Uncertainty separation via ensemble quantile regression

Navid Ansari, Hans-Peter Seidel, Vahid Babaei

Max Planck Institute for Informatics
nansari@mpi-inf.mpg.de, hpseidel@mpi-sb.mpg.de, vbabaei@mpi-inf.mpg.de

Abstract

This paper introduces a novel and scalable framework for uncertainty estimation and separation with applications in data driven modeling in science and engineering tasks where reliable uncertainty quantification is critical. Leveraging an ensemble of quantile regression (E-QR) models, our approach enhances aleatoric uncertainty estimation while preserving the quality of epistemic uncertainty, surpassing competing methods, such as Deep Ensembles (DE) and Monte Carlo (MC) dropout. To address challenges in separating uncertainty types, we propose an algorithm that iteratively improves separation through progressive sampling in regions of high uncertainty. Our framework is scalable to large datasets and demonstrates superior performance on synthetic benchmarks, offering a robust tool for uncertainty quantification in data-driven applications.

1 Introduction

In recent years, uncertainty estimation has become an essential aspect of machine learning, especially for applications that demand high reliability in decision-making, such as autonomous driving (Kendall and Gal 2017), and medical diagnosis (Lambrou, Papadopoulos, and Gammerman 2010; Yang et al. 2009). Accurate uncertainty estimation not only supports better model interpretability but also helps identify areas where models are likely to make errors, ensuring safety in high-stakes environments.

Uncertainty in machine learning is typically categorized into two types: *aleatoric* uncertainty, which originates from inherent noise or variability in the data, and *epistemic* uncertainty, which stems from limitations in the model's knowledge and can potentially be reduced with additional data (Hüllermeier and Waegeman 2021). Separating these two types of uncertainty is essential in applications like Bayesian optimization (Frazier 2018; Hernández-Lobato et al. 2017; Shahriari et al. 2015; Ansari et al. 2023), inverse design (Wijaya et al. 2024; Ansari et al. 2022), and active learning (Ren et al. 2021; Settles 2009; Kirsch, Van Amersfoort, and Gal 2019; Gal, Islam, and Ghahramani 2017), where understanding the nature of uncertainty influences strategic decisions.

For instance, in Bayesian optimization (BO) (Frazier 2018; Hernández-Lobato et al. 2017; Shahriari et al. 2015; Ansari et al. 2023) and active learning (Ren et al. 2021; Settles 2009; Kirsch, Van Amersfoort, and Gal 2019; Gal, Islam, and Ghahramani 2017), focusing on epistemic uncertainty directs computational resources toward regions where additional data could improve model performance. In engineering, inverse design is a critical task aimed at identifying design parameters that achieve a desired performance. Accurate characterization of uncertainty—both its magnitude and type— is essential for this process, as it enables the identification of designs that are not only optimized for performance but also robust and reliable under real-world conditions (Wijaya et al. 2024; Ansari et al. 2022).

While various methods exist for modeling uncertainty, most classical BO approaches relying on Guassian processes as surrogate models cannot effectively scale with increasing dataset size which is a common theme in engineering problems (Wang et al. 2016; Hernández-Lobato and Adams 2015; Snoek, Larochelle, and Adams 2012). For this reason there has been a lot of efforts in replacing GPs with Bayesian neural networks (Neal et al. 2011; Chen, Fox, and Guestrin 2014; Lakshminarayanan, Pritzel, and Blundell 2016; Gal and Ghahramani 2016a; Lee et al. 2017; Wilson et al. 2016). However Ansari et al. (2023) showed that most of these BNNs also do not scale very well with the size of the dataset. Among the techniques capable of handling large-scale data are Deep Ensembles (DE) (Lakshminarayanan, Pritzel, and Blundell 2017) and Monte Carlo (MC) dropout (Gal and Ghahramani 2016b). While the former can provide some degree of uncertainty separation MC dropout only predicts the epistemic uncertainty. However, even DE faces limitations in both aspects of uncertainty estimation quality:

- **Localization of uncertainty:** The first task in uncertainty estimation is identifying regions within the input space that may lead to unreliable predictions, regardless of the type of uncertainty. Effective localization aids in detecting areas where model predictions might not remain faithful to reality.
- **Separation of uncertainties:** The next task involves distinguishing between aleatoric and epistemic uncertainty across the input domain. This distinction is particularly valuable in applications such as Bayesian optimization and active learning, where it is beneficial to avoid exploring regions dominated by irreducible aleatoric noise.

We demonstrate that under certain conditions, existing methods can misidentify uncertainty types and provide inaccurate predictions. To address these issues, we introduce Algorithm 1 designed to enhance accuracy and reliability in uncertainty separation.

In Section 4, we apply DE and E-QR to both a toy problem and a synthetic mechanical problem to illustrate the challenges associated with uncertainty separation. We further demonstrate how Algorithm 1 effectively addresses these challenges, highlighting its advantages in accurately separating and quantifying uncertainty.

2 Related work and background

Uncertainty estimation in machine learning, particularly in deep learning models, has gained significant attention due to its importance in reliable decision-making. Separating epistemic and aleatoric uncertainty is critical in many applications. Deep Ensembles (DE), introduced by Lakshminarayanan, Pritzel, and Blundell (2017), provides uncertainty estimates by averaging predictions from independently trained neural networks. While DE can estimate both epistemic and aleatoric uncertainty, it sometimes fails to separate them effectively.

Monte Carlo (MC) dropout approximates Bayesian neural networks by applying dropout during training and inference (Gal and Ghahramani 2016b).

Quantile regression has emerged as a promising approach for uncertainty estimation, predicting the upper and lower quantiles of the target distribution instead of point estimates (Koenker and Bassett 1978). To ensemble such models, Fakoor et al. (2023) proposed a paradigm focused on aleatoric uncertainty capture but did not address epistemic uncertainty separation. Tagasovska and Lopez-Paz (2018) leveraged quantile regression ensembles to model epistemic uncertainty in high-dimensional spaces, while Hoel, Wolff, and Laine (2023) applied it to reinforcement learning in autonomous driving, demonstrating robustness in safety-critical scenarios. Additionally, Mallick, Balaprakash, and Macfarlane (2022) showcased its scalability and accuracy in separating uncertainties in spatiotemporal problems.

DE and Ensemble Quantile Regression (E-QR) are superior to MC dropout for epistemic uncertainty estimation, as both train independent models with varying parameters, unlike MC dropout, which uses stochastic variations of a single model (Gal and Ghahramani 2016a; Lakshminarayanan, Pritzel, and Blundell 2016; Koenker and Hallock 2001). While DE uses sub-networks trained with Negative Log Likelihood (NLL) loss to model aleatoric uncertainty, E-QR leverages pinball loss to predict quantiles (Romano, Patterson, and Candes 2019).

E-QR is simpler and more stable to train than DE. Pinball loss is less sensitive to initialization and learning rate compared to NLL loss, which is prone to gradient instabilities and overfitting on small datasets (Moustakides and Basioti 2019; Streit and Luginbuhl 1994). DE requires a two-step process—optimizing primary predictions (μ) followed by training uncertainty (σ) heads with NLL loss—doubling its computational effort compared to E-QR, which learns quantile predictions in a single step (Zhang 2020; Ansari et al. 2022). Moreover, E-QR provides superior aleatoric uncertainty predictions by directly modeling quantile intervals, avoiding parametric assumptions required in DE (Barron 2019).

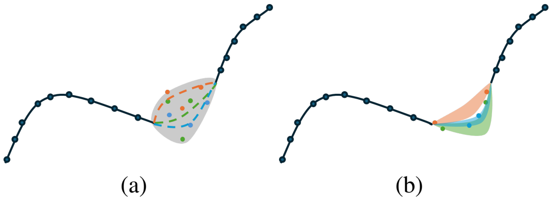

Figure 1: On the left figure, we observe how the lack of data in a region with aleatoric uncertainty causes each subnetwork to fit the noise differently, as they only access small subsets of the data. This overfitting results in the false reporting of epistemic uncertainty where none exists. On the right figure, we see that in regions lacking sufficient data, the fits for aleatoric uncertainty become unreliable, leading to incorrect report of aleatoric uncertainty.

3 Methodology

In Section 3.2, we discuss the shortcomings of scalable Bayesian Neural Networks (BNNs) in uncertainty separation and demonstrate how Algorithm 1 can be used to achieve robust uncertainty separation.

3.1 Challenges in uncertainty separation

Leakage of aleatoric uncertainty into epistemic uncertainty

Although epistemic uncertainty is intended to capture only uncertainty due to limited knowledge, in practice, it may inadvertently include aleatoric uncertainty when data is insufficient. This occurs because each subnetwork in an ensemble model is trained on a subsample of the data. If these subsamples are too small, subnetworks may overfit to their specific data rather than generalizing between data points, as expected with an L2 loss function. As a result, aleatoric uncertainty can "leak" into the epistemic uncertainty estimates. This phenomenon is common in both Deep Ensembles and Ensemble Quantile Regression methods.

Leakage of epistemic uncertainty in aleatoric uncertainty.

All models studied here rely on fitting mechanisms to model aleatoric uncertainty. DE uses the Negative Log Likelihood (NLL) loss, while E-QR employs the pinball loss to fit upper and lower quantiles. As a result, the accuracy of these predictions depends heavily on the availability of high-quality and sufficient data. In regions with high epistemic uncertainty, the fit of aleatoric uncertainty estimates can become arbitrarily. This issue is exacerbated by the use of bagging in sub-networks, where only a subsample of data is visible to models attempting to learn aleatoric uncertainty.

Knowing that the root cause of both problems is the lack of data, we propose a progressive sampling strategy in the next section. This strategy focuses on regions where uncertainty is detected, but the type of uncertainty (aleatoric or

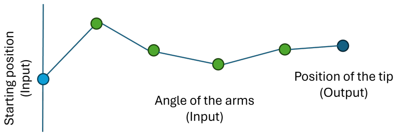

Figure 2: Multi joint robotic arm with a moving base and 4 rotating joints. The goal is to train a model that can predict the 2D position of the tip of the arm.

```
Algorithm 1: Uncertainty separation.
Input
$(X, Y)^0$    // Initial data set
$Q$    // Number of iterations of the main algorithm
$O$    // Number of outputs
$\Phi$    // Native Forward Process, e.g., a simulation
$T$    // Threshold value for Binarizing the uncertainty map
Output
$U_E$, $D_{UE}$ , $P_{UE}$    // Position and value of the separated epistemic regions.
$U_A$, $D_{UA}$ , $P_{UA}$    // Position and value of the separated aleatoric regions.

begin
    $dataset \leftarrow (\mathbf{X^0},\ \mathbf{Y^0})$
    $f^0_{BNN} \xLeftarrow{\text{train}} dataset$ // Train the BNN surrogate.
    $U_A, U_E \leftarrow f^0_{BNN}$ // calculate the uncertainties from the surrogate.
    $\min U_E, \max U_E, \min U_A, \max U_A \leftarrow \mathit{MIN\text{-}MAX}(U_E, U_A)$
        // Extracting the min-max of the uncertainties to be used for scaling.
    for $i \leftarrow 1$ to $Q$ do
        for $j \leftarrow 1$ to $O$ do
            $\overline{U_E}, \overline{U_A} \leftarrow \mathit{Normalizer}(\min U_E, \max U_E, \min U_A, \max U_A)$
                // Scaling and normalizing both total uncertainty and epistemic uncertainty calculated from the original dataset.
            $\overline{U^j_{total}} = \overline{U_E} + \overline{U_A}$ // Adding both normalized uncertainties to make sure all the uncertain regions are captured.
            $\overline{U^{total}_{total}} = \overline{U^{total}_{total}} \times \overline{U^j_{total}}$ // Multiplying all the total uncertainty maps for all outputs to make sure we only keep the ones that are mutual.
            $(X, Y)^i \leftarrow \mathit{Binarize}(\overline{U^{total}_{total}}, T)$ // Binarizing the uncertainty map to create a mask for extracting the next batch of data $(X, Y)^i$.
            $dataset \leftarrow (X, Y)^i$ // Append new data to the old.
            $f^i_{BNN} \xLeftarrow{\text{train}} dataset$ // Train the BNN surrogate.
            $U^i_A, U^i_E \leftarrow f^i_{BNN}$ // calculate the uncertainties from the surrogate.
        end
    end
    $U_A, D_{UA}, P_{UA} \leftarrow \overline{U^{total}_{total}}$ // After sufficient number of iterations the total uncertainty map only contains aleatoric uncertainty.
    $U_A, D_{UA}, P_{UA} \leftarrow \overline{U^{total}_{total}} \oplus \overline{U^0_{total}}$ // By comparing the final total uncertainty with the initial one we can determine the epistemic uncertain regions of the initial data.
end
```

epistemic) remains unclear. By progressively acquiring additional data in these regions, the model can refine its predictions and better separate the two types of uncertainty.

3.2 Reliable separation of epistemic and aleatoric uncertainty

To reliably separate aleatoric and epistemic uncertainties, we first identify regions where uncertainty is suspected and create a comprehensive uncertainty map by normalizing and combining all available uncertainty maps. In cases with multiple outputs, we focus on the common regions across each output's uncertainty map, as these shared uncertainties are more likely to originate from the data rather than local misfitings specific to individual outputs.

Once this common uncertainty map is established, we gather additional data in the uncertain regions and retrain the models. Iteratively repeating this process diminishes regions of epistemic uncertainty while areas of aleatoric uncertainty remain unchanged. By performing a logical XOR operation between the final uncertainty map and the initial one, we can isolate the initial epistemic uncertainty present in our initial dataset. Algorithm 1 presents the complete procedure.

4 Evaluation

4.1 Experiment setup

Toy In this experiment, we aim to fit the function

$$y_1 = \sin\left(\sqrt{x_1^2 + x_2^2}\right), y_2 = x{\cdot}\cos\left(\sqrt{x_1^2 + x_2^2}\right){\cdot}\cos(x_2)$$

We introduce 4 specific regions in the input space two on the top lacking data, and two on the bottom polluted with irreducible random noise modeled as $\mathcal{N}(0, 0.3)$ (Figure 3).

Multi-joint robot In this problem, a robotic arm with four rotatable joints and an adjustable base position on the wall ($x \in \mathbb{R}^5$) is considered. The goal is to predict the final 2D position of the arm's tip ($y \in \mathbb{R}^2$) given its joint angles (Ardizzone et al. 2019). To evaluate uncertainty separation, we conduct an experiment where the behavior of one joint is excluded from the dataset to test whether our model can recover the missing information and determine the type of uncertainty.

4.2 Aleatoric uncertainty leak into epistemic uncertainty prediction

Figure 3 illustrates the aleatoric and epistemic uncertainty calculated by the E-QR model for both outputs for the toy

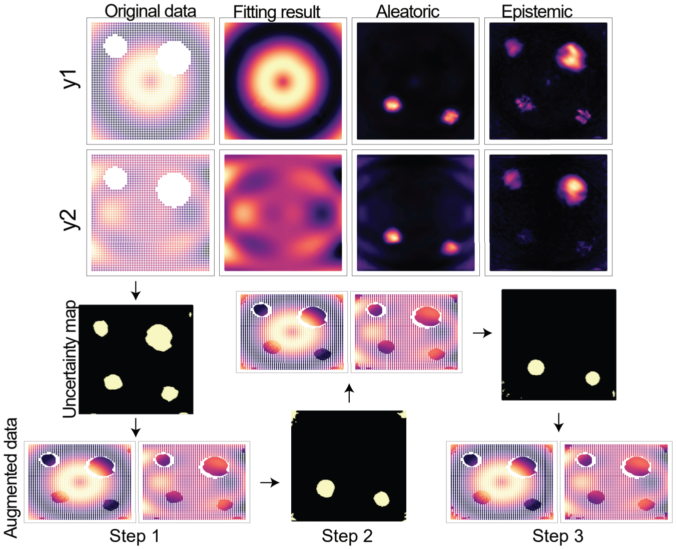

Figure 3: In the top figure, from left to right, we present the original training data, the model's predictions, and the aleatoric and epistemic uncertainty maps. The first row corresponds to the first output, while the second row corresponds to the second output of the model. The epistemic uncertainty map highlights four regions: two caused by a lack of data and two influenced by the leak of random noise. To achieve accurate separation, we apply Algorithm 1. After two iterations, only the regions with aleatoric uncertainty remain in the uncertainty map, confirming that the vanished uncertain areas were indeed epistemic. Note that the white dots are due to the low density of the training samples.

experiment. Notably, aleatoric uncertainty appears to leak into the epistemic uncertainty plot.

Using Algorithm 1, we can generate an uncertainty map. By iteratively focusing on uncertain regions and selectively filling them with additional data, the true aleatoric uncertainty regions recovers accurately in one iteration.

4.3 Epistemic uncertainty leak into aleatoric uncertainty prediction

In Figure 4, the robotic arm is not augmented with aleatoric noise; however, the behavior of one joint is excluded from the dataset over a range of angles. This setup induces epistemic uncertainty, as the model lacks information about the excluded range. While we expect the model to predict only epistemic uncertainty, the figure shows a leakage of epistemic uncertainty into the aleatoric uncertainty predictions.

Algorithm 1 addresses this issue, identifying the uncertainty as epistemic after four iterations. By adding data to the uncertain regions, the uncertainty is completely resolved, confirming that it was indeed epistemic.

5 Conclusion

This work introduces a novel framework for uncertainty separation using Ensemble Quantile Regression (E-QR), addressing the challenges of uncertainty leakage that lead existing methods to erroneous separations. By leveraging E-QR and Algorithm 1, we achieve robust separation of aleatoric and epistemic uncertainties while mitigating leakage issues. The proposed method is computationally efficient, scalable to large datasets, and validated through experiments on synthetic benchmarks. These results establish our framework as a reliable tool for uncertainty separation in scientific and engineering applications.

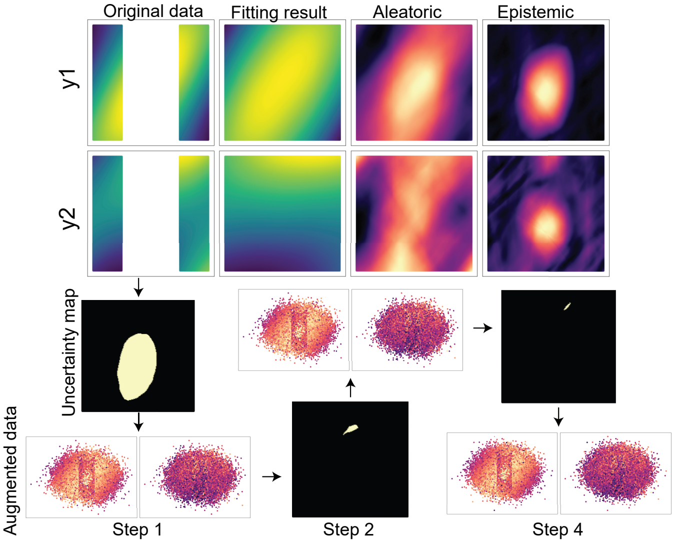

Figure 4: This experiment highlights the leakage of epistemic uncertainty into aleatoric uncertainty. The plot illustrates the cross-section of two out of four rotating joints and their effect on the 2D position of the tip. From left to right, we present the original training data, the model's predictions, and the aleatoric and epistemic uncertainty maps. Since no random noise is injected into this problem, we expect to observe only epistemic uncertainty. However, the aleatoric uncertainty map incorrectly reflects leaked epistemic uncertainty. Applying Algorithm 1 for four iterations resolves this issue, as the uncertainty vanishes when the uncertain regions are filled with additional data, confirming that the observed uncertainty was indeed epistemic.

References

Ansari, N.; Javanmardi, A.; Hüllermeier, E.; Seidel, H.-P.; and Babaei, V. 2023. Large-Batch, Iteration-Efficient Neural Bayesian Design Optimization. *arXiv preprint arXiv:2306.01095*.

Ansari, N.; Seidel, H.-P.; Vahidi Ferdowsi, N.; and Babaei, V. 2022. Autoinverse: Uncertainty aware inversion of neural networks. *Advances in Neural Information Processing Systems*, 35: 8675–8686.

Ardizzone, L.; Kruse, J.; Rother, C.; and Köthe, U. 2019. Analyzing Inverse Problems with Invertible Neural Networks. In *International Conference on Learning Representations*.

Barron, J. T. 2019. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4331–4339.

Chen, T.; Fox, E.; and Guestrin, C. 2014. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, 1683–1691. PMLR.

Fakoor, R.; Kim, T.; Mueller, J.; Smola, A. J.; and Tibshirani, R. J. 2023. Flexible model aggregation for quantile regression. *Journal of Machine Learning Research*, 24(162): 1–45.

Frazier, P. I. 2018. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

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

Gal, Y.; and Ghahramani, Z. 2016b. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 1050–1059.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, 1183–1192. PMLR.

Hernández-Lobato, J. M.; and Adams, R. 2015. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, 1861–1869. PMLR.

Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; and Aspuru-Guzik, A. 2017. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. In *International conference on machine learning*, 1470–1479. PMLR.

Hoel, C.-J.; Wolff, K.; and Laine, L. 2023. Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 24(6): 6030–6041.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3): 457–506.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Koenker, R.; and Bassett, G. 1978. Regression Quantiles. *Econometrica*, 46(1): 33–50.

Koenker, R.; and Hallock, K. F. 2001. Quantile regression. *Journal of economic perspectives*, 15(4): 143–156.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2016. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 6402–6413.

Lambrou, A.; Papadopoulos, H.; and Gammerman, A. 2010. Reliable confidence measures for medical diagnosis with evolutionary algorithms. *IEEE Transactions on Information Technology in Biomedicine*, 15(1): 93–99.

Lee, J.; Bahri, Y.; Novak, R.; Schoenholz, S. S.; Pennington, J.; and Sohl-Dickstein, J. 2017. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*.

Mallick, T.; Balaprakash, P.; and Macfarlane, J. 2022. Deep-ensemble-based uncertainty quantification in spatiotemporal graph neural networks for traffic forecasting. *arXiv preprint arXiv:2204.01618*.

Moustakides, G. V.; and Basioti, K. 2019. Training neural networks for likelihood/density ratio estimation. *arXiv preprint arXiv:1911.00405*.

Neal, R. M.; et al. 2011. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11): 2.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9): 1–40.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Settles, B. 2009. Active learning literature survey.

Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; and De Freitas, N. 2015. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Streit, R. L.; and Luginbuhl, T. E. 1994. Maximum likelihood training of probabilistic neural networks. *IEEE Transactions on neural networks*, 5(5): 764–783.

Tagasovska, N.; and Lopez-Paz, D. 2018. Frequentist uncertainty estimates for deep learning. *arXiv preprint arXiv:1811.00908*.

Wang, Z.; Hutter, F.; Zoghi, M.; Matheson, D.; and De Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55: 361–387.

Wijaya, K. T.; Ansari, N.; Seidel, H.-P.; and Babaei, V. 2024. TrustMol: Trustworthy Inverse Molecular Design via Alignment with Molecular Dynamics. *arXiv preprint arXiv:2402.16930*.

Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; and Xing, E. P. 2016. Deep kernel learning. In *Artificial intelligence and statistics*, 370–378. PMLR.

Yang, F.; Wang, H.-z.; Mi, H.; Lin, C.-d.; and Cai, W.-w. 2009. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics*, 10: 1–14.

Zhang, L. 2020. A Novel Penalized Log-likelihood Function
for Class Imbalance Problem.