Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks

Bálint Mucsányi [1] **Michael Kirchhof** [1] **Seong Joon Oh** [1 2]

Abstract

Uncertainty quantification, once a singular task, has evolved into a spectrum of tasks, including abstained prediction, out-of-distribution detection, and aleatoric uncertainty quantification. The latest goal is disentanglement: the construction of multiple estimators that are each tailored to one and only one source of uncertainty. This paper evaluates a wide spectrum of Bayesian, evidential, and deterministic methods across various uncertainty tasks on ImageNet. We find that, despite promising theoretical endeavors, disentanglement is not yet achieved in practice. Further, we reveal which uncertainty estimators excel at which specific tasks, providing insights for practitioners and guiding future research toward task-centric and disentangled uncertainty estimation methods. Our code is available on GitHub.

1. Introduction

When uncertainty quantification methods were first pioneered for deep learning (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017), their task was simple: giving one total uncertainty estimate. The recent demand for trustworthy machine learning (Mucsányi et al., 2023) created new requirements, mostly centering around disentangling the above predictive uncertainty into aleatoric (data-inherent and irreducible) and epistemic (model-centric and reducible) components (Depeweg et al., 2018; Valdenegro-Toro & Mori, 2022; Shaker & Hüllermeier, 2021). They serve different purposes: epistemic uncertainty is widely used for out-of-distribution detection (Mukhoti et al., 2023), and separated epistemic and aleatoric uncertainty estimates enable tasks like active learning (Lahlou et al., 2023).

One limitation of these recent endeavors is that they are primarily theoretical, supported by only toy or small-scale experiments (Shaker & Hüllermeier, 2021; Van Amersfoort et al., 2020; Mukhoti et al., 2023). Larger-scale benchmarks often evaluate methods w.r.t. only one source of uncertainty and do not test for undesirable side effects on other components (Galil et al., 2023a; Ovadia et al., 2019). While this approach allows for insights into the performance of a subset of methods on a selection of tasks, there is currently no study that evaluates which component(s) each method captures in practice and which it does not.

Our work establishes a comprehensive overview of this vast landscape of methods and tasks. We reimplement sixteen uncertainty quantification methods in up to eight ways and evaluate each on seven practically defined tasks on ImageNet-1k (Deng et al., 2009) and CIFAR-10 (Krizhevsky & Hinton, 2009), ranging from abstained prediction to out-of-distribution detection. We further study if recent uncertainty decomposition formulas separate the total uncertainty into two disentangled components for aleatoric and epistemic uncertainty as theoretically intended (Wimmer et al., 2023; Pfau, 2013; Depeweg et al., 2018). We find that disentanglement is unachieved in practice since most proposed pairs of estimators are highly internally correlated and fail to unmix aleatoric and epistemic uncertainty (Section 3.1). However, there are individual estimators that strongly reflect one type of uncertainty while being almost unrelated to the other (Sections 3.2 and 3.3), revealing a promising path to disentangled estimates in the future.

These findings emphasize the importance of specifying the particular task one wants to solve and developing uncertainty estimators tailored to it. We anticipate that our insights into the practical workings of uncertainty estimators will drive the field of uncertainty quantification toward developing robust and disentangled uncertainty estimators.

2. Benchmarked Methods

This section provides an overview of the benchmarked uncertainty estimators and disentanglement formulas. We reimplement all sixteen benchmarked methods as plug-and-play model wrappers available on Github. Details are provided in Appendix D.

[1]University of Tübingen, Germany [2]Tübingen AI Center, Germany. Correspondence to: Bálint Mucsányi <b dot h dot mucsanyi at gmail dot com>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

2.1. Uncertainty Quantification Methods

We consider a classification setting with a discrete label space of C classes. We benchmark the supervised uncertainty quantification methods studied in Kirchhof et al. (2023b) and evaluate an additional eight methods to encourage diversity and general applicability of our findings. In total, we study sixteen uncertainty quantification methods and provide a comprehensive evaluation of uncertainty disentanglement that is simple to extend with new methods. The benchmarked approaches can be categorized into distributional and deterministic methods.

2.1.1. Distributional methods

Distributional methods model a second-order predictive distribution $q(f(x) \mid x)$ over class probability vectors, abbreviated as $q(f)$.

Spectral-normalized Gaussian processes (SNGP) (Liu et al., 2020) represent the $q(f)$ distributions by approximating a Gaussian process (GP) over the classifier *output* aided by spectral normalization. We also benchmark the last-layer GP without spectral normalization. The last-layer **Laplace approximation** (Daxberger et al., 2021) models a Gaussian parameter distribution in a post-hoc fashion that induces the $q(f)$ distributions. Similarly, **latent heteroscedastic classifiers (HET-XL)** (Collier et al., 2023) predict a heteroscedastic Gaussian distribution over the pre-logit *embeddings*. Evidential deep learning methods for classification (Sensoy et al., 2018a; Charpentier et al., 2020a) directly learn a Dirichlet distribution over the output probability vectors. Following (Ulmer et al., 2023), we refer to the method of Sensoy et al. (2018a) as **EDL** and that of Charpentier et al. (2020a) as **PostNet**.

MC-Dropout (Gal & Ghahramani, 2016; Srivastava et al., 2014) and **deep ensembles** (Lakshminarayanan et al., 2017) do not construct second-order predictive distributions $q(f)$ explicitly. Instead, they sample from them by M repeated forward passes with randomly switched off activations or by training M models, respectively. **Shallow ensembles** (Lee et al., 2015) are lightweight approximations of deep ensembles with a shared backbone and M output heads.

Practical tasks like threshold-based rejection often need a scalar uncertainty output instead of a second-order predictive distribution $q(f)$. To this end, **uncertainty aggregators** compile the above distributions into scalar uncertainty estimates $u(x) \in \mathbb{R}$. Several methods exist for this aggregation, e.g., calculating the Bayesian Model Average (BMA) $\tilde{f}(x) := \mathbb{E}_{q(f)}[f(x)]$ and using its entropy as the uncertainty estimate $u(x)$ or quantifying the variance of $q(f)$. We consider eight aggregators detailed in Appendix F and, unless stated otherwise, use the best-performing one for each distributional method in the practical benchmarks. The effect of the chosen aggregators is studied in detail in Appendix G.

2.1.2. Deterministic Methods

Deterministic methods (Postels et al., 2022) directly output scalar uncertainty estimates $u(x)$ instead of modeling a second-order predictive distribution $q(f)$ over class probability vectors. **Loss prediction** (Yoo & Kweon, 2019; Lahlou et al., 2023; Kirchhof et al., 2023b) employs an additional MLP head for $u(x)$ that estimates the loss of the prediction $f(x)$, predicting a notion of (in-)correctness. **Correctness prediction** is a special variant for classification where $u(x)$ predicts how likely the predicted class is to be the correct class. **Deterministic uncertainty quantification (DUQ)** (Van Amersfoort et al., 2020) learns a latent density on the training set and outputs as $u(x)$ how close an input's embedding is to the mixture means. The **Mahalanobis** method (Lee et al., 2018) builds a similar latent mixture of Gaussians in a post-hoc fashion. The DDU method (Mukhoti et al., 2023) combines the spectral normalization of SNGPs with the latent density of the Mahalanobis method. **Temperature scaling** (Guo et al., 2017) post-hoc calibrates the predicted probability vectors with a temperature scalar. As a **baseline**, we use a deterministic single-point network trained with the cross-entropy loss.

2.2. Uncertainty Decomposition Formulas

So far, we only considered uncertainty estimators that (sometimes after aggregating) output a single $u(x)$. A second strain of literature outputs not only one estimate but decomposes the $q(f)$ of distributional methods into multiple estimators that each quantify one source of uncertainty, such as epistemic (caused by a lack of data, reducible) and aleatoric uncertainty (inherent in the generative process, irreducible) (Hora, 1996; Mucsányi et al., 2023). The estimators for each source should be disentangled: the aleatoric estimator should only capture aleatoric uncertainty, and the epistemic estimator should only reflect epistemic uncertainty. See Appendix J for more details. We benchmark two prominent approaches to obtain such pairs of estimators: the information-theoretical (IT) (Depeweg et al., 2018; Mukhoti et al., 2023; Wimmer et al., 2023) and the Bregman decomposition (Pfau, 2013; Gupta et al., 2022; Gruber & Buettner, 2023).

The IT decomposition decomposes the entropy of the predictive distribution $p(y \mid x) = \int p(y \mid x, f)\, dq(f)$ into an aleatoric and an epistemic component:

$$\underbrace{\mathbb{H}_{p(y|x)}(y)}_{\text{predictive}} = \underbrace{\mathbb{E}_{q(f)}\left[\mathbb{H}_{p(y|x,f)}(y)\right]}_{\text{aleatoric}} + \underbrace{\mathbb{I}_{p(y,f|x)}(y; f)}_{\text{epistemic}}, \quad (1)$$

where $\mathbb{H}_{p(y|x)}(y)$ is the entropy and $\mathbb{I}_{p(y,f|x)}(y; f)$ is the mutual information. Intuitively, the aleatoric component represents the spread of the labels that the plausible models

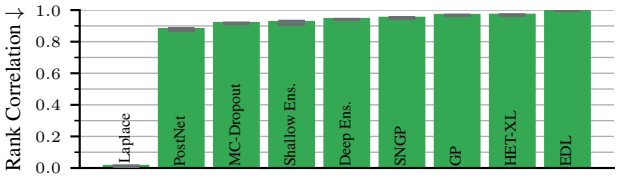


Figure 1. Eight out of nine distributional methods exhibit a severely high rank correlation (≥ 0.88) between the IT aleatoric and epistemic components when evaluated on ImageNet-ReaL. These methods violate a necessary condition of uncertainty disentanglement.

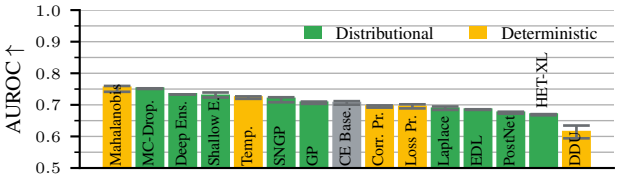


Figure 2. Distinguishing ID from OOD samples as a proxy for which uncertainties reflect epistemic uncertainty. OOD samples are perturbed by ImageNet-C corruptions of severity two. Mahalanobis, the best method, is a specialized estimator for this task.

in the posterior have on average, whereas the epistemic component only captures the disagreement of the predictions $p(y \mid x, f)$ among the models f.

The definition of the Bregman decomposition and corresponding disentanglement results are shown in Appendix E.

3. Experiments

With these different estimators at hand, we now investigate our main research questions: Does any approach give disentangled uncertainty estimators in practice? Furthermore, what type of uncertainty does each estimator capture in terms of practical tasks? The following sections answer these questions, with their titles highlighting the key observations of our experiments.

We reimplement and train each approach for 50 ImageNet-1k (Deng et al., 2009) epochs on a pretrained ResNet-50 with a training pipeline following Tran et al. (2022). Since the DUQ method has memory and stability issues on ImageNet, in Appendix C, we repeat experiments on CIFAR-10 (Krizhevsky & Hinton, 2009) with the Wide ResNet 28-10 architecture, following Liu et al. (2020). We track the validation performance to conduct early stopping and choose hyperparameters. We report mean, minimum, and maximum performance across five seeds. The benchmark took 1.5 GPU years on RTX 2080 Ti GPUs.

3.1. Uncertainty disentanglement often fails

We first study if the IT and Bregman decompositions yield disentangled estimators. Since they decompose the second-order predictive distributions $q(f)$ of distributional methods, deterministic methods are excluded in this section.

Fig. 1 reveals a simple failure: for six of the seven distributional methods, the IT decomposition leads to highly correlated aleatoric and epistemic estimates (rank corr. ≥ 0.92). While one should not expect estimators with a perfect decorrelation of zero, this by far exceeds the achievable correlation level we find in Appendix E.2. The correlation remains for Bregman decompositions (Appendix E) and also does not considerably lower when we add more epistemic uncertainty into the dataset (Appendix L.1). These severe internal correlations prohibit the estimators from capturing semantically different sources of uncertainty.

The only exception is the IT-decomposed Laplace method, whose aleatoric and epistemic estimators are entirely decorrelated. However, in Appendix L.2 we show that its epistemic estimator predicts epistemic uncertainty at a chance level. Therefore, according to the definition in Appendix J, no benchmarked method gives disentangled uncertainties.

3.2. OOD-ness is hard to detect

Let us continue with testing which estimators represent epistemic uncertainty, captured by an out-of-distribution (OOD) detection task (Gruber & Buettner, 2023; Mukhoti et al., 2023). We use ImageNet-C (Hendrycks & Dietterich, 2019) with corruptions of severity level two as OOD data. This is far enough out-of-distribution to deteriorate the accuracy by 26%, see also Appendix B.2. We measure via a binary classification AUROC if uncertainty estimates are higher on these OOD samples than on ID samples. From this point forward, we also consider deterministic methods.

As shown in Fig. 2, the best OOD detection, and thus the highest alignment with epistemic uncertainty, is achieved by the Mahalanobis method. This is a method developed specifically for OOD detection. It is also trained specifically for corruptions of severity two, and its advantage vanishes already when using OOD samples of severity three (Appendix M.3). The distance-awareness property of DDU – also following the latent density intuition – induced by spectral normalization does not seem a powerful enough regularizer to tell ID and OOD samples apart (AUROC $= 0.62$). This also highlights that the adversarial perturbations the Mahalanobis method employs are crucial for its performance and that uncertainty estimators must be specifically tailored to the task a practitioner intends to use them for rather than relying on high-level intuitions.

3.3. Aleatoric uncertainty alone is hard to quantify

The previous experiment isolated the epistemic capabilities of uncertainty estimates. Let us now evaluate how well the

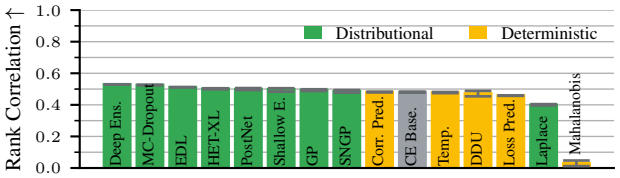


Figure 3. Rank correlation of uncertainty estimators and the GT aleatoric uncertainty on ImageNet, showing which uncertainties reflect aleatoric uncertainty. The entropy of the ImageNet-ReaL labels is used as GT aleatoric uncertainty.

benchmarked models predict aleatoric uncertainty. We follow Tran et al. (2022) and Kirchhof et al. (2023a;b) and use the disagreement of human annotators as ground truths for aleatoric uncertainty. ImageNet-ReaL (Beyer et al., 2020) queries multiple annotators for labels on each image. We showcase some examples in Appendix O. We use the entropy of the input-conditional soft-label distributions as GT aleatoric uncertainties.

To evaluate aleatoric uncertainty estimates, we calculate the rank correlation between the estimates and the GT label entropies across all images. Fig. 3 shows that most distributional methods perform above the cross-entropy baseline. Deep ensembles are most aligned with human uncertainties on average. On the other side of the spectrum, the Mahalanobis method is almost uncorrelated with aleatoric uncertainty. This is in fact a strength: Mahalanobis estimates reflect epistemic uncertainty while being independent from aleatoric uncertainty. Combining this with a second estimator for aleatoric uncertainty can pave the way for disentangled uncertainties. As a simple start, combining it with the CE baseline achieves a low rank correlation of 0.15 ± 0.01 between the two. We see this as a promising pathway to disentangled uncertainty estimators in the future.

3.4. Different tasks require different methods

The previous sections suggest that uncertainty tasks are not all solved by the same best methods. In this section, we investigate how correlated the performance of methods on different tasks is. In particular, we use the previous two tasks along with six more and measure the between-task Pearson correlations of the performance of all benchmarked methods (see also Appendix F). Rank correlation results are similar, see Appendix M.7.

Fig. 4 shows two clusters of metrics. The Brier score and log probability proper scoring rules, coupled with the ECE metric, are all recognized as *predictive uncertainty* metrics in the literature (Mucsányi et al., 2023). The high rank correlation among these metrics (rank corr. $\in [0.86, 0.94]$) evidences this claim empirically. The second cluster is the accuracy, abstinence, aleatoric uncertainty triad: interestingly, methods that capture human uncertainty well are also the most accurate, even though models are trained without access to human uncertainty soft labels. OOD detection is not in any of these clusters, underlining that it benchmarks a different type of uncertainty: epistemic uncertainty.

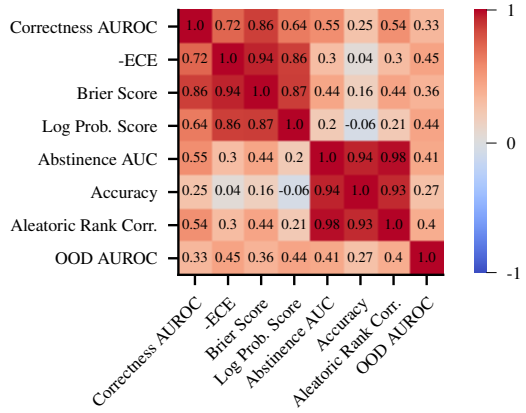


Figure 4. Pearson correlation of metric pairs across all methods and aggregators. Only some of the considered metrics have a very high correlation among methods on the ImageNet validation dataset: most capture different aspects of uncertainty methods. Spearman correlation results are similar, see Fig. M.13.

4. Limitations

Our benchmark only considers classification and two datasets as these are the only larger-scale datasets with a ground-truth notion of aleatoric uncertainties, i.e., multiple labels. Further aleatoric uncertainty ground-truths are an ongoing effort (Schmarje et al., 2022; Collins et al., 2022). Once such datasets are available, future research should aim to expand our uncertainty disentanglement investigations for uncertainties in regression (Upadhyay et al., 2023) or unsupervised learning (Kirchhof et al., 2023a).

5. Conclusion

We study how widely used uncertainty estimators and decomposition formulas perform on a comprehensive set of uncertainty quantification tasks. Our findings encourage a pragmatic reassessment of uncertainty quantification research. There is no general uncertainty; instead, uncertainty quantification covers a spectrum of tasks where the definition of the exact task heavily influences the optimal method and performance. Such a precise definition of tasks per estimator would also benefit constructing disentangled uncertainties. This shift could lead to the alignment of theoretical developments and intuitive descriptions about what particular types of uncertainty different methods (or aggregators) aim to capture, with tangible improvements on the benchmark tasks we consider.

References

Bengs, V., Hüllermeier, E., and Waegeman, W. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning (ICML)*, 2023.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1356–1367. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0eac690d7059a8de4b48e90f14510391-Paper.pdf.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33:1356–1367, 2020b.

Collier, M., Jenatton, R., Mustafa, B., Houlsby, N., Berent, J., and Kokiopoulou, E. Massively scaling heteroscedastic classifiers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sIoED-yPK9l.

Collins, K. M., Bhatt, U., and Weller, A. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 10, 2022.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193. PMLR, 2018.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Galil, I., Dabbah, M., and El-Yaniv, R. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023a.

Galil, I., Dabbah, M., and El-Yaniv, R. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers? In *International Conference on Learning Representations (ICLR)*, 2023b.

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., and Kauermann, G. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

Gruber, S. and Buettner, F. Uncertainty estimates of predictions via a general bias-variance decomposition. In *International Conference on Artificial Intelligence and Statistics*, pp. 11331–11354. PMLR, 2023.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Gupta, N., Smith, J., Adlam, B., and Mariet, Z. Ensembling over classifiers: a bias-variance perspective. *arXiv preprint arXiv:2206.10566*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hora, S. C. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223, 1996.

Kirchhof, M., Kasneci, E., and Oh, S. J. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. *International Conference on Machine Learning (ICML)*, 2023a.

Kirchhof, M., Mucsányi, B., Oh, S. J., and Kasneci, E. URL: A representation learning benchmark for transferable uncertainty estimates. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023b. URL https://openreview.net/forum?id=e9n4JjkmXZ.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Lahlou, S., Jain, M., Nekoei, H., Butoi, V. I., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=eGLdVRvvfQ. Expert Certification.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mucsányi, B., Kirchhof, M., Nguyen, E., Rubinstein, A., and Oh, S. J. Trustworthy machine learning. *arXiv preprint arXiv:2310.08215*, 2023.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Nado, Z., Band, N., Collier, M., Djolonga, J., Dusenberry, M., Farquhar, S., Filos, A., Havasi, M., Jenatton, R., Jerfel, G., Liu, J., Mariet, Z., Nixon, J., Padhy, S., Ren, J., Rudner, T., Wen, Y., Wenzel, F., Murphy, K., Sculley, D., Lakshminarayanan, B., Snoek, J., Gal, Y., and Tran, D. Uncertainty Baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9617–9626, 2019.

Pfau, D. A generalized bias-variance decomposition for bregman divergences. *Unpublished Manuscript*, 2013.

Postels, J., Segù, M., Sun, T., Sieber, L. D., Van Gool, L., Yu, F., and Tombari, F. On the practicality of deterministic epistemic uncertainty. In *International Conference on Machine Learning (ICML)*, 2022.

Schmarje, L., Grossmann, V., Zelenka, C., Dippel, S., Kiko, R., Oszust, M., Pastell, M., Stracke, J., Valros, A., Volkmann, N., et al. Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232, 2022.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018b.

Shaker, M. H. and Hüllermeier, E. Ensemble-based uncertainty quantification: Bayesian versus credal inference. In *PROCEEDINGS 31. WORKSHOP COMPUTATIONAL INTELLIGENCE*, volume 25, pp. 63, 2021.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Tran, D., Liu, J. Z., Dusenberry, M. W., Phan, D., Collier, M., Ren, J., Han, K., Wang, Z., Mariet, Z. E., Hu, H., Band, N., Rudner, T. G. J., Nado, Z., van Amersfoort, J., Kirsch, A., Jenatton, R., Thain, N., Buchanan, E. K., Murphy, K. P., Sculley, D., Gal, Y., Ghahramani, Z., Snoek, J., and Lakshminarayanan, B. Plex: Towards reliability using pretrained large model extensions. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=6x0gB9gOHFg.

Ulmer, D., Hardmeier, C., and Frellsen, J. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=xqS8k9E75c.

Upadhyay, U., Kim, J. M., Schmidt, C., Schölkopf, B., and Akata, Z. Likelihood annealing: Fast calibrated uncertainty for regression. *arXiv preprint arXiv:2302.11012*, 2023.

Valdenegro-Toro, M. and Mori, D. S. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Von Luxburg, U. and Schölkopf, B. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pp. 651–706. Elsevier, 2011.

Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations (ICLR)*, 2022.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pp. 2282–2292. PMLR, 2023.

Yang, Y., Mandt, S., Theis, L., et al. An introduction to neural data compression. *Foundations and Trends® in Computer Graphics and Vision*, 15(2):113–200, 2023.

Yoo, D. and Kweon, I. S. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Procedings of the British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

A. Related Works

This section discusses previous quantitative uncertainty studies and connects their findings to this work.

Disentanglement. The decomposition of aleatoric and epistemic uncertainties (Pfau, 2013; Depeweg et al., 2018) has recently been shown to have failure cases (Wimmer et al., 2023; Bengs et al., 2023; Gruber et al., 2023; Valdenegro-Toro & Mori, 2022). These papers make their claims theoretically or via toy problems like binary classification or 1D regression. Our results support this discussion with a practical and quantitative perspective. We find that uncertainty decompositions do not work in general and, if at all, depend greatly on the underlying method. Our findings thus encourage taking a more holistic view of disentanglement, including decomposition formulas, methods, and implementations. We further find that instead of relying on decomposition formulas, it is promising to combine separate methods, such as the CE baseline's predictive entropy and the Mahalanobis values, where each method handles a specific type of uncertainty, similar to Mukhoti et al. (2023).

Sensitivity to implementation. The finding that the performance of uncertainty estimators depends greatly on their implementation and that different design choices are better for different metrics is in line with recent benchmarks (Galil et al., 2023b; Kirchhof et al., 2023b). Our benchmark further shows that the aggregator function of distributional methods is a crucial component and that simply averaging outputs can be inferior to, e.g., averaging in the dual space (Gupta et al., 2022).

Robustness. Recent benchmarks on OOD detection and robustness (Nado et al., 2021; Ovadia et al., 2019; Postels et al., 2022; Galil et al., 2023a) have first highlighted robustness issues of uncertainty estimates. Our benchmark supports these findings on CIFAR-10, especially in the region where the OOD-ness is only slight yet already causes degradation of both the main task and the uncertainty estimator. The latter implies that uncertainty estimators either need to become more robust to distribution shifts (Kirchhof et al., 2023b) or be better able to detect subtle epistemic uncertainties. However, our experiments on ImageNet do not show robustness issues, which might be attributed to the vast space of natural images the ImageNet training dataset covers, leading to robustness. This further highlights the importance of the used dataset's properties.

Aleatoric uncertainty. While epistemic uncertainty is widely evaluated on the OOD detection proxy task, aleatoric uncertainty still lacks a standardized testing protocol. The current approaches seem to converge to soft labels, but nuances in how they are collected still need discussion (compare, for example, CIFAR-10H (Peterson et al., 2019) to CIFAR-10S (Collins et al., 2022) and CIFAR-10N (Wei et al., 2022)). An increasing number of uncertainty quantification approaches compare to such human GT notions of aleatoric uncertainty (Tran et al., 2022; Kirchhof et al., 2023a;b), indicating the interest in the field. Our benchmark shows that no method can give highly accurate aleatoric uncertainty estimates yet, stressing the need for benchmarks, methods, and training resources to develop along.

Predictive uncertainty and calibration. Contrary to aleatoric uncertainty alignment, calibration and predictive uncertainty benchmarks are starting to become saturated and, according to our experiments, the top performers are ready for deployment. This corroborates recent findings by Galil et al. (2023b). In comparison to this benchmark that compared model architectures, we compared sixteen different approaches on the same backbone.

B. Further Main Practical Results on ImageNet

B.1. Correctness prediction works across the board

This section shows the results of uncertainty quantification methods on additional popular tasks: correctness prediction and abstained prediction.

Let us now broaden the view beyond disentanglement to benchmark how well uncertainty estimators solve other practically relevant tasks. We start with correctness prediction, where the AUROC quantifies whether wrong predictions generally have higher uncertainties than correct predictions. Fig. B.1 shows that most uncertainty estimators perform within ± 0.014 of the cross-entropy baseline when predicting correctness. Modern methods like HET-XL do not outperform older methods like deep ensembles or MC-Dropout. Evidential deep learning methods perform significantly better than the rest of the methods. We see similar results when slightly altering the correctness metric to account for soft labels in Appendix M.1.

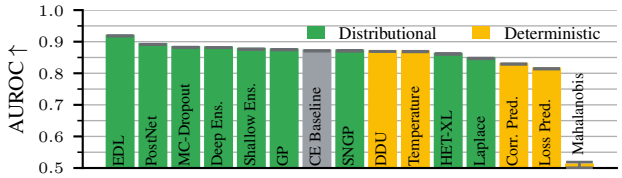


Figure B.1. ID correctness prediction results measured by the AUROC w.r.t. model correctness on the ImageNet validation set. The performance of methods on predicting correctness is saturated around the cross-entropy baseline, with most methods being within a 0.023 AUROC band. The evidential deep learning methods, EDL and PostNet, capture predictive uncertainty remarkably well. The Mahalanobis method is a specialized OOD detector that cannot differentiate between ID samples.

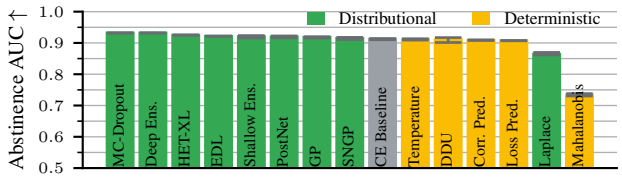


Figure B.2. ID abstained prediction results using the abstained accuracy AUC metric on the ImageNet validation dataset. All benchmarked methods apart from Mahalanobis and Laplace perform very well on the abstinence task (AUC ≥ 0.9). Similarly to correctness prediction in Fig. B.1, methods are saturated, with almost all of them being within a 0.03 AUC band. MC-Dropout and deep ensembles stand out as the best-performing methods (AUC ≥ 0.93).

A related task to correctness prediction called abstained prediction involves excluding the predictions for the $x\%$ most uncertain examples and calculating the model's accuracy on the remaining samples. We measure the Area Under Curve (AUC) for a curve that plots accuracy against increasing fractions of abstained samples from 0% to 100% on the $x \in \mathcal{X}$-axis. Informally, the AUC shows how many errors are prevented by removing supposedly uncertain samples. Fig. B.2 shows that the saturation is just as pronounced on the abstained prediction task. All uncertainty methods apart from Mahalanobis and Laplace obtain an AUC score greater than 0.91. Practically, this means that one can obtain a high classification accuracy by utilizing most evaluated uncertainty quantifiers to abstain from prediction on a tiny set of uncertain samples. While computationally expensive distributional methods have a slight edge, the cheaper deterministic methods also give considerable performance.

B.2. Uncertainties can generalize well to OOD settings

A necessary condition for the reliable deployment of uncertainty quantification methods is that their estimates remain effective when facing uncertain inputs. We test this by checking if their previous abstinence and correctness performances are preserved longer than the model's accuracy when increasing the OOD perturbation level. Only then can we trust them and, e.g., base the abstinence from prediction on these uncertainty estimates.

Fig. B.3 shows the correctness prediction AUROC, abstinence AUC, and model accuracy as we increasingly perturb the ImageNet validation samples and go OOD. The results show an almost constant correctness prediction curve, whereas the accuracy degrades to almost 20% at severity level five. Abstained prediction performance degrades together with accuracy, which is a fundamental property of the metric itself since the area under the accuracy is lower-bounded by the baseline accuracy. The AUC gain (i.e., AUC $-$ Accuracy) increases with the severity, showing that the uncertainty estimators even become better on the abstinence task as the severity increases. The tendencies are maintained when we normalize the metrics (solid lines) according to their random predictive performance (see Appendix M.2 for details). This holds for all methods except Mahalanobis, see Appendix M.2. This observation underlines the trustworthiness of existing uncertainty quantification methods on ImageNet OOD correctness prediction.

C. Conclusions do not always transfer among datasets

In this section, we offer a word of caution for using small-scale results as a proxy for large-scale performance. Appendix K repeats all experiments of the main paper and Appendix B on CIFAR-10, which is widely used in the uncertainty quantification literature (Van Amersfoort et al., 2020; Mukhoti et al., 2023; Gruber & Buettner, 2023) but can sometimes lead to

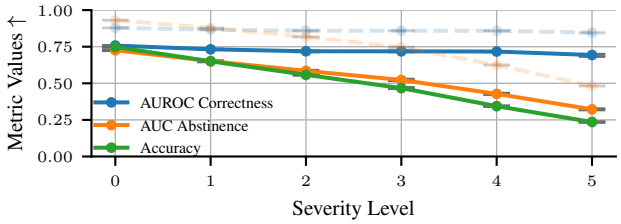


Figure B.3. Degradation of correctness prediction, abstained prediction, and accuracy metrics w.r.t. an increasing level of OOD-ness on the ImageNet validation dataset. As the images become more OOD due to ImageNet-C corruptions, the performance of uncertainty methods on correctness prediction degrades much slower than the model's accuracy. The shown MC-Dropout results are typical for all but the Mahalanobis method. Solid lines: metrics normalized to the $[0, 1]$ range w.r.t. the random predictor of the task. Dashed lines: unnormalized values.

Table C.1. Rank correlations of method rankings on different metrics for all combinations of methods and aggregators between CIFAR-10 and ImageNet. The rankings of approaches are considerably different between these two datasets.

Metric	Rank Corr. CIFAR-10 vs ImageNet
Correctness AUROC	0.622
Abstinence AUC	0.728
Log Prob. Score	0.479
Brier Score	0.428
Aleatoric Rank Corr.	0.245
ECE	0.482
OOD AUROC	0.018
Accuracy	0.531

conclusions not in line with the larger-scale ImageNet. We share some key differences below.

Disentanglement. Most methods show an almost perfect rank correlation between the components of the IT decomposition (see Fig. 1 for ImageNet and Fig. K.1 for CIFAR-10). On ImageNet, Laplace is the only method that *decorrelates* its IT epistemic and aleatoric estimates but its IT epistemic estimator predicts at a chance level – none of the methods achieve *disentanglement* with the IT decomposition. On CIFAR-10, Laplace has almost perfectly correlated estimates and while the SNGP-variants showcase low correlation, they notably underperform the cross-entropy baseline (Fig. L.4).

Aleatoric uncertainty. On CIFAR-10, methods are less aligned with human uncertainties (best rank corr. 0.40 vs 0.54 on ImageNet, Appendix K.5). This is peculiar, as the vast number of classes on ImageNet could introduce more noise into the human soft labels. For example, human labelers might unanimously recognize that the object in question is a dog, yet there may be variations in opinion regarding its specific breed. This indicates that the way soft labels are obtained, which is different in ImageNet-ReaL and CIFAR-10H (Beyer et al., 2020; Hendrycks & Dietterich, 2019), has a significant impact on the results.

Robustness. Correctness predictors are much more robust on ImageNet than on CIFAR-10, even though the drop in accuracy is very similar. Unlike on ImageNet, where the uncertainty estimators maintain a close to constant performance in predicting correctness as we go OOD (Fig. B.3), on CIFAR-10, correctness estimators deteriorate together with the model's accuracy (Fig. K.4). While robustness appears as a striking problem on CIFAR-10, it gets resolved by scaling to a larger dataset that covers the image manifold better.

Method rankings. The above discrepancies also influence the rankings of the approaches. Table C.1 shows the correlation of rankings on CIFAR-10 and ImageNet. Six out of eight metrics have substantially different rankings (rank. corr ≤ 0.531). This indicates that performance on CIFAR-10 should not be taken as an estimate for ImageNet performance.

These experiments underline that methods might show substantially different behaviors on large-scale datasets. We encourage, as best practice, to first scale the approaches to the final deployment domain (and define a precise task) instead of making fundamental design choices on toy datasets.

D. Details of Benchmarked Methods

We consider a classification setting with discrete label space $\{1, \dots, C\}$ of C classes and models that output a probability vector $f(x) \in \Delta^{C-1}$ for input $x \in \mathcal{X}$. The (pre-softmax) logits of the models are denoted by $\log \hat{f}(x)$.

We evaluate two classes of methods: direct prediction methods and distributional methods.

D.1. Direct Prediction Methods

Direct prediction methods output an uncertainty estimate $u(x)$ for input $x \in \mathcal{X}$, such as the estimated probability of the model's prediction to be correct.

D.1.1. Loss Prediction

Loss prediction (Upadhyay et al., 2023; Lahlou et al., 2023; Kirchhof et al., 2023b) employs an additional output head u^{lp} connected to the pre-logit layer that predicts the loss of the network's prediction on input $x \in \mathcal{X}$. The loss predictor head is trained in a supervised fashion by making $u^{\text{lp}}(x)$, the predicted loss, closer to the actual loss $\ell(f(x), y) = -\log \hat{f}_y(x)$. Precisely, we use the objective

$$\mathcal{L}^{\text{lp}} = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_{y_i}(x_i) + \lambda \left(u^{\text{lp}}(x_i) + \log \hat{f}_{y_i}(x_i) \right)^2, \tag{2}$$

where the risk predictor loss (squared Euclidean distance) is traded off with the label predictor loss (cross-entropy) with a hyperparameter $\lambda \in \mathbb{R}_+$.

Note that $Y \mid X = x$ is a random variable in the presence of aleatoric uncertainty. In expectation, Eq. (2) encourages $u^{\text{lp}}(x)$ to approximate the true *pointwise risk* $\mathcal{R}(f, x) = \mathbb{E}_{p^*(y|x)} \left[\ell(f(x), y) \right]$ at each input $x \in \mathcal{X}$.

D.1.2. Correctness Prediction

Correctness prediction is a variant of risk prediction that, instead of aiming to predict the risk of the network on input $x \in \mathcal{X}$, predicts the probability of correctness $p\left(\arg\max_{c \in \{1,\dots,C\}} f_c(x) = y \;\middle|\; x \right)$ on input $x \in \mathcal{X}$. This is achieved by using a sigmoid correctness predictor head h and using the objective

$$\mathcal{L}^{\text{cp}} = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_{y_i}(x_i) - \lambda \left(l_i \log h(x_i) + (1 - l_i) \log(1 - h(x_i)) \right), \tag{3}$$

where $l = \mathbf{I}\left[\arg\max_{c \in \{1,\dots,C\}} \log \hat{f}_c(x) = y \right] \; \forall i \in \{1, \dots, n\}$, and the correctness predictor loss (binary cross-entropy) is traded off with the label predictor loss (cross-entropy) with a hyperparameter $\lambda \in \mathbb{R}_+$. The uncertainty estimate is $u^{\text{cp}}(x) = 1 - h(x)$ (i.e., the probability of making an error).

D.1.3. Deterministic Uncertainty Quantification

The deterministic uncertainty quantification (DUQ) method of Van Amersfoort et al. (2020) learns a latent mixture-of-RBF density on the training set with a strictly proper scoring rule to capture the uncertainty in the prediction based on the Euclidean distance of the input's embedding to the mixture means. The training objective is

$$\mathcal{L}^{\text{duq}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C} \text{one\_hot}(y_i)_c \log K_c(x_i) + (1 - \text{one\_hot}(y_i)_c) \log(1 - K_c(x_i)), \tag{4}$$

where $K_c(x) = \exp\left(-\frac{1}{2\gamma} \left\| \log \hat{f}(x) - m_c \right\| \right)$ is the RBF value corresponding to class $c \in \{1, \dots, C\}$ identified by its mean vector m_c in the latent space. To facilitate minibatch training, Van Amersfoort et al. (2020) employ an exponential

moving average (EMA) to learn the mean vector using the following update rules:

$$n_c \leftarrow \gamma \cdot n_c + (1-\gamma)|\mathcal{B}_c| \tag{5}$$

$$M_c \leftarrow \gamma \cdot M_c + (1-\gamma) \sum_{(x,y)\in\mathcal{B}_c} W_c \log \hat{f}(x) \tag{6}$$

$$m_c \leftarrow \frac{M_c}{n_c}, \tag{7}$$

where $\mathcal{B}$ is a minibatch of samples and $\mathcal{B}_c = \{(x,y) \in \mathcal{B} \mid y = c\}\ \forall c \in \{1,\dots,C\}$. γ is the EMA parameter and W_c characterizes a linear mapping of the logits for each class.

To regularize the latent density and prevent feature collapse, Van Amersfoort et al. (2020) use the following gradient penalty added to $\nabla_\theta \mathcal{L}$:

$$\lambda \cdot \left(\left\| \nabla_x \sum_{c=1}^{C} K_c \right\|_2^2 - 1 \right)^2 \tag{8}$$

Each RBF component in the latent space corresponds to one class. The confidence output of the method is the maximal RBF value of the input over all classes. Therefore, the *uncertainty* estimate can be calculated as $u^{\text{duq}}(x) = 1 - \max_{c\in\{1,\dots,C\}} K_c(x)$.

The predicted class of the trained network is $\arg\max_{c\in\{1,\dots,C\}} K_c(x)$.

D.1.4. MAHALANOBIS

The Mahalanobis method (Lee et al., 2018) builds a post-hoc latent density for the training set in the latent space by calculating per-class means and covariances, and using the induced mixture-of-Gaussians as the latent density estimate. Such latent densities are estimated in multiple layers of the network. One layer's confidence estimate is the maximal Mahalanobis score (Gaussian log-likelihood) $K_\ell(x)$ over all classes:

$$K_{\ell,c}(x) = -\left(f_\ell(x) - \mu_{\ell,c}\right)^\top \Sigma_\ell^{-1} \left(f_\ell(x) - \mu_{\ell,c}\right) \tag{9}$$

$$K_\ell(x) = \max_{c\in\{1,\dots,C\}} K_{\ell,c}(x), \tag{10}$$

where f_ℓ is the ℓ-th layer's output,

$$\mu_{\ell,c} = \frac{1}{n_c} \sum_{i=1}^{n} \mathbf{I}\left[y_i = c\right] f_\ell(x_i) \tag{11}$$

is the centroid of the Gaussian for class $c \in \{1,\dots,C\}$ in layer $\ell \in \{1,\dots,L\}$, n_c is the number of samples with label c, and

$$\Sigma_\ell = \frac{1}{n} \sum_{c=1}^{C} \sum_{i=1}^{n} \mathbf{I}\left[y_i = c\right] \left(f_\ell(x) - \mu_{\ell,c}\right) \left(f_\ell(x) - \mu_{\ell,c}\right)^\top \tag{12}$$

is the tied covariance matrix used for all classes in layer $\ell \in \{1,\dots,L\}$.

To make the differences of latent embeddings of ID and OOD samples more pronounced, all samples are adversarially perturbed w.r.t. the maximal Mahalanobis score for each layer's confidence score:

$$\hat{x}^{(\ell)} = x + \epsilon \operatorname{sgn}\left(\nabla_x K_\ell(x)\right). \tag{13}$$

This perturbed sample is used to compute $K_\ell(\hat{x}^{(\ell)})$. Finally, a logistic regression OOD detector is learned on a held-out validation set of a balanced mix of ID and OOD samples to learn weights w_ℓ for each layer $\ell \in \{1,\dots,L\}$ using the L-dimensional inputs $\left[K_1\left(\hat{x}^{(1)}\right),\dots,K_L\left(\hat{x}^{(L)}\right)\right]^\top$. The final *uncertainty* estimate becomes $u^{\text{Mah}}(x) = \sum_{\ell=1}^{L} w_\ell K_\ell(\hat{x}^{(\ell)})$.

This is the only method in our benchmark that requires a mixed ID-OOD validation set for training the logistic regression OOD detector.

D.1.5. TEMPERATURE SCALING

Temperature scaling (Guo et al., 2017) post-hoc calibrates the predictive softmax distribution $f(x)$ by learning a temperature parameter $\tau \in \mathbb{R}_+$ on a held-out ID validation set after training and setting $f(x) := \text{softmax}\left(\log \hat{f}(x)/\tau\right)$. Guo et al. (2017) show that temperature scaling leads to improvements on both the ECE score and strictly proper scoring rules. To determine the optimal τ, we perform a grid search over $\tau \in \{0.1, 0.2, 0.3, \ldots, 10.1\}$ and choose the one that leads to the lowest NLL loss, following (Mukhoti et al., 2023).

D.1.6. DEEP DETERMINISTIC UNCERTAINTY

The Deep Deterministic Uncertainty (DDU) method (Mukhoti et al., 2023) applies the spectral normalization of SNGPs (Appendix D.2.1) to the hidden weights to establish a distance-aware latent space. It then fits a Mixture-of-Gaussians to this latent space based on (ID) training set statistics. Unlike the Mahalanobis method, it

1. does not use adversarial perturbations;
2. only builds a latent density in the pre-logit layer;
3. does not tie the covariance matrix across classes:

$$\pi_c = \frac{n_c}{n}; \tag{14}$$

$$\mu_c = \frac{1}{n_c} \sum_{i=1}^{n} \mathbf{I}\left[y_i = c\right] f_{L-1}(x_i); \tag{15}$$

$$\Sigma_c = \frac{1}{n_c - 1} \sum_{i=1}^{n} \mathbf{I}\left[y_i = c\right] \left(f_{L-1}(x_i) - \mu_c\right) \left(f_{L-1}(x_i) - \mu_c\right)^\top \tag{16}$$

for $c \in \{1, \ldots, C\}$ where $f_{L-1}(x)$ denotes the output of the pre-logit layer on input $x \in \mathcal{X}$. Finally, it uses a held-out ID validation set to apply temperature scaling to the logits.

Unlike the other methods we evaluate, the DDU method uses two uncertainty estimators, one for epistemic uncertainty and one for aleatoric uncertainty. The epistemic estimator is the negative log probability of the pre-logit on sample $x \in \mathcal{X}$ under the MoG:

$$u_{\text{eu}}^{\text{ddu}}(x) = -\log p\left(f_{L-1}(x) \mid \{\pi_c\}_{c=1}^{C}, \{\mu_c\}_{c=1}^{C}, \{\Sigma_c\}_{c=1}^{C}\right).$$

The aleatoric estimator is the entropy of the softmax predictive distribution:

$$u_{\text{au}}^{\text{ddu}}(x) = \mathbb{H}\left(f(x)\right).$$

Mukhoti et al. (2023) do not provide a predictive uncertainty estimator and the sum of the aleatoric and epistemic estimator is not a performant choice for this task, as the magnitude of the epistemic part is usually much larger than that of the aleatoric part in practice.

During training, we employ the cross-entropy loss to match the network's predicted probabilities to the (one-hot) ground-truth labels.

For a fair comparison of DDU with the other methods, we use the epistemic estimator for the OOD detection task following Mukhoti et al. (2023) and the best-performing one from Appendix F otherwise.

D.1.7. CROSS-ENTROPY BASELINE

As a baseline, we also benchmark a deterministic single-point network trained with the cross-entropy loss. While this is a deterministic method, one can also equate it to a degenerate Dirac delta distribution in parameter space: $q(\theta') = \delta(\theta - \theta')$, making it the simplest possible distributional method.

D.2. Distributional Methods

Distributional methods output a second-order input-conditional probability distribution over probability vectors $q(f(x) \mid x)$, abbreviated as $q(f)$.

D.2.1. SPECTRAL NORMALIZED GAUSSIAN PROCESS

Spectral normalized Gaussian processes (SNGP) (Liu et al., 2020) give an approximate Bayesian treatment to obtain uncertainty estimates using spectral normalization of the parameter tensors and a last-layer Gaussian process approximated by Fourier features. For an input $x \in \mathcal{X}$, it predicts a multivariate Gaussian distribution

$$\mathcal{N}\left(\beta\phi(x), \phi(x)^\top \left(\Phi^\top\Phi + I\right)^{-1} \phi(x) I\right), \tag{17}$$

where β is a learned parameter matrix that maps from the pre-logits to the logits, and $\phi(x) = \cos\left(W f_{L-1}(x) + b\right)$ is a random feature embedding of the input $x \in \mathcal{X}$ with $f_{L-1}(x)$ being a pre-logit embedding, W a fixed semi-orthogonal random matrix, and b a fixed random vector sampled from Uniform$(0, 2\pi)$. $\Phi^\top\Phi$ is the (unnormalized) empirical covariance matrix of the pre-logits of the training set. This is calculated during the last epoch. The multivariate Gaussian presented above can be Monte-Carlo sampled to obtain M logit vectors. During training, we calculate the BMA from the set of logits and use the cross-entropy loss to fit the BMA to the (one-hot) labels.

The method also applies spectral normalization to the hidden weights in each layer to satisfy input distance awareness. We treat whether to apply spectral normalization to the batch normalization modules and whether to use layer normalization in the GP layer as hyperparameters. We benchmark both SNGPs and their non-spectral-normalized variants (denoted by GP).

D.2.2. LATENT HETEROSCEDASTIC CLASSIFIER

Latent heteroscedastic classifiers (HET-XL) (Collier et al., 2023) construct a heteroscedastic Gaussian distribution in the pre-logit layer to model per-input uncertainties: $\mathcal{N}(\phi(x), \Sigma(x))$, where $\phi(x)$ is the learned input-conditional pre-logit mean and

$$\Sigma(x) = V(x)^\top V(x) + \mathrm{diag}(d(x)) \tag{18}$$

is an input-conditional full-rank covariance matrix. Both the low-rank term's $V(x)$ and the diagonal term's $d(x)$ are calculated as a linear function of the layer's output before the pre-logit layer.

One can Monte-Carlo sample the pre-logits from the above Gaussian distribution and obtain a set of logits by transforming each using the last linear layer of the network. During training, this set is used to calculate the Bayesian Model Average (BMA) whose argmax is the final prediction.

HET-XL uses a temperature parameter to scale the logits before calculating the BMA. This is chosen using a validation set. During training, we sample a set of logits, calculate the BMA, and use the cross-entropy loss to fit the BMA to the (one-hot) labels.

D.2.3. LAPLACE APPROXIMATION

The Laplace approximation (Daxberger et al., 2021) approximates a Gaussian posterior $q(\theta \mid \mathcal{D})$ over the network parameters for a Gaussian prior $p(\theta)$ and likelihood defined by the network architecture. It uses the maximum a posteriori (MAP) estimate as the mean and the inverse Hessian of the loss evaluated at the MAP as the covariance matrix:

$$\mathcal{N}\left(\theta_{\text{MAP}}, \left(\frac{\partial^2 \mathcal{L}(\mathcal{D};\theta)}{\partial\theta_i \partial\theta_j}\bigg|_{\theta_{\text{MAP}}}\right)^{-1}\right). \tag{19}$$

This is a post-hoc method applied to a point estimate network. Following the recommendation of Daxberger et al. (2021), we employ a last-layer KFAC Laplace approximation and find the prior variance using cross-validation. We draw network outputs using the GLM predictive on CIFAR-10, and the NN predictive on ImageNet because of the infeasibility of calculating the network Jacobian for the GLM due to extreme memory requirements (≈ 450 GB VRAM).

D.2.4. MC-DROPOUT

MC-Dropout (Srivastava et al., 2014) has been shown to be a variational approximation to a deep Gaussian process (Gal & Ghahramani, 2016). During training, only one logit vector per input is sampled and the cross-entropy loss is used. MC-Dropout in the realm of uncertainty quantification remains active during inference and is used to sample M logits by performing M forward passes. Therefore, it directly samples from $q(f)$ without characterizing it.

D.2.5. DEEP ENSEMBLE

Deep ensembles (Lakshminarayanan et al., 2017) are approximate model distributions that give rise to a mixture of Dirac deltas in parameter space: $q(\theta) = \frac{1}{M}\sum_{i=1}^{M}\delta(\theta - \theta^{(i)})$. Predominantly used to reduce the variance in the predictions and improve model accuracy, deep ensembles can also be used as approximators to the true distribution $p(\theta)$ induced by the randomness over datasets $\mathcal{D} := \{(x_i, y_i) \,|\, i \in \{1, \dots, n\}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ in the generative process $p(x, y)$.

We obtain a set of logits by performing a forward pass over all models. Similarly to MC-Dropout, deep ensembles do not explicitly parameterize the distribution over the predictions, they only sample from it. We ensemble five independently trained cross-entropy models.

D.2.6. SHALLOW ENSEMBLE

Shallow ensembles (Lee et al., 2015) are lightweight approximations of deep ensembles. They use a shared backbone and M output heads (often referred to as "experts"). With a single forward pass, one obtains M logit vectors per input. During training, the BMA of the M predictions is calculated and matched to the ground-truth labels.

D.2.7. EVIDENTIAL DEEP LEARNING

The seminal evidential deep learning method of Sensoy et al. (2018b) (denoted by EDL following Ulmer et al. (2023)) directly learns a second-order predictive distribution $q(f(x) \mid x)$ using closed-form Bayesian inference. In particular, it learns an input-conditional Dirichlet posterior $q(f(x) \mid x) = \mathrm{Dir}(\boldsymbol{\beta}(x))$ with a fixed Dirichlet (conjugate) prior $\mathrm{Dir}(\mathbf{1})$ representing a total lack of information and a categorical distribution over the classes as the likelihood. The logits of the network, $\log \hat{f}(x)$, are turned into pseudo-counts $\boldsymbol{\alpha}(x) \in \mathbb{R}_+^C$ using either the exp or the softplus activation function. The posterior distribution is obtained in closed form by setting $\boldsymbol{\beta}(x) = \boldsymbol{\alpha}(x) + \mathbf{1}$. The components of the IT decomposition can also be derived in closed form, see Ulmer et al. (2023) for details.

The loss of the EDL method has three components:

$$\mathcal{L}^{\text{edl}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\text{one\_hot}(y_i) - \frac{\boldsymbol{\beta}(x_i)}{S(x_i)}\right\|^2 + \sum_{c=1}^{C}\frac{\beta(x_i)_c(S(x_i) - \beta(x_i)_c)}{S^2(x_i)(S(x_i)+1)} + \lambda_t D_{\text{KL}}\left(\mathrm{Dir}(\tilde{\boldsymbol{\beta}}(x_i, y_i) \parallel \mathrm{Dir}(\mathbf{1})\right). \tag{20}$$

where $\tilde{\boldsymbol{\beta}}(x, y) = \text{one\_hot}(y) + (1 - \text{one\_hot}(y)) \odot \boldsymbol{\beta}$ is the Dirichlet parameter vector after removing the prediction corresponding to the label's index. The first term matches the mean of the Dirichlet posterior to the (one-hot) GT labels. The second term reduces the summed variance of each index $c \in \{1, \dots, C\}$ of the random variable distributed as the Dirichlet posterior. These two terms concentrate the Dirichlet density onto the one-hot label. The third term is a regularizer that drives all dimensions of the Dirichlet parameter vector toward a complete lack of knowledge except the one corresponding to the GT label. λ_t is the scheduled trade-off factor at step t. We use a linear up-scaling of λ_t from 0 to $\lambda_{\text{max}} \leq 1$. On CIFAR-10, $\lambda_{\text{max}} = 1$ is used following Sensoy et al. (2018b). Om ImageNet, this led to an overly strong regularizer that prohibited learning (as the regularizer's magnitude depends on the number of classes). We found $\lambda_{\text{max}} = 0.001$ to be a performant maximum trade-off factor for ImageNet.

D.2.8. POSTNET

The PostNet method of Charpentier et al. (2020b) builds upon the EDL method. PostNet also keeps the prior parameters fixed to $\mathbf{1}$, but instead of directly predicting pseudo-counts $\boldsymbol{\alpha}(x)$, they are calculated as $\alpha(x)_c = n_c \cdot p_\phi(z(x) \mid c)$ where $z(x)$ is the latent embedding of input $x \in \mathcal{X}$, n_c is the number of training samples of class $c \in \{1, \dots, C\}$ and $p_\phi(z(x) \mid c)$ is a class-conditional normalizing flow with parameters ϕ. Intuitively, the class-conditional normalizing flows give soft class membership indicators to each input and their indicators are weighted by the class size.

The PostNet method is trained with a regularized Uncertain Cross-Entropy (UCE) loss:

$$\mathcal{L}^{\text{postnet}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{p}\sim\mathrm{Dir}(\boldsymbol{\beta}(x_i))}\left[\mathrm{CE}\left(\mathbf{p}, \text{one\_hot}(y_i)\right)\right] - \lambda\mathbb{H}\left(\mathrm{Dir}(\boldsymbol{\beta}(x_i))\right). \tag{21}$$

While the first term drives $\mathrm{Dir}(\boldsymbol{\beta}(x))$ toward a Dirac distribution concentrated at the one-hot label, the second term maximizes the entropy of the Dirichlet posterior. The effect of each is determined by the trade-off factor λ.

E. Definition and Results of the Bregman Decomposition

Bregman decompositions (Pfau, 2013; Gupta et al., 2022; Lahlou et al., 2023; Gruber & Buettner, 2023) use not only the second-order predictive distribution $q(f)$ but also take the ground-truth (GT) generative process $p^*(x, y)$ into account. Bregman decompositions break up the expected loss of a model over all possible training datasets. This variability is approximated by $q(f)$:

$$\underbrace{\mathbb{E}_{q(f),p^*(y|x)}\left[D_F\left[\text{one\_hot}(y) \parallel f(x)\right]\right]}_{\text{predictive}} = \underbrace{\mathbb{E}_{p^*(y|x)}\left[D_F\left[\text{one\_hot}(y) \parallel f^*(x)\right]\right]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{q(f)}\left[D_F\left[\bar{f}(x) \parallel f(x)\right]\right]}_{\text{epistemic}} + \underbrace{D_F\left[f^*(x) \parallel \bar{f}(x)\right]}_{\text{bias}} \tag{22}$$

The loss D_F is a Bregman divergence like the Euclidean distance or the KL divergence. Since $f^*(x) = \mathbb{E}_{p^*(y|x)}\left[\text{one\_hot}(y)\right]$ is the Bayes predictor, the aleatoric uncertainty is the Bayes risk of the generative process, which is irreducible and independent of the $q(f)$ distribution. As this process is unknown in practice, we *estimate* the aleatoric term by $\mathbb{E}_{q(f)}\left[\mathbb{H}\left(f(x)\right)\right]$. Similarly to the IT decomposition, the epistemic uncertainty is the average distance of predictions $f(x)$ from their centroid $\bar{f}(x) = \arg\min_z \mathbb{E}_{q(f)}\left[D_F\left[z \parallel f(x)\right]\right]$. This average is calculated in a dual space, but in certain cases is equal to the BMA (Gupta et al., 2022). The Bregman decomposition has an additional term, the bias – an uncertainty source that subsumes the uncertainty about the function class (Von Luxburg & Schölkopf, 2011).

E.1. Correlation of Components and Limitations

Let us carry out the same experiments for Bregman as we did for the IT decomposition in Section 3.1 of the main paper. As the Bregman and DEUP decompositions (Equations 22 and 59) consider the *ground-truth* label distribution as the aleatoric component, we use the IT aleatoric uncertainty as an estimator of it. The Bregman decomposition includes a bias component, whose correlations we also investigate.

E.1.1. IMAGENET

Fig. E.1 shows that there is a considerable rank correlation between the Bregman ground-truth aleatoric and bias components but is not severe enough such that it prevents the theoretical possibility of disentangling them via estimators. On the right, we can also see that the IT and Bregman correlation results are very similar between the aleatoric and epistemic estimates.

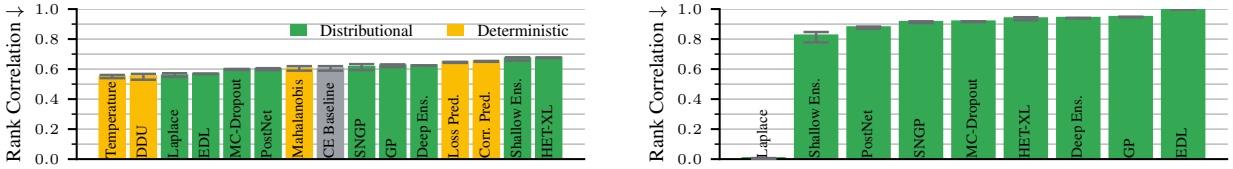


Figure E.1. Left. On ImageNet-ReaL, the rank correlation of the Bregman aleatoric and bias terms is between 0.58 and 0.7 for all distributional methods we benchmark. *Right.* The Bregman decomposition shows similar rank correlation results to the IT decomposition between the estimated aleatoric uncertainty and the epistemic component on the ImageNet validation dataset (Fig. 1).

E.1.2. CIFAR-10

We see in Fig. E.2 that the results using Bregman are virtually the same as those of the IT decomposition: most distributional methods exhibit very high rank correlations. When equipped with the IT decomposition estimators, the less-correlated methods (GP, SNGP, and shallow ensemble) all underperform the baseline (Fig. L.4) and the aleatoric uncertainty alignment is low across all methods. **Given the high ground-truth rank correlation of the aleatoric and bias components shown in Fig. E.2, there also seems to be a fundamental limitation in disentangling them.**

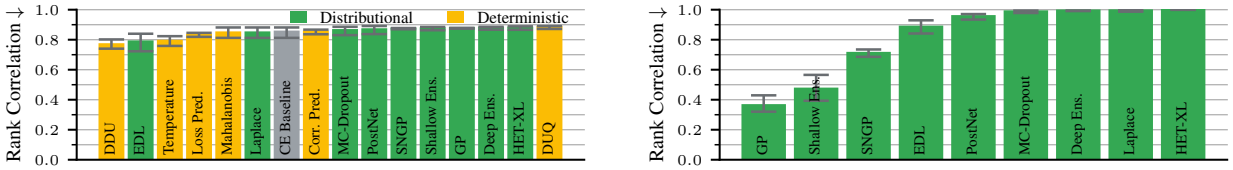


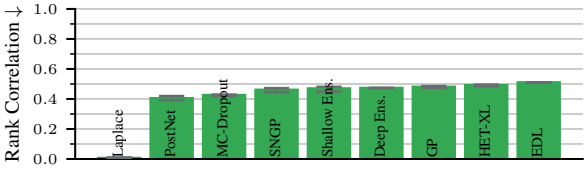
Figure E.2. Left. The rank correlation of the Bregman aleatoric and bias terms is above 0.78 for all methods we benchmark on CIFAR-10. *Right.* On CIFAR-10, the Bregman decomposition shows similar rank correlation results to the IT decomposition between the estimated aleatoric uncertainty and the epistemic component (Fig. K.1).

E.2. Some correlation is inevitable, but disentangling epistemic and aleatoric uncertainty is still feasible

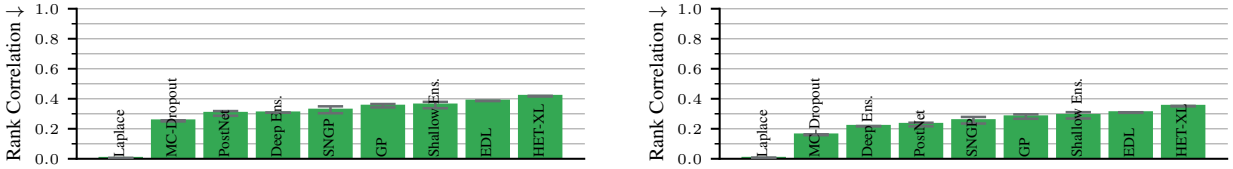
In this section, we present the rank correlation between the *ground-truth* aleatoric uncertainty and the models' epistemic uncertainties. This serves to capture an inevitable level of correlation between these uncertainty sources, as measured (and defined) by the Bregman decomposition.

E.2.1. ImageNet

ImageNet results on the correlation of the ground-truth aleatoric uncertainty and epistemic uncertainties of methods are shown in Fig. E.3. There is a positive rank correlation between these quantities, implying that some level of correlation is inevitable (but not such extreme values displayed in Fig. 1). In Section 3.3 of the main paper, we find that the Mahalanobis epistemic estimates and the CE baseline aleatoric estimates lead to reasonably well-performing yet decorrelated uncertainties. We hypothesize that the main reason is the explicit measurement of aleatoric and epistemic uncertainty at different parts of the computation graph. This, e.g., is lacking for the BMA decomposition: the aleatoric and epistemic estimates are generated from the same set of logits which limits diverse behaviors across estimators.



(a) ID results.



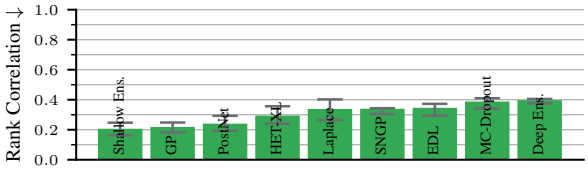
(b) OOD severity level one.

(c) OOD severity level two.

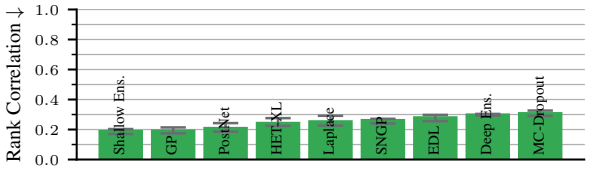
Figure E.3. On ImageNet, we find a positive rank correlation between the (ground-truth) aleatoric and epistemic components of the Bregman decomposition, implying that some level of correlation is inevitable when using this decomposition formula. However, this correlation is considerably lower than that between the aleatoric and epistemic *estimates* in Fig. 1. Only severity levels one and two are shown, as the GT aleatoric uncertainty values from the soft ImageNet-ReaL labels are only valid for these corruption levels – higher corruption would possibly lead to a shift in labeler votes.

E.2.2. CIFAR-10

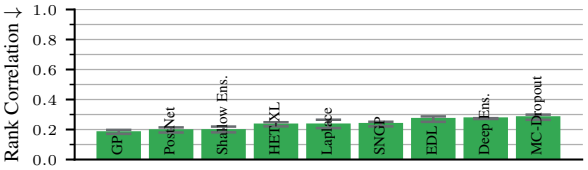
CIFAR-10 results on the correlation of the ground-truth aleatoric uncertainty and epistemic uncertainties of methods are shown in Fig. E.4. Similar to ImageNet, there is a positive rank correlation between these quantities, implying a (low) inevitable level of correlation between the uncertainty sources.



(a) ID results.



(b) OOD severity level one.



(c) OOD severity level two.

Figure E.4. On CIFAR-10, we find a positive rank correlation between the (ground-truth) aleatoric and epistemic components of the Bregman decomposition, implying that some level of correlation is inevitable when using this decomposition formula. However, this correlation is considerably lower than that between the aleatoric and epistemic *estimates* in Fig. K.1. Only severity levels one and two are shown, as the GT aleatoric uncertainty values from the soft CIFAR-10H labels are only valid for these corruption levels – higher corruption would possibly lead to a shift in labeler votes.

E.3. Alignment of Methods with the Bregman Bias

E.3.1. ImageNet

The rank correlation of benchmarked methods with the bias component of the Bregman decomposition is shown in Fig. E.5 for ID and OOD with severity two. The EDL method is most correlated with the Bregman bias but most methods exhibit a high rank correlation (≥ 0.8) (unlike CIFAR-10 below). **This suggests that uncertainty estimators are most aligned with the bias component of Bregman out of the three.** All methods become less correlated with bias with increasing severity, unlike CIFAR-10 below.

E.3.2. CIFAR-10

The rank correlation of benchmarked methods with the bias component of the Bregman decomposition is shown in Fig. E.6. EDL, DDU, and temperature scaling are notably more correlated with the Bregman bias component than the cross-entropy baseline. All methods become better correlated with bias with increasing severity.

F. Aggregators

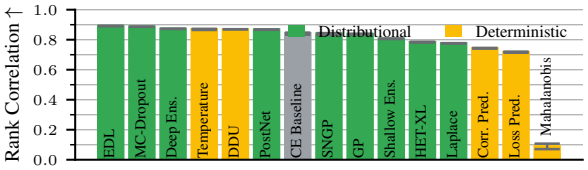
In practical applications, distributional methods output a discrete set of probability vectors $\{f^{(m)}(x)\}_{m=1}^{M}$ per input $x \in \mathcal{X}$. This set can be aggregated in several ways to construct an uncertainty estimate $u(x)$. Commonly used aggregators are the Bayesian Model Average (BMA):

$$\tilde{f}(x) = \frac{1}{M} \sum_{m=1}^{M} f^{(m)}(x), \tag{23}$$

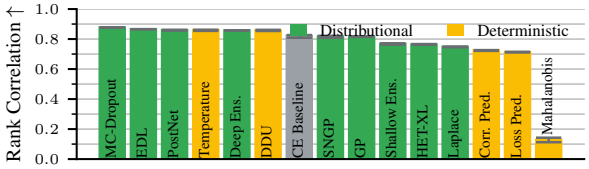
and the Bregman decomposition's central prediction term (Appendix E):

$$\bar{f}(x) = \text{softmax}\left(\frac{1}{M} \sum_{m=1}^{M} \log f^{(m)}(x) \right), \tag{24}$$

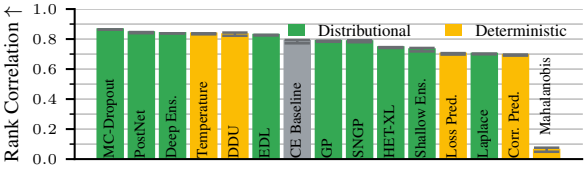
followed by taking their maximum probability, entropy, mutual information, or expected divergence (Mukhoti et al., 2023; Depeweg et al., 2018; Wimmer et al., 2023; Gupta et al., 2022; Gruber & Buettner, 2023). Similarly, one can take the expected maximum probability and expected entropy over the set of probability vectors (Mukhoti et al., 2023). These possible choices are detailed below with pointers to their use in the literature.



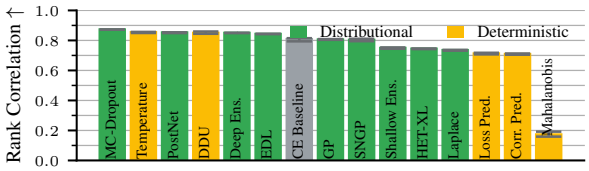
(a) ID rank correlation of methods with the Bregman decomposition's bias component.



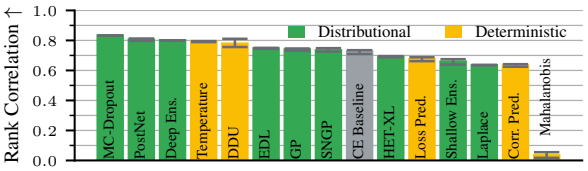
(b) Mixed ID and OOD rank correlation of methods with the Bregman bias using severity-one perturbations.



(c) OOD rank correlation of methods with the Bregman bias using severity-one perturbations.



(d) Mixed ID and OOD rank correlation of methods with the Bregman bias using severity-two perturbations.



(e) OOD rank correlation of methods with the Bregman bias using severity-two perturbations.

Figure E.5. Rank correlation with the Bregman bias component on the ImageNet validation dataset. Most methods exhibit a high rank correlation (≥ 0.8). When going more OOD, all methods become less correlated with bias. Only severity levels one and two are shown, as the GT bias values from the soft CIFAR-10H labels are only valid for these corruption levels – higher corruption would possibly lead to a shift in labeler votes.

F.1. Entropy-based Aggregators

According to the Source Coding Theorem, the entropy of the code is a fundamental and tight lower bound on the expected code word length for prefix-free symbol codes (Yang et al., 2023). The entropy is an expectation over the length of per-symbol codewords. For general distributions $p(x)$, it intuitively measures the spread or the "amount of surprise" in $p(x)$: a higher entropy indicates more stochasticity in the distribution. We consider three entropy-based aggregators of $\{f^{(m)}(x)\}_{m=1}^{M}$ per input $x \in \mathcal{X}$:

$$u(x) = \mathbb{H}\left(\tilde{f}(x)\right) \tag{25}$$

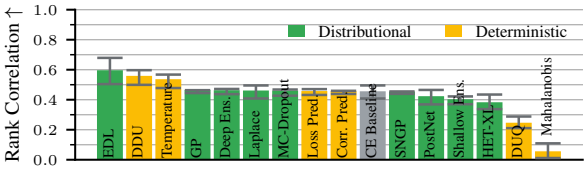
$$u(x) = \mathbb{H}\left(\bar{f}(x)\right) \tag{26}$$

$$u(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{H}\left(f^{(m)}(x)\right) \tag{27}$$

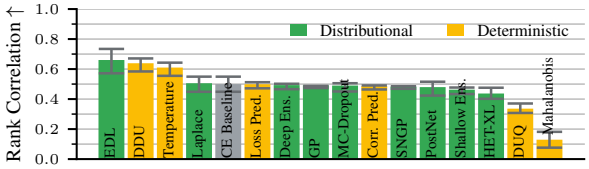
The entropy appears both in the IT and Bregman decompositions (Eq. (1), Eq. (22)). Eq. (25) is often cited to capture predictive (or total) uncertainty, whereas Eq. (27) is known to capture aleatoric uncertainty (Mukhoti et al., 2023; Depeweg et al., 2018; Wimmer et al., 2023). As $\bar{f}(x)$ is a central predictor similar to $\tilde{f}(x)$, its entropy aligns well with a notion of predictive uncertainty.

F.2. Maximum-probability-based Aggregators

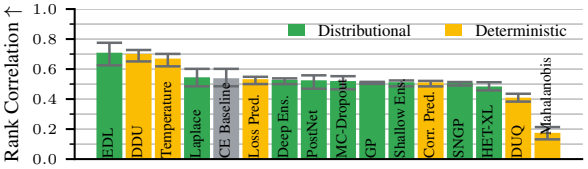
Maximum-probability-based aggregators are similar to entropy-based ones: a small maximum probability value in the prediction vector necessarily means that all entries are small, leading to a high spread and entropy. As uncertainty estimates are higher when the model is more uncertain by convention, one usually takes one minus the maximum probability as a



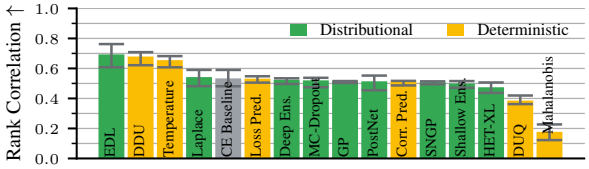
(a) ID rank correlation of methods with the Bregman decomposition's bias component.



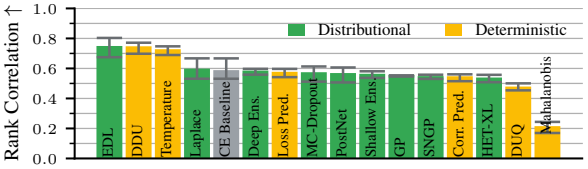
(b) Mixed ID and OOD rank correlation of methods with the Bregman bias using severity-one perturbations.



(c) OOD rank correlation of methods with the Bregman bias using severity-one perturbations.



(d) Mixed ID and OOD rank correlation of methods with the Bregman bias using severity-two perturbations.



(e) OOD rank correlation of methods with the Bregman bias using severity-two perturbations.

Figure E.6. EDL, DDU, and temperature scaling are significantly more correlated with the Bregman bias component than the cross-entropy baseline on CIFAR-10. When going more OOD, all methods become better correlated with bias. Only severity levels one and two are shown, as the GT bias values from the soft CIFAR-10H labels are only valid for these corruption levels – higher corruption would possibly lead to a shift in labeler votes.

notion of uncertainty. We consider three maximum-probability-based aggregators of $\{f^{(m)}(x)\}_{m=1}^{M}$ per input $x \in \mathcal{X}$:

$$u(x) = 1 - \max_{c \in \{1,\dots,C\}} \tilde{f}_c(x) \tag{28}$$

$$u(x) = 1 - \max_{c \in \{1,\dots,C\}} \bar{f}_c(x) \tag{29}$$

$$u(x) = 1 - \frac{1}{M} \sum_{m=1}^{M} \max_{c \in \{1,\dots,C\}} f_c^{(m)}(x) \tag{30}$$

The maximum-probability-based aggregators are restricted to the $[0, 1]$ range. This is particularly important for (strictly) proper scoring rules for the correctness of prediction (Mucsányi et al., 2023) and the notion of calibration, including the ECE and the reliability diagram (Nixon et al., 2019). Similarly to the entropy-based aggregators, Eq. (28) and Eq. (29) align with a notion of predictive uncertainty, whereas Eq. (30) is more aligned with a notion of aleatoric uncertainty.

F.3. Disagreement-based Aggregators

One can directly use the epistemic components of the Bregman and IT decompositions as they do not require a ground truth. In particular, one can use

$$u(x) = \mathbb{H}\left(\tilde{f}(x)\right) - \frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\left(f^{(m)}(x)\right), \quad f^{(m)} \sim q(f) \; \forall m \in \{1, \dots, M\}, \tag{31}$$

the (discretized) epistemic part of the IT decomposition which is the Jensen-Shannon Divergence (see Appendix H), or

$$u(x) = \frac{1}{M} \sum_{m=1}^{M} \left[D_{\text{KL}} \left(\bar{f}(x) \,\middle\|\, f^{(m)}(x) \right) \right], \quad f^{(m)} \sim q(f) \; \forall m \in \{1, \ldots, M\}, \tag{32}$$

the (discretized) epistemic part of the Bregman decomposition which is the expected divergence from the central predictor (see Appendix I). As both aggregators are divergences, they capture disagreement among a set of models. Thus, they are usually cited to be aligned with epistemic uncertainty (Mukhoti et al., 2023; Depeweg et al., 2018; Wimmer et al., 2023; Gupta et al., 2022; Gruber & Buettner, 2023).

Unless stated otherwise, we use the best-performing alternative for each distributional method in the benchmarks. For these methods, the model's prediction is always the most confident class of the BMA. For deterministic methods, we use their "canonical" uncertainty estimator introduced in Appendix D.1.

G. The behavior of the aggregators does not align with what the literature suggests

In this section, we collect per-aggregator results of specific methods to highlight that the best-performing aggregator often goes against what these aggregators intuitively aim to capture as described in Appendix F. Below, we provide a list of abbreviations used in the figures and connect them to the formulas in Appendix F. The "it" and "b" superscripts refer to the IT and Bregman decompositions, respectively. "AU", "PU", "EU", and "B" are shorthands for aleatoric, predictive, epistemic uncertainty, and bias, respectively.

$$\text{PU}^{\text{it}} \equiv \mathbb{H}\left(\tilde{f}(x)\right) \tag{33}$$

$$\text{AU}^{\text{it}} \equiv \frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\left(f^{(m)}(x)\right) \tag{34}$$

$$\text{EU}^{\text{it}} \equiv \mathbb{H}\left(\tilde{f}(x)\right) - \frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\left(f^{(m)}(x)\right) \tag{35}$$

$$\text{PU}^{\text{b}} \equiv \frac{1}{M} \sum_{m=1}^{M} \left[\text{CE}(f^*(x), f^{(m)}(x)) \right] \tag{36}$$

$$\text{AU}^{\text{b}} \equiv \mathbb{H}\left(f^*(x)\right) \tag{37}$$

$$\text{EU}^{\text{b}} \equiv \frac{1}{M} \sum_{m=1}^{M} \left[D_{\text{KL}} \left[\bar{f}(x) \parallel f^{(m)}(x) \right] \right] \tag{38}$$

$$\text{B}^{\text{b}} \equiv D_{\text{KL}} \left[f^*(x) \parallel \bar{f}(x) \right] \tag{39}$$

$$\mathbb{H}\left(\bar{f}\right) \equiv \mathbb{H}\left(\bar{f}(x)\right) \tag{40}$$

$$\mathbb{E}\left[\max f\right] \equiv 1 - \frac{1}{M} \sum_{m=1}^{M} \max_{c \in \{1,\ldots,C\}} f_c^{(m)}(x) \tag{41}$$

$$\max \tilde{f} \equiv 1 - \max_{c \in \{1,\ldots,C\}} \tilde{f}_c(x) \tag{42}$$

$$\max \bar{f} \equiv 1 - \max_{c \in \{1,\ldots,C\}} \bar{f}_c(x) \tag{43}$$

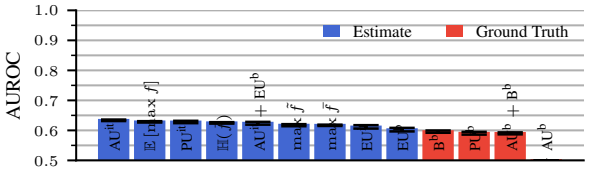
where $f^{(m)} \sim q(f) \; \forall m \in \{1, \ldots, M\}$.

G.1. Aleatoric and predictive aggregators are often best for OOD detection

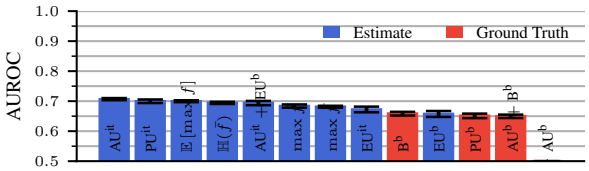
Let us consider the binary prediction task of distinguishing ID and OOD samples. Fig. G.1 and Fig. G.2 show the per-aggregator results on the OOD detection task for the GP and MC-Dropout methods, respectively. These figures highlight two important observations:

1. the best aggregator for the task varies among different methods, and

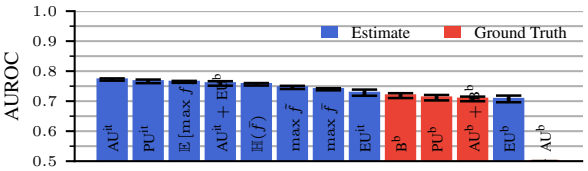
2. the disagreement-based epistemic uncertainty aggregators are not the best-performing ones for either of the methods.



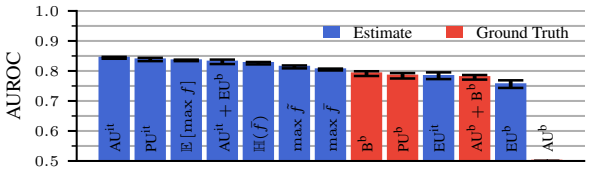
(a) GP AUROC OOD-ness with OOD severity level one.



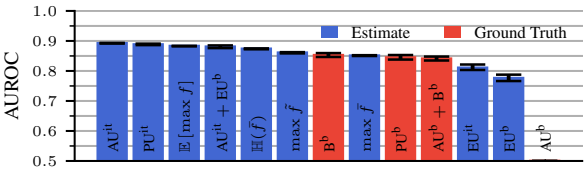
(b) GP AUROC OOD-ness with OOD severity level two.



(c) GP AUROC OOD-ness with OOD severity level three.



(d) GP AUROC OOD-ness with OOD severity level four.



(e) GP AUROC OOD-ness with OOD severity level five.

Figure G.1. OOD detection results of the GP method on the ImageNet validation dataset measured by the AUROC metric. The OOD detection performance of all aggregators increases steadily as we increase the severity of the perturbed half of the mixed dataset. However, the disagreement-based epistemic aggregators, $\mathrm{EU}^{\mathrm{it}}$ and EU^{b}, notably underperform the $\mathrm{AU}^{\mathrm{it}}$ aggregator, even though the epistemic aggregators are deemed more suitable for OOD detection in the literature.

G.2. Methods are not equally sensitive to the choice of aggregator on correctness prediction

Let us turn to the binary correctness prediction task. Fig. G.3 and Fig. G.4 show the per-aggregator results on the correctness prediction detection task for the HET-XL and Deep Ensemble methods, respectively. These figures show that HET-XL is considerably less sensitive to the choice of the aggregator, and the ranking of the aggregators is inconsistent between the two methods. While the epistemic aggregators are the worst-performing estimates for both methods, they are still reasonably indicative of correctness, challenging the common assumption that epistemic aggregators perform poorly in-distribution.

H. Special Form on the Information-Theoretical Decomposition for Discrete Posteriors

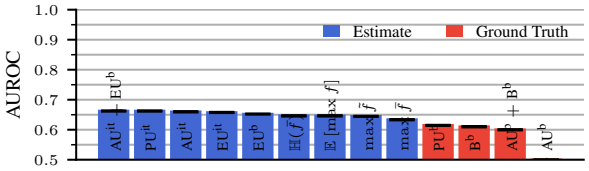
Below, we show that the information-theoretical (IT) decomposition (Depeweg et al., 2018) separates the entropy of the BMA into an expected entropy term and a Jensen-Shannon divergence term when considering discrete uniform distributions $q(f(x) \mid x) = \frac{1}{M} \sum_{m=1}^{M} \delta(f(x) - f^{(m)}(x))$ abbreviated as $q(f)$ in the main paper.

The IT decomposition treats the entropy of the predictive distribution $p(y \mid x) = \int p(y \mid f(x))\, dq(f(x) \mid x)$ as the predictive uncertainty metric and decomposes it into

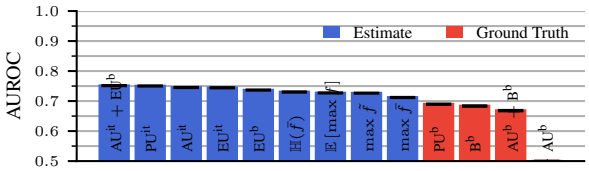
$$\underbrace{\mathbb{H}_{p(y|x)}\left(y\right)}_{\text{predictive}} = \underbrace{\mathbb{E}_{q(f(x)|x)}\left[\mathbb{H}_{p(y|x,f)}\left(y\right)\right]}_{\text{aleatoric}} + \underbrace{\mathbb{I}_{p(y,f(x)|x)}\left(y; f(x)\right)}_{\text{epistemic}}, \tag{44}$$

where $\mathbb{H}$ is the entropy and $\mathbb{I}$ is the mutual information.

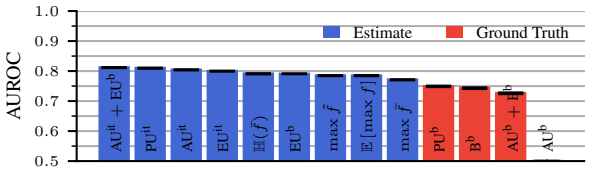
Under a discrete uniform approximate distribution $q(f)$, the predictive uncertainty is still the entropy of the BMA and the aleatoric uncertainty also stays the expected entropy of the probability vectors of non-zero measure. We only have to show that the mutual information takes the convenient form of the Jensen-Shannon divergence under such an approximate



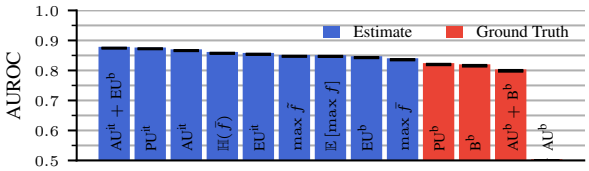
(a) MC-Dropout AUROC OOD-ness with OOD severity level one.



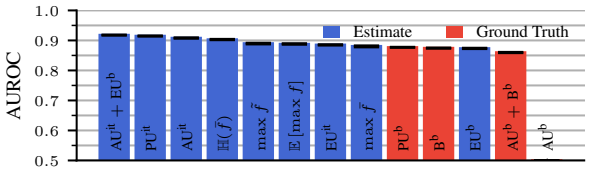
(b) MC-Dropout AUROC OOD-ness with OOD severity level two.



(c) MC-Dropout AUROC OOD-ness with OOD severity level three.



(d) MC-Dropout AUROC OOD-ness with OOD severity level four.



(e) MC-Dropout AUROC OOD-ness with OOD severity level five.

Figure G.2. OOD detection results of the MC-Dropout method on the ImageNet validation dataset measured by the AUROC metric. Similarly to the GP method in Fig. G.1, the disagreement-based epistemic aggregators underperform the $\mathrm{AU}^{\mathrm{it}}$ aggregator on the binary OOD detection task on the ImageNet validation set. Additionally, the ranking of aggregators seems random compared to Fig. G.1.

posterior. Using $p(y, f(x) \mid x) = p(y \mid f(x))q(f(x) \mid x)$, we have

$$\mathbb{I}_{p(y,f(x)|x)}\left(y; f(x)\right) = \sum_{y=1}^{C} \int \log \frac{p(y, f(x) \mid x)}{p(y \mid x)q(f(x) \mid x)} dp(y, f(x) \mid x) \tag{45}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \sum_{y=1}^{C} p\left(y \mid f^{(m)}(x)\right) \log \frac{p\left(y \mid f^{(m)}(x)\right)}{p(y \mid x)} \tag{46}$$

$$= -\frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\left(f^{(m)}(x)\right) - \sum_{y=1}^{C} \frac{1}{M} \sum_{m=1}^{M} p\left(y \mid f^{(m)}(x)\right) \log p(y \mid x) \tag{47}$$

$$= \mathbb{H}\left(\frac{1}{M} \sum_{m=1}^{M} f^{(m)}(x)\right) - \frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\left(f^{(m)}(x)\right) \tag{48}$$

which is the Jensen-shannon divergence of the distributions $p\left(y \mid f^{(m)}(x)\right), m \in \{1, \ldots, M\}$.

I. Special Form of the Bregman Decomposition for the Kullback-Leibler Divergence

Considering a Bregman divergence induced by the strictly convex function F as the loss function $\mathcal{L}$, the Bregman decomposition disentangles it into

$$\underbrace{\mathbb{E}_{q(f),p^*(y|x)}\left[D_F\left[\text{one\_hot}(y) \parallel f(x)\right]\right]}_{\text{predictive}} = \underbrace{\mathbb{E}_{p^*(y|x)}\left[D_F\left[\text{one\_hot}(y) \parallel f^*(x)\right]\right]}_{\text{aleatoric}} \tag{49}$$

$$+ \underbrace{\mathbb{E}_{q(f)}\left[D_F\left[\bar{f}(x) \parallel f(x)\right]\right]}_{\text{epistemic}} \tag{50}$$

$$+ \underbrace{D_F\left[f^*(x) \parallel \bar{f}(x)\right]}_{\text{bias}}, \tag{51}$$

where $f^*(x) = \mathbb{E}_{p^*(y|x)}\left[\text{one\_hot}(y)\right]$ is the Bayes predictor and $\bar{f}(x) = \arg\min_{z \in \Delta^{C-1}} \mathbb{E}_{q(f)}\left[D_F\left[z \parallel f(x)\right]\right]$ is the central predictor. The one_hot function converts a label $y \in \{1, \ldots, C\}$ into a vector that has only zero entries except for the yth one which is one.

When choosing $F(\cdot) = -\mathbb{H}(\cdot)$, we obtain $D_F\left[\cdot \parallel \cdot\right] = D_{\text{KL}}\left(\cdot \parallel \cdot\right)$. Consider the predictive uncertainty term. A one-hot vector's entropy is zero; therefore, the predictive uncertainty becomes $\mathbb{E}_{q(f),p^*(y|x)}\left[\text{CE}(\text{one\_hot}(y), f(x))\right]$. The aleatoric term takes a convenient form:

$$\mathbb{E}_{p^*(y|x)}\left[D_{\text{KL}}\left(\text{one\_hot}(y) \parallel f^*(x)\right)\right] = \mathbb{E}_{p^*(y|x)}\left[\sum_{i=1}^{C} y_i \log \frac{y_i}{f_i^*(x)}\right] = -\sum_{i=1}^{C} f_i^*(x) \log f_i^*(x) = \mathbb{H}\left(f^*(x)\right). \tag{52}$$

On datasets with multiple labels per input, this quantity is precisely the entropy of the (normalized) label distribution corresponding to the labeler votes.

To calculate $\bar{f}(x)$, we can proceed as follows.

$$\bar{f}(x) = \arg\min_{z \in \Delta^{C-1}} \mathbb{E}_{q(f)}\left[D_{\text{KL}}\left(z \parallel f(x)\right)\right] \tag{53}$$

$$= \arg\min_{z \in \Delta^{C-1}} \sum_{i=1}^{C} z_i \log z_i - \sum_{i=1}^{C} z_i \log\left(\exp\left(\mathbb{E}_{q(f)}\left[\log f_i(x)\right]\right)\right) \tag{54}$$

$$= \arg\min_{z \in \Delta^{C-1}} \sum_{i=1}^{C} z_i \log z_i - \sum_{i=1}^{C} z_i \log\left(\exp\left(\mathbb{E}_{q(f)}\left[\log f_i(x)\right]\right)\right) + \sum_{i=1}^{C} z_i \log\left(\sum_{j=1}^{C} \exp\left(\mathbb{E}_{q(f)}\left[\log f_j(x)\right]\right)\right) \tag{55}$$

$$= \arg\min_{z \in \Delta^{C-1}} \sum_{i=1}^{C} z_i \log z_i - \sum_{i=1}^{C} z_i \log \underbrace{\frac{\exp\left(\mathbb{E}_{q(f)}\left[\log f_i(x)\right]\right)}{\sum_{j=1}^{C} \exp\left(\mathbb{E}_{q(f)}\left[\log f_j(x)\right]\right)}}_{p_i :=} \tag{56}$$

$$= \arg\min_{z \in \Delta^{C-1}} D_{\text{KL}}(z \parallel p) \tag{57}$$

$$= p. \tag{58}$$

Therefore, $\bar{f}(x) = \text{softmax}\left(\mathbb{E}_{q(f)}\left[\log f(x)\right]\right)$, where $\log$ is applied elementwise.

I.1. DEUP is a Special Case of Bregman

As mentioned in Appendix E, a closely related formula to Bregman is the risk decomposition of Lahlou et al. (2023) where the predictive uncertainty is directly equated to the risk of a deterministic predictor $f\colon \mathcal{X} \to \mathcal{Y}$, not an expectation of risks over datasets or hypothesis distributions:

$$\underbrace{\mathcal{R}(f, x)}_{\text{predictive}} = \underbrace{\mathcal{R}(f^*, x)}_{\text{aleatoric}} + \underbrace{\mathcal{R}(f, x) - \mathcal{R}(f^*, x)}_{\text{bias}} \tag{59}$$

where $\mathcal{R}(f, x) = \mathbb{E}_{p(y|x)}[\mathcal{L}(f(x), y)]$ is the pointwise risk of f on $x \in \mathcal{X}$. When choosing $\mathcal{L}$ to be the squared Euclidean distance or the Kullback-Leibler divergence, Equation 59 becomes a special case of Equation 22 for a Dirac distribution $q(f') = \delta(f' - f)$ at an arbitrary predictor f. This formulation is desired when one wants the predictive uncertainty to be aligned with the risk of one particular predictor and not the expected risk over a hypothesis distribution.

J. Goals of Disentanglement

What does it mean to have disentangled uncertainty estimators? Consider two estimators $u^{(a)}(x_i), u^{(e)}(x_i)$ and *ground-truth* aleatoric and epistemic uncertainties $U^{(a)}(x_i), U^{(e)}(x_i)$ for each input x_i. The estimators $u^{(a)}$ and $u^{(e)}$ are decorrelated if

1. $u^{(a)}$ has low rank correlation with $U^{(e)}$ and
2. $u^{(e)}$ has low rank correlation with $U^{(a)}$.

Importantly, $u^{(a)}$ and $u^{(e)}$ having a severely high rank correlation prohibits disentanglement. Further, they are well-performing if

3. $u^{(a)}$ has high rank correlation with $U^{(a)}$ and
4. $u^{(e)}$ has high rank correlation with $U^{(e)}$.

We call uncertainty estimators disentangled when they are simultaneously decorrelated and well-performing.

Inspired by generalized bias-variance decompositions (Pfau, 2013; Gruber & Buettner, 2023), one may treat the training dataset $\mathcal{D}$ as a random variable sampled from the generative process $p(x, y)$ and record the variability of the trained predictor under dataset change. Following the Bregman decomposition, one may then define

$$U^{(e)}(x) := \mathbb{E}_{p(\mathcal{D})}\left[D_F\left[\bar{f}(x) \parallel f_{\mathcal{D}}(x)\right]\right] \tag{60}$$

with the corresponding central predictor $\bar{f}(x) = \arg\min_z \mathbb{E}_{p(\mathcal{D})}[D_F[z \parallel f_{\mathcal{D}}(x)]]$. As this is impossible to obtain in practical setups (or is too noisy to Monte Carlo approximate), we instead consider the proxy task of OOD detection for evaluating the disentanglement of aleatoric and epistemic uncertainties.

The above definition generalizes to any pair of different uncertainty sources, e.g., the Bregman bias and aleatoric uncertainties. The workaround of choosing a proxy task is not needed for the evaluation of the Bregman bias and aleatoric components' disentanglement.

K. CIFAR-10 Experiments

This section mirrors Section 3 of the main paper on the CIFAR-10 dataset and also includes CIFAR-10 variants of Appendix B. We want to understand the behavior of current uncertainty quantification methods. We first benchmark disentanglement formulas and then individual estimators on practical tasks.

K.1. Disentanglement formulas often fail

Fig. K.1 reveals a surprisingly simple failure of disentanglement formulas: In six of the nine distributional methods, the IT decomposition leads to aleatoric and epistemic estimates that are highly mutually correlated (rank corr. ≥ 0.97). This violates a necessary condition of disentanglement. The correlation remains even when we add more epistemic uncertainty into the dataset in Appendix L.1 or switch to Bregman estimates in Appendix E. Three posterior estimators, SNGP, GP, and shallow ensembles (but not deep ensembles) show a lower rank correlation between their aleatoric and epistemic components. From this, it follows that disentanglement cannot be thought about only on the level of decomposition formulas but needs to take the explicit implementation into account.

To find out whether the remaining three estimator pairs are not only decorrelated but also well-performing, we compare them to the above-defined aleatoric and epistemic ground truths in Appendix L.2. We find that they perform better than random but worse than specialized deterministic estimators for both tasks. In other words, combining two specialized estimators may practically perform better than deriving them by decomposing a single quantity.

K.2. Correctness prediction works across the board

Fig. K.2 shows that most uncertainty estimators perform within ± 0.015 of the cross-entropy baseline when predicting correctness ID, and modern methods like HET-XL do not outperform older methods like deep ensembles. We see a similar saturation (but a different ranking) when slightly altering the correctness metric to account for soft labels in Appendix M.1.

The saturation is even more pronounced on the abstained prediction task. All uncertainty methods apart from Mahalanobis obtain an AUC score greater than 0.99 in Fig. K.3. Practically, this means that one can obtain a close-to-perfect classification accuracy by abstaining from prediction on a tiny set of samples. In both tasks, the more computationally expensive distributional methods have a slight edge over deterministic methods.

The Mahalanobis method – being a specialized OOD detector – cannot predict the correctness of only in-distribution samples. DUQ, which also models the data density, falls behind the cross-entropy baseline as well on the correctness prediction task.

K.3. Uncertainties are only as robust as accuracy

For the reliable deployment of uncertainty quantification methods, their estimates should stay performant longer than the model itself when going OOD. Only then can we trust the uncertainty estimates and make downstream decisions based on them. At first glance, when plotting the raw performance values as dashed lines in Fig. K.4, this is the case. However, AUROCs have a random performance value of 0.5 and the random predictor's expected accuracy is the inverse of the number of classes. Normalizing these values out by $(\text{metric} - \text{rnd})/(1 - \text{rnd})$ for direct comparability, where rnd is the base value that a random predictor achieves on that metric (0.5 for AUROC, $1/C$ for classification accuracy), reveals that both correctness prediction and abstained prediction degrade consistently with the model's accuracy as the data becomes more and more OOD (solid lines). Results per method are reported in Appendix M.2. This observation shows that all benchmarked methods are incapable of generalizing better than the models, making their correctness predictions not trustworthy at higher perturbation levels.

K.4. OOD-ness is hard to detect

If both accuracy and uncertainty estimates deteriorate OOD, can current methods reliably detect OOD samples? To evaluate OOD detection performance, we use balanced mixtures of ID and OOD datasets. OOD samples are perturbed ID samples with severity level two where the models' accuracy already severely deteriorates according to Fig. K.4. The uncertainty estimators are tasked to predict which sample is OOD, i.e., OOD inputs should have higher uncertainty estimates. As shown in Fig. K.5, the Mahalanobis method, which in the previous tasks was far off, is by far the most performant on average in telling apart clean CIFAR-10 samples from perturbed ones. This shows that the OOD task indeed benchmarks a notably different uncertainty quantification capability than the previous predictive uncertainty tasks, namely epistemic uncertainty. We further investigate the correlations between the tasks in Appendix K.6.

One may expect that latent-density-based methods are naturally good at predicting OOD-ness since, for a well-behaved latent density, the embeddings of OOD inputs will lie far away from the class centroids. Interestingly, the worst-performing method also builds a latent density estimate: we cannot establish clear categories of methods that work and that do not. We hypothesize that the main power of the Mahalanobis method lies in its adversarial perturbation of the inputs and that adding this to DDU or DUQ would make them as good as Mahalanobis. We leave this experiment to future work.

The uncertainty methods show a steady increase in OOD detection performance as we increase the severity of the perturbed half of the dataset (see Appendix M.3). The advantage of the Mahalanobis method vanishes as the severity increases. At severity level five, most methods become saturated.

K.5. Aleatoric uncertainty alone is hard to quantify

Let us now benchmark how much they predict aleatoric uncertainties. Since we use the entropy of human annotator label distributions as ground truths, this could also be considered the alignment with human uncertainties. Fig. K.6 shows that most methods perform within the error bar of the cross-entropy baseline. Correctness prediction is most aligned *on average* with a notably small min-max error bar. This is reasonable since ID, the aleatoric uncertainty of the sample determines the network's correctness the most.[1] The latent density methods DUQ and Mahalanobis are particular outliers. Even though they are intended to capture aleatoric uncertainty by placing aleatorically uncertain samples in between class centroids,

[1] Note that we train with only one label per input.

Method ranking	Correctness AUROC	Abstinence AUC	Log Prob. Score	Brier Score	ECE	Aleatoric Rank Corr.	OOD AUROC
Deep Ensemble	1	1	1	1	6	2	2
MC-Dropout	2	2	4	2	7	4	12
CE Baseline	5	7	8	8	9	8	7
SNGP	3	6	13	12	2	3	6
GP	4	4	12	9	1	6	5
Mahalanobis	16	16	-	-	-	16	1
Shallow Ensemble	11	3	7	5	10	12	10
Laplace	6	8	3	7	3	9	8
HET-XL	8	5	6	4	11	14	11
Correctness Pred.	13	11	14	13	14	1	9
Loss Prediction	14	13	-	-	-	5	3
DUQ	15	15	9	3	8	15	16
Temperature	7	9	2	6	4	10	4
DDU	9	12	5	11	5	7	15
EDL	10	14	10	14	13	11	13
PostNet	12	10	11	10	12	13	14

Table K.1. Different tasks have different best-performing methods on CIFAR-10. The deep ensemble method is a good choice across the board but is expensive, see Appendix N.3. The best, second-best, and third-best methods are highlighted in gold, silver, and bronze, respectively. Note that differences between ranks can be very small, see the plots per task for details.

resulting in a low density (Van Amersfoort et al., 2020), they perform worse than the remaining methods. Mahalanobis even performs randomly. This reinforces that it is an epistemic, not an aleatoric uncertainty estimator.

K.6. Different tasks require different estimators

In the previous sections, we have hinted at the fact that the performance across methods is very similar on some tasks and dissimilar on others. In this section, we investigate the correlation among the previous practical tasks using a correlation matrix displayed in Fig. K.7. To construct the matrix, we consider all benchmarked methods with all uncertainty aggregators (see Appendix F) and calculate the correlation of their rankings on different metrics. We find that methods good in predicting correctness are good in abstaining from prediction and vice versa (corr. 0.96). The other cluster is formed by the two strictly proper scoring rules – the log probability and the Brier score – and the ECE metric. These metrics form a cluster that captures predictive uncertainty capabilities. Interestingly, methods that are good at detecting OOD samples are also quite aligned with aleatoric uncertainty. Accuracy is not correlated with any of the three clusters.

As there are different groups of tasks, there is no one-fits-all uncertainty estimator. Table K.1 demonstrates this by ranking all methods on all tasks. An uncertainty estimator has to be chosen or developed for the specific task a practitioner is interested in. If the task is unknown, deep ensembles offer a good compromise, but even MC-Dropout is a good starting point if the runtime costs of deep ensembles are too high.

K.7. Performance depends on implementation details

We conclude with an example where defining the specific task is of particular importance. SNGP shows a highly different performance depending on the dataset and task its aggregator is optimized for. Fig. K.8 shows that GPs provide the best-calibrated uncertainties ID, but already at the lowest OOD perturbation level drop to the cross-entropy baseline level, with the ECE jumping from 0.005 to 0.084. In Appendix M.4, we show that the best way to aggregate SNGP's second-order predictive distributions $q(f)$ into an uncertainty score $u(x)$ is different when optimizing for the ECE versus for correctness prediction. This is not a bug; we implemented SNGP three times from scratch with the same results. It rather goes to show

that subtle design choices greatly affect performance. Hence, we encourage, as best practice, to first define the uncertainty quantification task at hand and then develop a specialized uncertainty estimator.

L. Further Results on the Information-Theoretical Decomposition

L.1. OOD Generalization Performance

The ID correlation of the IT decomposition's components using different distributional methods is discussed in Appendix K.1 and Section 3.1 of the main paper. In this section, we focus on how this correlation changes as we go more and more OOD.

L.1.1. IMAGENET

Fig. L.1 shows the generalization performance of benchmarked distributional methods using the IT decomposition at severity levels one and two. Only Laplace shows decorrelated IT aleatoric and epistemic estimates – but when equipped with these estimators, it predicts epistemic uncertainty at a chance level. For the other methods, these sources generally become less correlated as we go more OOD. Balanced mixtures of ID and OOD samples lead to higher correlations, but the ranking of methods remains unchanged.

L.1.2. CIFAR-10

Results for severity levels two and five are shown in Fig. L.2. Only SNGP and GP show a reasonably low rank correlation between the IT aleatoric and epistemic components across all severities. These sources become slightly less correlated as we go more OOD– much less so than on ImageNet (Fig. L.1). Balanced mixtures of OOD samples lead to higher correlations, but the ranking of methods remains unchanged.

L.2. Performance of Decorrelated Methods using the Information-Theoretical Components

L.2.1. IMAGENET

In Section 3.1 of the main paper, we show that using Laplace, the aleatoric and epistemic components of the IT decomposition can become uncorrelated. In this section, we show that they do not perform well on the tasks they are made for. Fig. L.3 shows that Laplace does not match the cross-entropy baseline with its best estimator when Laplace uses the estimators of the IT decomposition—in fact, it even performs randomly on the OOD detection task. This limitation prohibits its practical use, and we cannot benefit from the less correlated components.

L.2.2. CIFAR-10

In Appendix K.1, we discuss that when using SNGP variants or shallow ensembles on CIFAR-10, the aleatoric and epistemic components of the IT decomposition can become more uncorrelated. In this section, we show that while these components might be less correlated, they are not performant on the tasks they are made for. Fig. L.4 shows that both the SNGP variants and the shallow ensemble method underperform the cross-entropy baseline when the SNGPs and the shallow ensemble use the estimators of the IT decomposition.

M. Further Practical Results

M.1. Correctness Prediction

M.1.1. IMAGENET

We show the correctness prediction performance of methods on OOD and mixed ID + OOD datasets in Fig. M.1. Evidential deep learning methods, EDL and PostNet, dominate on the correctness prediction task across all severity levels and mixtures. This becomes even more pronounced on highly OOD datasets. The performance of Mahalanobis increases on mixed datasets as models perform worse on OOD images than on ID ones and it can detect OOD samples well.

As we have access to validation sets with multiple labels per input, the notion of a "correct" prediction becomes unclear. In the previous plots, we focus on the canonical notion of correctness: whether the model predicts the one-hot label. A related notion of correctness is soft correctness: here, the model does not receive a binary reward for its prediction but rather a continuous number $c \in [0, 1]$: this gives the ground-truth probability of the predicted class being the correct one. We can

calculate a similar AUROC correctness score as before, but as the AUROC requires binary labels, one has to unroll the continuous correctness value: e.g., if $c = 0.8$ for an input $x \in \mathcal{X}$, one can represent this with 8 correct and 2 incorrect predictions on the same input $x \in \mathcal{X}$. The resulting AUROC has a theoretical limit strictly less than one when $c < 1$. The corresponding results are shown in Fig. M.2 that follows the same structure as Fig. M.1. Even though this metric also takes the GT aleatoric uncertainty into account, the ordering of the methods is unaffected compared to Fig. M.1.

M.1.2. CIFAR-10

We show the correctness prediction performance of methods on OOD and mixed ID + OOD datasets in Fig. M.3. We observe a consistent degradation of performance across methods on both dataset types. The only exception is the Mahalanobis method: here, we observe an increase in performance. For the mixed datasets, it can be explained by the fact that the models perform worse on OOD images than on ID ones, making the Mahalanobis method a suitable estimator of correctness. However, the result is unexpected for solely OOD samples.

The results on the soft AUROC metric (Appendix M.1.1) are shown in Fig. M.4 that follows the same structure as Fig. M.3. The saturation of methods is unchanged; however, the ordering is affected.

M.2. Performance Tendency for Increasing Severity

In Appendix B.2 of the main paper and Appendix K.3, we show the performance of MC-Dropout when going OOD, claiming that it is prototypical for other methods. Figs. M.5 and M.6 show for CIFAR-10 and ImageNet, respectively, that MC-Dropout is not an outlier and other methods show very similar generalization capabilities. In Figs. B.3, M.5 and M.6, we normalize the metrics using the formula $(\text{metric} - \text{rnd})/(1 - \text{rnd})$ for direct comparability, where rnd is the base value that a random predictor achieves on that metric (0.5 for AUROC, $1/C$ for classification accuracy).

M.3. OOD Detection

M.3.1. ImageNet

In Appendix K.4, we hint at the fact that nearly all methods show a steady increase in OOD detection performance as we increase the severity of the perturbed half of the dataset. Fig. M.7 shows how the performance of each method changes as we increase the severity level. We can see a steady increase in OOD detection performance for all methods. However, the specialized OOD detector, Mahalanobis, benefits less than the other methods. In particular, at severity level three, MC-Dropout becomes best on average, and shallow ensembles also overtake the Mahalanobis method. At severity level five, Mahalanobis becomes the *second-worst* OOD detector out of the benchmarked methods. This may be because Mahalanobis was trained to detect samples at severity level two and cannot generalize as well to higher severity levels as the other methods.

M.3.2. CIFAR-10

Fig. M.8 shows the OOD detection performance of the benchmarked methods as we increase the severity level. We can see a steady increase in the performance for all methods. However, the specialized OOD detector, Mahalanobis, benefits less than the other methods: At severity level four, deep ensembles become best on average, and GP also overtakes the Mahalanobis method at severity level five.

M.4. Sensitivity to the Choice of Aggregator on CIFAR-10

This section demonstrates that the choice of aggregator is of crucial importance for specific methods and tasks, showing results on the CIFAR-10 dataset. Some tasks, such as correctness prediction and abstained prediction tasks (ID), have highly correlated performances across methods. This is an intuitive result, as both tasks fundamentally require telling apart correct and incorrect samples. However, this is not always the case. OOD detection and the ECE score are two such tasks/metrics.

Considering that the ECE is closely connected to correctness but on a much more fine-grained scale than the binary correctness prediction task, one would expect that a high ECE score translates over to a high correctness AUROC score. Surprisingly, this is sometimes not the case for distributional methods, as shown in Fig. M.9. Importantly, the choice of aggregator (see Appendix F) per distributional method has a high influence on predictive power and the correlation of performances: SNGP variants, Laplace, deep ensemble, and EDL all have different optimal aggregators for correctness and

ECE. SNGP variants are particular outliers: they have the largest trade-off between the best possible performances in the two tasks.

M.5. ECE OOD Generalization on CIFAR-10

In Fig. K.8, we show that GP is the most calibrated ID on the CIFAR-10 dataset, but it also breaks down to the cross-entropy baseline level from severity level one. Fig. M.10 shows that the four methods that bring considerable improvements compared to the cross-entropy baseline as we increase the severity are deep ensemble, DDU, temperature scaling, and EDL ($\geq .05$ ECE improvement).

M.6. Log Probability Proper Scoring Rule

M.6.1. ImageNet

Results on the log probability proper scoring rule are shown in Fig. M.11. The EDL method consistently outperforms all other methods across all severity levels. The performance of all methods stays roughly constant as we increase the severity, which is in line with Fig. B.3 that also shows an almost constant correctness prediction performance on ImageNet.

M.6.2. CIFAR-10

In Fig. M.12, we present the methods' results on the log probability proper scoring rule considering the CIFAR-10 dataset. We find that deep ensemble, temperature scaling, and Laplace consistently outperform the others ID and on lower OOD severity levels. On severity level five, the EDL method surpasses the others.

M.7. Spearman Correlation Matrixon ImageNet

Fig. M.13 showcases the Spearman correlation matrix variant of Fig. 4. The results are stable over these different metrics.

N. Training and Implementation Details

For both datasets, we train and evaluate on an NVIDIA GeForce RTX 2080 Ti. We only use an NVIDIA A100 Tensor Core GPU for the construction of the Laplace approximation on ImageNet, owing to the VRAM requirements of this method.

N.1. CIFAR-10

For CIFAR-10, we follow the augmentations and training schedules of the uncertainty_baselines GitHub repository. In particular, we train a Wide ResNet 28-10 (Zagoruyko & Komodakis, 2016) for 200 epochs with a step decay schedule at $[60, 120, 160]$ epochs with decay rate 0.2. We use stochastic gradient descent with momentum 0.9 and a batch size of 128. Our training augmentation comprises a random crop using padding 2 and a random flip on the vertical axis with probability 0.5. The learning rate and weight decay hyperparameters are chosen by the Bayesian optimization scheme of Weights and Biases (Biewald, 2020). The additional hyperparameters of benchmarked methods are determined by either using values suggested by the original authors or including these in the hyperparameter sweep.

N.2. ImageNet

On ImageNet, we fine-tune a pretrained ResNet 50 (He et al., 2016) using the `resnet50.a1_in1k` parameters from the timm library as initialization. We fine-tune for 50 epochs following a cosine learning rate schedule (Loshchilov & Hutter, 2017) using the AdamW optimizer (Loshchilov & Hutter, 2019) and a learning rate warmup period of 5 epochs. We use a batch size of 128 with 16 accumulation steps, resulting in an effective batch size of 1024. The hyperparameters are chosen identically to those on CIFAR-10 (see Appendix N.1).

N.3. Runtime

Table N.1 and Table N.2 show statistics of the per-epoch runtime for each method on ImageNet and CIFAR-10, respectively. As Laplace, Mahalanobis, temperature scaling, and deep ensemble are post-hoc methods, their reported time comprises the construction of the method and its evaluation on various ID and OOD test datasets.

Table N.1. Summary of per-epoch times for the benchmarked methods on CIFAR-10. For methods trained from scratch (above the separator), the reported time includes both iterating through the training dataset and the ID evaluation. As Laplace, Mahalanobis, and Deep Ensemble are post-hoc methods (below separator), their reported time comprises the construction of the method and its evaluation on the ID and multiple OOD datasets. Methods are sorted by increasing mean per-epoch runtime, separately for trained and post-hoc methods.

Method	**Mean** (s)	**Min** (s)	**Max** (s)	**Std Dev** (s)
CE Baseline	90.1829	83.7423	122.5249	6.0592
GP	91.2036	83.9509	236.0735	9.5304
Correctness Prediction	91.8449	82.6371	133.5970	6.4087
Shallow Ensemble	91.9881	83.0224	119.2658	5.1190
Risk Prediction	94.7748	83.2490	123.5470	8.8957
SNGP	96.8125	88.4623	129.1145	6.4576
EDL	96.9129	82.2683	152.2057	15.2316
HET-XL	99.0390	89.4465	161.1186	8.7700
PostNet	115.8050	100.1761	444.4346	22.3231
DDU	119.5674	87.1077	203.7920	29.8407
MC-Dropout	134.0979	126.2741	188.3551	7.0563
DUQ	148.6540	137.8707	197.6619	7.9503
Temperature	205.6734	165.4589	303.5972	52.1396
Laplace	273.2982	250.8563	307.6892	22.9335
Mahalanobis	370.4277	360.0912	376.3072	5.7191
Deep Ensemble	865.0582	835.1872	904.2822	22.8003

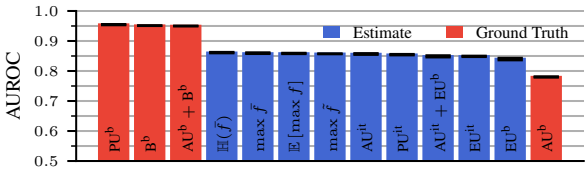
O. Visualization of Images and Label Distributions

This section displays both easy (low human uncertainty) ImageNet samples in Fig. O.1 and hard (high human uncertainty) ones in Fig. O.2 using the ImageNet-ReaL labels and ImageNet-C perturbations.

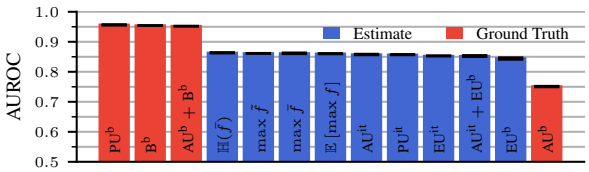
Figs. O.3 and O.4 give summary statistics of the label distributions of ImageNet-ReaL and CIFAR-10H, respectively.

Table N.2. Summary of per-epoch times for the benchmarked methods on ImageNet. For methods trained from scratch (above the separator), the reported time includes both iterating through the training dataset and the ID evaluation. As Laplace, Mahalanobis, and Deep Ensemble are post-hoc methods (below separator), their reported time comprises the construction of the method and its evaluation on the ID and multiple OOD datasets. Methods are sorted by increasing mean per-epoch runtime, separately for trained and post-hoc methods.

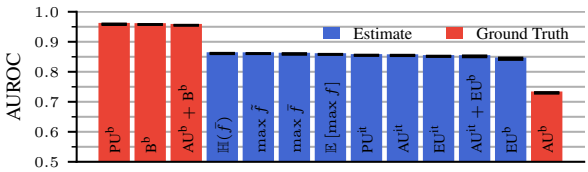
Method	**Mean** (s)	**Min** (s)	**Max** (s)	**Std Dev** (s)
CE Baseline	2646.6512	2566.9703	2899.1143	44.7298
DDU	2662.4925	2610.7394	2740.5103	23.8702
PostNet	2674.9798	2576.4275	3053.4624	129.0717
Risk Prediction	2692.1484	2637.5598	2908.5478	33.4657
Correctness Prediction	2732.9235	2562.6328	3284.8763	230.3123
Shallow Ensemble	2803.4469	2671.9823	3268.5573	181.0415
EDL	3020.7741	2610.7233	3761.1175	450.4668
GP	3059.7266	2645.5470	3880.5432	399.7528
MC-Dropout	3145.4667	3034.7784	3307.8100	80.4335
SNGP	3233.9454	3081.7184	3742.6995	145.4325
HET-XL	4018.7616	3915.7693	4214.7061	55.4737
Temperature	29267.2569	28410.3914	30294.3377	671.6607
Mahalanobis	33929.2063	32972.1129	35235.9114	956.6842
Laplace	52836.5020	52298.8008	53949.9782	588.1313
Deep Ensemble	161492.2153	161492.2153	161492.2153	0.0000



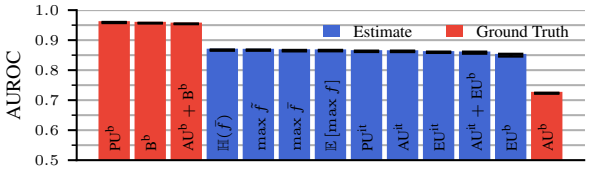
(a) HET-XL correctness AUROC on ID data.



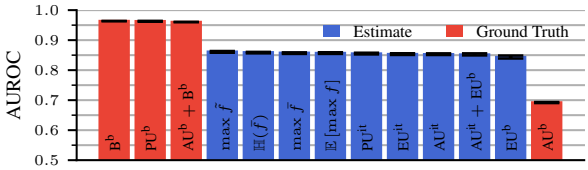
(b) HET-XL correctness AUROC on mixed ID and OOD data of severity level one.



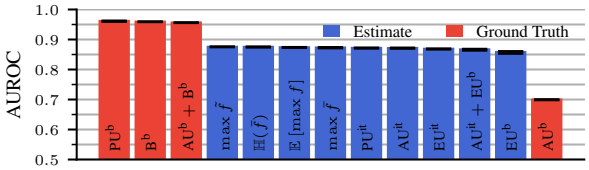
(c) HET-XL correctness AUROC on OOD data of severity level one.



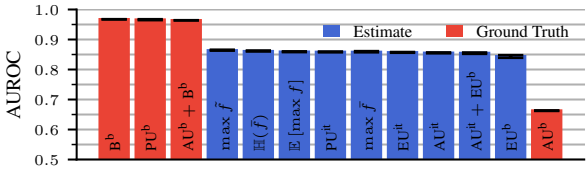
(d) HET-XL correctness AUROC on mixed ID and OOD data of severity level two.



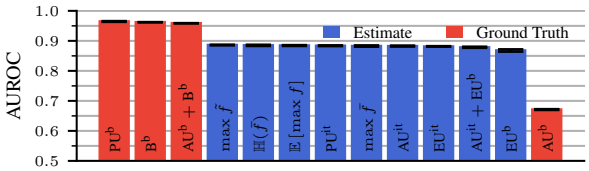
(e) HET-XL correctness AUROC on OOD data of severity level two.



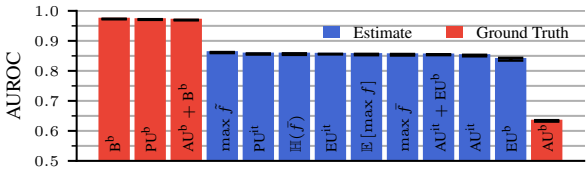
(f) HET-XL correctness AUROC on mixed ID and OOD data of severity level three.



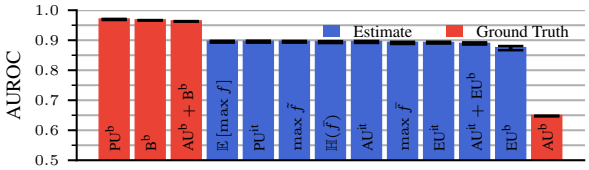
(g) HET-XL correctness AUROC on OOD data of severity level three.



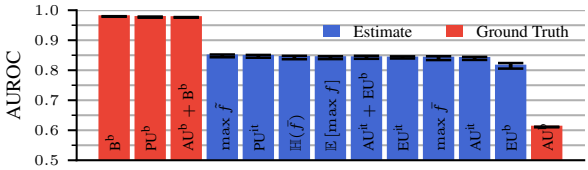
(h) HET-XL correctness AUROC on mixed ID and OOD data of severity level four.



(i) HET-XL correctness AUROC on OOD data of severity level four.

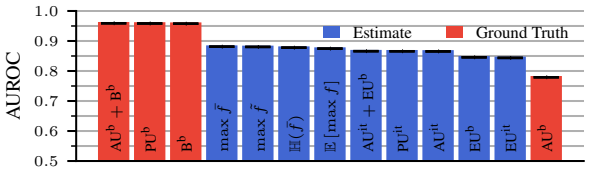


(j) HET-XL correctness AUROC on mixed ID and OOD data of severity level five.

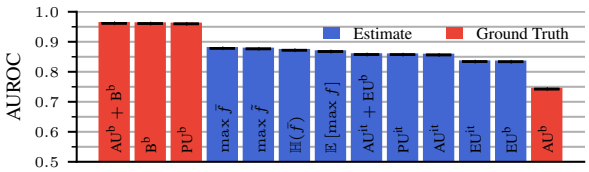


(k) HET-XL correctness AUROC on OOD data of severity level five.

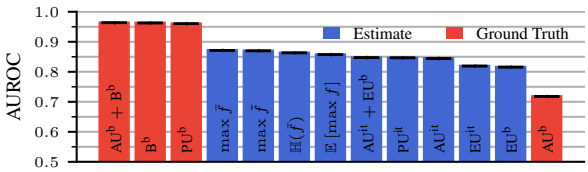
Figure G.3. For the HET-XL method, the correctness prediction performance is saturated across aggregators on the ImageNet validation dataset. The disagreement-based epistemic aggregators, EU^{it} and EU^b, are the worst-performing choices ID, but they are still reasonably indicative of correctness, challenging the common assumption that epistemic aggregators perform poorly in-distribution.



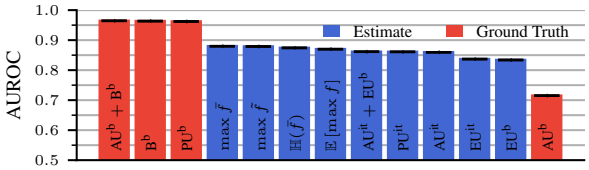
(a) Deep Ensemble correctness AUROC on ID data.



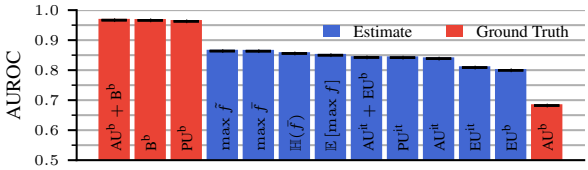
(b) Deep Ensemble correctness AUROC on mixed ID and OOD data of severity level one.



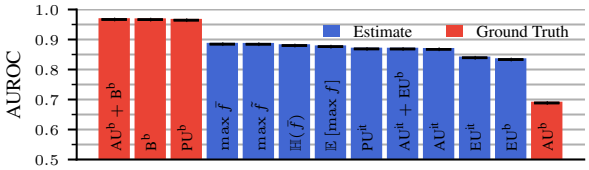
(c) Deep Ensemble correctness AUROC on OOD data of severity level one.



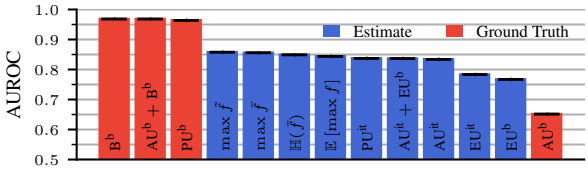
(d) Deep Ensemble correctness AUROC on mixed ID and OOD data of severity level two.



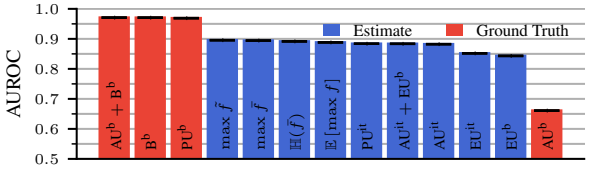
(e) Deep Ensemble correctness AUROC on OOD data of severity level two.



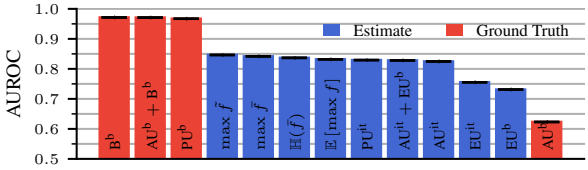
(f) Deep Ensemble correctness AUROC on mixed ID and OOD data of severity level three.



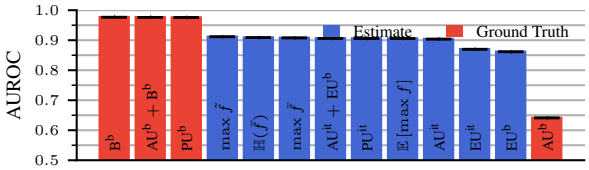
(g) Deep Ensemble correctness AUROC on OOD data of severity level three.



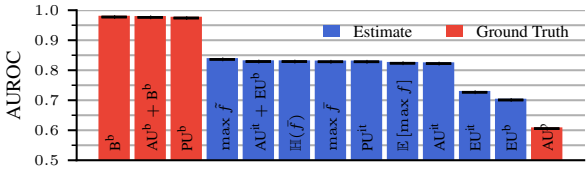
(h) Deep Ensemble correctness AUROC on mixed ID and OOD data of severity level four.



(i) Deep Ensemble correctness AUROC on OOD data of severity level four.



(j) Deep Ensemble correctness AUROC on mixed ID and OOD data of severity level five.



(k) Deep Ensemble correctness AUROC on OOD data of severity level five.

Figure G.4. For the Deep Ensemble method, similarly to the HET-XL method in Fig. G.3, the disagreement-based epistemic aggregators underperform the AU^{it} aggregator on the correctness prediction task on the ImageNet validation set ID, but are still reasonably performant. Additionally, the ranking of aggregators seems random compared to Fig. G.3.

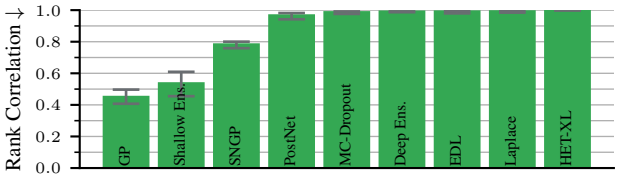


Figure K.1. Six out of nine distributional methods exhibit an almost perfect rank correlation (≥ 0.97) between the IT aleatoric and epistemic components when tested on CIFAR-10. Unlike in Fig. 1, the Laplace method shows an extreme rank correlation.

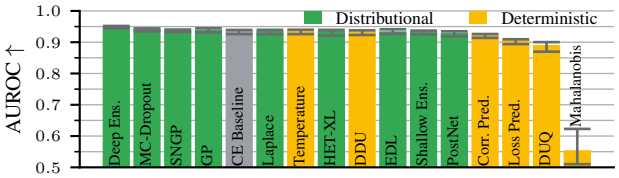


Figure K.2. On ID CIFAR-10 samples, the performance of methods on predicting correctness is saturated, with only deep ensembles and MC-Dropout achieving consistently better results (AUROC ≥ 0.94). The Mahalanobis method is a specialized OOD detector that cannot distinguish ID samples (AUROC $= 0.55$).

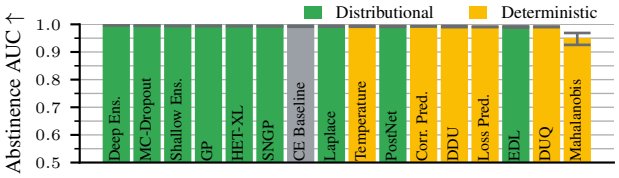


Figure K.3. ID abstained prediction results using the abstained accuracy AUC metric on the CIFAR-10 test dataset. On ID CIFAR-10 samples, most methods solve the abstinence task almost perfectly.

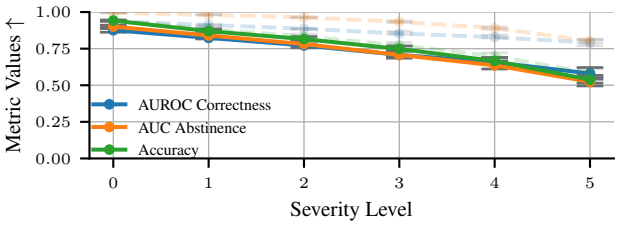


Figure K.4. Degradation of correctness prediction, abstained prediction, and accuracy metrics w.r.t. an increasing level of OOD-ness on the CIFAR-10 test dataset. Model accuracy and the performance of the uncertainty method degrade together as the samples become more OOD by corrupting the images with CIFAR-10C corruptions. The displayed method is MC-Dropout, whose results are representative of all other methods (except Mahalanobis). Solid lines correspond to metrics normalized to the $[0, 1]$ range w.r.t. the random predictor; dashed lines correspond to the unnormalized values.

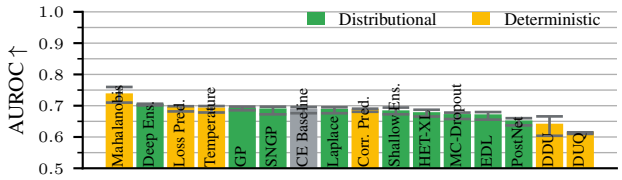


Figure K.5. OOD detection results on the CIFAR-10 test set, measured by the AUROC metric. OOD samples are perturbed by CIFAR-10C corruptions of severity level two. Only deep ensembles and Mahalanobis, a direct OOD detector, can distinguish ID and OOD samples considerably better than the cross-entropy baseline on CIFAR-10.

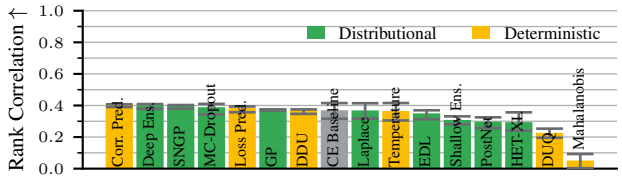


Figure K.6. Rank correlation with the soft input-conditional label distributions of CIFAR-10H corresponding to labeler votes. None of the methods have a considerably higher rank correlation with the ground-truth aleatoric uncertainty than the cross-entropy baseline.

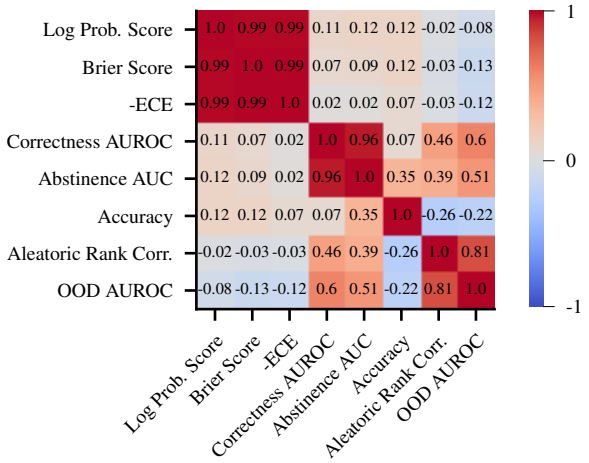


Figure K.7. Pearson correlation of metric pairs calculated over all (method, aggregator) pairs on CIFAR-10. The correlation of metrics is notably different from that of ImageNet (Fig. 4).

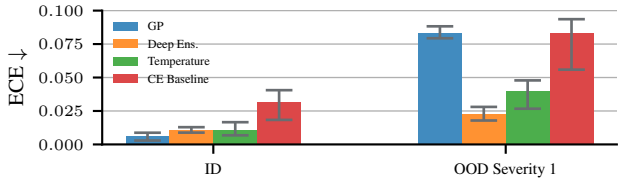
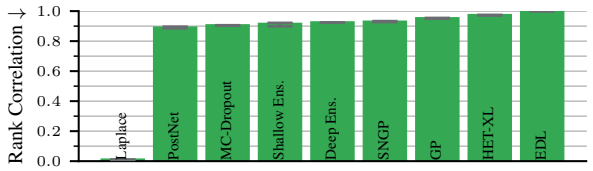
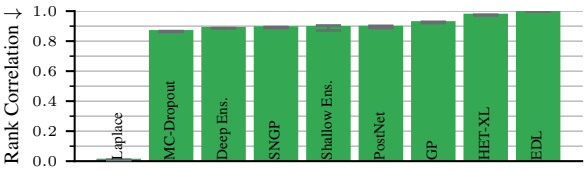


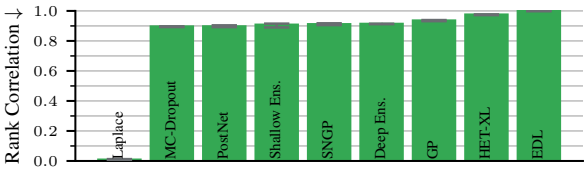
Figure K.8. On the CIFAR-10 test set, methods do not preserve their rankings on the ECE metric and GP degrades to the cross-entropy baseline already at severity level one.



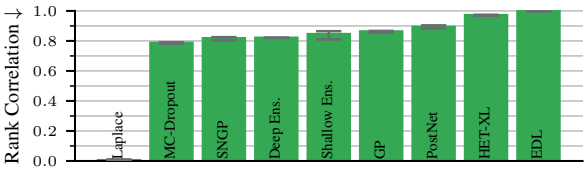
(a) Mixed ID and OOD with severity level one.



(b) OOD severity level one.

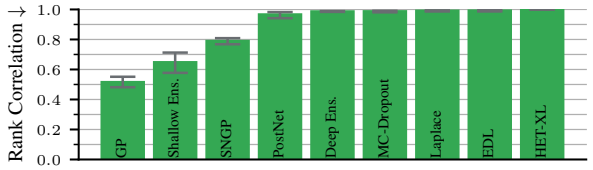


(c) Mixed ID and OOD with severity level two.

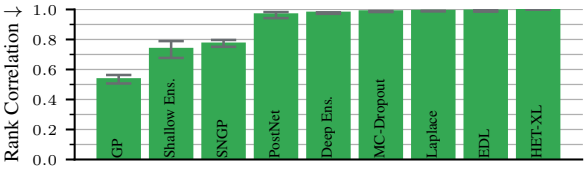


(d) OOD severity level two.

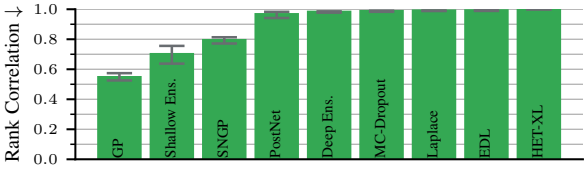
Figure L.1. OOD rank correlation results of the IT decomposition on ImageNet-ReaL using different distributional methods. Laplace has uncorrelated epistemic and aleatoric IT estimates across different OOD levels on the ImageNet validation set. Other methods do not show a significant drop in correlation even at severity level two. Only severity levels one and two are shown, as the GT aleatoric uncertainty values from the soft ImageNet-ReaL labels are only valid for these corruption levels – higher corruption would possibly lead to a shift in labeler votes.



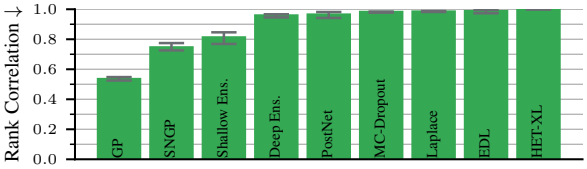
(a) Mixed ID and OOD with severity level one.



(b) OOD severity level one.

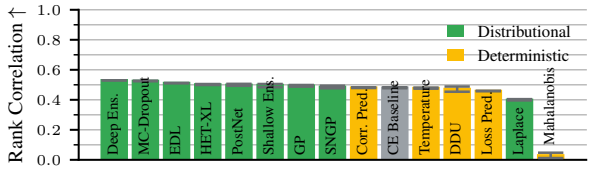


(c) Mixed ID and OOD with severity level two.

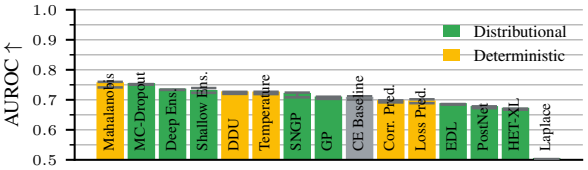


(d) OOD severity level two.

Figure L.2. OOD rank correlation results of the IT decomposition on CIFAR-10H using different distributional methods. GP shows a stably low rank correlation between the IT epistemic and aleatoric components across different OOD levels. Other methods do not show a significant drop in correlation even at severity level two. Only severity levels one and two are shown, as the GT aleatoric uncertainty values from the soft CIFAR-10H labels are only valid for these corruption levels – higher corruption would possibly lead to a shift in labeler votes.

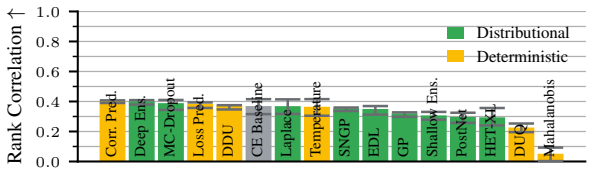


(a) Rank correlation with the ground-truth human uncertainties.

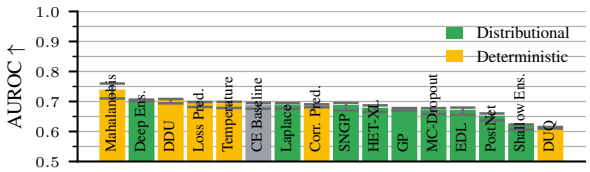


(b) AUROC of the OOD detection task using severity level two.

Figure L.3. On ImageNet, Laplace cannot match the cross-entropy baseline when using the estimators of the IT decomposition for the OOD detection and human uncertainty alignment tasks. All other methods are equipped with their best-performing estimator for the respective tasks, showing that specialized estimators work better.

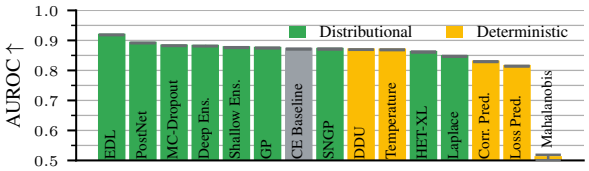


(a) Rank correlation between methods and the Bregman aleatoric component.

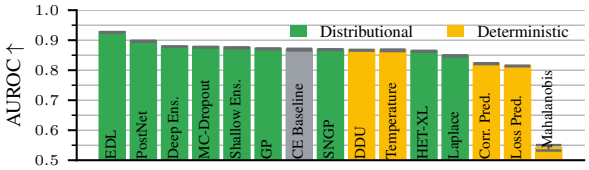


(b) AUROC of OOD detection performance of methods using perturbations of severity level two.

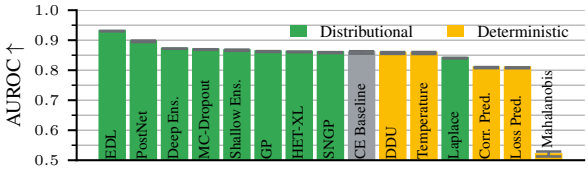
Figure L.4. SNGP, GP, and shallow ensemble underperform the cross-entropy baseline on CIFAR-10 when using the estimators of the IT decomposition for the OOD detection and human uncertainty alignment tasks. All other methods are equipped with their best-performing estimator for the respective tasks, showing that the IT decomposition is not practically beneficial.



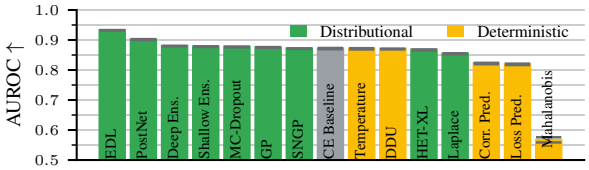
(a) ID results.



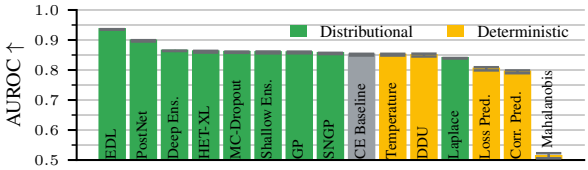
(b) Mixed ID and OOD severity level one.



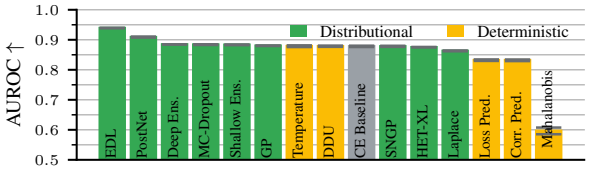
(c) OOD severity level one.



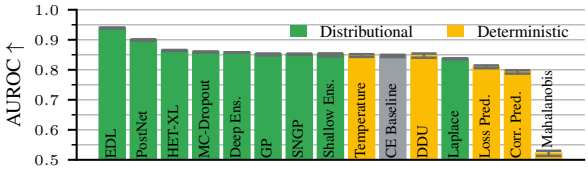
(d) Mixed ID and OOD severity level two.



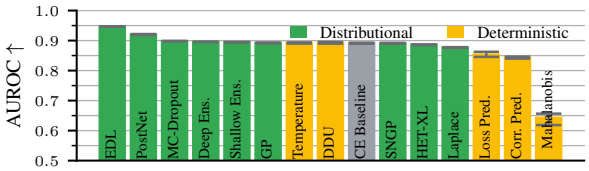
(e) OOD severity level two.



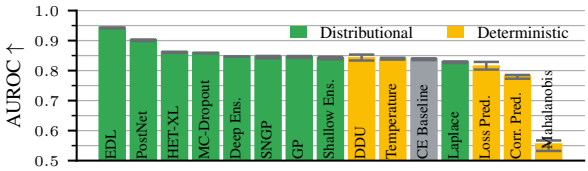
(f) Mixed ID and OOD severity level three.



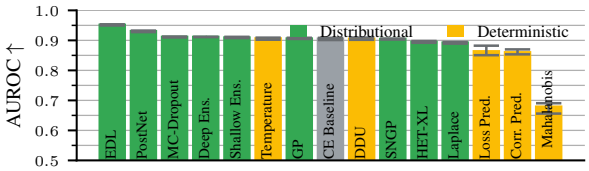
(g) OOD severity level three.



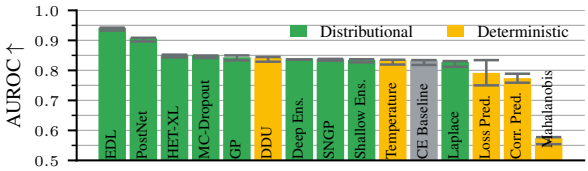
(h) Mixed ID and OOD severity level four.



(i) OOD severity level four.

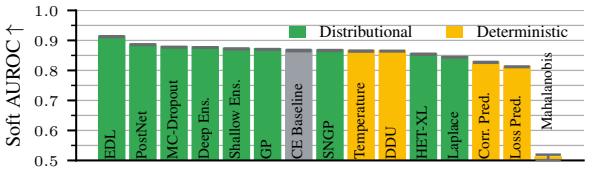


(j) Mixed ID and OOD severity level five.

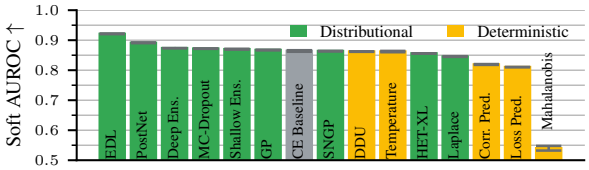


(k) OOD severity level five.

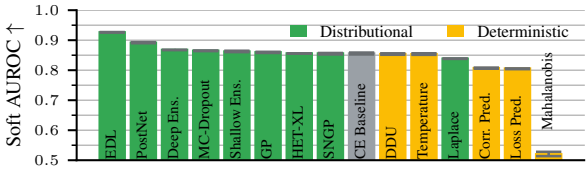
Figure M.1. On ImageNet, the evidential deep learning methods, EDL and PostNet, dominate on the correctness prediction task, which becomes even more pronounced as we go more and more OOD. The performance of Mahalanobis stably increases on mixed datasets, as models perform worse on OOD images than on ID ones, and it can detect OOD samples well.



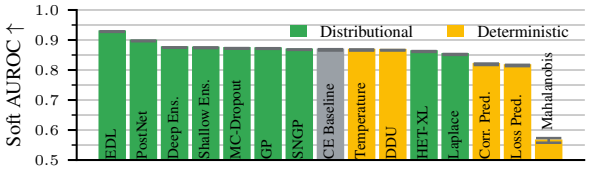
(a) ID results.



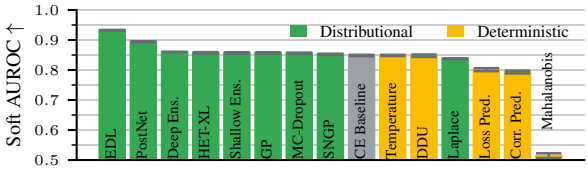
(b) Mixed ID and OOD severity level one.



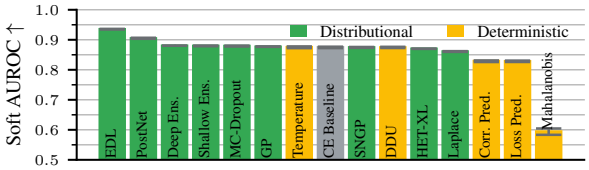
(c) OOD severity level one.



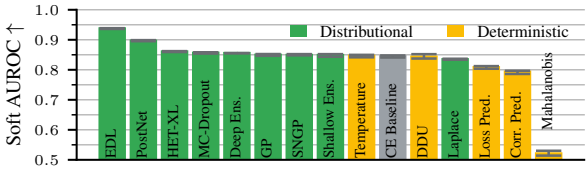
(d) Mixed ID and OOD severity level two.



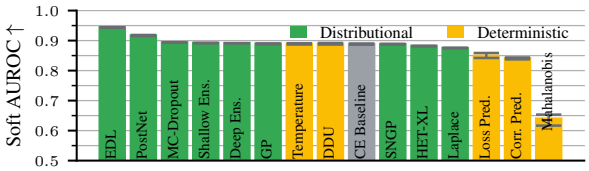
(e) OOD severity level two.



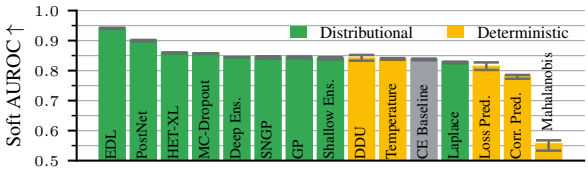
(f) Mixed ID and OOD severity level three.



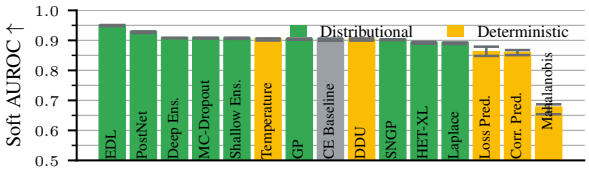
(g) OOD severity level three.



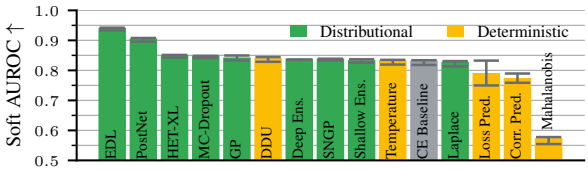
(h) Mixed ID and OOD severity level four.



(i) OOD severity level four.

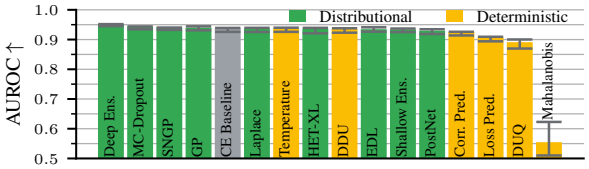


(j) Mixed ID and OOD severity level five.

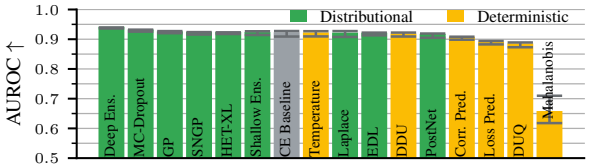


(k) OOD severity level five.

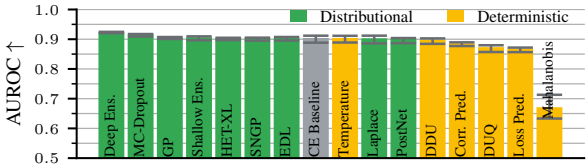
Figure M.2. Variant of Fig. M.1 where correctness is calculated w.r.t. the soft labels of ImageNet-ReaL. While many methods are saturated on the correctness prediction task, the rankings remain stable when switching from the usual notion of correctness to "soft correctness", where a prediction does not earn a binary indicator of correctness but rather the probability of the prediction's correctness.



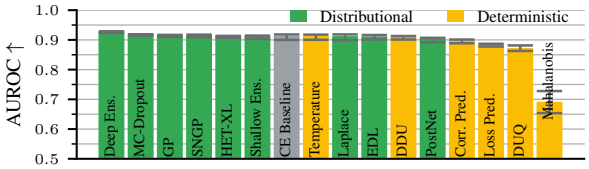
(a) ID results.



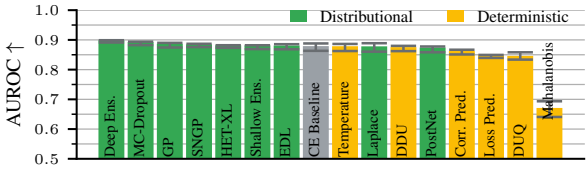
(b) Mixed ID and OOD severity level one.



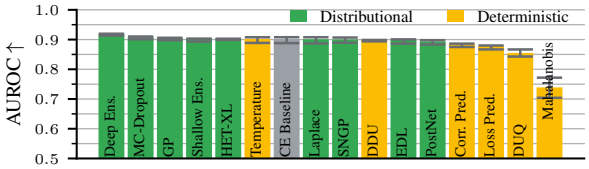
(c) OOD severity level one.



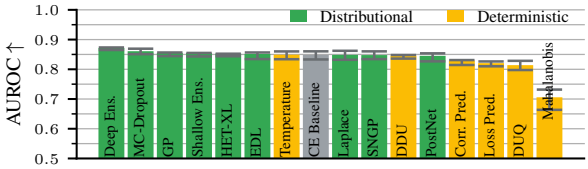
(d) Mixed ID and OOD severity level two.



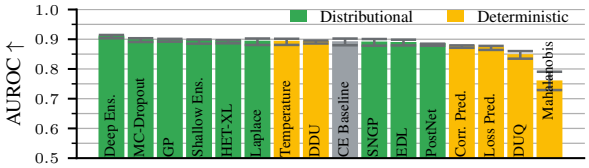
(e) OOD severity level two.



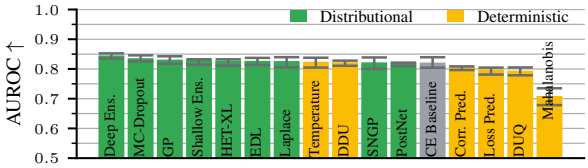
(f) Mixed ID and OOD severity level three.



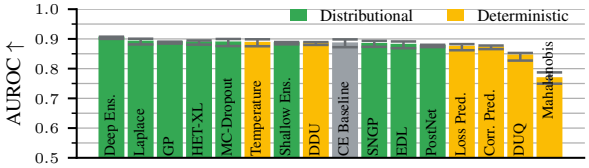
(g) OOD severity level three.



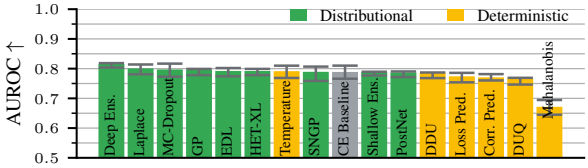
(h) Mixed ID and OOD severity level four.



(i) OOD severity level four.

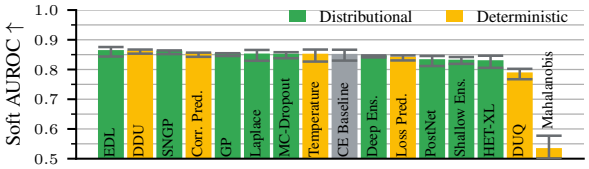


(j) Mixed ID and OOD severity level five.

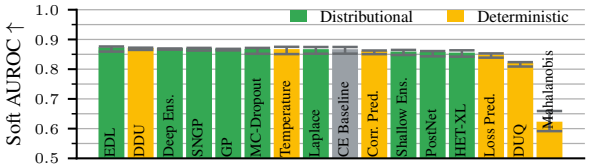


(k) OOD severity level five.

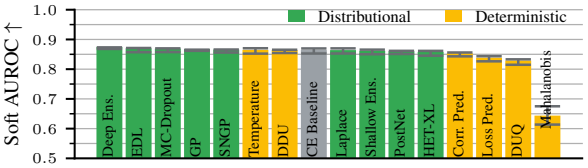
Figure M.3. On CIFAR-10, the performance of the methods consistently drops on the correctness prediction task, both on completely OOD datasets (right column) and on balanced mixtures of ID and OOD datasets (left column). A notable exception is the Mahalanobis method: here, we observe an increase in performance as we increase the level of severity. For the mixed datasets, this can be explained by the fact that the models perform worse on OOD images than on ID ones, making the Mahalanobis method a suitable estimator of correctness. However, the result is unexpected for solely OOD samples.



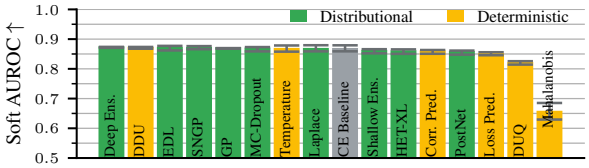
(a) ID results.



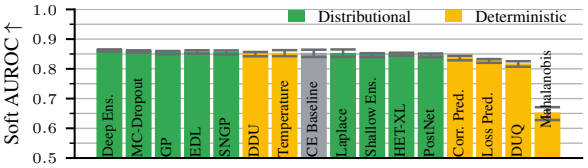
(b) Mixed ID and OOD severity level one.



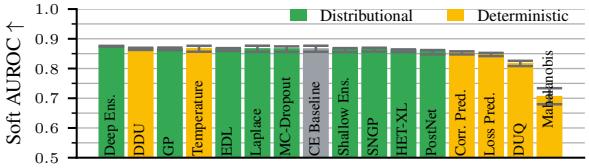
(c) OOD severity level one.



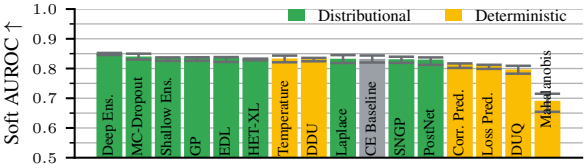
(d) Mixed ID and OOD severity level two.



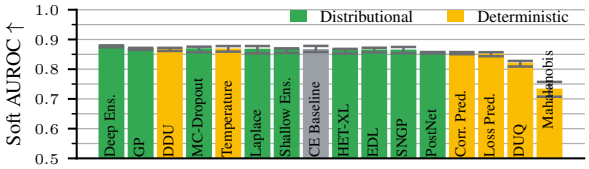
(e) OOD severity level two.



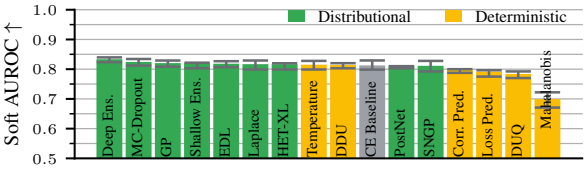
(f) Mixed ID and OOD severity level three.



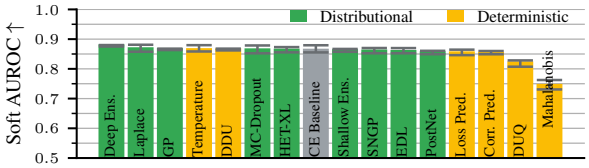
(g) OOD severity level three.



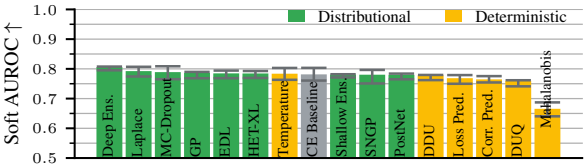
(h) Mixed ID and OOD severity level four.



(i) OOD severity level four.

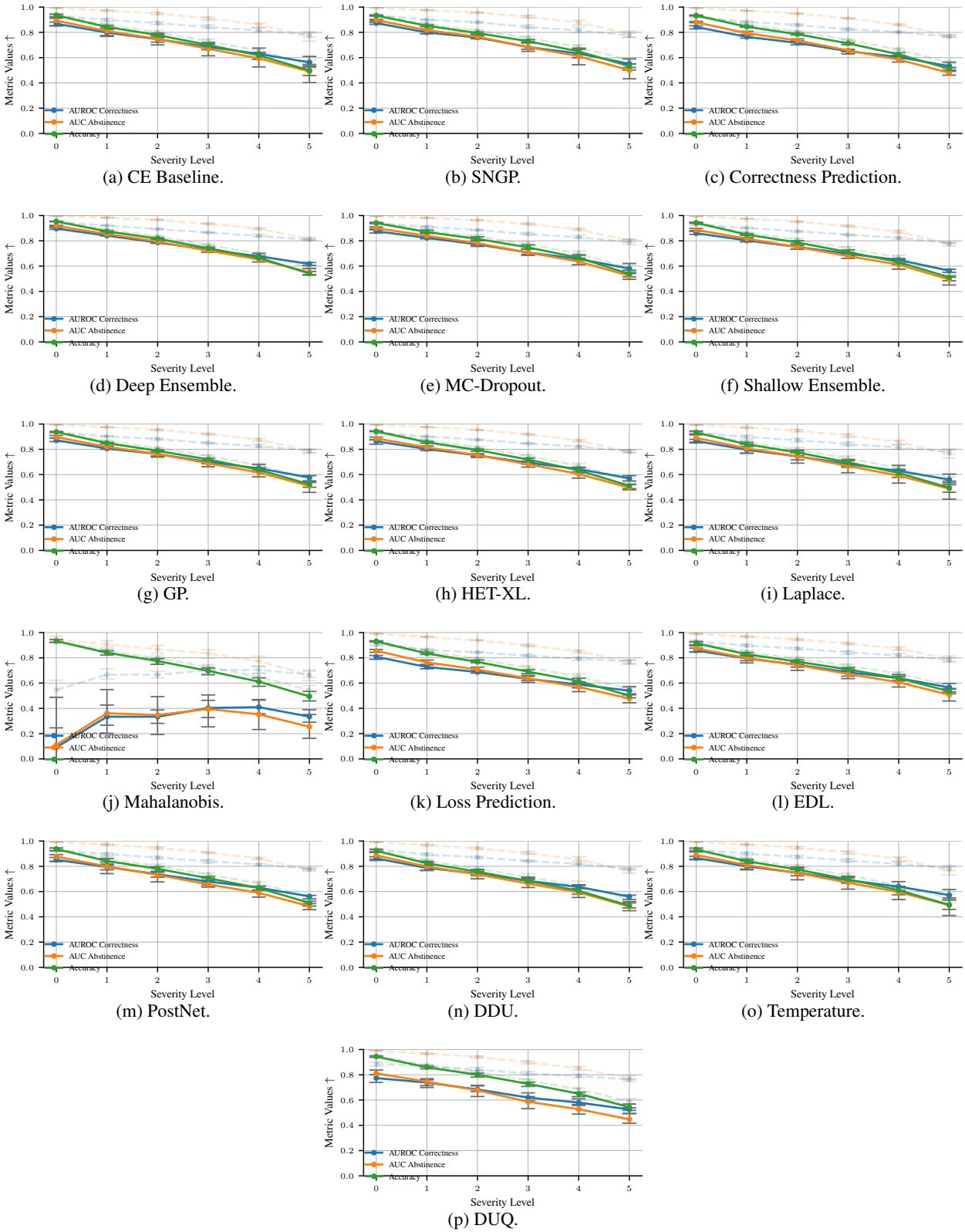


(j) Mixed ID and OOD severity level five.



(k) OOD severity level five.

Figure M.4. Variant of Fig. M.3 w.r.t. soft label correctness. Unlike Fig. M.2, slightly altering the evaluation criterion on the correctness prediction task changes the ranking of methods considerably. Soft AUROC uses the notion of "soft correctness", where a prediction does not earn a binary indicator of correctness but rather the probability of the prediction's correctness.



(a) CE Baseline. (b) SNGP. (c) Correctness Prediction.

(d) Deep Ensemble. (e) MC-Dropout. (f) Shallow Ensemble.

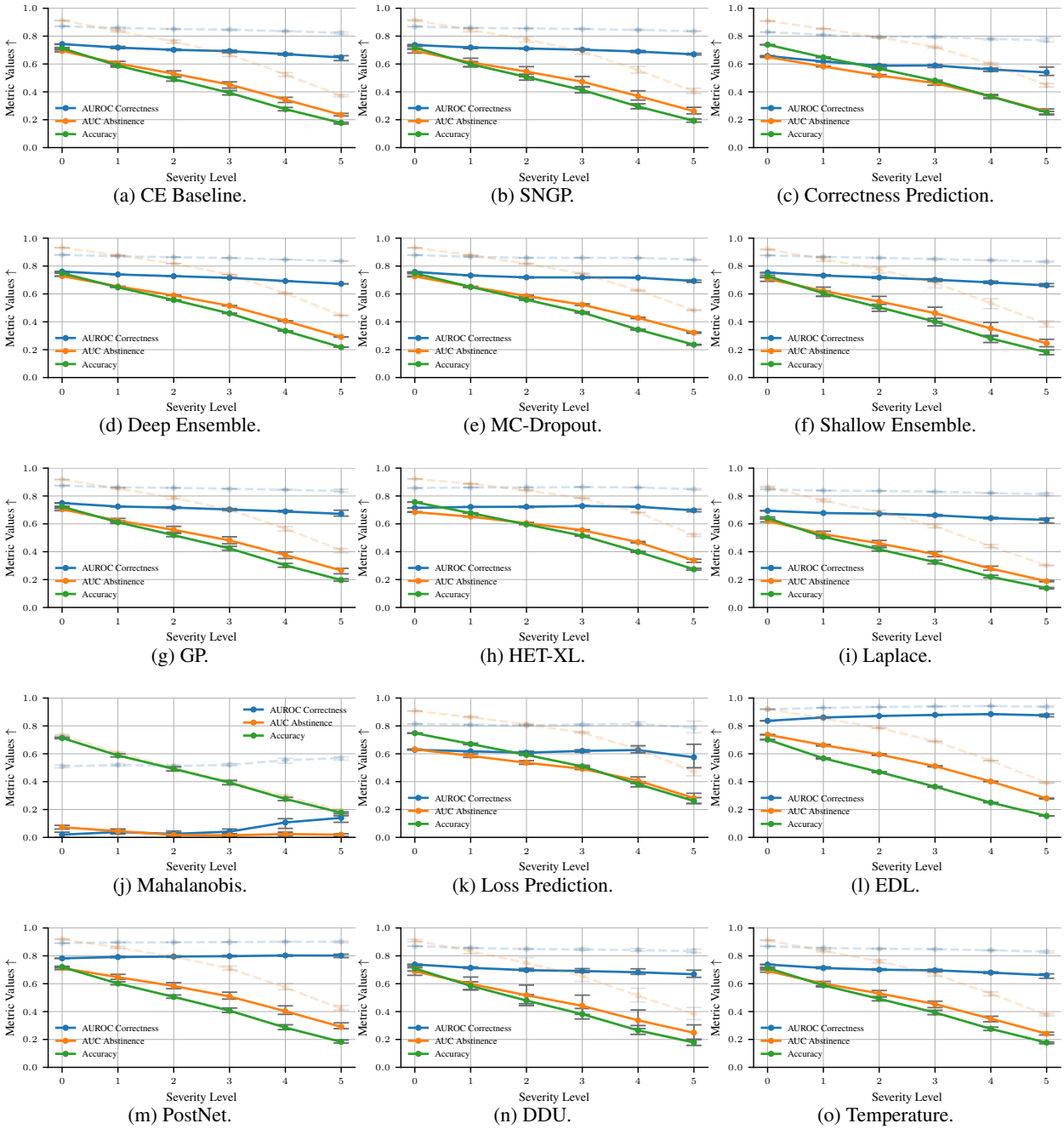
(g) GP. (h) HET-XL. (i) Laplace.

(j) Mahalanobis. (k) Loss Prediction. (l) EDL.

(m) PostNet. (n) DDU. (o) Temperature.

(p) DUQ.

Figure M.5. On CIFAR-10, all methods' performance deteriorates at the same rate as the model's accuracy on the correctness and abstinence tasks. The only exception is Mahalanobis, which is a specialized OOD detector.



(a) CE Baseline. (b) SNGP. (c) Correctness Prediction.

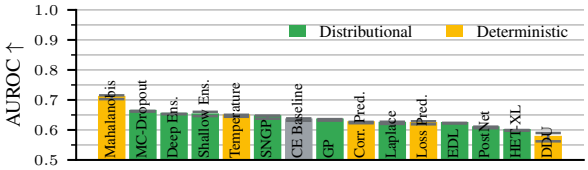
(d) Deep Ensemble. (e) MC-Dropout. (f) Shallow Ensemble.

(g) GP. (h) HET-XL. (i) Laplace.

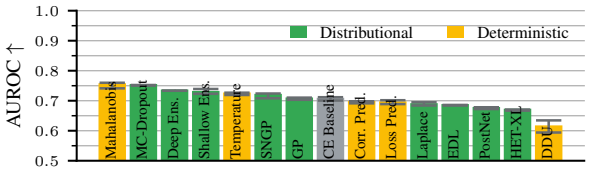
(j) Mahalanobis. (k) Loss Prediction. (l) EDL.

(m) PostNet. (n) DDU. (o) Temperature.

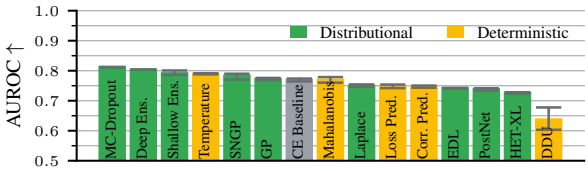
Figure M.6. On ImageNet, the estimate for predictive correctness is much more robust to OOD perturbations than the model's accuracy for all methods except Mahalanobis (which is a specialized OOD detector). The AUC abstinence score deteriorates at the same rate as the model's accuracy, which is an inherent property of the metric as the accuracy lower bounds the abstinence AUC metric.



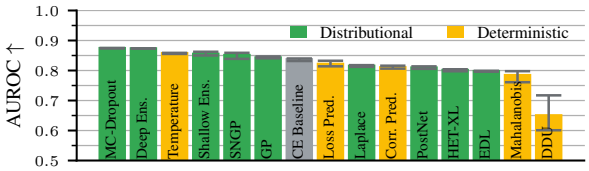
(a) AUROC OOD-ness with OOD severity level one.



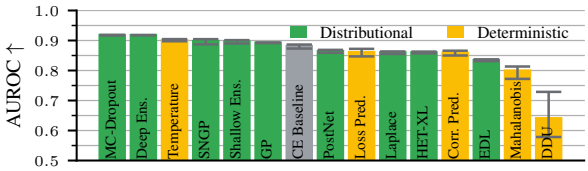
(b) AUROC OOD-ness with OOD severity level two.



(c) AUROC OOD-ness with OOD severity level three.

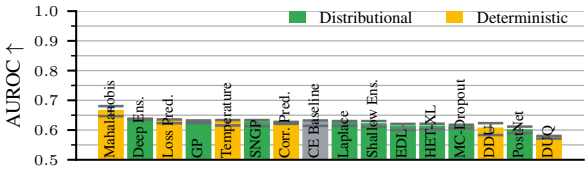


(d) AUROC OOD-ness with OOD severity level four.

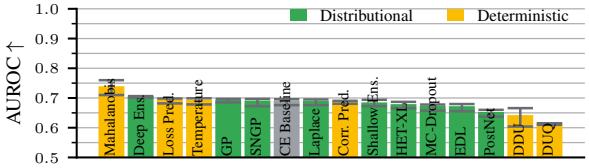


(e) AUROC OOD-ness with OOD severity level five.

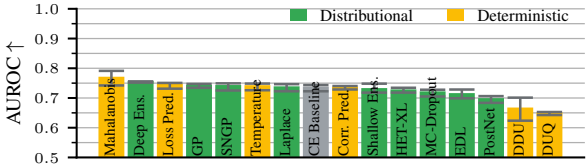
Figure M.7. The OOD detection performance of all methods increases steadily as we increase the severity of the perturbed half of the mixed dataset on the ImageNet validation dataset. However, the specialized OOD detector, Mahalanobis, generalizes worse than the other methods.



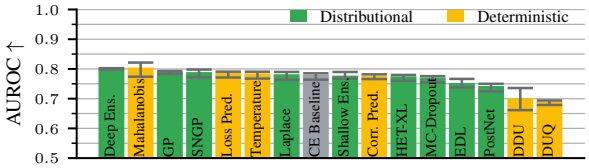
(a) AUROC OOD-ness with OOD severity level one.



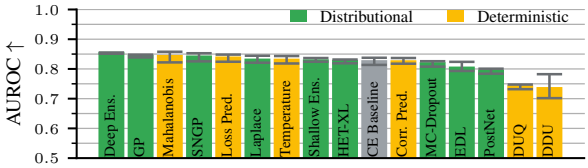
(b) AUROC OOD-ness with OOD severity level two.



(c) AUROC OOD-ness with OOD severity level three.



(d) AUROC OOD-ness with OOD severity level four.



(e) AUROC OOD-ness with OOD severity level five.

Figure M.8. On CIFAR-10, the OOD detection performance of all methods increases steadily as we increase the severity of the perturbed half of the mixed dataset. Mahalanobis generalizes worse than the other methods.

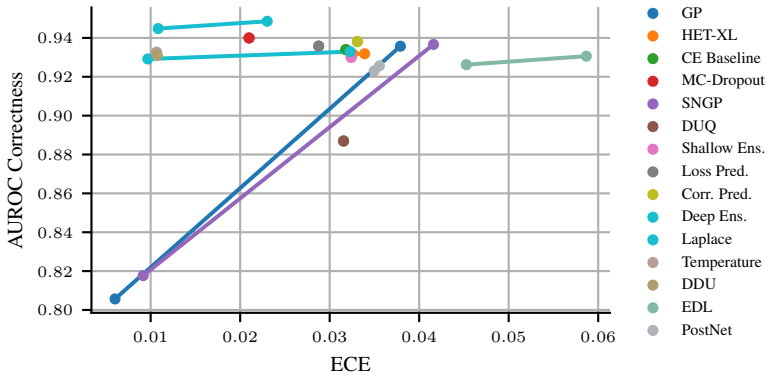


Figure M.9. Depending on what task we optimize the aggregator for, we obtain notably different results for SNGP variants, deep ensembles, and Laplace networks on CIFAR-10. Each color corresponds to one method. The point pairs per method show the performance of the method with an aggregator optimized for ECE and that optimized for the correctness AUROC.

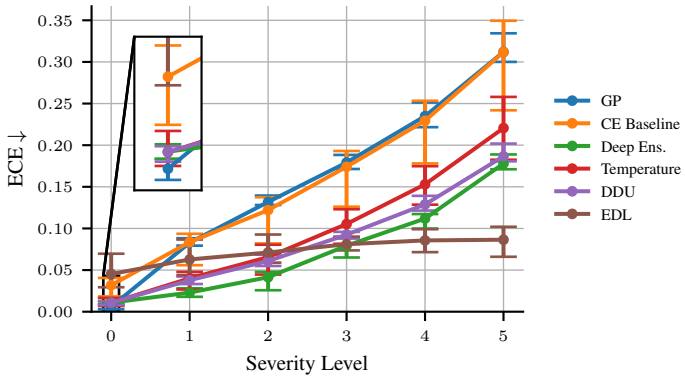
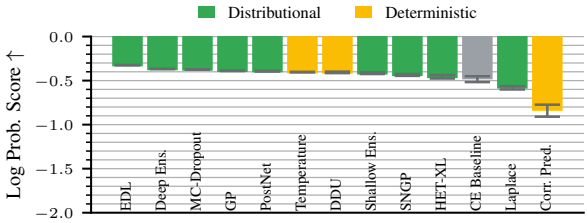
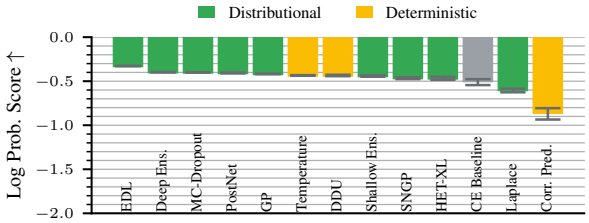


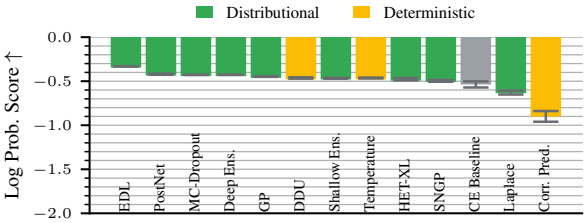
Figure M.10. On CIFAR-10, the EDL method is most robust to increasing the severity level of the CIFAR-10C perturbations on the ECE metric. While in-distribution the GP method performs best, it degrades to the cross-entropy baseline performance already at severity level one. The deep ensemble method is best for light perturbations.



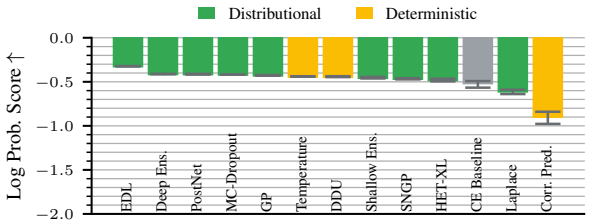
(a) ID results.



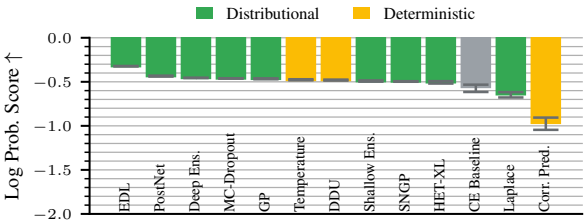
(b) Mixed ID and OOD severity level one.



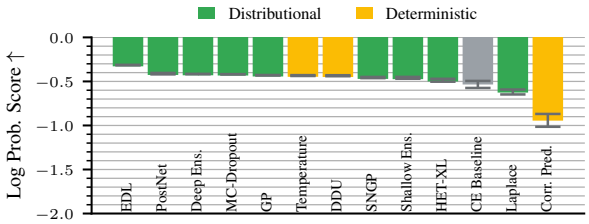
(c) OOD severity level one.



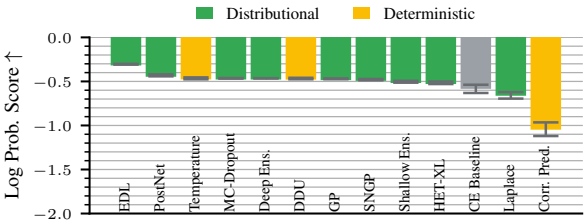
(d) Mixed ID and OOD severity level two.



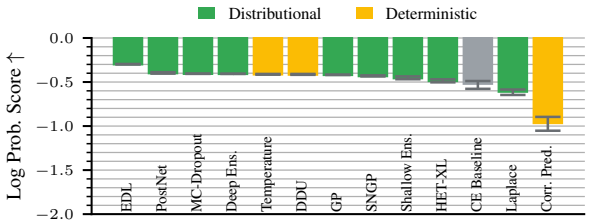
(e) OOD severity level two.



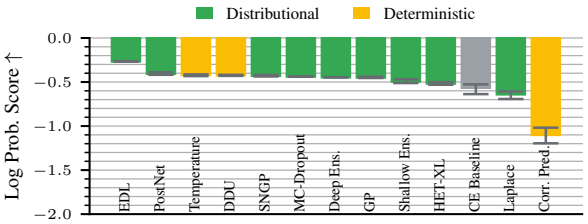
(f) Mixed ID and OOD severity level three.



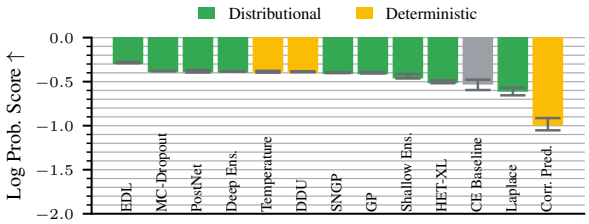
(g) OOD severity level three.



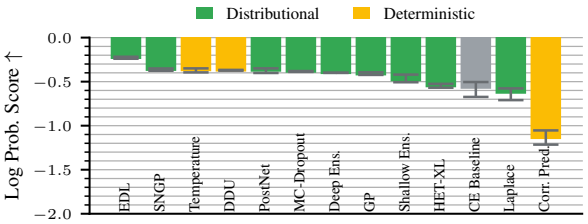
(h) Mixed ID and OOD severity level four.



(i) OOD severity level four.

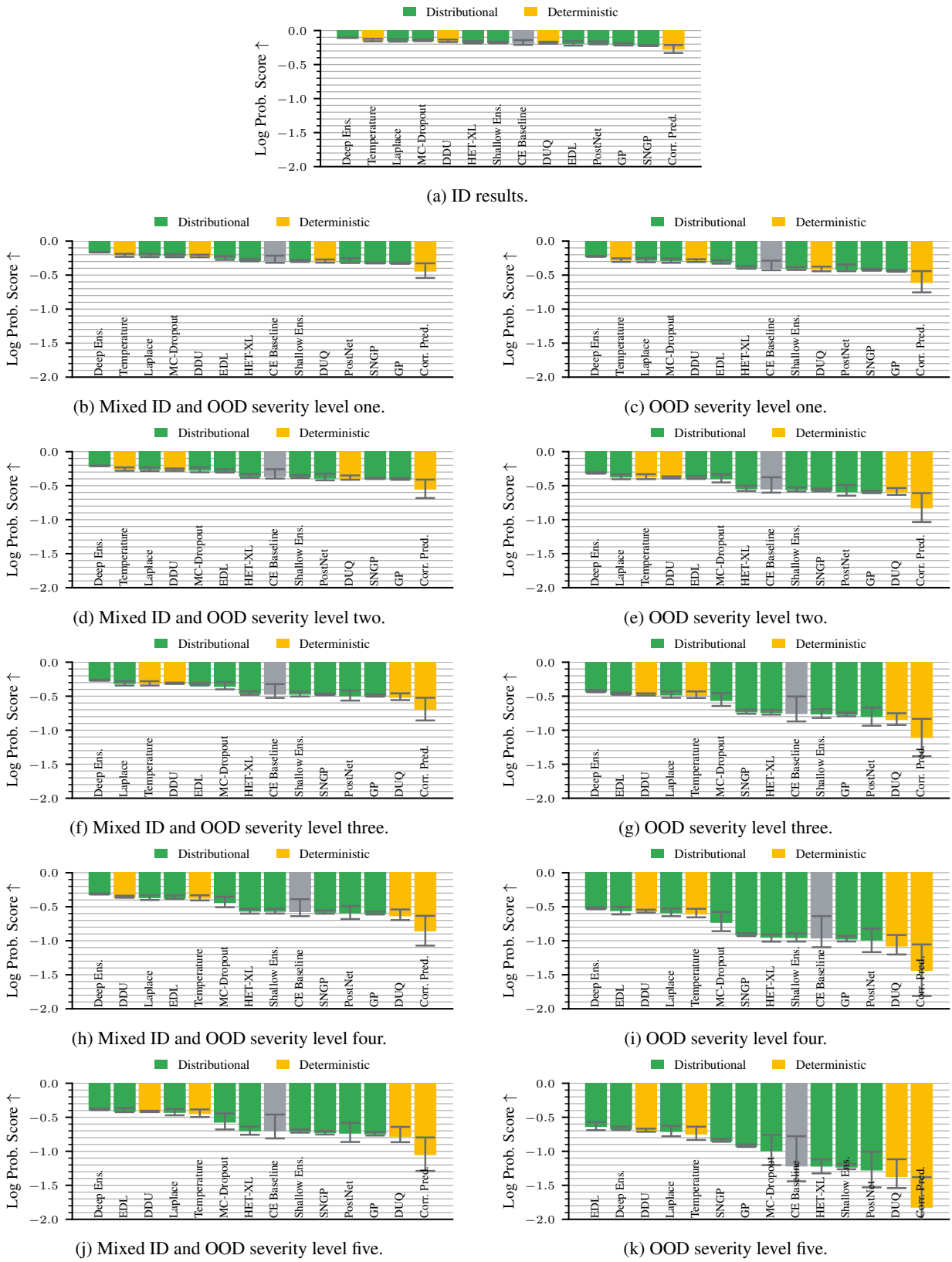


(j) Mixed ID and OOD severity level five.



(k) OOD severity level five.

Figure M.11. On ImageNet, most methods consistently outperform the cross-entropy baseline on average, both ID and OOD for all severity levels when evaluating on the log probability proper scoring rule. The EDL method performs best across all settings.



(a) ID results.

(b) Mixed ID and OOD severity level one.

(c) OOD severity level one.

(d) Mixed ID and OOD severity level two.

(e) OOD severity level two.

(f) Mixed ID and OOD severity level three.

(g) OOD severity level three.

(h) Mixed ID and OOD severity level four.

(i) OOD severity level four.

(j) Mixed ID and OOD severity level five.

(k) OOD severity level five.

Figure M.12. On CIFAR-10, the deep ensemble is a consistently robust method both ID and OOD for all severity levels when evaluating on the log probability proper scoring rule.

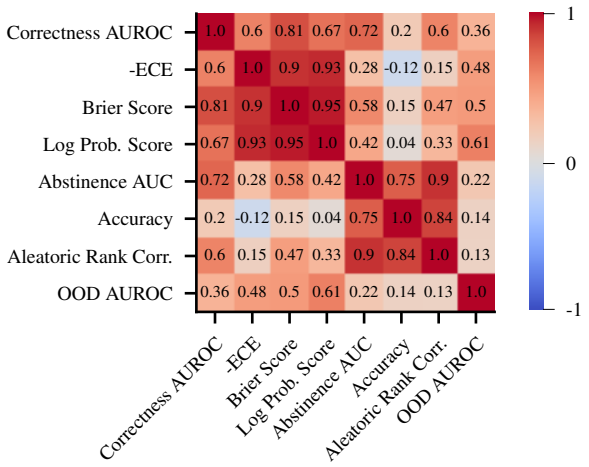


Figure M.13. Spearman correlation of metric pairs across all methods and aggregators on the ImageNet validation set. Only some of the considered metrics have a very high rank correlation among methods on the ImageNet validation dataset: most capture different aspects of uncertainty methods.

Figure O.1. Easy ImageNet-ReaL cases with no human disagreement on the labels. OOD samples are of severity two.

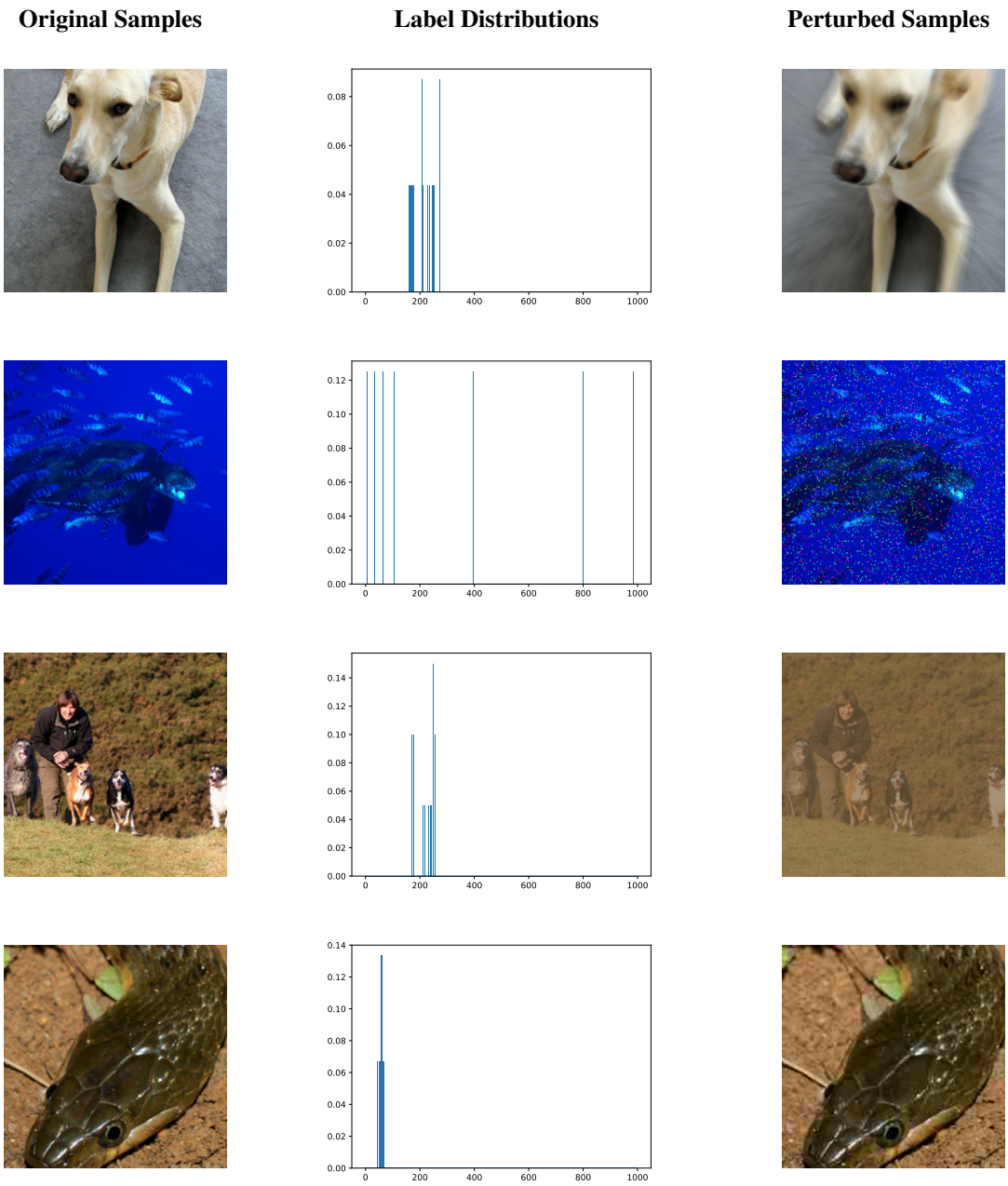


Figure O.2. Hard ImageNet-ReaL cases with high human uncertainty (i.e., high disagreement among annotators on the correct label). OOD samples are of severity two.

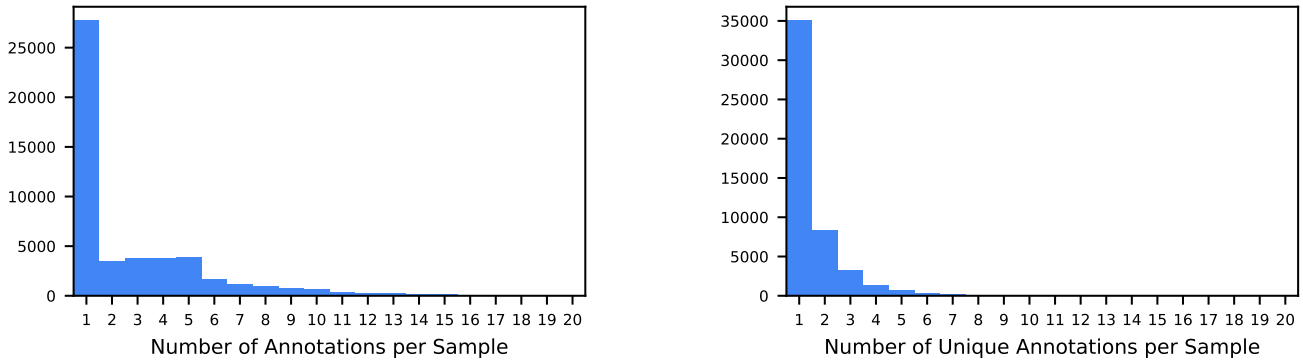


Figure O.3. Histograms of the label distributions of the ImageNet-ReaL validation set.

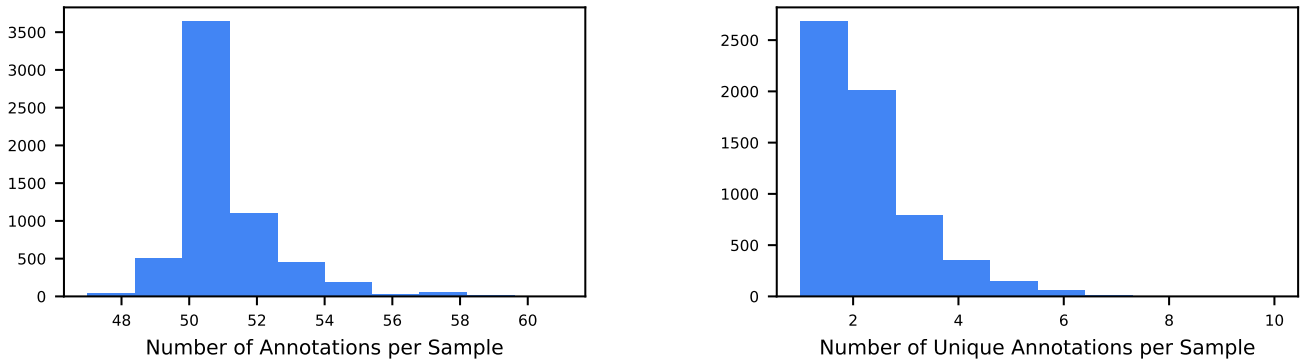


Figure O.4. Histograms of the label distributions of the CIFAR-10H validation set.