Supplementary Material

S1. Experimental Details

S1.1. Augmentations

The simple augmentation policy consists of a random crop and a horizontal flip, drawn from a widely used test-time augmentation policy in image classification [23]. The random crop pads the original image by 4 pixels and takes a 256x256 crop of the resulting image. The expanded augmentation consists of 12 augmentations; certain augmentations are stochastic, while others are deterministic. We design this set based on the augmentations included in AutoAugment [11]. We exclude certain augmentations, however, to exclude 1) redundancies among augmentations and thereby make the learned weights interpretable and 2) augmentations are unlikely to be label-preserving. In particular, we exclude CutOut (because it is clearly not label-preserving in many domains) and exclude brightness, contrast, saturation, and color for their overlap with color-jitter. We also exclude contrast, because it is already modified via autocontrast, and equalize and solarize for their overlap with autocontrast and invert. This leaves us the following augmentations:

- *Shear*: Shear an image by some number of degrees, sampled between [-10, 10] (stochastic).
- *Translate*: Samples a vertical shift (by fraction of image height) from [0, .1] (stochastic).
- *Rotate*: Samples a rotation (by degrees) from [-10, 10] (stochastic).
- *Autocontrast*: Maximizes contrast of images by remapping pixel values such that the the lowest becomes black and the highest becomes white (deterministic).
- *Invert*: Inverts the colors of an image (deterministic).
- *Blur*: Applies Gaussian blur with kernel size 5 (and default σ range of [.1, .2]) (stochastic).
- *Posterize*: Reduces the number of bits per channel to 4 (deterministic).
- *Color Jitter*: Randomly samples a brightness, contrast, and saturation adjustment parameter from the range [.9, 1.1] (stochastic).
- *Increase Sharpness*: Adjusts sharpness of image by a factor of 1.3 (deterministic).
- *Decrease Sharpness*: Adjusts sharpness of image by a factor of 0.7 (deterministic).
- *Random Crop*: Pads each image by 4 pixels, takes a 256x256 crop, and then proceeds to take a 224x224 center crop (stochastic).
- *Horizontal Flip*: Flips image horizontally (deterministic).

There are many possible expanded test-time augmentation policies; this particular policy serves as an illustrative example.

S1.2. Learning aggregation function

We learn $\hat{g}$ by minimizing the cross-entropy loss with respect to the true labels on the calibration set. Specifically, we learning the weights using SGD with a learning rate of .01, momentum of .9, and weight decay of 1e-4. We train each model for 50 epochs. There are natural improvements to our optimization, but this is not the focus of our work. Instead, our goal is to highlight the surprising effectiveness of TTA-Learned *without* the introduction of hyperparameter optimization. We train all models using a machine equipped with 4 Titan Xp GPUs, 2 Octa Intel Xeon E5-2620 CPUs, and 1TB of RAM.

S2. Supplementary Results

S2.1. Test-Time Augmentation and APS

TTA-Learned combined with the expanded augmentation policy produces the smallest set sizes when combined with APS, across the datasets considered (Table S1) and each base classifier (Table S3). In contrast to the results using RAPS, TTA-Learned does not significantly outperform TTA-Avg when combined with APS. The central reason is that the improvements TTA confers — namely, improved top-k accuracy — do not address the underlying sensitivity of APS to classes with low predicted probabilities. As Angelopoulos et al. [1] discuss, APS produces large prediction sets because of noisy estimates of small probabilities, which then end up included in the prediction sets. Both TTA-Learned and TTA-Avg smooth the probabilities: they reduce the number of low-probability classes by aggregating predictions over perturbations of the image. The benefit that both TTA-Learned and TTA-Avg add to APS is thus similar to how RAPS penalizes classes with low probabilities.

S2.2. Comparison to Top-1 and Top-5

We expand Table 1 to include the Top-1 and Top-5 baselines in Table S6. Unsurprisingly, neither outperform RAPS, and consequently none outperform the combination of RAPS, TTA-Learned, and the expanded augmentation policy.

S2.3. Comparison to minimizing focal loss

We expand Table 1 to include results for a variant of TTA-Learned which uses a focal loss in place of the cross-entropy loss. We conduct this exploration because empirically, the focal loss has been known to produce better-calibrated models. Table S7 reports our results. We see little difference between results when using a different loss function; RAPS+TTA-Leanred still outperforms RAPS + an average over the test-time augmentations, and RAPS alone. While this speaks to the method's flexibility to different loss functions, it is possible that the use of a loss function designed to reduce prediction set size could produce better

performance.

S2.4. Impact on coverage

We provide exact values of coverage for the main experiments here. In short, TTA-Learned combined with the expanded augmentation policy *never* worsens coverage, and in some cases, significantly improves it (although the improvements are small in magnitude). Coverage values for the RAPS experiment across coverage values and datasets can be found in Table S8 and coverage values for the RAPS experiment across base classifiers can be found in Table S9. Similarly, we provide coverage values for the APS experiment across datasets (Table S2) and across models (Table S3).

S2.5. Impact of different coverage guarantees and datasets

We replicate the class-specific analysis for ImageNet at a value of $\alpha = .05$ (Figure S5), iNaturalist (Figure S6), and CUB-Birds (Figure S7). All trends are consistent with results in the main text, save for one notable exception: when TTA-Learned is applied to CUB-Birds, prediction set sizes of the classes with the *smallest* prediction set sizes and classes that are *easier* to predict benefit most from TTA. The significance of the relationship between original prediction set size and TTA improvement disappears when conducted on an example level in this setting. This could be a result of class imbalance in the dataset; it is possible that the class-average prediction set size obscures important variation in CUB-Birds.

S2.6. Impact of augmentation policy size

We also analyze the impact of augmentation policy size on average prediction set size for CUB-Birds (Figure S2), to understand if additional augmentations may produce larger reductions in set size than we observe. Larger augmentation policies appear to provide an improvement to average prediction set size at $\alpha = .05$, but offer little improvement for $\alpha = .01$.

S2.7. Impact of TTA data split

Learning the test-time augmentation policy requires a set of labeled data *distinct* from those used to select the conformal threshold. This introduces a trade-off: more labeled data for test-time augmentation may result in more accurate weights, but a less accurate conformal threshold, and vice versa. We study this tradeoff empirically in the context of ImageNet and the expanded augmentation policy and show results in Figure S3. We find that, as more data is taken away from the conformal calibration set, variance in performance grows. This is in line with our intuition; we have fewer examples to approximate the distribution of conformal scores. However, at all percentages, test-time augmentation introduces a significant improvement in prediction set sizes over using all the labeled examples, and their original probabilities, to determine the threshold. This suggests that the benefits TTA confers outweigh the costs to the estimation of the conformal threshold, a practically useful insight to those who wish to apply conformal prediction in practice6

S2.8. Impact of calibration set size

We plot the relationship between calibration set size and average prediction set size in Figure S4 across two augmentation policies, two datasets, and two values of α. We see that TTA is more effective the larger the calibration set, in the context of ImageNet. In the context of CUB-Birds, it appears that TTA approaches equivalence with the conformal score alone as the calibration set size increases.

S2.9. Impact of different backbone architecture

Our results in the main text are limited to a single architecture (residual networks). Here, we provide evidence of generalizability to different architectures by replicating our ImageNet results using MobileNetV2, across a range of coverage guarantees and both augmentation policies (Table S11) and find consistent results, which support the versatility of the proposed method.

		Expanded Aug Policy			Simple Aug Policy		
Alpha	Method	ImageNet	iNaturalist	CUB-Birds	ImageNet	iNaturalist	CUB-Birds
0.01	APS	98.493 ± 3.075	131.681 ± 3.515	19.436 ± 0.995	98.493 ± 3.075	**131.681 ± 3.515**	**19.436 ± 0.995**
0.01	APS+TTA-Avg	**68.714 ± 2.856**	**84.546 ± 3.655**	**17.715 ± 1.523**	**92.027 ± 4.797**	145.401 ± 4.635	**19.152 ± 1.667**
0.01	APS+TTA-Learned	**69.009 ± 2.156**	**85.093 ± 2.768**	**17.766 ± 1.608**	**90.613 ± 6.421**	144.134 ± 4.371	**18.552 ± 1.326**
0.05	APS	19.820 ± 0.482	33.481 ± 0.786	5.921 ± 0.192	19.820 ± 0.482	**33.481 ± 0.786**	**5.921 ± 0.192**
0.05	APS+TTA-Avg	14.308 ± 0.279	**26.021 ± 0.282**	**4.870 ± 0.208**	**18.862 ± 0.498**	37.370 ± 0.735	6.306 ± 0.350
0.05	APS+TTA-Learned	**14.084 ± 0.241**	**26.289 ± 0.529**	**4.913 ± 0.145**	**19.119 ± 0.479**	36.940 ± 0.632	6.361 ± 0.480
0.10	APS	8.969 ± 0.158	16.755 ± 0.394	3.455 ± 0.164	8.969 ± 0.158	**16.755 ± 0.394**	**3.455 ± 0.164**
0.10	APS+TTA-Avg	**7.193 ± 0.101**	**14.583 ± 0.333**	**3.108 ± 0.114**	**8.787 ± 0.136**	18.300 ± 0.418	**3.609 ± 0.135**
0.10	APS+TTA-Learned	**7.215 ± 0.106**	**14.538 ± 0.395**	**3.046 ± 0.073**	**8.813 ± 0.180**	18.086 ± 0.420	3.638 ± 0.146

Table S1. We replicate our experiments across coverage levels and datasets using APS, another conformal score. TTA-Learned combined with the expanded augmentation policy produces the smallest set sizes across all comparisons. Interestingly, the simple augmentation policy is not as effective in the context of iNaturalist when using APS.

		Expanded Aug Policy			Simple Aug Policy		
Alpha	Method	ImageNet	iNaturalist	CUB-Birds	ImageNet	iNaturalist	CUB-Birds
0.01	APS	0.980 ± 0.001	0.986 ± 0.000	0.985 ± 0.001	0.980 ± 0.001	**0.986 ± 0.000**	**0.985 ± 0.001**
0.01	APS+TTA-Avg	**0.985 ± 0.001**	**0.989 ± 0.001**	**0.989 ± 0.002**	**0.981 ± 0.001**	0.987 ± 0.000	**0.986 ± 0.003**
0.01	APS+TTA-Learned	**0.985 ± 0.001**	**0.989 ± 0.001**	**0.990 ± 0.002**	**0.980 ± 0.002**	0.987 ± 0.000	**0.985 ± 0.002**
0.05	APS	0.931 ± 0.002	0.952 ± 0.001	0.945 ± 0.004	0.931 ± 0.002	**0.952 ± 0.001**	**0.945 ± 0.004**
0.05	APS+TTA-Avg	0.944 ± 0.002	**0.956 ± 0.001**	**0.949 ± 0.005**	**0.937 ± 0.002**	0.960 ± 0.001	0.949 ± 0.004
0.05	APS+TTA-Learned	**0.943 ± 0.002**	**0.957 ± 0.001**	**0.950 ± 0.005**	**0.937 ± 0.002**	0.959 ± 0.001	0.950 ± 0.005
0.10	APS	0.896 ± 0.002	0.923 ± 0.001	0.915 ± 0.006	0.896 ± 0.002	**0.923 ± 0.001**	**0.915 ± 0.006**
0.10	APS+TTA-Avg	**0.903 ± 0.002**	**0.930 ± 0.001**	**0.920 ± 0.007**	**0.905 ± 0.002**	0.933 ± 0.001	**0.922 ± 0.005**
0.10	APS+TTA-Learned	**0.904 ± 0.002**	**0.930 ± 0.001**	**0.918 ± 0.006**	**0.906 ± 0.002**	0.932 ± 0.001	0.922 ± 0.004

Table S2. Coverage values associated with experiments in Table S1. TTA-Learned produces significant improvements in coverage — larger in magnitude than in conjunction with RAPS — across when using the expanded augmentation policy. TTA-Learned produces no drops in coverage when using the simple augmentation policy, a nd produces improvements at $\alpha = .01$ and $\alpha = .05$.

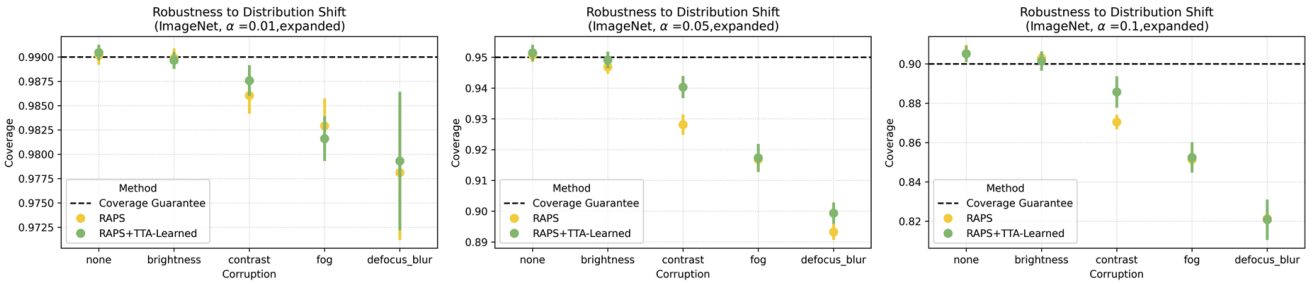

Figure S1. **Impact on coverage.** We plot achieved coverage for both RAPS and RAPS+TTA-Learned across several coverage guarantees and distribution shifts. As expected, distribution shift leads conformal predictors to not meet the coverage guarantee. In each case, the addition of TTA does not worsen coverage; in some cases (for example, given the contrast corruption and a coverage guarantee of 0.05) it even improves coverage.

		Expanded Aug Policy			Simple Aug Policy		
Alpha	Method	ResNet-50	ResNet-101	ResNet-152	ResNet-50	ResNet-101	ResNet-152
0.01	APS	98.493 ± 3.075	88.279 ± 4.121	79.231 ± 4.570	98.493 ± 3.075	88.279 ± 4.121	79.231 ± 4.570
0.01	APS+TTA-Avg	**68.714 ± 2.856**	**64.197 ± 2.336**	**62.885 ± 3.125**	**92.027 ± 4.797**	**77.344 ± 2.214**	**73.377 ± 3.600**
0.01	APS+TTA-Learned	**69.009 ± 2.156**	**64.852 ± 2.823**	**64.045 ± 3.398**	**90.613 ± 6.421**	**78.627 ± 4.101**	**74.571 ± 3.516**
0.05	APS	19.820 ± 0.482	15.830 ± 0.611	14.437 ± 0.591	19.820 ± 0.482	15.830 ± 0.611	**14.437 ± 0.591**
0.05	APS+TTA-Avg	14.308 ± 0.279	**11.085 ± 0.267**	**10.605 ± 0.373**	**18.862 ± 0.498**	**15.039 ± 0.405**	**14.206 ± 0.499**
0.05	APS+TTA-Learned	**14.084 ± 0.241**	**11.118 ± 0.209**	**10.595 ± 0.368**	**19.119 ± 0.479**	**15.011 ± 0.346**	**14.252 ± 0.486**
0.10	APS	8.969 ± 0.158	6.671 ± 0.175	6.134 ± 0.163	8.969 ± 0.158	**6.671 ± 0.175**	**6.134 ± 0.163**
0.10	APS+TTA-Avg	**7.193 ± 0.101**	**5.454 ± 0.098**	**5.111 ± 0.096**	**8.787 ± 0.136**	6.838 ± 0.143	6.309 ± 0.178
0.10	APS+TTA-Learned	**7.215 ± 0.106**	**5.490 ± 0.090**	**5.131 ± 0.061**	**8.813 ± 0.180**	6.826 ± 0.121	6.311 ± 0.123

Table S3. Results across base classifiers using APS alone, APS + TTA-Avg, and APS + TTA-learned in conjunction with the expanded augmentation policy (left) and simple augmentation policy (right). TTA-Learned and the expanded augmentation policy produce the smallest prediction sets (on average).

		Expanded Aug Policy			Simple Aug Policy		
Alpha	Method	ResNet-50	ResNet-101	ResNet-152	ResNet-50	ResNet-101	ResNet-152
0.01	APS	0.980 ± 0.001	0.979 ± 0.002	0.978 ± 0.002	**0.980 ± 0.001**	**0.979 ± 0.002**	**0.978 ± 0.002**
0.01	APS+TTA-Avg	**0.985 ± 0.001**	**0.985 ± 0.001**	**0.984 ± 0.001**	**0.981 ± 0.001**	**0.980 ± 0.001**	**0.978 ± 0.002**
0.01	APS+TTA-Learned	**0.985 ± 0.001**	**0.985 ± 0.001**	**0.984 ± 0.001**	**0.980 ± 0.002**	**0.980 ± 0.002**	**0.979 ± 0.002**
0.05	APS	0.931 ± 0.002	0.930 ± 0.002	0.929 ± 0.002	0.931 ± 0.002	0.930 ± 0.002	0.929 ± 0.002
0.05	APS+TTA-Avg	**0.944 ± 0.002**	**0.942 ± 0.001**	**0.942 ± 0.002**	**0.937 ± 0.002**	**0.935 ± 0.002**	**0.934 ± 0.002**
0.05	APS+TTA-Learned	0.943 ± 0.002	**0.942 ± 0.001**	**0.942 ± 0.002**	**0.937 ± 0.002**	**0.935 ± 0.001**	**0.934 ± 0.002**
0.10	APS	0.896 ± 0.002	0.892 ± 0.002	0.893 ± 0.002	0.896 ± 0.002	0.892 ± 0.002	0.893 ± 0.002
0.10	APS+TTA-Avg	**0.903 ± 0.002**	**0.901 ± 0.001**	**0.902 ± 0.001**	**0.905 ± 0.002**	**0.903 ± 0.001**	**0.903 ± 0.002**
0.10	APS+TTA-Learned	**0.904 ± 0.002**	**0.902 ± 0.001**	**0.902 ± 0.001**	**0.906 ± 0.002**	**0.903 ± 0.002**	**0.903 ± 0.002**

Table S4. Coverage values for APS and TTA variants of APS across base classifiers, using ImageNet. TTA-Learned or TTA-Avg in combination with the expanded augmentation policy significantly improve coverage in every comparison.

	Expanded Aug Policy			Simple Aug Policy		
Method	ResNet50	ResNet101	ResNet152	ResNet50	ResNet101	ResNet152
Original	0.761 ± 0.002	0.773 ± 0.001	0.783 ± 0.002	0.761 ± 0.002	0.773 ± 0.001	0.783 ± 0.002
TTA-Avg	0.764 ± 0.002	0.778 ± 0.001	0.788 ± 0.002	0.77 ± 0.002	0.783 ± 0.001	0.792 ± 0.002
TTA-Learned	0.771 ± 0.002	0.785 ± 0.001	0.793 ± 0.002	0.771 ± 0.002	0.784 ± 0.001	0.793 ± 0.002

Table S5. **TTA effect on classifier performance.** We report differences in classifier performance using a learned test-time augmentation policy compared to a simple average (TTA-Avg) and no test-time augmentation (Original). TTA-Learned offers small improvements over a simpler average and the original model across architectures. FILL IN THE REST, explain how TTA's improvement to Top-1 accuracy alone is small, and does not fully explain the value of test-time augmentation to conformal prediction.

		ImageNet		iNaturalist		CUB-Birds	
Alpha	Method	Prediction Set Size	Empirical Coverage	Prediction Set Size	Empirical Coverage	Prediction Set Size	Empirical Coverage
0.01	Top-1	1.000 ± 0.000	0.761 ± 0.002	1.000 ± 0.000	0.766 ± 0.001	1.000 ± 0.000	0.804 ± 0.008
0.01	Top-5	5.000 ± 0.000	0.928 ± 0.001	5.000 ± 0.000	0.915 ± 0.001	5.000 ± 0.000	0.959 ± 0.003
0.01	RAPS	37.751 ± 2.334	0.990 ± 0.001	61.437 ± 6.067	0.990 ± 0.001	15.293 ± 2.071	0.990 ± 0.001
0.01	RAPS+TTA-Avg	35.600 ± 2.200	0.991 ± 0.001	57.073 ± 5.914	0.990 ± 0.001	13.111 ± 2.470	0.991 ± 0.002
0.01	RAPS+TTA-Learned	31.248 ± 2.177	0.990 ± 0.001	53.195 ± 4.884	0.990 ± 0.001	14.045 ± 1.323	0.991 ± 0.002
0.05	Top-1	1.000 ± 0.000	0.761 ± 0.002	1.000 ± 0.000	0.766 ± 0.001	1.000 ± 0.000	0.804 ± 0.008
0.05	Top-5	5.000 ± 0.000	0.928 ± 0.001	5.000 ± 0.000	0.915 ± 0.001	5.000 ± 0.000	0.959 ± 0.003
0.05	RAPS	5.637 ± 0.357	0.951 ± 0.002	7.991 ± 1.521	0.954 ± 0.002	3.624 ± 0.361	0.955 ± 0.007
0.05	RAPS+TTA-Avg	5.318 ± 0.113	0.951 ± 0.001	7.067 ± 0.344	0.952 ± 0.002	3.116 ± 0.210	0.954 ± 0.007
0.05	RAPS+TTA-Learned	4.889 ± 0.168	0.952 ± 0.001	6.682 ± 0.447	0.954 ± 0.002	3.571 ± 0.576	0.957 ± 0.007
0.10	Top-1	1.000 ± 0.000	0.761 ± 0.002	1.000 ± 0.000	0.766 ± 0.001	1.000 ± 0.000	0.804 ± 0.008
0.10	Top-5	5.000 ± 0.000	0.928 ± 0.001	5.000 ± 0.000	0.915 ± 0.001	5.000 ± 0.000	0.959 ± 0.003
0.10	RAPS	2.548 ± 0.074	0.906 ± 0.004	2.914 ± 0.116	0.907 ± 0.003	2.038 ± 0.153	0.919 ± 0.014
0.10	RAPS+TTA-Avg	2.470 ± 0.071	0.905 ± 0.005	2.740 ± 0.026	0.908 ± 0.002	1.780 ± 0.139	0.912 ± 0.014
0.10	RAPS+TTA-Learned	2.312 ± 0.054	0.905 ± 0.004	2.625 ± 0.043	0.909 ± 0.003	1.893 ± 0.187	0.919 ± 0.016

Table S6. **Comparison to Top-1 and Top-5 baselines.** Results comparing performance against Top-K baselines. In each setting, conformal prediction produces either smaller set sizes, higher coverage, or both compared to the Top-K baselines.

		Expanded Aug Policy		Simple Aug Policy	
Alpha	Method	ImageNet	CUB-Birds	ImageNet	CUB-Birds
0.01	RAPS+TTA-Learned+Focal	32.612 ± 3.799	13.416 ± 1.991	31.230 ± 1.510	15.503 ± 2.364
0.01	RAPS+TTA-Learned+Conformal	32.257 ± 3.608	13.776 ± 2.198	31.716 ± 2.078	14.432 ± 2.184
0.01	RAPS+TTA-Learned+CE	31.248 ± 2.177	14.045 ± 1.323	32.702 ± 2.409	13.803 ± 1.734
0.05	RAPS+TTA-Learned+Focal	4.906 ± 0.195	3.194 ± 0.202	4.956 ± 0.239	3.313 ± 0.331
0.05	RAPS+TTA-Learned+Conformal	4.867 ± 0.122	3.302 ± 0.312	4.996 ± 0.405	3.412 ± 0.406
0.05	RAPS+TTA-Learned+CE	4.889 ± 0.168	3.571 ± 0.576	5.040 ± 0.176	3.290 ± 0.186
0.10	RAPS+TTA-Learned+Focal	2.363 ± 0.085	1.791 ± 0.102	2.308 ± 0.045	1.860 ± 0.131
0.10	RAPS+TTA-Learned+Conformal	2.308 ± 0.068	1.865 ± 0.163	2.330 ± 0.072	1.868 ± 0.122
0.10	RAPS+TTA-Learned+CE	2.312 ± 0.054	1.893 ± 0.187	2.362 ± 0.065	1.840 ± 0.106

Table S7. **Alternate training objectives.** Results across datasets for two augmentation policies and three coverage specifications using a focal loss. We set γ to be 1, in line with prior work [14]. Each entry corresponds to the average prediction set size across 10 calibration/test splits. Both the focal and conformal loss do not outperform the cross-entropy loss; for simplicity, we report all results using the cross-entropy loss.

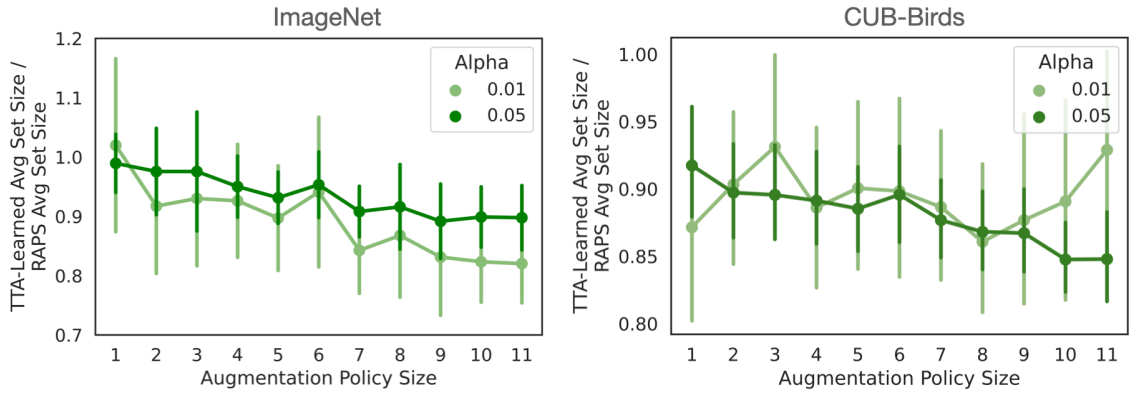

Figure S2. **Impact of augmentation policy size.** We see that larger policy sizes translate to a greater improvement (in terms of the ratio of average prediction set sizes using RAPS+TTA-Learned to average prediction set sizes using RAPS alone) for $\alpha = .05$. For $\alpha = .01$, there is no clear trend.

		Expanded Aug Policy			Simple Aug Policy		
Alpha	Method	ImageNet	iNaturalist	CUB-Birds	ImageNet	iNaturalist	CUB-Birds
0.01	RAPS	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**
0.01	RAPS+TTA-Avg	**0.991 ± 0.001**	**0.990 ± 0.001**	**0.991 ± 0.002**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.991 ± 0.002**
0.01	RAPS+TTA-Learned	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.991 ± 0.002**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.002**
0.05	RAPS	**0.951 ± 0.002**	**0.954 ± 0.002**	**0.955 ± 0.007**	**0.951 ± 0.002**	**0.954 ± 0.002**	**0.955 ± 0.007**
0.05	RAPS+TTA-Avg	**0.951 ± 0.001**	**0.952 ± 0.002**	**0.954 ± 0.007**	**0.951 ± 0.001**	**0.953 ± 0.003**	**0.957 ± 0.004**
0.05	RAPS+TTA-Learned	**0.952 ± 0.001**	**0.954 ± 0.002**	**0.957 ± 0.007**	**0.951 ± 0.002**	**0.952 ± 0.002**	**0.956 ± 0.007**
0.10	RAPS	**0.906 ± 0.004**	**0.907 ± 0.003**	**0.919 ± 0.014**	**0.906 ± 0.004**	**0.907 ± 0.003**	**0.919 ± 0.014**
0.10	RAPS+TTA-Avg	**0.905 ± 0.005**	**0.908 ± 0.002**	**0.912 ± 0.014**	**0.905 ± 0.004**	**0.908 ± 0.002**	**0.915 ± 0.010**
0.10	RAPS+TTA-Learned	**0.905 ± 0.004**	**0.909 ± 0.003**	**0.919 ± 0.016**	**0.907 ± 0.004**	**0.908 ± 0.003**	**0.913 ± 0.011**

Table S8. **Comparison of achieved coverage.** Coverage values for RAPS, RAPS+TTA-Avg, and RAPS+TTA-Learned across datasets and coverage values. RAPS+TTA-Learned never decreases the coverage achieved by RAPS alone, and in some cases, improves it significantly (as in the case of ImageNet and iNaturalist).

		Expanded Aug Policy			Simple Aug Policy		
Alpha	Method	ResNet-50	ResNet-101	ResNet-152	ResNet-50	ResNet-101	ResNet-152
0.01	RAPS	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**
0.01	RAPS+TTA-Avg	**0.991 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**
0.01	RAPS+TTA-Learned	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**
0.05	RAPS	**0.951 ± 0.002**	**0.952 ± 0.002**	**0.952 ± 0.002**	**0.951 ± 0.002**	**0.952 ± 0.002**	**0.952 ± 0.002**
0.05	RAPS+TTA-Avg	**0.951 ± 0.001**	**0.951 ± 0.001**	**0.952 ± 0.002**	**0.951 ± 0.001**	**0.952 ± 0.002**	**0.952 ± 0.002**
0.05	RAPS+TTA-Learned	**0.952 ± 0.001**	**0.952 ± 0.002**	**0.952 ± 0.002**	**0.951 ± 0.002**	**0.952 ± 0.002**	**0.952 ± 0.002**
0.10	RAPS	**0.906 ± 0.004**	**0.906 ± 0.004**	0.906 ± 0.002	**0.906 ± 0.004**	0.906 ± 0.004	0.906 ± 0.002
0.10	RAPS+TTA-Avg	**0.905 ± 0.005**	0.905 ± 0.002	0.908 ± 0.002	**0.905 ± 0.004**	**0.908 ± 0.004**	**0.910 ± 0.002**
0.10	RAPS+TTA-Learned	**0.905 ± 0.004**	**0.907 ± 0.003**	**0.911 ± 0.002**	**0.907 ± 0.004**	**0.908 ± 0.004**	**0.910 ± 0.002**

Table S9. **Comparison of coverage across base classifiers.** Coverage values for TTA variants of conformal prediction compared to RAPS alone, across different base classifiers on ImageNet. TTA-Learned preserves coverage across all comparisons and significantly improves upon the achieved coverage using ResNet-101 with RAPS (granted, the magnitude of this improvement is small).

Alpha	Method	ImageNet	iNaturalist	CUB-Birds
0.01	RAPS	0.0112 ± 0.0043	0.0207 ± 0.0043	0.0076 ± 0.0031
0.01	RAPS+TTA-Learned	0.0113 ± 0.0067	0.0247 ± 0.0027	0.0046 ± 0.0026
0.05	RAPS	0.2134 ± 0.0348	0.0609 ± 0.0217	0.0112 ± 0.0105
0.05	RAPS+TTA-Learned	0.3338 ± 0.0994	0.0899 ± 0.0520	0.0350 ± 0.0412
0.10	RAPS	0.1318 ± 0.0696	0.0852 ± 0.0151	0.2218 ± 0.1260
0.10	RAPS+TTA-Learned	0.3198 ± 0.0977	0.1008 ± 0.0058	0.1931 ± 0.1208

Table S10. **Effect of test-time augmented conformal prediction on adaptivity.** We show results in the context of ResNet-50 and RAPS, across several coverage guarantees. We compute size-stratified coverage violation (SSCV) for each run as described in Sec. 5, and report the mean and standard deviation of SSCV across runs here. Test-time augmentation does not significantly diminish adaptivity at each coverage guarantee considered (assessed via a two-sample t-test, $p > 0.05$).

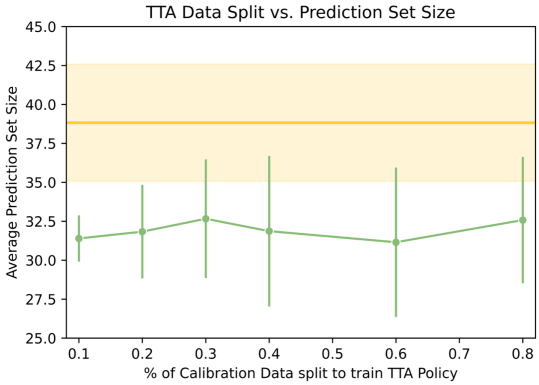

Figure S3. **Robustness to size of dataset used to train test-time augmentation policy.** We plot the percentage of data used to train the TTA policy on the x-axis and the average prediction set size on the y-axis. Error bars describe variance over 10 random splits of the calibration and test set. We can make two observations: 1) as the data used to train the TTA policy increases and the data used to estimate the conformal threshold decreases, variance in performance grows and 2) across a wide range of data splits, learned TTA policies (green) introduce improvements to achieved prediction set sizes compared to the original probabilities (gold). These results also suggest that relatively little training data is required to learn a useful test-time augmentation policy; in this case, 2-3 images per class, or 10% of the available labeled data.

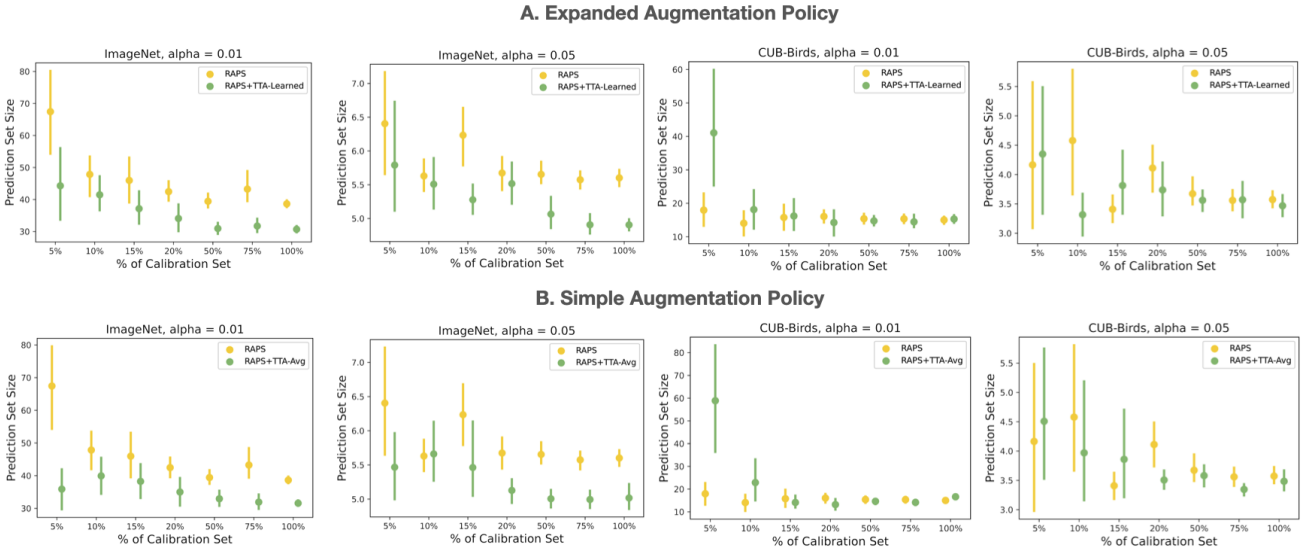

Figure S4. **Impact of calibration set size.** We plot the relationship between calibration set size and average prediction set size across two values of alpha, two augmentation policies, and two datasets (ImageNet and CUB-Birds). For ImageNet, larger calibration set sizes correlate with larger and more consistent improvements from the addition of TTA, where the improvement flattens out for calibration set sizes larger than 50%, or 12,500 images (12-13 per class). TTA does appear to be able to improve average prediction set size even with a calibration set size of 1,250 (5% of original ImageNet calibration set size). For CUB-Birds, a dataset on which TTA does not perform as well, we see that TTA performs comparably to RAPS alone the larger the calibration set.

α	Method	ImageNet (Expanded)	ImageNet (Simple)
0.01	RAPS	52.332 ± 8.970	52.332 ± 8.970
0.01	RAPS+TTA-Avg	45.604 ± 1.515	42.431 ± 1.516
0.01	RAPS+TTA-Learned	40.872 ± 1.377	40.843 ± 1.707
0.05	RAPS	8.872 ± 0.417	8.872 ± 0.417
0.05	RAPS+TTA-Avg	8.304 ± 0.322	7.945 ± 0.861
0.05	RAPS+TTA-Learned	7.723 ± 0.916	7.609 ± 1.027
0.10	RAPS	3.677 ± 0.104	3.677 ± 0.104
0.10	RAPS+TTA-Avg	3.480 ± 0.056	3.298 ± 0.069
0.10	RAPS+TTA-Learned	3.321 ± 0.289	3.348 ± 0.275

Table S11. **Replicated results on MobileNetV2.** We observe trends similar to those reported to in the main text in the context of MobileNetV2. In short, RAPS combined with a learned test-time augmentation policy (RAPS+TTA-Learned) produces the smallest set sizes across the considered coverage guarantees ($\alpha \in \{0.01, 0.05, 0.10\}$) and augmentation policies.

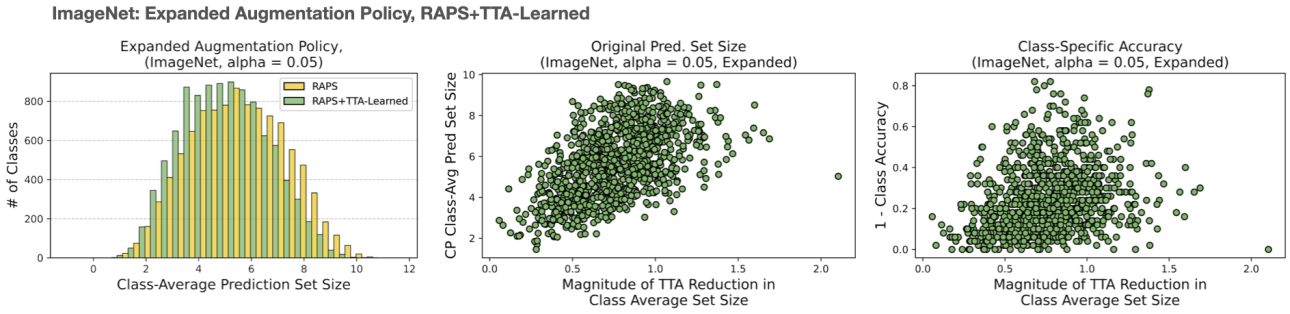

Figure S5. **Class-specific performance for ImageNet**, for a coverage of 95% $\alpha = .05$. Using the expanded augmentation policy RAPS+TTA-Learned produces a noticeable shift in class-average prediction set sizes to the left. There is a significant correlation between original prediction set size and improvements from TTA (middle) and between class difficulty and improvements from TTA (right).

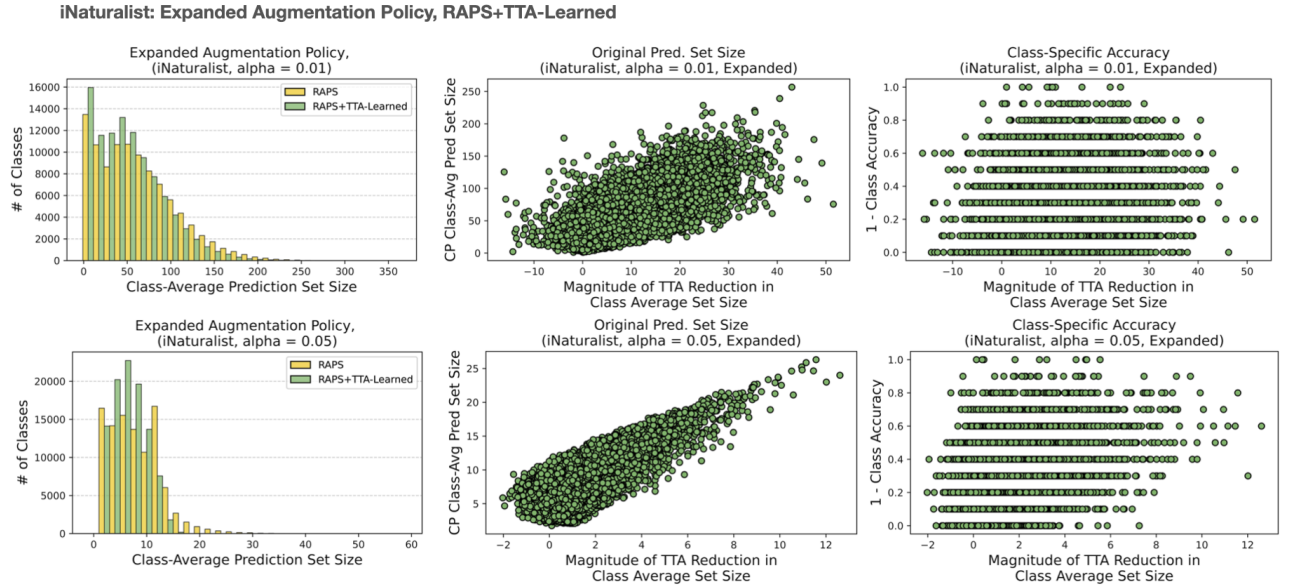

Figure S6. **Class-specific performance for iNaturalist**, for $\alpha = .01$ (top) and $\alpha = .05$ (bottom). We see a consistent relationship between TTA improvements and original class-average prediction set size (middle) and class difficulty (right). Estimates of class-specific accuracy on iNaturalist are quite noisy because there are 10 images per class (which produces distinct accuracy bands).

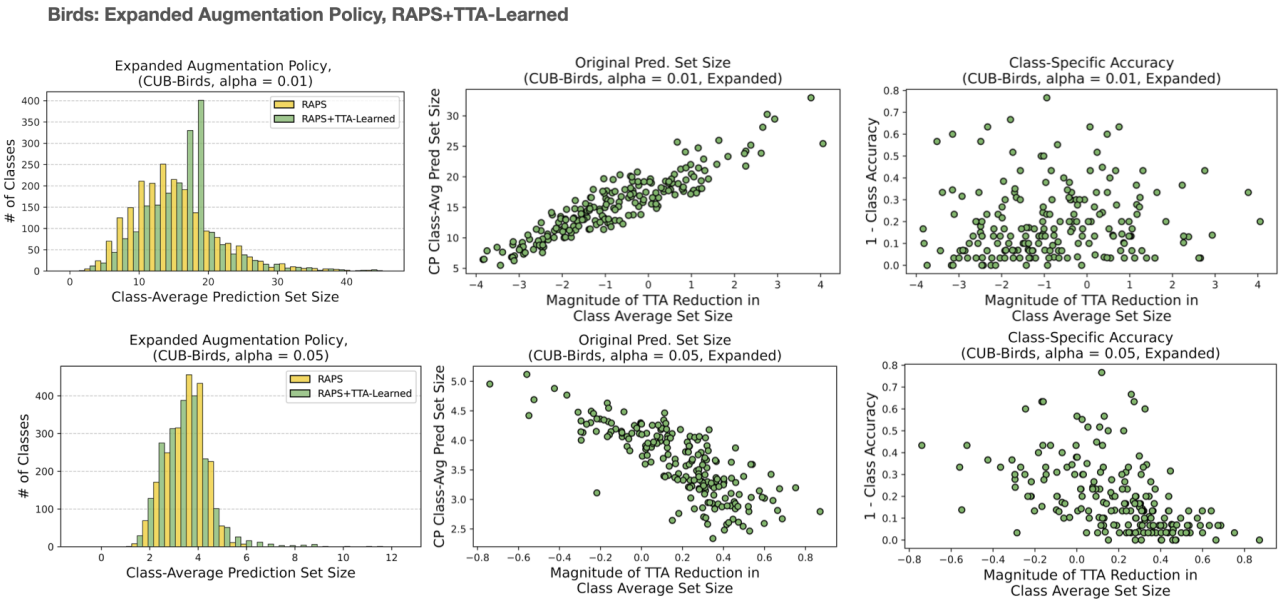

Figure S7. **Class-specific performance for CUB-Birds**, for $\alpha = .01$ (top) and $\alpha = .05\%$ (bottom). These graphs show an example for which TTA-Learned does *not* produce improvements in average prediction set size (computed across all examples). Interestingly, behavior on a class-specific level is different between $\alpha = .01$ and $\alpha = .05$. For $\alpha = .01$, results are consistent with other datasets: classes which originally receive large prediction set sizes and classes which are more difficult benefit most from the addition of TTA. For $\alpha = .05$, the exact opposite is true. While a majority of classes are hurt by TTA, classes that benefit from TTA are easier and receive smaller prediction set sizes.