	OOD K. α_0/var.	OOD K. MI.	OOD F. α_0/var.	OOD F. MI.	OODom K. α_0/var.	OODom K. MI.	OODom F. α_0/var.	OODom F. MI
Ensemble	*97.19±0.0	*97.44±0.0	97.53±0.1	97.69±0.1	42.36±0.3	42.38±0.3	37.85±1.1	37.86±1.1
RKL-PN	54.11±3.4	54.9±3.3	72.54±3.6	73.33±3.5	8.94±0.0	8.94±0.0	8.96±0.0	8.96±0.0
RKL-PN w/ F.	78.4±4.8	78.73±4.8	***100.0±0.0**	***100.0±0.0**	9.08±0.1	9.08±0.1	87.49±5.0	87.49±5.0
PostN	**96.04±0.2**	**96.05±0.2**	98.17±0.2	98.17±0.2	***100.0±0.0**	***100.0±0.0**	***100.0±0.0**	***100.0±0.0**

Table A: OOD detection (MNIST). MI and α_0 (Dirichlet) / variance (Ensemble) results are highly correlated.

	Acc.	Alea. Conf.	Epist. Conf.	Brier	OOD S. Alea.	OOD S. Epist.	OODom S. Alea.	OODom S. Epist.
Ensemble	*91.34±0.0	*99.1±0.0	98.77±0.0	17.69±0.1	*80.1±0.3	75.14±0.2	21.1±3.1	24.42±3.7
RKL-PN	60.05±0.7	85.63±0.8	82.11±1.3	70.84±0.9	50.97±3.9	55.37±4.3	56.16±1.4	51.33±2.4
RKL-PN w/ C100	88.18±0.1	95.44±0.3	94.15±0.3	79.99±2.0	56.67±2.1	73.37±2.3	57.06±1.7	50.31±1.4
PostNet	**90.05±0.1**	**98.87±0.0**	***98.82±0.0**	***15.44±0.1**	**76.04±0.4**	***75.57±0.4**	***87.65±0.3**	***92.13±0.5**

Table B: Results (VGG16) on CIFAR10 on classic split. RKL-PN w/ C100 uses CIFAR100 as training OOD.

Uncertainty metrics (R1). Based on R1's comments we also evaluated the models based on mutual information (Tab. A). MI is highly correlated with both α_0 and variance with barely score changes.

AUROC vs APR (R1). Both metrics have been used by prior works to assess OOD detection performance [20,4,A,B]. Theoretically, the two metrics bring similar information [C]. In practice, APR is preferred when working with imbalanced classes (such as anomaly detection) since AUROC might lead to too optimistic results [D]. For these reasons, we decided to use APR.

Flow on input vs latent space (R1). As shown in existing work, distinguishing between CIFAR10 and SVHN is not trivial [4, 24, E]. We attribute the strong performance of PostNet to the dim. reduction and the classification task (Sect. 2.2). Similar conclusions have been drawn in [E].

PostNet CIFAR10 acc. (R1). PostNet provides both good uncertainty estimates and accuracy (Fig. 5, 7, 8, 9, 10). In our paper we use 5 random splits (60%, 20%, 20%). Based on R1's comments, we also trained on the classic split (79%, 5%, 16%). PostNet achieves $\sim$90% accuracy (Tab. B). Experiments using random splits lead to better estimates of the true model performance. We made a proper comparison focused on small number of classes similar to [20] and enrich experiments with tabular, shifts and OODom settings. We agree that results on more classes are interesting future work.

PriorNet acc. w/o OOD (R1, R3). Specifying training OOD data is unrealistic (l.70-71) and is unlikely to generalize to all other OOD datasets (l.72-74). We demonstrate these issues with practical results (Tab. 3, Tab. B). Indeed, the results of PriorNets deteriorate when the OOD data used at training time (e.g. noise/KMNIST/CIFAR100) differs from the OOD data at test time (e.g. FMNIST/SVHN). Still, we also report results for PriorNet with true OOD on MNIST and CIFAR10, where it obtains $\sim$99% and $\sim$89% accuracy, respectively. This is similar to reported results in [21], ruling out the possibility of under-trained or mis-specified models.

Ensemble baseline (R1, R3). We provide results of Ensemble in Fig. 5, 7, 8, 9, 10 in app. and additionally on CIFAR10 with VGG16 in Tab. B. Ensemble has a high training cost which justified a specific treatment. Note that Tab. 4 aims at comparing models training a single network (l.298), this is why here ensemble is not included. Ensemble achieves good performance except for tabular left-out classes and OODom datasets where PostNet shows substantially better results.

Flow choice (R3). In our experiments (app. Fig. 5, 7, 8, 9, 10), both flows (e.g. PostN Rad (6) and PostN IAF (6)) achieve good performance on the four datasets, even though the flow depth can impact the performance. Using MoG leads to weaker performance. Note that the No-Flow model outputs α which are directly used to compute the Bayesian loss (no likelihood with NF or MoG).

Stronger baselines (R3). We compare PostNet to recent Dirichlet-based SoTA methods (2018 and newer). We also consider Drop-Out and Ensembles, which are strong baselines [20, 21, 33].

Dataset shifts (R4). Fig. 4 shows that PostNet assigns lower confidence to larger dataset shifts (l.264).

Related work (R1, R2, R3). We will include suggestions and correct the misleading EDL statement. In particular, we will explain connections between RKL and the Bayesian loss.

[A] Hendrycks et al. "Deep Anomaly Detection with Outlier Exposure". ICLR 2019.
[B] Hendrycks et al. "A Baseline for Detecting Misclassified & OOD Examples in Neural Networks". ICLR 2017.
[C] Jesse et al. Davis. "The Relationship Between Precision-Recall and ROC Curves". ICML 2006.
[D] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". PloS one 2015.
[E] Kirichenko et al. "Why Normalizing Flows Fail to Detect OOD Data". Arxiv 2020.