

Multi-Teacher Knowledge Distillation via Teacher-Informed Mixture Priors

Luyang Fang¹, Yongkai Chen², Jiazhang Cai¹, Ping Ma^{1,*}, Wenxuan Zhong^{1,*}

¹Department of Statistics, University of Georgia

²Department of Statistics, Harvard University

*Corresponding authors

Abstract

Knowledge distillation is a powerful method for model compression, enabling the efficient deployment of complex deep learning models (teachers), including large language models. However, its underlying statistical mechanisms remain unclear, and uncertainty evaluation is often overlooked, especially in real-world scenarios requiring diverse teacher expertise. To address these challenges, we introduce *Multi-Teacher Bayesian Knowledge Distillation* (MT-BKD), where a distilled student model learns from multiple teachers within the Bayesian framework. Our approach leverages Bayesian inference to capture inherent uncertainty in the distillation process. We introduce a teacher-informed prior, integrating external knowledge from teacher models and task-specific training data, offering better generalization, robustness, and scalability. Additionally, an entropy-based weighting mechanism adaptively adjusts each teacher’s influence, allowing the student to combine multiple sources of expertise effectively. MT-BKD enhances the interpretability of the student model’s learning process, improves predictive accuracy, and provides uncertainty quantification. We validate MT-BKD on both synthetic and real-world tasks, including protein subcellular location prediction and image classification. Our experiments show improved performance and robust uncertainty quantification, highlighting the strengths of our MT-BKD framework.

Keywords: Uncertainty Quantification; Bayesian Priors; Large Language Models; Protein Subcellular Prediction; Image Classification

1 Introduction

The rapid advancements in deep learning have led to the development of increasingly complex models with billions of parameters, achieving state-of-the-art performance across numerous tasks (Kondratyuk et al., 2023; Dosovitskiy et al., 2020). However, deploying these large models in real-world applications is often infeasible due to significant computational and storage demands. Knowledge distillation (KD) offers a powerful solution by transferring knowledge from a large, pre-trained *teacher* model to a smaller, more efficient *student* model (Zhao et al., 2022; Hinton et al., 2015; Latif et al., 2023). By training the student to replicate the teacher’s predictions, KD produces lightweight models with competitive performance, suitable for resource-constrained environments (Gou et al., 2021). This approach is particularly prevalent in the development of large language models, where vendors routinely release both full-scale and lightweight versions, with the latter often trained using KD (Touvron et al., 2024).

Many real-world scenarios involve learning from multiple teacher models with diverse expertise. For instance, a model trained on structured data may provide insights different from one specialized in unstructured data. Multi-teacher KD extends KD by leveraging this diversity to improve student learning. Recent research addresses this need with data-dependent weighting schemes that balance each teacher’s influence, enabling the student to capture multiple teachers’ strengths (You et al., 2017; Zhao et al., 2024).

Despite the empirical success of Multi-teacher KD, significant challenges remain. First, the statistical insights of KD and their effects on student-model improvement remain unclear. Although some studies (Phuong and Lampert, 2019; Menon et al., 2021) have explored its impact, they primarily evaluate improvements in predictive performance. A deeper investigation into the statistical and theoretical aspects is crucial for understanding

its full potential. Second, most existing methods prioritize improving prediction accuracy but lack quantification of uncertainty in its prediction. This limitation reduces their ability to assess the reliability of the student model’s predictions, restricting their applicability in high-stakes domains where interpretability and robustness are crucial. Moreover, current approaches often treat teachers independently, combining their outputs through weighted averages without considering potential interactions among the teachers. This simplification can lead to suboptimal knowledge transfer, preventing the student from fully leveraging the combined expertise of multiple teachers.

To address these challenges, we propose a Bayesian framework for multi-teacher knowledge distillation, introducing a teacher-informed mixture prior that integrates guidance from multiple teachers into one probabilistically coherent model. By integrating both teacher knowledge and task-specific data, this prior captures diverse external expertise while remaining closely aligned with the task. This integration of knowledge offers better generalization and robustness than purely data-driven priors used in classical Bayesian analysis, especially in the case of limited training data. The mixture prior also reflects teacher interactions, allowing the data to favor the teachers that best explain the observations naturally. Additionally, we employ an entropy-based weighting scheme that dynamically adjusts each teacher’s influence on a per-sample basis, ensuring that the most knowledgeable teacher has the greatest impact. Beyond improving the distillation process, our Bayesian formulation naturally equips the student model with uncertainty quantification capabilities. By sampling from the student’s posterior via stochastic gradient Langevin dynamics ([Welling and Teh, 2011](#)), we derive robust uncertainty measures and estimate credible intervals. These uncertainty estimates improve interpretability by quantifying the model’s confidence in its predictions, which is crucial for reliable decision-making in critical applications.

Our contributions are summarized as follows:

- **Uncertainty Quantification:** We develop Bayesian inference tools for the student model to provide uncertainty estimates that improve interpretability and enable robust decision-making.
- **Teacher-informed Prior:** We introduce a Bayesian KD framework that integrates the external knowledge of teacher models and training data to provide the prior, offering better generalization, robustness, and scalability.
- **Unified Multi-teacher Framework:** We leverage a mixture distribution to combine the knowledge of multiple teachers, which captures their interactions, and ensures effective knowledge transfer. Our entropy-driven, sample-specific weighting mechanism prioritizes important teachers for each data point, optimizing the student’s learning process.
- **Validation:** We demonstrate the effectiveness of our approach using synthetic and real-world datasets, highlighting its ability to enhance both predictive performance and uncertainty quantification.

2 Related Research

Knowledge distillation (KD) has been widely studied for its effectiveness and limitations. The original work by [Hinton et al. \(2015\)](#) transfers knowledge by minimizing the KL divergence between teacher and student outputs. Later studies explored strategies like mimicking intermediate-layer representations ([Zagoruyko and Komodakis, 2016](#); [Phuong and Lampert, 2019](#); [Shen et al., 2022](#)). To leverage diverse teacher expertise, researchers have proposed multi-teacher KD strategies. ([You et al., 2017](#); [Zhang et al., 2022](#); [Liu et al., 2020](#); [Chen et al., 2022](#); [Gui et al., 2023](#); [Zhao et al., 2024](#)). For instance, [You et al. \(2017\)](#) align

intermediate-layer representations across teachers using relative dissimilarity constraints and aggregate their guidance through a voting mechanism. Zhao et al. (2024) combine multi-teacher KD with adversarial training by employing specialized teachers for clean and adversarial data, enhancing both accuracy and robustness. However, existing multi-teacher KD approaches predominantly focus on improving predictive accuracy and often overlook interactions between teachers or the uncertainty in the student model’s predictions.

Efforts to incorporate Bayesian principles into KD have led to methods such as *Bayesian Dark Knowledge* (Korattikara Balan et al., 2015), which distills the teacher model’s posterior predictive distribution into a compact student model, enabling uncertainty quantification. Subsequent extensions (Wang et al., 2018; Malinin et al., 2019; Vadera et al., 2020) refine this framework, but these methods require access to the teacher model’s posterior predictive distribution, which is often impractical for large-scale pre-trained models.

Formulating data-based priors has a long history in Bayesian statistics (Robbins, 1992). Some studies incorporate information from previous work to construct priors (Ibrahim et al., 2015). In particular, the power prior is derived by raising the likelihood of historical data to a specified power. However, its properness requires the covariate matrix of either historical or current data to have full column rank (Chen et al., 2000). Another approach is to use synthetic data to define the prior distribution by training a simpler model and generating samples from its predictive distribution (Huang et al., 2020; Bernardo, 1979). In cases with limited expert knowledge, this method is effective. However, with powerful teacher models, MT-BKD can integrate both expert insights and current data to create the prior.

3 Preliminaries

We begin by introducing the preliminaries of deep neural network models, followed by an overview of knowledge distillation and its extension to multi-teacher knowledge distillation.

Deep Neural Network Model. For the convenience of presentation, we introduce our method in the context of classification problems. In the classification problems, we are given the training sample $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \Omega_{\mathbf{x}} \subset \mathbb{R}^m$ is the predictor and $\mathbf{y}_i = (y_{i1}, \dots, y_{iK})^T$ is the corresponding label represented in the form of indicator vector. Specifically, if the i^{th} data point belongs to class k , we have $y_{ik} = 1$ and $y_{ij} = 0$ for all $j \neq k$, where $i = 1, \dots, N, j = 1, \dots, K$. Each data point is independently collected across $i \in \{1, \dots, N\}$. Our goal is to find a function $\mathbf{h} : \mathbb{R}^m \rightarrow [0, 1]^K$ from a function class that approximates the conditional probability $\mathbb{P}(y_{ik} = 1 \mid \mathbf{x}_i)$ for each $k \in \{1, \dots, K\}$. Specifically, we seek a function $\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}) = (h_1(\mathbf{x}_i; \boldsymbol{\theta}), \dots, h_K(\mathbf{x}_i; \boldsymbol{\theta}))^T$, where $h_k(\mathbf{x}_i; \boldsymbol{\theta})$ corresponds to the approximation of $\mathbb{P}(y_{ik} = 1 \mid \mathbf{x}_i)$. The function class can be very general, ranging from logistic regressions (Faraway, 2016), nonlinear regressions (Bates and Watts, 1988), to deep neural networks (Goodfellow et al., 2016). In this article, we focus on finding $\mathbf{h}$ in the function class of deep neural networks, including feed-forward networks, convolutional neural nets, transformers, etc, and please see Zhang et al. (2021) for more details. For example, a feed-forward network utilizes functions from the following compositional function class:

$$\mathbf{h}(\mathbf{x}; \boldsymbol{\theta}) = \boldsymbol{\sigma}_L(\mathbf{W}_L \boldsymbol{\sigma}_{L-1}(\mathbf{W}_{L-1} \cdots \boldsymbol{\sigma}_2(\mathbf{W}_2 \boldsymbol{\sigma}_1(\mathbf{W}_1 \mathbf{x})))), \quad (1)$$

where $\boldsymbol{\theta} = \{\mathbf{W}_1, \dots, \mathbf{W}_L\}$ are the parameters of the model, and $\boldsymbol{\sigma}_l, l = 1, \dots, L$ are some fixed nonlinear activation functions. A popular choice of $\boldsymbol{\sigma}_l(\cdot)$ for $l = 1, \dots, L - 1$ is the ReLU (Rectified Linear Unit) function, which sets negative values to zero while keeping positive values unchanged, applied element-wise in multivariate settings. In classification

problems, for the final layer L , researchers often use a soft-max function (Fan et al., 2020).

As $\mathbf{h}(\cdot; \cdot)$ is determined by the neural network structure, the unknown parameter is $\boldsymbol{\theta}$.

To estimate $\boldsymbol{\theta}$, we minimize the empirical risk:

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{1}{N} \sum_{i=1}^N \text{CE}(\mathbf{y}_i, \mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})), \quad (2)$$

where $\text{CE}(\mathbf{y}_i, \mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})) = -\sum_{k=1}^K y_{ik} \log(h_k(\mathbf{x}; \boldsymbol{\theta}))$ is the cross-entropy loss between the probability vector $\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})$ and $\mathbf{y}_i$.

Knowledge Distillation and its Multi-Teacher Variant. Knowledge distillation (KD), introduced by Hinton et al. (2015), is a procedure where a simpler model (student) learns from a larger model (teacher). In the classical setup of KD, a pre-trained complex deep neural network model serves as the teacher. For each data point i in the training set $\mathcal{D}$, the teacher model M_t outputs a probability vector $\mathbf{p}_i = (p_{i1}, \dots, p_{iK})^T$, where p_{ij} represents the predicted class probability that $\mathbf{x}_i$ belongs to class j . Mathematically, p_{ij} approximates $\mathbb{P}(y_{ik} = 1 \mid \mathbf{x}_i)$ according to model M_t . In this work, teacher models are treated as closed-source models, i.e., their structure and parameters are unknown. This is common in practice as many popular pre-trained large language models, e.g., ChatGPT and Claudia, are closed-source.

In contrast, the student model M_s is transparent, defined by a deep neural network with a known structure $\mathbf{h}(\cdot; \cdot)$ and unknown parameter $\boldsymbol{\theta}$. The student model M_s is trained using both the training sample $\mathcal{D}$ and the corresponding teacher model’s predicted class probabilities $\mathbf{p} = \{\mathbf{p}_i\}_{i=1, \dots, N}$. The student model M_s outputs a probability vector $\mathbf{q}_i = (q_{i1}, \dots, q_{iK})^T$ for data point i , where $\mathbf{q}_i = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})$ and q_{ij} represents the predicted class probability that $\mathbf{x}_i$ belongs to class j . KD measures the discrepancy between the student

and teacher models with

$$\tilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}) = \frac{1}{N} \sum_{i=1}^N \text{CE}(\mathbf{p}_i, \mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K p_{ik} \log(h_k(\mathbf{x}_i; \boldsymbol{\theta})), \quad (3)$$

which is the sample mean of the cross-entropy $\text{CE}(\mathbf{p}_i, \mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}))$ across the training data. To leverage the information from both the training data and the teacher model’s predictions, KD estimates the parameter $\boldsymbol{\theta}$ by minimizing the objective function,

$$\mathcal{L}^{\text{KD}}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}, \lambda) = \mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) + \lambda \tilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}), \quad (4)$$

where $\mathcal{L}^{\text{KD}}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}, \lambda)$ combines the losses in Equation (2) and Equation (3) with a tuning parameter $\lambda \in (0, \infty)$ to balance their contributions.

In practice, a student can learn from multiple teachers, with the collective guidance of these teachers providing a comprehensive framework beneficial for training the student network. Consider G pre-trained teacher models, denoted by $M_t^{(g)}$ for $g = 1, \dots, G$, we have the predicted class probabilities, denoted by $\mathbf{p}^{(g)} = \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}$, where $\mathbf{p}_i^{(g)} = (p_{i1}^{(g)}, \dots, p_{iK}^{(g)})^T = M_t^{(g)}(\mathbf{x}_i)$. A natural way to formulate the multi-teacher KD loss is:

$$\mathcal{L}^{\text{Multi-KD}}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}, \lambda) = \mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) + \lambda \sum_{g=1}^G w^{(g)} \tilde{\mathcal{L}}^{(g)}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}^{(g)}), \quad (5)$$

where $w^{(g)} \in (0, \infty)$ is the weight assigned to $M_t^{(g)}$, $\sum_{g=1}^G w^{(g)} = 1$, and $\tilde{\mathcal{L}}^{(g)}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}^{(g)})$ is analogously defined as in Equation (3). Extensions, including assigning sample-wise weights and distilling intermediate-layer representations, have also been explored (You et al., 2017; Fukuda et al., 2017; Wu et al., 2019).

While existing research has advanced our understanding of KD and its extensions to

multi-teacher scenarios, significant gaps remain in capturing interactions among teachers and providing uncertainty quantification for the model predictions. In this work, we bridge these gaps by introducing a Bayesian framework for multi-teacher KD that explicitly models teacher-student interactions through a mixture prior, enabling both robust knowledge integration and uncertainty-aware learning. We provide details in the following section.

4 Bayesian Knowledge Distillation

Unlike the previous formulations of multi-teacher KD, in this work, we develop multi-teacher KD in a Bayesian paradigm, which naturally enables uncertainty quantifications.

4.1 Bayesian Model with a Teacher-Informed Prior

In our proposed Bayesian model, since each observation belongs to exactly one class, we assume that the class label $\mathbf{y}_i$ follows a multinomial distribution given the predictor $\mathbf{x}_i$,

$$\mathbf{y}_i|\mathbf{x}_i \sim \text{Multinomial}(1, \mathbf{q}_i), \quad (6)$$

where $i = 1, \dots, N$, and the multinomial probability is provided by a deep neural network, i.e., $\mathbf{q}_i = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})$.

In the Bayesian paradigm, parameter $\boldsymbol{\theta}$ is assumed to be random. Our goal is to develop a proper prior distribution. A straightforward approach might be to directly place a prior on $\boldsymbol{\theta}$ as commonly did in Bayesian generalized linear models, see Chapter 14 in [Gelman et al. \(2013\)](#). However, in our study, the teacher models are close-source, meaning the teacher models' architectures and parameters are inaccessible, with only access to their outputs. Even in the case that the teacher models are open-source, since the teacher and student

models differ in structure, the teacher models' knowledge cannot be directly encoded into a prior for the parameters of the student model. Without any further information, the only option for us to impose objective (aka noninformative) prior for $\boldsymbol{\theta}$, which is not satisfiable. As an alternative, we propose using the teacher models' predicted class probabilities to build the prior, referred to as the teacher-informed prior (TIP), for $\boldsymbol{\theta}$. We set the following mixture distribution (McLachlan and Peel, 2000) as our TIP for $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{\theta} \mid \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}^{g=1, \dots, G}) \propto \prod_{i=1}^N \sum_{g=1}^G w_i^{(g)} f(\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}) \mid \mathbf{p}_i^{(g)}), \quad (7)$$

where $\propto$ denotes proportionality, G is the number of mixture components, $w_i^{(g)}$ are the weights assigned to the component g on data point $\mathbf{x}_i$, $f(\cdot \mid \mathbf{p}_i^{(g)})$ are probability density functions. In this work, we consider $f(\cdot \mid \mathbf{p}_i^{(g)})$ is the density function of a Dirichlet distribution $Dir(\mathbf{1}_K + \lambda \mathbf{p}_i^{(g)})$, where λ is a tuning parameter weighting the contribution of teacher model predictions. Hence, we have

$$f(\mathbf{q}_i \mid \mathbf{p}_i^{(g)}) = \frac{1}{B(\mathbf{1}_K + \lambda \mathbf{p}_i^{(g)})} \prod_{k=1}^K (q_{ik})^{\lambda p_{ik}^{(g)}}, \quad (8)$$

where $B(\cdot)$ is the multivariate Beta function. Note that the mode of $f(\cdot \mid \mathbf{p}_i^{(g)})$ is $\mathbf{p}_i^{(g)}$ for $\lambda > 0$. Furthermore, we show that $\pi(\boldsymbol{\theta} \mid \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}^{g=1, \dots, G})$ is a proper prior with the following proposition. We relegate the proof to Appendix C.

Proposition 4.1. *Consider the probability density function $f(\mathbf{q} \mid \mathbf{p}_i^{(g)})$ as defined in Equation (8) with a constant $\lambda > 0$. Assuming that the parameters of the student model lie in a compact space, then $\pi(\boldsymbol{\theta} \mid \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}^{g=1, \dots, G})$ is a proper prior.*

According to our best knowledge, the proposed TIP is novel in the statistical literature. The proposed TIP has four distinguishing features: (1) Broad knowledge integration. It

is a prior integration of the information from both teacher models and task-specific training data. The teacher models are typically pre-trained using extraneous data, and our training data are fed into the teacher models to yield predicted probabilities. Therefore, it captures extensive external knowledge while maintaining direct relevance to the task. (2) Statistical robustness. It reduces overfitting risk by incorporating external information, making it more robust than purely data-driven priors used in classical Bayesian analysis. This is particularly advantageous with limited training data. (3) Adaptability. It offers adaptability by incorporating multiple teacher models with automatic weight adjustment based on model quality, and the prior can be updated when better teacher models become available. (4) Practically user-friendly. It relies less on expert elicitation and is scalable to different domains, including texts, images, and numerical data. The key innovation is the principled integration of external knowledge (LLMs) with empirical data, offering better generalization than either approach alone.

In the general case with multiple teachers, the posterior density of $\boldsymbol{\theta}$ is,

$$\pi(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}, \lambda, \mathbf{h}(\cdot; \cdot)) \propto \exp \left\{ \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log(h_k(\mathbf{x}_i; \boldsymbol{\theta})) + \sum_{i=1}^N \log \left[\sum_{g=1}^G w_i^{(g)} \frac{1}{B(\mathbf{1}_k + \lambda \mathbf{p}_i^{(g)})} \prod_{k=1}^K (h_k(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}} \right] \right\}. \quad (9)$$

Please note that when $G = 1$, Equation (9) reduces to a single-teacher case (Fang et al., 2024). We present the discussion of this special case in Appendix B.1.

Applying the mixture prior defined in Equation (7) leads to improved results compared to the traditional multi-teacher knowledge distillation (KD) strategy described in Equation (5). Specifically, the traditional multi-teacher KD strategy is equivalent to independently deriving posterior distributions from each teacher model and combining them using logarithmic opinion pooling (Genest et al., 1986; Garthwaite et al., 2005). We establish the

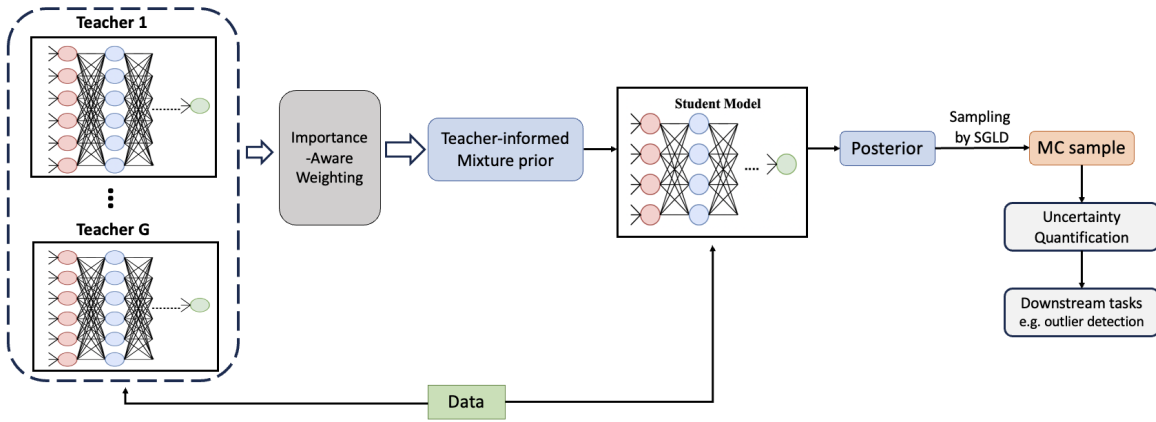


Figure 1: **The multiple teacher Bayesian knowledge distillation (MT-BKD) framework.** A teacher-informed prior is established for the student model’s parameters based on the predicted probabilities from multiple teacher models, and the posterior distribution is derived. An importance-aware weighting mechanism balances contributions from the teachers. The stochastic Gradient Langevin Dynamics (SGLD) method is then applied to generate Monte Carlo samples from the posterior. Uncertainty quantification of the predictions and the subsequent downstream tasks, such as outlier detection, can be achieved accordingly.

following theorem, with the proof presented in Appendix C.

Theorem 4.2. *Given the prior distribution defined in Equation (7) and Equation (8), the traditional multi-teacher KD strategy in Equation (5) is equivalent to deriving each teacher model’s posterior independently and combining them using logarithmic opinion pooling.*

Traditional multi-teacher KD algorithms typically provide only point estimates, overlooking the uncertainty in model parameters. In contrast, our proposed MT-BKD method estimates the full posterior distribution, equipping the student model with uncertainty quantification capabilities. This is particularly important when the posterior distribution is multimodal. Point estimates in such cases may fall between modes, failing to capture the true distribution and potentially leading to misleading conclusions. In addition, we employ a mixture prior approach in MT-BKD. We begin with a mixture prior with fixed weights representing our initial beliefs about each component. Once data is observed, we compute the posterior distribution based on this mixture prior. Although the initial prior

weights remain fixed, the posterior naturally shifts in favor of components that better explain the data, balancing our assumptions with empirical evidence. If multiple priors fit the data well, the posterior may be multimodal, reflecting multiple plausible parameter configurations. By maintaining a mixture of priors, we preserve these distinct possibilities while allowing the data to emphasize those that are most consistent. Conversely, combining individual posteriors via logarithmic opinion pooling first updates each prior with data to obtain individual posteriors and then combines them via a weighted geometric mean of their log-transformed values. As a result, each prior retains its predetermined weight in the final combined posterior, regardless of its performance on the data.

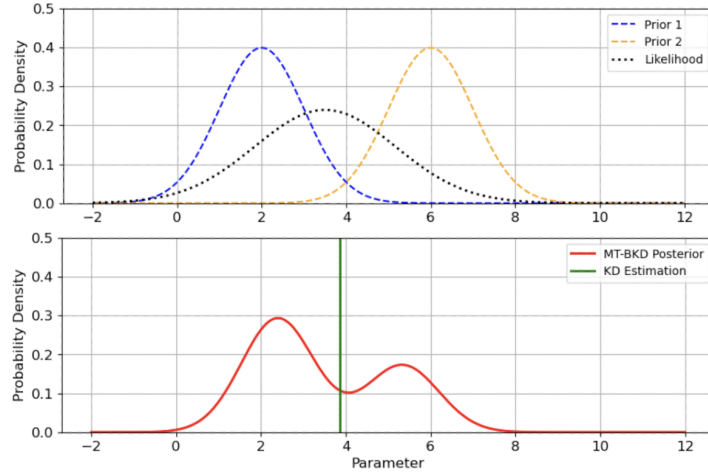


Figure 2: Comparison of posterior distributions obtained through MT-BKD and the estimate of multi-teacher KD. The figure illustrates the prior distributions, the data distribution, the resulting posterior distributions derived from MT-BKD (red line), and the estimation achieved through multi-teacher KD (green line).

We illustrate the advantage of MT-BKD using a toy example with two Gaussian priors centered at $\mu_1 = 2.0$ and $\mu_2 = 6.0$ ($\sigma^2 = 1.0$). The observed data has a mean of 3.5, which lies between these priors, indicating no strong preference for either. Figure 2 shows that MTBKD (red line) yields a bimodal posterior, accurately reflecting the balanced support for both priors while slightly favoring the one better supported by the data. In contrast, traditional multi-teacher KD (green line) collapses to a point estimate, overlooking the

bimodal nature implied by the priors and the data.

4.2 Importance-Aware Distillation

Equation (7) highlights the critical role of the weights $w_i^{(g)}$ in distilling knowledge from multiple teacher models into a unified student model. Since teacher networks vary in their effectiveness across different data, identifying and leveraging these differences is essential for successfully integrating multiple teachers. We define $w_i^{(g)}$ for the teacher model $M_t^{(g)}$ as

$$w_i^{(g)} = \frac{\xi(H(\mathbf{p}_i^{(g)}))}{\sum_{g=1}^G \xi(H(\mathbf{p}_i^{(g)}))}, \quad (10)$$

where $H(\cdot)$ represents the entropy, calculated as $H(\mathbf{p}_i^{(g)}) = -\sum_k p_{ik}^{(g)} \log(p_{ik}^{(g)})$, and ξ is a monotonically decreasing function. Intuitively, from the perspective of information theory, information entropy is used to describe the degree of randomness and uncertainty of information, so it is reasonable to apply information entropy to quantify the knowledge scale of the teacher model on the given data point. Thus, we assign higher distillation weights to teacher models with lower entropy predictions. This allows the student model to learn more strongly from the teachers that have developed greater expertise on each specific data point, leading to more effective knowledge transfer.

We now discuss the effect of λ . In the extreme case of $\lambda = 0$, $f(\mathbf{q} | \mathbf{p}_i^{(g)})$ becomes the density of a symmetric Dirichlet distribution, $Dir(\mathbf{1}_K)$. This implies that there is no prior knowledge favoring one category over another when classifying each data point. For the case of $\lambda \rightarrow \infty$, we derive the following proposition with proof in Appendix C,

Proposition 4.3. *Consider the probability density function $f(\mathbf{q} | \mathbf{p}_i^{(g)})$ as defined in Equation (8), as $\lambda \rightarrow \infty$, we have $f(\mathbf{q} | \mathbf{p}_i^{(g)}) \rightarrow \delta(\mathbf{q} - \mathbf{p}_i^{(g)})$, where $\delta(\cdot)$ is the multivariate*

Dirac delta function.

The above proposition shows that, as λ approaches infinity, the student model assigns nonzero probability only when its predicted class probabilities exactly match the teacher’s, aligning with original knowledge distillation (KD) that fully trusts the teacher’s probabilities at infinitely large λ . In summary, a large λ indicates strong reliance on the teacher’s predictions, while a small λ suggests limited teacher influence. In practice, λ can be determined by selecting candidate values based on domain expertise and selecting the best one through cross-validation (Shao, 1993).

5 Uncertainty Quantification via Bayesian Predictive Inference

In practice, it is very common to use the trained student model to analyze confidential data. Analyzing confidential data using pre-trained teacher models as presented in the previous section may result in information leaking (Ray, 2023). In this section, we develop a method to quantify prediction uncertainty using the trained student model alone.

5.1 Measurement of Prediction Uncertainty

For a new data point $\tilde{\mathbf{x}}$, we denote the trained student model’s prediction as $\tilde{\mathbf{q}} = \mathbf{h}(\tilde{\mathbf{x}}; \boldsymbol{\theta})$. Suppose that the label of $\tilde{\mathbf{x}}$ is $\tilde{\mathbf{y}}$. Instead of estimating the predictive posterior distribution (Gelman et al., 1995) of the unobserved label $\tilde{\mathbf{y}}$, we introduce a single-valued metric to gauge our confidence in the model’s prediction.

A commonly used confidence measure for classification is the deviance (Spiegelhalter et al., 2014), defined as $\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}}) = -2 \sum_{k=1}^K \tilde{y}_k \log(\tilde{q}_k)$, where $\tilde{y}_k$ and $\tilde{q}_k$ are the k th el-

ements of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{q}}$. This deviance can be interpreted as a generalized sum of squared residuals for $\tilde{\mathbf{q}}$. A higher deviance value suggests greater uncertainty in the model’s predictions. To make predictive inferences about $\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}})$, we use its posterior predictive distribution. Under the multinomial model in Equation (6) and the posterior distribution of $\boldsymbol{\theta}$ in Equation (9), we have

$$\pi(\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}}) | \mathcal{D}, \mathbf{p}, \lambda, \mathbf{h}(\cdot; \cdot)) = \int \pi(\boldsymbol{\theta} | \mathcal{D}, \mathbf{p}, \lambda, \mathbf{h}(\cdot; \cdot)) p(\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}}) | \boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (11)$$

where $p(\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}}) | \boldsymbol{\theta})$ is the conditional probability density function of $\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}})$ given $\boldsymbol{\theta}$. Finally, the posterior mean deviance is used to quantify the model’s prediction uncertainty.

5.2 Posterior Sampling

For the high dimension of the parameter $\boldsymbol{\theta}$ and the complex structure of $\mathbf{h}(\cdot; \boldsymbol{\theta})$, it may not be feasible to obtain an analytical solution for Equation (11). Hence, we apply the stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011), a variant of Langevin Monte Carlo, to efficiently sample from the complex, high-dimensional posterior distributions. SGLD achieves this by seamlessly integrating the general approach of stochastic gradient descent with Langevin dynamics.

Suppose $\boldsymbol{\theta}^{(j-1)}$ is sampled in iteration $j - 1$. Then in the j -th step, given a mini-batch of m data points $\mathcal{D}^{(j)} = \{(\mathbf{x}_i^{(j)}, \mathbf{y}_i^{(j)})\}_{i=1, \dots, m}$ and the class probability $\{\mathbf{p}_i^{(g)(j)}\}_{i=1, \dots, m}^{g=1, \dots, G}$ predicted by the teacher model, SGLD generates the sample from posterior using gradient updates plus Gaussian noise,

$$\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)} + \tau \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^{(j-1)} | \mathcal{D}^{(j)}, \{\mathbf{p}_i^{(g)(j)}\}_{i=1, \dots, m}^{g=1, \dots, G}, \lambda, \mathbf{h}) + \sqrt{2\tau} \boldsymbol{\xi}^{(j)},$$

Algorithm 1 Multi-Teacher Bayesian Knowledge Distillation (MT-BKD).

- Input:** $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N, \mathbf{h}(\cdot; \cdot), \tau, \lambda, r$. New dataset $\mathcal{T} = \{(\tilde{\mathbf{x}}_s, \tilde{\mathbf{y}}_s)\}_{s=1}^S$
1. Get the output $\mathbf{p}^{(g)}$, $g = 1, \dots, G$, of each teacher model for each data point in $\mathcal{D}$.
 2. Build the mix prior as in Equation (7) using weights defined in Equation (10).
 3. Calculate the posterior distribution as in Equation (9).
 4. Generate Monte Carlo sample of $\boldsymbol{\theta}$:
 - At iteration j^{th} with a subset of m data points $\mathcal{D}^{(j)} = \{(\mathbf{x}_i^{(j)}, \mathbf{y}_i^{(j)})\}_{i=1}^m$,
 - Sample $\xi^{(j)} \sim N(0, I)$,
 - Sample $\boldsymbol{\theta}^{(j)}$ using SGLD.
 5. Predict the uncertainty $\{\tilde{\Delta}_s\}_{s=1}^S$ for each data point in $\mathcal{T}$ using Equation (12).
 6. Construct credible intervals $\{[0, \tau_s]\}_{s=1}^S$ for the deviance of the prediction on $\mathcal{T}$.
- Output:** $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^r$, $\{\tilde{\Delta}_s\}_{s=1}^S$, and $\{[0, \tau_s]\}_{s=1}^S$.
-

where τ is the step size, and $\xi^{(j)}$ is randomly sampled from $N(0, I)$. SGD optimizes the log-likelihood to guide the sampling process toward regions of higher probability, while Langevin dynamics introduces controlled noise into the parameter updates, ensuring convergence to the full posterior distribution rather than merely its mode. Specifically, in the limit of $j \rightarrow \infty$ and $\tau \rightarrow 0$, the probability distribution of $\boldsymbol{\theta}^{(j)}$, denoted as $\rho^{(j)}$, converges to the posterior distribution of $\boldsymbol{\theta}$. By using gradient information and introducing controlled noise, SGLD becomes more efficient in handling high-dimensional data (Girolami and Calderhead, 2011). Furthermore, SGLD’s use of mini-batches for gradient computation alleviates the computational burden associated with optimizing over entire datasets, making it particularly well-suited for large-scale datasets.

Since the distribution of MC sample $\boldsymbol{\theta}^{(j)}$ converges to the posterior distribution of $\boldsymbol{\theta}$ as $j \rightarrow \infty$, it allows for a precise estimation of the characteristics of $\boldsymbol{\theta}$. Consequently, it fosters a comprehensive analysis and understanding of our model covering various aspects. One aspect that we are particularly interested in is the model’s prediction on new data.

5.3 Posterior Mean Deviance and Credible Interval

With the posterior predictive distribution of the deviance in Equation (11), we propose the posterior mean deviance as a measure of predictive uncertainty and provide the credible interval of deviance.

With the MC sample $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^r$, we can estimate the posterior mean of $\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}})$ by,

$$\begin{aligned}\tilde{\Delta} &= \frac{1}{r} \sum_{j=1}^r E\left(\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}}) \mid \boldsymbol{\theta}^{(j)}\right) \\ &= -\frac{2}{r} \sum_{j=1}^r \sum_{k=1}^K \tilde{q}_k^{(j)} \log(\tilde{q}_k^{(j)}),\end{aligned}$$

where $\tilde{q}_k^{(j)}$ is the k th element of $\tilde{\mathbf{q}}^{(j)} = \mathbf{h}(\tilde{\mathbf{x}}; \boldsymbol{\theta}^{(j)})$.

Now, we construct the credible interval for the deviance of the prediction. For the deviance $\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}})$, its $1 - \alpha$ credible interval, denoted by $CI = [0, \tau]$, satisfies

$$E(\mathbf{1}_{\{\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}}) \leq \tau\}}) = 1 - \alpha, \quad (12)$$

where α is the credible level. We can empirically estimate τ by minimizing $L(\tau)$ where

$$\begin{aligned}L(\tau) &= \left| \frac{1}{r} \sum_{j=1}^r E(\mathbf{1}_{\{\text{dev}(\tilde{\mathbf{y}}, \tilde{\mathbf{q}}) \leq \tau\}} \mid \boldsymbol{\theta}^{(j)}) - (1 - \alpha) \right| \\ &= \left| \frac{1}{r} \sum_{j=1}^r \sum_{k=1}^K \tilde{q}_k^{(j)} \mathbf{1}_{\{-2 \log(\tilde{q}_k^{(j)}) \leq \tau\}} - (1 - \alpha) \right|. \quad (13)\end{aligned}$$

The function $L(\tau)$ denotes the difference between the left-hand side and the right-hand side of Equation (12).

Constructing a credible interval on deviance, rather than directly on the prediction $\mathbf{q}$, presents a notable advantage. It enables the calculation of the coverage rate of the constructed interval using the true label. This is beneficial since, in most cases, the true

labels of the testing dataset are readily available for evaluation, whereas the true class probabilities are typically elusive. The MT-BKD algorithm is summarized in Algorithm 1.

6 Theoretical Analysis

Here, we develop an asymptotic convergence of our proposed MT-BKD estimate. Without the loss of generality, we only consider the binary classification problem here. Let μ represent the predictor $\mathbf{x}$'s true density function in the training sample of the student model while $\mu^{(g)}$ represents the predictor's true density function in the training sample of the g th teacher model for $g = 1, \dots, G$. Suppose the true parameter for the student model defined by Equation (1) and (6) is $\boldsymbol{\theta}_0$. To define a good estimate, we consider the excess ϕ -risk with the log loss,

$$\mathcal{E}^\phi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = \mathcal{E}^\phi(\hat{\boldsymbol{\theta}}_n) - \mathcal{E}^\phi(\boldsymbol{\theta}_0), \quad (14)$$

where $\mathcal{E}^\phi(\boldsymbol{\theta}) = E_{\mathbf{x}, \mathbf{y}}[\phi(\mathbf{y}^T \text{logit}[\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})])]$, and $\phi(x) = \log(1 + \exp(-x))$.

We define the quadratic functional $V(f) = \int_{\Omega_{\mathbf{x}}} f^2 \mu(\mathbf{x}) d\mathbf{x}$.

The following conditions are assumed to hold.

(A1) For any $\mathbf{x} \in \Omega_{\mathbf{x}}$, $\mu(\mathbf{x}) > 0$, $\max_{1 \leq g \leq G} \mu^{(g)}(\mathbf{x}) > 0$.

(A2) The minimizer $\boldsymbol{\theta}_n^*$ in Equation (2) yield a consistent estimate of $\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$, i.e., the error $\int_{\Omega_{\mathbf{x}}} [\mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_n^*) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]^2 \mu(\mathbf{x}) d\mathbf{x} = O_p(C_1(n))$ and $\lim_{n \rightarrow \infty} C_1(n) = 0$.

(A3) There exist a positive constant $\varepsilon^{(g)}$ such that $\int_{\Omega_{\mathbf{x}}} [M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]^2 \mu^{(g)}(\mathbf{x}) d\mathbf{x} \leq \varepsilon^{(g)}$ for $g = 1, \dots, G$.

Condition (A1) essentially requires that for any $\mathbf{x} \in \Omega_{\mathbf{x}}$, there exists at least one of the training samples' population distributions corresponding to the teacher models that assign a nonzero density to it. Condition (A2) can be proven under several regularity

conditions, e.g., $\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$ is a (p, C) -smooth generalized hierarchical interaction model (Bauer and Kohler, 2019). Without loss of generality, we use $C_1(n)$ to represent its convergence rate while the detailed forms under different scenarios can be found in the existing literature (Györfi et al., 2006; Horowitz and Mammen, 2007; Bauer and Kohler, 2019; Schmidt-Hieber, 2020). Condition (A3) requires that the estimation errors of the teacher models be finite.

Theorem 6.1 (Convergence of MT-BKD). *The estimate of $\boldsymbol{\theta}$ by maximizing Equation(5) yields a consistent estimate of $\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$ with the optimal convergence rate upper bounded by*

$$C_2(n) \wedge G \cdot V \left(\left\{ \sum_{g=1}^G \left[\frac{\mu^{(g)}(\mathbf{x})}{\mu(\mathbf{x})\varepsilon^{(g)}} \right]^{\frac{1}{2}} \right\}^{-1} \right), \quad (15)$$

where $C_2(n) \asymp C_1(n)$.

We find the estimation error of MT-BKD could be potentially reduced compared to if the second term of Equation (15) is smaller than $C_2(n)$. Smaller $\varepsilon^{(g)}$ leads to a smaller value of the second term.

Remark. If $\mu^{(g)} = \mu$ for $g = 1, \dots, G$, the second term of Equation (15) equals to $G \left(\sum_{g=1}^G \sqrt{1/\varepsilon^{(g)}} \right)^{-2}$.

7 Simulation

We conduct simulation studies to evaluate the performance of the proposed MT-BKD method in classification tasks under two scenarios with varying dimensions and class counts. Model performance is assessed using classification accuracy and mean square error (MSE) on class probability p , where MSE for MT-BKD compares the predicted posterior mode with the true p . Beyond overall accuracy, we focus on the model’s ability to quantify uncertainty. Mean deviance, as discussed in Section 5, is used to measure prediction uncertainty, and

its relationship with the original data is examined. We further analyze coverage rates across different confidence levels and sample sizes, offering a comprehensive evaluation of the method’s uncertainty quantification.

We compare MT-BKD with four benchmark methods: (1) the teacher model; (2) the original KD method; (3) a combination of original KD with a Bayesian neural network (BNN) (Blundell et al., 2015); and (4) a combination of original KD with Monte Carlo dropout (Dropout) (Gal and Ghahramani, 2016). Details on the implementation of BNN and Dropout are provided in Appendix D. For each distillation method, we evaluate both weighted distillation across multiple teachers, as described in Section 4.2, and distillation with equal teacher weights.

7.1 Simulation 1

In this simulation, we consider binary classification problems. Let $p(\mathbf{x}) = \frac{\exp(\eta(\mathbf{x}))}{1+\exp(\eta(\mathbf{x}))}$, and $Y \sim \text{Bernoulli}(p(\mathbf{x}))$ with $\eta(\mathbf{x}) = -0.5 + 0.1x_1 + 0.8x_1^2 - 0.8x_2^2$. We generate a sample of size 20,000 by letting $x_1 \sim \text{Unif}(-3, 3.5)$ and $x_2 \sim \text{Unif}(-3, 3)$. The top left panel of Fig. 3 shows the probability distribution $p(\mathbf{x})$ mapped across x_1 and x_2 coordinates, with probability values indicated by color intensity.

To simulate teacher specialization, we divide the dataset into three domains b_1 , b_2 , and b_3 , as shown in the top left panel of Fig. 3. We have Teacher 1 primarily trains on domain b_1 , with limited exposure to the other domains, while Teacher 2 specializes in b_3 . As shown in Table 1, Teacher 1 excels in b_1 , whereas Teacher 2 achieves high accuracy in b_3 . We then train a student model on data evenly drawn from all three domains, distilling knowledge from both teachers. Details on model structure are in Appendix D.

Accuracy and MSE. Table 1 presents the overall and domain-specific classification accu-

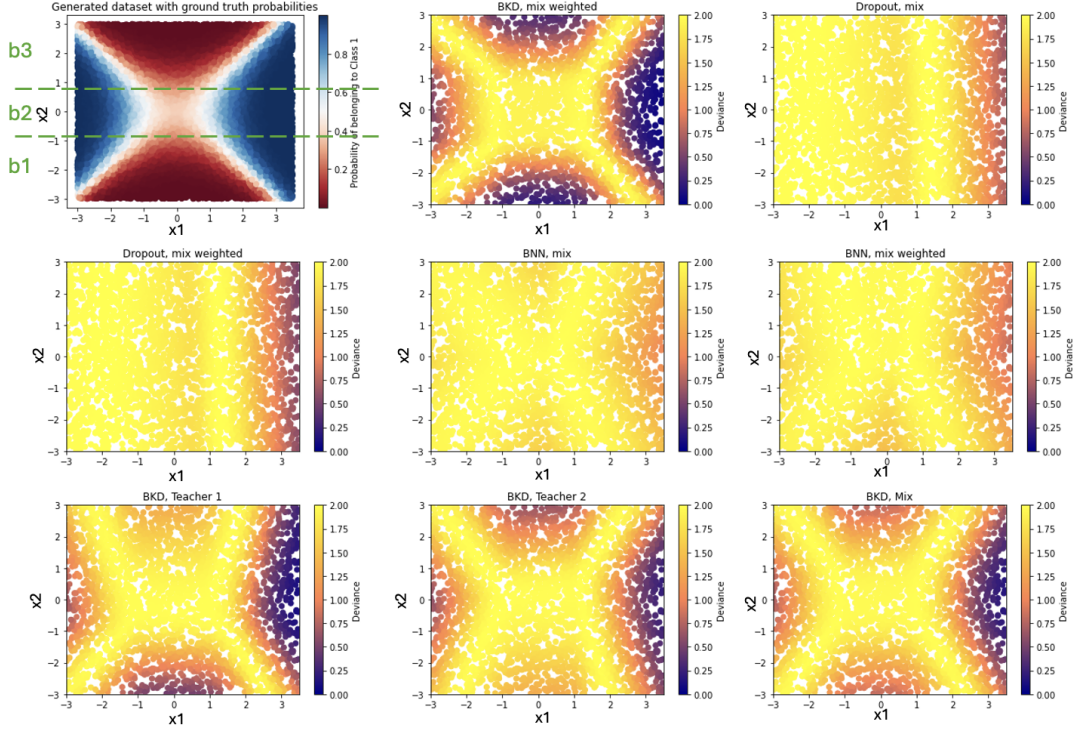


Figure 3: Top left panel: Ground truth probability distribution $p(\mathbf{x})$ with three domains marked as b_1 , b_2 and b_3 . Remaining panels: Predicted uncertainty (measured as deviance) for each data point across various methods.

racy of all methods. We have the following findings: (1) Despite their smaller model sizes, student models trained with KD methods achieve performance comparable to or even exceeding that of their teacher models. (2) MT-BKD outperforms original KD, Dropout, and BNN methods, achieving the highest overall accuracy (0.83) and excelling particularly in domains b_1 (0.87) and b_3 (0.85). (3) The proposed weighted distillation strategy consistently surpasses the equal-weight approach, highlighting its effectiveness in enhancing model performance. Additionally, MSE results for class probability p demonstrate that the proposed MT-BKD method outperforms all other methods, with the weighted version achieving the lowest MSE.

Uncertainty quantification. We evaluate the inference performance of MT-BKD and compare it to Dropout and BNN. For clarity, we also include results from the single-teacher BKD method to assess individual teacher contributions. Fig. 3 compares the uncertainty

Acc	Teachers		Orig KD		MT-BKD		Dropout		BNN	
	T1	T2	equal	weighted	equal	weighted	equal	weighted	equal	weighted
Overall	0.66	0.67	0.72	0.72	0.79	0.83	0.72	0.73	0.77	0.73
b_1	0.84	0.41	0.77	0.75	0.83	0.87	0.75	0.76	0.78	0.78
b_2	0.78	0.79	0.66	0.69	0.78	0.78	0.67	0.68	0.80	0.80
b_3	0.39	0.83	0.74	0.74	0.79	0.85	0.73	0.73	0.76	0.63
MSE	0.14	0.14	0.08	0.08	0.02	0.01	0.08	0.08	0.07	0.07

Table 1: Accuracy and MSE comparison for Simulation 1. We evaluate models on: overall accuracy, accuracy on data from each domain b_i with $i = 1, 2, 3$, and overall MSE for class probability p . The best results in each scenario are highlighted in bold.

quantification performance of various methods, with the top left subplot displaying the ground truth probability distribution. In this setting, where $Y \sim \text{Bernoulli}(p)$, the variance of each observation is highly correlated with $p(1-p)$, making the class probability p a key determinant of uncertainty. Predictions for data points with p near 0.5 are inherently more uncertain, while predictions for p close to 0 or 1 are more certain. Consequently, the white region ($p \approx 0.5$) in the top left subplot should exhibit the highest uncertainty, whereas the dark blue/red regions ($p \approx 0/1$) should show the lowest. Among the methods, weighted MT-BKD (top middle) aligns best with this expected uncertainty pattern.

The bottom-left subplots illustrate BKD’s performance with individual teachers. The student model trained solely with Teacher 1 exhibits clear uncertainty contrasts (yellow vs. dark purple) in domain b_1 but blurred boundaries in b_3 . In contrast, the model trained with Teacher 2 shows sharp uncertainty patterns in b_3 but weak distinction in b_1 , reflecting each teacher’s domain expertise. The weighted MT-BKD effectively integrates both teachers’ strengths, resulting in improved uncertainty estimation across all domains.

Coverage rate. With the credible interval constructed using Equation (13), We evaluate the coverage rate at three standard credible levels (0.85, 0.90, 0.95). As shown in Fig. 4, the weighted MT-BKD method achieves strong performance, with empirical coverage closely matching the nominal credible levels, indicating its effectiveness in quantifying uncertainty. A strong competitor is the weighted BNN method, which achieves coverage performance

Acc	Teachers			Orig KD		MT-BKD		Dropout		BNN	
	T1	T2	T3	equal	weighted	equal	weighted	equal	weighted	equal	weighted
Overall	0.56	0.61	0.53	0.85	0.85	0.85	0.87	0.86	0.86	0.84	0.85
b_1	0.96	0.00	0.33	0.84	0.84	0.91	0.92	0.83	0.83	0.77	0.79
b_2	0.90	0.92	0.05	0.90	0.90	0.89	0.89	0.92	0.92	0.91	0.90
b_3	0.35	0.91	0.93	0.90	0.91	0.86	0.88	0.91	0.92	0.90	0.90
b_4	0.58	0.53	0.51	0.79	0.78	0.86	0.86	0.81	0.80	0.81	0.82
b_5	0.00	0.67	0.83	0.85	0.84	0.76	0.82	0.82	0.81	0.81	0.82
MSE	0.07	0.07	0.08	0.02	0.02	0.01	0.00	0.02	0.01	0.02	0.01

Table 2: Accuracy and MSE comparison for Simulation 2. We evaluate models by overall accuracy, accuracy on data from each domain b_i with $i = 1, \dots, 5$, and overall MSE for class probability p . The best results in each scenario are highlighted in bold.

comparable to the weighted MT-BKD method.

7.2 Simulation 2

In this simulation, we consider the following multi-class scenario. Let $Y \sim \text{Multinomial}(\mathbf{p}(\mathbf{x}))$, and $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), p_3(\mathbf{x}), p_4(\mathbf{x}), p_5(\mathbf{x}))^T$, where $p_i(\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{\sum_{j=1}^5 f(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma})}$, $i = 1, \dots, 5$. We have $\boldsymbol{\mu}_1 = (0, 0, 0, 0, 0)^T$, $\boldsymbol{\mu}_2 = (3, 3, 3, 3, 3)^T$, $\boldsymbol{\mu}_3 = (0, 0, 1, 3, 2)^T$, $\boldsymbol{\mu}_4 = (2, 0, 1, 2, 1)^T$, $\boldsymbol{\mu}_5 = (2, 2, 1, 0, 1)^T$, and $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{5 \times 5}$. We simulate 8,000 sample from each $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, \dots, 5$, and generate labels accordingly.

We consider three teacher models T_1 , T_2 , T_3 , each specializing in data drawn from $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1, 2, 3$. Table 2 shows the overall classification accuracy, MSE, and domain-specific performance on domain b_i , which contains data generated from $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, \dots, 5$. The results align with those from Simulation 1, demonstrating that the proposed MT-BKD method outperforms all other approaches. In particular, the weighted distillation strategy for combining multiple teachers surpasses the equal-weight strategy in accuracy. The advantages of the MT-BKD method are further highlighted in its uncertainty quantification capabilities, as shown by the coverage rates in Fig. 4 (b). MT-BKD with the weighting strategy closely matches the assigned credible levels, whereas other methods exhibit significantly larger errors, with discrepancies of up to 10%.

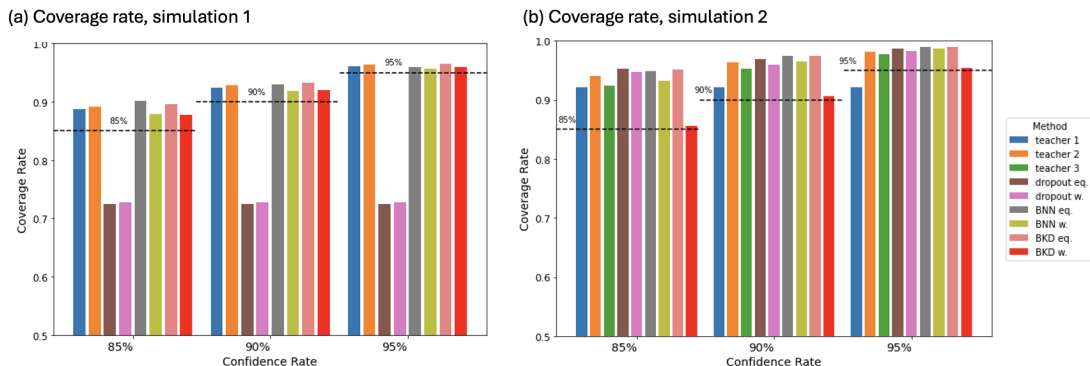


Figure 4: Comparison of coverage rate of (a) simulation 1 and (b) simulation 2 at three standard credible levels (0.85, 0.90, 0.95).

8 Application

We evaluate the performance of MT-BKD on two tasks: eukaryotic protein subcellular localization prediction and image classification. In each task, a compact student model is distilled from teacher models, and we report the student model’s uncertainty predictions. Details of the model architecture and training procedures are provided in Appendix E.

8.1 Protein Subcellular location prediction

We apply MT-BKD to predict eukaryotic protein subcellular localization, a key factor in understanding protein function and disease mechanisms. Protein localization provides the necessary context for function, while mislocalization can lead to diseases such as cancer, neurodegenerative, and metabolic disorders (Hung and Link, 2011). Accurate localization prediction is therefore vital for advancing both basic research and therapeutic development.

We analyze a subset of the DeepLoc2 dataset (Thumulari et al., 2022), derived from UniProt (release 2021_03) (UniProt Consortium, 2018), which classifies proteins into ten subcellular locations: Cytoplasm, Nucleus, Extracellular, Cell Membrane, Mitochondrion, Plastid, Endoplasmic Reticulum, Lysosome/Vacuole, Golgi Apparatus, and Peroxisome. We focus on sequences from Metazoa, Fungi, or Viridiplantae that range from 40 to 1000

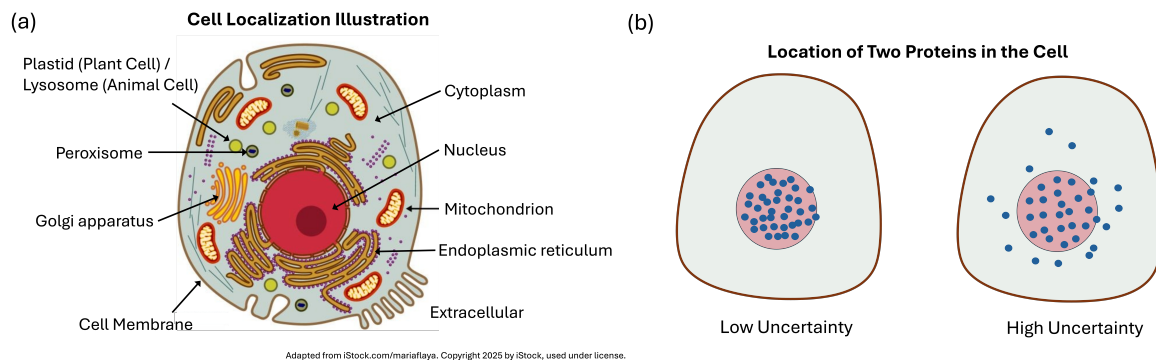


Figure 5: Data description. (a) Ten eukaryotic subcellular compartments for the localization task. (b) Protein-specific predictive uncertainty. Low-uncertainty cases support routine annotation, while high-uncertainty cases suggest ambiguous or multi-compartment behavior and merit follow-up.

Location	Number of proteins	Sublocations
Nucleus	1243	Envelope, inner and outer membrane, matrix, lamina, chromosome, nucleus speckle
Cytoplasm	955	Cytoplasm (cytosol and cytoskeleton)
Extracellular	717	Extracellular
Mitochondrion	515	Envelope, inner and outer membrane, matrix, intermembrane space
Cell membrane	619	Apical, apicolateral, basal, basolateral, lateral, cell membrane, cell projection
Endoplasmic reticulum	334	ER membrane and lumen, microsome, rough ER, smooth ER, Sarcoplasmic reticulum
Plastid	247	Plastid membrane, stroma and thylakoid
Golgi apparatus	142	Golgi apparatus membrane and lumen
Lysosome/Vacuole	171	Contractile, lytic and protein storage vacuole, vacuole lumen and membrane, lysosome lumen and membrane
Peroxisome	57	Peroxisome matrix and membrane

Table 3: Number of proteins in each location and sublocations that were grouped together under the same main location.

amino acids, excluding proteins assigned to multiple locations. From this filtered set, we randomly select 5,000 sequences. Table 3 summarizes the subcellular locations and corresponding sequence counts. The data is then split into training, validation, and test sets in a 6:2:2 ratio. We adopt two teacher models with high-capacity protein sequence representations: ESM-2-650M (Lin et al., 2023) (650 million parameters) and ProtT5-XL-half (120 million parameters), alongside a lightweight student model, ESM-2-8M (8 million parameters). A classifier is added after the feature extraction layers of each model to adapt their pre-trained embeddings to subcellular location prediction.

We explore the performance of MT-BKD in uncertainty quantification. Fig. 6 shows

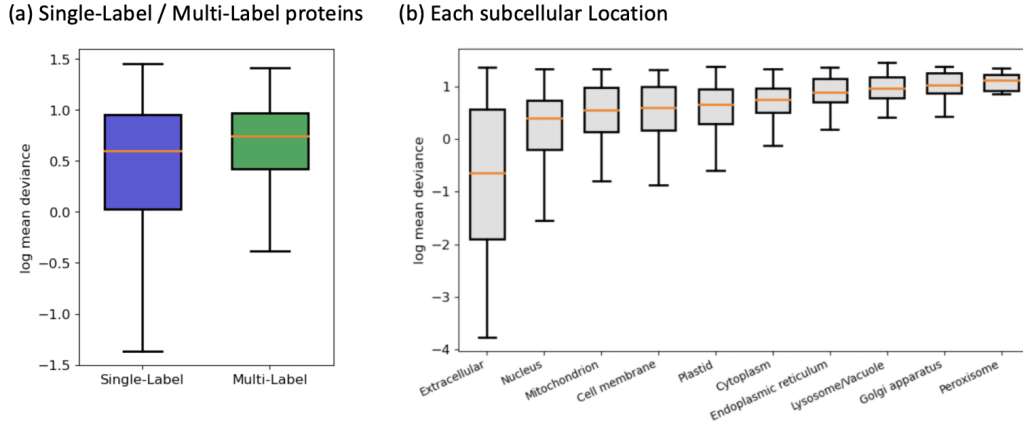


Figure 6: Distribution of log-transformed mean deviance. (a) The first box shows results for the test set, containing only single-label proteins. The second box presents results for a separate dataset of multiple-label proteins. (b) Results for individual protein classes within the single-label test set.

the log-transformed mean deviance distribution. In subfigure (a), the first box shows results from the test set of exclusively single-label proteins, aligning with the training data. The second box displays results from a separate multi-label protein dataset for comparison. As expected, uncertainty is higher for multi-label proteins. This increased uncertainty likely comes from the greater complexity of multiple-label proteins, making them harder to predict. Moreover, since our model was trained only on single-label proteins, reduced performance on multi-label data is unsurprising. The elevated uncertainty appropriately indicates the model’s recognition of its own predictive limitations. Furthermore, we evaluate the model’s prediction confidence across different classes. In Fig. 6 (b), each box corresponds to a specific class in the test set. MT-BKD exhibits the highest confidence for Extracellular, Nuclear, and Mitochondrial, and the lowest for Golgi apparatus, Lysosomes, and Peroxisomes. Such observations align with biological evidence that locations like Extracellular, Nuclear, and Mitochondrial are typically easier to identify with high confidence (Yogev and Pines, 2011; Lu et al., 2021; Owji et al., 2018). This is because of their distinct, well-characterized targeting signals that are consistent across species and highly conserved, i.e., they remain very similar or unchanged across different species over long evolutionary

timescales. In particular, nuclear proteins contain one or more short, positively charged nuclear localization signals (NLS) recognized by importins (Dingwall and Laskey, 1991; Lu et al., 2021), while mitochondria-bound proteins typically possess an N-terminal targeting sequence Yogev and Pines (2011), and extracellular proteins are generally marked by a clear N-terminal signal peptide that directs them through the secretory pathway (Owji et al., 2018). On the contrary, locations such as the Golgi apparatus, lysosomes, and peroxisomes present challenges due to more diverse and less conserved signals. Specifically, Golgi proteins often undergo complex modification processes, such as glycosylation, which can vary between species and even cell types (He et al., 2024). Lysosomal targeting relies on tags like mannose-6-phosphate, but not all lysosomal proteins consistently carry this motif (Braulke and Bonifacino, 2009). Peroxisomal proteins are similarly difficult to predict due to a varied array of peroxisomal targeting signals (PTS1 and PTS2) with structural nuances (Schäfer et al., 2004). These biological facts support the biological basis for varying certainty in subcellular localization predictions given by MT-BKD.

The uncertainty scores achieved by MT-BKD can be leveraged to prioritize experimental validation of borderline predictions, guide model refinement by highlighting where additional data or improved feature engineering is needed, and ultimately enhance our understanding of how subcellular localization signals shape protein distribution.

8.2 Image Classification

In this study, we utilized the Digit-Five dataset, which contains digit images from multiple domains, each presenting unique characteristics (Peng et al., 2019). There are 10 classes corresponding to digits ranging from 0 to 9 in each domain. Specifically, we focused on three colored variants of the dataset, where each color scheme corresponds to a distinct do-

main, thereby introducing variations in data distribution and visual representation. These colored variants aim to simulate real-world challenges in domain adaptation by altering pixel intensity patterns, backgrounds, or foreground colors, which can significantly impact model performance.

Each teacher model was trained to specialize in one of the colored datasets, effectively becoming an expert in its respective domain. Using the proposed MT-BKD approach, we trained a single student model to generalize across all three color domains, enabling it to learn from the unique expertise of each teacher. The teacher models use a ResNet architecture with approximately 25.6 million parameters, while the student model employs a convolutional neural network with around 0.1 million parameters, which is 256 times fewer than the teacher models. Additional details about the data and models can be found in Appendix E.

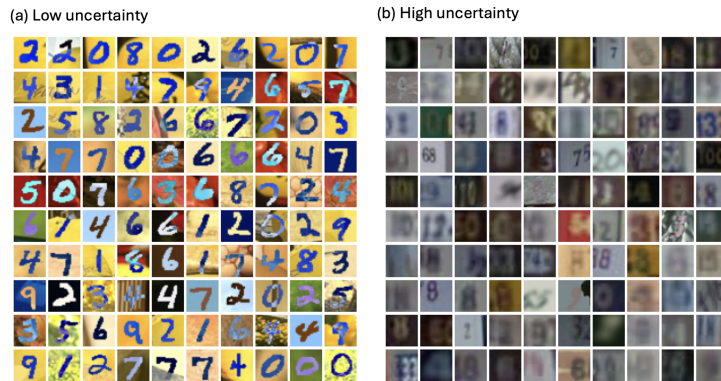


Figure 7: Left panel showcases images with the lowest uncertainty, while the bottom panel showcases images with the highest uncertainty.

In Fig. 7, the images with the lowest and highest uncertainty are shown on the left and right panels, respectively. The images with higher uncertainty are noticeably more difficult to classify. The clear association between high uncertainty and the inherent difficulty in image recognition validates the performance of MT-BKD in uncertainty quantification.

9 Conclusion

In this work, we develop a Multiple-teacher Bayesian Knowledge Distillation (MT-BKD) framework aimed at compressing the teacher model and providing a suite of Bayesian inference tools for the student model’s uncertainty quantification. MT-BKD constructs a teacher-informed prior by integrating external knowledge from multiple teacher models with task-specific training data for the student model’s parameters, from which a posterior is derived. We show that minimizing KD loss is related to estimating the posterior mode in MT-BKD, offering a clearer interpretation of how KD operates and suggesting new ways to enhance KD methods. The developed MT-BKD framework contributes to the growing demand for reliable, privacy-aware, and resource-efficient machine learning models in real-world applications.

We evaluate the performance of MT-BKD using both synthetic and real datasets, focusing on classification accuracy and uncertainty quantification ability. The results demonstrate that MT-BKD achieves performance comparable to the teacher model with the advantage of a significantly smaller size. A key strength of MT-BKD is its ability to precisely quantify prediction uncertainty, as demonstrated by three main findings. First, we confirm that mean deviance is a reliable metric for uncertainty by examining its relationship to the true class probability p and the data’s intrinsic structure (see Fig. 3). Second, an analysis of coverage rates highlights the robustness of MT-BKD. Third, in experiments on real datasets, visualizations and distribution analyses of uncertainty estimates across classes further affirm the reliability of MT-BKD’s uncertainty measures.

MT-BKD could be extended to regression and data generation tasks by imposing suitable teacher-informed priors. For regression, one can adopt a Gaussian prior whose mean aligns with the teacher models’ predictions. For score-based generative models, similarly,

a Gaussian prior on the score function can be set to have its mean match the teacher-estimated scores.

References

- Bates, D. M. and D. G. Watts (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley Series in Probability and Statistics. Wiley.
- Bauer, B. and M. Kohler (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics* 47(4), 2261–2285.
- Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 41(2), 113–128.
- Blundell, C., J. Cornebise, K. Kavukcuoglu and D. Wierstra (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR.
- Braulke, T. and J. S. Bonifacino (2009). Sorting of lysosomal proteins. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1793(4), 605–614.
- Chen, D., J.-P. Mei, H. Zhang, C. Wang, Y. Feng and C. Chen (2022). Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11933–11942.
- Chen, M.-H., J. G. Ibrahim and Q.-M. Shao (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference* 84(1-2), 121–137.
- Dingwall, C. and R. A. Laskey (1991). Nuclear targeting sequences—a consensus? *Trends in biochemical sciences* 16, 478–481.

- Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Fan, J., C. Ma and Y. Zhong (2020). A selective overview of deep learning. *Statistical science: a review journal of the Institute of Mathematical Statistics* 36(2), 264.
- Fang, L., Y. Chen, W. Zhong and P. Ma (2024). Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12935–12956. PMLR.
- Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (Second Edition ed.). Chapman & Hall/CRC Texts in Statistical Science. CRC Press.
- Fukuda, T., M. Suzuki, G. Kurata, S. Thomas, J. Cui and B. Ramabhadran (2017). Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pp. 3697–3701.
- Gal, Y. and Z. Ghahramani (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR.
- Garthwaite, P. H., J. B. Kadane and A. O’Hagan (2005). Statistical methods for eliciting probability distributions. *Journal of the American statistical Association* 100(470), 680–701.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton: Chapman and Hall/CRC.

- Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.
- Genest, C., K. J. McConway and M. J. Schervish (1986). Characterization of externally bayesian pooling operators. *The Annals of Statistics*, 487–501.
- Girolami, M. and B. Calderhead (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73(2), 123–214.
- Goodfellow, I., Y. Bengio and A. Courville (2016). *Deep Learning*. MIT Press.
- Gou, J., B. Yu, S. J. Maybank and D. Tao (2021). Knowledge distillation: A survey. *International Journal of Computer Vision* 129(6), 1789–1819.
- Gui, S., Z. Wang, J. Chen, X. Zhou, C. Zhang and Y. Cao (2023). Mt4mtl-kd: a multi-teacher knowledge distillation framework for triplet recognition. *IEEE Transactions on Medical Imaging*.
- Györfi, L., M. Kohler, A. Krzyzak and H. Walk (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.
- He, M., X. Zhou and X. Wang (2024). Glycosylation: mechanisms, biological functions and clinical implications. *Signal Transduction and Targeted Therapy* 9(1), 194.
- Hinton, G., O. Vinyals, J. Dean and others (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Horowitz, J. L. and E. Mammen (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions.

- Huang, D., N. Stein, D. B. Rubin and S. Kou (2020). Catalytic prior distributions with application to generalized linear models. *Proceedings of the National Academy of Sciences* 117(22), 12004–12010.
- Hung, M.-C. and W. Link (2011). Protein localization in disease and therapy. *Journal of cell science* 124(20), 3381–3392.
- Ibrahim, J. G., M.-H. Chen, Y. Gwon and F. Chen (2015). The power prior: theory and applications. *Statistics in medicine* 34(28), 3724–3749.
- Kondratyuk, D., L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung et al. (2023). Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.
- Korattikara Balan, A., V. Rathod, K. P. Murphy and M. Welling (2015). Bayesian dark knowledge. *Advances in neural information processing systems* 28.
- Latif, E., L. Fang, P. Ma and X. Zhai (2023). Knowledge distillation of LLM for education. *arXiv preprint arXiv:2312.15842*.
- Lin, Z., H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379(6637), 1123–1130.
- Liu, Y., W. Zhang and J. Wang (2020). Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing* 415, 106–113.
- Lu, J., T. Wu, B. Zhang, S. Liu, W. Song, J. Qiao et al. (2021). Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell communication and signaling* 19(1), 60.
- Malinin, A., B. Mlodozieniec and M. Gales (2019). Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*.

- McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley-Interscience.
- Menon, A. K., A. S. Rawat, S. Reddi, S. Kim and S. Kumar (2021). A statistical perspective on distillation. In *International Conference on Machine Learning*, pp. 7632–7642. PMLR.
- Owji, H., N. Nezafat, M. Negahdaripour, A. Hajiebrahimi and Y. Ghasemi (2018). A comprehensive review of signal peptides: Structure, roles, and applications. *European journal of cell biology* 97(6), 422–441.
- Peng, X., Q. Bai, X. Xia, Z. Huang, K. Saenko and B. Wang (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415.
- Phuong, M. and C. Lampert (2019). Towards understanding knowledge distillation. In *International conference on machine learning*, pp. 5142–5151. PMLR.
- Ray, S. (2023, May). Samsung bans chatgpt among employees after sensitive code leak. *Forbes*. Published May 2, 2023.
- Robbins, H. E. (1992). An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer.
- Schäfer, A., D. Kerssen, M. Veenhuis, W.-H. Kunau and W. Schliebs (2004). Functional similarity between the peroxisomal pts2 receptor binding protein pex18p and the n-terminal half of the pts1 receptor pex5p. *Molecular and cellular biology* 24(20), 8895–8906.
- Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function.

- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association* 88(422), 486–494.
- Shen, Y., L. Xu, Y. Yang, Y. Li and Y. Guo (2022). Self-distillation from the last mini-batch for consistency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11943–11952.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin and A. Linde (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76(3), 485–493.
- Thummuluri, V., J. J. Almagro Armenteros, A. R. Johansen, H. Nielsen and O. Winther (2022). Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research* 50(W1), W228–W234.
- Touvron, H., T. Lavril, G. Izacard, X. Martinet, H. Jegou, E. Grave et al. (2024, July). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- UniProt Consortium, T. (2018). Uniprot: the universal protein knowledgebase. *Nucleic acids research* 46(5), 2699–2699.
- Vadera, M., B. Jalaian and B. Marlin (2020). Generalized Bayesian posterior expectation distillation for deep neural networks. In *Conference on Uncertainty in Artificial Intelligence*, pp. 719–728. PMLR.
- Wang, K.-C., P. Vicol, J. Lucas, L. Gu, R. Grosse and R. Zemel (2018). Adversarial distillation of Bayesian neural network posteriors. In *International conference on machine learning*, pp. 5190–5199. PMLR.

- Welling, M. and Y. W. Teh (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688.
- Wu, M.-C., C.-T. Chiu and K.-H. Wu (2019). Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2202–2206. IEEE.
- Yogev, O. and O. Pines (2011). Dual targeting of mitochondrial proteins: mechanism, regulation and function. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1808(3), 1012–1020.
- You, S., C. Xu, C. Xu and D. Tao (2017). Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1285–1294.
- Zagoruyko, S. and N. Komodakis (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.
- Zhang, A., Z. C. Lipton, M. Li and A. J. Smola (2021). *Dive into Deep Learning*. Cambridge University Press.
- Zhang, H., D. Chen and C. Wang (2022). Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4498–4502. IEEE.
- Zhao, B., Q. Cui, R. Song, Y. Qiu and J. Liang (2022). Decoupled knowledge distillation.

In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962.

Zhao, S., X. Wang and X. Wei (2024). Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (01), 1–14.

APPENDIX

Here is an outline of the Appendix.

- Appendix [A](#) summarizes the notations used in the main text.
- Appendix [B](#) discusses the single teacher situation.
- Appendix [C](#) gives the proof of theoretical results.
- Appendix [D](#) presents supplementary details for simulation experiments.
- Appendix [E](#) presents supplementary details for real data analysis.

A Notation

In Table 4, we summarize the notations in the main text.

Notation	Interpretation
$\mathcal{F}$	function classes
$\mathbb{P}$	population distribution
$\mathcal{D}, \mathcal{T}$	dataset
$\mathcal{P}$	distribution class
$\mathbf{h}$	function representing the student neural network
$\boldsymbol{\theta}$	parameters of the student neural network
$\boldsymbol{\theta}^*$	estimate of $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}$	Monte Carlo sample of $\boldsymbol{\theta}$
$\mathbf{x}$	vector representing m -dimensional feature
y	class label
$\mathbf{y}$	one-hot encoding of class label y
$\mathbf{p}$	output of the tea neural network, representing the predicted class probability
$\mathbf{q}$	output of the student neural network, representing the predicted class probability
N	sample size of the training dataset $\mathcal{D}$
n	sample size of the testing dataset $\mathcal{T}$
l	likelihood function
∇	gradient operator
λ	weight parameter
dev	deviance

Table 4: Notation table

B Additioanl Discussion

B.1 Single Teacher Case

In the case where $G = 1$, meaning there is only a single teacher model, we present the following theorem, which establishes the connection between our established Bayesian posterior and the original KD framework.

Theorem B.1. *Assume the prior distributions are defined in Equation (7) and (8). When there is only one teacher model, represented by $G = 1$, the posterior mode of $\boldsymbol{\theta}$ is the minimizer of KD, i.e., $\boldsymbol{\theta}_{KD}^*$.*

Proof. The negative log-transformed posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned}
-l(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}, \lambda, \mathbf{h}(\cdot; \cdot)) &= - \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log(h_k(\mathbf{x}_i; \boldsymbol{\theta})) \\
&\quad - \lambda \sum_{i=1}^N \sum_{k=1}^K p_{ik} \log(h_k(\mathbf{x}_i; \boldsymbol{\theta})) + c, \\
&= \mathcal{L}^{\text{KD}}(\mathbf{h}(\cdot, \boldsymbol{\theta}) \mid \mathcal{D}, \mathbf{p}, \lambda) + c,
\end{aligned} \tag{B.1}$$

where c is a constant. It is easy to verify that the posterior mode of $\boldsymbol{\theta}$ is the minimizer of the KD loss in Equation (4). $\square$

In other words, optimizing the KD loss to search for parameters $\boldsymbol{\theta}$ in the student model is tantamount to achieving the maximum likelihood estimation of the parameters $\boldsymbol{\theta}$.

C Proof of Theoretical Results

Proposition 4.1. *Consider the probability density function $f(\mathbf{q} | \mathbf{p}_i^{(g)})$ as defined in Equation (8) with a constant $\lambda > 0$. Assuming that the parameters of the student model lie in a compact space, then $\pi(\boldsymbol{\theta} | \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}^{g=1, \dots, G})$ is a proper prior.*

Proof. According to Equation (7) and (8), for some constant $C > 0$, we have

$$\begin{aligned} \pi(\boldsymbol{\theta} | \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}^{g=1, \dots, G}) &= C \prod_{i=1}^N \sum_{g=1}^G f(\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}) | \mathbf{p}_i^{(g)}) \\ &= C \prod_{k=1}^K \sum_{g=1}^G b_{ig} \prod_{i=1}^N (\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}}, \end{aligned} \quad (\text{C.1})$$

where $b_{ig} \triangleq \frac{1}{B(\mathbf{1}_K + \lambda \mathbf{p}_i^{(g)})} > 0$ is a constant. Taking integration, we get

$$\int \pi(\boldsymbol{\theta} | \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}^{g=1, \dots, G}) d\boldsymbol{\theta} = C \int \prod_{k=1}^K \sum_{g=1}^G b_{ig} \prod_{i=1}^N (\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}} d\boldsymbol{\theta}. \quad (\text{C.2})$$

Since $0 \leq \mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}) \leq 1$, we have $0 \leq \prod_{k=1}^K \sum_{g=1}^G b_{ig} \prod_{i=1}^N (\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}} \leq 1$ for $1 \leq k \leq K$, $1 \leq g \leq G$ and $1 \leq i \leq N$. Given that $\boldsymbol{\theta}$ lies in a compact space and $\prod_{k=1}^K \sum_{g=1}^G b_{ig} \prod_{i=1}^N (\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}}$ is bounded, we have

$$\int \prod_{k=1}^K \sum_{g=1}^G b_{ig} \prod_{i=1}^N (\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}} d\boldsymbol{\theta} < \infty. \quad (\text{C.3})$$

Thus, we can easily verify that $\pi(\boldsymbol{\theta} | \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}^{g=1, \dots, G})$ is a proper prior. $\square$

Proposition 4.3. *Consider the probability density function $f(\mathbf{q} | \mathbf{p}_i^{(g)})$ as defined in Equation (8), as $\lambda \rightarrow \infty$, we have*

$$f(\mathbf{q} | \mathbf{p}_i^{(g)}) \longrightarrow \delta(\mathbf{q} - \mathbf{p}_i^{(g)}), \quad (\text{C.4})$$

where $\delta(\cdot)$ is the multivariate Dirac delta function.

Proof. For the sake of brevity, we consider the general formula $f(\mathbf{q} | \mathbf{p}, \lambda)$, where $\mathbf{q} = (q_1, q_2, \dots, p_K)^T$ and $\mathbf{p} = (p_1, p_2, \dots, p_K)^T$. WLOG, we consider the situation where $p_k \neq 0$ for $k \in 1, \dots, K$. We have

$$f(\mathbf{q} | \mathbf{p}, \lambda) = \frac{1}{B(\mathbf{1}_K + \lambda \mathbf{p})} \prod_{k=1}^K (q_k)^{\lambda p_k} = \frac{\Gamma(\lambda + K)}{\prod_{k=1}^K (\Gamma(\lambda p_k + 1))} \prod_{k=1}^K (q_k)^{\lambda p_k}. \quad (\text{C.5})$$

We start with the situation when $\lambda p_k, k = 1, \dots, K$, are all integers. We have

$$f(\mathbf{q} \mid \mathbf{p}, \lambda) = \frac{(\lambda + K - 1)!}{\prod_{k=1}^K (\lambda p_k)!} \cdot \prod_{k=1}^K (q_k)^{\lambda p_k}. \quad (\text{C.6})$$

According to the Stirling formula that $n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$, we have

$$\begin{aligned} f(\mathbf{q} \mid \mathbf{p}, \lambda) &\sim \frac{\lambda^{K-1} \cdot \sqrt{\lambda} \left(\frac{\lambda}{e}\right)^\lambda}{\prod_{k=1}^K \sqrt{\lambda p_k} \left(\frac{\lambda p_k}{e}\right)^{\lambda p_k}} \cdot \prod_{k=1}^K (q_k)^{\lambda p_k} \\ &\sim \lambda^{\frac{K-1}{2}} \cdot \prod_{k=1}^K \left(\frac{q_k}{p_k}\right)^{\lambda p_k}. \end{aligned} \quad (\text{C.7})$$

Write $g(\mathbf{q}) = \prod_{k=1}^K \left(\frac{q_k}{p_k}\right)^{p_k}$, we now get the maximization of $g(\mathbf{q})$ w.r.t. $\mathbf{q}$. Since we have the constraint that $\sum_{k=1}^K q_k = 1$, we use the Lagrange multiplier to find the maximum of $g(\mathbf{q})$. The Lagrangian function is defined as

$$\mathcal{L}(q_1, \dots, q_K, c) \equiv \log g + c \left(1 - \sum_{k=1}^K q_k\right), \quad (\text{C.8})$$

where $\log g = \sum_{k=1}^K p_k (\log q_k - \log p_k)$.

To solve

$$\nabla_{q_1, \dots, q_K, c} g(q_1, \dots, q_K, c) = 0, \quad (\text{C.9})$$

we have

$$\begin{cases} \frac{p_k}{q_k} - c = 0, & \text{for } k = 1, \dots, K \\ \sum_{k=1}^K q_k = 1 \end{cases} \quad (\text{C.10})$$

since $\sum_{k=1}^K p_k = 1$, we can easily get that the solution is

$$\begin{cases} q_k = p_k, & \text{for } k = 1, \dots, K \\ c = 1. \end{cases} \quad (\text{C.11})$$

That is, we have $g_{\max}(\mathbf{q}) = 1$ when Equation (C.11) holds, and this leads to $f(\mathbf{q} \mid \mathbf{p}, \lambda) \sim \lambda^{\frac{K-1}{2}} \cdot 1^\lambda \xrightarrow{\lambda \rightarrow \infty} \infty$.

For $\forall \mathbf{q} \neq \mathbf{p}$, we have $g(\mathbf{q}) < 1$, which leads to $f(\mathbf{q} \mid \mathbf{p}, \lambda) \sim \lambda^{\frac{K-1}{2}} \cdot (g(\mathbf{q}))^\lambda \xrightarrow{\lambda \rightarrow \infty} 0$. Thus, as $\lambda \rightarrow \infty$, we have

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \int f(\mathbf{q} \mid \mathbf{p}, \lambda) d\mathbf{q} &= 1, \\ f(\mathbf{q} \mid \mathbf{p}, \lambda) &= 0 \text{ if } \mathbf{q} - \mathbf{p} \neq 0. \end{aligned} \quad (\text{C.12})$$

That is, $f(\mathbf{q} \mid \mathbf{p}, \lambda) \longrightarrow \delta(\mathbf{q} - \mathbf{p})$ as $\lambda \rightarrow \infty$.

If $\exists k$ s.t. λp_k is not an integer. According to the Stirling formula for the gamma function that $\Gamma(z) \sim \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z$, following the same analysis, we can also get $f(\mathbf{q} \mid \mathbf{p}, \lambda) \longrightarrow \delta(\mathbf{q} - \mathbf{p})$ as $\lambda \rightarrow \infty$.

□

Theorem 4.2. Given the prior distribution defined in Equation (7) and (8), the traditional multi-teacher KD strategy in Equation (5) is equivalent to deriving the posterior distributions for individual teacher models separately and combining them using logarithmic opinion pooling.

Proof. Based on a single g^{th} teacher model, we have the prior distribution $\pi_{(g)}(\boldsymbol{\theta}; \{\mathbf{p}_i^{(g)}\}_{i=1, \dots, N}) \propto \prod_{i=1}^N f(\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta}); \mathbf{p}_i^{(g)})$, where $g = 1, \dots, G$. The posterior density is given by

$$\pi_{(g)}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}^{(g)}, \lambda, \mathbf{h}(\cdot, \cdot)) \propto \prod_{i=1}^N \prod_{k=1}^K h_k(\mathbf{x}_i; \boldsymbol{\theta})^{y_{ik}} \cdot \prod_{i=1}^N \frac{1}{B(\mathbf{1}_k + \lambda \mathbf{p}_i^{(g)})} \prod_{k=1}^K (h_k(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}}. \quad (\text{C.13})$$

Combining the posterior distributions based on each teacher model through a logarithmic pooling strategy, we have the combined posterior

$$\pi_{\text{pooled}}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}, \lambda, \mathbf{h}(\cdot, \cdot)) \propto \prod_{g=1}^G \pi_{(g)}(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}^{(g)}, \lambda, \mathbf{h}(\cdot, \cdot))^{w^{(g)}}, \quad (\text{C.14})$$

where $w^{(g)}$ is the weight for the g^{th} teacher model. Thus, the negative log-transformed posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned} -l(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{p}, \lambda, \mathbf{h}(\cdot, \cdot)) &= - \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log(h_k(\mathbf{x}_i; \boldsymbol{\theta})) \\ &\quad - \lambda \sum_{g=1}^G w^{(g)} \sum_{i=1}^N \sum_{k=1}^K p_{ik}^{(g)} \log(h_k(\mathbf{x}_i; \boldsymbol{\theta})) + c, \\ &= \mathcal{L}^{\text{Multi-KD}}(\mathbf{h}(\cdot, \boldsymbol{\theta}) \mid \mathcal{D}, \mathbf{p}, \lambda) + c, \end{aligned} \quad (\text{C.15})$$

where c is a constant. That is, finding the mode of the combined posterior, obtained by merging the posterior distributions from each teacher model using a logarithmic pooling strategy, is equivalent to minimizing the multi-teacher KD loss described in Equation (5). $\square$

Theorem 6.1. The estimate of $\boldsymbol{\theta}$ by maximizing Equation (5) yields a consistent estimate of $\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$ with the optimal convergence rate upper bounded by

$$C_2(n) \wedge G \cdot V \left(\left\{ \sum_{g=1}^G \left[\frac{\mu^{(g)}(\mathbf{x})}{\mu(\mathbf{x}) \varepsilon^{(g)}} \right]^{\frac{1}{2}} \right\}^{-1} \right), \quad (\text{C.16})$$

where $C_2(n) \asymp C_1(n)$.

(A1) For any $\mathbf{x} \in \Omega_{\mathbf{x}}$, $\mu(\mathbf{x}) > 0$, $\max_{1 \leq g \leq G} \mu^{(g)}(\mathbf{x}) > 0$.

(A2) The minimizer $\hat{\boldsymbol{\theta}}_n$ in Equation (2) yield a consistent estimate of $\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})$, i.e., the error $\int_{\Omega_{\mathbf{x}}} [\mathbf{h}(\mathbf{x}; \hat{\boldsymbol{\theta}}_n) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]^2 \mu(\mathbf{x}) d\mathbf{x} = O_p(C_1(n))$ and $\lim_{n \rightarrow \infty} C_1(n) = 0$.

(A3) There exist a positive constant $\varepsilon^{(g)}$ such that $\int_{\Omega_{\mathbf{x}}} [M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]^2 \mu^{(g)}(\mathbf{x}) d\mathbf{x} \leq \varepsilon^{(g)}$ for $g = 1, \dots, G$.

Proof. Let $\hat{\boldsymbol{\theta}}_n$ be the maximize of Equation (5). We define

$$\mathcal{E}_n^\varphi(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \varphi(\mathbf{y}_i^T \text{logit}[\mathbf{h}(\mathbf{x}_i; \boldsymbol{\theta})])$$

Notice that $\mathcal{E}_n^\varphi(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ and

$$\begin{aligned} \mathcal{E}^\varphi(\boldsymbol{\theta}) &= E_{\mathbf{x}}(KL[\mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]) \\ &= E_{\mathbf{x}, \mathbf{y}}(CE[\mathbf{y}, \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]) - E_{\mathbf{x}}(CE[\mathbf{h}(\mathbf{x}_0; \boldsymbol{\theta}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)]) \end{aligned}$$

For g th teacher

$$\begin{aligned} CE[\mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})] &\leq CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})] + \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2 \log \|\mathbf{h}(\mathbf{x}; \boldsymbol{\theta})\|_2 \\ &\leq CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})] + D_1 \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2 \end{aligned}$$

Without loss of generality, we consider the function

$$J(\boldsymbol{\theta}; \mathbf{x}) = \log \left[\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp(\lambda CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]) \right],$$

where $\tilde{w}^{(g)}(\mathbf{x}) = w^{(g)}(\mathbf{x})/B \left(\mathbf{1}_k + \lambda M_t^{(g)}(\mathbf{x}) \right)$.

suppose we set $\tilde{w}^{(g)}(\mathbf{x}) \propto 1/\exp(\lambda CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)])$. Let $W(\mathbf{x}) = \sum_{g=1}^G \tilde{w}^{(g)}$. Since $\sum_{g=1}^G w^{(g)} = 1$ and $B \left(\mathbf{1}_k + \lambda M_t^{(g)}(\mathbf{x}) \right)$ is bounded, we may assume $\frac{1}{W(\mathbf{x})} < D_2$ for a constant $D_2 > 0$. By Jensen's inequality, we have

$$\begin{aligned} J(\boldsymbol{\theta}_0; \mathbf{x}) &= -\log \left[\frac{W(\mathbf{x})}{G} \sum_{g=1}^G \exp(-\lambda CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)]) \right] \\ &\leq \lambda \frac{W(\mathbf{x})}{G} \sum_{g=1}^G CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)] \end{aligned}$$

By Hölder's inequality, we have

$$\begin{aligned} &\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp(CE[\mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]) \\ &= \sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp(CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]) \exp([M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)]^T \log \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})) \\ &\leq \left[\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp(\lambda CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})]) \right]^{\frac{1}{\lambda}} \left[\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp(\lambda [M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)]^T \log \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})) \right]^{\frac{1}{\lambda}} \\ &\leq \exp J(\boldsymbol{\theta}; \mathbf{x}) \left[\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp([M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)]^T \log \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})) \right] \\ &\leq \exp J(\boldsymbol{\theta}; \mathbf{x}) \exp(T_1(\mathbf{x}, \boldsymbol{\theta})) \end{aligned}$$

$$\begin{aligned}
\mathcal{E}^\varphi(\boldsymbol{\theta}) &\leq E_{\mathbf{x}} \left[\frac{1}{W(\mathbf{x})} \sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})] + \sum_{g=1}^G \frac{\tilde{w}^{(g)}(\mathbf{x})}{W(\mathbf{x})} \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2 \right] \\
&\leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] + \sum_{g=1}^G \int \frac{\tilde{w}^{(g)}(\mathbf{x})}{W(\mathbf{x})} \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2 \mu(\mathbf{x}) d\mathbf{x} \\
&\leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] + \sum_{g=1}^G \sqrt{\int \frac{[\tilde{w}^{(g)}(\mathbf{x})\mu(\mathbf{x})/W(\mathbf{x})]^2}{\mu^{(g)}(\mathbf{x})} d\mathbf{x} \int \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2^2 \mu^{(g)}(\mathbf{x}) d\mathbf{x}} \\
&\leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] + \sum_{g=1}^G \sqrt{\varepsilon^{(g)} \int \frac{[\tilde{w}^{(g)}(\mathbf{x})\mu(\mathbf{x})/W(\mathbf{x})]^2}{\mu^{(g)}(\mathbf{x})} d\mathbf{x}},
\end{aligned}$$

We have

$$\begin{aligned}
&\log \left[\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp([M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)]^T \log \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})) \right] \\
&\leq \log \left[\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp(\|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2^2/2 + \|\log \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})\|_2^2/2) \right] \\
&\leq D_1 \log \left[\sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) \exp(\|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2^2/2) \right]
\end{aligned}$$

By Jensen's inequality, we have

$$J(\boldsymbol{\theta}; \mathbf{x}) \geq \sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})].$$

Let $W(\mathbf{x}) = \sum_{g=1}^G \tilde{w}^{(g)}$. Since $\sum_{g=1}^G w^{(g)} = 1$ and $B(\mathbf{1}_k + \lambda M_t^{(g)}(\mathbf{x}))$ is bounded, we may assume $\frac{1}{W(\mathbf{x})} < D_2$ for a constant $D_2 > 0$. We have

$$\begin{aligned}
\mathcal{E}^\varphi(\boldsymbol{\theta}) &\leq E_{\mathbf{x}} \left[\frac{1}{W(\mathbf{x})} \sum_{g=1}^G \tilde{w}^{(g)}(\mathbf{x}) CE[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta})] + \sum_{g=1}^G \frac{\tilde{w}^{(g)}(\mathbf{x})}{W(\mathbf{x})} \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2 \right] \\
&\leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] + \sum_{g=1}^G \int \frac{\tilde{w}^{(g)}(\mathbf{x})}{W(\mathbf{x})} \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2 \mu(\mathbf{x}) d\mathbf{x} \\
&\leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] + \sum_{g=1}^G \sqrt{\int \frac{[\tilde{w}^{(g)}(\mathbf{x})\mu(\mathbf{x})/W(\mathbf{x})]^2}{\mu^{(g)}(\mathbf{x})} d\mathbf{x} \int \|M_t^{(g)}(\mathbf{x}) - \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)\|_2^2 \mu^{(g)}(\mathbf{x}) d\mathbf{x}} \\
&\leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] + \sum_{g=1}^G \sqrt{\varepsilon^{(g)} \int \frac{[\tilde{w}^{(g)}(\mathbf{x})\mu(\mathbf{x})/W(\mathbf{x})]^2}{\mu^{(g)}(\mathbf{x})} d\mathbf{x}},
\end{aligned}$$

Notice that if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we have $\mathcal{E}^\varphi(\boldsymbol{\theta}_0) \leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}_0; \mathbf{x})]$.

Suppose we set $\tilde{w}^{(g)}(\mathbf{x}) \propto \sqrt{\mu^{(g)}(\mathbf{x})/\varepsilon^{(g)}}$, we have

$$\mathcal{E}^\varphi(\boldsymbol{\theta}) \leq D_2 E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] + G \cdot V\left(\frac{\sqrt{\mu(\mathbf{x})}}{\sum_{g=1}^G \sqrt{\mu^{(g)}(\mathbf{x})/\varepsilon^{(g)}}}\right)$$

Furthermore, we have

$$E_{\mathbf{x}}[J(\boldsymbol{\theta}_0; \mathbf{x})] \leq \log \sum_{g=1}^G E_{\mathbf{x}} \left[\tilde{w}^{(g)}(\mathbf{x}) \exp(\lambda C E[M_t^{(g)}(\mathbf{x}), \mathbf{h}(\mathbf{x}; \boldsymbol{\theta}_0)]) \right],$$

For convenience, we denote $\frac{1}{N} \sum_{i=1}^N \log \left[\sum_{g=1}^G w_i^{(g)} \frac{1}{B(\mathbf{1}_k + \lambda \mathbf{p}_i^{(g)})} \prod_{k=1}^K (h_k(\mathbf{x}_i; \boldsymbol{\theta}))^{\lambda p_{ik}^{(g)}} \right]$ by $\hat{E}_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})|\mathcal{D}]$. Let $T_1 = V\left(\frac{\sqrt{\mu(\mathbf{x})}}{\sum_{g=1}^G \sqrt{\mu^{(g)}(\mathbf{x})/\varepsilon^{(g)}}}\right)$

Hence, we have

$$\begin{aligned} (1 + \lambda) \mathcal{E}^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) &\leq \mathcal{E}^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) + \lambda E_{\mathbf{x}}[J(\hat{\boldsymbol{\theta}}_n; \mathbf{x})] + \lambda T_1 - \lambda \mathcal{E}^\varphi(\boldsymbol{\theta}_0) \\ &= \mathcal{E}^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) - \mathcal{E}_n^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) + \mathcal{L}^{Multi-KD}(\hat{\boldsymbol{\theta}}_n) - \mathcal{L}(\boldsymbol{\theta}_0|\mathcal{D}) \\ &\quad + \lambda E_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})] - \lambda \hat{E}_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})|\mathcal{D}] + \lambda T_1 - \lambda \mathcal{E}^\varphi(\boldsymbol{\theta}_0) \\ &\leq \left| \mathcal{E}^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) - \mathcal{E}_n^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \right| + \mathcal{L}^{Multi-KD}(\boldsymbol{\theta}_0) - \mathcal{L}(\boldsymbol{\theta}_0|\mathcal{D}) \\ &\quad + \lambda \left| E_{\mathbf{x}}[J(\boldsymbol{\theta}_0; \mathbf{x})] - \hat{E}_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})|\mathcal{D}] \right| + \lambda T_1 - \lambda \mathcal{E}^\varphi(\boldsymbol{\theta}_0) \\ &\leq \left| \mathcal{E}^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) - \mathcal{E}_n^\varphi(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \right| + \lambda \hat{E}_{\mathbf{x}}[J(\boldsymbol{\theta}_0; \mathbf{x})|\mathcal{D}] - \lambda E_{\mathbf{x}}[J(\boldsymbol{\theta}_0; \mathbf{x})] \\ &\quad + \lambda \left| E_{\mathbf{x}}[J(\boldsymbol{\theta}_0; \mathbf{x})] - \hat{E}_{\mathbf{x}}[J(\boldsymbol{\theta}; \mathbf{x})|\mathcal{D}] \right| + \lambda T_1 + \lambda E_{\mathbf{x}}[J(\boldsymbol{\theta}_0; \mathbf{x})] - \lambda \mathcal{E}^\varphi(\boldsymbol{\theta}_0) \end{aligned}$$

□

D Simulation

D.1 Details on Model Structure

Simulation 1: The teacher models employ a Multilayer Perceptron (MLP) architecture consisting of five hidden layers. These layers have 7, 10, 12, 10, and 5 nodes respectively. The model uses the ReLU activation function. Different dropout rates are applied for two teacher models. The student model employs an MLP architecture consisting of one hidden layer with 5 nodes.

Simulation 2: The teacher models employ the MLP architecture consisting of five hidden layers with 7, 15, 12, 10, and 7 nodes respectively. The model uses the ReLU activation function and incorporates a dropout rate of 0.2. The student model utilizes an MLP architecture with a single hidden layer comprising 10 nodes, activated by the ReLU function.

D.2 Implementation Details of BNN and Dropout

To implement the method integrating the original KD and BNN, we set the BNN model as the student. We train this model using variational inference with details described in

Table 5: Model Structure. The number of parameters for each respective model is indicated in parentheses.

	Teacher models			Student Model
	$M_T^{(1)}$	$M_T^{(2)}$	$M_T^{(3)}$	M_s
Simulation 1	MLP-1 (430)	MLP-1 (430)	-	MLP-S1 (27)
Simulation 2	MLP-2 (600)	MLP-2 (600)	MLP-2 (600)	MLP-S2 (115)
Protein	ESM-2-650M (650M)	ProtT5-XL-half (120M)	-	ESM-2-8M (8M)
Digit-5	CNN-L (2M)	Resnet-50 (26M)	Resnet-50 (26M)	CNN-S (0.1M)

Shridhar et al. (2019), and incorporate the KD loss to distill knowledge from the teacher model. An important hyperparameter when training the BNN model is the weight of the CE loss. We perform a grid search over the values $\{0.01, 0.05, 0.1, 0.2\}$ to find the parameter that yields the highest accuracy. For the CIFAR-10 and CIFAR-100 datasets, due to computational constraints, we add the Bayesian layers only after the feature extraction blocks.

To implement the method integrating the original KD and Monte Carlo dropout, we enable dropout during the testing/inference stage to obtain multiple stochastic predictions.

E Real Data Analysis

E.1 Details on Protein Subcellular Location Prediction

ESM-2 model: The ESM-2 model (Evolutionary Scale Modeling v2) (Lin et al., 2023), developed by Meta AI, is a state-of-the-art protein language model trained on large-scale protein sequence datasets, including UniRef50 (Suzek et al., 2015). Using a Transformer architecture, it captures intricate sequence patterns and relationships, excelling in tasks such as structure prediction, mutation effect prediction, and protein design. We use the ESM-2-650M variant, which contains 650 million parameters and offers improved performance and scalability over its predecessor, as teacher model 1. The ESM-2-8M variant, which contains 8 million parameters, is used as the student model.

ProtT5 model: The ProtT5 model is a protein language model based on the T5 architecture and trained on UniRef50. It captures both local and global sequence features, excelling in tasks such as structure prediction, function annotation, and sequence classification. We use the ProtT5-XL-half model as teacher model 2, which is a half-precision encoder-only variant optimized for efficient and accurate protein sequence embeddings.

During training, the feature extraction layers of these teacher models are frozen, and only the classifier layers are fine-tuned using our dataset.

E.2 Details on Digit-5

We focused on three colored variants of Digit-5: **SVHN**, **MNIST-M**, and **SynthDigits**. **SVHN** (Street View House Numbers) is a dataset of digit images obtained from real-world house number signs in Google Street View (Netzer et al., 2011). It includes a training set of 73,257 examples and a test set of 26,032 examples. Each example is a 32×32 colored image associated with a label of 10 classes, representing the digits 0 through 9. **MNIST-M** is a modified version of the original MNIST dataset, where grayscale digit images are

superimposed on randomly chosen colored background patches (Peng et al., 2019). The dataset retains the original MNIST structure, with a training set of 55,000 examples and a test set of 10,000 examples. Each example is a 28×28 colored image associated with a label of 10 classes. **SynthDigits** is a synthetic dataset consisting of 25,000 training images and 10,000 test images, where each example is a 32×32 colored image of a digit from 0 to 9 (Ganin and Lempitsky, 2015). The images are generated with diverse fonts, colors, and background patterns, providing a visually rich and challenging dataset for digit recognition tasks.

Model Structure: The teacher model employs the 50-layer ResNet architecture as described by He et al. (2016). The ResNet50 model has 25.6 million parameters. The student model employs a CNN architecture with two convolutional layers and a subsequent fully connected layer. The first convolutional layer has 16 output channels, batch normalization, ReLU activation, and max pooling. The second convolutional layer follows a similar pattern but with 8 output channels. The convolutional layer is then followed by fully connected layers, consisting of a hidden layer with 60 nodes and an output layer with 10 nodes. The student model has approximately 0.1 million parameters, which is 256 times fewer than the teacher models.

References

- Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. (2011). Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada.
- Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415.
- Shridhar, K., Laumann, F., and Liwicki, M. (2019). A comprehensive guide to Bayesian convolutional neural network with variational inference. arxiv 2019. *arXiv preprint arXiv:1901.02731*.
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2015). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.