On Information-Theoretic Measures of Predictive Uncertainty

Kajetan Schweighofer[1] **Lukas Aichberger**[1] **Mykyta Ielanskyi**[1] **Sepp Hochreiter**[1,2]

[1]ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University Linz, Austria
[2]NXAI GmbH, Linz, Austria

Abstract

Reliable estimation of predictive uncertainty is crucial for machine learning applications, particularly in high-stakes scenarios where hedging against risks is essential. Despite its significance, there is no universal agreement on how to best quantify predictive uncertainty. In this work, we revisit core concepts to propose a framework for information-theoretic measures of predictive uncertainty. Our proposed framework categorizes predictive uncertainty measures according to two factors: **(I)** The predicting model **(II)** The approximation of the true predictive distribution. Examining all possible combinations of these two factors, we derive a set of predictive uncertainty measures that includes both known and newly introduced ones. We extensively evaluate these measures across a broad set of tasks, identifying conditions under which certain measures excel. Our findings show the importance of aligning the choice of uncertainty measure with the predicting model on in-distribution (ID) data, the limitations of epistemic uncertainty measures for out-of-distribution (OOD) data, and that the disentanglement between measures varies substantially between ID and OOD data. Together, these insights provide a more comprehensive understanding of predictive uncertainty measures, revealing their implicit assumptions and relationships.

1 INTRODUCTION

Integrating machine learning models into high-stakes scenarios, such as autonomous driving or managing critical healthcare systems, introduces substantial risks. To hedge against these risks, we need to quantify the uncertainty associated with each prediction to prevent models from making decisions that carry significant risk and uncertainty. In such cases, it is better to defer uncertain decisions to human experts or opt for a safer, though potentially less advantageous, alternative decision. Consequently, it is vital to employ reliable measures of predictive uncertainty and provide estimates for them when implementing machine learning models for decision making in high-stakes applications.

The entropy of the posterior predictive distribution has become the standard information-theoretic measure to assess predictive uncertainty [Houlsby et al., 2011, Gal, 2016, Depeweg et al., 2018, Smith and Gal, 2018, Mukhoti et al., 2023]. Despite its widespread use, this measure has drawn criticism [Malinin and Gales, 2021, Wimmer et al., 2023], prompting the proposal of alternative information-theoretic measures [Malinin and Gales, 2021, Schweighofer et al., 2023b,a] and frameworks of uncertainty measures based on proper scoring rules that encompass information-theoretic measures [Kotelevskii et al., 2025, Hofman et al., 2024b]. However, it remains hard to gauge in which practical settings one measure should be favored over another.

We show that all these measures can be interpreted as approximations of the same measure, the cross-entropy between the predictive distributions of the predicting model and the true model. The predicting model is used to predict on new data, while the true model generated the dataset. However, since we do not know the true model, this fundamental measure is intractable to compute. Therefore, we consider different assumptions about the predicting model and how the true predictive distribution is approximated. This gives rise to our proposed framework to categorize information-theoretic measures of predictive uncertainty. Our framework includes existing measures, introduces new ones, and clarifies their relationships by eliciting implicit assumptions. In sum, our contributions are as follows:

- We present a unifying framework to categorize predictive uncertainty measures according to assumptions about the predicting model and how the true predictive distribution is approximated. This framework encompasses existing measures, but also suggests new ones, clarifies their rela-

tionship, and guides the selection of appropriate measures.

- We evaluate the measures suggested by our framework across a broad set of tasks, identifying specific conditions under which certain measures outperform others in practice. Notably, for tasks depending on which model predicts, it is best to align the uncertainty measure with the predicting model. Furthermore, we assess the level of disentanglement between measures, finding a pronounced dependency on the distribution the new data is coming from.

2 QUANTIFYING PREDICTIVE UNCERTAINTY

We consider the canonical classification setting with inputs $\boldsymbol{x} \in \mathbb{R}^D$ and targets $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all K possible targets. The dataset $\mathcal{D}$ is given, sampled i.i.d. according to the data generating distribution. We consider deep neural networks as a class of probabilistic models that map an input $\boldsymbol{x}$ to the $K-1$ dimensional probability simplex $\Delta^{K-1} = \{\boldsymbol{\theta} \in \mathbb{R}^K \mid \theta_k \geq 0 \; \forall k, \sum_{k=1}^K \theta_k = 1\}$. This mapping is defined as $f_{\boldsymbol{w}} : \mathbb{R}^D \to \Delta^{K-1}$ for a model with parameters $\boldsymbol{w}$. The output of this mapping defines the distribution parameters of a categorical distribution, in the following referred to as the model's predictive distribution $p(y \mid \boldsymbol{x}, \boldsymbol{w}) = \mathrm{Cat}(y; f_{\boldsymbol{w}}(\boldsymbol{x})) = \mathrm{Cat}(y; \boldsymbol{\theta})$.

The predictive distribution of a probabilistic model represents the uncertainty inherent in its predictions. When the probability mass is uniformly distributed across all possible outcomes, it denotes complete uncertainty about the prediction, whereas concentration on a single class indicates complete certainty. If we have access to the true data-generating model, denoted by parameters $\boldsymbol{w}^*$, the predictive distribution $p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)$ captures the inherent and irreducible uncertainty in the prediction, often referred to as *aleatoric uncertainty* (AU) [Gal, 2016, Kendall and Gal, 2017]. This assumes that the chosen model class can accurately represent the true predictive distribution, thus $p(y \mid \boldsymbol{x}) = p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)$, which is a common and often necessary assumption [Hüllermeier and Waegeman, 2021]. We will discuss how to interpret AU if we do not have access to $\boldsymbol{w}^*$ in Sec. 2.2. The information-theoretic entropy $\mathrm{H}(\cdot)$ [Shannon, 1948] of the true predictive distribution is a natural and universally accepted measure of AU, defined as

$$\mathrm{H}(p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)) \;:=\; \mathrm{E}_{p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)} \left[-\log p(y \mid \boldsymbol{x}, \boldsymbol{w}^*) \right] \quad (1)$$

However, we generally don't know the true model and have to choose parameters $\boldsymbol{w}$ out of all possible ones. Consequently, uncertainty arises due to the lack of knowledge about the true parameters of the model. This is called *epistemic uncertainty* (EU) [Apostolakis, 1990, Helton, 1993, 1997, Gal, 2016, Smith and Gal, 2018]. An effective measure of predictive uncertainty should be consistent with Eq. (1) and capture both AU and EU, typically assumed to sum up to a *total uncertainty* (TU).

2.1 CURRENT STANDARD MEASURE: ENTROPY OF POSTERIOR PREDICTIVE DISTRIBUTION

Given a dataset $\mathcal{D}$ and prior $p(\boldsymbol{w})$ on the model parameters, Bayes' theorem yields the posterior distribution $p(\boldsymbol{w} \mid \mathcal{D})$. The posterior distribution denotes the probability that the parameters $\boldsymbol{w}$ match the true parameters $\boldsymbol{w}^*$ of the model that generated the dataset $\mathcal{D}$. Instead of committing to a single model, the posterior distribution allows marginalizing over all possible models, which is known as Bayesian model averaging. This gives rise to the posterior predictive distribution

$$p(y \mid \boldsymbol{x}, \mathcal{D}) \;=\; \mathrm{E}_{p(\boldsymbol{w} \mid \mathcal{D})} \left[p(y \mid \boldsymbol{x}, \boldsymbol{w}) \right] \,. \quad (2)$$

The entropy of the posterior predictive distribution is the currently most widely accepted approach to measure predictive uncertainty [Houlsby et al., 2011, Gal, 2016, Depeweg et al., 2018, Smith and Gal, 2018, Hüllermeier and Waegeman, 2021, Mukhoti et al., 2023]. According to a well-known result from information theory [Cover and Thomas, 2006], this entropy can be additively decomposed into the conditional entropy and the mutual information $\mathrm{I}(\cdot)$ between y and $\boldsymbol{w}$:

$$\begin{aligned} &\underbrace{\mathrm{H}(p(y \mid \boldsymbol{x}, \mathcal{D}))}_{\text{TU}} \quad (3) \\ &= \underbrace{\mathrm{E}_{p(\boldsymbol{w} \mid \mathcal{D})} \left[\mathrm{H}(p(y \mid \boldsymbol{x}, \boldsymbol{w})) \right]}_{\text{AU}} + \underbrace{\mathrm{I}(p(y, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D}))}_{\text{EU}} \,. \end{aligned}$$

Eq. (3) is equivalent to a decomposition of expected cross-entropy $\mathrm{CE}(\cdot \; ; \; \cdot)$ into conditional entropy and expected KL-divergence $\mathrm{KL}(\cdot \parallel \cdot)$ [Schweighofer et al., 2023b,a]:

$$\begin{aligned} &\underbrace{\mathrm{E}_{p(\boldsymbol{w} \mid \mathcal{D})} \left[\mathrm{CE}(p(y \mid \boldsymbol{x}, \boldsymbol{w}) \; ; \; p(y \mid \boldsymbol{x}, \mathcal{D})) \right]}_{\text{TU}} \quad (4) \\ &\quad = \underbrace{\mathrm{E}_{p(\boldsymbol{w} \mid \mathcal{D})} \left[\mathrm{H}(p(y \mid \boldsymbol{x}, \boldsymbol{w})) \right]}_{\text{AU}} \\ &\quad + \underbrace{\mathrm{E}_{p(\boldsymbol{w} \mid \mathcal{D})} \left[\mathrm{KL}(p(y \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(y \mid \boldsymbol{x}, \mathcal{D})) \right]}_{\text{EU}} \,. \end{aligned}$$

If the parameters of the true model are known, EU vanishes and Eq. (3) as well as Eq. (4) simplify to Eq. (1), thus are consistent with it. However, the entropy of the posterior predictive distribution has been found to be inadequate for specific scenarios, such as autoregressive predictions [Malinin and Gales, 2021] or for a given predicting model [Schweighofer et al., 2023b] and was criticised on grounds of not fulfulling certain expected theoretical properties [Wimmer et al., 2023]. In response, alternative information-theoretic measures [Malinin and Gales, 2021, Schweighofer et al., 2023b,a] and frameworks of uncertainty measures based on proper scoring rules [Kotelevskii et al., 2025, Hofman et al., 2024b] have been introduced. We seek to give a unified framework for these measures that is interpretable and can guide the selection of uncertainty measures in practice. Therefore, we next propose a fundamental, yet generally intractable, predictive uncertainty measure, where all of these measures are special cases under specific assumptions.

2.2 OUR PROPOSED MEASURE: CROSS-ENTROPY BETWEEN SELECTED AND TRUE PREDICTIVE DISTRIBUTION

An effective measure of TU should be consistent with Eq. (1) and should incorporate EU. Given this, we propose to measure predictive uncertainty with the cross-entropy between the predictive distributions of a selected predicting model and the true model. Let $p(y \mid \boldsymbol{x}, \cdot)$ be the predictive distribution of any selected model for some new input $\boldsymbol{x}$, which we will refer to as the *predicting model*. We will examine different cases for the predicting model later; for now, it suffices to consider it to be a specific model with parameters $\boldsymbol{w}$. The cross-entropy between the predictive distributions of the predicting model and the true model is given by

$$\begin{aligned} &\underbrace{\mathrm{CE}(p(y \mid \boldsymbol{x}, \cdot)\ ;\ p(y \mid \boldsymbol{x}, \boldsymbol{w}^*))}_{\text{TU}} \\ &:= \mathrm{E}_{p(y \mid \boldsymbol{x}, \cdot)}\left[-\log p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)\right] \\ &= \underbrace{\mathrm{H}(p(y \mid \boldsymbol{x}, \cdot))}_{\text{AU}} + \underbrace{\mathrm{KL}(p(y \mid \boldsymbol{x}, \cdot) \parallel p(y \mid \boldsymbol{x}, \boldsymbol{w}^*))}_{\text{EU}} . \end{aligned} \tag{5}$$

If the predictive distribution of the predicting model is equal to the predictive distribution of the true model, the EU is zero by definition of the KL-divergence and Eq. (5) simplifies to Eq. (1). Thus, as expected, if the parameters of the true model are known, the EU vanishes. Eq. (5) is a fundamental, though generally intractable, measure of predictive uncertainty. To obtain tractable measures, assumptions about the predicting model and about how to approximate the true predictive distribution are necessary. This gives rise to our framework, which we introduce in detail in Sec. 3. Notably, the opposite order of the arguments in Eq. (5) leads to the same framework but differs in the interpretation of AU and EU. However, we find that Eq. (5) is more useful in practice, allowing to align the uncertainty measure to the predicting model. For details on the alternative measure see Apx. A.5.

Interpretation of AU and EU. An important distinction compared to previous work is in our interpretation of AU and EU, which aligns with the understanding of Apostolakis [1990], Helton [1993, 1997], Schweighofer et al. [2023b] as follows. The AU is not generally understood as a property of the true predictive distribution, but of the selected predicting model used to make a prediction. Thus, it is the uncertainty that arises due to predicting with the selected probabilistic model. The EU is defined as the additional uncertainty due to predicting with the selected predicting model instead of the true model. Thus, it is the additional uncertainty that arises due to selecting a model from the given model class.

3 PROPOSED FRAMEWORK

Our proposed measure of predictive uncertainty (Eq. (5)) allows for different assumptions about (I) the selected predicting model and (II) how to approximate the true predictive distribution. We consider three different assumptions for each, denoted as `(A,B,C)` for the predicting model and `(1,2,3)` for the approximation of the true predictive distribution. This results in nine distinct predictive uncertainty measures within our proposed framework. An overview of all measures is given in Tab. 1, summarizing possible measures of total uncertainty (TU), as well as their respective aleatoric uncertainty (AU) and epistemic uncertainty (EU).

`(A,B,C)`: PREDICTING MODEL

One can make different choices about the model used during inference for predicting the class of a new input. The most obvious choice of a predicting model is `(A)` a pre-selected given model with parameters $\boldsymbol{w}$. This is the standard case in machine learning, where model parameters are selected, e.g. by maximizing the likelihood on the training dataset or downloaded from a model hub.

Another widely used approach is `(B)` the Bayesian model average (see Eq. (2)). Here, instead of predicting with a single model, the predictive distribution is marginalized over all possible models according to their posterior probability. In practice, exact marginalization is generally intractable and therefore approximated by posterior sampling.

Finally, it is possible to `(C)` consider every possible model as the predicting model, weighted by their posterior probabilities. This might seem counterintuitive, as it means that the predicting model is not fixed but is sampled anew for each prediction. Nevertheless, the AU of the resulting uncertainty measures, $\mathrm{E}_{p(\boldsymbol{w} \mid \mathcal{D})}\left[\mathrm{H}(p(y \mid \boldsymbol{x}, \boldsymbol{w}))\right]$, is the best approximation of the AU under the true model for a given posterior distribution [Schweighofer et al., 2023a]. However, as pointed out by Wimmer et al. [2023], it is neither a lower nor an upper bound on the AU under the true model and is highly dependent on the posterior distribution.

`(1,2,3)`: APPROXIMATION OF THE TRUE PREDICTIVE DISTRIBUTION

One can also make different choices about how to approximate the true predictive distribution. The simplest but probably biased choice to approximate the true predictive distribution is `(1)` the predictive distribution under a single given model with parameters $\tilde{\boldsymbol{w}}$. Although this might be a poor approximation, it might be the only feasible choice in specific settings. For example, it is used in speculative decoding [Stern et al., 2018, Leviathan et al., 2023], where a small model is used to predict and whose predictive distribution is compared against a large model that serves as ground truth. Furthermore, $\tilde{\boldsymbol{w}}$ does not necessarily approximate the true model, but instead serves as a reference model of interest, such as a previously used or competitor model, whose discrepancy from the predicting model should be captured.

Table 1: **Our proposed framework of information-theoretic measures of predictive uncertainty.** Each measure denotes a different approximation of the fundamental measure given by Eq. (5) for different assumptions about the predicting model and how the true model is approximated. For brevity, we define $p_{\boldsymbol{w}} \coloneqq p(y \mid \boldsymbol{x}, \boldsymbol{w})$, $p_{\mathcal{D}} \coloneqq p(y \mid \boldsymbol{x}, \mathcal{D})$, and $\mathrm{E}_{\boldsymbol{w}} \coloneqq \mathrm{E}_{p(\boldsymbol{w}\mid\mathcal{D})}$ (the same for $\tilde{\boldsymbol{w}}$). Expressions with the same cell coloring are equivalent to each other. For each measure, TU additively decomposes into AU and EU by $\mathrm{CE}(p\ ;\ q) = \mathrm{H}(p) + \mathrm{KL}(p \parallel q)$.

	Predicting model	Approximation of the true predictive distribution		
		(1) $\tilde{\boldsymbol{w}}$	(2) $\mathrm{E}_{\tilde{\boldsymbol{w}}}$	(3) $\tilde{\boldsymbol{w}} \sim p(\tilde{\boldsymbol{w}} \mid \mathcal{D})$
TU	(A) $\boldsymbol{w}$	$\mathrm{CE}(p_{\boldsymbol{w}}\ ;\ p_{\tilde{\boldsymbol{w}}})$	$\mathrm{CE}(p_{\boldsymbol{w}}\ ;\ p_{\mathcal{D}})$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{CE}(p_{\boldsymbol{w}}\ ;\ p_{\tilde{\boldsymbol{w}}})\right]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$\mathrm{CE}(p_{\mathcal{D}}\ ;\ p_{\tilde{\boldsymbol{w}}})$	$\mathrm{CE}(p_{\mathcal{D}}\ ;\ p_{\mathcal{D}})$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{CE}(p_{\mathcal{D}}\ ;\ p_{\tilde{\boldsymbol{w}}})\right]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{CE}(p_{\boldsymbol{w}}\ ;\ p_{\tilde{\boldsymbol{w}}})\right]$	$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{CE}(p_{\boldsymbol{w}}\ ;\ p_{\mathcal{D}})\right]$	$\mathrm{E}_{\boldsymbol{w}}\left[\,\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{CE}(p_{\boldsymbol{w}}\ ;\ p_{\tilde{\boldsymbol{w}}})\right]\right]$
AU	(A) $\boldsymbol{w}$	$\mathrm{H}(p_{\boldsymbol{w}})$	$\mathrm{H}(p_{\boldsymbol{w}})$	$\mathrm{H}(p_{\boldsymbol{w}})$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$\mathrm{H}(p_{\mathcal{D}})$	$\mathrm{H}(p_{\mathcal{D}})$	$\mathrm{H}(p_{\mathcal{D}})$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{H}(p_{\boldsymbol{w}})\right]$	$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{H}(p_{\boldsymbol{w}})\right]$	$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{H}(p_{\boldsymbol{w}})\right]$
EU	(A) $\boldsymbol{w}$	$\mathrm{KL}(p_{\boldsymbol{w}} \parallel p_{\tilde{\boldsymbol{w}}})$	$\mathrm{KL}(p_{\boldsymbol{w}} \parallel p_{\mathcal{D}})$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{KL}(p_{\boldsymbol{w}} \parallel p_{\tilde{\boldsymbol{w}}})\right]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$\mathrm{KL}(p_{\mathcal{D}} \parallel p_{\tilde{\boldsymbol{w}}})$	$\cancelto{0}{\mathrm{KL}(p_{\mathcal{D}} \parallel p_{\mathcal{D}})}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{KL}(p_{\mathcal{D}} \parallel p_{\tilde{\boldsymbol{w}}})\right]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{KL}(p_{\boldsymbol{w}} \parallel p_{\tilde{\boldsymbol{w}}})\right]$	$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{KL}(p_{\boldsymbol{w}} \parallel p_{\mathcal{D}})\right]$	$\mathrm{E}_{\boldsymbol{w}}\left[\,\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{KL}(p_{\boldsymbol{w}} \parallel p_{\tilde{\boldsymbol{w}}})\right]\right]$

Another possibility is to use (2) the posterior predictive distribution. Although intuitively appealing, Schweighofer et al. [2023a] criticized this as there is no guarantee that the posterior predictive distribution approaches the true predictive distribution, even for a perfect estimate of the posterior predictive distribution. Furthermore, there are degenerate cases where the posterior predictive distribution can't be represented by any model with non-vanishing posterior probability. However, it is often a well-performing approximation empirically for expressive models such as neural networks. Moreover, (2) is the only option that guarantees finite EU and TU (for cases (B2) and (C2)).

Finally, perhaps the most intuitive approach is to consider (3) all possible models according to their posterior probability. Since any model could be the true model under the posterior, we should consider the mismatch between the predictive distribution of the predicting model and that of all possible models, weighted by their posterior probability.

3.1 RELATIONSHIPS BETWEEN MEASURES

Importantly, the AU of all uncertainty measures depend only on the predicting model and does not depend on the approximation of the true predictive distribution. Thus, they are the same for cases (1), (2) and (3). Furthermore, the AU of case (B) is an upper bound of the AU of case (C), i.e. $\mathrm{H}(p(y \mid \boldsymbol{x}, \mathcal{D})) \geq \mathrm{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[\mathrm{H}(p(y \mid \boldsymbol{x}, \boldsymbol{w}))\right]$, which follows from Eq. (3) as the mutual information is non-negative. Notably, Kotelevskii et al. [2025] showed that this inequality more generally holds for any proper scoring rule. Due to the linearity in the first argument of the cross-entropy, the TU for cases (B) and (C) are equal. Furthermore, as already discussed, the AU for cases (B) and (C) differ by the mutual information $\mathrm{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[\mathrm{KL}(p(y \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(y \mid \boldsymbol{x}, \mathcal{D}))\right]$. Therefore, the EU for cases (B) and (C) also differ by this factor. This is trivial to see for cases (B2) and (C2), where the EU of case (B2) cancels to zero and the EU of case (C2) is the mutual information. For cases (B3) and (C3), this was already mentioned by [Malinin and Gales, 2021] and a proof was given by [Schweighofer et al., 2023a], which we include for completeness in Apx. A.1, together with a version for cases (B1) and (C1).

4 RELATED WORK

Information-theoretic measures. The standard measure (Eq. (4)) introduced by Houlsby et al. [2011] and popularized, for instance, by Gal [2016], Depeweg et al. [2018], Smith and Gal [2018] is the measure (C2). In the context of autoregressive predictions, Malinin and Gales [2021] introduced measure (B3), due to the feasibility of a Monte Carlo (MC) approximation compared to the standard measure (C2). Schweighofer et al. [2023b] introduced measure (A3) together with a posterior sampling algorithm that is explicitly taylored to this measure. Schweighofer et al. [2023a] introduced measure (C3) as an improvement over the standard measure (C2) for certain settings and discussed (B3) in the appendix. Hofmann et al. [2024] and Kotelevskii et al. [2025] consider a broad set of proper scoring rules, including log, Brier, zero-one, and spherical score,

to derive uncertainty measures. Kotelevskii et al. [2025] base their framework on pointwise risk and subsequent Bayesian estimation, which, for the particular case of the log score, yields measures (B2), (B3), (C2) and (C3). Hofman et al. [2024b] discuss measure (C3) for the particular case of the log score. See Apx. A.5 for more details on how their framework relates to our framework. Information-theoretic measures have also been considered for uncertainty estimation for large language models, i.e. autoregressive prediction. Due to the large model sizes, recent work has focussed on AU for case (A) [Kuhn et al., 2023, Aichberger et al., 2025].

Alternative measures. There are also other measures of predictive uncertainty, not based on information-theoretic quantities. Kotelevskii et al. [2022] discusses the connection between pointwise risk and uncertainty and provides a nonparametric estimator thereof based on the Nadaraya-Watson kernel. Depeweg et al. [2018] introduced variance-based measures, based on the law of total variance. This perspective was recently developed further for specific settings [Duan et al., 2024, Sale et al., 2023b]. Furthermore, Sale et al. [2024b] introduced label-wise measures of predictive uncertainty, formulating both information-theoretic and variance-based measures. Another idea recently proposed by Sale et al. [2024a] is quantifying uncertainty through distances to reference (second-order) distributions (for TU, AU, and EU) denoting complete certainty. Thus, the higher the distance from the reference distribution, the more uncertain the prediction. All measures discussed so far operate on a distributional representation of uncertainty. Orthogonal to that, there are also set-based approaches [Hüllermeier et al., 2022, Sale et al., 2023a, Hofman et al., 2024a].

5 EXPERIMENTS

Next, we empirically evaluate the performance and characteristics of the uncertainty measures in our proposed framework. Specifically, we are interested in three different aspects. First, we investigate if there is merit in aligning the uncertainty measures with the predicting model. Second, we investigate the performance of the different measures for detecting distributional mismatch. Finally, we investigate the disentanglement between measures assessing their rank correlation and active learning performance.

Datasets. Our experiments are performed on the CIFAR10/100 [Krizhevsky and Hinton, 2009], SVHN [Netzer et al., 2011], Tiny-ImageNet (TIN) [Le and Yang, 2015] and LSUN [Yu et al., 2015] datasets. For TIN, we resize the inputs to 32x32 to match the other datasets. We train models on all datasets except LSUN, which is used solely as an OOD dataset.

Models and training. We used Deep Ensembles [Lakshminarayanan et al., 2017] to approximate posterior expectations through samples (10), as they are the de facto gold standard [Ovadia et al., 2019, Izmailov et al., 2021, Kotelevskii et al., 2025]. For example, the posterior predictive distribution given by Eq. (2) is approximated with N samples:

$$p(y \mid \boldsymbol{x}, \mathcal{D}) \approx \frac{1}{N} \sum_{n=1}^{N} p(y \mid \boldsymbol{x}, \boldsymbol{w}_n) \,, \tag{6}$$

where $\boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})$. We used three different model architectures for our experiments: ResNet-18 [He et al., 2016], DenseNet-169 [Huang et al., 2017] and RegNet-Y 800MF [Radosavovic et al., 2020]. Individual models were trained for 100 epochs using SGD with momentum of 0.9 with a batch size of 256 and an initial learning rate of 1e-2. Furthermore, a standard combination of linear (from factor 1 to 0.1) and cosine annealing schedulers was used. The main paper's results are for ResNet-18. Results for the two other architectures, found in Apx. B.3, are consistent to those.

Predictive uncertainty measures. We consider all measures proposed by our framework, see Tab. 1. For example, the (total) measure (A1) is referred to as TU (A1) with its aleatoric component as AU (A) and its epistemic component as EU (A1). Here, AU (A) is used instead of AU (A1) to emphasize the independence from the approximation of the true predictive distribution. Also, we group equivalent TU measures. For example TU (B1) and TU (C1) are merged to TU (B/C1).

5.1 ALIGNING THE UNCERTAINTY MEASURE WITH THE PREDICTING MODEL

The widely regarded selective prediction and misclassification tasks essentially evaluate the correlation of an uncertainty measure with the correctness of the prediction. However, it is not a priori clear which model is used to predict and if the uncertainty measure needs to be aligned to this choice, which we investigate in the following experiments.

Selective prediction. In this task, the model's predictions are limited to a specific subset, and its performance is evaluated on that subset. Therefore, we sampled models on the CIFAR10/100, SVHN and TIN datasets obtain uncertainty estimates and, importantly, predictions on their respective test datasets. We evaluated the accuracy for a subset of predictions of *(i) the single model*, *(ii) the average model* and *(iii) some model according to the posterior distribution* to investigate the impact of aligning the measure of uncertainty with the predicting model - (A) for setting (i), (B) for setting (ii) and (C) for setting (iii). The single model for (i) is the first of the sampled models. The average model for (ii) is defined by Eq. (6), averaging over all sampled models. For (iii), one model from the sampled models was randomly selected for each prediction. Note that (iii), also called the fully Bayesian setting, is rather unrealistic in practice. Selecting a new model according to the posterior for every prediction does not improve over a single fixed model (i) in expectation, if the performance of individual sampled models

Table 2: **Selective prediction under different predicting models.** AUARC for different predicting models, i.e., the single model, the average model and a model according to the posterior, using different predictive uncertainty measures as score. We highlight the two best measures for each setting; TU (B/C2) and AU (B) are equivalent, thus both get highlighted. For the single predicting model, TU (A3) and EU (A3) perform best, in the other two settings, TU (B/C3) and TU (B/C2) (AU (B)). Results are averaged over all datasets with statistics over five runs.

Measures:	TU						AU			EU								Random
Prediction	A1	A2	A3	B/C1	**B/C2**	B/C3	A	**B**	C	A1	A2	A3	B1	B3	C1	C2	C3	Baseline
Single	87.96	88.10	88.24	88.07	88.12	88.22	87.50	88.12	87.95	87.30	88.11	88.23	86.58	87.60	87.31	88.12	87.85	79.67
Model	±0.10	±0.10	±0.11	±0.09	±0.09	±0.09	±0.11	±0.09	±0.08	±0.08	±0.07	±0.11	±0.10	±0.11	±0.10	±0.11	±0.11	±0.10
Average	89.28	89.52	89.46	90.17	90.23	90.30	89.20	90.23	90.11	88.63	89.15	89.17	88.73	89.52	89.36	90.07	89.78	83.02
Model	±0.06	±0.08	±0.08	±0.05	±0.05	±0.05	±0.08	±0.05	±0.06	±0.04	±0.05	±0.09	±0.03	±0.05	±0.04	±0.06	±0.06	±0.08
Acc. to	87.74	88.00	87.96	88.74	88.79	88.90	87.61	88.79	88.63	87.06	87.67	87.73	87.20	88.22	87.94	88.75	88.48	80.53
Posterior	±0.08	±0.06	±0.07	±0.09	±0.08	±0.07	±0.06	±0.08	±0.08	±0.07	±0.07	±0.06	±0.12	±0.08	±0.11	±0.06	±0.07	±0.06

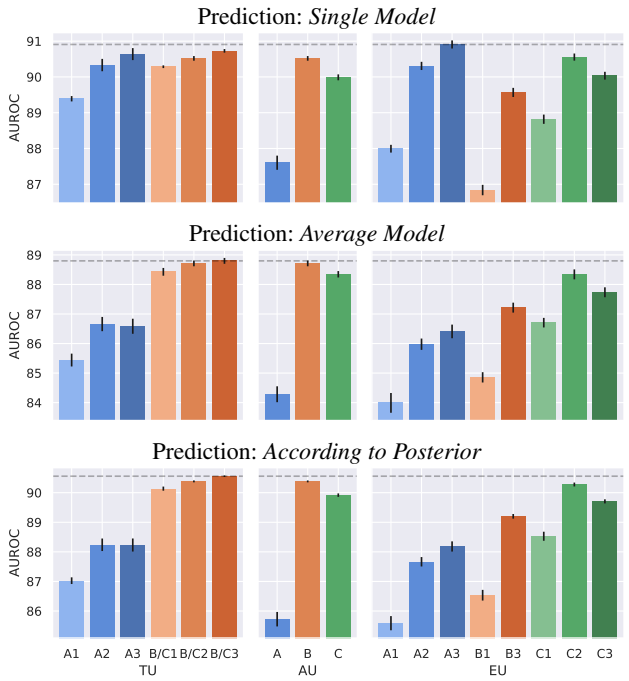


Figure 1: **Misclassification detection under different predicting models.** AUROC for distinguishing correct from incorrect predictions under different predicting models, using the different proposed measures of uncertainty as score. EU (A3) performs best when predicting with the single model. For the other cases, TU (B/C3) performs best. AUROCs are averages over all datasets. Statistics over five runs.

is comparable. Generally, it is also more convenient to stick to a single model.

We evaluated subsets ranging from the most certain 50% of datapoints to the entire dataset. To compare how well uncertainty measures rank data for selecting those subsets, we used the area under the accuracy rejection curve (AUARC) as the performance metric. Results are provided in Tab. 2, showing that the optimal measure depends on the model used for prediction. For (i), TU (A3) performs best, followed by EU (A3), while for (ii) and (iii), TU (B/C3) performs best. Notably, uncertainty measures perform essentially the same for settings (ii) and (iii). As we will show in Sec. 5.3, uncertainty measures for (B) and (C) are highly correlated on ID data, which explains their similar performance in settings (ii) and (iii). Additional details and plots of accuracy rejection curves are provided in Apx. B.2.

Misclassification detection. This task evaluates the ability of an uncertainty measure to distinguish between the set of correct and incorrectly predicted samples. The setup in this experiment is identical to the selective prediction setup. We consider the AUROC for distinguishing between correctly and incorrectly predicted datapoints for the different proposed measures of predictive uncertainty as scoring functions. Alternative metrics commonly used to evaluate misclassification detection, such as AUPR or FPR@TPR95, were also considered in our experiments. Those induced the same ordering of uncertainty measures, thus we report the AUROC for experiments of this type.

The results are provided in Fig. 1. We average over the four considered datasets and report means and standard deviations over five independent runs. The results for individual datasets are reported in Fig. 10 - Fig. 12 in the appendix. For detecting the misclassification of (i) the single model, EU (A3) performs best. However, when predicting with (ii) the average or (iii) a model according to the posterior, TU (B/C3) performs best. This mirrors the results of the selective prediction task, confirming the importance of aligning the uncertainty measure to the predicting model.

Adversarial example detection. Additionally, we explored the efficacy of uncertainty measures for detecting adversarial examples on a specific model under FGSM [Goodfellow et al., 2015] and PGD [Madry et al., 2018] attacks. Importantly, we do not intend to claim any level of adversarial robustness to these attacks, but use them as a tool to understand the behavior of the uncertainty measures resulting from our framework. Results and detailed discussion are provided in Apx. B.4 due to space limitations.

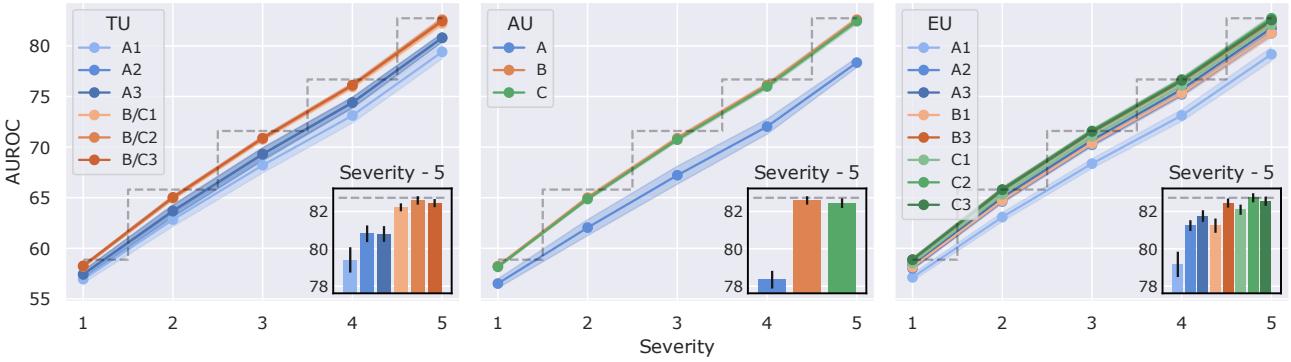


Figure 2: **Distribution shift detection on CIFAR10-C.** AUROC for distinguishing between clean and corrupted test datapoints, using the different proposed measures of uncertainty as score. Black dashed line shows the maximum AUROC over all measures per severity. Insets shows detailed results for the highest severity. Statistics over five runs.

5.2 DETECTING DISTRIBUTIONAL MISMATCH

The first set of experiments aimed to assess the correlation between an uncertainty measure and prediction correctness. This is typically evaluated on i.i.d. test data drawn from the same distribution as the training dataset $\mathcal{D}$. However, there is no guarantee that the model will perform reliably on test data from a different distribution. Therefore, an effective uncertainty measure should ideally assign high uncertainty to such out-of-distribution (OOD) samples.

Out-of-distribution detection. We considered standard OOD detection dataset pairings from the literature [Mukhoti et al., 2023, Hofmann et al., 2024] for our experiments. Again, we sampled models on CIFAR10/100, SVHN and TIN. Then, we use the respective test datasets as ID dataset and the test datasets of the remaining datasets, as well as LSUN, as OOD datasets. Thus, we consider ID/OOD dataset pairings CIFAR10/CIFAR100, CIFAR10/SVHN, and so on; a total of 16 pairings. We compare the AUROC for distinguishing between ID and OOD datapoints for each measure within our framework as a scoring function. Alternative commonly used metrics such as the AUPR and the FPR@TPR95 were also considered. However, since they induced the same ranking of uncertainty measures, we report the AUROC for the OOD detection experiments.

Results are provided in Fig. 3, showing that TU `(B/C2)` and TU `(B/C3)` perform best in this task. We observe, that across all measures, TU and AU perform better than EU, which is contrary to assumptions commonly formulated in the literature [e.g., Mukhoti et al., 2023, Mucsányi et al., 2024]. We find, that only on a single ID / OOD dataset combination (TIN/SVHN, see Fig. 16 in the appendix) out of the 16 pairings we considered, an EU measure performs best. One possible explanation for the strong performance of AU `(B)` and `(C)` could be low noise for these datasets.

Distribution shift detection. The considered dataset pairings represent comparatively strong OOD-ness. Therefore, we additionally investigated the performance of the uncertainty measures for detecting distribution shifts on CIFAR10 using the CIFAR10-C [Hendrycks and Dietterich, 2019] dataset. The different levels of corruption provided by this dataset can be considered as increasing levels of OOD-ness. We utilized the 15 main corruptions and excluded the four additional corruptions designated for hyperparameter tuning. We report averages over all 15 corruptions in Fig. 2 for the five severity levels of corruption provided, showing the AUROC of distinguishing between clean and corrupted versions of the CIFAR10 test dataset using the different uncertainty measures. The results show that EU measures are more effective than AU or TU measures at intermediate severities but become equally effective for the highest severity.

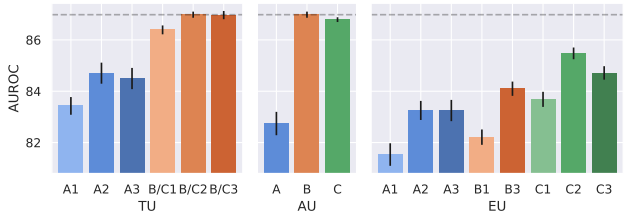


Figure 3: **OOD detection.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. TU `(B/C2)` and TU `(B/C3)` perform best. AUROCs are averaged over all ID / OOD combinations. Statistics over five runs.

5.3 DISENTANGLEMENT OF MEASURES

There has been a recent surge of interest in the disentanglement of AU and EU [Valdenegro-Toro and Mori, 2022, Mukhoti et al., 2023, Mucsányi et al., 2024]. Furthermore, concerns about the additive decomposition TU = AU + EU has been raised [Wimmer et al., 2023]. Therefore, we investigate the correlation of our measures on both ID and OOD data and conclude with active learning experiments, one of the main tasks requiring disentangled EU measures.

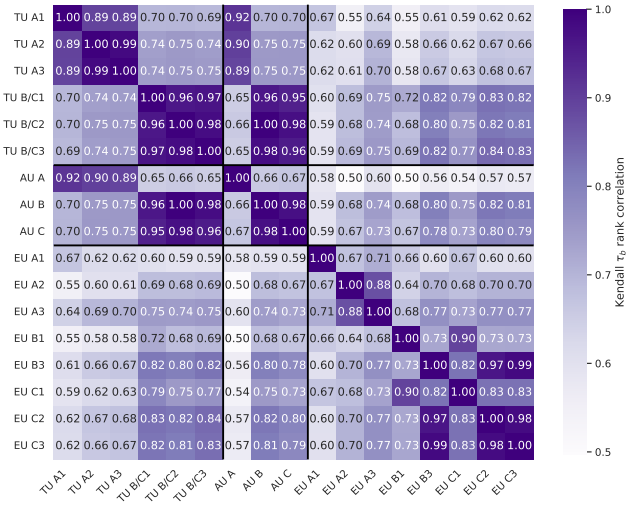


Figure 4: **Correlation of uncertainty measures on ID dataset (CIFAR10).** High rank correlation blocks exist.

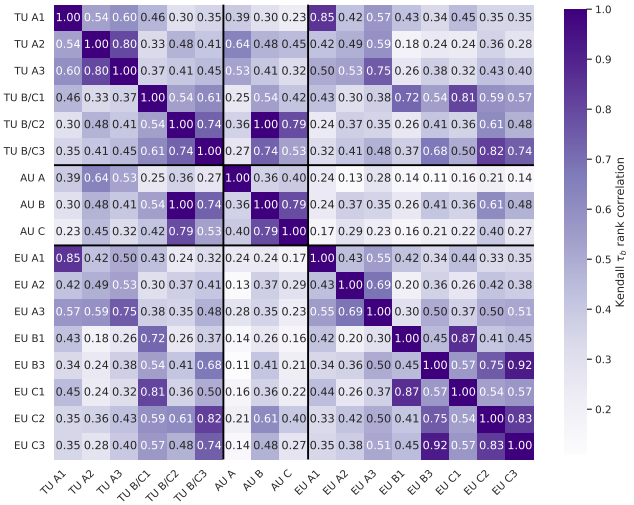


Figure 5: **Correlation of uncertainty measures on OOD dataset (LSUN).** Rank correlations are very low overall.

Rank correlation of measures. Following Mucsányi et al. [2024], we investigate the rank correlation of the uncertainty measures in our proposed framework. We consider the rank correlation (Kendall's τ_b to account for ties) of measures on both ID and OOD data. Models are sampled on CIFAR10, thus the ID dataset is the CIFAR10 test dataset, and we consider LSUN as the OOD dataset.

The results for the ID dataset are shown in Fig. 4. We observe, that for TU and AU measures, there are two highly correlated blocks. Those measures that depend on a single predicting model and the rest. While there is a stark contrast between considering the single predicting model and the rest, there are no real differences between considering the average model as predicting model and each model according to the posterior distribution. Furthermore, the approximation of the true predictive distribution `(1,2,3)` does not seem to play a major role for TU and AU on the ID dataset. For EU, there is a block of highly correlated measures, `(B3)`, `(C2)` and `(C3)`. This can be explained due to their additive relationship (see Apx. A.1). The same holds for the strong correlation between EU `(B1)` and `(C1)`, which share a similar relationship (again, see Apx. A.1). Finally, EU `(A2)` and `(A3)` have a high correlation. Although they do not share a similar relationship as the other blocks of high correlation for EU measures, EU `(A3)` is an upper bound of `(A2)`.

The results for the OOD dataset are shown in Fig. 5. Here, we do not observe similarly high rank correlations as for the ID dataset. The strongest rank correlation exhibited is those for EU `(B3)`, `(C2)` and `(C3)`, as well as EU `(B1)` and `(C1)`. Note that TU `(B/C3)` and AU `(B)` are equivalent, thus perfectly correlated. In general, we observe a very distinct disentanglement of uncertainty measures in our proposed framework. Although there are blocks of high correlation for ID data, we mostly do not observe such strong correlations for OOD data.

Active Learning. Finally, we investigated the proposed framework of uncertainty measures on active learning tasks. Given the computational complexity of the active learning setting, we utilized different datasets and models as for the rest of our experiments. Specifically, we used the MNIST [Lecun et al., 1998] and FMNIST [Xiao et al., 2017] datasets and a small CNN. Details on the network architecture and training procedure are provided in Apx. B.5. We obtained five posterior samples (ensemble members) in each acquisition round. The average over the predictive distributions of those sampled models, the approximated posterior predictive, was used to calculate the accuracies for each acquisition step. The same sampled models were used to approximate the respective uncertainty measures as acquisition functions to select the next datapoints from the pool dataset to transfer to the training dataset.

For MNIST, we started with 20 datapoints in the training dataset and the remaining 49,980 datapoints in the pool dataset. Those 20 datapoints were balanced so that two datapoints from each class were contained. In each iteration, the five samples with the highest uncertainty were selected from the pool dataset and added to the training dataset. We considered TU, AU and EU for measures `(B2)`, `(B3)`, `(C2)` and `(C3)` as acquisition functions, as well as random selection as a baseline. We did not investigate measures `(A1)`, `(A2)`, `(A3)`, `(B1)` and `(C1)` due to the long runtimes of the experiments, but would expect them to perform worse than those considered in light of the other experiments we conducted. However, an interesting situation could be EU `(A1)` when training a single model on the dataset in the current iteration and comparing the model from the previous iteration. Future work should investigate this setting, e.g. in transfer learning scenarios.

The results are given in Fig. 6. We observe that EU `(C2)`, the mutual information, leads to the best performance for

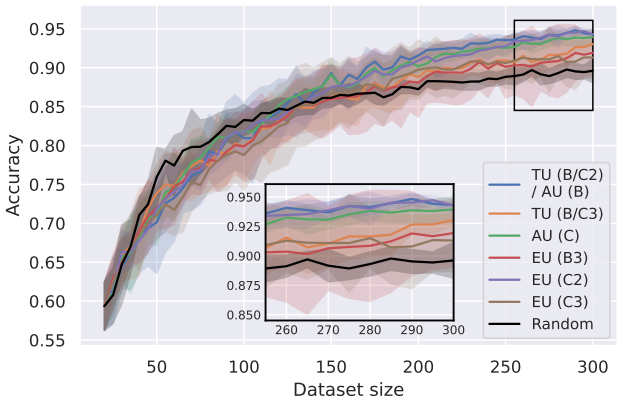


Figure 6: **Active learning on MNIST.** EU (C3) and EU (B3) perform worst overall. The random baseline performs strong for small dataset sizes. The accuracy is those of the posterior predictive. Statistics over five runs.

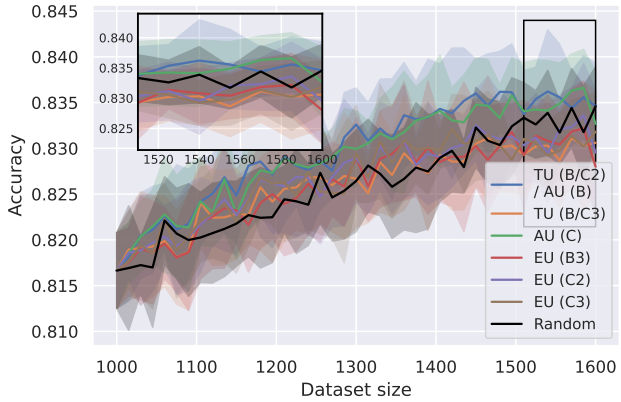


Figure 7: **Active learning on FMNIST.** All three EU measures perform worse than TU (B/C2) / AU (B) and AU (C). The accuracy is those of the posterior predictive. Statistics over ten runs.

any of the EU measures. Interestingly, we find TU (B/C2) which is identical to AU (B) to be equally well performing. The same is found for TU (B/C3). EU (B3) and EU (C3) are found to be the worst performing measures as acquisition functions, contrary to the sentiment that EU estimators should perform best in this task [e.g. Gal et al., 2017, Mukhoti et al., 2023]. In contrast, AU (C), performs very well as an acquisition function. The random sampling baseline is also very effective as an acquisition function until the training dataset size reaches around 100 samples, more effective than any of the considered uncertainty measures. We hypothesize that until a certain dataset size, models sampled from the posterior are not specified enough and provide no good signal of what datapoints to add next, in fact even the contrary as they are performing worse than random. This behavior would be interesting to investigate in more details in future experiments.

For FMNIST, we started with 1000 datapoints in the training dataset and the remaining 49,000 datapoints in the pool dataset. The initial training dataset was balanced such that 100 datapoints from each class were included. In each iteration, the 15 samples with the highest uncertainty are selected from the pool dataset and added to the training dataset. As for the MNIST experiment, we considered TU, AU and EU for measures (B2), (B3), (C2) and (C3) as acquisition functions, as well as random selection as baseline.

The results are provided in Fig. 7. We find that all EU measures including the mutual information EU (C2), are outperformed by TU (B/C2) which is identical to AU (B), and AU (C). Until around 1500 samples, the random baseline performs roughly on par with the EU measures, afterwards it even performs slightly better. Although both datasets are relatively simple, FMNIST proved notably more challenging for the active learning pipeline and has higher variance between runs. This is likely due to its increased complexity and less clear-cut class boundaries.

6 CONCLUSION

In this work, we introduced a unifying framework to categorize information-theoretic measures of predictive uncertainty. By systematically considering the choice of the predicting model and the approximation of the true predictive distribution, we derived a family of uncertainty measures that encompasses existing measures and introduces new ones. This framework clarifies the relationships between different uncertainty measures and their implicit assumptions, offering a more structured way of understanding them and applying them in practice.

Our empirical evaluations demonstrated the importance of aligning the uncertainty measure with the predicting model. We found that different measures perform optimally, depending on the predicting model, emphasizing the need for careful selection based on the intended application. Furthermore, our results challenge prevailing assumptions in the literature, such as the effectiveness of EU measures for OOD detection, showing that AU often plays a dominant role. We also investigated the degree of disentanglement between the measures of our proposed framework. Our findings indicate that, while there are strong correlations between specific uncertainty measures for ID data, correlations weaken significantly in OOD scenarios. This suggests that the disentanglement of measures may be highly context-dependent, which should be investigated in depth in the future. Future work should also explore extensions of our framework to autoregressive models. Here, uncertainty estimation plays a crucial role in addressing hallucinations, particularly a subset known as confabulations, and in ensuring reliable text generation [Xiao and Wang, 2021, Abbasi-Yadkori et al., 2024, Aichberger et al., 2025, Farquhar et al., 2024].

Overall, our work provides a clear conceptual foundation for understanding predictive uncertainty measures and offers actionable insights for their selection for a given task.

Acknowledgements

We thank Andreas Kirsch for constructive feedback on an earlier version of this work.

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects FWF AIRI FG 9-N (10.55776/FG9), AI4GreenHeatingGrids (FFG- 899943), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01), FWF Bilateral Artificial Intelligence (10.55776/COE12). We thank NXAI GmbH, Audi AG, Silicon Austria Labs (SAL), Merck Healthcare KGaA, GLS (Univ. Waterloo), TÜV Holding GmbH, Software Competence Center Hagenberg GmbH, dSPACE GmbH, TRUMPF SE + Co. KG.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. To Believe or Not to Believe Your LLM: Iterative Prompting for Estimating Epistemic Uncertainty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. Improving Uncertainty Estimation through Semantically Diverse Language Generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

George Apostolakis. The Concept of Probability in Safety Assessments of Technological Systems. *Science*, 1990.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux - Effortless Bayesian Deep Learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.

Ruxiao Duan, Brian Caffo, Harrison X. Bai, Haris I. Sair, and Craig Jones. Evidential Uncertainty Quantification: A Variance-Based Perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*, 2024.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective. *ArXiv*, 1912.02757, 2019.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015.

Sebastian Gruber and Florian Buettner. Uncertainty Estimates of Predictions via a General Bias-Variance Decomposition. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. PMLR, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Jon C. Helton. Risk, Uncertainty in Risk, and the EPA Release Limits for Radioactive Waste Disposal. *Nuclear Technology*, 1993.

Jon C. Helton. Uncertainty and Sensitivity Analysis in the Presence of Stochastic and Subjective Uncertainty. *Journal of Statistical Computation and Simulation*, 1997.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Lara Hoffmann and Clemens Elster. Deep Ensembles from a Bayesian Perspective. *ArXiv*, 2105.13283, 2021.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying Aleatoric and Epistemic Uncertainty: A Credal Approach. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024a.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying Aleatoric and Epistemic Uncertainty with Proper Scoring Rules. *ArXiv*, 2404.12215, 2024b.

Claus Hofmann, Simon Lucas Schmid, Bernhard Lehner, Daniel Klotz, and Sepp Hochreiter. Energy-based Hopfield Boosting for Out-of-Distribution Detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *ArXiv*, 1112.5745, 2011.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 2021.

Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. PMLR, 2022.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. Nonparametric Uncertainty Quantification for Single Deterministic Neural Network. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. From Risk to Uncertainty: Generating Predictive Uncertainty Measures via Bayesian Estimation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. *University of Toronto*, 2009.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct Epistemic Uncertainty Prediction. *Transactions on Machine Learning Research*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

Ya Le and Xuan Yang. Tiny ImageNet Visual Recognition Challenge. *CS 231N*, 2015.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based Learning applied to Document Recognition. *Proceedings of the IEEE*, 1998.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast Inference from Transformers via Speculative Decoding. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023.

David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 1992.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, 2018.

Andrey Malinin and Mark Gales. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*, 2021.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. Deep Deterministic Uncertainty: A New Simple Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading Digits in Natural Images with Unsupervised Feature Learning. *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you Trust your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollar. Designing Network Design Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A Scalable Laplace Approximation for Neural Networks. In *International Conference on Learning Representations*, 2018.

Yusuf Sale, Michele Caprio, and Eyke Höllermeier. Is the Volume of a Credal Set a good Measure for Epistemic Uncertainty? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. PMLR, 2023a.

Yusuf Sale, Paul Hofman, Lisa Wimmer, Eyke Hüllermeier, and Thomas Nagler. Second-Order Uncertainty Quantification: Variance-Based Measures. *ArXiv*, 2401.00276, 2023b.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-Order Uncertainty Quantification: A Distance-Based Approach. In *Forty-first International Conference on Machine Learning*, 2024a.

Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise Aleatoric and Epistemic Uncertainty Quantification. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, 2024b.

Aydin Sarraf and Yimin Nie. RGAN: Rényi Generative Adversarial Network. *SN Computer Science*, 2021.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. Introducing an Improved Information-Theoretic Measure of Predictive Uncertainty. *ArXiv*, 2311.08309, 2023a.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. Quantification of Uncertainty with Adversarial Models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023b.

Claude Elwood Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 1948.

Lewis Smith and Yarin Gal. Understanding Measures of Uncertainty for Adversarial Example Detection. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2018.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise Parallel Decoding for Deep Autoregressive Models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

Ferenc Cole Thierrin, Fady Alajaji, and Tamás Linder. Rényi Cross-Entropy Measures for Common Distributions and Processes with Memory. *Entropy*, 2022.

Matias Valdenegro-Toro and Daniel Saromo Mori. A Deeper Look Into Aleatoric and Epistemic Uncertainty Disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. The Case for Shifting the Rényi Entropy. *Entropy*, 2019.

Tim van Erven and Peter Harremos. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 2014.

Andrew G Wilson and Pavel Izmailov. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures? In *Uncertainty in Artificial Intelligence*. PMLR, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *ArXiv*, 1708.07747, 2017.

Yijun Xiao and William Yang Wang. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *ArXiv*, 1506.03365, 2015.

On Information-Theoretic Measures of Predictive Uncertainty (Supplementary Material)

Kajetan Schweighofer[1] **Lukas Aichberger**[1] **Mykyta Ielanskyi**[1] **Sepp Hochreiter**[1,2]

[1]ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University Linz, Austria
[2]NXAI GmbH, Linz, Austria

A TECHNICAL DETAILS

In this section, we provide additional details on the relationship between epistemic components as well as the MC approximations for all measures. Furthermore, we discuss potential generalizations of our framework to the Rényi cross-entropy and other proper scoring rules. Finally, we discuss the alternative order of arguments for Eq. (5) and its implications for interpreting the resulting uncertainty measures as well as the regression setting.

A.1 RELATIONSHIPS BETWEEN EPISTEMIC COMPONENTS

Schweighofer et al. [2023a] proved the relationship that the sum of the EU of C2 and B3 is equivalent to the EU of C3. For completeness, we provide a version of the proof as follows:

$$
\begin{aligned}
&\overbrace{\mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{KL}(p(y \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(y \mid \boldsymbol{x}, \mathcal{D}))\right]}^{\text{EU (C2) - Mutual Information}} + \overbrace{\mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{KL}(p(y \mid \boldsymbol{x}, \mathcal{D}) \parallel p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]}^{\text{EU (B3)}} && (7)\\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log \frac{p(y \mid \boldsymbol{x}, \boldsymbol{w})}{p(y \mid \boldsymbol{x}, \mathcal{D})}\right]\right] + \mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log \frac{p(y \mid \boldsymbol{x}, \mathcal{D})}{p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})}\right]\right] && (8)\\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \boldsymbol{w})\right] - \mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right]\right] + && (9)\\
&\quad \mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right] - \mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right] \\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \boldsymbol{w})\right]\right] - \cancel{\mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right]} + && (10)\\
&\quad \cancel{\mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right]} - \mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right] \\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \boldsymbol{w})\right]\right] - && (11)\\
&\quad \mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right]\right] \\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \boldsymbol{w})\right]\right]\right] - && (12)\\
&\quad \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right]\right] \\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log \frac{p(y \mid \boldsymbol{x}, \boldsymbol{w})}{p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})}\right]\right]\right] && (13)\\
&= \underbrace{\mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{KL}(p(y \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right]}_{\text{EU (C3)}}, && (14)
\end{aligned}
$$

which is what we wanted to show. The step from (9) to (10) is due to additivity and linearity of expectations. The step from (11) to (12) is due to the fact that we can insert the expectation $\mathrm{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}$ in the first term as it does not depend on $\tilde{\boldsymbol{w}}$ and due to the fact that $p(\tilde{\boldsymbol{w}} \mid \mathcal{D}) = p(\boldsymbol{w} \mid \mathcal{D})$.

□

Furthermore, a similar proof can be constructed for EU `(C1)` = EU `(C2)` + EU `(B1)` as follows:

$$
\begin{aligned}
&\overbrace{\mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{KL}(p(y \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(y \mid \boldsymbol{x}, \mathcal{D}))\right]}^{\text{EU (C2) - Mutual Information}} + \overbrace{\mathrm{KL}(p(y \mid \boldsymbol{x}, \mathcal{D}) \parallel p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))}^{\text{EU (B1)}} && (15)\\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log \frac{p(y \mid \boldsymbol{x}, \boldsymbol{w})}{p(y \mid \boldsymbol{x}, \mathcal{D})}\right]\right] + \mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log \frac{p(y \mid \boldsymbol{x}, \mathcal{D})}{p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})}\right] && (16)\\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \boldsymbol{w})\right] - \mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right]\right] + && (17)\\
&\quad \mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right] - \mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right] \\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \boldsymbol{w})\right]\right] - \cancel{\mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right]} + && (18)\\
&\quad \cancel{\mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \mathcal{D})\right]} - \mathrm{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[\log p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right] \\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \boldsymbol{w})\right]\right] - \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right] && (19)\\
&= \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}\left[\log \frac{p(y \mid \boldsymbol{x}, \boldsymbol{w})}{p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})}\right]\right] && (20)\\
&= \underbrace{\mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{KL}(p(y \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(y \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]}_{\text{EU (C1)}}, && (21)
\end{aligned}
$$

which is what we wanted to show. Again, the step from (17) to (18) is due to additivity and linearity of expectations. The linearity property is used to get to (19), after which elementary algebra leads to the result.

□

In the same fashion, it is possible to construct a proof for EU `(C2)` = EU `(C2)` + EU `(B2)`. However, as we know that EU `(B2)` = 0, this is trivial.

A.2 MONTE CARLO APPROXIMATIONS

The measures we proposed through our framework, except for measure `(A1)`, incorporate posterior expectations $\mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}[\cdot]$. These are generally intractable to calculate exactly and are thus approximated through samples drawn from the distribution - a Monte Carlo approximation of the expectation. In this section we provide those approximations explicitly and discuss efficient ways to implement them, utilizing relationships between individual measures.

We assume that the posterior $p(\boldsymbol{w} \mid \mathcal{D})$ models to predict are drawn from and the posterior $p(\tilde{\boldsymbol{w}} \mid \mathcal{D})$ approximations of the true model are drawn from are the same. However, in practice it is generally the case that models for averaging are selected based on their accuracy on a validation set, or more generally that they are selected in a way optimal for predicting well. When sampling potential true models that are likely under the data, the functional diversity of samples is often of concern, e.g. as done with the sampling algorithm in Schweighofer et al. [2023b]. This can be seen as either having different posteriors due to different priors or having different algorithms to obtain samples from the same posterior. However, if the samples are not the same, many of the relationships in Sec. 3.1 are no longer valid, thus also the coloring in Tab. 1. The provided implementation is able to handle both the case where different samples are used for the MC approximation as well as the case where the same samples are used and thus more efficient computation of TU `(C3)` and EU `(C3)` is possible. The MC approximations for a single set of samples $\{\boldsymbol{w}_n\}_{n=1}^N$ from $p(\boldsymbol{w} \mid \mathcal{D}) = p(\tilde{\boldsymbol{w}} \mid \mathcal{D})$ are given by

TU `(A2)`:

$$
\begin{aligned}
\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})) &= \mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ \mathrm{E}_{\tilde{\boldsymbol{w}}}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]) \\
&\approx \mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ \frac{1}{N}\sum_{n=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})
\end{aligned} \tag{22}
$$

TU `(A3)`:

$$
\begin{aligned}
&\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \\
&\quad \approx \frac{1}{N}\sum_{n=1}^{N} \mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})
\end{aligned} \tag{23}
$$

TU (B/C1):

$$\begin{aligned}\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})) \ &= \ \mathrm{E}_{\boldsymbol{w}}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \\ &\approx \ \frac{1}{N}\sum_{n=1}^{N} \mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{24}$$

TU (B/C2):

$$\begin{aligned}\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})) \ &= \ \mathrm{E}_{\boldsymbol{w}}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ \mathrm{E}_{\tilde{\boldsymbol{w}}}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right])\right] \\ &\approx \ \frac{1}{N}\sum_{n=1}^{N} \mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)\ ;\ \frac{1}{N}\sum_{n=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{25}$$

TU (B/C3):

$$\begin{aligned}\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{CE}((p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \ &= \ \mathrm{E}_{\boldsymbol{w}}\left[\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right] \\ &\approx \ \frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{n'=1}^{N} \mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)\ ;\ p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_{n'})), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{26}$$

AU (B):

$$\begin{aligned}\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})) \ &= \ \mathrm{H}(\mathrm{E}_{\boldsymbol{w}}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right]) \\ &\approx \ \mathrm{H}(\frac{1}{N}\sum_{n=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{27}$$

AU (C):

$$\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))\right] \ \approx \ \frac{1}{N}\sum_{n=1}^{N} \mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D}) \tag{28}$$

EU (A2):

$$\begin{aligned}\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})) \ &= \ \mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel \mathrm{E}_{\tilde{\boldsymbol{w}}}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]) \\ &\approx \ \mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel \frac{1}{N}\sum_{n=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{29}$$

EU (A3):

$$\begin{aligned}&\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \\ &\approx \ \frac{1}{N}\sum_{n=1}^{N} \mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{30}$$

EU (B1):

$$\begin{aligned}\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})) \ &= \ \mathrm{KL}(\mathrm{E}_{\boldsymbol{w}}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right] \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})) \\ &\approx \ \mathrm{KL}(\frac{1}{N}\sum_{n=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{31}$$

EU (B3):

$$\begin{aligned}\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \ &= \ \mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{KL}(\mathrm{E}_{\boldsymbol{w}}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right] \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \\ &\approx \ \frac{1}{N}\sum_{n=1}^{N} \mathrm{KL}(\frac{1}{N}\sum_{n=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D})\end{aligned} \tag{32}$$

EU (C1):

$$\begin{aligned} &\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \\ &\quad \approx \frac{1}{N} \sum_{n=1}^{N} \mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D}) \end{aligned} \tag{33}$$

EU (C2):

$$\begin{aligned} \mathrm{E}_{\boldsymbol{w}}\left[\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right] \;&=\; \mathrm{E}_{\boldsymbol{w}}\left[\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel \mathrm{E}_{\tilde{\boldsymbol{w}}}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right])\right] \\ \approx \frac{1}{N} \sum_{n=1}^{N} \mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n) \parallel \frac{1}{N} \sum_{n=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n)), &\qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D}) \end{aligned} \tag{34}$$

EU (C3):

$$\begin{aligned} &\mathrm{E}_{\boldsymbol{w}}\left[\mathrm{E}_{\tilde{\boldsymbol{w}}}\left[\mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right] \\ &\quad \approx \frac{1}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=1}^{N} \mathrm{KL}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_n) \parallel p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_{n'})), \qquad \boldsymbol{w}_n \sim p(\boldsymbol{w} \mid \mathcal{D}) \end{aligned} \tag{35}$$

A.3 GENERALIZATION TO RENYI CROSS-ENTROPY

In this section we review the Rényi cross-entropy which is a generalization of the cross-entropy discussed in the main paper. This allows to directly transfer our proposed measure of predictive uncertainty in Eq. (5) and the framework we introduced based on it (overview in Tab. 1) to other instances of Rényi cross-entropies.

Let us start with the Rényi entropy, which was proposed as a generalization of the Shannon entropy, in that for the limit of the Rényi parameter $\alpha \to 1$ the Rényi entropy becomes the Shannon entropy. For two discrete distributions p and q on the same support $\mathcal{Y}$ it is defined as

$$\mathrm{H}_\alpha(p) \;=\; \frac{1}{1-\alpha} \log \sum_i p_i^\alpha \tag{36}$$

Similarly, the Rényi divergence is a generalization of the Kullback-Leibler (KL) divergence, in that for the limit of the Rényi parameter $\alpha \to 1$ the Rényi divergence becomes the KL divergence. It is defined as

$$\mathrm{D}_\alpha(p \,||\, q) \;=\; \frac{1}{\alpha-1} \log \sum_i p_i^\alpha \; q_i^{1-\alpha} \tag{37}$$

Note that there are also versions of both for continuous distributions, basically exchanging the sum with an integral. However, the resulting Rényi differential entropy shares the same deficiencies as the Shannon differential entropy.

What is left is defining the Rényi cross-entropy. Motivated by the additive decomposition of Shannon cross-entropy into the entropy and KL divergence, Sarraf and Nie [2021] proposed to define the Rényi cross-entropy as

$$\mathrm{CE}_\alpha(p\,;\,q) \;:=\; \mathrm{H}_\alpha(p) \;+\; \mathrm{D}_\alpha(p \,||\, q) \tag{38}$$

Multiple closed form solutions for different values of α are already known for the Rényi entropy and divergence, making this a very simple solution. Furthermore, Valverde-Albacete and Peláez-Moreno [2019] introduced a closed form solution, which has been simplified to the following form by Thierrin et al. [2022]:

$$\mathrm{CE}_\alpha(p\,;\,q) \;:=\; \frac{1}{1-\alpha} \log \sum_i p_i \; q_i^{\alpha-1} \tag{39}$$

Furthermore, Thierrin et al. [2022] proposes closed form solutions for this form of the Rényi differential cross-entropy for various continuous distributions.

In the following we stick to the definition of the Rényi cross-entropy by Sarraf and Nie [2021] (Eq. (38)) and state the respective entropy and divergence for special cases of α. By defining the arbitrary discrete distributions as $p := p(y \mid \boldsymbol{x}, \cdot)$ and $q := p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)$ each value of α yields a variant our proposed measure of predictive uncertainty (Eq. (5)), giving rise to variants of our proposed framework.

$\alpha = 0$: The measure of entropy is called the Hartley or max-entropy, which is the cardinality of possible events $\mathcal{Y}$. It is given by

$$\mathrm{H}_0(p) := \log |\mathcal{Y}| \,. \tag{40}$$

The divergence is called max-divergence and is given by

$$\mathrm{D}_0(p \,||\, q) := -\log Q(\{i : p_i > 0\}) \,. \tag{41}$$

$\alpha = \frac{1}{2}$: The measure of entropy is referred to as Bhattacharyya-entropy. It is given by

$$\mathrm{H}_{\frac{1}{2}}(p) := 2 \log \sum_i \sqrt{p_i} \,. \tag{42}$$

The divergence is called Bhattacharyya-divergence (minus twice the logarithm of the Bhattacharyya coefficient) and is given by

$$\mathrm{D}_{\frac{1}{2}}(p \,||\, q) := -2 \log \sum_i \sqrt{p_i q_i} \,. \tag{43}$$

$\alpha = 1$: This case is the well known Shannon-entropy, given by

$$\mathrm{H}_1(p) = \mathrm{H}(p) := -\sum_i p_i \log p_i \,. \tag{44}$$

The divergence is known as Kullback-Leibler divergence, given by

$$\mathrm{D}_1(p \,||\, q) = \mathrm{KL}(p \,\|\, q) := \sum_i p_i \log \frac{p_i}{q_i} \,. \tag{45}$$

$\alpha = 2$: This case is called the collision entropy, which is closely related to the index of coincidence. It is given by

$$\mathrm{H}_2(p) := -\log \sum_i p_i^2 \,. \tag{46}$$

The corresponding divergence is based upon the chi-square divergence

$$\mathrm{D}_2(p \,||\, q) := \log \left(\sum_{i=1}^{N} \frac{p_i^2}{q_i} \right) = \log \left(1 + \sum_{i=1}^{N} \frac{(p_i - q_i)^2}{q_i} \right) \,. \tag{47}$$

$\alpha = \infty$: The entropy is known as the min-entropy. It is given by

$$\mathrm{H}_\infty(p) := -\log \max_i p_i \,. \tag{48}$$

The divergence

$$\mathrm{D}_\infty(p \,||\, q) := \log \sup_i \frac{p_i}{q_i} \,. \tag{49}$$

Notes. Realizations of Renyi entropy satisfy the inequalities

$$\mathrm{H}_0(p) \geq \mathrm{H}_1(p) \geq \mathrm{H}_2(p) \geq \mathrm{H}_\infty(p) \tag{50}$$

Also Theorem 7 in van Erven and Harremos [2014] states that Renyi divergences are continuous in the order of α.

Table 3: **Our proposed framework applied under the zero-one score.** Each measure denotes a different instantiation of our proposed measure given by Eq. (5), but using the zero-one score instead of the cross-entropy (log score) for different assumptions about the predicting model and how the true model is approximated. For brevity, we define $p_{\boldsymbol{w}} \coloneqq p(y \mid \boldsymbol{x}, \boldsymbol{w})$, $p_{\mathcal{D}} \coloneqq p(y \mid \boldsymbol{x}, \mathcal{D})$, $\mathrm{E}_{\boldsymbol{w}} \coloneqq \mathrm{E}_{p(\boldsymbol{w}\mid\mathcal{D})}$ (the same for $\tilde{\boldsymbol{w}}$) and $p_\bullet(\widehat{p_\circ}) \coloneqq p(y = \operatorname{argmax} p(y \mid \boldsymbol{x}, \circ) \mid \boldsymbol{x}, \bullet)$. Expressions with the same cell coloring are equivalent to each other. Each measure of TU additively decomposes into AU and EU.

	Predicting model	Approximation of true predictive distribution		
		(1) $\tilde{\boldsymbol{w}}$	(2) $\mathrm{E}_{\tilde{\boldsymbol{w}}}$	(3) $\tilde{\boldsymbol{w}} \sim p(\tilde{\boldsymbol{w}} \mid \mathcal{D})$
TU	(A) $\boldsymbol{w}$	$1 - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})$	$1 - p_{\boldsymbol{w}}(\widehat{p_{\mathcal{D}}})$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[1 - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$1 - p_{\mathcal{D}}(\widehat{p_{\tilde{\boldsymbol{w}}}})$	$1 - p_{\mathcal{D}}(\widehat{p_{\mathcal{D}}})$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[1 - p_{\mathcal{D}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\,[1 - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]$	$\mathrm{E}_{\boldsymbol{w}}\,[1 - p_{\boldsymbol{w}}(\widehat{p_{\mathcal{D}}})]$	$\mathrm{E}_{\boldsymbol{w}}\,[\,\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[1 - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]]$
AU	(A) $\boldsymbol{w}$	$1 - \max p_{\boldsymbol{w}}$	$1 - \max p_{\boldsymbol{w}}$	$1 - \max p_{\boldsymbol{w}}$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$1 - \max p_{\mathcal{D}}$	$1 - \max p_{\mathcal{D}}$	$1 - \max p_{\mathcal{D}}$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$1 - \mathrm{E}_{\boldsymbol{w}}\,[\max p_{\boldsymbol{w}}]$	$1 - \mathrm{E}_{\boldsymbol{w}}\,[\max p_{\boldsymbol{w}}]$	$1 - \mathrm{E}_{\boldsymbol{w}}\,[\max p_{\boldsymbol{w}}]$
EU	(A) $\boldsymbol{w}$	$\max p_{\boldsymbol{w}} - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})$	$\max p_{\boldsymbol{w}} - p_{\boldsymbol{w}}(\widehat{p_{\mathcal{D}}})$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[\max p_{\boldsymbol{w}} - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$\max p_{\mathcal{D}} - p_{\mathcal{D}}(\widehat{p_{\tilde{\boldsymbol{w}}}})$	$\cancelto{0}{\max p_{\mathcal{D}} - p_{\mathcal{D}}(\widehat{p_{\mathcal{D}}})}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[\max p_{\mathcal{D}} - p_{\mathcal{D}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\,[\max p_{\boldsymbol{w}} - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]$	$\mathrm{E}_{\boldsymbol{w}}\,[\max p_{\boldsymbol{w}} - p_{\boldsymbol{w}}(\widehat{p_{\mathcal{D}}})]$	$\mathrm{E}_{\boldsymbol{w}}\,[\,\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[\max p_{\boldsymbol{w}} - p_{\boldsymbol{w}}(\widehat{p_{\tilde{\boldsymbol{w}}}})]]$

A.4 GENERALIZATION TO OTHER PROPER SCORING RULES

Another perspective on our measure of uncertainty (Eq. (5)) is the interpretation of Kotelevskii et al. [2025] and Hofman et al. [2024b], which consider Bayesian approximations of multiple proper scoring rules as measures of uncertainty. They consider the zero-one, Brier, and Spherical score in addition to the log-score, which is the cross-entropy upon which the information-theoretic measures we discussed in the main paper are based (see Eq. (5)). Note however, that they consider the opposite order of arguments (see Apx. A.5), which leads to similar measures due to symmetry, yet has a different interpretation. For the zero-one score, the resulting framework of measures according to our interpretation is given in Tab. 3, for the Brier score in Tab. 4 and for the spherical score it is given in Tab. 5.

A.5 ALTERNATIVE MEASURE

The reverse order of the arguments for the cross-entropy in Eq. (5), that is, $\mathrm{CE}(p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)\ ;\ p(y \mid \boldsymbol{x}, \cdot))$, gives rise to an alternative measure that is consistent with Eq. (1). This measure, also known as "pointwise risk" under the log score at an input (point) $\boldsymbol{x}$, has been considered as a measure of predictive uncertainty [Gruber and Buettner, 2023, Lahlou et al., 2023, Kotelevskii et al., 2025, Hofman et al., 2024b]. Kotelevskii et al. [2025] introduced a framework based on similar ideas for Bayesian approximation as we consider in our framework (i.e. settings `(B)`, `(C)`, `(2)` and `(3)`) and is the work closest to our work. They introduce a framework based on decompositions of proper scoring rules (pointwise risk measures), where the logarithmic proper scoring rule is a special case, and subsequent Bayesian approximation. However, they do not cover settings `(A)` and `(1)` in our framework, introduce their framework based on the notion of pointwise risk and focus on different empirical aspects of their resulting measures.

We argue that our proposed measure (Eq. (5)) has a more meaningful interpretation than $\mathrm{CE}(p(y \mid \boldsymbol{x}, \boldsymbol{w}^*)\ ;\ p(y \mid \boldsymbol{x}, \cdot))$. Our measure considers the uncertainty inherent to predicting with the selected model, plus the uncertainty due to any potential mismatch with the true model. The alternative measure considers the uncertainty inherent to predicting with the true model, plus the uncertainty due to any potential mismatch with the selected model. However, we generally don't know the true model, thus can't actually use it to predict and have to resort to an approximation of the true model anyways.

Table 4: **Our proposed framework applied under the Brier score.** Each measure denotes a different instantiation of our proposed measure given by Eq. (5), but using the Brier score instead of the cross-entropy (log score) for different assumptions about the predicting model and how the true model is approximated. For brevity, we define $p_{\boldsymbol{w}} \coloneqq p(y \mid \boldsymbol{x}, \boldsymbol{w})$, $p_{\mathcal{D}} \coloneqq p(y \mid \boldsymbol{x}, \mathcal{D})$, and $\mathrm{E}_{\boldsymbol{w}} \coloneqq \mathrm{E}_{p(\boldsymbol{w}\mid\mathcal{D})}$ (the same for $\tilde{\boldsymbol{w}}$). The 2-norm is defined as $\|p(y \mid \boldsymbol{x}, \bullet)\|_2 \coloneqq \sqrt{\sum_{k=1}^{K} p(y = k \mid \boldsymbol{x}, \bullet)^2}$. Expressions with the same cell coloring are equivalent to each other. Each measure of TU additively decomposes into AU and EU.

	Predicting model	Approximation of true predictive distribution		
		(1) $\tilde{\boldsymbol{w}}$	(2) $\mathrm{E}_{\tilde{\boldsymbol{w}}}$	(3) $\tilde{\boldsymbol{w}} \sim p(\tilde{\boldsymbol{w}} \mid \mathcal{D})$
TU	(A) $\boldsymbol{w}$	$1 - \|p_{\boldsymbol{w}}\|_2^2 + \|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2$	$1 - \|p_{\boldsymbol{w}}\|_2^2 + \|p_{\boldsymbol{w}} - p_{\mathcal{D}}\|_2^2$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[1 - \|p_{\boldsymbol{w}}\|_2^2 + \|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$1 - \|p_{\mathcal{D}}\|_2^2 + \|p_{\mathcal{D}} - p_{\tilde{\boldsymbol{w}}}\|_2^2$	$1 - \|p_{\mathcal{D}}\|_2^2 + \cancelto{0}{\|p_{\mathcal{D}} - p_{\mathcal{D}}\|_2^2}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[1 - \|p_{\mathcal{D}}\|_2^2 + \|p_{\mathcal{D}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\,[1 - \|p_{\boldsymbol{w}}\|_2^2 + \|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]$	$\mathrm{E}_{\boldsymbol{w}}\,[1 - \|p_{\boldsymbol{w}}\|_2^2 + \|p_{\boldsymbol{w}} - p_{\mathcal{D}}\|_2^2]$	$\mathrm{E}_{\boldsymbol{w}}\,[\,\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[1 - \|p_{\boldsymbol{w}}\|_2^2 + \|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]]$
AU	(A) $\boldsymbol{w}$	$1 - \|p_{\boldsymbol{w}}\|_2^2$	$1 - \|p_{\boldsymbol{w}}\|_2^2$	$1 - \|p_{\boldsymbol{w}}\|_2^2$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$1 - \|p_{\mathcal{D}}\|_2^2$	$1 - \|p_{\mathcal{D}}\|_2^2$	$1 - \|p_{\mathcal{D}}\|_2^2$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\,[1 - \|p_{\boldsymbol{w}}\|_2^2]$	$\mathrm{E}_{\boldsymbol{w}}\,[1 - \|p_{\boldsymbol{w}}\|_2^2]$	$\mathrm{E}_{\boldsymbol{w}}\,[1 - \|p_{\boldsymbol{w}}\|_2^2]$
EU	(A) $\boldsymbol{w}$	$\|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2$	$\|p_{\boldsymbol{w}} - p_{\mathcal{D}}\|_2^2$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[\|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$\|p_{\mathcal{D}} - p_{\tilde{\boldsymbol{w}}}\|_2^2$	$\cancelto{0}{\|p_{\mathcal{D}} - p_{\mathcal{D}}\|_2^2}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[\|p_{\mathcal{D}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}}\,[\|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]$	$\mathrm{E}_{\boldsymbol{w}}\,[\|p_{\boldsymbol{w}} - p_{\mathcal{D}}\|_2^2]$	$\mathrm{E}_{\boldsymbol{w}}\,[\,\mathrm{E}_{\tilde{\boldsymbol{w}}}\,[\|p_{\boldsymbol{w}} - p_{\tilde{\boldsymbol{w}}}\|_2^2]]$

A.6 REGRESSION

For a probabilistic regression model, e.g. under a Gaussian assumption, the distribution parameters are estimated, i.e. mean and variance for the Gaussian predictive distribution. The model is then trained by minimizing the negative log-likelihood under the training dataset.

Many works follow Depeweg et al. [2018] and utilize a variance decomposition for uncertainty quantification, where the AU is the expected variance and the EU is the variance of means, where expectation and variance are over the model posterior. However, Depeweg et al. [2018] also consider the uncertainty measure given by Eq. (3), using differential entropies for the continuous predictive distributions. The same can be done in order to adapt our framework in Tab. 1 for continuous predictive distributions.

Nevertheless, there are two important drawbacks one need to consider when doing this. First, differential entropy can be unbounded, depending on the nature of the predictive distribution. For the example of a Gaussian, it can be between $-\infty$ and ∞. In addition, it is not invariant to a change of variables, making it a relative rather than an absolute measure. Second, the posterior predictive distribution as defined in Eq. (2) is generally a mixture of individual distributions, unlike in the discrete case. This makes MC approximations of the resulting measures more involved.

Table 5: **Our proposed framework applied under the spherical score.** Each measure denotes a different instantiation of our proposed measure given by Eq. (5), but using the spherical score instead of the cross-entropy (log score) for different assumptions about the predicting model and how the true model is approximated. For brevity, we define $p_{\boldsymbol{w}} \coloneqq p(y \mid \boldsymbol{x}, \boldsymbol{w})$, $p_{\mathcal{D}} \coloneqq p(y \mid \boldsymbol{x}, \mathcal{D})$, and $\mathrm{E}_{\boldsymbol{w}} \coloneqq \mathrm{E}_{p(\boldsymbol{w}|\mathcal{D})}$ (the same for $\tilde{\boldsymbol{w}}$). The 2-norm is defined as $\|p_\bullet\|_2 \coloneqq \sqrt{\sum_{k=1}^{K} p(y = k \mid \boldsymbol{x}, \bullet)^2}$. Furthermore, the scalar product is defined as $\langle p_\bullet, p_\circ \rangle \coloneqq \sum_{k=1}^{K} p(y = k \mid \boldsymbol{x}, \bullet) \cdot p(y = k \mid \boldsymbol{x}, \circ)$. Expressions with the same cell coloring are equivalent to each other. Each measure of TU additively decomposes into AU and EU.

	Predicting model	Approximation of true predictive distribution		
		(1) $\tilde{\boldsymbol{w}}$	(2) $\mathrm{E}_{\tilde{\boldsymbol{w}}}$	(3) $\tilde{\boldsymbol{w}} \sim p(\tilde{\boldsymbol{w}} \mid \mathcal{D})$
TU	(A) $\boldsymbol{w}$	$1 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2}$	$1 - \frac{\langle p_{\boldsymbol{w}}, p_{\mathcal{D}} \rangle}{\|p_{\mathcal{D}}\|_2}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}} \left[1 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$1 - \frac{\langle p_{\mathcal{D}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2}$	$1 - \frac{\langle p_{\mathcal{D}}, p_{\mathcal{D}} \rangle}{\|p_{\mathcal{D}}\|_2}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}} \left[1 - \frac{\langle p_{\mathcal{D}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}} \left[1 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right]$	$\mathrm{E}_{\boldsymbol{w}} \left[1 - \frac{\langle p_{\boldsymbol{w}}, p_{\mathcal{D}} \rangle}{\|p_{\mathcal{D}}\|_2} \right]$	$\mathrm{E}_{\boldsymbol{w}} \left[\mathrm{E}_{\tilde{\boldsymbol{w}}} \left[1 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right] \right]$
AU	(A) $\boldsymbol{w}$	$1 - \|p_{\boldsymbol{w}}\|_2$	$1 - \|p_{\boldsymbol{w}}\|_2$	$1 - \|p_{\boldsymbol{w}}\|_2$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$1 - \|p_{\mathcal{D}}\|_2$	$1 - \|p_{\mathcal{D}}\|_2$	$1 - \|p_{\mathcal{D}}\|_2$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}} \left[1 - \|p_{\boldsymbol{w}}\|_2 \right]$	$\mathrm{E}_{\boldsymbol{w}} \left[1 - \|p_{\boldsymbol{w}}\|_2 \right]$	$\mathrm{E}_{\boldsymbol{w}} \left[1 - \|p_{\boldsymbol{w}}\|_2 \right]$
EU	(A) $\boldsymbol{w}$	$\|p_{\boldsymbol{w}}\|_2 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2}$	$\|p_{\boldsymbol{w}}\|_2 - \frac{\langle p_{\boldsymbol{w}}, p_{\mathcal{D}} \rangle}{\|p_{\mathcal{D}}\|_2}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}} \left[\|p_{\boldsymbol{w}}\|_2 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right]$
	(B) $\mathrm{E}_{\boldsymbol{w}}$	$\|p_{\mathcal{D}}\|_2 - \frac{\langle p_{\mathcal{D}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2}$	$\cancelto{0}{\|p_{\mathcal{D}}\|_2 - \frac{\langle p_{\mathcal{D}}, p_{\mathcal{D}} \rangle}{\|p_{\mathcal{D}}\|_2}}$	$\mathrm{E}_{\tilde{\boldsymbol{w}}} \left[\|p_{\mathcal{D}}\|_2 - \frac{\langle p_{\mathcal{D}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right]$
	(C) $\boldsymbol{w} \sim p(\boldsymbol{w} \mid \mathcal{D})$	$\mathrm{E}_{\boldsymbol{w}} \left[\|p_{\boldsymbol{w}}\|_2 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right]$	$\mathrm{E}_{\boldsymbol{w}} \left[\|p_{\boldsymbol{w}}\|_2 - \frac{\langle p_{\boldsymbol{w}}, p_{\mathcal{D}} \rangle}{\|p_{\mathcal{D}}\|_2} \right]$	$\mathrm{E}_{\boldsymbol{w}} \left[\mathrm{E}_{\tilde{\boldsymbol{w}}} \left[\|p_{\boldsymbol{w}}\|_2 - \frac{\langle p_{\boldsymbol{w}}, p_{\tilde{\boldsymbol{w}}} \rangle}{\|p_{\tilde{\boldsymbol{w}}}\|_2} \right] \right]$

B EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

In this section, we provide additional empirical results of our evaluation of the proposed framework of uncertainty measures. The code to reproduce our experiments is available at `https:\github.com/ml-jku/uncertainty-measures`.

B.1 ILLUSTRATIVE EXAMPLE

Here, we provide an illustrative synthetic example often discussed in the literature [Wimmer et al., 2023, Schweighofer et al., 2023a, Sale et al., 2023b]. We consider a predictor defined as a Bernoulli distribution leading to the predictive distribution $p(y \mid \theta)$. Thus, there is no model involved for mapping from the input space to the Bernoulli parameter. The only free parameter is the Bernoulli parameter. Therefore, the posterior distribution is defined as $p(\theta \mid \mathcal{D}) = p(\mathcal{D} \mid \theta)p(\theta)/p(\mathcal{D})$. To examplify our framework, we consider a Beta posterior distribution $Beta(\theta; 2, 3)$. The true Bernoulli parameter θ^* is not known.

Results are shown in Fig. 8, depicting what is considered as predicting model (green) and what is compared to as approximation of the true model. The green line for measures `(A1/2/3)` and the violet line for measures `(A/B/C1)` were chosen arbitrarily, but different to the expected Bernoulli parameter to exemplify the differences between measures.

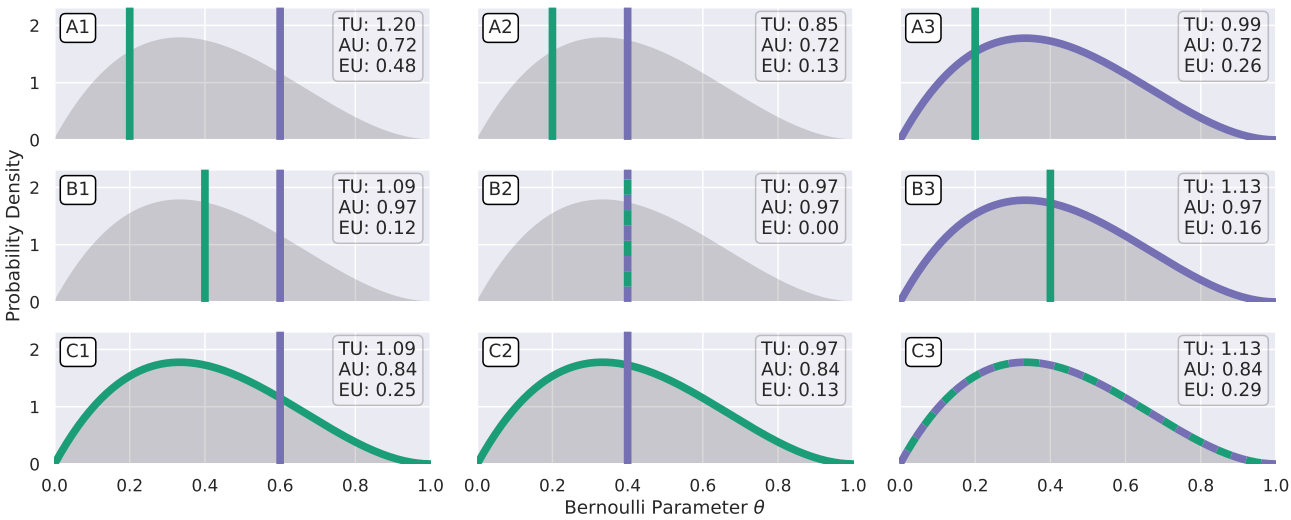


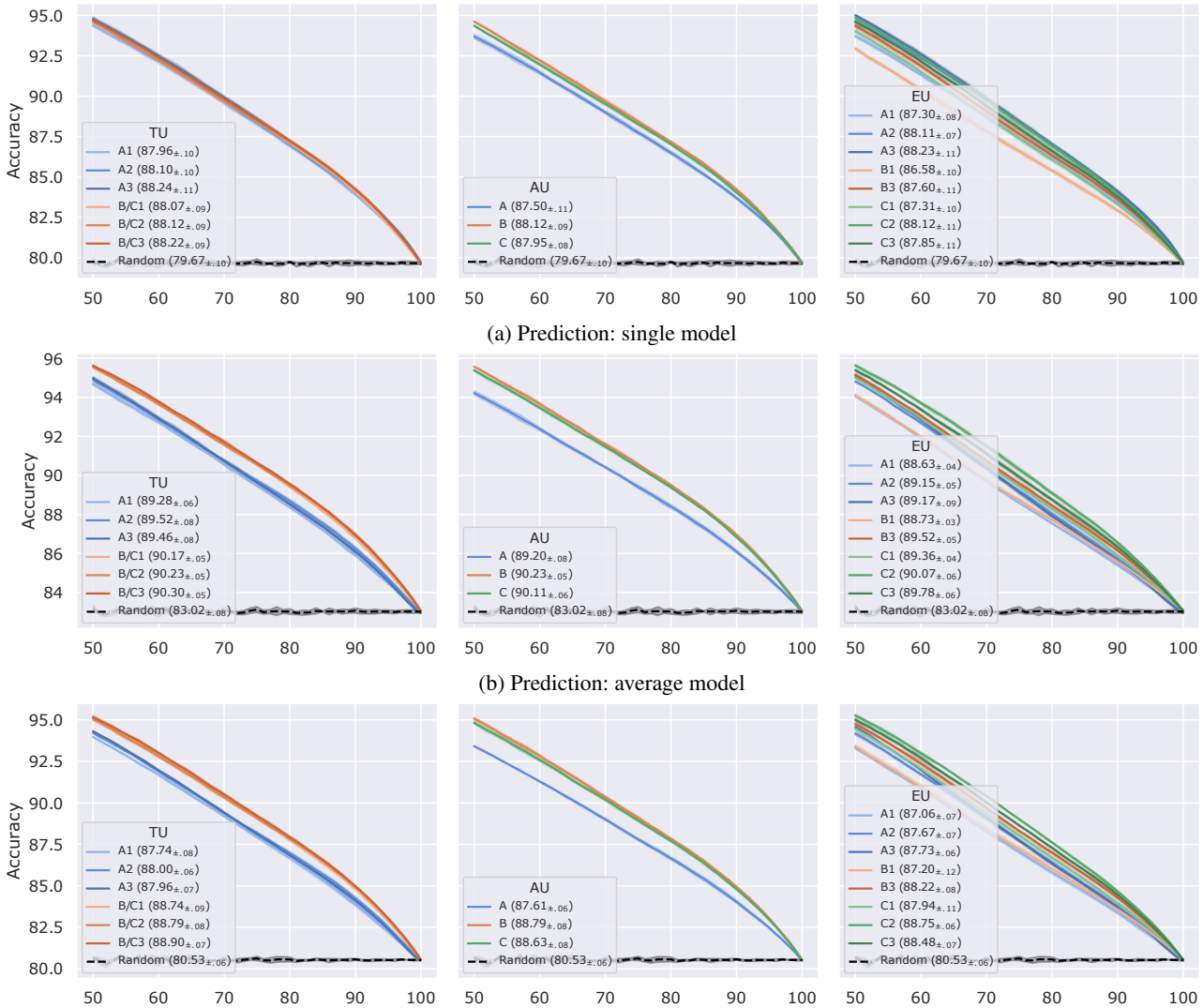
Figure 8: Uncertainty measures given by the predicting model and the approximation of the true model. We consider the posterior distribution $Beta(\theta; 2, 3)$, shaded in gray.

B.2 DETAILED RESULTS ON MAIN EXPERIMENTS

The results for selective prediction in the main paper only provide the AUARC, but not individual accuracy-rejection curves, which we provide in this section. Furthermore, results for misclassification detection and OOD detection in the main paper show aggregate performances over multiple datasets to provide more robust conclusions about the performance of individual measures of uncertainty. In this section, we provide individual results for completeness.

Selective prediction. We provide detailed additional results for selective prediction as discussed in the main paper. The accuracy rejection curves and their AUARCs are shown in Fig. 9. For predicting using a single model, the best measure is TU `(A3)`. Overall, measures that consider the single model as predicting model perform well overall for measures of TU, AU and EU. For predicting using the average model, TU `(B/C3)` performs best. Finally, the results for predicting under a model according to the posterior are very similar to the results under the average model.

Misclassification detection. The detailed results for misclassification detection are given in Fig. 10 for a single predicting model, in Fig. 11 for the average predicting model as well as in Fig. 12 for predicting with a model according to the posterior. Although there are nuanced differences between datasets, conclusions translate very well between them.



(a) Prediction: single model

(b) Prediction: average model

(c) Prediction: according to posterior

Figure 9: **Selective prediction results.** Accuracies per fraction of datapoints the respecitive predictive model model predicts on, as well as AUARC (tabulated in legend) using different proposed measures of uncertainty as score. Accuracies are averaged over all datasets. Means and standard deviations are calculated using five runs.

OOD detection. The detailed results for OOD detection for CIFAR10 as ID dataset are given in Fig. 13, for CIFAR100 as ID dataset in Fig. 14, for SVHN as ID dataset in Fig. 15 and for TIN as ID dataset in Fig. 16. We observe the highest variability of experiments for TIN as ID dataset, where there is high variability depending on the OOD dataset. For the other ID datasets, the different OOD datasets lead to very similar results.

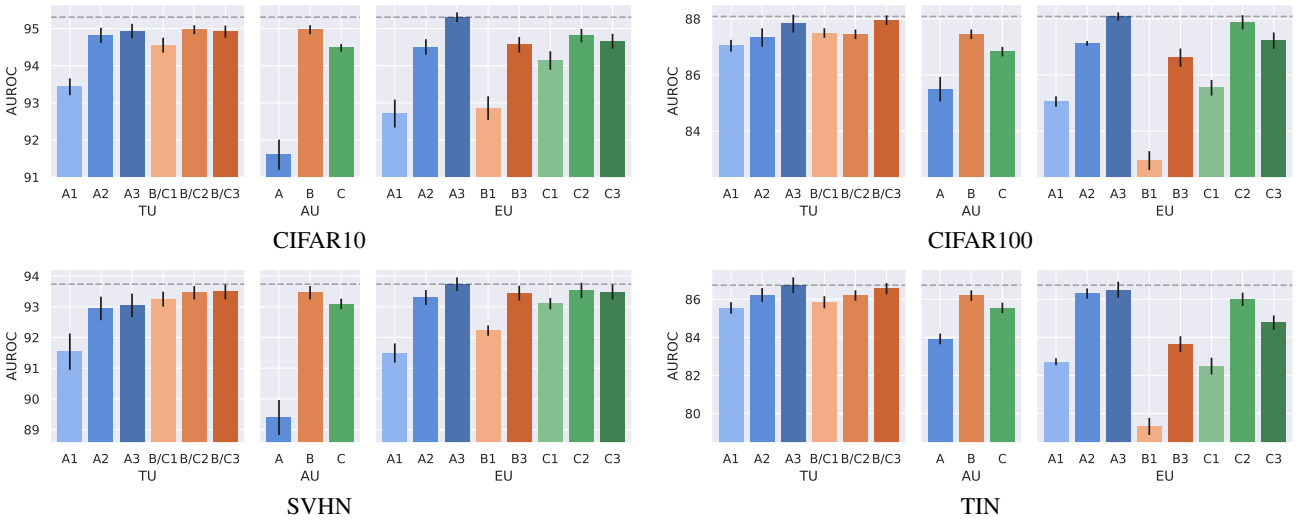


Figure 10: **Misclassification detection results under (i) single predicting model.** AUROC for distinguishing between correctly and incorrectly predicted datapoints under a single predicting model, using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.

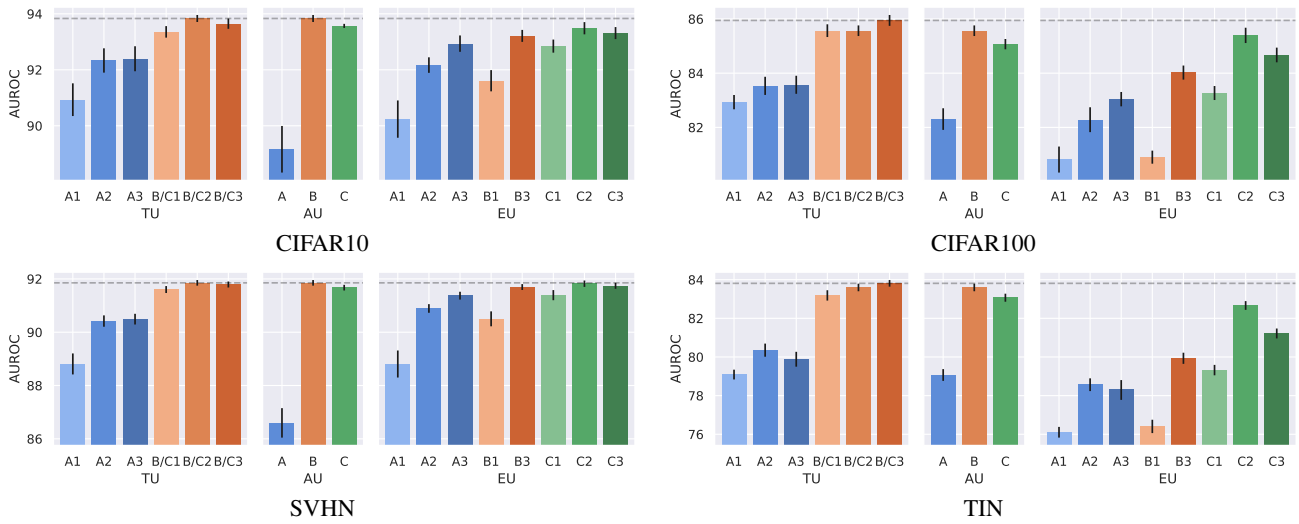


Figure 11: **Misclassification detection results under (ii) average predicting model.** AUROC for distinguishing between correctly and incorrectly predicted datapoints under the average predicting model, using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.

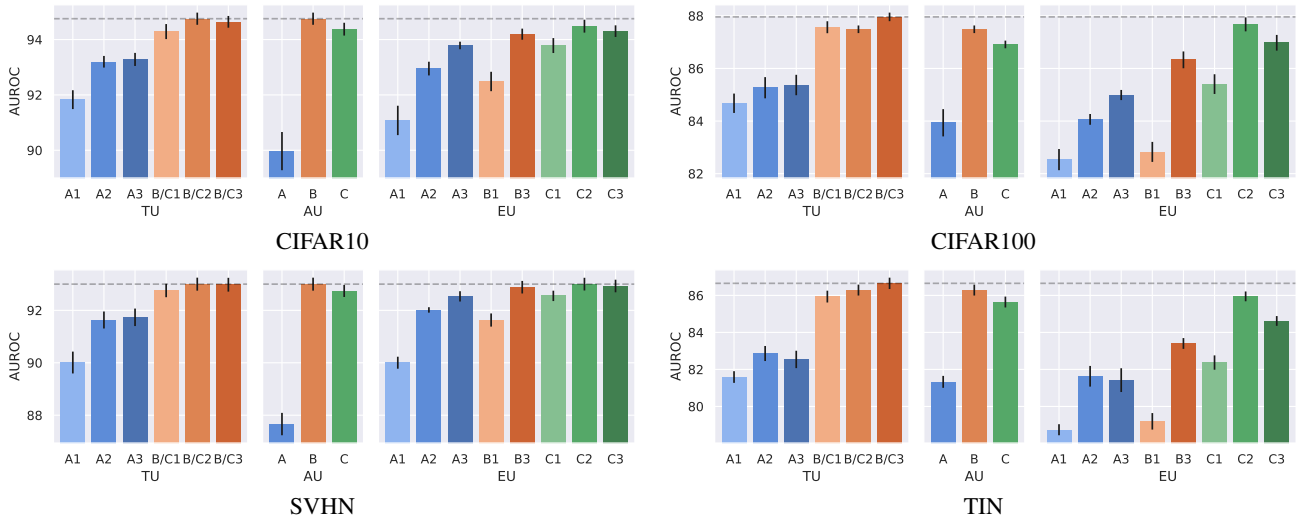


Figure 12: **Misclassification detection results under (iii) a model according to the posterior.** AUROC for distinguishing between correctly and incorrectly predicted datapoints under a model according to the posterior, using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.

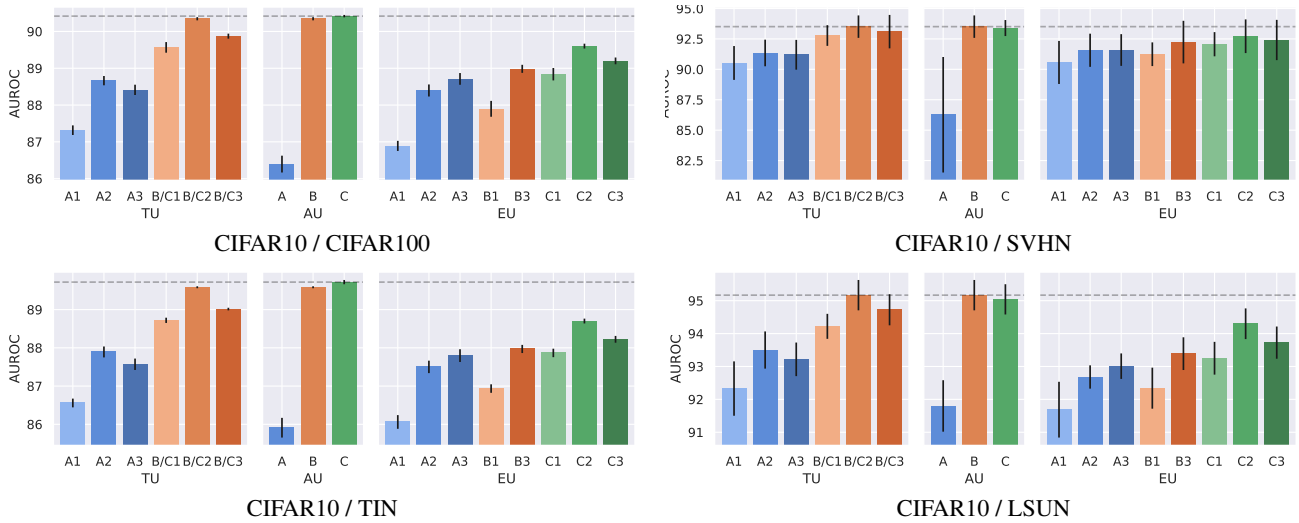


Figure 13: **OOD detection results for CIFAR10.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.

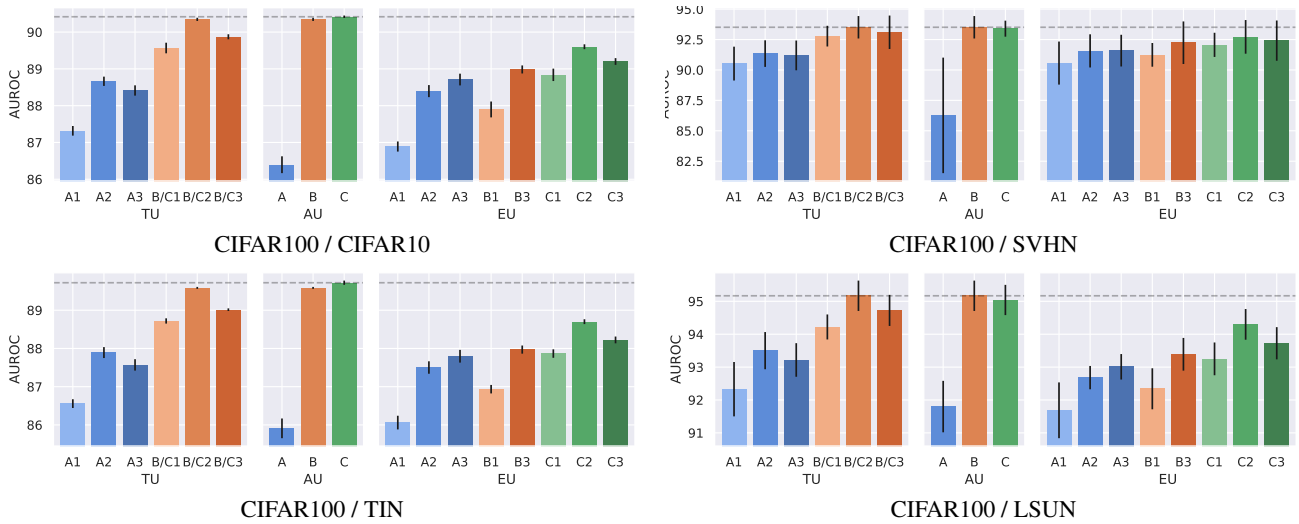


Figure 14: **OOD detection results for CIFAR100.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.

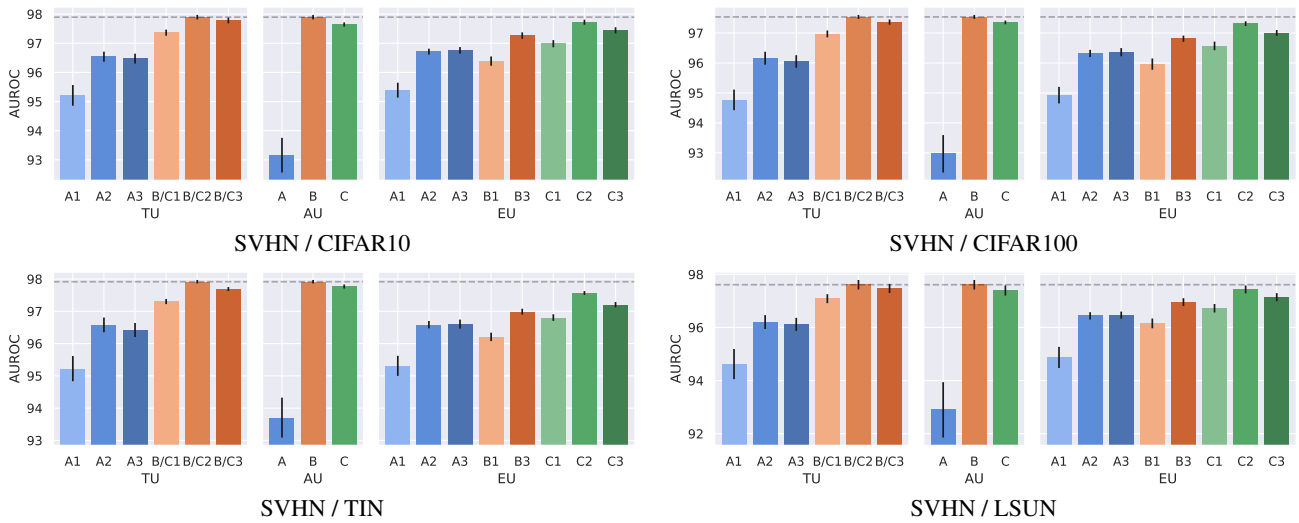


Figure 15: **OOD detection results for SVHN.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.

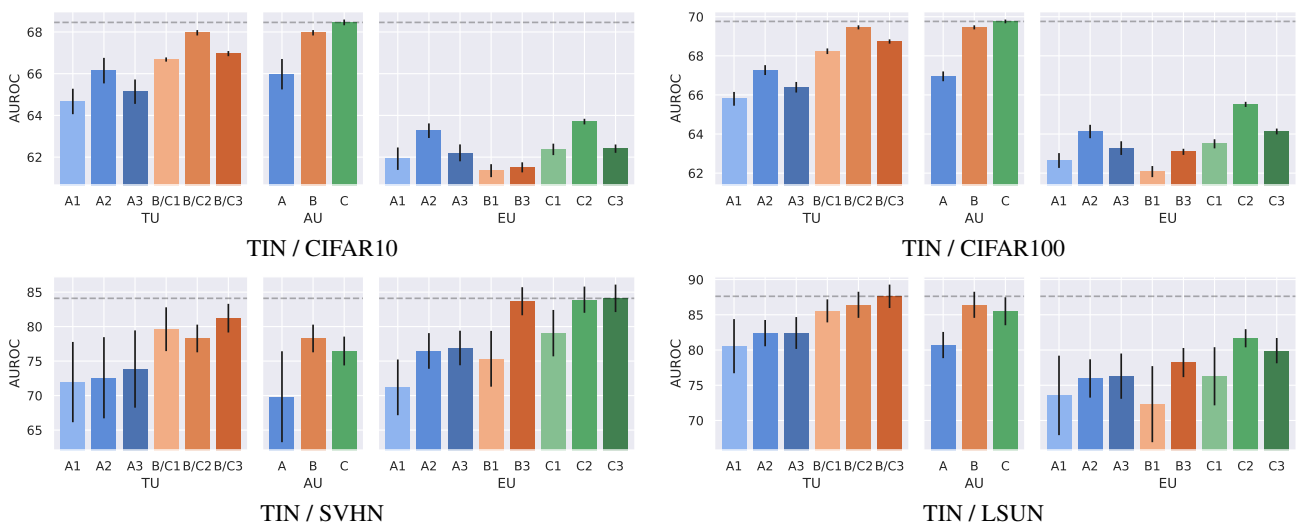


Figure 16: **OOD detection results for TIN.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.

B.3 DIFFERENT NETWORK ARCHITECTURE

We want to assess the influence of the network architecture on the ranking of the results. To that end, we also trained DEs of DenseNet169 and RegNet-Y 800MF, using the same training recipe as for ResNet-18 described in Sec. 5. A comparison of the sampled models is given in Fig. 17. We observe that ResNet-18 performs a bit better than the two other models, with RegNet-Y 800MF being the worst models in terms of negative log-likelihood (NLL) and accuracies. In terms of AU and EU, we observe only minor differences in the upper tails of the distributions for CIFAR100 and TIN. For CIFAR10 and SVHN, we do not observe differences. Next, we analyze the influence of the network architecture on the misclassification and OOD detection tasks.

Misclassification detection. The results for misclassification detection using DEs with different model architectures are given in Fig. 18. We observe no major differences for different models (per column) under a given predicting model (per row).

OOD detection. The results for OOD detection using DEs with different model architectures are given in Fig. 19. We observe that the AU `(C)` is the best measure for DenseNet-169 and RegNet-Y 800MF, while it is AU `(B)` which is equivalent to TU `(B/C2)` for ResNet-18. However, the general trends are the same across all architectures.

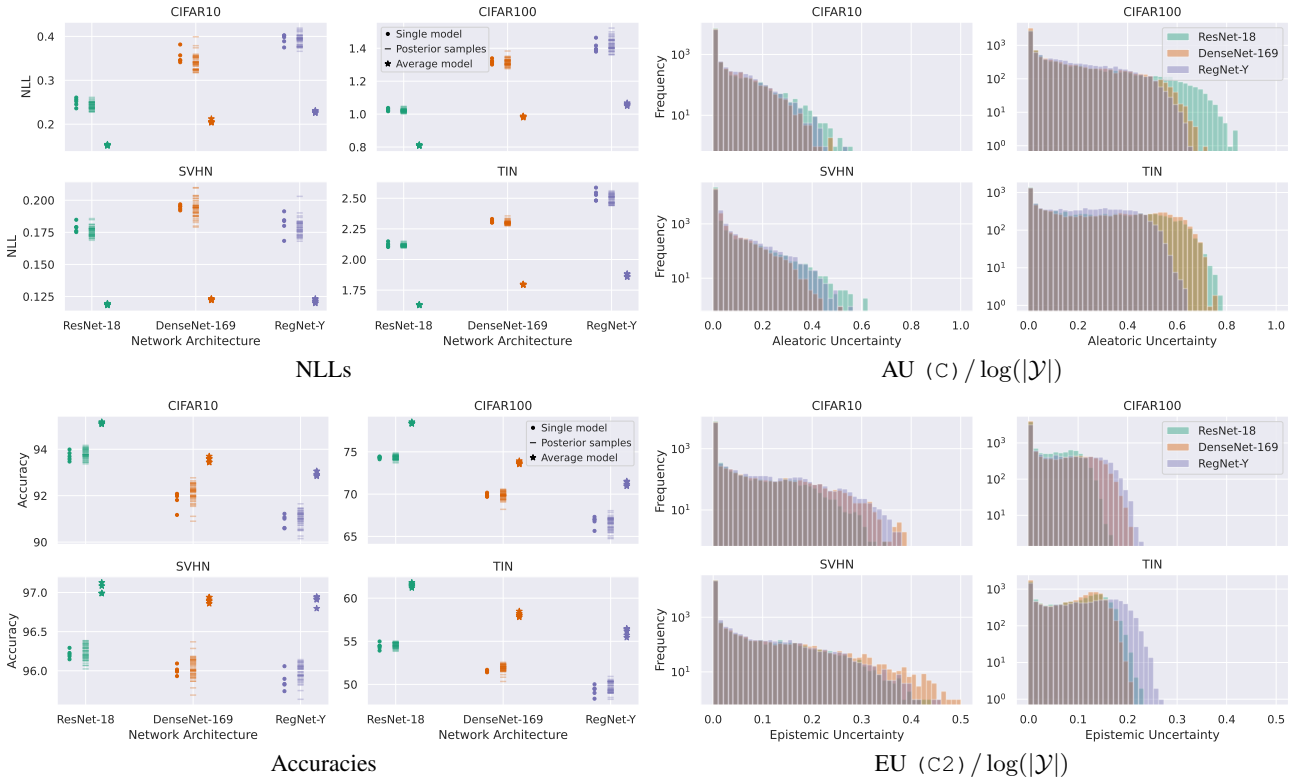
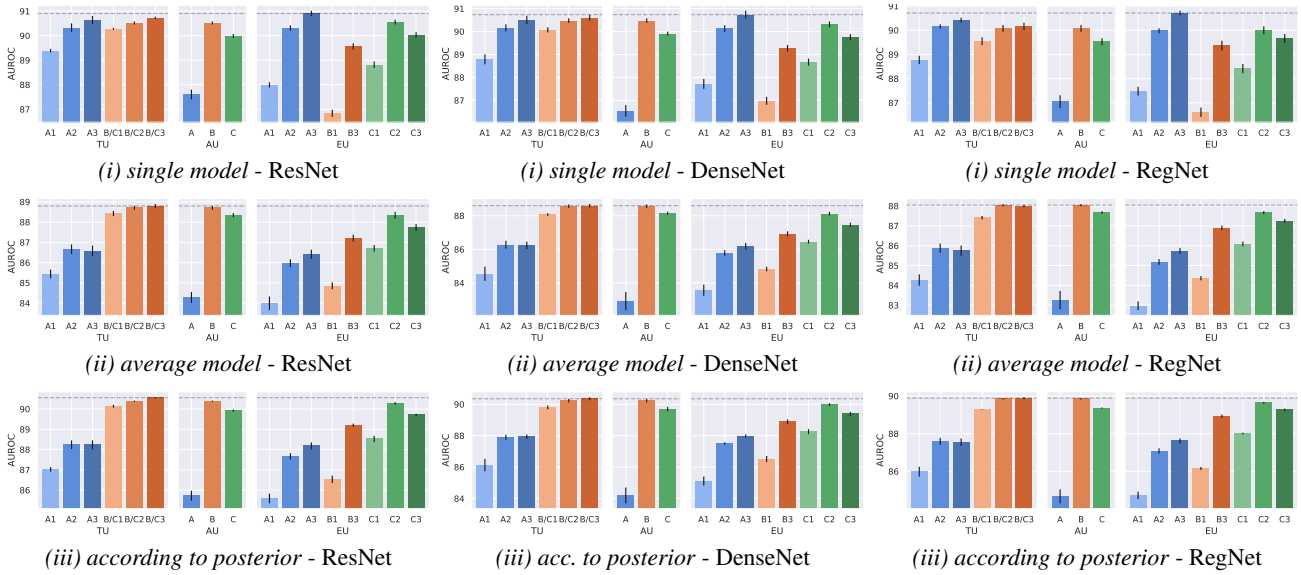


Figure 17: **Comparison of network architectures.** Results are obtained on the test split of the respective dataset. We compare the NLLs and accuracies for different models obtained through DEs on ResNet-18, DenseNet-169 and RegNet-Y 800MF. The single model is randomly selected among all sampled models. We depict all models sampled in five runs. Furthermore, the normalized AU `(C)` and the normalized EU `(C2)` are given per sampling method. All three network architectures lead to similar results on all considered datasets.

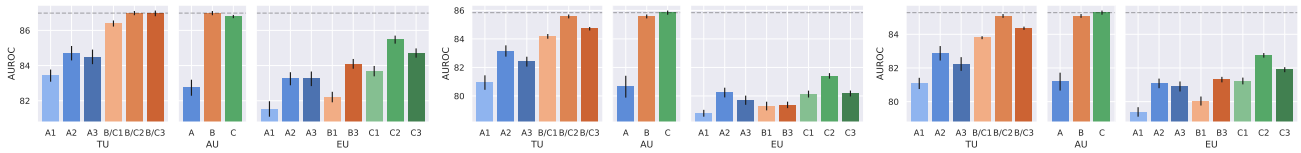


(i) single model - ResNet *(i) single model* - DenseNet *(i) single model* - RegNet

(ii) average model - ResNet *(ii) average model* - DenseNet *(ii) average model* - RegNet

(iii) according to posterior - ResNet *(iii) acc. to posterior* - DenseNet *(iii) according to posterior* - RegNet

Figure 18: **Misclassification detection results for DE with different model architectures and under different predicting models.** AUROC for distinguishing between correctly and incorrectly predicted samples under different predicting models, using the different proposed measures of uncertainty as score. Means and standard deviations are calculated using five runs.



ResNet-18 DenseNet-169 RegNet-Y 800MF

Figure 19: **OOD detection results for DE with different model architectures.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. AUROCs are averaged over all ID / OOD combinations. Means and standard deviations are calculated using five runs.

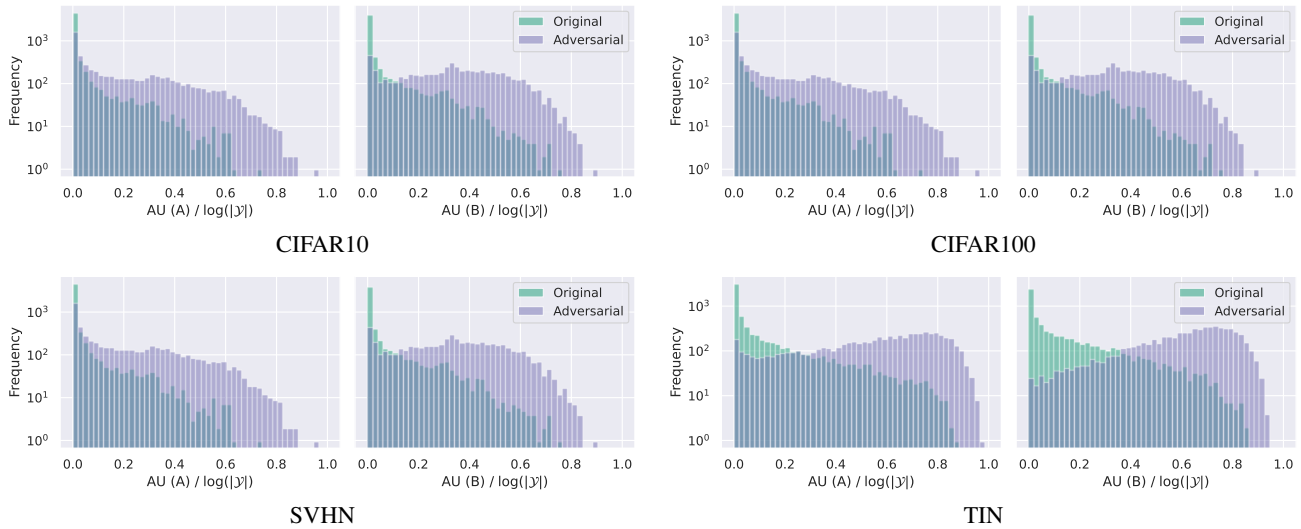


Figure 20: **Histogram of AU (A) and AU (B) for original and adversarial datapoints obtained through applying FGSM.** AUs are normalized with $\log(|\mathcal{Y}|)$ to be more comparable across datasets with different number of classes.

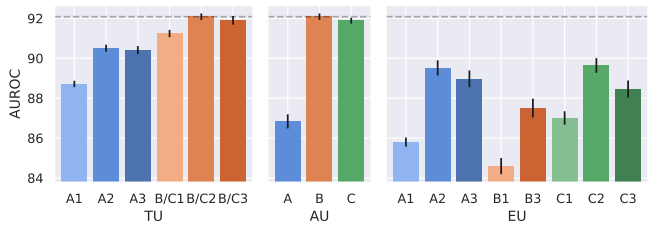


Figure 21: **Adversarial example detection (FGSM).** Means and standard deviations are calculated on five runs.

B.4 ADVERSARIAL EXAMPLE DETECTION

We want to investigate the effect of adversarially created inputs on the uncertainty estimates. Throughout this experiments, we consider adversarial attacks on the single network. However, it would also be possible to attack the average model, albeit more computationally expensive. As adversarial examples are known to transfer well between models of similar architecture [Goodfellow et al., 2015], results for attacking the average model are expected to be relatively similar to those presented here.

We consider two different adversarial attacks, FGSM [Goodfellow et al., 2015] and PGD under infinity norm perturbation [Madry et al., 2018]. For our experiments, we only consider the subset of the test datasets that are predicted correctly. This we refer to as the *original* dataset. Then we apply the adversarial attacks the datapoints in the original dataset and select those datapoints where the model was successfully fooled to predict incorrectly. This we refer to as the *adversarial* dataset. We utilize the different uncertainty scores to calculate the AUROC of distinguishing between the original and the adversarial dataset, akin to the OOD detection experiments reported in the main paper. We also investigated the AUPR and FPR@TPR95 as alternative metrics, which lead to equivalent conclusions.

FGSM. We start with the results obtained through the FGSM attack with $\epsilon = 8/255$. Histograms of the AU (A), the entropy of the predictive distribution of the single attacked model and the AU (B), the entropy of the predictive distribution under the average model, are shown in Fig. 20. We observe a shift towards higher AUs for the adversarial datapoints compared to the original datapoints. This shift appears more pronounced for AU (B), which makes sense as the adversarial examples have been obtained with the single model.

The main results are shown in Fig. 21, denoting the AUROC of distinguishing between the original and the adversarial datapoints using the different measures of uncertainty as score. We observe qualitatively very similar results to the OOD detection experiments, in that TU and AU measures for cases (B) and (C) are the most effective. Surprisingly, EU measures underperform for adversarial example detection, compared to TU and AU measures.

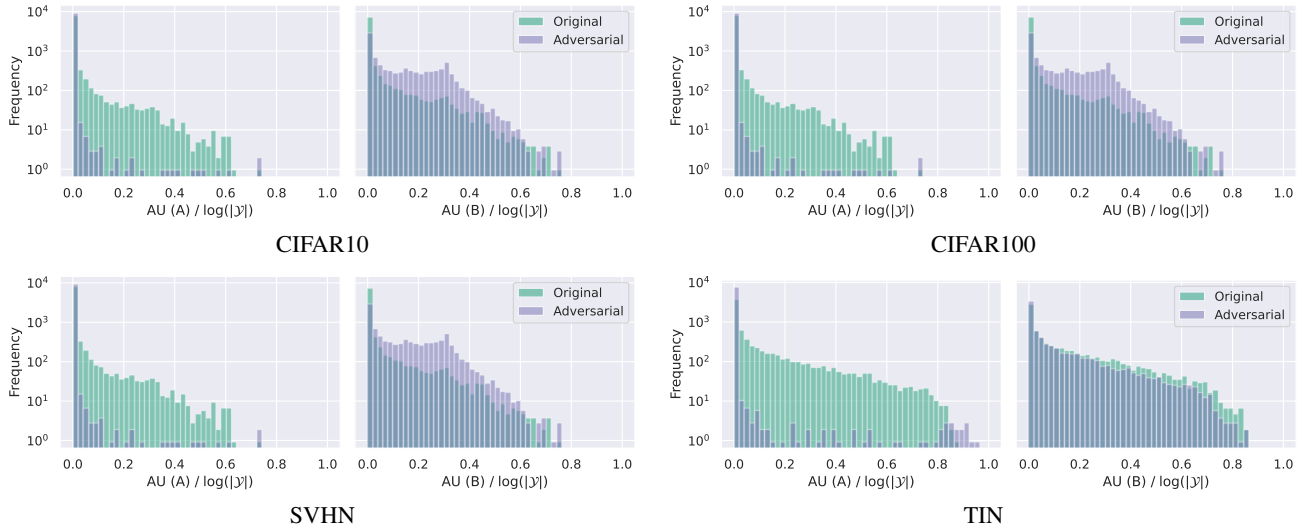


Figure 22: **Histogram of AU (A) and AU (B) for original and adversarial datapoints obtained through applying L_∞-PGD.** AUs are normalized with $\log(|\mathcal{Y}|)$ to be more comparable across datasets with different number of classes.

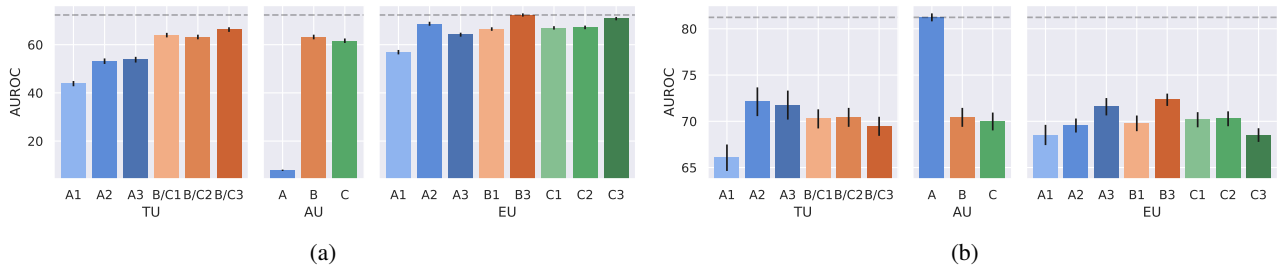


Figure 23: **Adversarial example detection (L_∞-PGD)**. (a) Assuming higher uncertainty for adversarial examples. (b) Assuming lower uncertainty for adversarial examples. Means and standard deviations are calculated on five runs.

PGD. Next, we conduct the same investigation using the L_∞-PGD attack with $\epsilon = 8/255$. Histograms of the AU (A), the entropy of the predictive distribution of the single attacked model and the AU (B), the entropy of the predictive distribution under the average model, are shown in Fig. 22. We observe a shift towards **lower** AU (A) for the adversarial datapoints compared to the original datapoints. However for for AU (B), adversarial datapoints exhibit slightly higher values than the original datapoints.

The results are given in Fig. 23(a), denoting the AUROC of distinguishing between the original and the adversarial datapoints using the different measures of uncertainty as score. We observe that all measures except TU (A1) and AU (A) perform better than random. The very bad performance of AU (A) stems from the fact that adversarial datapoints exhibit lower uncertainties than the original datapoints (see Fig. 22). However, contrary to the experiments with FGSM, measures of EU perform best for PGD.

The two experiments for adversarial example detection were conducted under the assumption that adversarial datapoints should exhibit higher uncertainty than the original datapoints. Finally, we investigate a special variant of our experiments with L_∞-PGD adversarial examples, where we assume that adversarial datapoints exhibit **lower** uncertainty than the original datapoints. The results are shown in Fig. 23(b). We observe that using AU (A) leads to the best results in this setting. However, these results do not help to attain a mechanism for adversarial robustness, as we leverage additional side information that the single model was fooled into being very confident about the adversarial examples. Attackers could add constraints on the deviation between the AU (A) under the original and the adversarial datapoint in an improved version of the L_∞-PGD attack, rendering this detection mechanism useless.

B.5 ACTIVE LEARNING

In this section, we provide additional details on the active learning experiments described in the main paper.

The network architective of the utilized small CNN are: 5x5 conv [1 to 6 channels], 2x2 max-pool, 5x5 conv [6 to 12 channels], 2x2 max-pool, two linear layers with hidden size 32 and a final output linear layer; ReLU activations after each max-pool and linear layer except the last as well as dropout with dropout rate 0.2 between linear layers. The training of the models utilized the Adam optimizer [Kingma and Ba, 2015] for 50 epochs with a learning rate of 1e-3, a batch size of 32 and l2 weight decay of 1e-4. Early stopping was performed on the official validation split of the respective datasets, the evaluation of the performance per step was conducted on the official test splits. Although the size of the training dataset increases each step, the effective size, thus the number of gradient steps per epoch, was kept constant at 1000 for the MNIST experiments and 1600 for the FMNIST experiments.

B.6 ALTERNATIVE POSTERIOR SAMPLING METHODS

In addition to Deep Ensembles (DE) used as posterior sampling method in the main paper, we investigate alternative posterior sampling methods. Specifically, we consider the Laplace Approximation (LA) [MacKay, 1992] on the last layer with Kronecker-factored approximate curvature [Ritter et al., 2018] using the implementation of Daxberger et al. [2021]. Furthermore, we consider MC Dropout (MCD) [Gal and Ghahramani, 2016]. As for DE, we sample 10 models with the alternative posterior sampling methods for the MC approximations of posterior expectations. Measures based on a single model (combinations with `(A)`) use the maximum a posteriori (MAP) model for LA, and the model without dropout activated for MCD as the predicting models.

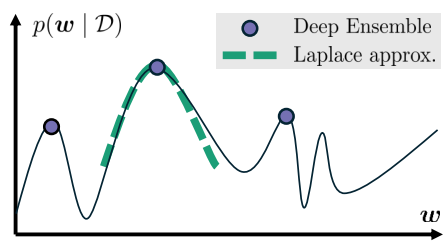


Figure 24: Posterior sampling methods.

There is a distinction between multi- and single-basin posterior sampling techniques [Wilson and Izmailov, 2020], sometimes also referred to as multi- and single-mode approaches [Hoffmann and Elster, 2021]. We refer to them as global and local posterior sampling techniques for simplicity. In this categorization, DE is a global method, while LA and MCD are local methods [Fort et al., 2019], see Fig. 24 for an illustration. We hypothesize that different methods for posterior sampling have a strong impact on which uncertainty measure performs well empirically, especially given whether they are global or local methods.

In the following, we first analyze the characteristics of posterior samples drawn with those different sampling methods. Then we investigate the importance of aligning the uncertainty measure with the predicting model, detecting distributional mismatch and the disentanglement of measures on these additional posterior sampling methods.

B.6.1 Characteristics of Posterior Samples

To better understand the performance of different posterior sampling methods, we examine the characteristics of their sampled models. The results in Fig. 25 show that these methods perform differently across datasets. For the global sampling method DE, the average model consistently outperforms individual sampled models with a lower negative log-likelihood (NLL) and higher accuracy across all datasets. In contrast, for local sampling methods LA and MCD, individual sampled models exhibit higher NLL than both the single model and the average model. Additionally, the accuracy of individual sampled models is lower than that of the single model. Specifically, for MCD, the single model's accuracy is comparable to the average model, while for LA, the single model's accuracy exceeds that of the average model.

We further analyze the predictive uncertainties estimated by different posterior sampling methods using measure `(C2)`, which incorporates posterior samples and is upper-bounded. To ensure comparability between datasets, we normalize the uncertainties by the maximal predictive uncertainty TU `(C2)`, equal to the entropy of the uniform distribution $\log(|\mathcal{Y}|)$. The results in the two right plots in Fig. 25 show that these methods yield similar distributions of uncertainties for CIFAR10 and SVHN. However, for CIFAR100 and TIN, DE exhibits many more datapoints with very low EU and AU.

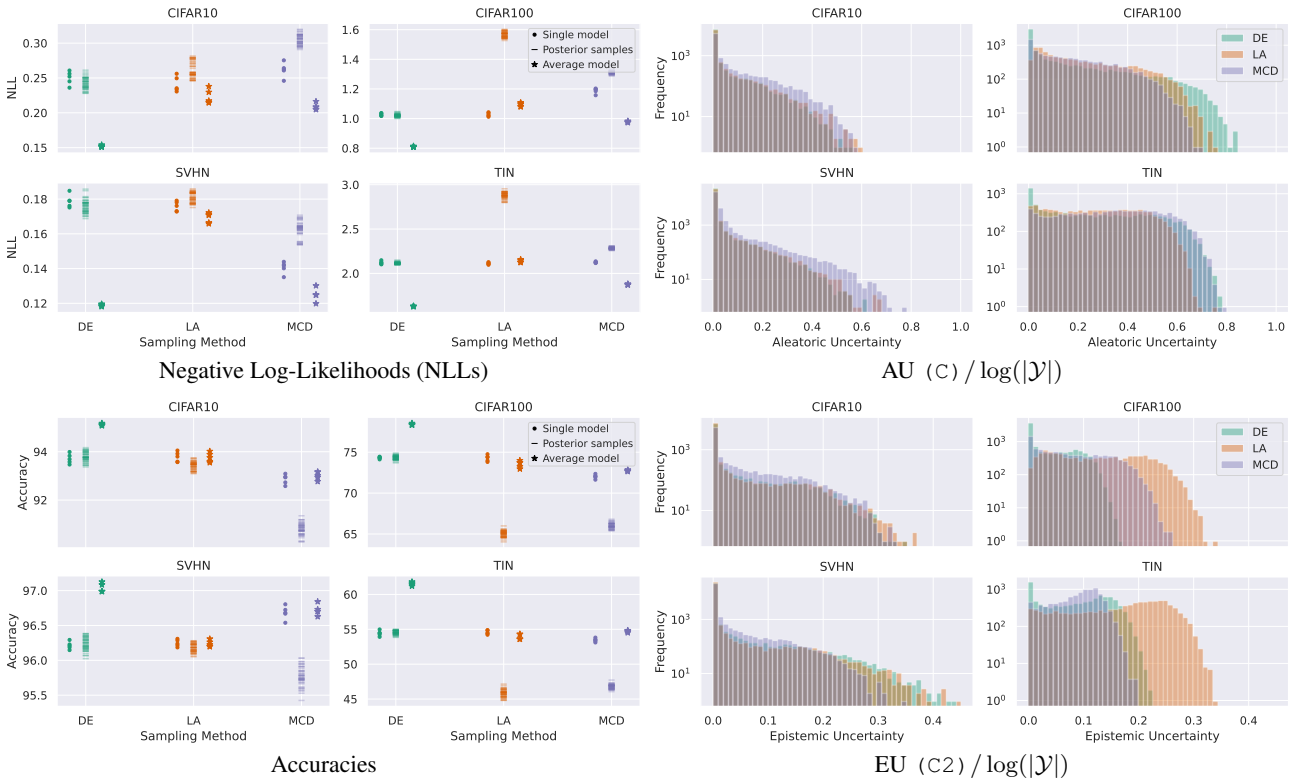


Figure 25: **Comparison of posterior sampling methods.** Results are obtained on the test split of the respective dataset. We compare the NLLs and accuracies for different models obtained through DE, LA and MCD. Similarly, the normalized AU (C) and the normalized EU (C2) are shown per sampling method. All three methods yield similar results for CIFAR10 and SVHN, but differ greatly on CIFAR100 and TIN. Models sampled using LA have higher NLL and lower accuracy. Furthermore, they lead to higher EU and lack predictions with very low AU. Additionally, the average model does not improve over the single model in terms of NLL and accuracy for those two datasets. The results in the two left plots show single models, posterior samples and average models of five runs, those in the two right plots show uncertainties for a single run.

B.6.2 Aligning the Uncertainty Measure with the Predicting Model

Selective Prediction. The results for LA are provided in Tab. 6, for MCD in Tab. 7. In contrast to the global posterior sampling method DE, we do not find that it is not directly important to align the uncertainty measure to the predicting model for LA and MCD. We find that no matter the predicting model, it is best to use measures for a single predicting model (A). This is likely due to the strong dependence of the sampling space on the original model, i.e. the MAP for LA and the model without dropout for MCD. An indicator for this is, that AU (A) is better than AU (B) or (C) for all settings, which was never the case for DE (see Tab. 2).

Misclassification Detection. For misclassification detection, we have similar findings as for selective prediction. The results for LA are shown in Fig. 26, those for MCD in Fig. 27. Again, the measures for (A) perform best, regardless of the predicting model. The reason for this behavior is likely again the same as for misclassification, that the original model, i.e. the MAP for LA and the model without dropout for MCD, has a strong dependence on the sampling space for those posterior sampling methods. In accordance with this perspective, we find that AU (A) is better than AU (B) or (C) for all settings, which was never the case for DE (see Tab. 2).

B.6.3 Detecting Distributional Mismatch

OOD Detection. The results for LA are provided in Fig. 28, the results for MCD in Fig. 29. For MCD, the results are largely comparable to the results for DE in the main paper (see Fig. 3), where TU (B/C2) which is equivalent to AU (B) and TU (B/C3) perform best. The EU measures perform very poor in general, worse than most TU and AU measures. For

Table 6: **Selective prediction under different predicting models for LA.** AUARC for different predicting models, i.e., the single model, the average model and a model according to the posterior, using different predictive uncertainty measures as score. We highlight the two best measures for each setting. For the single predicting model, and the average model, TU (A3) performs best and both TU (A1) and AU (A) second best. For models sampled according to the posterior, TU (A3) performs best, followed by TU (A2). Results are averaged over all datasets with statistics over five runs.

Measures:	TU						AU			EU								Random
Prediction	A1	A2	A3	B/C1	**B/C2**	B/C3	A	**B**	C	A1	A2	A3	B1	B3	C1	C2	C3	Baseline
Single	87.53	87.51	87.59	86.87	87.15	87.18	87.53	87.15	87.03	79.81	83.65	86.72	80.67	85.47	84.22	86.54	85.94	79.72
Model	±0.14	±0.14	±0.15	±0.15	±0.15	±0.16	±0.14	±0.15	±0.14	±0.16	±0.20	±0.18	±0.22	±0.20	±0.21	±0.17	±0.18	±0.13
Average	87.35	87.34	87.44	86.68	86.96	87.00	87.35	86.96	86.83	79.44	83.47	86.60	80.37	85.28	84.01	86.37	85.76	79.36
Model	±0.17	±0.17	±0.18	±0.18	±0.18	±0.19	±0.17	±0.18	±0.17	±0.18	±0.20	±0.19	±0.22	±0.22	±0.22	±0.19	±0.21	±0.16
Acc. to	85.10	85.15	85.26	84.67	84.86	84.99	85.10	84.86	84.66	76.63	81.63	84.69	78.17	83.39	82.16	84.53	83.91	76.56
Posterior	±0.12	±0.11	±0.13	±0.12	±0.11	±0.13	±0.12	±0.11	±0.11	±0.14	±0.12	±0.14	±0.17	±0.16	±0.17	±0.14	±0.15	±0.08

Table 7: **Selective prediction under different predicting models for MCD.** AUARC for different predicting models, i.e., the single model, the average model and a model according to the posterior, using different predictive uncertainty measures as score. We highlight the two best measures for each setting. Regardless of the predicting model, TU (A2) performs best, followed by TU (A3). Results are averaged over all datasets with statistics over five runs.

Measures:	TU						AU			EU								Random
Prediction	A1	A2	A3	B/C1	**B/C2**	B/C3	A	**B**	C	A1	A2	A3	B1	B3	C1	C2	C3	Baseline
Single	86.97	87.15	87.11	85.59	86.56	86.42	86.97	86.56	86.53	78.81	84.57	85.86	81.53	84.30	82.83	85.40	84.76	78.75
Model	±0.07	±0.07	±0.07	±0.05	±0.06	±0.06	±0.07	±0.06	±0.06	±0.08	±0.08	±0.09	±0.05	±0.09	±0.06	±0.07	±0.08	±0.09
Average	87.24	87.42	87.35	86.04	86.97	86.81	87.24	86.97	86.94	79.38	84.90	86.11	82.04	84.68	83.30	85.77	85.13	79.30
Model	±0.06	±0.05	±0.05	±0.06	±0.04	±0.04	±0.06	±0.04	±0.05	±0.07	±0.06	±0.05	±0.08	±0.04	±0.07	±0.05	±0.04	±0.07
Acc. to	84.93	85.22	85.21	84.06	84.88	84.81	84.93	84.88	84.78	76.47	83.03	84.29	80.01	82.94	81.42	84.00	83.39	76.36
Posterior	±0.07	±0.07	±0.08	±0.08	±0.05	±0.05	±0.07	±0.05	±0.05	±0.14	±0.11	±0.09	±0.09	±0.06	±0.09	±0.07	±0.07	±0.12

LA, we observe that all measures for TU and AU perform nearly identical and much better than EU meaures. In sum, we surprisingly find that EU measures never outperform the best TU and AU measures for all posterior sampling methods.

Distribution Shift Detection. The results for distribution shift detection for LA are given in Fig. 30, the results for MCD in Fig. 31. For LA, all TU and AU measures perform nearly identically, outperforming all EU measures in all severities. For MCD, we find that except for the highest severity, the best TU, AU, and EU measures perform roughly on par.

B.6.4 Disentanglement of Measures

The results for LA are provided in Fig. 32 and Fig. 33, the results for MCD in Fig. 34 and Fig. 35. For MCD, the results are largely consistent with the findings for DE in the main paper for the OOD data. For ID data, we find even more correlation between measures, not only within TU, AU, and EU measures, but also between those. For LA, we find that the TU and AU measures are all highly correlated with the OOD data compared to other posterior sampling methods. This explains the very similar performance of all TU and AU measures on OOD detection with LA as a posterior sampling method. For the ID data, the results for LA are very similar to the results with DE in the main paper.

B.6.5 Conclusions

Overall, the behavior of the global posterior sampling method DE discussed in the main paper and the local posterior sampling methods LA and MCD discussed here differ in quite some dimensions.

First, we find that for LA the sampled models often perform poorly, especially for SVHN and TIN with more classes. Here, the average model does not actually improve over the single models, as is the case for DE and MCD. This might be a limitation of the particular LA used in this work.

Second, there does not appear to be a benefit in aligning the uncertainty measure to the predicting model for neither LA nor

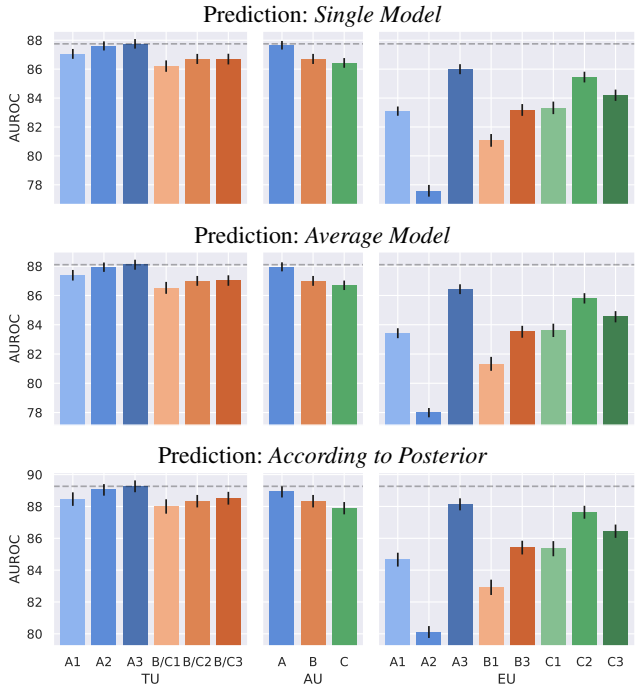


Figure 26: **Misclassification detection under different predicting models for LA.** AUROC for distinguishing correct from incorrect predictions under different predicting models, using the different proposed measures of uncertainty as score. For all settings, TU (A3) performs best. AUROCs are averages over all datasets. Statistics over five runs.

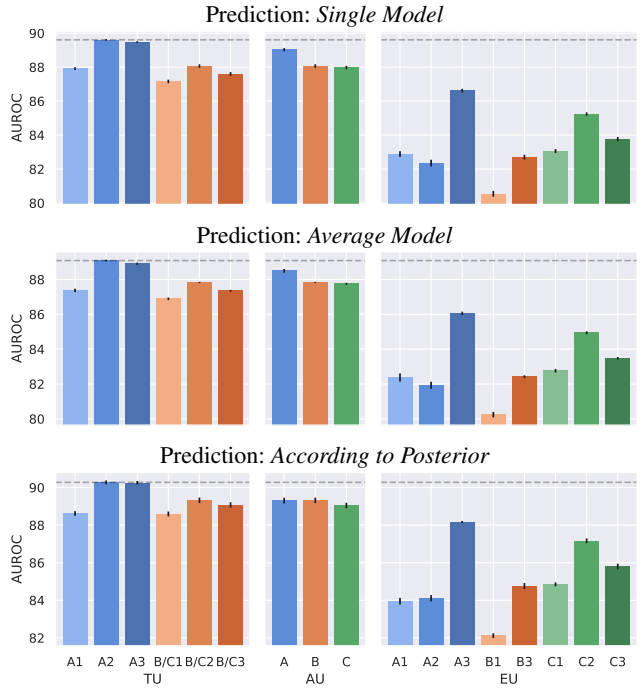


Figure 27: **Misclassification detection under different predicting models for MCD.** AUROC for distinguishing correct from incorrect predictions under different predicting models, using the different proposed measures of uncertainty as score. For all settings, TU (A2) performs best. AUROCs are averages over all datasets. Statistics over five runs.

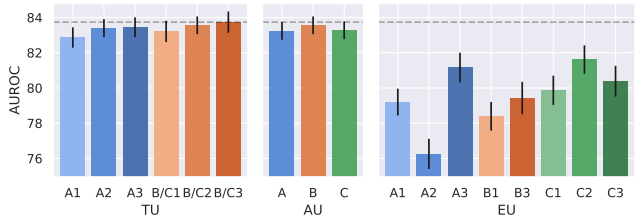


Figure 28: **OOD detection for LA.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. TU (B/C2) and TU (B/C3) perform best. AUROCs are averaged over all ID / OOD combinations. Statistics over five runs.

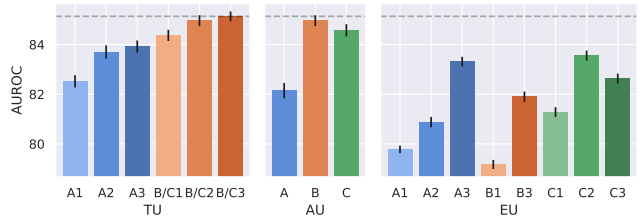


Figure 29: **OOD detection for MCD.** AUROC for distinguishing between ID and OOD datapoints using the different proposed measures of uncertainty as score. TU (B/C2) and TU (B/C3) perform best. AUROCs are averaged over all ID / OOD combinations. Statistics over five runs.

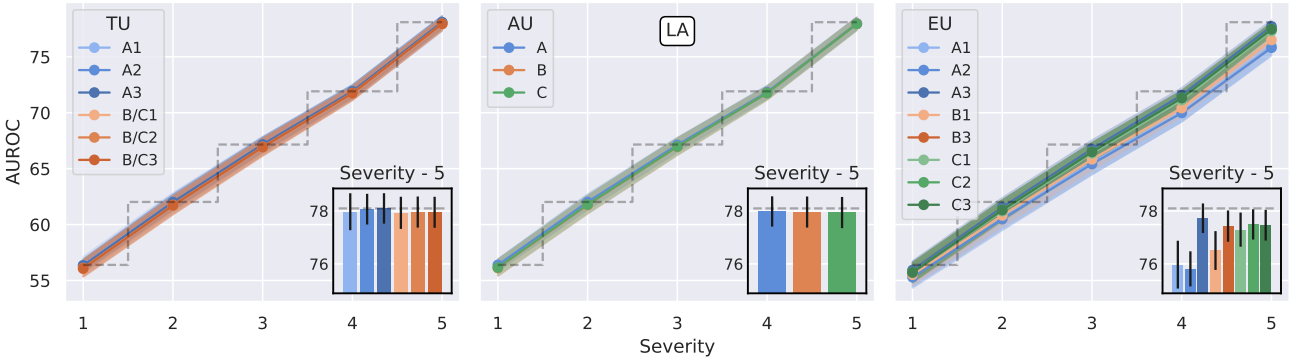


Figure 30: **Distribution shift detection on CIFAR10-C for LA.** AUROC for distinguishing between clean and corrupted test datapoints, using the different proposed measures of uncertainty as score. Black dashed line shows the maximum AUROC over all measures per severity. Insets shows detailed results for the highest severity. Statistics over five runs.

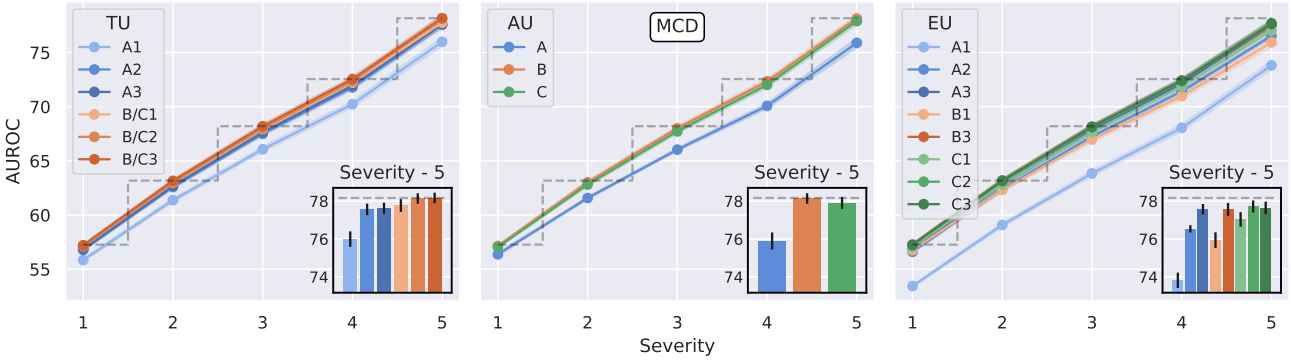


Figure 31: **Distribution shift detection on CIFAR10-C for MCD.** AUROC for distinguishing between clean and corrupted test datapoints, using the different proposed measures of uncertainty as score. Black dashed line shows the maximum AUROC over all measures per severity. Insets shows detailed results for the highest severity. Statistics over five runs.

MCD. We hypothesize that this is due to the fact that local methods probe local sensitivity around the original model, i.e. the MAP for LA and the model without dropout for MCD, leading to uncertainty measures based on this original model performing very well. For the global method DE, there is no such original model that dominates the sampling space.

Third, for detecting distributional mismatch, MCD largely leads to the same results as DE. However, for LA we find that all TU and AU measures perform on very similar levels. This is also evident when we look at the rank correlations of measures. Here we find a strong correlation between all the TU and AU measures on the OOD data for LA. This is probably because the models sampled with LA do not strongly disagree with the original model on the OOD data. Other than that, we find that the results match those presented for DE in the main paper.

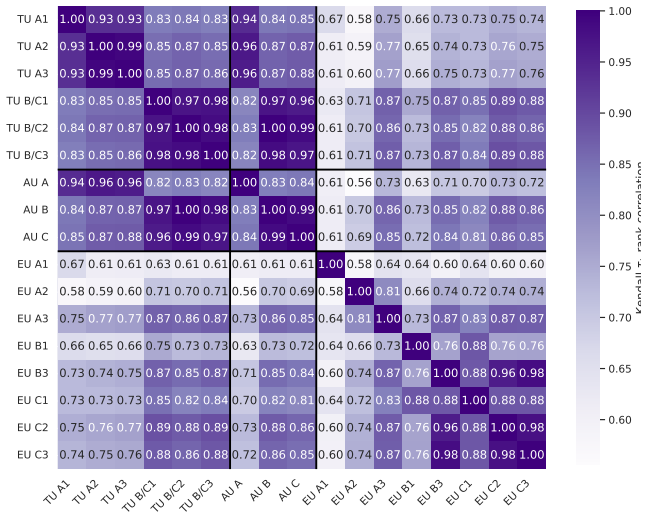


Figure 32: **Correlation of uncertainty measures on ID dataset (CIFAR10) for LA.** High rank correlation blocks exist.

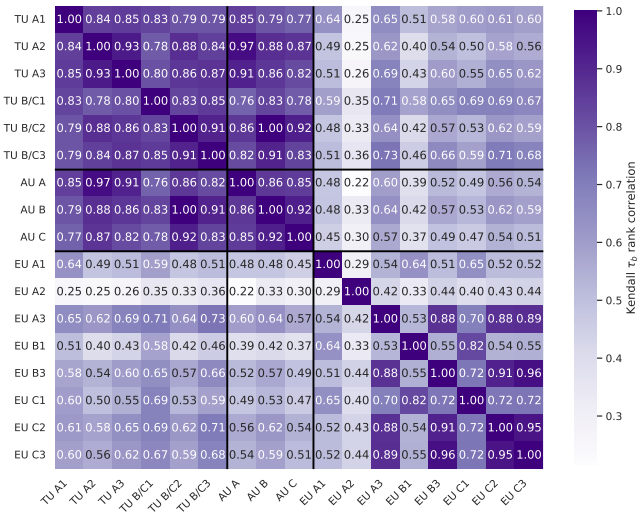


Figure 33: **Correlation of uncertainty measures on OOD dataset (LSUN) for LA.** All TU and AU measures are highly correlated.

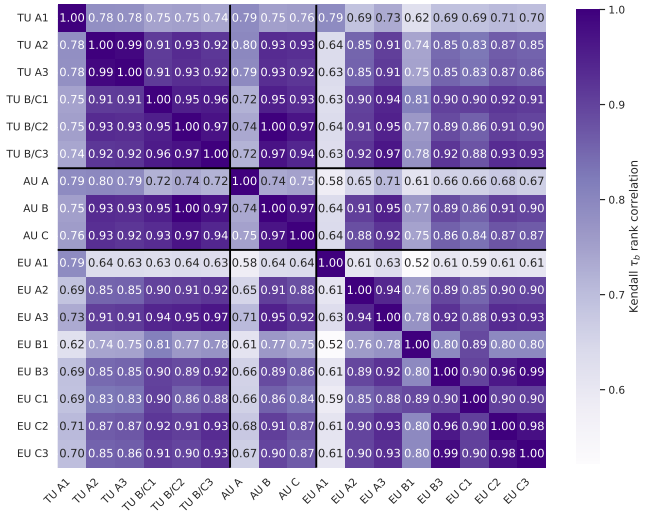


Figure 34: **Correlation of uncertainty measures on ID dataset (CIFAR10) for MCD.** High rank correlation blocks exist.

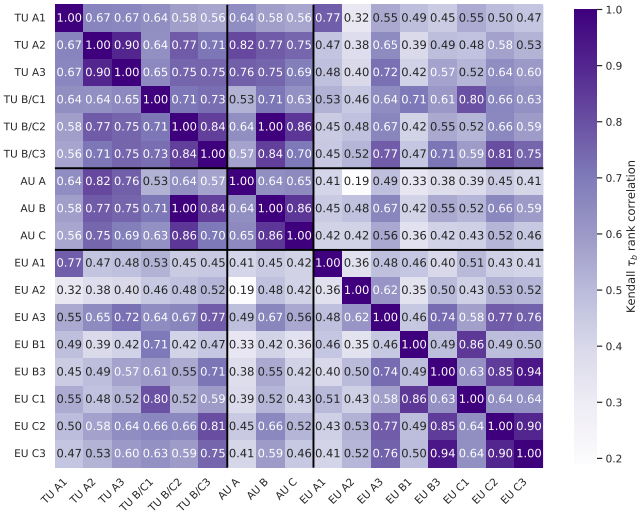


Figure 35: **Correlation of uncertainty measures on OOD dataset (LSUN) for MCD.** Rank correlations are very low overall.