Finer Disentanglement of Aleatoric Uncertainty Can Accelerate Chemical Histopathology Imaging

Ji-Hun Oh[†], Kianoush Falahkheirkhah[†], and Rohit Bhargava[†,‡]

[†] University of Illinois Urbana-Champaign, Urbana, IL, USA
[‡] CZ Biohub Chicago, LLC, Chicago, IL, USA
{jihunoh2, kf4, rxb}@illinois.edu

Abstract. Label-free chemical imaging holds significant promise for improving digital pathology workflows, but data acquisition speed remains a limiting factor. To address this gap, we propose an adaptive strategy—initially scan the low information (LI) content of the entire tissue quickly, identify regions with high aleatoric uncertainty (AU), and selectively re-image them at better quality to capture higher information (HI) details. The primary challenge lies in distinguishing between high-AU regions mitigable through HI imaging and those that are not. However, since existing uncertainty frameworks cannot separate such AU subcategories, we propose a fine-grained disentanglement method based on post-hoc latent space analysis to unmix resolvable from irresolvable high-AU regions. We apply our approach to streamline infrared spectroscopic imaging of breast tissues, achieving superior downstream segmentation performance. This marks the first study focused on fine-grained AU disentanglement within dynamic image spaces (LI-to-HI), with novel application to streamline histopathology. Code will be made public.

Keywords: Digital pathology · Label-free imaging · Aleatoric and epistemic uncertainty · Uncertainty quantification and disentanglement

1 Introduction

Histopathology, the practice of diagnosing and treating cancer, relies on special dyes like hematoxylin and eosin to highlight tissue and cellular architecture. While this process is well-established, challenges remain—recurrent costs of labor and reagents, data quality issues from staining artifacts, lab variability, and tissue damage, and interpretation subjectivity. The desire to overcome these limitations has driven interest in label-free imaging [10,16,18,33,40], which extracts intrinsic sample information without binding dyes. In particular, infrared (IR) chemical imaging leverages the fundamental vibrational spectroscopic fingerprints of biomolecules—including proteins, lipids, nucleic acids, and collagen—to quantify their spatial distributions. Pairing this rich molecular detail with deep neural net (DNN) tools like image segmenters enhances digital pathology, potentially offering scalability and robustness [2].

However, acquiring such images with high-information (HI) content can take hours depending on technology and tissue size, posing challenges in low-resource
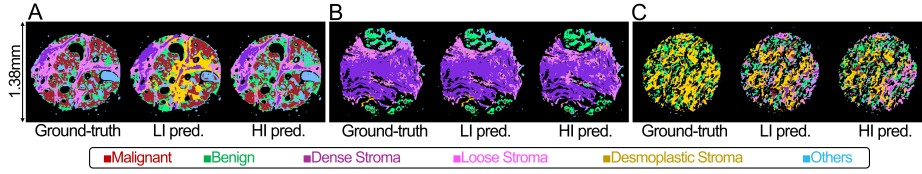
Fig. 1: **Impact of data quality.** We show segmentation results of LI and HI IR images of breast tissues (§4.1). We observe discrepant behavior—A is accurate only at HI, B is performant at both HI and LI, while C is subpar throughout.

or time-sensitive environments. A practical workaround is to minimize the data collected, such as employing fewer spectral frequencies or sparse scanning. Unfortunately, such low-information (LI) images often induce ambiguity (*i.e.*, aleatoric uncertainty, AU), degrading downstream DNN performance (sample A in Fig. 1). On the brighter side, not all data suffer from this issue, with LI being sufficient (B in Fig. 1). This observation suggests an adaptive strategy that balances throughput and accuracy: Scan the entire sample at LI, identify critical regions, and then re-image these areas at HI. The crux lies in selecting regions with high AU while avoiding conflation with epistemic uncertainty (EU)—e.g., when data is out-of-distribution (OOD). While recent studies have made strides in disentangling these two types of uncertainty [8,14,21,36,45], there remains another issue: *not all AU is resolvable at HI* (C in Fig. 1). This calls for a further partition of high-AU data into resolvable *vs.* irresolvable cases, minimizing the unnecessary sampling of the latter. No current uncertainty framework achieves this, as they do not go beyond the EU-AU dichotomy.

Our contribution is two-fold: **(i)** We propose a scalable, fine-grained disentanglement method for categorizing AU by resolvability, based on post-hoc latent space analysis. This is the first uncertainty work to tackle dynamic settings with inconstant input spaces (LI⇒HI) leading to reducible AU. **(ii)** We put forth a new application of uncertainties in medical contexts, beyond uses for trustworthy AI [1,38], active learning [30,37], and semi-supervised learning [50]. In IR imaging of breast tissues, we show multifold gains in downstream segmentation tasks, compared to random re-scanning, by targeting resolvable high-AU pixels. This is the first demonstration of experimental design optimization using uncertainty guidance at a granular level.

2 Background and Preliminaries

We consider a supervised classification setting, denoting input $\boldsymbol{x} \in X$ and discrete label $y \in Y$ among C classes. Empirical risk minimization (ERM) over training set $\mathcal{D}$ yields a DNN model $w : X \to \Delta_{C-1}$ *s.t.* $w \in \mathcal{H}$, where $\boldsymbol{\pi} \in \Delta_{C-1}$ is the probability simplex, and $\mathcal{H}$ is the hypothesis space. Let w^* denote the true data-generating model, or the Bayes predictor with minimal risk over $X \times Y$. The total uncertainty is the error between the predicted $y = \texttt{one\_hot}(\boldsymbol{\pi}) =$

`one_hot`($w(\boldsymbol{x})$) and the observed true label $y^* \sim \text{Cat}(\boldsymbol{\pi^*}) = \text{Cat}(w^*(\boldsymbol{x}))$. This can be broken down to epistemic (EU) and aleatoric (AU) parts [19]—**EU** arises from $w \neq w^*$, reflecting the learner's ignorance, and stems from model misspecification due to inductive biases ($\mathbb{E}[w] \neq w^*$) and the ERM variance ($\mathbb{V}[w] > 0$). On the other hand, **AU** reflects the intrinsic fuzziness or stochasticity in the data itself, quantified by the Shannon entropy $\mathbb{H}[\boldsymbol{\pi^*}]$.

Estimations. AU is estimated by some form of $\mathbb{H}[\boldsymbol{\pi}]$, which holds when $\boldsymbol{\pi} \approx \boldsymbol{\pi^*}$. This is true for low-EU data constrained by ERM, such that $w(\boldsymbol{x}) \approx w^*(\boldsymbol{x})$, $\forall w$, despite $w \neq w^*$. Thus, low EU is a necessity for AU estimations [19, 51]. EU estimation is more elusive due to its ill-defined nature. Bayesian methods learn the posterior distribution of w [3, 11, 25, 31, 53], with information-theoretic decomposition of total entropy into conditional entropy (AU) and mutual information (EU) [8, 21, 44], but this relies on improper metrics [51] and neglects the EU's bias term [24]. Bregman decomposition [14] addresses these issues but still struggles to unmix EU/AU [34], highlighting the limitations of such formulations [20, 34, 38, 48, 49, 51]. Additionally, they face scalability issues from multiple DNN Monte Carlo forward passes to perform marginalization. To address this, **deterministic methods**, such as evidential learning [6], excess risk predictors [24], and latent space estimators [26, 36, 42, 46], measure EU through a single DNN pass. Among these, a recent benchmark [34] shows that latent spaces are especially effective at isolating EU by measuring post-hoc[1] "distance" to low-EU samples in $\mathcal{D}$. This assumes similar data cluster in DNN latent spaces—a property synonymous with well-learned representations—thereby explaining the strong generalization of latent space EU estimators, irrespective of architecture or training method.

Static *vs.* Dynamic. Most deep learning frameworks operate in a *static* joint space $X \times Y$, where AU is deemed irreducible (*cf.* EU is reducible via additional data or better choice of $\mathcal{H}$). However, in a *dynamic* LI$\Rightarrow$HI scenario, AU at LI can be resolved by transitioning to a richer, higher-dimensional HI space. However, this shift increases EU, as more complex inputs are inherently harder to learn ("curse of dimensionality"). See §2.3 in [19].

Fine-Grained Disentanglement. A few studies have explored EU subcategories. For example, [32, 45] divide EU into OOD *vs.* miscalibrated in-distribution cases, while [47, 52] differentiate OOD uncertainty by semantic *vs.* covariate shifts. Our study differs by focusing on finer AU subcategories.

3 Uncertainty Quantification Method

Problem definition. Suppose two domains, LI and HI, have shared label space but distinct DNN models trained on identical tissue samples that differ only in imaging quality. We operate primarily in LI and reserve HI data queries for uncertain LI samples that become certain in HI; our goal is to design a decision function that effectively performs this task.

---

[1] Some works incorporate regularization like spectral norm for bi-Lipschitzness [36] to prevent feature collapse, though studies without it are also successful [26, 46].

Fig. 2: **Visualization of uncertainty categories.** Left to right: static scenarios at LI and HI, and dynamic scenario with LI⇒HI transition (ours).

3.1   Static Uncertainty Taxonomy

We begin by considering uncertainty estimates in a static scenario, applying to both LI and HI domains. Let $w = \mathtt{softmax} \circ g \circ f$, where f is the feature extractor and g is the final layer such that $\boldsymbol{z} = f(\boldsymbol{x})$. The generalized formalism for latent space EU and AU estimates is given by:

$$EU(\boldsymbol{x}) \triangleq d(\boldsymbol{z}; \mathcal{D}_v),$$

$$AU(\boldsymbol{x}) \triangleq \begin{cases} \mathbb{H}[\boldsymbol{\pi}] & \text{if} \quad EU(\boldsymbol{x}) < \tau_{EU}, \\ \text{n/a} & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{D}_v \subseteq \mathcal{D}$ is a training subset (or validation set, if allowed), and $d(\cdot; \mathcal{D}_v)$ is a distance function $w.r.t.$ $\{f(\boldsymbol{x_v}) \mid \boldsymbol{x_v} \in \mathcal{D}_v\}$, where larger values indicate a higher EU. We define a decision function $\Phi_{\mathrm{St.}}(\cdot)$ with three categories:

$$\Phi_{\mathrm{St.}}(\boldsymbol{x}) \triangleq \begin{cases} \text{Certain: } \mathtt{C} & \text{if} \quad \big(EU(\boldsymbol{x}) < \tau_{EU}\big) \wedge \big(AU(\boldsymbol{x}) < \tau_{AU}\big), \\ \text{Uncertain} \begin{cases} \text{Aleatoric: } \mathtt{UA} & \text{if} \quad \big(EU(\boldsymbol{x}) < \tau_{EU}\big) \wedge \big(AU(\boldsymbol{x}) \geqslant \tau_{AU}\big), \\ \text{Epistemic: } \mathtt{UE} & \text{otherwise} \quad EU(\boldsymbol{x}) \geqslant \tau_{EU}, \end{cases} \end{cases} \tag{2}$$

with thresholds τ_{EU} and τ_{AU} governing category boundaries. While these are, in principle, user-defined hyperparameters, we follow literature-guided protocols to methodologically set them (described later). We schematically delineate $\Phi_{\mathrm{St.}}$ for the LI and HI spaces in Fig. 2. As per [19], EU and AU boundaries tighten in HI, resulting in larger and smaller fractions of $\mathtt{UE}$ and $\mathtt{UA}$ data, respectively.

3.2   Dynamic Uncertainty Taxonomy

We now consider the transition from LI to HI, with notations denoted by superscripts L and H, respectively. As evident from Fig. 2, not all $\mathtt{UA}$ data at LI are $\mathtt{C}$ at HI (*i.e.*, resolvable)—some become $\mathtt{UE}$, while others remain as $\mathtt{UA}$. We seek only the resolvable cases, grouping the rest as irresolvable. Here, we define a new fine-grained uncertainty taxonomy, $\Phi^{\mathrm{Dy.}}(\cdot)$, $w.r.t.$ LI data $\boldsymbol{x}^{\mathrm{L}}$:

$$\Phi_{\mathrm{Dy.}}(\boldsymbol{x}^{\mathrm{L}}) \triangleq \begin{cases} \text{Cert.: } \mathtt{C} & \text{if} \quad \Phi^{\mathrm{L}}_{\mathrm{St.}}(\boldsymbol{x}^{\mathrm{L}}) = \mathtt{C}, \\ \text{Uncert.} \begin{cases} \text{Al.} \begin{cases} \text{Res.: } \mathtt{UAR} & \text{if} \quad \big(\Phi^{\mathrm{L}}_{\mathrm{St.}}(\boldsymbol{x}^{\mathrm{L}}) = \mathtt{UA}\big) \wedge \big(\Phi^{\mathrm{H}}_{\mathrm{St.}}(\boldsymbol{x}^{\mathrm{H}}) = \mathtt{C}\big), \\ \text{Irres.: } \mathtt{UAI} & \text{if} \quad \big(\Phi^{\mathrm{L}}_{\mathrm{St.}}(\boldsymbol{x}^{\mathrm{L}}) = \mathtt{UA}\big) \wedge \big(\Phi^{\mathrm{H}}_{\mathrm{St.}}(\boldsymbol{x}^{\mathrm{H}}) \neq \mathtt{C}\big), \end{cases} \\ \text{Ep.: } \mathtt{UE} & \text{otherwise} \quad \Phi^{\mathrm{L}}_{\mathrm{St.}}(\boldsymbol{x}^{\mathrm{L}}) = \mathtt{UE}. \end{cases} \end{cases} \tag{3}$$

During test-time, explicitly using this framework is infeasible as it requires foreknowledge of HI data from $\Phi^{\mathrm{H}}_{\mathrm{St.}}(\boldsymbol{x}^{\mathrm{H}})$. Instead, following similar approach used for

EU estimation, we approximate the decision criteria of classes $\forall \phi \in \{\texttt{UAR}, \texttt{UAI}\}$ by leveraging latent space distances in the LI domain, as follows:

$$\Phi_{\text{Dy.}}(\boldsymbol{x}^{\text{L}}) \approx \phi \quad \text{if} \quad \arg\min_\phi d(\boldsymbol{x}^{\text{L}}; \mathcal{D}_v^\phi) = \phi,$$
$$s.t. \quad \mathcal{D}_v^\phi \triangleq \{\boldsymbol{x}_{\boldsymbol{v}}^{\text{L}} \mid \Phi_{\text{Dy.}}(\boldsymbol{x}_{\boldsymbol{v}}^{\text{L}}) = \phi, \forall \boldsymbol{x}_{\boldsymbol{v}} \in \mathcal{D}_v\}. \tag{4}$$

That is, we assign $\texttt{UAR}$ *vs.* $\texttt{UAI}$ by evaluating the proximity to respective LI prototypes within $\mathcal{D}_v$, for which we can compute Eq. 3, since both their LI and HI data are available. The surrogate Eq. 4 obviates the need for HI test data, thereby enabling blind uncertainty inference directly in LI, while operating post-hoc with minimal overhead—an advantage in digital pathology, where data can sum up to TBs. However, since we aim to emulate Eq. 3, its performance is upper-bounded by it. This hinges on static uncertainties at LI and HI (Eqs. 1-2), and while we opt for latent space formulations due to their state-of-the-art performance [34], it is not restricted to this backbone. Unless specified otherwise, we use the popular Mahalanobis distance (MD) [9, 26, 41] for all $d(\cdot; \cdot)$ metrics, measuring proximity to nearest Gaussian class centroid.

3.3   Application in Adaptive Imaging

Selecting class-$\texttt{UAR}$ LI data for HI query yields the final adaptive prediction:

$$\boldsymbol{\pi}^{\text{A}} = \underbrace{\mathbb{1}[\Phi_{\text{Dy.}}(\boldsymbol{x}^{\text{L}}) \neq \texttt{UAR}] \cdot w^{\text{L}}(\boldsymbol{x}^{\text{L}})}_{\text{Existing default LI data}} + \underbrace{\mathbb{1}[\Phi_{\text{Dy.}}(\boldsymbol{x}^{\text{L}}) = \texttt{UAR}] \cdot w^{\text{H}}(\boldsymbol{x}^{\text{H}})}_{\text{Newly queried HI data}}. \tag{5}$$

Note that not all data are certain—only $\texttt{C}$ and, for well-approximated Eq. 4, $\texttt{UAR}$. Let T^{H} and T^{L} represent the imaging complexity in HI and LI, respectively. The cost of Eq. 5 is then $T^{\text{A}} = T^{\text{L}} + P_{\boldsymbol{x} \sim X}(\Phi_{\text{Dy.}}(\boldsymbol{x}^{\text{L}}) = \texttt{UAR}) \cdot T^{\text{H}}$. Still, this may be prohibitive with larger $\texttt{UAR}$ ratios or tighter budgets. In such circumstances, we chose by ascending values of $d(\boldsymbol{x}^{\text{L}}; \mathcal{D}_v^{\texttt{UAR}})$, stopping once the allowable query limit is reached; this prioritizes data with higher "confidence" of belonging to $\texttt{UAR}$.

4   Experiments

4.1   Experimental setting

Here, we describe the conducted experiments to showcase our approach.

Dataset. We use the IR dataset of a breast cancer tissue microarray consisting of 101 cores from 47 patients, provided by [33]. We select 10 IR spectral bands at 1.1µm/pixel resolution as HI, simulating the LI counterparts by reducing the number of bands to 4 and spatially downsampling by $10\times$—a $250\times$ difference in imaging complexity. The downstream task is 6-class segmentation by epithelial and stromal types, with an 80-20% training-test split. Since segmentation is dense per-pixel classification, all notations and concepts transfer.

Implementations. We train four models[2] per domain (with differing seeds), resulting in 16 adaptive imaging outcomes per pair. As the task is segmentation,

---

[2] Attention-Unet architecture [39], trained for ~5K iterations using a batch size of 16, focal loss [28], pixel dropout, and AdamW [29] with an initial learning rate of 1e-4.
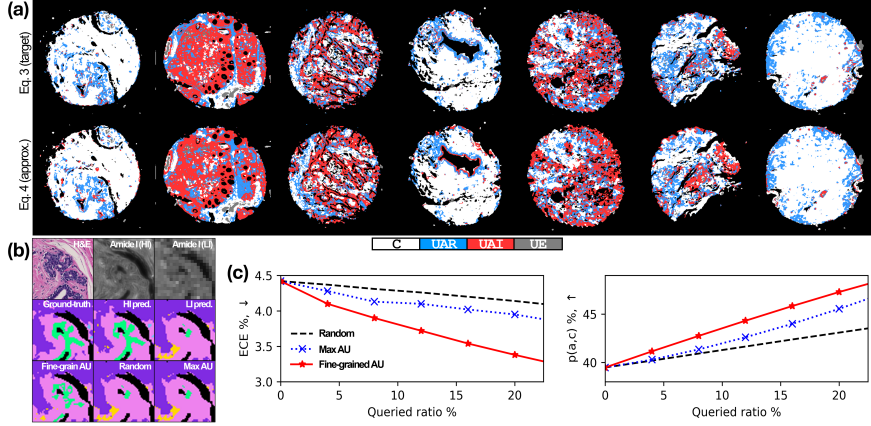
Fig. 3: **Results. (a)** Comparison of $\Phi_{\text{Dy.}}$: Eq. 3 (top row) *vs.* Eq. 4 (bottom row). **(b)** Example segmentation results via adaptive imaging using different query strategies. **(c)** Segmentation performance metrics under tighter query budgets.

we generate dense uncertainty maps from the final feature map, with adaptive imaging occurring at the pixel level. To reduce uncertainty inference latency, we downscale these maps by $4\times$. We set implementation hyperparameters as so: For $\mathcal{D}_v$, we draw $\sim$6K training pixels stratified by class and core, following the smallest sampling ratios in [22, 26, 46] for stringent conditions. τ_{EU} is set to achieve a 95% true positive rate in the low-EU training set, as per routine OOD detection evaluation protocols [5,15,26,46]. τ_{AU} is calibrated to maximize the number of accurate C data, matching the $P(accurate, certain)$ metric [35]. We also tested transformer architectures [4] and hyperparameter configurations, reaching the same conclusion (omitted due to space constraints).

Baselines. We compare against two query baselines: *random*, and *maximum AU* among samples where $\Phi_{\text{St.}}^{\text{L}}(\boldsymbol{x}^{\text{L}}) = \text{UA}$. These baselines serve to prove our core claims: the random baseline asserts the value of prioritizing high-AU pixels, while the monolithic max-AU baseline tests whether finer subcategorization yields additional benefit. Although alternative (non-uncertainty-based) sampling strategies exist, we do not consider them here due to space constraints.

4.2   Results & Discussion

Next, we present our experimental findings, with in-depth analysis.

Fine-grained AU disentanglement via latent spaces is feasible. As a sanity check, we first examine whether Eq. 3 using latent spaces can approximate Eq. 4. We compare the uncertainty class maps for selected test cores in Fig. 3-(a), observing a strong agreement between the two, with a $\sim$60% mean F1 score for UAR and UAI. This affirms the feasibility of our approach: *latent spaces can not only capture EU extent but also disentangle fine-grained AU subtypes.* We hypothesize this is due to UAR and UAI being low-EU categories, resulting

Table 1: **Comparison of query methods.** The best method is in bold. The LI/HI performance metrics are as follows—F1%: $53_{\pm0.8}/60.5_{\pm1.3}$, ECE%: $4.3_{\pm0.4}/2.1_{\pm0.8}$, and $P(a,c)$%: $41.6_{\pm2}/54.6_{\pm2.5}$.

Method	Unconstrained, $T^A \approx 50$			Tighter constraints, $1 < T^A < 50$		
	F1%,↑	ECE%,↓	$P(a,c)$%,↑	$\int$F1%,↑	$\int$ECE%,↓	$\int P(a,c)$%,↑
Random	$54.46_{\pm0.55}$	$3.77_{\pm0.39}$	$44.04_{\pm1.38}$	$53.70_{\pm0.58}$	$4.06_{\pm0.35}$	$42.75_{\pm1.54}$
Max AU	$\mathbf{55.80_{\pm0.54}}$	$3.51_{\pm0.46}$	$46.51_{\pm1.59}$	$\mathbf{54.45_{\pm0.49}}$	$3.89_{\pm0.36}$	$43.59_{\pm1.64}$
Fine-grained AU	$54.90_{\pm0.81}$	$\mathbf{3.16_{\pm0.53}}$	$\mathbf{49.84_{\pm1.52}}$	$53.97_{\pm0.77}$	$\mathbf{3.38_{\pm0.46}}$	$\mathbf{47.98_{\pm1.75}}$

in robust, accurate latent representations—facilitating separation via a suitable distance function $d(\cdot;\cdot)$. Notably, we made no assumptions about the LI/HI domains, rendering our approach versatile and extendable to any uncertainty taxonomy, provided their uncertainty class prototypes are accessible.

Querying by AU resolvability is efficacious. In Fig. 3-(b), we show an example of a segmented region with a fixed adaptive query budget, where our method approaches the ground-truth most proficiently. We quantify three segmentation metrics in Tab. 1, with arrows indicating the desired direction: F1 (↑), Expected Calibration Error (ECE, ↓) [23], and the above $P(a,c)$ (↑) [35]. We evaluate two budget scenarios—*(i) Unconstrained*, allowing an arbitrary number of queries (Eq. 5). In our dataset, the proportion of UAR was ∼20%, corresponding to a cost $T^A = 50$ when denoting T^L and T^H as 1 and 250 (a.u.). *(ii) Constrained*, with tighter budgets. We compute the metrics spanning $1 < T^A < 50$ (Fig. 3-(c)) and report the area-under-curve (AUC), denoted by $\int$. The results show that our method achieves superior ECE and $P(a,c)$, with gains over LI of 1.14/8.21% and 0.92/6.35% for $T^A = 50$ and $1 < T^A < 50$, respectively. Although these improvements may appear modest due to the upper bound imposed by HI performance, they constitute substantial relative gains—exceeding the random and max-AU baselines by **2.1–5.7×** and **1.4–3.2×**, respectively. While the max-AU marginally outperforms in F1, its inferior ECE and $P(a,c)$ indicate a tendency to select high-AU samples that, while becoming more accurate, remain poorly calibrated or uncertain (i.e., do not cross the τ_{AU} threshold); arguably, re-imaging such samples are less valuable.

Assumption violations cause failures. Recall that our approach hinges on two core premises—(i) the strong base performance of static LI and HI uncertainties, and (ii) the well-approximation of Eq. 4. A failure in either premise undermines our method. To demonstrate this, we conduct ablation experiments: (i) We replace the latent space static uncertainties with an inferior alternative—the information-theoretic EU/AU estimates, which are popular in medical fields [17], despite known limitations [20, 34, 38, 48, 49, 51]. Specifically, we adopt Deep Ensemble (DE) [25], where the four DNNs per domain are treated as constituents. We use the same MD-based Eq. 4, but with averaged distances across DE members. (ii) We compute MD using only the feature's 1$^{\text{st}}$ principal component (PC) in Eq. 4, anticipating diminished performance due to the loss of

Table 2: **Other configurations.** First row is the original version, followed by ablations and design variations. For ablations, we highlight in red if there is a significant drop in performance. No ±std. is reported for the DE ablation, as we aggregate all models. We also report uncertainty inference latency $(s/mm^2, \downarrow)$ on an i9-9900KF CPU in parentheses for the original method and design variations.

Method	$\tau, \approx 0$	AUCC,↓	Apx.F1%,↑	$\int$F1%,↑	$\int$ECE%,↓	$\int P(a,c)$%,↑
Original; MD w/o DR (0.97)	-0.94±5.68	2.46±0.70	60.39±1.27	53.97±0.77	3.38±0.46	47.98±1.75
Abl. (i): $\Phi_{St.}$ w/ DE	30.88	6.58	65.38	55.58	6.33	44.95
Abl. (ii): d in Eq. 4 w/ PC1	-0.94±5.68	2.46±0.70	49.08±0.52	53.78±0.61	3.88±0.34	42.99±1.59
KNN w/o DR (3.11)	7.92±5.16	2.88±1.25	60.70±1.79	53.70±0.77	3.22±0.43	48.03±1.47
MD w/ DR (0.82)	-9.24±11.2	3.14±1.18	59.65±2.58	53.85±0.77	3.33±0.43	47.24±1.26

information; this modification is applied solely to Eq. 4, leaving the static latent space uncertainties (Eqs. 1-2) unaltered.

In Tab. 2, we report the segmentation metrics for these ablations. Since the exact number of UAR data slightly varied between methods, we report only the constrained budget scenario to ensure a fair comparison. Additionally, we evaluate the satisfaction of hypotheses (i)-(ii) as follows: (i) Following [34], we adopt two metrics to assess static uncertainties, averaged across LI and HI—rank correlation (Kendall's τ, where ≈0 is desired) between EU and AU to measure the degree of entanglement, and the AUC of Calibration-Coverage (AUCC, ↓), where we abstain from predictions by decreasing EU and compute the ECE of the remaining. (ii) As previously done, we report the mean F1 (↑) score of UAR and UAI approximations. As expected, targeted ablations cause a drop in the corresponding metric(s). Consequently, we observe decreased segmentation performance, highlighting the importance of maintaining assumptions.

Robustness to design variations. Alternative distances for $d(\cdot; \cdot)$, such as KNN [46], are also viable, while some works apply dimensionality reduction (DR) to mitigate the curse of dimensionality [12, 27]. We also explore these options in Tab. 2. Setting $k = 100$ for KNN, we observed slightly better performance due to its nonparametric nature. However, the gap is small, leading us to choose the more efficient MD with an uncertainty latency of $0.97s/mm^2$. For a typical whole-slide tissue image covering a $1cm^2$ area, this translates to 1-2 minutes, which can be reduced by further downsampling the uncertainty map, making it highly practical. Note, however, that we expect KNN to significantly outperform MD for heterogeneous or long-tailed latent distributions. For DR, we tried multiple manifold learning algorithms like Isomap, LLE, and Parametric UMAP [43], reporting only the latter, which performed the best. However, it provided no significant improvement, likely because our final feature vector dimension is only 64, typical in segmentation DNNs. As such, the benefit of DR is weaker, potentially offset by information loss. However, we anticipate that it will be more crucial in image classification extensions of our task, where feature

dimensionality is much larger (512+). Nonetheless, all variants performed well, supporting the robustness of our approach.

4.3   Concluding Comments

Having shown proof-of-concept, we conclude with several avenues for future research. (i) Even finer disentanglement: In this work, LI degradation occurred along two axes—chemical (via band reduction) and spatial (via sparse sampling)—which we abstracted into a binary UAR vs. UAI distinction. However, not all UAR data are equal; some can be more cheaply resolved by correcting just one dimension. This suggests that adaptive imaging could be further optimized through even finer (2+) AU disentanglement. (ii) Algorithmic improvements: One direction is to enhance static uncertainty estimates, e.g., through graph-based OOD/failure detectors [13, 22] for EU and heteroscedastic DNNs [7, 21] for AU. Another question is how to optimize $d(\cdot; \cdot)$ for fine-grained separation tasks. (iii) Variable LI/HI cost differential: While we assume a fixed cost ratio, reducing LI cost could enable larger query budgets and potentially improve net re-imaging quality. Although our approach assumes no specific cost structure and is thus broadly applicable, further analysis is needed to determine optimal strategies under varying cost regimes. (iv) Large-scale validations: Due to the lack of publicly available label-free histology datasets, we limited our study to just one. We plan to pursue further validation upon the release of more datasets.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Bernhardt, M., et al.: Failure detection in medical image classification: A reality check and benchmarking testbed. TMLR (2022)
2. Bhargava, R.: Digital histopathology by infrared spectroscopic imaging. Annu. Rev. Anal. Chem. (2023)
3. Blundell, C., et al.: Weight uncertainty in neural network. In: ICML (2015)
4. Cao, H., et al.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV (2022)
5. Chan, R., et al.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: ICCV (2021)
6. Charpentier, B., et al.: Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In: NeurIPS (2020)
7. Collier, M., et al.: Massively scaling heteroscedastic classifiers. In: ICLR (2023)

8. Depeweg, S., et al.: Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In: ICML (2018)
9. Dua, R., et al.: Task agnostic and post-hoc unseen distribution detection. In: WACV (2023)
10. Fereidouni, F., et al.: Microscopy with ultraviolet surface excitation for rapid slide-free histology. Nat. Biomed. Eng. (2017)
11. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML (2016)
12. Ghosal, S.S., et al.: How to overcome curse-of-dimensionality for out-of-distribution detection? In: AAAI (2024)
13. Goodge, A., et al.: Lunar: Unifying local outlier detection methods via graph neural networks. In: AAAI (2022)
14. Gruber, S., Buettner, F.: Uncertainty estimates of predictions via a general bias-variance decomposition. In: AISTATS (2023)
15. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
16. Hoover, E.E., Squier, J.A.: Advances in multiphoton microscopy technology. Nat. Photon. (2013)
17. Huang, L., et al.: A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods. MedIA (2024)
18. Huang, L., et al.: Rapid, label-free histopathological diagnosis of liver cancer based on raman spectroscopy and deep learning. Nat. Commun. (2023)
19. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. (2021)
20. de Jong, I.P., et al.: How disentangled are your classification uncertainties? arXiv (2024)
21. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NeurIPS (2017)
22. Kim, J.H., et al.: Neural relation graph: a unified framework for identifying label noise and outlier data. In: NeurIPS (2024)
23. Kumar, A., et al.: Verified uncertainty calibration. In: NeurIPS (2019)
24. Lahlou, S., et al.: Deup: Direct epistemic uncertainty prediction. TMLR (2023)
25. Lakshminarayanan, B., et al.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017)
26. Lee, K., et al.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)
27. Li, X., et al.: Characterizing submanifold region for out-of-distribution detection. IEEE TKDE (2024)
28. Lin, T.Y., et al.: Focal loss for dense object detection. In: ICCV (2017)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
30. Ma, S., et al.: Breaking the barrier: Selective uncertainty-based active learning for medical image segmentation. In: ICASSP (2024)
31. Maddox, W.J., et al.: A simple baseline for bayesian uncertainty in deep learning. In: NeurIPS (2019)
32. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. In: NeurIPS (2018)
33. Mittal, S., et al.: Simultaneous cancer and tumor microenvironment subtyping using confocal infrared microscopy for all-digital molecular histopathology. PNAS (2018)
34. Mucsányi, B., et al.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In: NeurIPS (2024)

35. Mukhoti, J., Gal, Y.: Evaluating bayesian deep learning methods for semantic segmentation. arXiv (2018)
36. Mukhoti, J., et al.: Deep deterministic uncertainty: A new simple baseline. In: CVPR (2023)
37. Nath, V., et al.: Diminishing uncertainty within the training pool: Active learning for medical image segmentation. IEEE TMI (2020)
38. Oh, J.H., et al.: Are we ready for out-of-distribution detection in digital pathology? In: MICCAI (2024)
39. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. In: MIDL (2018)
40. Park, Y., et al.: Quantitative phase imaging in biomedicine. Nat. Photon. (2018)
41. Podolskiy, A., et al.: Revisiting mahalanobis distance for transformer-based out-of-domain detection. In: AAAI (2021)
42. Postels, J., et al.: On the practicality of deterministic epistemic uncertainty. In: ICML (2022)
43. Sainburg, T., et al.: Parametric umap embeddings for representation and semisupervised learning. Neural Comput. (2021)
44. Smith, L., Gal, Y.: Understanding measures of uncertainty for adversarial example detection. In: UAI (2018)
45. Sun, H., et al.: What is flagged in uncertainty quantification? latent density models for uncertainty categorization. In: NeurIPS (2023)
46. Sun, Y., et al.: Out-of-distribution detection with deep nearest neighbors. In: ICML (2022)
47. Tian, J., et al.: Exploring covariate and concept shift for out-of-distribution detection. In: NeurIPS W. (2021)
48. Ulmer, D., Cinà, G.: Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection. In: UAI (2021)
49. Valdenegro-Toro, M., Mori, D.S.: A deeper look into aleatoric and epistemic uncertainty disentanglement. In: CVPR W. (2022)
50. Wang, K., et al.: Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation. In: MICCAI (2021)
51. Wimmer, L., et al.: Quantifying aleatoric and epistemic uncertainty in ml: Are conditional entropy and mutual information appropriate measures? In: UAI (2023)
52. Yang, J., et al.: Full-spectrum out-of-distribution detection. IJCV (2023)
53. Zhang, R., et al.: Cyclical stochastic gradient mcmc for bayesian deep learning. In: ICLR (2020)