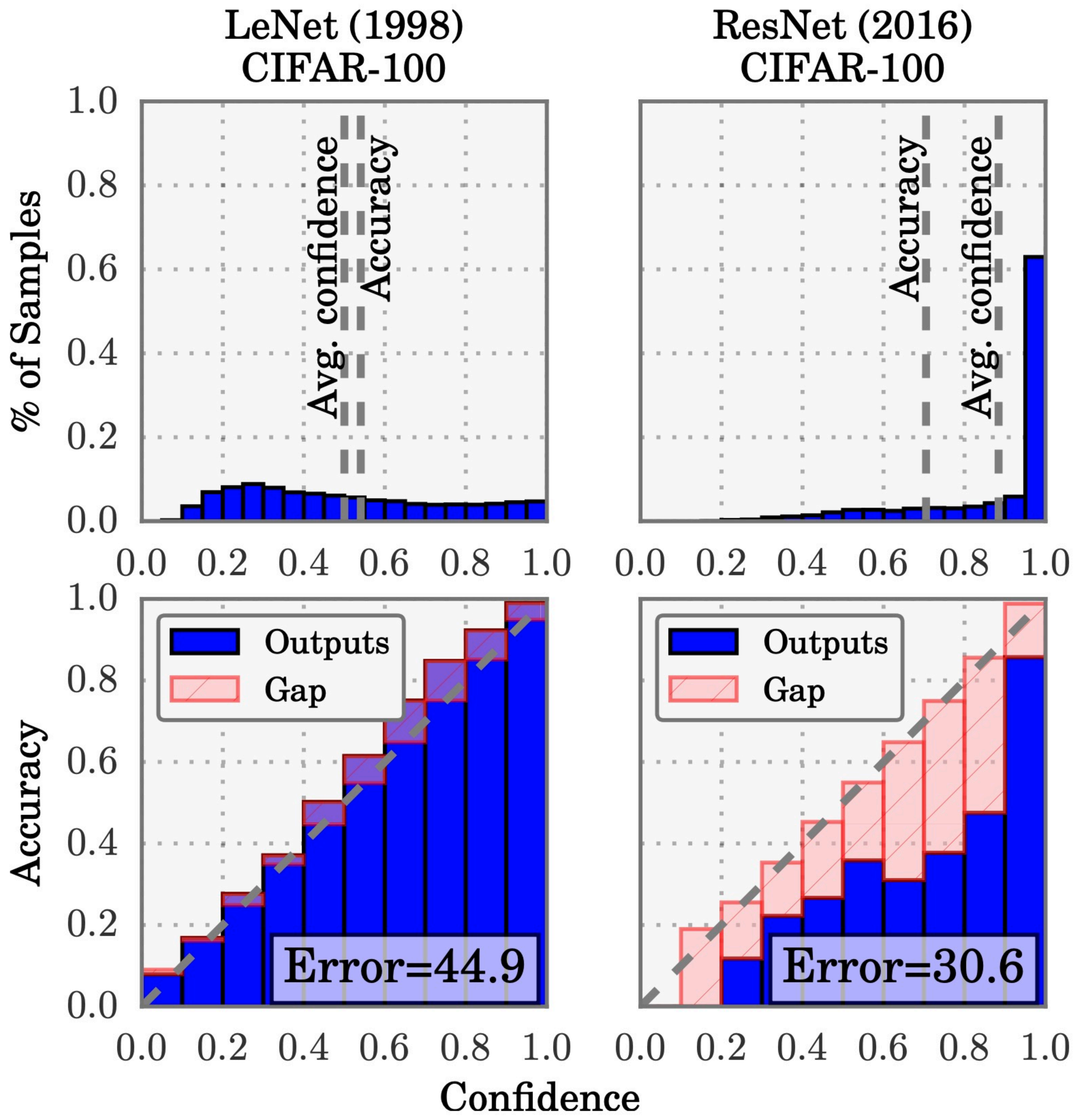
Uncertainty in data
(Aleatoric)

Uncertainty in the model
(**Epistemic**)

When the hypothesis function class is "simple" we can build generalization bound that underscore our confidence in average prediction

What is calibration

- Calibration error:
  - Difference between confidence (predicted probability) and accuracy



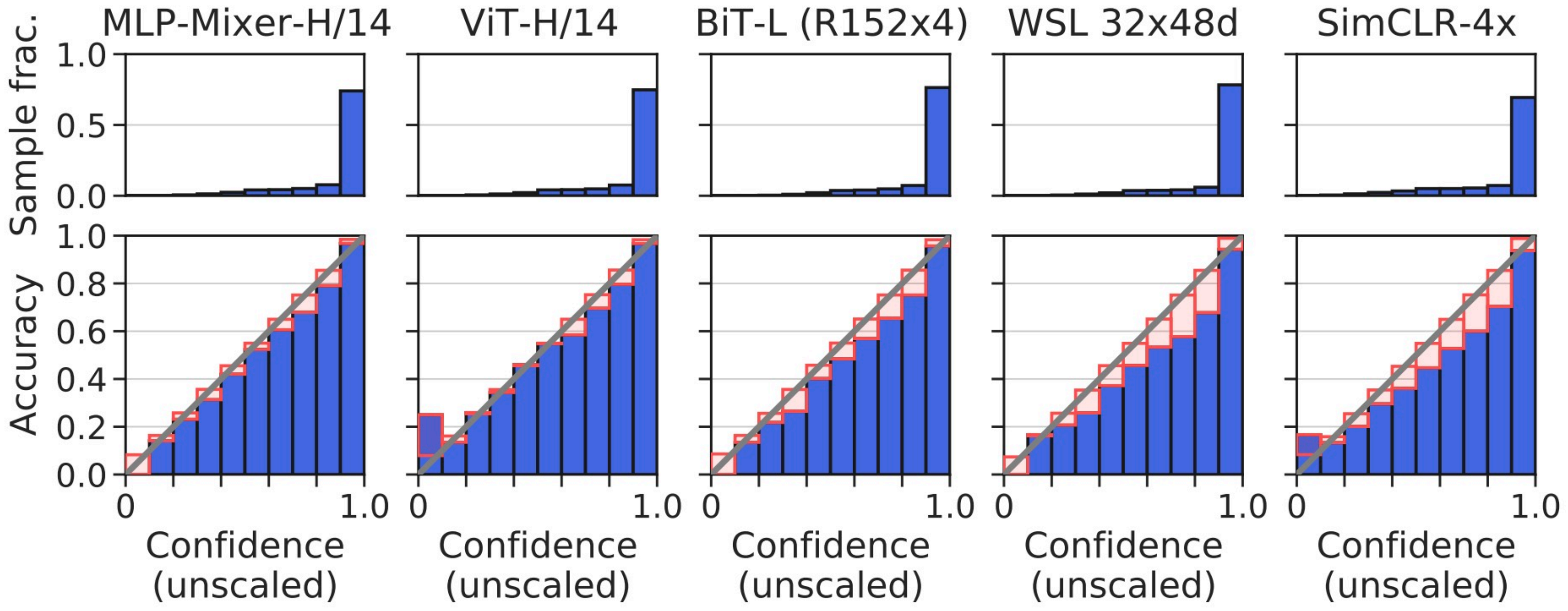
Calibration

- Measure degree of miscalibration: Expected Calibration Error (ECE)

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

- Break it into bins based on top predicted probability
  - $\text{accuracy}(B_i) = \frac{1}{|B_i|}\sum_{j \in B_i}[y_j \in \text{arg}\max f(x_j)]$ $\quad$ $\text{confidence}(B_i) = \frac{1}{|B_i|}\sum_{j \in B_i}\max f(x_j)$

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



Calibration

- The model is calibrated if

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

- A more practical condition is

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

- Measure degree of miscalibration: Expected Calibration Error (ECE)

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

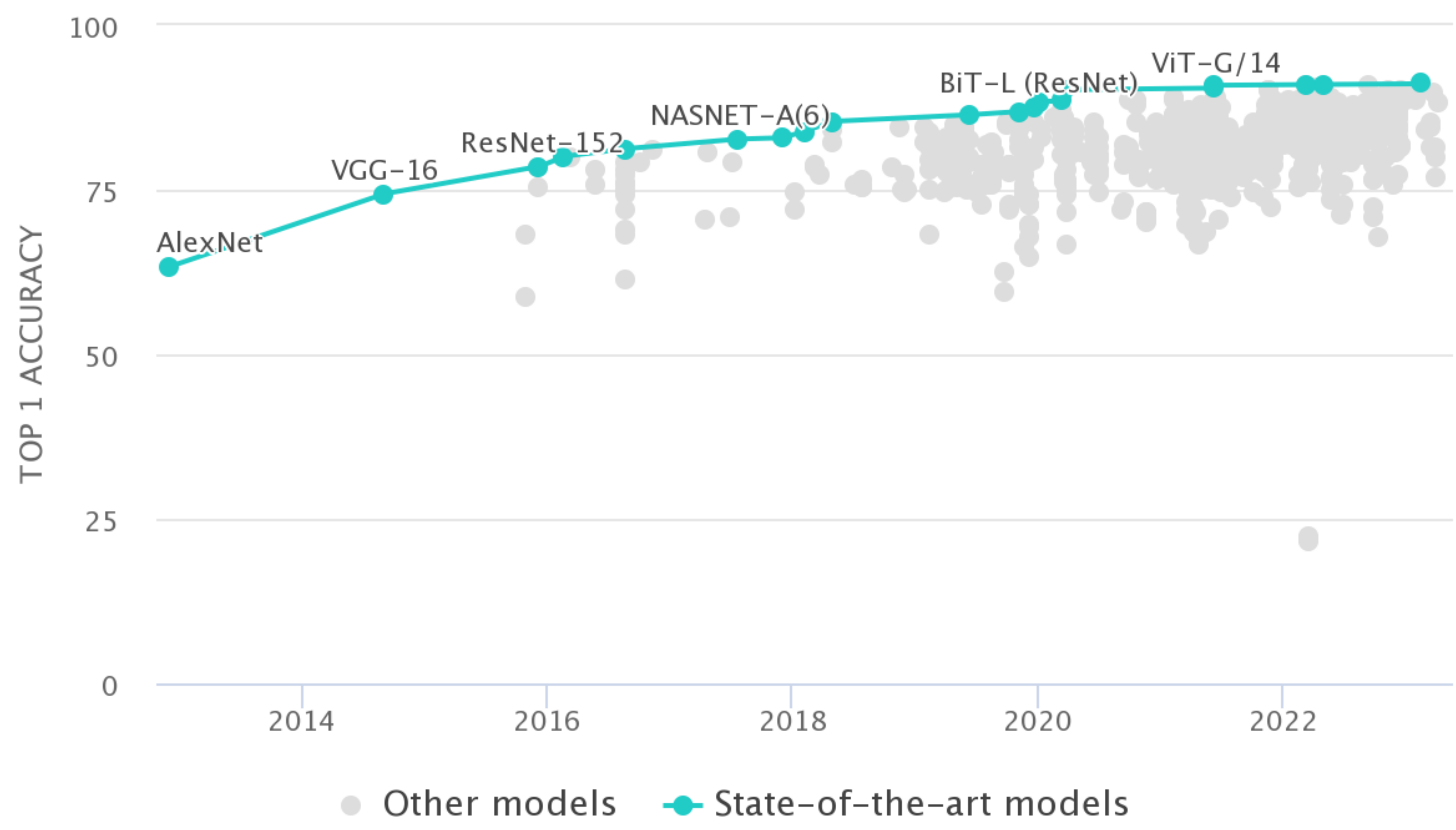
Calibration

Temperature scaling

- $$\hat{q}_i = \max_k \ \sigma_{\mathrm{SM}}(\mathbf{z}_i/T)^{(k)}.$$
    - T->0 , collapses to a point mass
    - T->1, recover the original probability
    - T->∞, approach to 1/K
- T is optimized with respect to NLL on the validation set

Recent developments

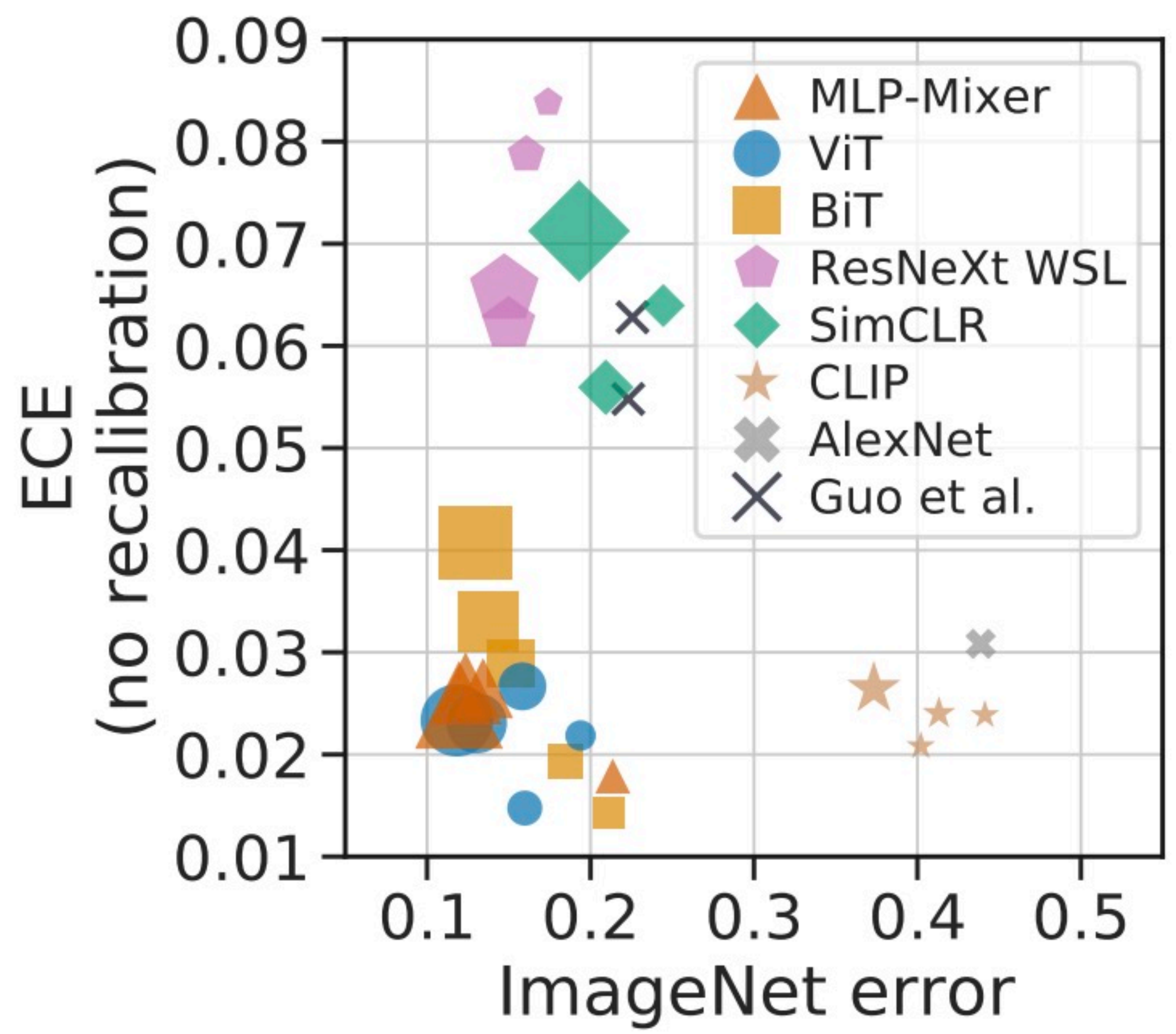
- Large-scale preparing
  - Big transfer (BiT)
- Weakly supervised pretraining
  - ResNext-WSL
- Unsupervised pretraining
  - SimCLR
- Non-convolutional architectures
  - Vision Transformer (ViT)
  - MLP-Mixer



In-distribution calibration

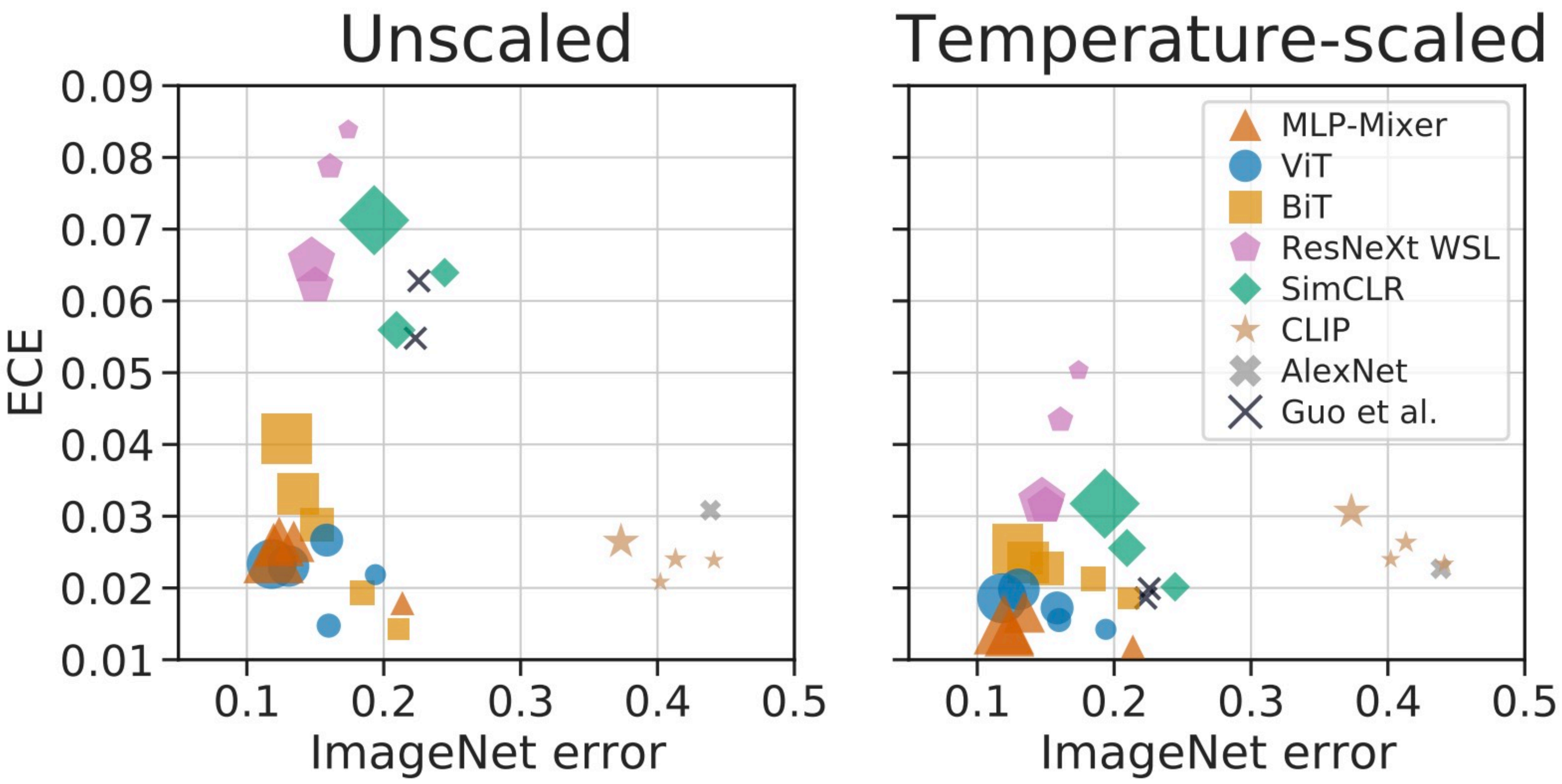
- Estimating calibration:
  - Expected Calibration Error (ECE)
  - In relation to classification error

Some modern neural network families are both highly accurate and well-calibrated.



Family differences

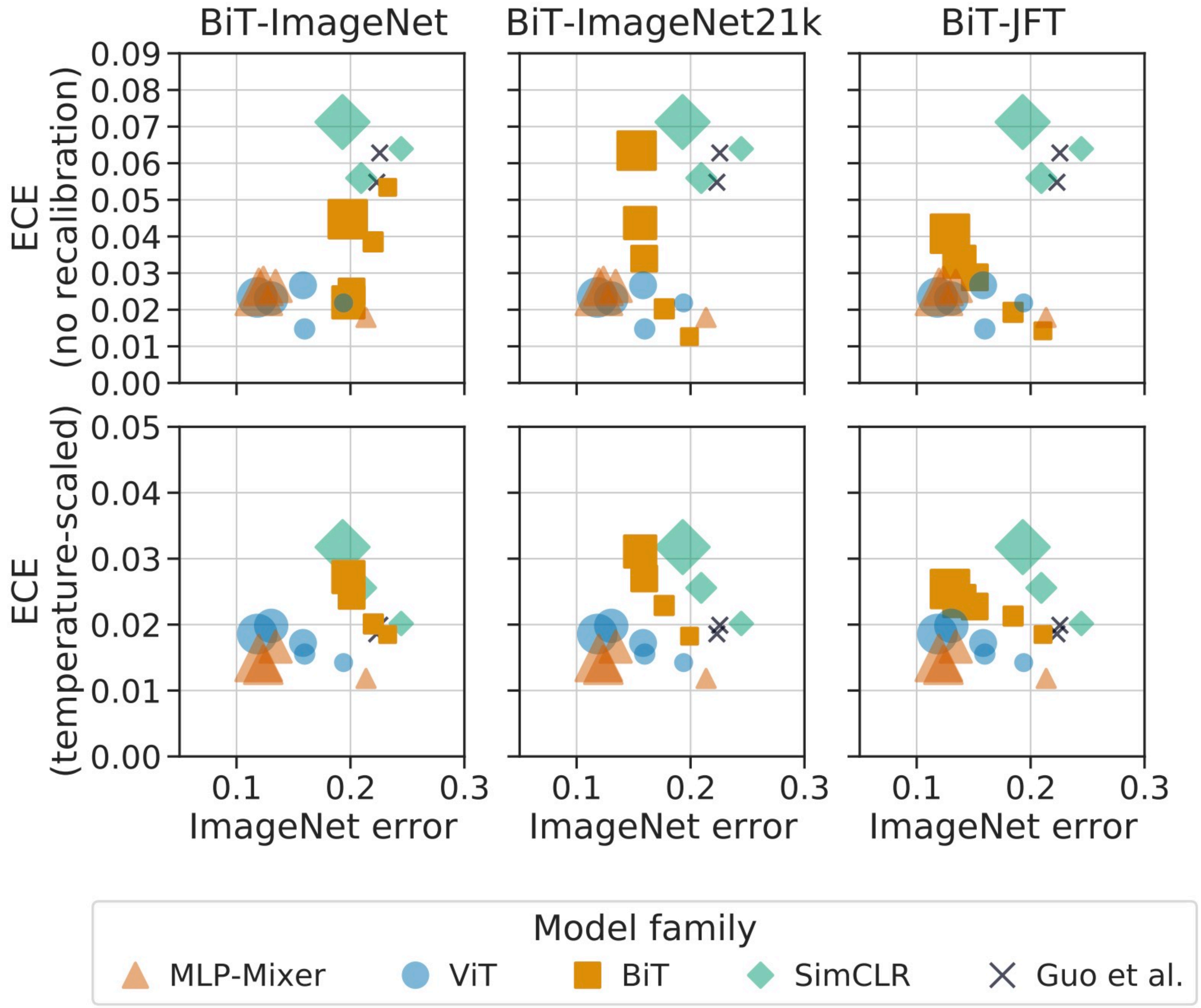
- Temperature scaling improves calibration and reveals consistent differences between model families.
- Temperature also reveals consistency with prior work
- Families occupy different Pareto sets



What explains family differences

- Model size? No.
- Pretraining dataset size? No.
- Pretraiing duration? No.

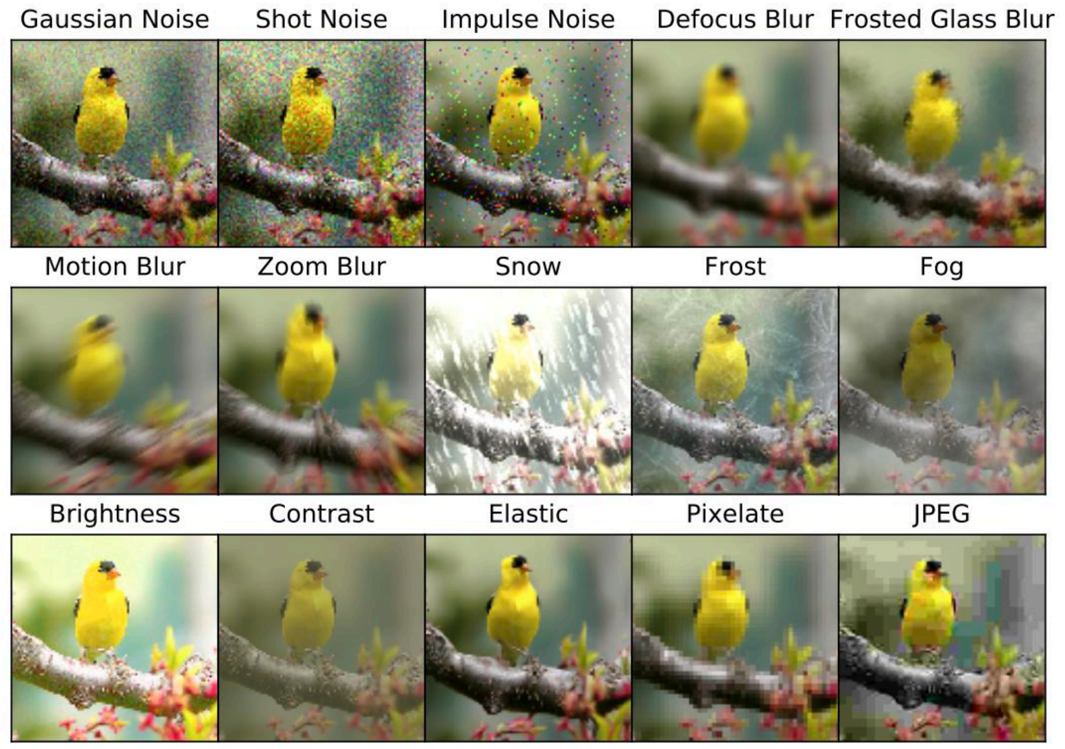
- Architecture? Likely.
- Other differences?Maybe.



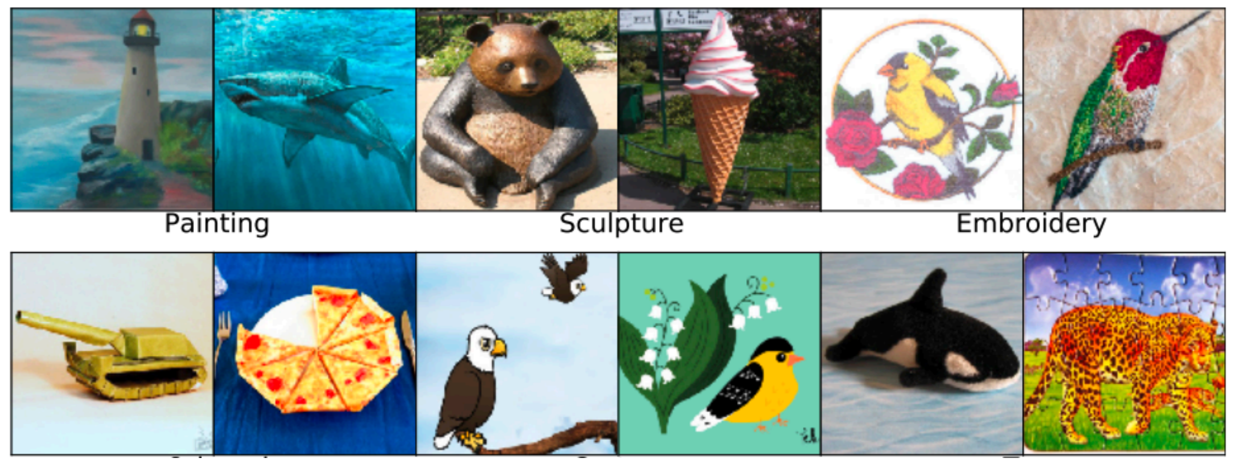
Out-of-distribution calibration

OOD datasets

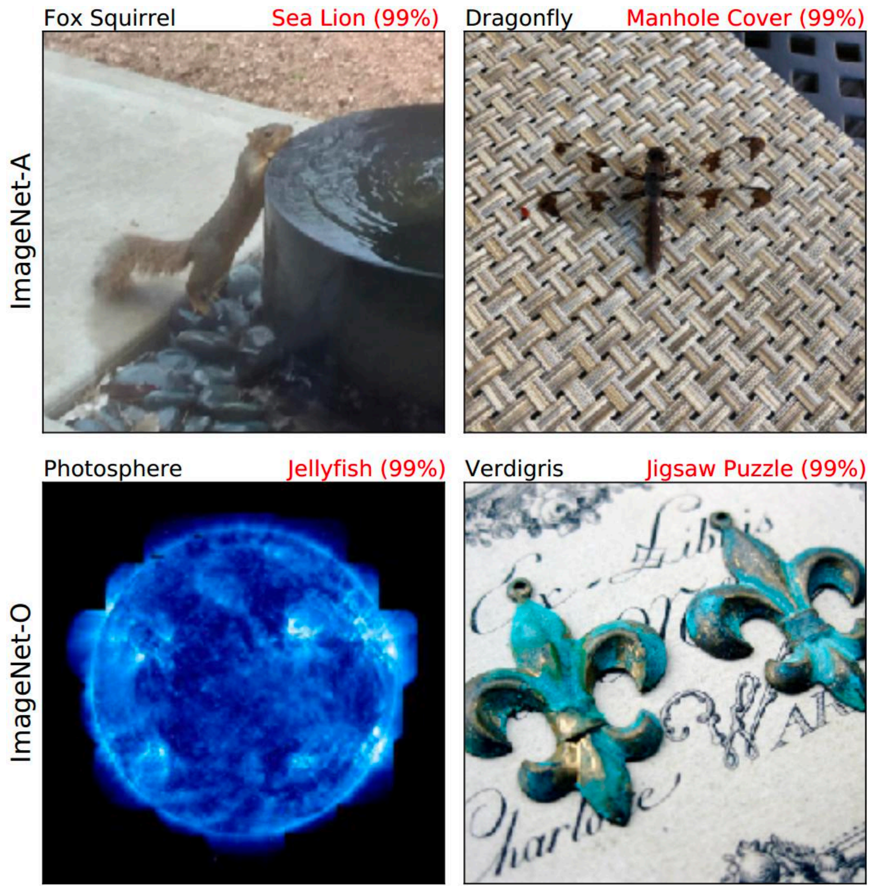
1. ImageNetV2 (Recht et al., 2019) is a new ImageNet test set collected by closely following the original ImageNet labeling protocol.
2. ImageNet-C (Hendrycks & Dietterich, 2019) consists of the images from ImageNet, modified with synthetic perturbations such as blur, pixelation, and compression artifacts 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.



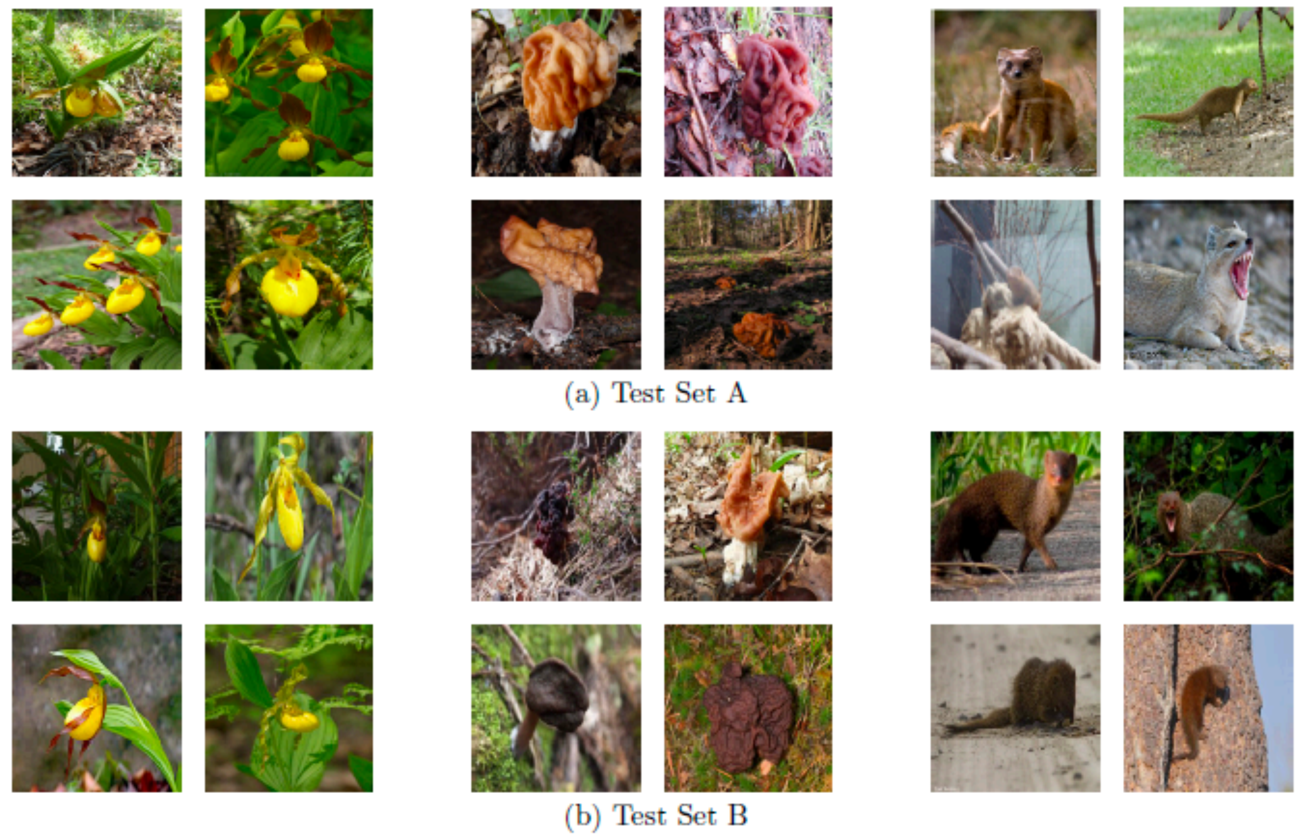
ImagNet-C



ImagNet-R



ImagNet-A

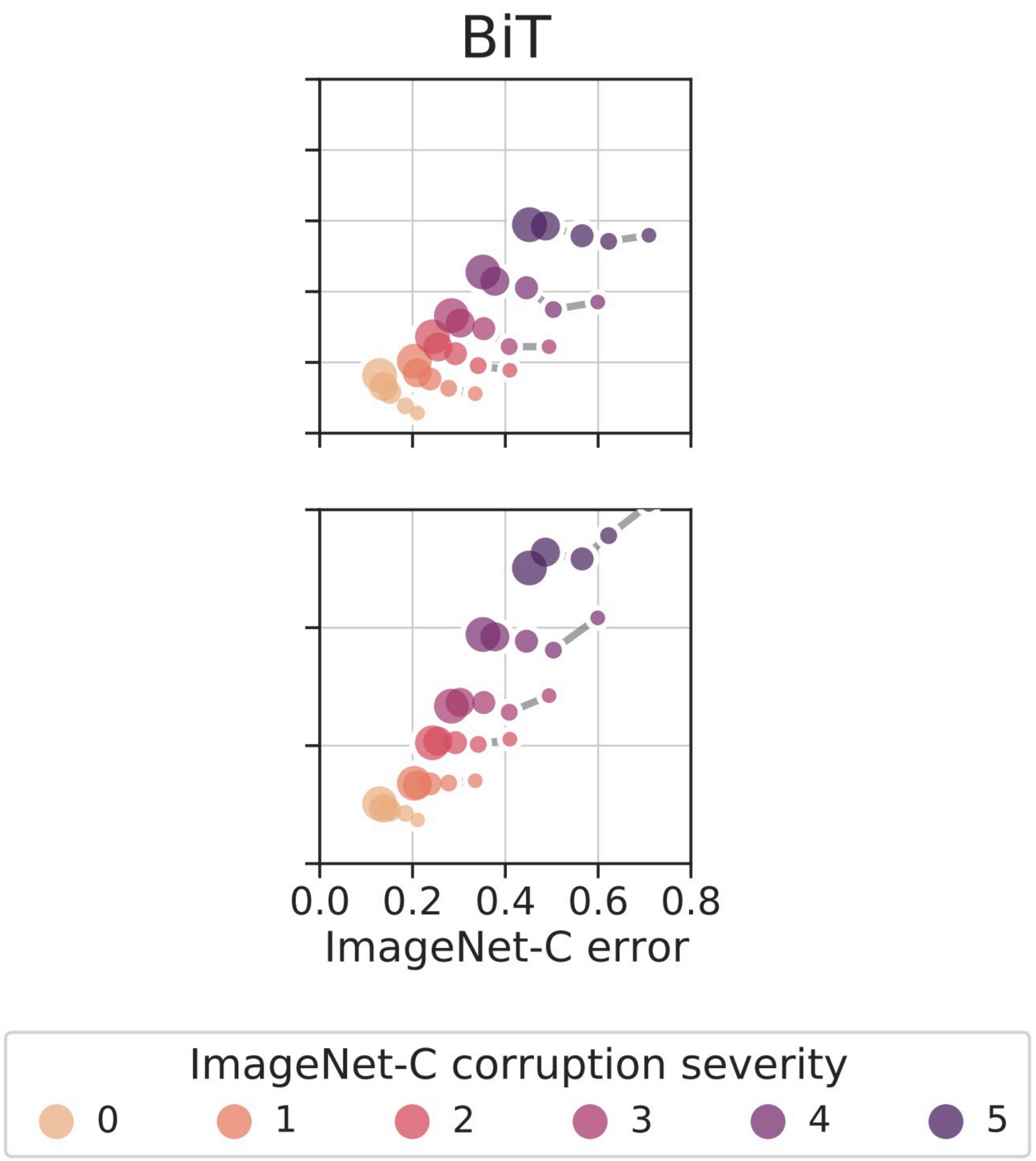


ImagNetV2

Out-of-distribution calibration

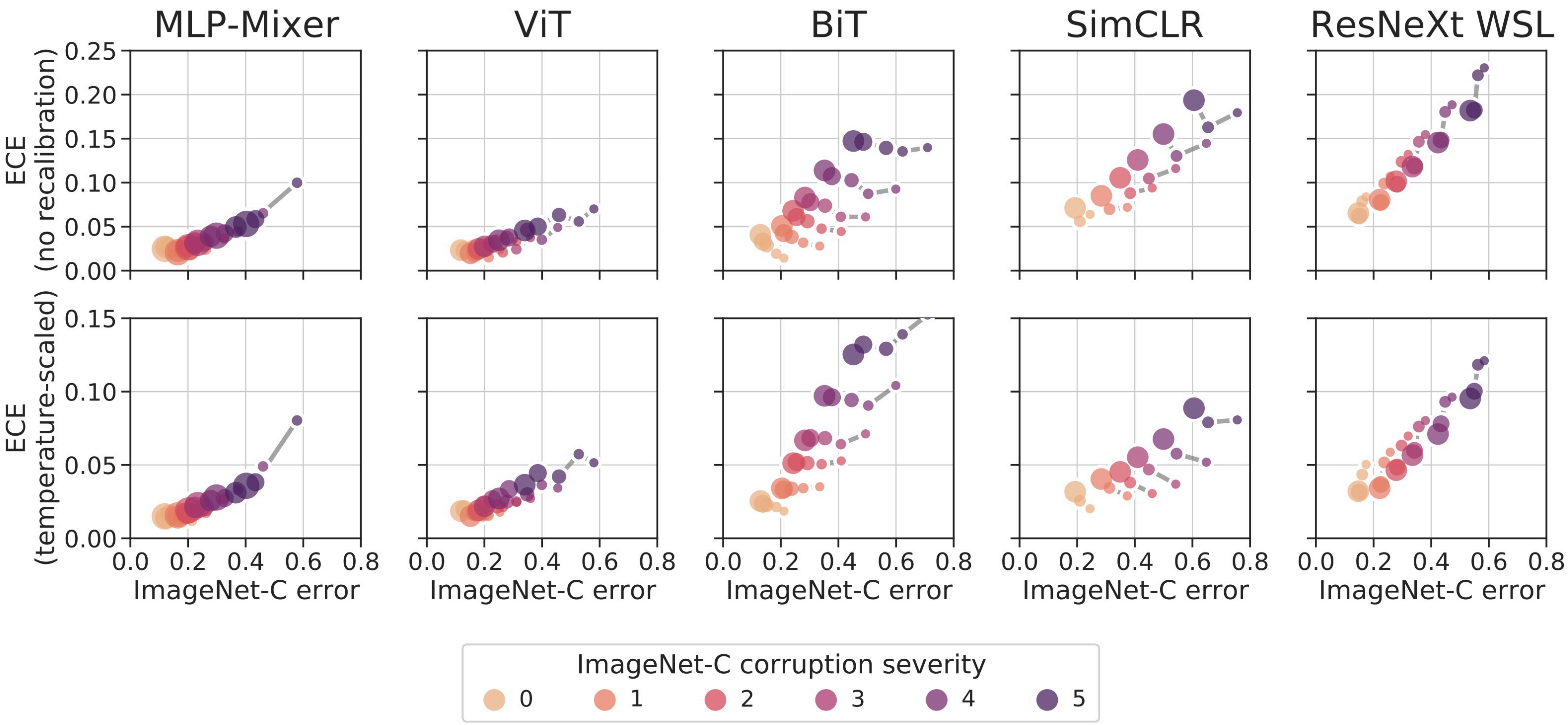
Calibration under distribution shift

- ImageNet-C:
  - Both classification error and calibration error increase under distribution shift.
  - Larger models tend to be more robust to distribution shift



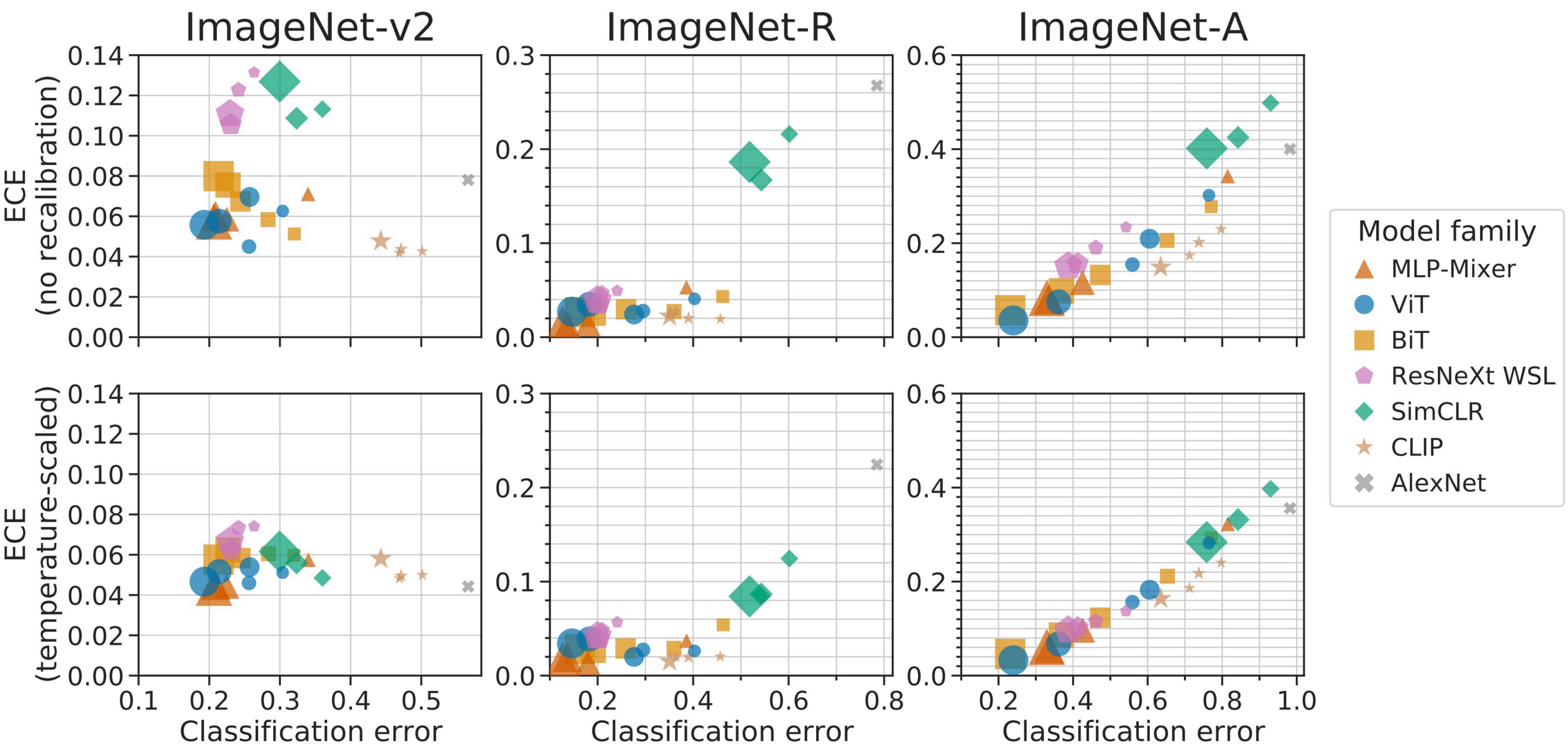
Out-of-distribution calibration

Calibration under distribution shift



Out-of-distribution calibration

Natural out-of-distribution benchmarks



Discussion

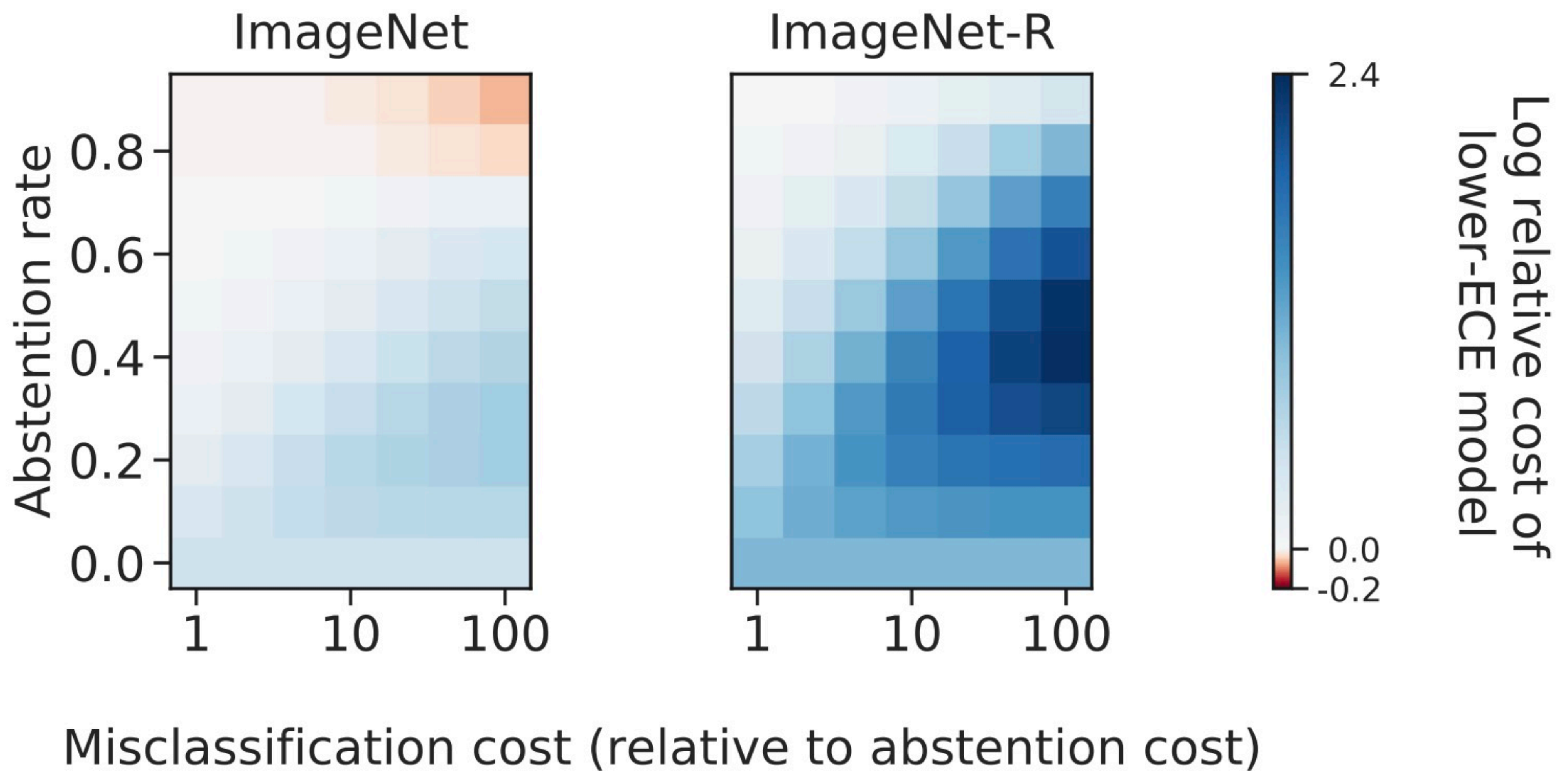
Trading off accuracy and calibration

- With families, there is an accuracy-calibration tradeoff.
- Which model variant should a practitioner choose?

Discussion

It depends on the task

- A decision cost function can relate accuracy and calibration
- In a selective prediction scenario, accuracy tends to outweigh calibration for the observed model differences



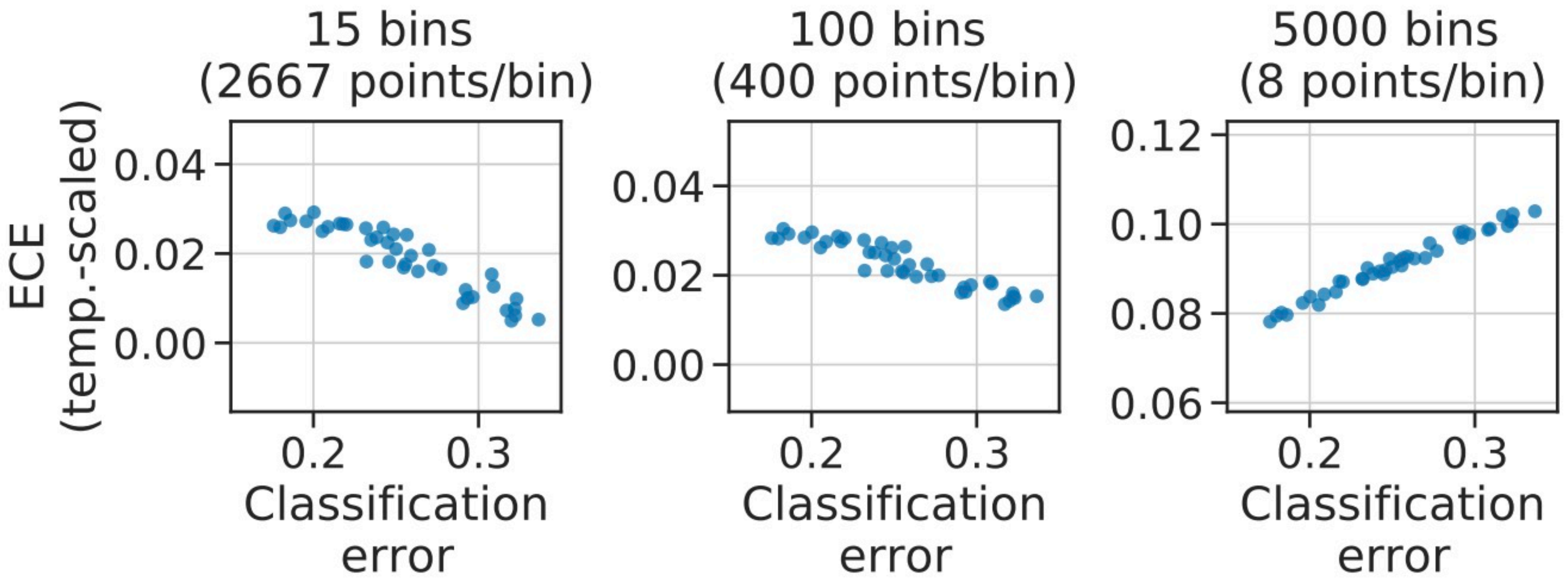
- Choose the more accurate model

Discussion

Estimator bias

- ECE estimators are biased.
- Bias depends on accuracy.
- Prudent choice of binning strategy minimize bias

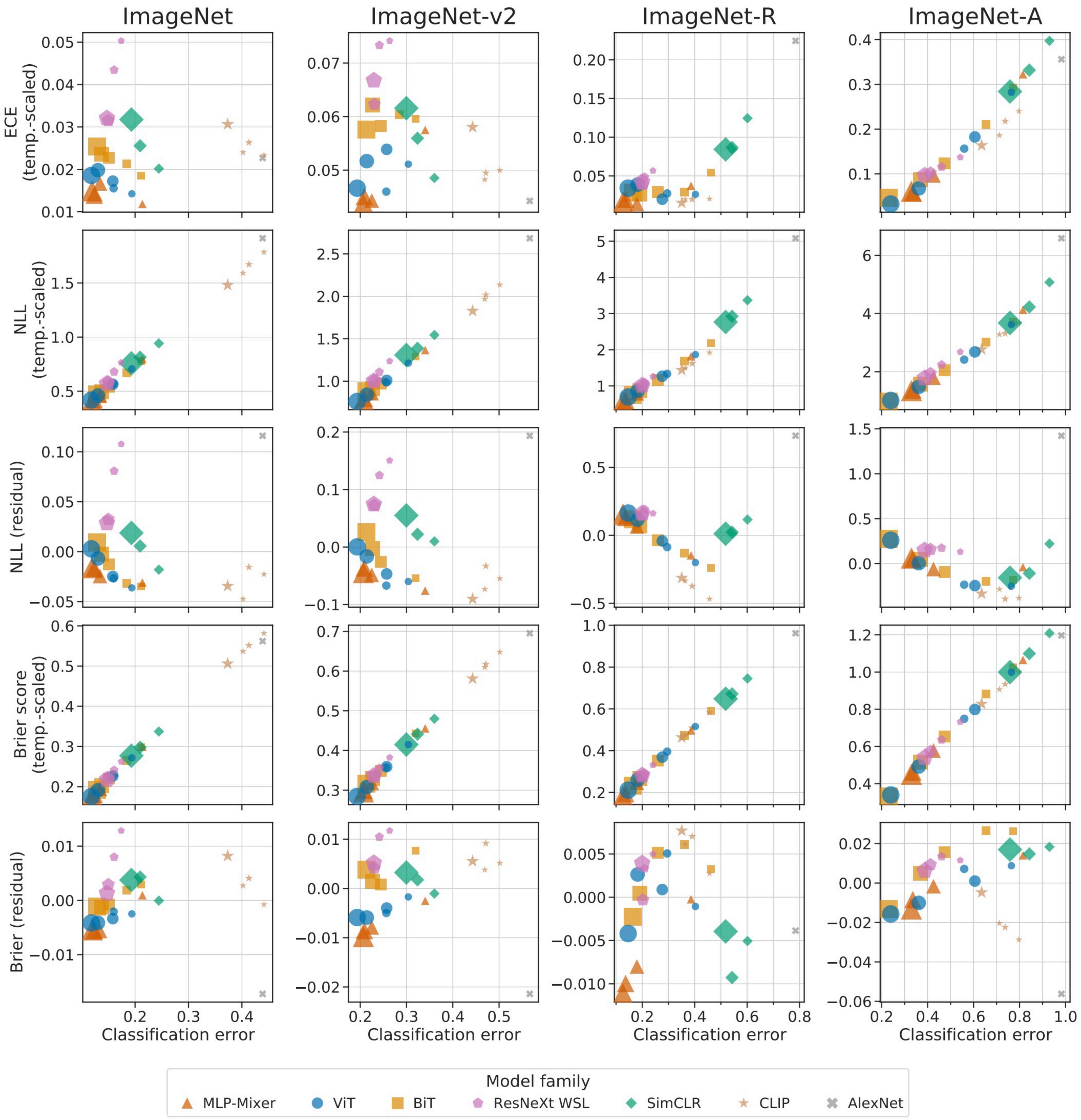
$$\frac{1}{n_i}\big(\mathbb{V}[A] + \mathbb{V}[C] - 2\mathrm{Cov}[C, A]\big),$$



Discussion

Alternative ECE variants

- Tested ECE estimator variants:
  - Equal-width binning
  - Equal-mass binning
  - Various bin sizes
  - Various normalization functions
  - All-label ECE
  - Class-wise ECE



- Results are qualitatively consistent