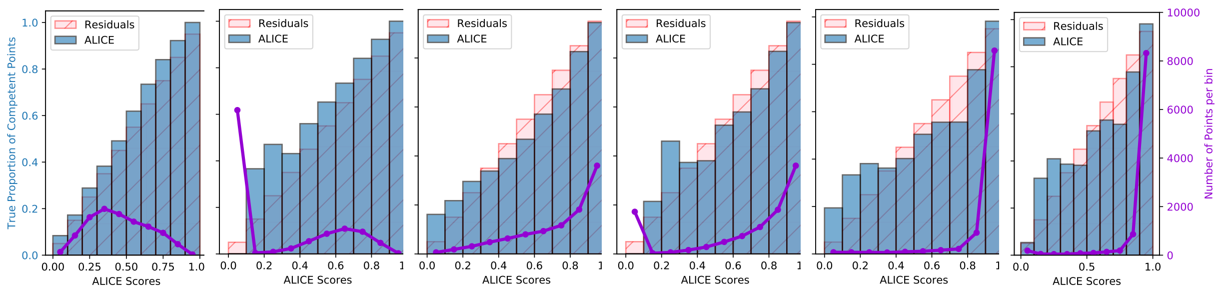
The authors would like to thank the reviewers for their thoughtful comments. Our responses are below:

We have replaced the calibration experiment. In this new experiment, we aim to show that the ALICE score matches its semantic meaning: for all points with ALICE score of p, we expect p of them to be truly competent. To show this, we bin the ALICE scores into tenths ([0.0 - 0.1), [0.1 - 0.2), ..., [0.9, 1.0)) and plot the true proportion of competent points for each bin as a histogram. Note that a perfect competence estimation would result in these histograms roughly resembling a $y = x$ curve. We visualize the difference between our competence estimator and perfect competence estimation by showing these residuals as well as the number of points in each bin in Figure 1. Note that ALICE is relatively well-calibrated at all stages of training and for all error functions tested. We would like to make clear that all mentions of the word *interpretable* refer to this interpretability of the ALICE score — *not* the interpretability of any machine learning model's predictions, as we state in Line 115.



(a) $\mathcal{E}_{0\text{-}1}$, epochs = 1 **(b)** $\mathcal{E}_{\text{xent}}$, epochs = 1 **(c)** $\mathcal{E}_{0\text{-}1}$, epochs = 5 **(d)** $\mathcal{E}_{\text{xent}}$, epochs = 5 **(e)** $\mathcal{E}_{0\text{-}1}$, epochs = 50 **(f)** $\mathcal{E}_{\text{xent}}$, epochs = 50

Figure 1: ALICE score calibration of ResNet32 trained on CIFAR10. At 50 epochs we reach max validation accuracy. Full experimental details are in the final version.

We have replaced the last paragraph of section 4 for futher clarity. It now reads: "Note that this metric only evaluates how well each estimator *orders* the test points based on competence, and does not consider the actual value of the score. We test this since some competence estimators (e.g. TrustScore) only seek to *rank* points based on competence and do not care what the magnitude of the final score is. As a technical detail, this means that we cannot parametrize the computation of Average Precision by ϵ (since some estimators don't output scores in the range [0, 1]), and must instead parametrize each estimator's AP computation separately by thresholding on that estimator's output."

We have redone the distributional uncertainty experiment to follow standard out-of-distribution (OOD) detection experiments. We train ResNet32 on CIFAR10 (in-distribution) and estimate the distributional competence (Line 241) of images from SVHN (OOD). The results are below in Table 1. We have also revised Table 1 in the paper to have underfit, wellfit, and overfit models for each model. The missing results are denoted below in Table 2. We have omitted RF (O) since our random forest did not overfit. Further experimental details will be in the camera-ready version.

Table 1: mAP for Distributional Uncertainty ($\mathcal{E} = \mathcal{E}_{\mathcal{D}}$).

Split	Softmax	TrustScore	Abl. ALICE	ALICE
10/90	.458 ±0.056	.518 ±0.039	.100 ±0.000	**.868 ±0.014**
30/70	.693 ±0.034	.721 ±0.026	.300 ±0.000	**.946 ±0.007**
50/50	.816 ±0.020	.833 ±0.015	.500 ±0.000	**.970 ±0.003**
70/30	.901 ±0.010	.910 ±0.008	.700 ±0.000	**.985 ±0.002**
90/10	.970 ±0.003	.972 ±0.002	.900 ±0.000	**.997 ±0.001**

Table 2: Continuation of Table 1 from the paper. ($\mathcal{E} = \mathcal{E}_{\text{xent}}$).

Model	Softmax	TrustScore	Abl. ALICE	ALICE
MLP (W)	.989 ±.005	.929 ±.044	.958 ±.042	**.998 ±.001**
MLP (O)	.532 ±.062	.768 ±.064	.576 ±.033	**.996 ±.003**
RF (W)	.998 ±.002	.898 ±.025	.923 ±.016	**.999 ±.000**
SVM (U)	.995 ±.003	.626 ±.046	.496 ±.069	**1.00 ±.000**
SVM (W)	**1.00 ±.000**	.931 ±.048	.963 ±.038	**1.00 ±.000**

We have modified all mentions of PAC Learning to clarify that our method is inspired, not derived, from PAC methods.

We have added to lines 39-41 to articulate prior work and motivate our usage of the three types of uncertainty and the limitations of these works. It now reads: "Previous attempts to explicitly model these three factors require out-of-distribution data, or are not scalable to high dimensional datasets or deep networks [1]".

We have edited the notation to distinguish between a finite "label space" $\mathcal{C}$ and the associated unit simplex, or "distributional space" $\mathcal{Y}$, and have revised parts of the paper to accomodate (e.g. Eq. 0, the unnumbered before Eq. 1).

On Line 70, we remove Eq. 1 and clarify: "The relaxation of the prediction error leads to the generalized notion of δ**-competence**, which we define as $p(\mathcal{E} < \delta | x, \hat{f})$. Confidence can be recovered by setting $\mathcal{E} = \mathcal{E}_{0\text{-}1}$ and $\delta \in (0, 1)$."

We have revised Line 219 to specifically link model confidence and softmax together so that the role of softmax in line 246 is clear (another score that does not require ground truth to compute and can be treated as a competence estimator).

Several other minor comments (clarifying definitions e.g. "pointwise rankings," "inverse true error," "randomness" etc., clarifying ambiguity between calibration and interpretability) have been assimilated.

[1] Malinin, Andrey, and Mark Gales. "Predictive uncertainty estimation via prior networks." Advances in Neural Information Processing Systems. 2018.; Yarin Gal, "Uncertainty in Deep Learning", Ph.D. thesis, University of Cambridge, 2016; Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," International Conference on Learning Representations, 2018.