

---

# Backward-Compatible Prediction Updates: A Probabilistic Approach

---

Frederik Träuble<sup>1\*†</sup>

Julius von Kügelgen<sup>1,3\*</sup>

Matthäus Kleindessner<sup>2</sup>

Francesco Locatello<sup>2</sup>

Bernhard Schölkopf<sup>2</sup>

Peter Gehler<sup>2†</sup>

<sup>1</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany

<sup>2</sup>Amazon Tübingen, Germany

<sup>3</sup>Department of Engineering, University of Cambridge, United Kingdom

## Abstract

When machine learning systems meet real world applications, accuracy is only one of several requirements. In this paper, we assay a complementary perspective originating from the increasing availability of pre-trained and regularly improving state-of-the-art models. While new improved models develop at a fast pace, downstream tasks vary more slowly or stay constant. Assume that we have a large unlabelled data set for which we want to maintain accurate predictions. Whenever a new and presumably better ML models becomes available, we encounter two problems: (i) given a limited budget, which data points should be re-evaluated using the new model?; and (ii) if the new predictions differ from the current ones, should we update? Problem (i) is about compute cost, which matters for very large data sets and models. Problem (ii) is about maintaining consistency of the predictions, which can be highly relevant for downstream applications; our demand is to avoid negative flips, i.e., changing correct to incorrect predictions. In this paper, we formalize the Prediction Update Problem and present an efficient probabilistic approach as answer to the above questions. In extensive experiments on standard classification benchmark data sets, we show that our method outperforms alternative strategies along key metrics for backward-compatible prediction updates.

## 1 Introduction

The machine learning (ML) community develops new models at a fast pace: for example, just in the past year, the state-of-the-art on ImageNet has changed at least five times [9, 10, 31, 50, 51]. As reproducibility has increasingly been scrutinized [33, 34, 42], it is now common practice to release pre-trained models upon publication. In this work we take the perspective of an owner of an unlabelled data set who is interested in keeping the best possible predictions at all times. When a new pre-trained model is released, we face what we refer to as the *Prediction Update Problem*: (i) decide which points in the data set to re-evaluate with the new model, and (ii) integrate the new, possibly contradicting, predictions. For this task, we postulate the following three desiderata:

1. The prediction updates should improve overall accuracy.
2. The prediction updates should avoid introducing new errors.
3. The prediction updates should be as cheap as possible since the target data set could be huge.

---

\*Work done while FT and JvK were interning at Amazon.

†Correspondence to: frederik.traeuble@tuebingen.mpg.de and pgehler@amazon.com

We consider the setting in which the target data set for which we wish to maintain predictions is fully unlabelled (i.e., the ground-truth labels are unknown) and may come from a different distribution than the one on which models have been pre-trained, but with overlap in the label space. This is a transductive or semi-supervised problem, but, due to computational constraints, we avoid any model fitting or fine-tuning and rely solely on the predictions of the pre-trained models that are released over time. Typically, these models exhibit increased performance on their labelled training domain (e.g., the ImageNet validation or test set) as evidence for being good candidates for re-evaluation.

Clearly, one goal of updating the predictions stored for the target data set is to improve overall performance, e.g., top-k accuracy for classification. At the same time, the stored predictions may form an intermediate step in a larger ML pipeline or are accessible to users. This is the reason for our second desideratum: we would like to be *backward-compatible*, i.e., new predictions should not flip previously correct predictions (*negative flips*). Finally, we aim to reduce computational cost during inference and to avoid evaluating the entire data set which may be prohibitive in practice and unnecessary if we are already somewhat certain about a prediction.

In this paper, we motivate and formalize the *Prediction Update Problem* and describe its relation to various relevant research areas like ensemble learning, domain adaption, active learning, and others. We propose a probabilistic approach that maintains a posterior distribution over the unknown true labels by combining all previous model re-evaluations. Based on these uncertainty estimates, we devise an efficient *selection strategy* which only chooses those examples with highest posterior label entropy for re-evaluation in order to reduce computational cost. Furthermore, we consider different prediction-update strategies to decide whether to change the stored predictions, taking asymmetric costs for negative and positive flips into account. Using the task of image classification as a case study, we perform extensive experiments on common benchmarks (ImageNet, CIFAR10, and ObjectNet) and demonstrate that our approach achieves competitive accuracy and introduces much fewer negative flips across a range of computational budgets, thus showing that our three desiderata are not necessarily at odds.

**Contributions** We highlight the following contributions:

- We introduce the *Prediction Update Problem* which addresses some common, but previously unaddressed challenges faced in real world ML systems (§ 2).
- We propose a probabilistic, model-agnostic approach for the *Prediction Update Problem*, based on Bayesian belief estimates of the true label combined with an efficient selection and different prediction-update strategies (§ 3).
- We contextualise this understudied problem setting as well as our method with related work (§ 4 & § 5) and discuss several extensions and limitations (§ 4).
- We demonstrate that our approach successfully outperforms alternative approaches and accomplishes all our desiderata in experiments across multiple common benchmark datasets (CIFAR-10, ImageNet, and ObjectNet) and practically relevant scenarios (§ 6).<sup>3</sup>

## 1.1 Backward-Compatible Prediction Systems

In real world ML applications, empirical performance is only one of several requirements. When humans interact with automatic predictions, they will start to build mental models of how these models operate and whether and when their predictions can be trusted. This is described as Human-AI teams by [1] who argue to “make the human factor a first-class consideration of AI updates”.

An example from [1] is autopilot functionality in cars for which drivers will build expectations in which driving situations the autopilot is safe to engage. It is important not to violate these assumptions when updating the models over the air. AI assisted medical decision processes are another example of a high stake application where medical professionals need to understand when systems can be trusted.

Consider the example of automatically tagging images in a user’s photo collection. Those tags are used for example in photo search. As models progress, the overall accuracy on all uploaded images may increase, but for any single user the experience can deteriorate if previously correct searches now show wrong results. Even worse, if errors fluctuate over the user’s photo collection as the result of prediction-updates, the user’s trust will be eroded. This “cost” is asymmetric and the negative experience may outweigh the benefit of better predictions on other images.

---

<sup>3</sup>All software and assets we use are open source and under MIT, Apache or Creative Commons Licenses.

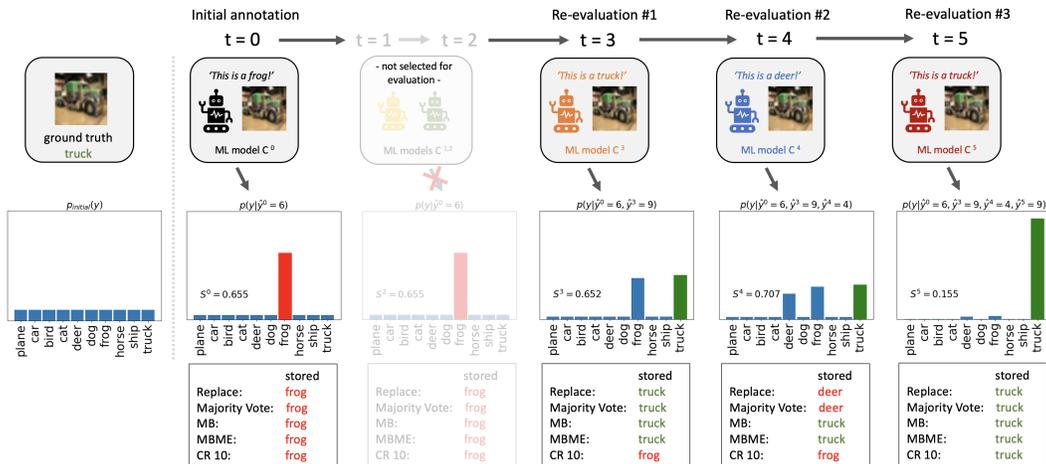


Figure 1: **Overview of our proposed Bayesian approach to the Prediction Update Problem.** Starting from a uniform prior, we maintain a posterior distribution  $p(y = k | \hat{y}^{0:t})$  (middle) over the unknown true label  $y$  of an unlabelled sample which takes the predictions  $\hat{y}^{0:t}$  from new ML models  $C^t$  (top) arriving over time  $t = 0, \dots, T$  into account. Given a limited compute budget  $B^t$ , we re-evaluate those samples with highest posterior label entropy  $S^t$  at each time step, e.g., the example shown is first selected in time step  $t = 3$  after the initial annotation at  $t = 0$ . We then consider different strategies for deciding whether to update the stored prediction (bottom) based on our changed beliefs. Note that the non-probabilistic baselines “Replace” (always update to last prediction) and “Majority Vote” (resolve ties by using the last prediction) incorrectly update the stored prediction from “truck” to “deer” in step  $t = 4$ . Our strategies (MB, MBME, CR-10) which rely on the estimated label posterior, on the other hand, avoid such a *negative flip*, which is one of our key goals.

In contrast to carefully curated and labelled ML benchmarks, many real-world data sets are magnitudes larger (up to billions of samples) and entirely unlabelled. Having no feedback which predictions are correct is a common scenario: consider any type of private data such as health data, photo collections, or personal information. Because the data is private, we can neither train on it, nor collect feedback, nor observe the effect of predictions. On the other hand, such data is valuable to an individual: she has an interest to keep it up to date with the best possible predictions. Since it is of little consolation to her if an update of the model improves predictions on average but on her data it gets worse, the update costs are asymmetric. Service providers often rely on models pre-trained on a different data set, and the desire to be backward-compatible arises naturally in this setting [1, 39, 43, 53].

This is an understudied problem where progress will have large impact. ML systems are becoming pervasive, and their accuracy will continue to increase. Being able to seamlessly transfer them to existing data will be crucial for real-world ML systems.

## 2 The Prediction Updates Problem Setting

**Target Data Set** We are given a large, unlabelled target data set  $\mathcal{D}^{\text{targ}} = \{\mathbf{x}_n\}_{n=1}^N$  comprising  $N$  independent and identically distributed (i.i.d.) observations  $\mathbf{x}_n \in \mathcal{X} \subseteq \mathbb{R}^d$  drawn from a target distribution  $\mathbb{P}_{\mathbf{X}}^{\text{targ}}$ . The ground-truth labels  $y_n \in \mathcal{Y} = \{1, \dots, K\}$  distributed according to  $\mathbb{P}_{\mathbf{Y}|\mathbf{X}}^{\text{targ}}$  are *not observed*. Note that we are particularly interested in a scenario where  $N$  may be extremely large.

**Models** Over time  $t = 0, 1, \dots, T$  we successively gain access to classifiers  $C^0, C^1, \dots, C^T : \mathcal{X} \rightarrow \mathcal{Y}$  which have been trained on a labelled data set  $\mathcal{D}^{\text{src}}$  from a potentially different source distribution  $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{src}}$  over  $\mathcal{X} \times \mathcal{Y}$ . For simplicity, we assume that the observation space  $\mathcal{X}$  and label space  $\mathcal{Y}$  are shared. We consider both the standard scenario where the models  $\{C^t\}_{t=0}^T$  are trained on a labelled set from the same domain ( $\mathbb{P}^{\text{src}} = \mathbb{P}^{\text{targ}}$ ); and the transfer scenario where we deploy a model trained on a labelled ML benchmark to a different data set ( $\mathbb{P}^{\text{src}} \neq \mathbb{P}^{\text{targ}}$ ). We assume that  $\{C^t\}_{t=0}^T$  are improving in performance on the training data set. Therefore, denoting by  $A_t$  the estimated accuracy of  $C^t$  on  $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{src}}$ , we have  $A^t \leq A^{t+1} \forall t$ . As motivating example, consider an object recognition task in the wild and let  $C_t$  be the winning entry of the ImageNet competition in year  $t$ .

**Labelling** To relate the source and target distributions and to justify applying  $\{C^t\}_{t=0}^T$  to our target data set  $\mathcal{D}^{\text{targ}}$ , we make the commonly used covariate shift assumption  $\mathbb{P}_{Y|\mathbf{X}}^{\text{targ}} = \mathbb{P}_{Y|\mathbf{X}}^{\text{src}}$ , i.e. the conditional label distribution is shared across source and target distributions [40, 45].<sup>4</sup> We denote the *predicted* label by  $C^t$  for  $\mathbf{x}_n$  by  $\hat{y}_n^t = C^t(\mathbf{x}_n)$  and the *stored* prediction for  $\mathbf{x}_n$  after time step  $t$  by  $l_n^t$ . The target data set is then initially fully labelled by  $C^0$ , i.e.,  $l_n^0 := \hat{y}_n^0$ .

**Objective** As new classifiers  $\{C^t\}_{t \geq 1}$  become available, our main objective is to maintain the best estimates  $\{l_n^t\}_{n=1}^N$  on our target data set at all times and improve overall accuracy, while, at the same time, maintaining backward compatibility by minimising the number of *negative flips*, i.e., the number of previously correctly stored predictions that are incorrectly changed. The key challenge is that no ground truth labels for our target data set are available, so that we have no feedback on which predictions are correct and which are wrong. For each test sample  $\mathbf{x}_n$  and each time step  $t \geq 1$ , we thus need to decide whether or not to update the previously stored prediction  $l_n^{t-1}$  based on the current and previous model predictions  $\hat{y}_n^t$  and  $\hat{y}_n^{0:t-1}$ , respectively.

**Limited Evaluation Budget** Re-evaluating all samples (so-called *backfilling*) can be very costly and requires significant resources. Since we consider  $N$  to be very large, we also consider a limited budget of at most  $B^t \leq N$  sample re-evaluations for step  $t$ . We thus additionally need to decide how to allocate this budget and select a subset of samples to be re-evaluated by  $C^t$  at every step.

### 3 Our Method

Having specified the setting, we next describe our proposed method for the Prediction Update Problem. We start by providing a Bayesian approach for maintaining and updating our beliefs about the unknown true labels as new predictions become available (§ 3.1), followed by describing strategies for selecting candidate samples for re-evaluation (§ 3.2) and for updating the stored predictions based on our changed beliefs (§ 3.3). Our framework is summarised in Figure 1.

#### 3.1 Bayesian Approach

Since the true labels  $\{y_n\}_{n=1}^N$  are unknown to us, we treat them as random quantities over which we maintain uncertainty estimates. We then perform Bayesian reasoning to update our beliefs as new evidence in the form of predictions  $\hat{y}_n^t$  from newly-available classifiers  $C^t$  arrives over time  $t = 1, \dots, T$ . In standard Bayesian notation, the true labels  $y_n$  thus take the role of unknown parameters  $\theta$  and the predictions  $\hat{y}_n^t$  of data  $x$ . Since  $\mathcal{D}^{\text{targ}}$  is sampled i.i.d., we reason about each label  $y_n$  independently of the others, i.e., the following is the same for all  $n$ .

**Prior** Lacking label information on the target data set, we choose a uniform prior over  $\mathcal{Y}$  for all  $y_n$ , i.e.,  $p(y_n = k) = 1/K, \forall k \in \mathcal{Y}$ . If (estimates of) the class probabilities on  $\mathcal{D}^{\text{targ}}$  are available, we may instead use these as a more informative prior.

**Likelihood** Next, we need to specify a likelihood function  $p(\hat{y}_n^{0:T} | y_n = k)$  for the observed model predictions  $\hat{y}_n^{0:T}$  given a value  $k$  of the true label  $y_n$ . We make the following simplifying assumption.

**Assumption 1** (Conditionally independent classifiers). *The different classifiers’ predictions  $\hat{y}_n^{0:T}$  are conditionally independent given the true label  $y_n$ , i.e., the likelihood factorises as*

$$p(\hat{y}_n^{0:T} | y_n = k) = \prod_{t=0}^T p(\hat{y}_n^t | y_n = k). \quad (1)$$

In a standard Bayesian setting, this corresponds to the assumption of conditionally independent observations given the parameters; we refer to § 4 for further discussion. The main advantage of Assumption 1 is that the factors  $p(\hat{y}_n^t | y_n = k)$  on the RHS of (1) have a natural interpretation: these are the (normalised) confusion matrices  $\pi^t$  of the classifiers  $C^t$ , i.e., we denote by

$$\pi^t(i, k) := p(\hat{y}^t = i | y = k),$$

the probability that  $C^t$  predicts class  $i$  given that the true label is  $k$ , which is the same for all  $n$ ; see below and § 4 for more details on how we estimate  $\pi^t$  in practice.

<sup>4</sup>In the context of image classification, this means that images of the same object under different environmental conditions (i.e., different  $\mathbb{P}_{\mathbf{X}}$ ) will always share the same label (i.e., same  $\mathbb{P}_{Y|\mathbf{X}}$ ). Note that such covariate shift may also lead to changes in  $\mathbb{P}_Y$  (label/target shift) and/or  $\mathbb{P}_{\mathbf{X}|Y}$  (conditional shift) [44, 55].

**Posterior** At every time step  $t \geq 0$ , we can then compute our posterior belief about the true label  $y_n$  given model predictions  $\hat{y}_n^{0:t}$  according to Bayes rule,

$$p(y_n = k | \hat{y}_n^{0:t}) = \frac{\pi^t(\hat{y}_n^t, k) p(y_n = k | \hat{y}_n^{0:t-1})}{\sum_{i \in \mathcal{Y}} \pi^t(\hat{y}_n^t, i) p(y_n = i | \hat{y}_n^{0:t-1})} \quad (2)$$

where we have used Assumption 1 to write  $p(\hat{y}_n^t | y_n = k, \hat{y}_n^{0:t-1}) = p(\hat{y}_n^t | y_n = k) = \pi^t(\hat{y}_n^t, k)$ . The posterior at step  $t - 1$  acts as prior for step  $t$ , so we do not have to store all previous predictions.

**Estimating Confusion Matrices** In practice,  $\pi^t$  are generally not known and we instead use their (maximum likelihood) estimates  $\hat{\pi}^t$  from the source distribution. If the number of classes  $K$  is large compared to the amount of labelled source data,<sup>5</sup> we only estimate the diagonal elements  $\hat{\pi}_{kk}^t$  (i.e., the class-specific accuracies) and set the  $K(K - 1)$  off-diagonal elements to be constant,

$$\hat{\pi}^t(i, k) = \frac{1 - \hat{\pi}^t(k, k)}{K - 1} \quad \forall i \neq k,$$

so that  $\sum_{i=1}^K \hat{\pi}^t(i, k) = 1 \quad \forall k \in \mathcal{Y}$ . We refer to § 4 for further discussion on the estimation of  $\pi^t$ .

### 3.2 Selecting Candidates for Re-evaluation

Given the label posteriors computed according to (2), we compute the Shannon entropies [38]

$$S_n^t = - \sum_{k \in \mathcal{Y}} p(y_n = k | \hat{y}_n^{0:t}) \log p(y_n = k | \hat{y}_n^{0:t}),$$

which provide a simple measure of uncertainty in the true label  $y_n$  after step  $t$ . We then select and re-evaluate the  $B^t$  samples with highest posterior label entropy  $S_n^t$  to update our beliefs.

### 3.3 Prediction-Update Strategies

Finally, we need a strategy for deciding whether and how to update the previously stored prediction  $l_n^{t-1}$  based on our new beliefs. We consider three such prediction-update strategies.

**MaxBelief (MB)** The simplest approach is to always update based on the maximum a posteriori belief, i.e.,  $l_n^t := \hat{l}_n^t = \operatorname{argmax}_{k \in \mathcal{Y}} p(y_n = k | \hat{y}_n^{0:t})$ . We refer to this strategy as MaxBelief (MB).

**MaxBeliefMinEntropy (MBME)** A slightly more sophisticated approach is to also take the change in posterior entropy into account and only update when it has decreased:

$$l_n^t := \begin{cases} \hat{l}_n^t & \text{if } S_n^t < S_n^{t-1}, \\ l_n^{t-1} & \text{otherwise.} \end{cases}$$

We refer to this strategy as MaxBeliefMinEntropy (MBME).

**CostRatio (CR)** So far, we have not taken the assumed larger penalty for negative flips into account. We therefore now develop a third approach based on asymmetric flip costs. We denote the cost of a negative flip (NF) by  $c^{\text{NF}} > 0$  and that of a positive flip (PF) by  $c^{\text{PF}} < 0$ .

We need to decide whether to update the previously stored prediction  $l_n^{t-1}$  based on our updated beliefs  $p(y_n = k | \hat{y}_n^{0:t})$ . Denote the MAP label estimate after step  $t$  by  $\hat{l}_n^t = \operatorname{argmax}_{k \in \mathcal{Y}} p(y_n = k | \hat{y}_n^{0:t})$ . If  $\hat{l}_n^t = l_n^{t-1}$  there is no reason to change the stored prediction. Suppose that  $\hat{l}_n^t \neq l_n^{t-1}$ . We then need to reason about the (estimated) positive and negative flip probabilities when changing the stored prediction from  $l_n^{t-1}$  to  $\hat{l}_n^t$ . A positive flip (PF) occurs if  $\hat{l}_n^t$  is the correct label (and hence  $l_n^{t-1}$  is not), and, vice versa, a negative flip occurs if  $l_n^{t-1}$  is correct (and hence  $\hat{l}_n^t$  is not):

$$\hat{p}_n^{\text{PF}}(l_n^{t-1} \rightarrow \hat{l}_n^t) = p(y_n = \hat{l}_n^t | \hat{y}_n^{0:t}), \quad \hat{p}_n^{\text{NF}}(l_n^{t-1} \rightarrow \hat{l}_n^t) = p(y_n = l_n^{t-1} | \hat{y}_n^{0:t}).$$

If neither  $l_n^{t-1}$  nor  $\hat{l}_n^t$  are the correct label, the flip is inconsequential which we assume incurs zero cost. The estimated cost of changing the stored prediction from  $l_n^{t-1}$  to  $\hat{l}_n^t$  is thus:

$$\hat{c}(l_n^{t-1} \rightarrow \hat{l}_n^t) = c^{\text{NF}} \hat{p}_n^{\text{NF}}(l_n^{t-1} \rightarrow \hat{l}_n^t) + c^{\text{PF}} \hat{p}_n^{\text{PF}}(l_n^{t-1} \rightarrow \hat{l}_n^t).$$

<sup>5</sup>For example, on ImageNet we have  $K = 1000$  which would require estimating 1 million parameters.

We only want to change the prediction if  $\hat{c}(l_n^{t-1} \rightarrow \hat{l}_n^t) < 0$ , i.e.,

$$\frac{\hat{p}_n^{\text{PF}}(l_n^{t-1} \rightarrow \hat{l}_n^t)}{\hat{p}_n^{\text{NF}}(l_n^{t-1} \rightarrow \hat{l}_n^t)} = \frac{p(y_n = \hat{l}_n^t | \hat{y}_n^{0:t})}{p(y_n = l_n^{t-1} | \hat{y}_n^{0:t})} > -\frac{c^{\text{NF}}}{c^{\text{PF}}} \quad (3)$$

leading to the following update rule:

$$l_n^t := \begin{cases} \hat{l}_n^t, & \text{if } \hat{l}_n^t = l_n^{t-1}, \\ \hat{l}_n^t, & \text{if } \hat{l}_n^t \neq l_n^{t-1} \wedge \hat{c}(l_n^{t-1} \rightarrow \hat{l}_n^t) < 0, \\ l_n^{t-1} & \text{otherwise.} \end{cases}$$

Note that (3) has an intuitive interpretation: we only want to update the currently stored prediction (thus potentially risking a negative flip) if our belief in a different label is larger than that in the current one by a factor exceeding  $|c^{\text{NF}}/c^{\text{PF}}|$ . We therefore refer to this strategy as CostRatio (CR).

## 4 Discussion: Extensions and Limitations

We discuss current limitations of our method and propose extensions to address them in future work.

**Soft vs. Hard Labels** Our approach presented in § 3 assumes deterministic classifiers which output hard labels, i.e., only the most likely class. This allows for maximum flexibility and a wide range of classifier models that can be used in conjunction with this method. However, our Bayesian framework can easily be adapted to also allow for probabilistic classifiers which output soft labels, i.e., vectors of class probabilities. Since deep neural networks are known to have unreliable uncertainty estimates [14, 25, 26, 47], we deliberately choose to work with hard labels. If, however, well-calibrated probabilistic classifiers are available (and can be scaled to huge data sets), taking this additional information into account will likely lead to more accurate posterior estimates and thus better performance.

**Assumption of Conditionally-Independent Classifiers** Since the models  $\{C^t\}$  are typically trained and developed on the same data and may even build on insights from prior architectures, our assumption of conditionally independent predictions on  $\mathcal{D}^{\text{targ}}$  does likely not hold exactly in practice. It should therefore rather be understood as an approximation that enables tractable posterior inference. Our experiments (§ 6) suggest that it is a useful approximation that yields competitive performance. Properly incorporating estimated model correlations may yield further improvements. We refer the interested reader to Appendix A.1 for a more detailed discussion.

**Confusion Matrix Estimates** Unless labelled data from  $\mathbb{P}^{\text{targ}}$  is available, the confusion matrices  $\{\pi^t\}$  need to be estimated from  $\mathbb{P}^{\text{src}}$ . This is only an approximation because they may change as a result of  $\mathbb{P}_{\mathbf{X}}^{\text{src}} \neq \mathbb{P}_{\mathbf{X}}^{\text{targ}}$ , and taking such shifts into account could yield more accurate posterior estimates. For this, one may use ideas from the field of *unsupervised domain adaptation* [12, 28, 45]. One could use an importance-weighting approach [40] to give more weight to points which are representative of  $\mathbb{P}_{\mathbf{X}}^{\text{targ}}$  when estimating  $\pi^t$  from  $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{src}}$ . As an example, in further experiments in the supplement we studied estimating the off-diagonal elements using Laplace smoothing [13, 35].

**Other Selection Strategies** Consider an ambiguous image that could be either a zucchini or a cucumber [4]. Such a sample would have large label entropy and could thus potentially be selected for re-evaluation again and again. To overcome this hurdle, one could decompose label uncertainty into epistemic (reducible) and aleatoric (irreducible) uncertainty [8, 19] and only re-evaluate samples with high aleatoric uncertainty, i.e., those with high expected information gain [24]. Such considerations also play a role in the field of *active learning* [37, 54].

**Growing Data Set Size** Our method is not constrained to fixed dataset sizes and can accommodate for the addition of new data. New samples can be added at any time using a uniform prior over labels. Given their high initial entropy, they would then be naturally selected for (re-)evaluation first.

**Adaptive Budgets** Currently, we consider a fixed local budget of  $B^t$  re-evaluations at every time step. A possible extension would be to allow for a global budget of  $B^{\text{total}}$  evaluations spread over all time-steps, i.e., to devise a strategy for deciding whether to (a) keep re-evaluating or (b) save budget for the next better model, potentially using techniques from reinforcement learning [46].

**On the Cost of “Neutral” Flips** For simplicity, we have assumed that “neutral” flips (i.e., changing a label estimate from an incorrect to a different incorrect one) bear no cost. However, as motivated in § 1, it is well conceivable that even such neutral flips have a cost due to the potential to disrupt downstream robustness. If this is the case, it can easily be incorporated into our CR update strategy.

## 5 Related Work

Besides the aforementioned connections, our problem setting bears resemblance to several other areas of ML. In the following, we discuss the main differences and commonalities.

**Backward Compatibility** The term was first introduced by Bansal et al. [1] in the context of humans making decisions based on an AI’s prediction (e.g., medical expert systems or driver supervision in semi-autonomous vehicles). They contextualise that even though an AI’s predictive performance might increase overall, *incompatible* predictions in updated models severely hurt overall performance and trust, and propose to penalize negative flips w.r.t an older model when training a newer model. Yan et al. [53] show that with standard training, there can be a significant number of negative flips, even if the two models only differ in their random initializations. They then reduce the number of negative flips by giving more weight to training points that are correctly classified by the reference model, which they call ‘positive-congruent training’. Previous work on backward-compatible learning is concerned with training a *new* model. Here, we focus on updating the stored predictions rather than updating the stored models. This makes our approach more generally applicable and complements the use-cases of backward-compatible learning. Backward compatibility was further studied empirically by Srivastava et al. [43] who emphasize that this also causes problems for large multi-component AI systems. They propose two key metrics to characterize backward compatibility: (i) Backward Trust Compatibility (BTC), first mentioned in [1], measuring the fraction of predictions that are still predicted correctly after a model update; and (ii) Backward Error Compatibility (BEC), which corresponds to the probability that an incorrect prediction after an update is not new.

**Ensemble Learning** Ensemble methods aim to combine several ML models into a single model with higher performance than each of the individual models. Common techniques are boosting [11], bagging [6], or Bayesian model averaging [16]. Our approach falls into the latter category. We compute the posterior probability (2) in the same way as the well-known Naive Bayes combiner [23]. The classifier corresponding to our MB strategy goes back to at least Nitzan and Paroush [27] and has been thoroughly analyzed [3]. There are also Bayesian techniques that avoid Assumption 1, but these either make some parametric assumptions [20] or assume a very special form of dependence [5].

## 6 Experiments

We now evaluate our Bayesian approach to the Prediction Update Problem against different baselines using the task of image classification as a case study.

### 6.1 Experimental Setup

**Data Sets** We use the three widely accepted benchmark data sets ImageNet1K [7] (1K classes, 50k validation set), ObjectNet [2] (313 classes, 50k validation set) and CIFAR-10 [21] (10 classes, 10k validation set). To imitate our assumed setting of deploying pre-trained models to an unlabelled target data set, we only use the corresponding validation sets as  $\mathcal{D}^{\text{targ}}$ . The ground truth labels are only used post-hoc to compute performance metrics and are not seen during the  $T$  update steps. Of the 313 classes in ObjectNet, 113 are shared with ImageNet, corresponding to a subset of 18,547 images. ObjectNet images exhibit more realistic variations than those in ImageNet. It only has a test set and thus constitutes a challenging transfer scenario for object recognition models. We deploy ImageNet-pretrained models both on ImageNet and on the above subset of ObjectNet, thus simulating the cases that the source and target distributions are the same or different, respectively. For the former, we split the ImageNet validation set in half and use one half to estimate  $\pi^t$  and the other as  $\mathcal{D}^{\text{targ}}$ . For the latter, we estimate  $\pi^t$  from the full ImageNet validation set and evaluate on ObjectNet. We thus assume the covariate shift case, where the conditional label distribution is shared across source and target distributions.

**Models & Architectures** To emulate the setting of sequentially improving classifiers arriving over time, we use the following 17 models and architectures with many of them setting a new “state-of-the-art” on ImageNet at the time they were first introduced: AlexNet [22]; VGG-11, 13, 16, and 19 [41]; ResNet-18, 34, 50, 101, and 152 [15]; SqueezeNet [18]; GoogLeNet [48]; InceptionV3 [49]; MobileNetV2 [36]; DenseNet-121 and 169 [17], and ResNeXt-101 32x8d [52]. For ease of reproducibility, we use pre-trained models from the torchvision model zoo [29] and [32].

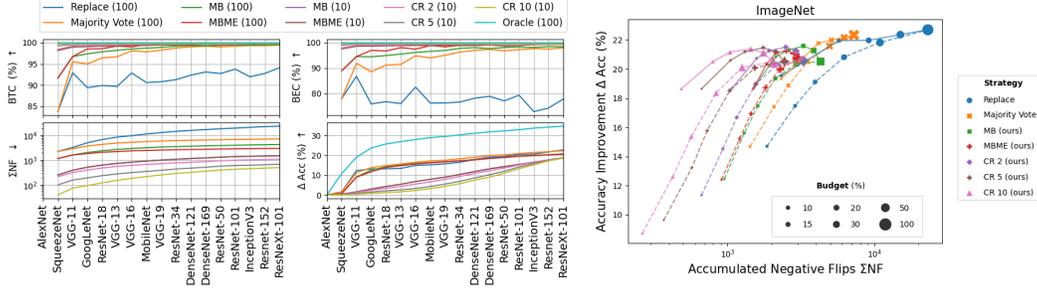


Figure 2: **Left:** Temporal evolution for ImageNet  $\rightarrow$  ImageNet over  $T = 16$  prediction-update steps for a subset of strategies and budgets. **Right:** Comparison of prediction-update strategies across different budgets after  $T = 16$  prediction-update steps. Dashed lines correspond to the ablation using a random selection strategy.

**Performance Metrics** Recall that our goal is to: (i) improve overall accuracy, (ii) avoid negative flips, and (iii) use as few re-evaluations as possible. To assess these different aspects, we report the following metrics: (i) *final accuracy* of the stored predictions (**Acc**) and *accuracy improvement* over the initial accuracy of  $C^0$  ( $\Delta\text{Acc}$ ); (ii) the *cumulative number of negative flips* from time  $t = 0$  to  $T$  ( $\Sigma \text{NF}$ ), the *average negative flip rate* experienced per iteration, i.e.,  $\frac{\Sigma \text{NF}}{N \cdot T}$  (**NFR**), and the ratio of accumulated positive to negative flips (**PF/NF**); (iii) the evaluation budget available to each strategy as percentage of the data set size, i.e., a budget of 10 means that 10% of all samples can be re-evaluated at each time step:  $B^t = 0.1N, \forall t$ ; finally, we measure the connective backward compatibility between (i) and (ii) via Backward Trust Compatibility (**BTC**) and Backward Error Compatibility (**BEC**) [43]. We refer to Appendix A.2 for additional details.

**Baselines and Oracle** We compare our method against two baselines: (i) **Replace** always updates the stored prediction with that predicted by the most recent classifier (a.k.a. *backfilling*); (ii) **Majority Vote** takes into account previous model predictions and updates the stored prediction according to the majority prediction—in case of a tie, the prediction of the most recent classifier is chosen. For reference, we also compare our method against an **Oracle**, which performs a prediction update if and only if this would lead to a positive flip; it thus incurs zero negative flips by definition (knowing the ground truth label). We emphasize that, in practice, we do not have that information in our setting.

**Selection- and Prediction-Update Strategies** For all methods, we select  $B^t \leq N$  samples using the posterior label entropy selection strategy from § 3.2, but also compare with randomly selecting samples for re-evaluation. For all experiments involving the random selection strategies we run the same experiment for five different random seeds each and report the average. We use the prediction-update strategies MB, MBME and CR from § 3.3 and consider cost ratios of  $|c^{\text{NF}}/c^{\text{PF}}| \in \{2, 5, 10\}$  for the latter (e.g., CR 2).

## 6.2 Results for ImageNet $\rightarrow$ ImageNet

In Fig. 2 (left), we show the temporal evolution of backwards compatibility scores, negative flips and accuracy gains for prediction-updates on the ImageNet validation set for a subset of strategies and budgets. A complete account of final performances with additional metrics is shown in Tab. 1 (left).

For the evolution of  $\Delta\text{Acc}$  in Fig. 2 (left), we observe that, unsurprisingly, strategies with 100% budget experience a more rapid gain in accuracy than those with 10%. Among the budget-constrained strategies, the CR strategy with large cost ratio shows the slowest increase, which makes sense as it requires a substantial change in posterior belief for updating a stored prediction and is thus more conservative. Interestingly, however, the *final* accuracies only differ marginally across both strategies and budgets which is also apparent from the minor differences in the  $\Delta\text{Acc}$  column of Tab. 1. For the evolution of  $\Sigma \text{NF}$  in Fig. 2 (right), we observe a clear separation of strategies with a natural ordering from least conservative (Replace) to most conservative (CR 10). These relative differences stay mostly constant over time as NFs appear to accumulate approximately linearly (note the log-scale). We find roughly an order of magnitude difference in  $\Sigma \text{NF}$  between the best non-probabilistic baseline (Majority Vote) and the best Bayesian method (CR 10). Especially for small budgets of up to 30%, our Bayesian strategies clearly dominate the non-probabilistic baselines both in terms of accuracy and flip metrics, as can be seen from Tab. 1 and Fig. 2 (right). Moreover, the CR strategy appears to

Table 1: Results for ImageNet  $\rightarrow$  ImageNet (left) and ImageNet  $\rightarrow$  ObjectNet (right): all metrics refer to final performance for the improving model sequence from Fig. 2 and Fig. 3 respectively. The character **E** or **R** in front of the strategy indicates that selection for re-evaluation is based on the entropy criterion or sampled randomly.

	Strategy	Avg. BTC $\uparrow$	Avg. BEC $\uparrow$	Acc (%) $\uparrow$	$\Delta$ Acc (%) $\uparrow$	$\Sigma$ NF $\downarrow$	NFR (%) $\downarrow$	PF / NF $\downarrow$
	Oracle	100	100	91.2	34.7	0	0	-
Budget = 100%	R-Replace	91.37	77.71	<b>79.2</b>	<b>22.7</b>	24214	6.05	1.2
	R-Majority Vote	97.18	93.95	78.9	22.3	7352	1.84	1.8
	MB	98.32	96.45	77.1	20.5	4378	1.09	2.2
	MBME	98.78	97.69	77.3	20.7	3057	0.76	2.7
	CR 2	98.72	97.19	77.1	20.6	3368	0.84	2.5
	CR 5	99.06	97.82	77.1	20.5	2520	0.63	3
CR 10	<b>99.22</b>	<b>98.15</b>	77	20.5	<b>2112</b>	<b>0.53</b>	<b>3.4</b>	
Budget = 30%	R-Replace	97.56	94.53	77.4	20.8	6546.4	1.64	1.8
	R-Majority Vote	98.6	97.18	77.1	20.8	3616.4	0.9	2.4
	E-Replace	96.53	91.01	<b>78.5</b>	<b>22</b>	9708	2.43	1.6
	E-Majority Vote	98.03	95.63	<b>78.5</b>	<b>22</b>	5232	1.31	2.1
	EMB	98.71	97.25	78.1	21.6	3375	0.84	2.6
	EMBME	98.98	98.04	77.8	21.2	2577	0.64	3.1
E-CR 2	99.02	97.86	78	21.5	2578	0.64	3.1	
E-CR 5	99.32	98.43	78	21.5	1831	0.46	3.9	
E-CR 10	<b>99.44</b>	<b>98.67</b>	77.9	21.4	<b>1517</b>	<b>0.38</b>	<b>4.5</b>	
Budget = 10%	R-Replace	99.22	98.63	71.3	14.7	1958.4	0.49	2.9
	R-Majority Vote	99.4	98.98	71.2	14.7	1481.4	0.37	3.5
	E-Replace	99.04	98.12	<b>76.1</b>	<b>19.5</b>	2468	0.62	3
	E-Majority Vote	99.06	98.18	75.9	19.3	2417	0.6	3
	EMB	99.38	98.89	75.3	18.8	1557	0.39	4
	EMBME	99.38	98.92	75.2	18.7	1533	0.38	4
E-CR 2	99.55	99.22	75.3	18.7	1118	0.28	5.2	
E-CR 5	99.72	99.51	75.2	18.6	700	0.18	7.7	
E-CR 10	<b>99.79</b>	<b>99.64</b>	75.2	18.6	<b>515</b>	<b>0.13</b>	<b>10.1</b>	

	Strategy	Avg. BTC $\uparrow$	Avg. BEC $\uparrow$	Acc (%) $\uparrow$	$\Delta$ Acc (%) $\uparrow$	$\Sigma$ NF $\downarrow$	NFR (%) $\downarrow$	PF / NF $\downarrow$
	Oracle	100	100	50.5	42.6	0	0	-
Budget = 100%	R-Replace	72.65	92.61	<b>31.9</b>	<b>24</b>	16669	5.62	1.3
	R-Majority Vote	89.99	98.02	29.6	21.6	4690	1.58	1.9
	MB	94.46	98.96	29.1	21.2	2477	0.83	2.6
	MBME	95.86	99.34	28.6	20.6	1599	0.54	3.4
	CR 2	95.92	99.21	29	21	1876	0.63	3.1
	CR 5	97.18	99.41	28.8	20.8	1372	0.46	3.8
CR 10	<b>97.82</b>	<b>99.54</b>	28.7	20.8	<b>1084</b>	<b>0.37</b>	<b>4.6</b>	
Budget = 30%	R-Replace	92.19	98.26	<b>29</b>	<b>21</b>	4070.6	1.37	2
	R-Majority Vote	94.91	99.02	27.3	19.4	2346.6	0.79	2.5
	E-Replace	91.75	98.14	<b>29</b>	<b>21</b>	4316	1.45	1.9
	E-Majority Vote	93.54	98.76	28.2	20.3	2970	1	2.3
	EMB	96.24	99.35	27.8	19.9	1565	0.53	3.4
	EMBME	96.64	99.48	26.9	18.9	1280	0.43	3.7
E-CR 2	97.42	99.55	27.7	19.7	1074	0.36	4.4	
E-CR 5	98.43	99.71	27.4	19.4	689	0.23	6.2	
E-CR 10	<b>98.91</b>	<b>99.79</b>	27.1	19.2	<b>504</b>	<b>0.17</b>	<b>8.1</b>	
Budget = 10%	R-Replace	97.49	99.6	22.1	14.2	996.6	0.34	3.6
	R-Majority Vote	97.86	99.68	21.6	13.6	808.8	0.27	4.1
	E-Replace	97.5	99.6	<b>23.7</b>	<b>15.7</b>	996	0.34	3.9
	E-Majority Vote	97.5	99.6	<b>23.7</b>	<b>15.7</b>	996	0.34	3.9
	EMB	98.08	99.72	22.7	14.8	696	0.23	4.9
	EMBME	98.08	99.72	22.7	14.8	696	0.23	4.9
E-CR 2	98.68	99.83	20.7	12.8	427	0.14	6.6	
E-CR 5	99.32	99.92	18.2	10.3	197	0.07	10.7	
E-CR 10	<b>99.57</b>	<b>99.95</b>	17.2	9.2	<b>122</b>	<b>0.04</b>	<b>15</b>	

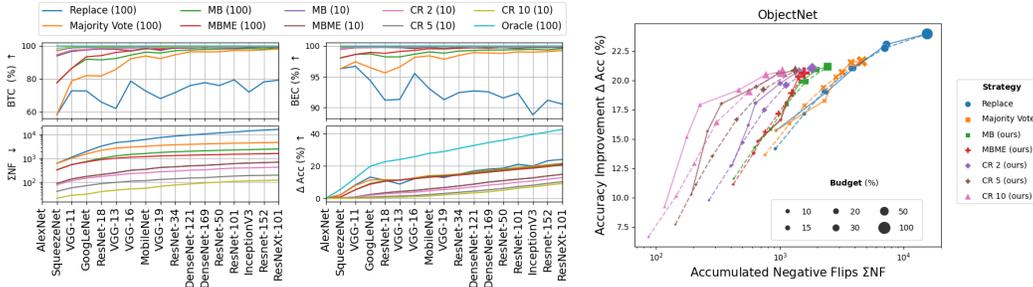


Figure 3: **Left:** Temporal evolution for ImageNet  $\rightarrow$  ObjectNet over  $T = 16$  prediction-update steps for a subset of strategies and budgets. **Right:** Comparison of all strategies after  $T = 16$  on ObjectNet.

provide control over the number of negative flips via its cost-ratio hyperparameter without adversely affecting final accuracy across a range of budgets, as already observed for a budget of 10% in Fig. 2. Interestingly, the update rules seem to be optimal when evaluating on less than 100% budget. We attribute this to posterior approximation errors on ImageNet, which is being supported by extensive ablations in the supplement. Regarding backward compatibility (our ultimate goal), we find that **BTC and BEC** scores reliably outperform the baselines across all budgets. In particular, the CR 10 strategy seems to be especially suitable with scores close to 100%, i.e., oracle performance.

**Summary** Our method appears to successfully fulfill the three desiderata for backward-compatible prediction-updates in an i.i.d. setting. In particular, our CR strategy seems like the most promising candidate to (i) maintain high accuracy gains and (ii) introduce very few negative flips, when (iii) given only a small compute budget for re-evaluations.

### 6.3 Results for ImageNet $\rightarrow$ ObjectNet

Results for prediction-updates on ObjectNet are presented (similarly to § 6.2) in Fig. 3 and Tab. 1 (right). This transfer setting constitutes a much more challenging task. Nevertheless, we observe very similar behaviour to that discussed in § 6.2 and thus only point out the main differences. First, we note that - despite the smaller target data set - the difference in negative flips across different strategies and budgets is even larger on ObjectNet. For example, we observe a reduction in  $\Sigma$  NF of more than two orders of magnitude between Replace (100) and CR (10), and about one order when comparing the two for the same budget. At the same time, differences in accuracy across strategies are also slightly more pronounced, especially for the smallest budget of 10%. Here, the more conservative CR strategies yield lower accuracy gains while MB and MBME maintain competitive accuracy gains. Our strategies are again clearly dominating in terms of backward compatibility w.r.t. BTC and BEC. We remark that these results are agnostic to any potential differences in the label space: they are based on a posterior over all 1000 ImageNet classes whereas ObjectNet only contains a subset of 113 of these classes.

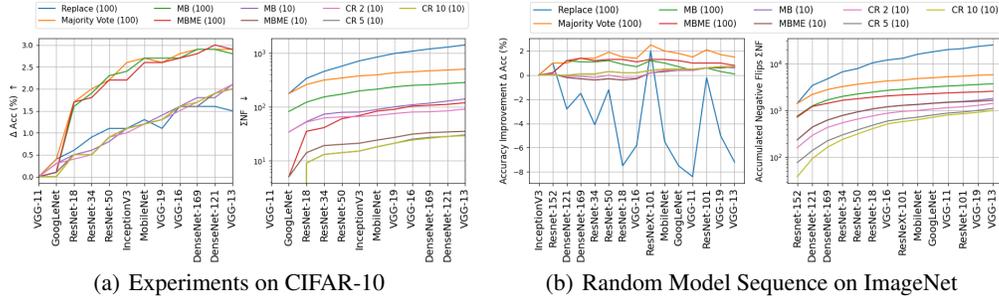


Figure 4: **Left:** Temporal evolution for CIFAR-10 over  $T = 11$  prediction-update steps for a subset of strategies and budgets. Our approaches achieve competitive overall accuracy gains (even at 10% of the budget) while introducing only a fraction of negative flips. **Right:** Experiment on ImageNet with random model sequence suggesting robustness of our approach to situations where an ordering by performance of may not be available.

## 6.4 Further Experiments and Ablations

**Results for CIFAR-10** In Fig. 4 (a) we show the corresponding results on the CIFAR-10 dataset. Here, pre-trained models exhibit a higher level of accuracy ( $\approx 93 - 95\%$ ) and we thus emulate an arguably more realistic scenario with models being released more frequently, thus with smaller accuracy differences. Our method shows very similar trends as we have worked out on ImageNet and ObjectNet. Interestingly, there is one novel characteristic: due to the presumably less steep increase in accuracy from one model to its successor and fewer class categories, we can form very accurate posterior beliefs which results in accuracy gains of all our methods that even outperform the accuracy-optimizing baselines concerning desideratum (i). A more comprehensive analysis of these experiments can be found in Appendix A.5.

**Role of the Selection Strategy** We also conduct a comparison between our entropy selection and the random baseline for all our methods across a range of budgets - see dashed lines in scatter plots. We find that random selection leads to substantially smaller accuracy gains, but also to fewer negative flips which is intuitive since random selection more often chooses “easy” samples for re-evaluation.

**Robustness to Random & Adversary Model Sequences** We have assumed that the models  $C^t$  are *improving* over time. We thus also consider the scenario where  $C^t$  arrive in a *random* or *adversarial* (i.e., strictly deteriorating) order. For the random order (see Fig. 4 (b)) we find that our methods - unlike, e.g., the replace strategy - achieve strict increases in accuracy while introducing much fewer negative flips. Even in the adversarial case, our methods improve accuracy during the initial steps with much fewer negative flips over the entire history. These findings suggest robustness of our approach to situations where an ordering by performance of  $C^t$  may not be available. We refer to Appendix A.3 for more details.

**Reducing Re-Evaluations Matters at Scale** The re-evaluations of a sample using deep neural network based models clearly dominate computational cost as compared to our method. As an example, the forward pass using the public ImageNet PyTorch models takes up to 550 (biggest VGG and ResNets) times longer than the unoptimized implementation of our method backbone. We analyze this in more detail in Appendix A.9. For very large data sets and with new models generally increasing in size, reducing the inference budget  $B$  is of crucial importance, emphasizing the relevance of desideratum 3.

## 7 Conclusion

The Prediction Update Problem appears frequently in practice and can take different forms. It relates to many different subfields of ML that we have discussed in § 4 and § 5, and there are interesting extensions (structured prediction, adaptive budgets) and improvements (modeling data set structure, across-dataset similarity, domain adaptation, calibration techniques) that need to be worked out. In this work, we have studied the classification case and proposed a Bayesian update rule based on simple assumptions. Empirically, we find improvements along the dimensions we set out to achieve, and we hope that progress on this problem will democratize ML usage even further as it lowers the bar for benefitting from the tremendous progress in model design seen over the last years.

## Acknowledgements

The authors would like to thank Matthias Bethge, Thomas Brox, Chris Russell, Karsten Roth, Nasim Rahaman, Alex Smola, Yuanjun Xiong and Stefano Soatto for helpful discussions and feedback.

## References

- [1] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019.
- [2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.
- [3] Daniel Berend and Aryeh Kontorovich. A finite sample analysis of the naive bayes classifier. *J. Mach. Learn. Res.*, 16(1), 2015.
- [4] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.
- [5] Philip J Boland, Frank Proschan, and Yung Liang Tong. Modelling dependence in simple and indirect majority systems. *Journal of Applied Probability*, 1989.
- [6] Leo Breiman. Bagging predictors. *Machine learning*, 24(2), 1996.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
- [8] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2), 2009.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.
- [10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. 2021.
- [11] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 1997.
- [12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 2015.
- [13] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4), 1953.
- [14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [16] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, 1999.
- [17] 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.

- [18] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [19] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.
- [20] Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, 2012.
- [21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.
- [23] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.
- [24] Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 1956.
- [25] David JC MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1), 1995.
- [26] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
- [27] Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 1982.
- [28] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 2009.
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
- [30] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.
- [31] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V Le. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*, 2020.
- [32] Huy Phan. huyvnphan/pytorch\_cifar10, jan 2021.
- [33] Joelle Pineau, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. ICLR2018 reproducibility challenge. In *ICLR workshop on Reproducibility in Machine Learning*, 2018.
- [34] Joelle Pineau, Koustuv Sinha, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. ICLR reproducibility challenge 2019. *ReScience C*, 5(2), 2019.
- [35] Herbert Robbins. An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Symp. Math. Statist. Probab.*, 1956, volume 1, 1956.
- [36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2 inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

- [37] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison, Department of Computer Sciences*, 2009.
- [38] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3), 1948.
- [39] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [40] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2), 2000.
- [41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.
- [42] Koustuv Sinha, Joelle Pineau, Jessica Forde, Rosemary Nan Ke, and Hugo Larochelle. Neurips 2019 reproducibility challenge. *ReScience C*, 6(2), 2020.
- [43] Megha Srivastava, Besmira Nushi, Ece Kamar, Shital Shah, and Eric Horvitz. An empirical analysis of backward compatibility in machine learning systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.
- [44] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.
- [45] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.
- [46] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [47] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.
- [48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
- [49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [50] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy: Fixefficientnet. *Advances in Neural Information Processing Systems 32*, 2019.
- [51] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [52] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
- [53] Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. *Conference on Computer Vision and Pattern Recognition*, 2021.
- [54] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from imperfect labelers. *Advances in Neural Information Processing Systems 29*, 2016.
- [55] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning (ICML)*, pages 819–827. PMLR, 2013.

## A Additional Experiments and Ablation Studies

In this supplement, we present additional results and ablations of experiments referred to in § 6 of the main paper.

### A.1 Assumption of Conditionally-Independent Classifiers

In the case of two given classifiers our assumption implies that  $\forall n, \forall (t, t') : I(\hat{Y}_n^t; \hat{Y}_n^{t'} | Y_n) = 0$ . To quantify deviations of Assumption 1, one could therefore measure this conditional mutual information, but note that this in itself is challenging on finite data and would again inherit model assumptions through the choice of the estimator. Since pairwise independence does not imply mutual independence, even having these quantities might not be sufficient for a long classifier sequence. An alternative approach could thus be to measure the Maximum Mean Discrepancy (MMD) between the LHS and RHS of Assumption 1, which can be viewed as a conditional version of the d-dimensional Hilbert Schmidt Independence Criterion (dHSIC), see [30].

Importantly, all of the above approaches aim to quantify and subsequently incorporate correlations between classifier predictions, and thus rely on having large amounts of labelled data to properly estimate them which is almost never the case in practice. Since our goal was to devise an approach that can be useful in practice, when the number of classes, observations, and models is typically very large, we believe that the model choice behind Assumption 1 is reasonable; it enables both tractable inference and parameter estimation.

### A.2 Backward Compatibility Scores

For the reported BTC and BEC scores we follow the definition proposed in [43] with  $h_1$  and  $h_2$  being the predictors at time step 1 and time step 2. Translated to our setup, we associate  $h_1$  with our estimated and stored prediction at time step 1 whereas  $h_2$  corresponds to the updated prediction dataset at time step 2:

**Backward Trust Compatibility (BTC) score.** “The ratio of points in a held-out test set (e.g.,  $D_{test}$ ) that predicted correctly among all points  $h_1$  had already predicted correctly.” [43]

$$\text{BTC} = \frac{\sum_{i=1}^{|D|} \mathbf{1}[h_1(x_i) = y_i, h_2(x_i) = y_i]}{\sum_{i=1}^{|D|} \mathbf{1}[h_1(x_i) = y_i]} \quad (4)$$

**Backward Error Compatibility (BEC) score.** “The proportion of points in a held-out test set that  $h_2$  predicted incorrectly, out of which  $h_1$  also predicted incorrectly, thus capturing the probability that a mistake made by  $h_2$  is not new.” [43]

$$\text{BEC} = \frac{\sum_{i=1}^{|D|} \mathbf{1}[h_1(x_i) \neq y_i, h_2(x_i) \neq y_i]}{\sum_{i=1}^{|D|} \mathbf{1}[h_2(x_i) \neq y_i]} \quad (5)$$

### A.3 Robustness to Random and Adversarial Model Sequences

As already eluded to in § 6.4, in our main experiments we have assumed that the models  $C^t$  are *improving* over time. In general, this assumption may be justified by having observed superior performance of a new model for the source domain on which they were trained, prior to deciding to use it for re-evaluation on the target data set. However, such information may not always be available.

To check for robustness against violations of the assumption of improving classifiers, we also consider scenarios where  $C^t$  arrive in a *random* or *adversarial* (i.e., strictly deteriorating) order with respect to their accuracy. To avoid unrealistically large fluctuations in the accuracy of new classifiers, we removed AlexNet and SqueezeNet as the most poorly performing models from the *random* model sequence for this set of experiments. Results on ImageNet in the form of the temporal evolution of accuracy improvement and accumulated negative flips across different strategies and budgets are shown in Fig. 5.

For the random ordering in Fig. 5 (a), we find that our methods—unlike, e.g., the replace strategy—achieve strict increases in *final* accuracy while introducing much fewer negative flips, e.g., almost an order of magnitude fewer for our CR strategies, compared to the Majority Vote baseline. Even in the adversarial case in Fig. 5 (b), our methods improve accuracy during the initial steps and introduce much fewer negative flips over the entire history. These findings suggest robustness of our approach to situations where an ordering by performance of  $C^t$  may not be available.

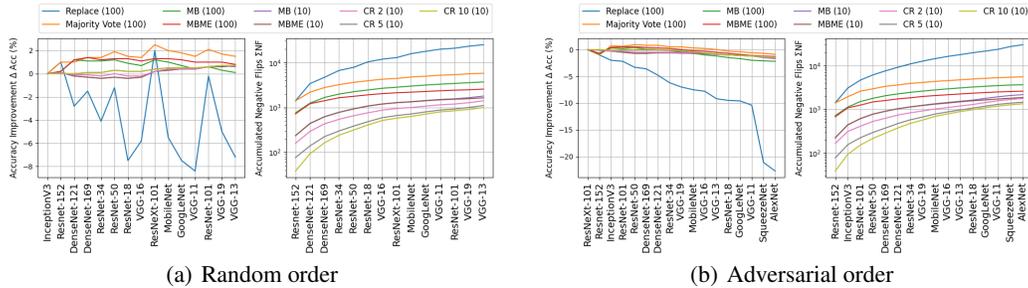


Figure 5: Temporal evolution of accuracy improvement and accumulated negative flips on ImageNet for two scenarios where the models  $C^t$  do not exhibit improving performance, but instead arrive in (a) random or (b) adversarial order. As can be seen, our methods are robust in both these cases while introducing much fewer negative flips than the non-probabilistic baselines.

#### A.4 Role of Confusion Matrix Estimates

For the results of our main experiments reported in § 6.2 and § 6.3, we estimated only the diagonal elements of the confusion matrices and set the off-diagonal elements to a constant, c.f. § 3.1. To investigate the role of the estimation procedure for the confusion matrices  $\pi^t$ , we also report results for the case where we estimate the *full* confusion matrix of each classifier (i.e., including the off-diagonal elements). To avoid off-diagonal elements which are estimated to be zero due to the large number of classes and the limited size of the validation set split, we use a Laplace-smoothed version of the maximum likelihood estimate, i.e., we add a one to each count and normalise accordingly, as suggested in § 4.

Results of this comparison between different confusion matrix estimators for ImageNet and ObjectNet are shown in Fig. 6. (The corresponding results for CIFAR-10 are shown in Fig. 7 and are discussed in more detail in A.5.) As is apparent from the comparison on ImageNet and ObjectNet in Fig. 6, we indeed observe generally improved performance from estimating full confusion matrices, as speculated in § 4.

In terms of final accuracy, we observe a different behaviour across different budgets. For large budgets, estimating full confusion matrices leads to substantial accuracy gains on ImageNet and to roughly equal or slightly improved accuracy on ObjectNet. For small budgets, on the other hand, we find similar or smaller accuracy gains for full confusion matrix estimates; importantly, this effect is most pronounced for the more conservative CR prediction-update strategies. In terms of negative flips, we observe slight to major reductions across all budgets and strategies with full confusion matrix estimates compared to only estimating the diagonals, i.e., the class-specific accuracies.

We believe that this behaviour is intuitive and can be explained as follows. Estimating the full confusion matrices with smoothed counts generally yields smaller diagonal elements and larger off-diagonal elements. This, in turn, means that our posterior beliefs (in particular, the ratio between two consecutive MAP values) change less drastically after new re-evaluations, so that the more conservative CR strategies update fewer labels and thus experience smaller and slower accuracy gains when only few samples can be re-evaluated. At the same time, the resulting posterior estimates are likely more accurate which is consistent with larger gains for large budgets and the reduced number of negative flips. In summary, better confusion matrix estimates result in a further reduction in the number of negative flips across the board. For small budgets and conservative prediction-update strategies, however, this comes with a trade-off in overall accuracy gain.

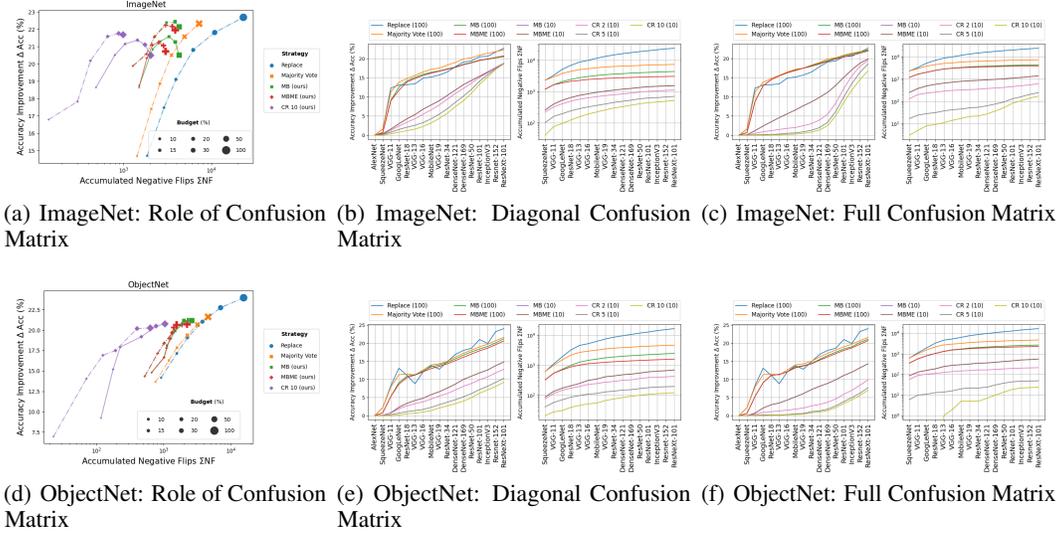


Figure 6: Investigation into the role of different confusion matrix estimates on ImageNet (top) and ObjectNet (bottom). Dashed lines in the scatter plots (a) and (d) represent results for when the full confusion matrices (i.e., including the off-diagonals) are estimated from smoothed counts on a split of the validation set, and solid lines represent results under the less accurate confusion matrix estimates where only diagonal elements (i.e., the class-specific accuracies) are estimated, as in the experiments of the main paper.

### A.5 Experiments on CIFAR-10

CIFAR-10 is an image classification dataset comprising images of objects from 10 distinct classes (airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, trucks). It contains a training set of 50,000 images and a test set of size 10,000. It is thus a rather simple classification dataset compared to ImageNet and ObjectNet but has been a key driver for advancing ML models and computer vision before the introduction of ImageNet. Using the CIFAR-10 test set as our target dataset, we can study the Prediction Updates Problem in an i.i.d. setting with a small number of classes.

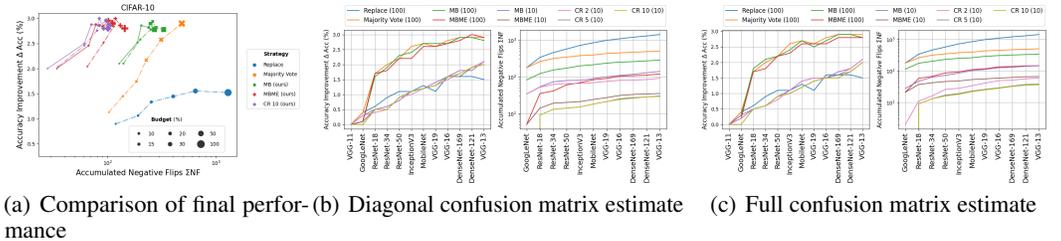


Figure 7: Summary of prediction-updates on CIFAR-10. Similar to Fig. 6, dashed lines in (a) represent results with full confusion matrices estimated from smoothed counts and solid lines represent using a less accurate confusion matrix estimate where only diagonals are estimated from data. We observe very similar behaviour to that described for ImageNet and ObjectNet in § 6.2 and § 6.3, respectively. The most interesting difference compared to our experiments on ImageNet and ObjectNet, is that, on CIFAR-10, our methods achieve substantially larger accuracy gains compared with the baselines, even for small budgets.

Similar to our experiments on ImageNet and ObjectNet reported in § 6 where we used models which had been pre-trained on ImageNet,<sup>6</sup> for our experiments on CIFAR-10, we use models which have instead been pre-trained on CIFAR-10, available from an open source github repository.<sup>7</sup> This repository contains pre-trained models for a subset of the same architectures listed in § 6.1, excluding

<sup>6</sup>ImageNet pretrained models: [link](#)  
<sup>7</sup>CIFAR-10 pretrained models: [link](#)

AlexNet, SqueezeNet, ResNet-101, ResNeXt; the released VGG pre-trained models represent variants with batch normalization. We order these CIFAR-10 pre-trained models according to their validation accuracies reported in the above repository. Compared to ImageNet and ObjectNet, pre-trained models on CIFAR-10 exhibit a much higher level of accuracy (between 93% and 95%) with much smaller differences between the best and worst model. As a consequence, our experiment on CIFAR-10 emulates a scenario in which incoming new classifiers exhibit smaller performance gains. This might reflect a practical prediction-updates setting with a higher frequency of incoming new models.

The results of our experiments on CIFAR-10 are summarised in Fig. 7. Our method shows very similar trends w.r.t. all update strategies and metrics as we have worked out on ImageNet and ObjectNet. Interestingly, there is one novel characteristic when it comes to the accuracy gains: due to the presumably less steep increase in accuracy from one model to its successor and fewer class categories, we can form more accurate posterior beliefs which results in accuracy gains for our CR 10 with smallest budget that outperforms the backfilling Replace (100) baseline. The overall accuracy gain similarly remains competitive with all Majority Vote variants while exhibiting only a fraction of the negative flips. Again, BTC and BEC strongly outperform all baselines along every update step for all our methods as can be seen in Fig. 8

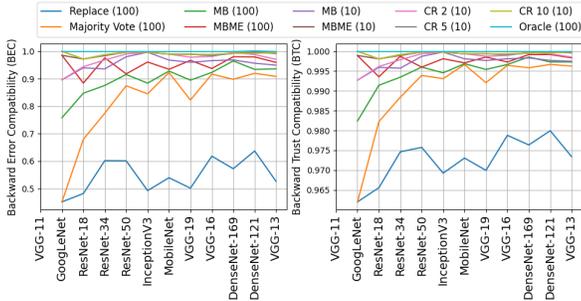


Figure 8: BTC and BEC evolution on CIFAR10 with utilizing diagonal confusion matrix elements only.

Figure 8

### A.6 On Peaking Accuracies

In Fig. 2 (right, solid lines), we observed a peak in  $\Delta\text{Acc}$  at 30% budget for all our strategies when performing prediction-updates on ImageNet. This suggests that our update rules are no longer beneficial on this dataset when re-evaluating more than the top 30% of samples with highest posterior entropy. From our ablation experiments on different confusion matrix estimates shown in Fig. 6 (a), we observe that this peaking phenomenon is much less pronounced and occurs only at larger budgets of  $B \geq 50\%$  when estimating the full confusion matrix based on the half of the validation set on which we do not evaluate. We therefore conjecture that this behaviour may be related to inaccurate estimates of the posterior resulting from our approximation of the unknown confusion matrices (and possibly the assumption of conditionally independent classifiers).

To further investigate this hypothesis, we performed an ablation where we use the *full* validation set both to estimate the confusion matrix (with smoothing) and for evaluation—i.e., we do not use a 50-50 split of the validation set as in Fig. 6 and all experiment in the main paper—thus matching almost exactly<sup>8</sup> the confusion counts statistics of the evaluation set. As a result, we can expect our estimates of the posterior distribution to be much more accurate in this case.

The results are shown in Fig. 9 in dashed lines, compared to the peaking behaviour from our default method in solid lines for reference. Apart from substantial gains in overall accuracy, we find that the peaking phenomenon disappears almost entirely, with only a very slight drop in accuracy remaining between budgets of 50% and 100% for some strategies. This seems to confirm our hypothesis that the main source of the peaking behaviour are approximation errors in our posterior estimates resulting from inaccurate confusion matrices.

### A.7 Role of the Selection Strategy

We also conduct an ablation study on the role of the selection strategy for samples to be re-evaluated as was being mentioned in § 6.4. In particular we conduct a comparison between our posterior entropy selection and the random selection (without replacement) baseline for all our methods across a range of budgets—see Fig. 10 for these ablation results on all three datasets CIFAR-10, ImageNet and

<sup>8</sup>up to the one count smoothing effect.

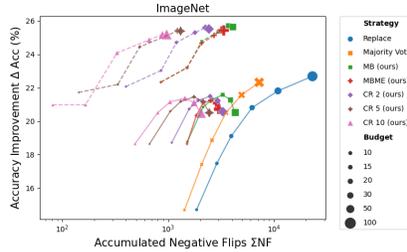


Figure 9: Further investigation into the peaking phenomenon observed for prediction-updates on ImageNet in Fig. 2 (right). Solid lines are as in Fig. 2 (right), i.e., using a 50:50 split of the validation set, using one half to estimate the diagonal elements of the confusion matrices and the other half as our target evaluation set; dashed lines show the results of our ablation experiment where we use the full validation set to both estimate the exact statistics of the confusion matrix and to evaluate the different methods. Since the evaluation set for the latter is twice as large, we scaled the reported negative flips by 0.5 for ease of comparison.

ObjectNet. Along all datasets, we find that random selection leads to substantially smaller accuracy gains, but also results in fewer negative flips. We can explain these fewer negative flips by the fact that our random selection more often chooses “easy” samples for re-evaluation. Under a random selection criteria, there is no preference in re-evaluating more “controversial” samples w.r.t previous predictions. This effect becomes particularly pronounced for small budgets. Our finding suggests that the selection strategy is indeed a relevant component, with room for improvement as discussed in § 4. This is also highlighting the potential control capability of this component regarding our three desiderata.

### A.8 Evolution of the Stored Label Distribution

For additional insights which samples are getting selected for re-evaluation by our entropy selection criterion, we also show some characteristic count distribution of correctly and incorrectly stored predictions along various update steps. See Fig. 11 for this distribution after every second update step on ImageNet using the MB strategy with a compute budget of  $B = 10$  (%).

### A.9 Reducing re-evaluations matters at scale

To see that the re-evaluations of an example image using deep neural network based models clearly dominate computational cost as compared to our method backbone that involved computing the approximate posterior, label entropy and update strategy details, we measure the time for the plain model estimation vs the time for posterior update on the same Intel Xeon (Cascade Lake) CPU for the sake of comparison. We summarize our measurement in Tab. 2 for all used ImageNet models, showing the multiples in compute time required by the model inference (per image on average) compared to the timing per image of our fully unoptimized method implementation, i.e. the posterior update backbone. For very large data sets and with new models generally increasing in size, reducing the inference budget  $B$  is of crucial importance, emphasizing the relevance of desideratum 3.

### A.10 Result Tables of Main Experiments

For sake of completeness and reproducibility, we also provide a complete quantitative account of all the main experiments and ablations on all three datasets discussed throughout this paper in Tab. 3 and Tab. 4.

### A.11 On Utilizing Uncalibrated Soft Labels

Finally, we tested our method on incorporating the uncalibrated soft labels provided as output in the pre-trained ImageNet models by repeating the ImageNet and ObjectNet experiment (in combination with full confusion matrix estimation). As already discussed in § 4, deep neural networks are known to have unreliable uncertainty estimates [14, 25, 26, 47] and we therefore do not expect these

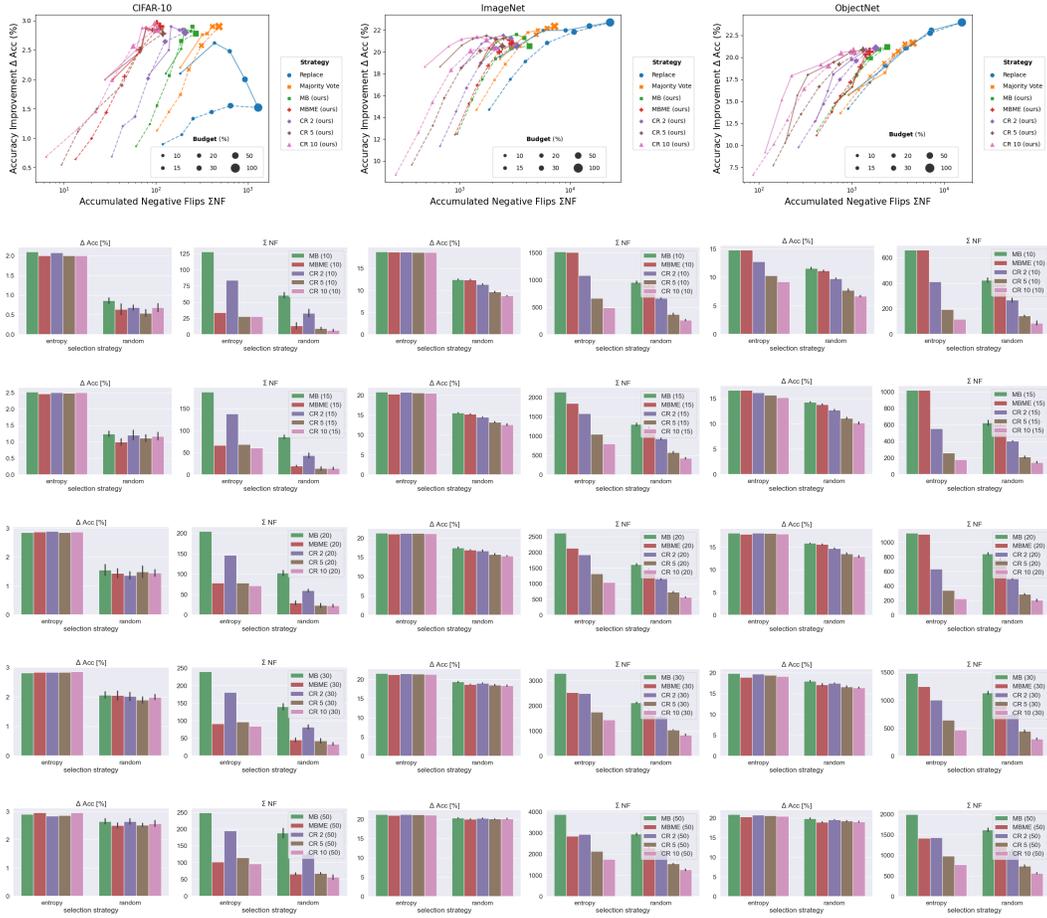


Figure 10: Ablation results for comparing random (dashed lines in scatter plots) and entropy-ranked (solid lines in scatter plots) selection strategies on CIFAR-10 (left column), ImageNet (middle column) and ObjectNet (right column). Different rows of the bar plots correspond to, from top to bottom,  $B = 10, 15, 20, 30, 50\%$ . Note that at 100% budgets, the two selection strategies are identical as all samples are re-evaluated at each step. We find that random selection leads to substantially smaller accuracy gains, but also fewer negative flips. This is intuitive as random selection more often chooses “easy” samples for re-evaluation. The effect is particularly pronounced for small budgets.

Table 2: Time measurements of a single forward pass versus the mean average time to compute the posterior update (0.406 milliseconds) on the same computational resources for comparability. For reference we extrapolated how long it would take to infer the predictions from 1B images.

Model	model re-evaluation [milliseconds]	ratio inference vs posterior update	inference for 1B images [days]
alexnet	29,6	73	343
squeezenet	32,6	80	378
vgg11	125,5	309	1452
googlenet	65,6	162	759
resnet18	31,5	78	364
vgg13	164,1	404	1899
vgg16	191,3	471	2215
mobilenet	23,1	57	267
vgg19	223,0	550	2582
resnet34	50,8	125	587
densenet121	74,4	183	862
densenet169	93,5	230	1082
resnet50	70,6	174	818
resnet101	114,0	281	1319
inceptionv3	115,8	285	1340
resnet152	158,2	390	1831
resnext101 32x8d	206,6	509	2391

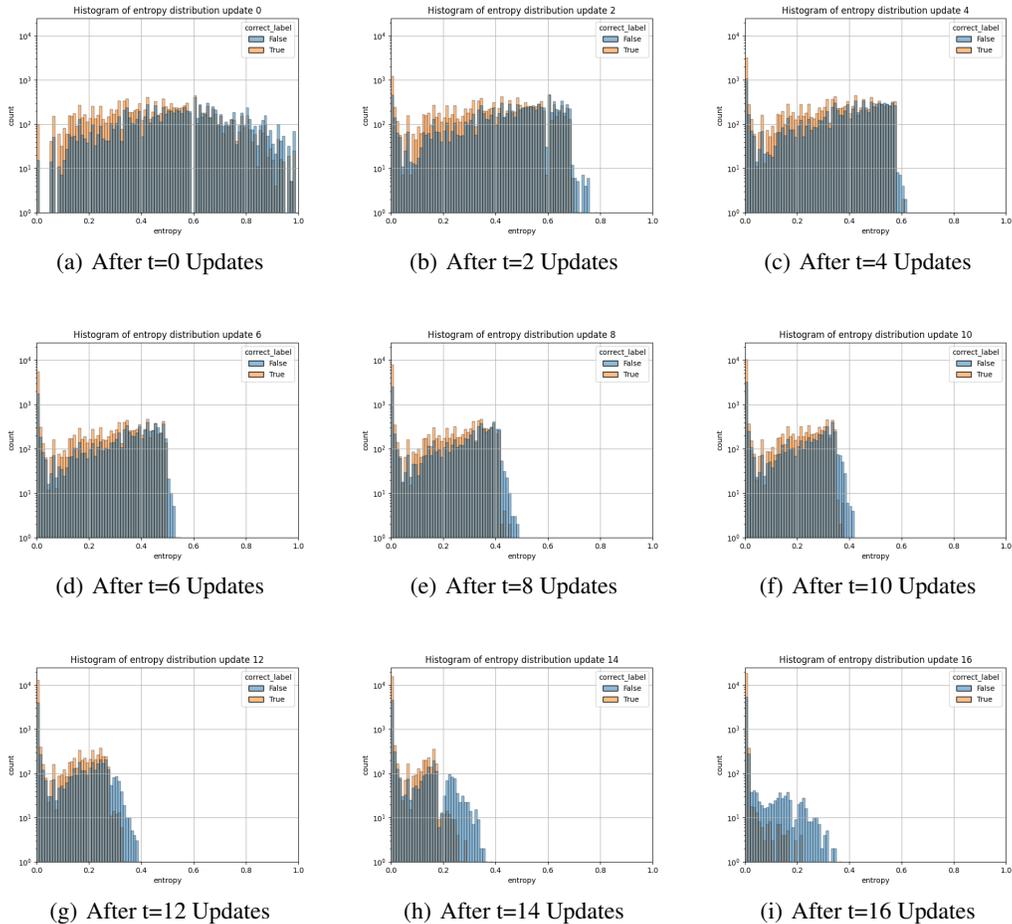


Figure 11: Count distribution evolution of correctly and incorrectly stored predictions on ImageNet using the MB update strategy under a budget of  $B = 10$ . Samples with currently stored predictions being false are depicted with blue, and correctly stored predictions with orange. We observe that under this MB strategy, samples with highest entropy tend to store false predictions, thus benefitting from re-evaluation particularly.

results to be representative of what our method could achieve if it has access to truly calibrated soft labels. We now utilize the full output softmax vector  $\mathbf{p}^{sl} = C^t(\mathbf{x}_n)$  and refine our likelihood estimate in eq. (2) by multiplying the confusion matrix coefficients with the soft label vector, i.e.  $\pi^t(\hat{y}_n^t, k) \rightarrow \sum_i p_i^{sl} \pi^t(i, k)$ . We show the temporal evolution compared to the hard-label implementation using only diagonal confusion matrix estimates in Fig. 12 and a full account over the final performance metrics and all selection strategies and budgets in Tab. 5 and Tab. 6. Overall the results are slightly less conclusive and less consistent, most likely due to the known uncalibrated soft labels, emphasizing the strength of our method when only having deterministic labels. Specifically, on CIFAR-10, there are additional accuracy gains but often we accumulate more negative flips and BTC, BEC are typically slightly worse. On ImageNet, accuracy gains are sometimes worse without finding a clear pattern when this is the case, but negative flips are sometimes significantly lower. BTC, BEC seem to be better on the majority of update strategies and labels. On ObjectNet, we likewise find no clear patten in the accuracy gains. However, accumulated negative flips are generally much lower and both, BTC and BEC, even higher (very close to 1.0) but we want to emphasize that they were already very high for the ObjectNet experiments from the main paper.

Table 3: Full results of all experiments on **ImageNet (left)** and **ObjectNet (right)** under the standard setting of an **improving sequence of models** and **estimating confusion matrices on a separate split** of the data (i.e., not on the target data set) as explained in § 6.1. The first number in each cell refers to estimating only the diagonal elements (i.e., the class-specific accuracies) of confusion matrices, and the **second number (in brackets)** refers to **estimating the full confusion matrices** with smoothing. The character **E** in front of a budget indicates that selection for re-evaluation is based on our posterior label **entropy criterion** (e.g., E30 refers to selecting the top 30% samples with highest entropy), and the character **R** indicates selecting a **randomly** sampled subset without replacement. Note that entropy-based selection requires a posterior and is thus not applicable for the baselines Replace and Majority Vote.

Strategy	Budget (%)	Acc (%)	ΔAcc (%)	Σ NF	NFR (%)	PF / NF	Avg. BTC	Avg. BEC
I00	Oracle	91.2 (91.2)	34.7 (34.7)	0 (0)	0.0 (0.0)	NaN (NaN)	100.0 (100.0)	100.0 (100.0)
R100	Replace	79.2 (79.2)	22.7 (22.7)	24214 (24214)	6.05 (6.05)	1.2 (1.2)	91.37 (91.37)	77.71 (77.71)
R100	Majority Vote	78.9 (78.9)	22.3 (22.3)	7352 (7352)	1.84 (1.84)	1.8 (1.8)	97.18 (97.18)	93.95 (93.95)
E100	MBM	77.1 (78.7)	20.5 (22.2)	30228 (19528)	1.09 (1.09)	2.2 (2.3)	98.32 (98.32)	96.40 (98.4)
E100	MBME	77.3 (78.5)	20.7 (22.0)	3057 (3910)	0.76 (0.98)	2.7 (2.4)	98.78 (98.48)	97.69 (96.88)
E100	CR 2	77.1 (78.5)	20.6 (22.0)	3368 (2342)	0.84 (0.59)	2.5 (3.3)	98.72 (99.12)	97.19 (98.0)
E100	CR 5	77.1 (78.5)	20.5 (21.8)	2520 (1300)	0.63 (0.33)	3.0 (2.9)	99.06 (99.52)	97.82 (98.86)
E100	CR 10	77.0 (78.2)	20.5 (21.7)	2112 (1023)	0.53 (0.26)	3.4 (6.1)	99.22 (99.62)	98.15 (99.11)
R50	Replace	78.4 (78.4)	21.8 (21.8)	11478.8 (11478.8)	2.87 (2.87)	1.5 (1.5)	95.82 (95.82)	89.88 (89.88)
R50	Majority Vote	78.1 (78.1)	21.6 (21.6)	5048.2 (5048.2)	1.26 (1.26)	2.1 (2.1)	98.07 (98.07)	95.92 (95.92)
R50	MBM	78.9 (77.7)	21.4 (21.4)	30228 (19528)	1.09 (1.09)	2.7 (2.7)	98.32 (98.32)	97.62 (97.7)
R50	MBME	76.5 (77.5)	20.0 (20.9)	2325.4 (2917.6)	0.58 (0.73)	3.1 (2.8)	99.08 (98.87)	98.27 (97.72)
R50	CR 2	76.8 (77.4)	20.1 (20.8)	2333.6 (1572.8)	0.56 (0.39)	3.3 (4.3)	99.14 (99.4)	98.22 (98.74)
R50	CR 5	76.6 (77.0)	20.1 (20.4)	1603.8 (840.6)	0.4 (0.21)	4.1 (4.1)	99.39 (99.68)	98.69 (99.32)
R50	CR 10	76.7 (76.7)	20.1 (20.2)	1314.6 (659.2)	0.33 (0.16)	4.8 (8.6)	99.51 (99.75)	98.92 (99.46)
E50	Replace	78.9 (79.6)	22.4 (23.1)	15732 (15503)	3.93 (3.88)	1.4 (1.4)	94.41 (94.53)	85.37 (85.23)
E50	Majority Vote	78.5 (78.1)	22.1 (22.6)	6318 (6420)	1.58 (1.46)	1.9 (1.9)	97.67 (97.58)	94.74 (94.56)
E50	MBM	77.8 (79.0)	21.3 (22.4)	3969 (3941)	0.99 (0.99)	2.3 (2.4)	98.48 (98.5)	96.78 (96.71)
E50	MBME	77.6 (78.7)	21.1 (22.2)	2904 (2836)	0.73 (0.91)	2.8 (2.5)	98.35 (98.6)	97.79 (97.05)
E50	CR 2	77.8 (78.8)	21.2 (22.2)	3014 (2128)	0.75 (0.53)	2.8 (3.6)	98.86 (99.21)	97.5 (98.17)
E50	CR 5	77.7 (78.6)	21.1 (22.0)	2214 (1168)	0.55 (0.29)	3.4 (5.7)	99.18 (99.57)	98.09 (98.96)
E50	CR 10	77.7 (78.3)	21.1 (21.8)	1832 (913)	0.46 (0.23)	3.9 (7.0)	99.33 (99.67)	98.4 (99.19)
R20	Replace	77.4 (77.4)	20.8 (20.8)	6546.4 (6546.4)	1.64 (1.64)	1.8 (1.8)	97.56 (97.56)	94.53 (94.53)
R20	Majority Vote	77.1 (77.1)	20.5 (20.5)	3616.4 (3616.4)	0.9 (0.9)	2.4 (2.4)	98.6 (98.6)	97.18 (97.18)
R20	MBM	75.9 (76.3)	19.4 (19.7)	2186.2 (2200.4)	0.55 (0.55)	3.2 (3.2)	99.14 (99.14)	98.35 (98.34)
R20	MBME	75.3 (76.2)	18.7 (19.6)	1839.6 (2183.8)	0.46 (0.55)	3.5 (3.2)	99.26 (99.14)	98.65 (98.36)
R20	CR 2	75.5 (75.5)	19.0 (18.9)	1607.0 (1092.4)	0.4 (0.27)	4.0 (5.3)	99.37 (99.57)	98.8 (99.0)
R20	CR 5	75.1 (74.7)	18.5 (18.2)	1087.2 (564.8)	0.27 (0.14)	4.3 (9.1)	99.58 (99.78)	99.17 (99.57)
R20	CR 10	74.9 (74.2)	18.4 (17.7)	875.4 (434.2)	0.22 (0.11)	6.2 (11.2)	99.66 (99.83)	99.33 (99.68)
E20	Replace	78.5 (79.3)	22.2 (22.8)	9708 (8900)	2.43 (2.23)	1.6 (1.6)	96.53 (96.82)	91.01 (91.69)
E20	Majority Vote	78.5 (79.2)	22.0 (22.6)	5232 (4889)	1.31 (1.28)	2.1 (2.1)	98.03 (98.12)	95.63 (95.79)
E20	MBM	78.1 (78.9)	21.6 (22.4)	3375 (3180)	0.84 (0.8)	2.6 (2.8)	98.71 (98.79)	97.25 (97.37)
E20	MBME	77.8 (78.8)	21.2 (22.2)	2577 (3071)	0.64 (0.77)	3.1 (2.8)	98.98 (98.83)	98.04 (97.49)
E20	CR 2	78.0 (78.7)	21.5 (22.1)	2578 (1664)	0.64 (0.42)	3.1 (4.3)	99.02 (99.37)	97.86 (98.58)
E20	CR 5	78.0 (78.3)	21.5 (21.8)	1831 (890)	0.46 (0.22)	3.9 (7.1)	99.32 (99.67)	98.43 (99.21)
E20	CR 10	77.9 (78.1)	21.4 (21.6)	1517 (707)	0.38 (0.18)	4.5 (8.6)	99.44 (99.74)	98.67 (99.37)
R10	Replace	75.7 (75.7)	19.1 (19.1)	4171.0 (4171.0)	1.04 (1.04)	2.1 (2.1)	98.41 (98.41)	96.71 (96.71)
R10	Majority Vote	75.4 (75.4)	18.9 (18.9)	2690.2 (2690.2)	0.67 (0.67)	2.8 (2.8)	98.95 (98.95)	97.99 (97.99)
R10	MBM	74.0 (74.2)	17.5 (17.6)	1662.8 (1655.4)	0.42 (0.41)	3.6 (3.7)	99.34 (99.34)	98.81 (98.81)
R10	MBME	73.5 (74.1)	16.9 (17.6)	1480.6 (1654.0)	0.37 (0.41)	3.9 (3.7)	99.4 (99.34)	98.97 (98.82)
R10	CR 2	73.2 (72.5)	16.7 (15.9)	1205.8 (786.2)	0.3 (0.32)	4.5 (4.6)	99.52 (99.68)	99.15 (99.45)
R10	CR 5	72.3 (71.3)	15.8 (14.7)	768.0 (388.8)	0.19 (0.11)	6.1 (10.5)	99.69 (99.84)	99.46 (99.67)
R10	CR 10	71.9 (70.5)	15.4 (14.0)	603.8 (280.2)	0.15 (0.07)	7.4 (13.5)	99.76 (99.89)	99.57 (99.81)
E10	Replace	78.5 (79.0)	22.0 (22.5)	6191 (5830)	1.55 (1.36)	1.9 (2.0)	97.74 (98.02)	94.47 (95.16)
E10	Majority Vote	78.3 (78.7)	21.8 (22.1)	4295 (3788)	1.07 (0.95)	2.3 (2.5)	98.37 (98.56)	96.46 (96.89)
E10	MBM	77.9 (78.1)	21.3 (21.6)	2700 (2371)	0.68 (0.59)	3.0 (3.3)	98.96 (99.09)	97.36 (98.11)
E10	MBME	77.6 (78.1)	21.1 (21.6)	2183 (2371)	0.55 (0.59)	3.4 (3.3)	99.13 (99.09)	98.39 (98.11)
E10	CR 2	77.8 (77.8)	21.3 (21.3)	1999 (1197)	0.5 (0.3)	3.7 (5.4)	99.23 (99.54)	98.41 (99.04)
E10	CR 5	77.7 (77.4)	21.2 (20.8)	1383 (599)	0.35 (0.15)	4.8 (9.7)	99.47 (99.77)	98.87 (99.49)
E10	CR 10	77.7 (78.7)	21.2 (20.2)	1101 (661)	0.28 (0.12)	5.8 (11.9)	99.58 (99.83)	99.08 (99.62)
R10	Replace	71.3 (71.3)	14.7 (14.7)	1958.4 (1958.4)	0.49 (0.49)	2.9 (2.9)	99.22 (99.22)	98.63 (98.63)
R10	Majority Vote	71.2 (71.2)	14.7 (14.7)	1481.4 (1481.4)	0.37 (0.37)	3.5 (3.5)	99.4 (99.4)	98.98 (98.98)
R10	MBM	69.0 (69.0)	12.5 (12.4)	991.4 (963.6)	0.25 (0.24)	4.1 (4.2)	99.59 (99.6)	99.25 (99.36)
R10	MBME	68.9 (69.0)	12.4 (12.4)	944.8 (976.8)	0.24 (0.24)	4.3 (4.2)	99.61 (99.6)	99.38 (99.35)
R10	CR 2	67.9 (66.0)	11.4 (9.5)	703.6 (432.2)	0.18 (0.11)	5.0 (6.5)	99.71 (99.82)	99.54 (99.73)
R10	CR 5	66.2 (64.2)	9.6 (7.6)	393.4 (147.0)	0.1 (0.04)	7.1 (14.0)	99.84 (99.84)	99.75 (99.91)
R10	CR 10	65.3 (63.5)	8.8 (7.0)	282.6 (104.8)	0.07 (0.03)	8.7 (17.6)	99.88 (99.96)	99.82 (99.93)
E10	Replace	76.1 (77.3)	19.5 (20.8)	2468 (2320)	0.62 (0.58)	3.0 (3.2)	99.04 (99.11)	98.12 (98.12)
E10	Majority Vote	75.2 (76.4)	18.7 (19.9)	1557 (1394)	0.39 (0.35)	4.0 (4.6)	99.38 (99.44)	98.92 (98.98)
E10	MBM	75.3 (76.3)	18.7 (19.5)	1118 (618)	0.28 (0.15)	5.2 (9.8)	99.55 (99.9)	99.22 (99.46)
E10	MBME	75.2 (76.4)	18.8 (19.9)	1553 (1394)	0.38 (0.35)	4.0 (4.6)	99.38 (99.44)	98.92 (98.98)
E10	CR 2	75.3 (76.3)	18.7 (19.5)	1118 (618)	0.28 (0.15)	5.2 (9.8)	99.55 (99.9)	99.22 (99.46)
E10	CR 5	75.2 (75.2)	18.6 (18.6)	700 (246)	0.18 (0.06)	7.7 (19.9)	99.72 (99.9)	99.51 (99.81)
E10	CR 10	75.2 (73.3)	18.6 (16.8)	515 (170)	0.13 (0.04)	10.1 (25.7)	99.79 (99.93)	99.64 (99.91)
I00	Oracle	50.5 (50.5)	4.2 (4.2)	0 (0)	0.0 (0.0)	NaN (NaN)	100.0 (100.0)	99.99 (99.99)
R100	Replace	31.9 (31.9)	24.0 (24.0)	16669 (16669)	5.62 (5.62)	1.3 (1.3)	72.65 (72.65)	92.61 (92.61)
R100	Majority Vote	29.6 (29.6)	21.6 (21.6)	4690 (4690)	1.58 (1.58)	1.9 (1.9)	89.99 (89.99)	98.02 (98.02)
E100	MBM	24.7 (28.5)	19.8 (21.3)	2477 (2875)	0.83 (0.83)	3.2 (3.1)	96.15 (95.9)	98.96 (98.88)
E100	MBME	28.6 (28.7)	20.6 (20.7)	1599 (2267)	0.54 (0.76)	3.4 (2.7)	95.86 (94.63)	99.34 (99.06)
E100	CR 2	29.0 (28.8)	21.0 (20.9)	1876 (1611)	0.63 (0.54)	3.1 (5.4)	95.92 (96.54)	99.21 (99.32)
E100	CR 5	28.4 (28.5)	20.8 (20.6)	1372 (635)	0.46 (0.29)	3.8 (5.5)	97.18 (98.28)	99.41 (99.63)
E100	CR 10	28.7 (28.2)	20.8 (20.3)	1084 (660)	0.37 (0.22)	4.6 (6.7)	97.82 (98.68)	99.54 (99.72)
R50	Replace	30.7 (30.7)	22.7 (22.7)	7583.8 (7583.8)	2.56 (2.56)	1.6 (1.6)	86.63 (86.63)	96.69 (96.69)
R50	Majority Vote	28.6 (28.6)	20.7 (20.7)	3281.6 (3281.6)	1.11 (1.11)	2.2 (2.2)	93.01 (93.01)	98.62 (98.62)
R50	MBM	27.8 (27.7)	19.9 (19.8)	4662 (1780.4)	0.86 (0.6)	3.2 (3.1)	96.15 (95.9)	99.1 (99.26)
R50	MBME	26.9 (27.4)	18.9 (19.4)	1280.4 (1682.0)	0.43 (0.57)	3.7 (3.1)	96.77 (96.94)	99.48 (99.31)
R50	CR 2	27.5 (27.3)	19.6 (19.4)	1165.4 (923.0)	0.39 (0.31)	4.1 (4.9)	97.3 (97.86)	99.52 (99.62)
R50	CR 5	27.1 (26.6)	19.2 (18.7)	780.2 (460.6)	0.26 (0.16)	5.6 (6.4)	98.26 (98.98)	99.67 (99.8)
R50	CR 10	27.0 (26.2)	19.1 (18.3)	598.8 (353.2)	0.2 (0.12)	6.9 (10.6)	98.7 (99.22)	99.75 (99.85)
E50	Replace	31.0 (31.3)	23.1 (23.4)	7882 (7730)	2.66 (2.6)	1.5 (1.6)	86.26 (86.78)	96.53 (96.99)
E50	Majority Vote	28.5 (28.4)	21.5 (21.7)	3895 (3759)	1.31 (1.27)	2.1 (2.1)	91.73 (92.2)	98.36 (98.49)
E50	MBM	28.9 (29.1)	20.9 (21.2)	2079 (2081)	0.7 (0.7)	2.9 (2.9)	95.23 (95.34)	99.13 (99.13)
E50	MBME	28.2 (28.6)	20.3 (20.7)	1452 (1944)	0.49 (0.66)	3.6 (3.0)	96.18 (95.52)	99.41 (99.19)
E50	CR 2	28.7 (28.7)	20.8 (20.8)	1506 (1136)	0.51 (0.38)	3.6 (4.4)	96.64 (97.54)	99.37 (99.52)
E50	CR 5	28.5 (28.4)	20.6 (20.4)	1049 (569)	0.35 (0.19)	4.6 (7.7)	97.83 (98.88)	99.55 (99.75)
E50	CR 10	28.4 (28.1)	20.5 (20.2)	828 (430)	0.28 (0.14)	5.6 (9.7)	98.35 (99.16)	99.64 (99.81)
R20	Replace	29.0 (29.0)	21.0 (21.0)	4070.6 (4070.6)	1.37 (1.37)	2.0 (2.0)	92.19 (92.19)	98.26 (98.26)
R20	Majority Vote	27.3 (27.3)	19.4 (19.4)	2346.6 (2346.6)	0.79 (0.79)	2.5 (2.5)	94.91 (94.91)	99.02 (99.02)
R20	MBM	25.9 (25.9)	18.0 (17.9)	1185.8 (1218.6)	0.4 (0.41)	3.8 (3.7)	97.12 (97.01)	99.52 (99.5)
R20	MBME	25.1 (25.7)	17.2 (17.6)	1008.4 (1208.0)	0.34 (0.41)	4.2 (3.7)	97.41 (97.03)	99.59 (99.51)
R20	CR 2	25.4 (24.9)	17.5 (16.9)	782.2 (601.2)	0.26 (0.2)	5.2 (6.2)	98.05 (98.46)	99.68 (99.76)
R20	CR 5	2						

Table 4: Full results of all experiments on **CIFAR10** under the standard setting of an **improving sequence of models** and **estimating confusion matrices on a separate split** of the data (i.e., not on the target data set) as explained in § 6.1. The first number in each cell refers to estimating only the diagonal elements (i.e., the class-specific accuracies) of confusion matrices, and the **second number (in brackets)** refers to **estimating the full confusion matrices** with smoothing. The character **E** in front of a budget indicates that selection for re-evaluation is based on our posterior label **entropy criterion** (e.g., E30 refers to selecting the top 30% samples with highest entropy), and the character **R** indicates selecting a **randomly** sampled subset without replacement. Note that entropy-based selection requires a posterior and is thus not applicable for the baselines Replace and Majority Vote.

Strategy (Budget %)	Acc (%)	$\Delta$ Acc (%)	$\Sigma$ NF	NFR (%)	PF / NF	Avg. BTC	Avg. BEC
I00:Oracle	99.3 (99.3)	6.1 (6.1)	0 (0)	0.0 (0.0)	NaN (NaN)	100.0 (100.0)	100.0 (100.0)
R100:Replace	94.7 (94.7)	1.5 (1.5)	1418 (1418)	2.58 (2.58)	1.1 (1.1)	97.26 (97.26)	54.79 (54.79)
R100:Majority Vote	96.1 (96.1)	2.9 (2.9)	502 (502)	0.91 (0.91)	1.3 (1.3)	99.03 (99.03)	81.94 (81.94)
E100:MB	95.9 (95.9)	2.8 (2.8)	284 (340)	0.52 (0.62)	1.5 (1.4)	99.45 (99.34)	89.64 (87.71)
E100:MBME	96.1 (96.0)	2.9 (2.8)	119 (146)	0.22 (0.27)	2.2 (2.0)	99.77 (99.72)	95.28 (94.47)
E100:CR 2	96.0 (96.0)	2.8 (2.8)	219 (170)	0.4 (0.31)	1.6 (1.8)	99.58 (99.67)	91.69 (93.28)
E100:CR 5	95.9 (95.9)	2.8 (2.8)	132 (126)	0.24 (0.23)	2.1 (2.1)	99.75 (99.76)	94.65 (94.98)
E100:CR 10	96.0 (96.0)	2.9 (2.8)	112 (103)	0.2 (0.19)	2.3 (2.4)	99.79 (99.8)	95.49 (95.92)
R50:Replace	94.7 (94.7)	1.6 (1.6)	711.4 (711.4)	1.29 (1.29)	1.1 (1.1)	98.62 (98.62)	77.8 (77.8)
R50:Majority Vote	95.7 (95.7)	2.6 (2.6)	327.4 (327.4)	0.6 (0.6)	1.4 (1.4)	99.37 (99.37)	89.06 (89.06)
R50:MB	95.8 (95.8)	2.7 (2.6)	196.4 (227.8)	0.36 (0.41)	1.7 (1.6)	99.62 (99.56)	93.41 (92.47)
R50:MBME	95.7 (95.6)	2.5 (2.5)	69.8 (90.6)	0.13 (0.16)	2.8 (2.4)	99.87 (99.83)	97.58 (96.93)
R50:CR 2	95.8 (95.8)	2.6 (2.6)	133.8 (104.2)	0.24 (0.19)	2.0 (2.3)	99.74 (99.8)	95.32 (96.24)
R50:CR 5	95.7 (95.7)	2.5 (2.5)	73.2 (69.0)	0.13 (0.13)	2.7 (2.8)	99.86 (99.87)	97.4 (97.59)
R50:CR 10	95.7 (95.7)	2.6 (2.5)	61.0 (60.6)	0.11 (0.11)	3.1 (3.1)	99.88 (99.88)	97.83 (97.85)
E50:Replace	95.2 (95.2)	2.0 (2.0)	1019 (1012)	1.85 (1.84)	1.1 (1.1)	98.04 (98.05)	64.97 (64.69)
E50:Majority Vote	96.1 (96.0)	2.9 (2.9)	428 (419)	0.78 (0.76)	1.3 (1.3)	99.18 (99.19)	84.48 (84.84)
E50:MB	96.1 (96.0)	2.9 (2.8)	260 (289)	0.47 (0.53)	1.6 (1.5)	99.5 (99.44)	90.63 (89.8)
E50:MBME	96.1 (96.0)	3.0 (2.9)	105 (134)	0.19 (0.24)	2.4 (2.1)	99.8 (99.74)	96.04 (95.09)
E50:CR 2	96.0 (96.0)	2.8 (2.8)	206 (149)	0.37 (0.27)	1.7 (1.9)	99.6 (99.71)	92.4 (94.27)
E50:CR 5	96.0 (95.9)	2.9 (2.8)	123 (112)	0.22 (0.2)	2.2 (2.2)	99.76 (99.79)	95.18 (95.67)
E50:CR 10	96.1 (96.0)	3.0 (2.8)	99 (95)	0.18 (0.17)	2.5 (2.5)	99.81 (99.82)	96.21 (96.35)
R30:Replace	94.6 (94.6)	1.4 (1.4)	439.6 (439.6)	0.8 (0.8)	1.2 (1.2)	99.15 (99.15)	86.74 (86.74)
R30:Majority Vote	95.3 (95.3)	2.2 (2.2)	238.4 (238.4)	0.43 (0.43)	1.5 (1.5)	99.54 (99.54)	92.54 (92.54)
R30:MB	95.2 (95.3)	2.1 (2.1)	146.2 (163.2)	0.27 (0.3)	1.7 (1.7)	99.72 (99.68)	95.46 (94.91)
R30:MBME	95.2 (95.2)	2.0 (2.1)	47.0 (64.2)	0.09 (0.12)	3.2 (2.6)	99.91 (99.88)	98.47 (97.92)
R30:CR 2	95.2 (95.2)	2.0 (2.0)	88.8 (64.6)	0.16 (0.12)	2.1 (2.6)	99.83 (99.88)	97.16 (97.92)
R30:CR 5	95.1 (95.0)	1.9 (1.9)	45.6 (47.0)	0.08 (0.09)	3.1 (3.0)	99.91 (99.91)	98.54 (98.49)
R30:CR 10	95.2 (95.1)	2.0 (2.0)	35.4 (35.4)	0.06 (0.06)	3.8 (3.8)	99.93 (99.93)	98.85 (98.85)
E30:Replace	95.6 (95.4)	2.5 (2.2)	688 (689)	1.25 (1.25)	1.2 (1.2)	98.68 (98.68)	75.18 (75.12)
E30:Majority Vote	95.9 (95.9)	2.8 (2.8)	380 (395)	0.69 (0.72)	1.4 (1.4)	99.27 (99.24)	86.44 (85.83)
E30:MB	96.0 (96.1)	2.8 (2.9)	253 (277)	0.46 (0.5)	1.6 (1.5)	99.51 (99.47)	91.14 (90.39)
E30:MBME	96.0 (96.2)	2.8 (3.0)	96 (123)	0.17 (0.22)	2.5 (2.2)	99.82 (99.76)	96.45 (95.68)
E30:CR 2	96.0 (96.2)	2.8 (3.0)	192 (147)	0.35 (0.27)	1.7 (2.0)	99.63 (99.72)	93.12 (94.5)
E30:CR 5	96.0 (96.2)	2.8 (3.0)	103 (98)	0.19 (0.18)	2.4 (2.6)	99.8 (99.81)	96.07 (96.36)
E30:CR 10	96.0 (96.2)	2.9 (3.0)	89 (86)	0.16 (0.16)	2.6 (2.7)	99.83 (99.83)	96.67 (96.85)
R20:Replace	94.5 (94.5)	1.3 (1.3)	274.6 (274.6)	0.5 (0.5)	1.2 (1.2)	99.47 (99.47)	91.94 (91.94)
R20:Majority Vote	94.9 (94.9)	1.7 (1.7)	195.2 (195.2)	0.35 (0.35)	1.4 (1.4)	99.62 (99.62)	94.06 (94.06)
R20:MB	94.7 (94.8)	1.6 (1.7)	107.0 (120.4)	0.19 (0.22)	1.7 (1.7)	99.79 (99.77)	96.86 (96.43)
R20:MBME	94.6 (94.6)	1.4 (1.4)	31.0 (45.4)	0.06 (0.08)	3.3 (2.6)	99.94 (99.91)	99.09 (98.67)
R20:CR 2	94.5 (94.5)	1.4 (1.4)	65.0 (48.6)	0.12 (0.09)	2.0 (2.4)	99.87 (99.91)	98.08 (98.55)
R20:CR 5	94.7 (94.7)	1.5 (1.5)	26.6 (27.4)	0.05 (0.05)	3.8 (3.7)	99.95 (99.95)	99.19 (99.16)
R20:CR 10	94.6 (94.6)	1.5 (1.4)	24.6 (24.4)	0.04 (0.04)	4.0 (3.9)	99.95 (99.95)	99.25 (99.26)
E20:Replace	95.8 (95.7)	2.6 (2.5)	462 (478)	0.84 (0.87)	1.3 (1.3)	99.11 (99.08)	83.35 (83.11)
E20:Majority Vote	95.9 (96.0)	2.8 (2.8)	336 (347)	0.61 (0.63)	1.4 (1.4)	99.35 (99.33)	88.2 (87.94)
E20:MB	96.0 (95.9)	2.9 (2.8)	217 (263)	0.39 (0.48)	1.7 (1.5)	99.58 (99.49)	92.45 (91.02)
E20:MBME	96.0 (96.0)	2.9 (2.8)	82 (115)	0.15 (0.21)	2.8 (2.2)	99.84 (99.78)	97.02 (96.04)
E20:CR 2	96.1 (95.9)	2.9 (2.8)	156 (130)	0.28 (0.24)	1.9 (2.1)	99.7 (99.75)	94.53 (95.27)
E20:CR 5	96.0 (95.9)	2.9 (2.8)	84 (90)	0.15 (0.16)	2.7 (2.5)	99.84 (99.83)	96.86 (96.71)
E20:CR 10	96.0 (95.9)	2.9 (2.8)	75 (80)	0.14 (0.15)	2.9 (2.7)	99.86 (99.85)	97.22 (97.11)
R10:Replace	94.1 (94.1)	0.9 (0.9)	129.2 (129.2)	0.23 (0.23)	1.3 (1.3)	99.75 (99.75)	96.31 (96.31)
R10:Majority Vote	94.3 (94.3)	1.1 (1.1)	109.4 (109.4)	0.2 (0.2)	1.5 (1.5)	99.79 (99.79)	96.86 (96.86)
R10:MB	94.0 (94.2)	0.9 (1.0)	64.0 (70.8)	0.12 (0.13)	1.7 (1.7)	99.88 (99.86)	98.23 (98.0)
R10:MBME	93.8 (93.8)	0.6 (0.6)	14.2 (21.4)	0.03 (0.04)	3.3 (2.4)	99.97 (99.96)	99.61 (99.41)
R10:CR 2	93.8 (93.8)	0.7 (0.7)	35.2 (23.8)	0.06 (0.04)	2.0 (2.4)	99.93 (99.95)	99.03 (99.34)
R10:CR 5	93.7 (93.7)	0.5 (0.6)	11.0 (11.0)	0.02 (0.02)	3.5 (3.6)	99.98 (99.98)	99.69 (99.69)
R10:CR 10	93.8 (93.8)	0.7 (0.7)	6.8 (7.0)	0.01 (0.01)	6.0 (5.9)	99.99 (99.99)	99.81 (99.8)
E10:Replace	95.3 (95.4)	2.1 (2.2)	206 (201)	0.37 (0.37)	1.5 (1.5)	99.6 (99.61)	93.45 (93.65)
E10:Majority Vote	95.4 (95.4)	2.2 (2.2)	200 (200)	0.36 (0.36)	1.6 (1.6)	99.61 (99.61)	93.67 (93.69)
E10:MB	95.3 (95.3)	2.1 (2.1)	140 (145)	0.25 (0.26)	1.8 (1.7)	99.73 (99.72)	95.59 (95.45)
E10:MBME	95.2 (95.2)	2.0 (2.0)	35 (66)	0.06 (0.12)	3.9 (2.5)	99.93 (99.87)	98.92 (98.02)
E10:CR 2	95.2 (95.2)	2.1 (2.1)	91 (59)	0.17 (0.11)	2.1 (2.7)	99.82 (99.89)	97.26 (98.11)
E10:CR 5	95.2 (95.2)	2.0 (2.0)	30 (38)	0.05 (0.07)	4.3 (3.7)	99.94 (99.93)	99.05 (98.79)
E10:CR 10	95.2 (95.2)	2.0 (2.0)	29 (36)	0.05 (0.07)	4.4 (3.8)	99.94 (99.93)	99.09 (98.86)

Table 5: Full results of all experiments on **CIFAR-10 (left)** and **ImageNet (right)** under the standard setting of an **improving sequence of models** and **estimating confusion matrices on a separate split** of the data (i.e., not on the target data set) as explained in § 6.1. The first number in each cell refers to estimating only the diagonal elements (i.e., the class-specific accuracies) of confusion matrices, and the **second number (in brackets)** refers to **estimating the full confusion matrices** with smoothing and incorporating soft labels. The character **E** in front of a budget indicates that selection for re-evaluation is based on our posterior label entropy criterion (e.g., E30 refers to selecting the top 30% samples with highest entropy), and the character **R** indicates selecting a **randomly** sampled subset without replacement. Note that entropy-based selection requires a posterior and is thus not applicable for the baselines Replace and Majority Vote.

Strategy	Budget %	Acc (%)	ΔAcc (%)	Σ NF	NFR (%)	PF / NF	Avg. BTC	Avg. BEC
100Oracle	99.3 (99.3)	6.1 (6.1)	0 (0)	0.0 (0.0)	NaN (NaN)	100.0 (100.0)	100.0 (100.0)	
R100Replace	94.7 (94.7)	1.5 (1.5)	1418 (1418)	2.58 (2.58)	1.1 (1.1)	97.26 (97.26)	54.79 (54.79)	
R100Majority Vote	96.1 (96.1)	2.9 (2.9)	502 (502)	0.91 (0.91)	1.3 (1.3)	99.03 (99.03)	81.94 (81.94)	
E100MB	95.9 (96.2)	2.8 (3.0)	284 (252)	0.52 (0.46)	1.5 (1.6)	99.45 (99.52)	89.64 (89.69)	
E100MBME	96.1 (96.1)	2.9 (3.0)	119 (173)	0.22 (0.31)	2.2 (1.9)	99.77 (99.67)	95.28 (92.98)	
E100CR 2	96.0 (96.1)	2.8 (3.0)	219 (170)	0.4 (0.31)	1.6 (1.9)	99.58 (99.68)	91.69 (92.98)	
E100CR 5	95.9 (96.2)	2.8 (3.0)	132 (103)	0.24 (0.19)	2.1 (2.5)	99.75 (99.8)	94.65 (95.68)	
E100CR 10	96.0 (96.2)	2.9 (3.0)	112 (73)	0.2 (0.13)	2.3 (3.1)	99.79 (99.86)	95.49 (96.99)	
R50Replace	94.7 (94.7)	1.6 (1.6)	714 (714)	1.29 (1.29)	1.1 (1.1)	96.62 (96.62)	77.8 (77.8)	
R50Majority Vote	95.7 (95.7)	2.6 (2.6)	327.4 (327.4)	0.6 (0.6)	1.4 (1.4)	99.37 (99.37)	89.06 (89.06)	
R50MB	95.8 (96.0)	2.7 (2.8)	106.4 (155.0)	0.36 (0.28)	1.7 (1.9)	99.62 (99.7)	93.41 (94.27)	
R50MBME	95.7 (95.9)	2.5 (2.7)	69.8 (106.0)	0.13 (0.19)	2.8 (2.3)	99.87 (99.8)	97.58 (96.2)	
R50CR 2	95.9 (96.0)	2.6 (2.8)	133.8 (99.8)	0.24 (0.18)	2.0 (2.4)	99.74 (99.81)	93.2 (96.2)	
R50CR 5	95.7 (96.0)	2.5 (2.8)	73.2 (59.2)	0.13 (0.11)	2.7 (3.4)	99.86 (99.89)	97.4 (97.7)	
R50CR 10	95.7 (95.9)	2.6 (2.7)	61.0 (40.8)	0.11 (0.07)	3.1 (4.3)	99.88 (99.8)	97.83 (98.47)	
E50Replace	95.2 (95.2)	2.0 (2.0)	1019 (1012)	1.85 (1.84)	1.1 (1.1)	98.04 (98.05)	64.97 (64.69)	
E50Majority Vote	96.1 (96.0)	2.9 (2.9)	428 (419)	0.78 (0.76)	1.3 (1.3)	99.18 (99.19)	84.48 (84.84)	
E50MB	96.1 (96.2)	2.9 (3.0)	260 (257)	0.47 (0.47)	1.6 (1.6)	99.5 (99.51)	90.63 (89.36)	
E50MBME	96.1 (96.1)	3.0 (3.0)	105 (170)	0.19 (0.31)	2.4 (1.9)	99.69 (99.67)	96.04 (93.08)	
E50CR 2	96.0 (96.1)	2.8 (3.0)	206 (170)	0.37 (0.31)	1.7 (1.9)	99.6 (99.68)	92.4 (92.92)	
E50CR 5	96.0 (96.2)	2.9 (3.0)	123 (102)	0.22 (0.19)	2.2 (2.5)	99.76 (99.81)	93.18 (95.72)	
E50CR 10	96.1 (96.2)	3.0 (3.0)	99 (71)	0.18 (0.13)	2.5 (3.1)	99.81 (99.86)	96.21 (97.07)	
R30Replace	94.6 (94.6)	1.4 (1.4)	439.6 (439.6)	0.8 (0.8)	1.2 (1.2)	99.15 (99.15)	86.74 (86.74)	
R30Majority Vote	95.3 (95.3)	2.2 (2.2)	238.4 (238.4)	0.43 (0.43)	1.5 (1.5)	99.54 (99.54)	92.54 (92.54)	
R30MB	95.2 (95.6)	2.1 (2.4)	166.2 (117.6)	0.27 (0.21)	1.7 (2.0)	99.72 (99.77)	95.46 (95.92)	
R30MBME	95.2 (95.6)	2.0 (2.4)	47.0 (88.4)	0.09 (0.16)	3.2 (2.4)	99.91 (99.83)	98.47 (96.94)	
R30CR 2	95.2 (95.5)	2.0 (2.4)	88.8 (72.6)	0.16 (0.15)	2.1 (2.6)	99.83 (99.86)	97.16 (97.5)	
R30CR 5	95.1 (95.4)	1.9 (2.3)	45.6 (40.8)	0.08 (0.07)	3.1 (3.8)	99.91 (99.92)	98.54 (98.61)	
R30CR 10	95.2 (95.4)	2.0 (2.2)	35.4 (28.8)	0.06 (0.05)	3.8 (4.8)	99.93 (99.94)	98.85 (99.03)	
E30Replace	95.6 (95.4)	2.5 (2.2)	688 (689)	1.25 (1.25)	1.2 (1.2)	98.68 (98.68)	78.18 (75.12)	
E30Majority Vote	95.9 (95.9)	2.8 (2.8)	380 (395)	0.69 (0.72)	1.4 (1.4)	99.27 (99.24)	86.44 (85.83)	
E30MB	96.0 (96.3)	2.8 (3.1)	253 (230)	0.46 (0.42)	1.6 (1.7)	99.51 (99.56)	91.14 (90.39)	
E30MBME	96.0 (96.2)	2.8 (3.0)	96 (163)	0.17 (0.3)	2.5 (1.9)	99.82 (99.69)	96.45 (93.3)	
E30CR 2	96.0 (96.2)	2.8 (3.1)	192 (160)	0.35 (0.29)	1.7 (2.0)	99.63 (99.69)	93.12 (93.28)	
E30CR 5	96.0 (96.2)	2.8 (3.1)	103 (94)	0.19 (0.17)	2.4 (2.6)	99.8 (99.82)	96.07 (96.01)	
E30CR 10	96.0 (96.3)	2.9 (3.1)	89 (57)	0.16 (0.11)	2.6 (3.8)	99.83 (99.89)	96.67 (97.65)	
R20Replace	94.5 (94.5)	1.3 (1.3)	274.6 (274.6)	0.5 (0.5)	1.2 (1.2)	99.47 (99.47)	91.94 (91.94)	
R20Majority Vote	94.9 (94.9)	1.7 (1.7)	195.2 (195.2)	0.35 (0.35)	1.4 (1.4)	99.62 (99.62)	94.06 (94.06)	
R20MB	94.7 (95.3)	1.6 (2.2)	107.0 (83.2)	0.19 (0.15)	1.7 (2.3)	99.79 (99.84)	96.86 (97.29)	
R20MBME	94.6 (95.3)	1.4 (2.2)	31.0 (62.0)	0.06 (0.11)	3.3 (2.7)	99.94 (99.88)	99.09 (97.98)	
R20CR 2	94.5 (95.2)	1.4 (2.0)	65.0 (56.4)	0.12 (0.11)	2.0 (2.8)	99.87 (99.89)	98.08 (98.17)	
R20CR 5	94.7 (95.2)	1.5 (2.0)	26.6 (25.6)	0.05 (0.05)	3.8 (4.9)	99.95 (99.95)	99.19 (99.16)	
R20CR 10	94.6 (94.9)	1.5 (1.8)	24.8 (19.6)	0.04 (0.04)	4.0 (5.6)	99.95 (99.96)	99.25 (99.28)	
E20Replace	95.8 (95.7)	2.6 (2.5)	462 (478)	0.84 (0.87)	1.3 (1.3)	99.11 (99.08)	83.35 (83.11)	
E20Majority Vote	95.9 (96.0)	2.8 (2.8)	336 (347)	0.61 (0.63)	1.4 (1.4)	99.35 (99.33)	88.2 (87.94)	
E20MB	96.0 (96.0)	2.9 (2.8)	217 (207)	0.39 (0.36)	1.7 (1.7)	99.5 (99.6)	92.45 (91.74)	
E20MBME	96.0 (95.9)	2.8 (2.8)	82 (148)	0.15 (0.27)	2.8 (1.9)	99.84 (99.72)	97.02 (94.17)	
E20CR 2	96.1 (96.0)	2.9 (2.8)	156 (133)	0.28 (0.24)	1.9 (2.1)	99.7 (99.75)	94.53 (94.67)	
E20CR 5	96.0 (96.0)	2.9 (2.9)	84 (80)	0.15 (0.15)	2.7 (2.8)	99.84 (99.85)	96.86 (96.73)	
E20CR 10	96.0 (96.1)	2.9 (2.9)	75 (55)	0.14 (0.11)	2.9 (3.7)	99.86 (99.89)	97.22 (97.89)	
R10Replace	94.1 (94.1)	0.9 (0.9)	129.2 (129.2)	0.23 (0.23)	1.3 (1.3)	99.75 (99.75)	96.31 (96.31)	
R10Majority Vote	94.3 (94.3)	1.1 (1.1)	109.4 (109.4)	0.2 (0.2)	1.5 (1.5)	99.79 (99.79)	96.86 (96.86)	
R10MB	94.0 (94.7)	0.9 (1.6)	64.0 (51.0)	0.12 (0.09)	1.7 (2.5)	99.8 (99.8)	95.23 (98.49)	
R10MBME	93.8 (94.6)	0.6 (1.4)	14.2 (36.2)	0.03 (0.07)	3.3 (2.9)	99.97 (99.93)	99.61 (98.93)	
R10CR 2	93.8 (94.7)	0.7 (1.5)	35.2 (32.4)	0.06 (0.06)	2.0 (3.3)	99.93 (99.94)	99.03 (99.03)	
R10CR 5	93.7 (94.3)	0.5 (1.1)	11.0 (14.0)	0.02 (0.03)	3.5 (5.1)	99.98 (99.97)	99.69 (99.59)	
R10CR 10	93.8 (94.3)	0.7 (1.1)	6.8 (7.4)	0.01 (0.01)	6.0 (8.5)	99.99 (99.99)	99.81 (99.78)	
E10Replace	95.3 (95.4)	2.1 (2.2)	206 (201)	0.37 (0.37)	1.5 (1.5)	99.6 (99.61)	93.45 (93.65)	
E10Majority Vote	95.4 (95.4)	2.2 (2.2)	200 (200)	0.36 (0.36)	1.6 (1.6)	99.61 (99.61)	93.67 (93.69)	
E10MB	95.3 (95.7)	2.1 (2.5)	140 (120)	0.25 (0.22)	1.8 (2.0)	99.73 (99.77)	95.59 (95.48)	
E10MBME	95.2 (95.6)	2.0 (2.5)	35 (99)	0.06 (0.18)	3.9 (2.5)	99.93 (99.81)	98.92 (96.32)	
E10CR 2	95.2 (95.7)	2.1 (2.5)	91 (78)	0.17 (0.14)	2.1 (2.6)	99.82 (99.85)	97.26 (97.08)	
E10CR 5	95.2 (95.6)	2.0 (2.4)	30 (52)	0.05 (0.09)	4.3 (3.3)	99.94 (99.9)	99.05 (98.04)	
E10CR 10	95.2 (95.5)	2.0 (2.3)	29 (37)	0.05 (0.07)	4.4 (4.1)	99.94 (99.93)	99.09 (98.73)	
100Oracle	91.2 (91.2)	34.7 (34.7)	0 (0)	0.0 (0.0)	NaN (NaN)	100.0 (100.0)	100.0 (100.0)	
R100Replace	79.2 (79.2)	22.7 (22.7)	24214 (24214)	6.05 (6.05)	1.2 (1.2)	91.37 (91.37)	77.71 (77.71)	
R100Majority Vote	78.9 (78.9)	23.2 (22.3)	7352 (7352)	1.84 (1.84)	1.8 (1.8)	97.18 (97.18)	93.95 (93.95)	
E100MB	77.1 (77.6)	20.5 (21.0)	4378 (3463)	1.09 (0.87)	2.2 (2.5)	98.32 (98.65)	96.45 (97.25)	
E100MBME	77.3 (77.4)	20.7 (20.8)	3057 (3207)	0.76 (0.8)	2.7 (2.6)	98.78 (98.74)	97.69 (97.51)	
E100CR 2	77.1 (77.3)	20.6 (20.8)	3368 (1648)	0.84 (0.41)	2.5 (4.2)	98.72 (99.36)	97.19 (98.7)	
E100CR 5	77.1 (77.0)	20.5 (20.5)	2520 (990)	0.63 (0.25)	3.0 (6.2)	99.06 (99.63)	97.82 (99.19)	
E100CR 10	77.0 (76.8)	20.5 (20.3)	2112 (788)	0.53 (0.2)	3.4 (7.4)	99.22 (99.71)	98.19 (99.54)	
R50Replace	78.4 (78.4)	21.8 (21.8)	11478.8 (11463.8)	2.87 (2.87)	1.5 (1.5)	98.82 (98.83)	89.88 (89.89)	
R50Majority Vote	78.1 (78.1)	21.6 (21.6)	5048.2 (5048.8)	1.26 (1.26)	2.1 (2.1)	98.07 (98.07)	95.92 (95.92)	
R50MB	76.9 (76.9)	20.4 (20.4)	3022.8 (2509.8)	0.76 (0.63)	2.7 (3.0)	98.3 (98.3)	97.62 (98.05)	
R50MBME	76.5 (76.7)	20.0 (20.1)	2325.4 (2437.2)	0.58 (0.61)	3.1 (3.1)	99.08 (99.05)	98.27 (98.12)	
R50CR 2	78.8 (78.5)	20.2 (19.4)	1616.1 (1203.8)	0.56 (0.31)	2.5 (2.6)	98.4 (98.4)	97.92 (98.66)	
R50CR 5	76.6 (76.1)	20.1 (19.5)	1603.8 (700.4)	0.4 (0.18)	4.1 (8.0)	99.39 (99.73)	98.69 (99.44)	
R50CR 10	76.7 (75.8)	20.1 (19.2)	1314.6 (530.0)	0.33 (0.13)	4.8 (10.1)	99.51 (99.8)	98.92 (99.57)	
E50Replace	78.9 (79.8)	22.4 (23.2)	15732 (16138)	3.93 (4.03)	1.4 (1.4)	98.41 (94.3)	85.37 (84.73)	
E50Majority Vote	78.7 (79.2)	22.2 (22.6)	6318 (6282)	1.58 (1.57)	1.9 (1.9)	97.6 (97.64)	94.74 (94.67)	
E50MB	77.8 (77.9)	21.3 (21.4)	3969 (3212)	0.99 (0.8)	2.3 (2.7)	98.48 (98.77)	96.78 (97.4)	
E50MBME	77.6 (77.6)	21.1 (21.1)	2904 (2972)	0.73 (0.74)	2.8 (2.8)	98.85 (98.85)	97.79 (97.66)	
E50CR 2	77.8 (77.7)	21.2 (21.1)	3014 (1547)	0.75 (0.39)	2.8 (4.4)	98.86 (99.41)	97.5 (98.75)	
E50CR 5	77.7 (77.3)	21.2 (20.8)	2214 (900)	0.55 (0.22)	3.4 (6.8)	99.18 (99.67)	98.09 (99.24)	
E50CR 10	77.7 (77.1)	21.1 (20.6)	1832 (709)	0.46 (0.18)	3.9 (8.3)	99.33 (99.74)	98.4 (99.88)	
R30Replace	77.4 (77.3)	20.8 (20.8)	6546.4 (6537.8)	1.64 (1.63)	1.8 (1.8)	97.56 (97.56)	94.53 (94.55)	
R30Majority Vote	77.1 (77.1)	20.5 (20.5)	3616.4 (3613.6)	0.9 (0.9)	2.4 (2.4)	98.6 (98.6)	97.18 (97.18)	
R30MB	75.9 (75.9)	19.4 (19.4)	2186.2 (1860.6)	0.55 (0.47)	3.2 (3.6)	99.14 (99.27)	98.35 (98.6)	
R30MBME	75.3 (75.8)	18.7 (19.3)	1859.6 (1855.6)	0.46 (0.46)	3.5 (3.6)	99.26 (99.27)	98.65 (98.61)	
R30CR 2	75.5 (75.3)	19.0 (18.7)	1607.0 (918.2)	0.4 (0.23)	4.0 (6.1)	99.37 (99.64)	98.85 (99.31)	
R30CR 5	75.1 (74.5)	18.5 (17.9)	1077.2 (488.4)	0.27 (0.12)	5.3 (10.2)	99.58 (99.81)	99.17 (99.63)	
R30CR 10	74.9 (73.9)	18.4 (17.3)	875.4 (341.8)	0.22 (0.09)	6.2 (13.7)	99.66 (99.87)	99.33 (99.74)	
E30Replace	78.5 (79.1)	22.0 (22.6)	9708 (8764)	2.43 (2.19)	1.6 (1.6)	96.53 (96.87)	91.01 (91.87)	
E30Majority Vote								

Table 6: Full results of all experiments on **ObjectNet** under the standard setting of an **improving sequence of models** and **estimating confusion matrices on a separate split** of the data (i.e., not on the target data set) as explained in § 6.1. The first number in each cell refers to estimating only the diagonal elements (i.e., the class-specific accuracies) of confusion matrices, and the **second number (in brackets)** refers to **estimating the full confusion matrices** with smoothing and incorporating soft labels. The character **E** in front of a budget indicates that selection for re-evaluation is based on our posterior label **entropy criterion** (e.g., E30 refers to selecting the top 30% samples with highest entropy), and the character **R** indicates selecting a **randomly** sampled subset without replacement. Note that entropy-based selection requires a posterior and is thus not applicable for the baselines Replace and Majority Vote.

Strategy (Budget %)	Acc (%)	$\Delta$ Acc (%)	$\Sigma$ NF	NFR (%)	PF / NF	Avg. BTC	Avg. BEC
100:Oracle	50.5 (50.5)	42.6 (42.6)	0 (0)	0.0 (0.0)	NaN (NaN)	100.0 (100.0)	99.99 (99.99)
R100:Replace	31.9 (31.9)	24.0 (24.0)	16669 (16669)	5.62 (5.62)	1.3 (1.3)	72.65 (72.65)	92.61 (92.61)
R100:Majority Vote	29.6 (29.6)	21.6 (21.6)	4690 (4690)	1.58 (1.58)	1.9 (1.9)	89.99 (89.99)	98.02 (98.02)
E100:MB	29.1 (29.6)	21.2 (21.7)	2477 (1984)	0.83 (0.67)	2.6 (3.0)	94.46 (95.55)	98.96 (99.16)
E100:MBME	28.6 (29.1)	20.6 (21.1)	1599 (1774)	0.54 (0.6)	3.4 (3.2)	95.86 (95.83)	99.34 (99.26)
E100:CR 2	29.0 (29.2)	21.0 (21.2)	1876 (931)	0.63 (0.31)	3.1 (5.2)	95.92 (97.9)	99.21 (99.61)
E100:CR 5	28.8 (28.7)	20.8 (20.7)	1372 (543)	0.46 (0.18)	3.8 (8.1)	97.18 (98.86)	99.41 (99.77)
E100:CR 10	28.7 (28.4)	20.8 (20.5)	1084 (422)	0.37 (0.14)	4.6 (10.0)	97.82 (99.16)	99.54 (99.82)
R50:Replace	30.7 (30.7)	22.7 (22.7)	7583.8 (7588.8)	2.56 (2.56)	1.6 (1.6)	86.63 (86.61)	96.69 (96.69)
R50:Majority Vote	28.6 (28.7)	20.7 (20.7)	3281.6 (3291.2)	1.11 (1.11)	2.2 (2.2)	93.01 (93.01)	98.62 (98.62)
R50:MB	27.8 (28.7)	19.9 (20.7)	1676.6 (1410.0)	0.56 (0.48)	3.2 (3.7)	96.13 (96.83)	99.3 (99.41)
R50:MBME	26.9 (28.3)	18.9 (20.3)	1280.4 (1360.8)	0.43 (0.46)	3.7 (3.8)	96.77 (96.89)	99.48 (99.43)
R50:CR 2	27.5 (28.1)	19.6 (20.2)	1165.4 (663.6)	0.39 (0.22)	4.1 (6.6)	97.3 (98.51)	99.52 (99.72)
R50:CR 5	27.2 (27.3)	19.2 (19.4)	780.2 (361.4)	0.26 (0.12)	5.6 (10.9)	98.26 (99.23)	99.67 (99.85)
R50:CR 10	27.0 (26.7)	19.1 (18.8)	599.8 (253.0)	0.2 (0.09)	6.9 (14.8)	98.7 (99.47)	99.75 (99.89)
E50:Replace	31.0 (31.4)	23.1 (23.4)	7882 (7329)	2.66 (2.47)	1.5 (1.6)	86.26 (87.93)	96.53 (96.75)
E50:Majority Vote	29.5 (29.4)	21.5 (21.5)	3895 (3397)	1.31 (1.14)	2.0 (2.2)	91.73 (93.32)	98.36 (98.55)
E50:MB	28.9 (30.0)	20.9 (22.1)	2079 (1610)	0.7 (0.54)	2.9 (3.5)	95.23 (96.66)	99.13 (99.31)
E50:MBME	28.2 (29.5)	20.3 (21.5)	1452 (1534)	0.49 (0.52)	3.6 (3.6)	96.18 (96.75)	99.41 (99.35)
E50:CR 2	28.7 (29.6)	20.8 (21.7)	1506 (763)	0.51 (0.26)	3.6 (6.3)	96.64 (98.41)	99.37 (99.67)
E50:CR 5	28.5 (29.0)	20.6 (21.1)	1049 (414)	0.35 (0.14)	4.6 (10.5)	97.83 (99.21)	99.55 (99.82)
E50:CR 10	28.4 (28.7)	20.5 (20.7)	828 (309)	0.28 (0.1)	5.6 (13.4)	98.35 (99.43)	99.64 (99.86)
R30:Replace	29.0 (29.0)	21.0 (21.1)	4070.6 (4074.0)	1.37 (1.37)	2.0 (2.0)	92.19 (92.21)	98.26 (98.26)
R30:Majority Vote	27.3 (27.3)	19.4 (19.4)	2346.6 (2348.8)	0.79 (0.79)	2.5 (2.5)	94.91 (94.88)	99.02 (99.02)
R30:MB	25.9 (27.3)	18.0 (19.4)	1185.8 (1024.8)	0.4 (0.35)	3.8 (4.5)	97.12 (97.58)	99.52 (99.58)
R30:MBME	25.1 (27.1)	17.2 (19.2)	1008.4 (988.6)	0.34 (0.33)	4.2 (4.6)	97.41 (97.67)	99.59 (99.59)
R30:CR 2	25.4 (26.5)	17.5 (18.5)	782.2 (482.8)	0.26 (0.16)	5.2 (8.1)	98.05 (98.87)	99.68 (99.8)
R30:CR 5	24.6 (25.2)	16.6 (17.2)	476.2 (236.2)	0.16 (0.08)	7.5 (14.6)	98.8 (99.45)	99.81 (99.9)
R30:CR 10	24.4 (24.5)	16.4 (16.5)	331.6 (154.2)	0.11 (0.05)	10.2 (20.9)	99.17 (99.65)	99.86 (99.93)
E30:Replace	29.0 (30.0)	21.0 (22.1)	4316 (3761)	1.45 (1.27)	1.9 (2.1)	91.75 (93.51)	98.14 (98.35)
E30:Majority Vote	28.2 (28.2)	20.3 (20.2)	2970 (2359)	1.0 (0.79)	2.3 (2.6)	93.54 (95.37)	98.76 (99.0)
E30:MB	27.8 (28.9)	19.9 (20.9)	1565 (1015)	0.53 (0.34)	3.4 (4.8)	96.24 (97.78)	99.35 (99.57)
E30:MBME	26.9 (28.9)	18.9 (20.9)	1280 (1015)	0.43 (0.34)	3.7 (4.8)	96.64 (97.78)	99.48 (99.57)
E30:CR 2	27.7 (28.1)	19.7 (20.1)	1074 (496)	0.36 (0.17)	4.4 (8.5)	97.42 (98.9)	99.55 (99.79)
E30:CR 5	27.4 (26.9)	19.4 (19.0)	689 (230)	0.23 (0.08)	6.2 (16.3)	98.43 (99.52)	99.71 (99.9)
E30:CR 10	27.1 (25.9)	19.2 (18.0)	504 (140)	0.17 (0.05)	8.1 (24.8)	98.91 (99.71)	99.79 (99.94)
R20:Replace	27.0 (27.1)	19.1 (19.1)	2453.8 (2436.4)	0.83 (0.82)	2.4 (2.5)	94.82 (94.84)	98.97 (98.98)
R20:Majority Vote	25.8 (25.8)	17.8 (17.9)	1658.8 (1663.4)	0.56 (0.56)	3.0 (3.0)	96.17 (96.16)	99.32 (99.32)
R20:MB	23.8 (25.3)	15.8 (17.3)	887.2 (744.0)	0.3 (0.25)	4.3 (5.3)	97.74 (98.15)	99.64 (99.7)
R20:MBME	23.5 (25.5)	15.6 (17.5)	793.8 (736.0)	0.27 (0.25)	4.6 (5.4)	97.93 (98.18)	99.68 (99.7)
R20:CR 2	22.6 (24.3)	14.7 (16.4)	530.0 (339.0)	0.18 (0.11)	6.2 (10.0)	98.57 (99.14)	99.79 (99.86)
R20:CR 5	21.5 (22.7)	13.5 (14.8)	304.2 (160.8)	0.1 (0.05)	9.3 (18.1)	99.14 (99.6)	99.88 (99.93)
R20:CR 10	20.9 (21.5)	12.9 (13.6)	215.6 (98.8)	0.07 (0.03)	12.1 (26.6)	99.4 (99.74)	99.91 (99.96)
E20:Replace	26.9 (29.3)	19.0 (21.3)	2588 (2149)	0.87 (0.72)	2.4 (2.8)	94.54 (96.08)	98.91 (99.07)
E20:Majority Vote	26.2 (27.5)	18.3 (19.6)	2390 (1742)	0.81 (0.59)	2.4 (3.1)	94.8 (96.61)	99.0 (99.26)
E20:MB	26.0 (27.7)	18.1 (19.7)	1169 (763)	0.39 (0.26)	3.9 (5.8)	97.01 (98.26)	99.53 (99.68)
E20:MBME	25.8 (27.7)	17.8 (19.7)	1131 (763)	0.38 (0.26)	3.9 (5.8)	97.06 (98.26)	99.54 (99.68)
E20:CR 2	26.0 (26.6)	18.1 (18.6)	670 (353)	0.23 (0.12)	6.0 (10.8)	98.14 (99.16)	99.73 (99.85)
E20:CR 5	26.0 (25.1)	18.0 (17.2)	369 (132)	0.12 (0.04)	10.1 (25.2)	98.97 (99.7)	99.85 (99.94)
E20:CR 10	25.9 (24.2)	17.9 (16.2)	248 (73)	0.08 (0.02)	14.4 (42.2)	99.34 (99.83)	99.9 (99.97)
R10:Replace	22.1 (22.1)	14.2 (14.2)	996.6 (997.2)	0.34 (0.34)	3.6 (3.6)	97.49 (97.5)	99.6 (99.6)
R10:Majority Vote	21.6 (21.5)	13.6 (13.5)	808.8 (820.4)	0.27 (0.28)	4.1 (4.1)	97.86 (97.83)	99.68 (99.67)
R10:MB	19.5 (21.0)	11.6 (13.1)	450.2 (371.2)	0.15 (0.13)	5.8 (7.6)	98.7 (98.95)	99.82 (99.85)
R10:MBME	19.1 (20.8)	11.1 (12.8)	446.4 (381.6)	0.15 (0.13)	5.6 (7.2)	98.69 (98.92)	99.83 (99.85)
R10:CR 2	17.7 (19.8)	9.8 (11.9)	287.2 (177.2)	0.1 (0.06)	7.3 (13.4)	99.11 (99.48)	99.89 (99.93)
R10:CR 5	15.6 (17.6)	7.7 (9.7)	151.6 (75.4)	0.05 (0.03)	10.4 (24.9)	99.5 (99.77)	99.94 (99.97)
R10:CR 10	14.6 (15.8)	6.7 (7.9)	94.6 (36.0)	0.03 (0.01)	14.1 (41.8)	99.68 (99.89)	99.96 (99.99)
E10:Replace	23.7 (25.8)	15.7 (17.9)	996 (683)	0.34 (0.23)	3.9 (5.9)	97.5 (98.46)	99.6 (99.72)
E10:Majority Vote	23.7 (25.4)	15.7 (17.5)	996 (663)	0.34 (0.22)	3.9 (5.9)	97.5 (98.48)	99.6 (99.72)
E10:MB	22.7 (24.4)	14.8 (16.5)	696 (364)	0.23 (0.12)	4.9 (9.4)	98.08 (99.03)	99.72 (99.85)
E10:MBME	22.7 (24.4)	14.8 (16.5)	696 (364)	0.23 (0.12)	4.9 (9.4)	98.08 (99.03)	99.72 (99.85)
E10:CR 2	20.7 (22.7)	12.8 (14.8)	427 (176)	0.14 (0.06)	6.6 (16.6)	98.68 (99.52)	99.83 (99.93)
E10:CR 5	18.2 (19.4)	10.3 (11.4)	197 (50)	0.07 (0.02)	10.7 (43.4)	99.32 (99.86)	99.92 (99.98)
E10:CR 10	17.2 (15.9)	9.2 (7.9)	122 (28)	0.04 (0.01)	15.0 (53.5)	99.57 (99.92)	99.95 (99.99)