

Conformal Annotation with Randomized Expert Review: Application to Online Cell-Type Labeling

Xuhe Zhang¹, Zhi Huang², and Ying Jin^{*1,3}

¹Applied Mathematics and Computational Science, University of Pennsylvania

²Department of Pathology and Laboratory Medicine, University of Pennsylvania

³Department of Statistics and Data Science, University of Pennsylvania

Abstract

Artificial intelligence (AI) models can accelerate cell-type annotation in digital pathology, but safely scaling their use requires determining—with error control—which model-generated labels can be accepted without expert review. This is challenging in evolving workflows because cell morphology and type composition may change across tissue regions and slides, making one-time calibration unreliable. We propose *Conformal Annotation with Randomized Expert review under Distribution shift* (CARED), a distribution-free framework that leverages randomized expert audit to achieve long-run error control for AI-assisted annotation under distribution shift. For each incoming batch, CARED uniformly selects a subset of cases for expert review and uses the acquired labels to calibrate later decisions. Among unreviewed cases, CARED determines a subset of them to safely skip expert labeling, while suggesting candidate labels for the others for downstream resolution. We prove that this workflow provides long-run control of cumulative error over both the full stream and the cells outside of the audit sample, without stationarity or exchangeability assumptions. We further enhance the framework by introducing class-conditional guarantees to ensure error control within important but rare cell types, together with a utility module to target more automatic resolutions, smaller candidate sets, and greater biological coherence, without compromising validity. Experiments on pancreas, lung, and ovary annotation streams, together with a stream constructed from real expert annotations, show that CARED controls the cumulative errors while a substantial fraction of cells (up to 70%) bypass expert review.

Keywords: Conformal inference; Trustworthy medical AI; Distribution shift; AI annotation

1 Introduction

Expert labeling underlies a broad range of scientific, engineering, and medical tasks. Experts examine and assign labels that guide downstream analyses or decisions on individual cases; the resulting labeled cases also form the datasets used to train and evaluate predictive models (Deng et al., 2009; Gulshan et al., 2016). Artificial intelligence (AI) models can accelerate this process by producing predictions at scale which, when sufficiently accurate, can serve as labels (Goel et al., 2023; Gilardi et al., 2023), thereby reducing or even bypassing expert review. However, accepting those “labels” without expert review leaves the prediction errors fundamentally unobserved (Northcutt et al., 2021; Bernhardt et al., 2022). This creates a tension between reducing expert effort and maintaining label quality (Candès et al., 2025; Huang et al., 2025).

We study this problem in image-based cell-type annotation, an important task in biomedical research and digital pathology (Ergen et al., 2024; Hou and Ji, 2024). In this task, a digitalized pathology slide contains many cells, each of which must be assigned a cell type to support downstream analysis (Angus et al., 2025;

*Correspondence to yjinstat@wharton.upenn.edu

Kim et al., 2026; Gaffney, 2025). Emerging AI systems now enable collaboration between pathologists and cell-level AI models: as pathologists proceed across a slide image, the model proposes labels for the cells, which can be reviewed by the pathologists to assign the labels. For example, the TissueLab platform (Li et al., 2025a) developed by the author(s) allows pathologists to navigate the slides and annotate cells with the support of an in-house AI model (Xu et al., 2025). Such collaborative systems streamline expert annotation, but every model proposal must still be inspected and finalized by a pathologist. Here, our goal is to identify cells whose AI predictions can be accepted without expert review while still ensuring long-run error control.

Recent work has begun to provide statistical guarantees for deciding which model-generated labels can be accepted without expert review. Assuming access to i.i.d. (or a uniform random subset of) labeled data, PAC Labels (Candès et al., 2025) calibrates a criterion for accepting AI labels in the unlabeled data while controlling the overall error rate, whereas Conformal Labeling (Huang et al., 2025) uses selective inference (Jin and Candès, 2023) to select AI labels for unlabeled data while controlling the false discovery rate among the selected labels. Their success relies on the data generating process ensuring the AI labeling errors in the labeled data are representative of the unknown errors in the new samples.

However, as a pathologist moves across regions of a whole-slide image (WSI) and then across slides, the cells may differ in both morphology and cell-type composition. Additional variation can arise from changes in specimen preparation, imaging devices, and data processing (Tellez et al., 2019; Stacke et al., 2020; Howard et al., 2021; Aubreville et al., 2023). Therefore, a rule calibrated on earlier data may fail to provide valid error control as annotation proceeds. Figure 1 visualizes such workflow shift using the NuClass model of Xu et al. (2025). We partition a pancreas WSI containing more than 10^5 cells into 206 approximately equal-sized spatial regions. We then order the regions to mimic a pathologist’s traversal across neighboring fields, with adaptive changes in viewing resolution; panel (a) shows this annotation order. For this diagnostic, we use the held-out label of each cell to compute the NuClass score assigned to its true class. Panel (b) reports the within-fraction empirical α -quantiles of these scores, separately for each cell type. The dashed rose curve in each panel, read against the right vertical axis, gives the cell-type proportion within each fraction. The composition of cell types varies substantially, irregularly, and at times abruptly across batches. Moreover, the in-class distributions also vary substantially: if the cells within the same type are identically distributed across batches, the quantile curves would remain approximately constant (up to sampling variation). Instead, the curves fluctuate markedly over the annotation order. These within-cell-type distributional changes, together with the composition shifts, make one-time calibration unreliable.

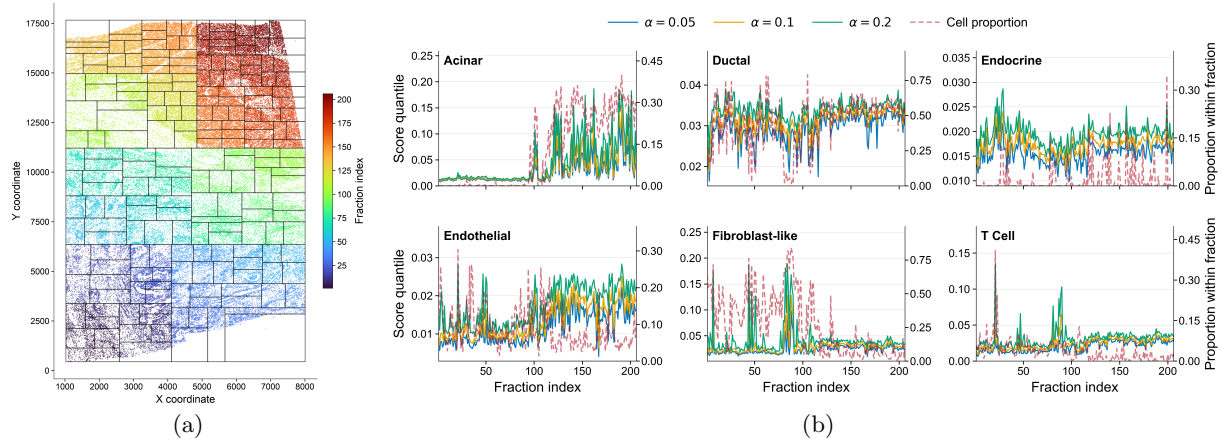


Figure 1: Evidence of workflow shift in image-based cell-type annotation. A pancreas whole-slide image containing more than 10^5 cells is partitioned into 206 approximately equal-sized spatial fractions. (a) Spatial locations of cells, with color indicating the order in which the fractions enter the annotation workflow. (b) Empirical α -quantiles of the NuClass score assigned to the true class, shown separately by cell type for $\alpha \in \{0.05, 0.10, 0.20\}$. The dashed rose curve in each panel (right vertical axis) shows the corresponding cell-type proportion. Its irregular and, at times, abrupt changes reveal shifts in cell-type composition, while the variation in the solid quantile curves reveals within-cell-type score-distribution changes over the workflow.

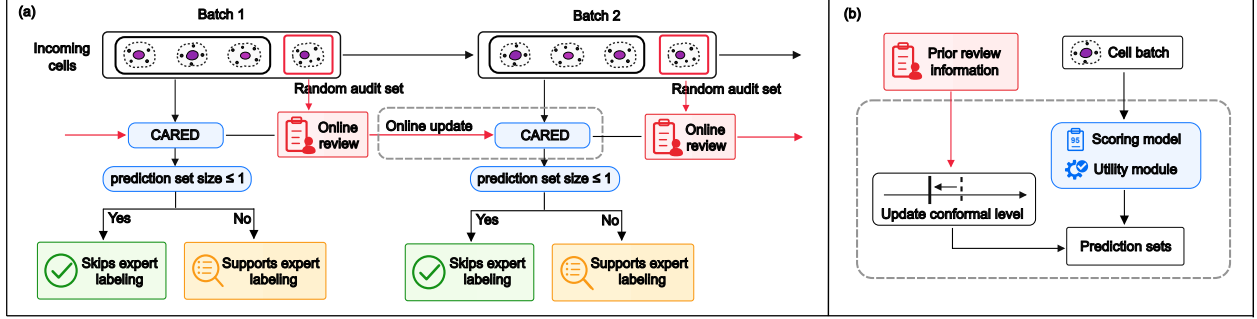


Figure 2: Schematic of the CARED workflow. (a) Incoming cell batches are processed sequentially: CARED forms prediction sets, cells with set size at most one skip expert labeling, and a random audit set is reviewed to update the conformal level for subsequent batches. (b) Zoom-in of the prediction set generation process highlighted by the gray dashed box in panel (a), where prior review information updates the conformal level and the scoring model with the utility module produces prediction sets for the incoming batch.

1.1 Overview and contributions

We introduce *Conformal Annotation with Randomized Expert review under Distribution shift* (CARED), an online framework that uses conformal prediction sets and randomized expert audits to automate a subset of annotations in an evolving data stream. CARED is designed for cell-type annotation, but its core construction applies to other multiclass labeling workflows in which limited expert feedback is available.

The first ingredient of CARED is to derive long-run control of cumulative annotation error from coverage of prediction sets (Vovk et al., 2005). At round t , let D_t denote the incoming unlabeled batch. For each cell $x \in D_t$, the model scores are converted to a prediction set $\hat{C}_t(x) \subseteq \mathcal{Y}$ of plausible cell types, where $\mathcal{Y}$ denotes the label space. CARED selects cells with $|\hat{C}_t(x)| \leq 1$ for automatic annotation; otherwise, the set is returned to the pathologist as a candidate list for later resolution. We refer to the fraction of cells that bypass expert review as the *skipped ratio*. Let $I_t(x)$ denote whether an automated annotation is produced, let $y(x)$ denote the true cell type, and let $\hat{y}_t(x)$ denote the automated annotation when it exists. Then

$$\mathbf{1}\{\hat{y}_t(x) \neq y(x)\} \leq \mathbf{1}\{y(x) \notin \hat{C}_t(x)\} \quad \text{whenever } I_t(x) = 1.$$

Thus, the fraction of prediction sets failing to cover the true labels (miscoverage) over the unreviewed stream upper bounds the long-run fraction of unreviewed cells that receive erroneous automatic annotations.

A natural approach to maintain this guarantee as the annotation workflow changes is to recalibrate the prediction sets in an online fashion. Adaptive conformal inference (ACI) (Gibbs and Candès, 2021; Gibbs and Candès, 2024) provides a starting point: it updates prediction sets using newly observed labels and provides long-run control of cumulative miscoverage regardless of the distributional changes. The standard ACI, however, requires observing the true label after every prediction. Applied directly to annotation, it would request expert feedback for every incoming cell, leaving the burden of pathologist review unchanged.

CARED resolves this tension by proposing a novel random review mechanism. After prediction sets $\hat{C}_t(x)$ are constructed for $x \in D_t$, a subset $A_t \subseteq D_t$ with fraction $p \in (0, 1)$ is sampled uniformly for expert review (audit). The reviewed cells provide unbiased feedback for the whole batch. We prove that updating with the audited labels is sufficient to provide long-run control of cumulative prediction-set miscoverage over both the full stream and the cells not selected for audit. Such a guarantee is ensured by the random review mechanism only, requiring no exchangeability or stationarity of the data stream, and allows the batches and prediction models to evolve arbitrarily over time. An overview of CARED is in Figure 2.

For cell-type annotation, we further enhance the core framework in two ways. First, we use class-conditional calibration (Vovk et al., 2003) to achieve the same long-run guarantee within each true cell type, ensuring accuracy in clinically important but rare and challenging cell types. Second, we develop a utility-aware module that updates the AI scores to favor more accepted annotations, smaller candidate sets, and biologically coherent outputs.

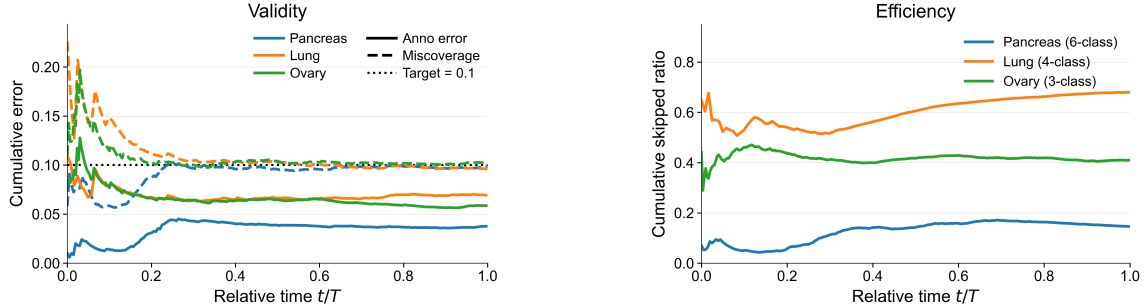


Figure 3: Validity and efficiency of CARED across three tissue-specific tasks using the global NuClass score source (Section 4). The x -axis indicates the fraction of processed batches in the stream. (a) Cumulative annotation error (solid lines) and miscoverage (dashed lines) for the three tasks; the dotted horizontal line denotes the target error level $\alpha = 0.1$. (b) Fraction of cases bypassing expert review among all incoming ones for the three tasks at $\alpha = 0.1$.

We instantiate CARED with NuClass (Xu et al., 2025) in image-based cell-type annotation and evaluate it on streams constructed from spatially partitioned WSIs across multiple tissue types, together with a stream based on real expert annotations from TissueLab (Li et al., 2025a). Our experiments emulate various workflow shifts in real annotation processes to examine the long-run error control, reduction in expert review, prediction-set utility, and sensitivity to the audit budget and target error level. Figure 3 previews the main validity and efficiency results for CARED. Although only a fraction ($p = 0.1$) of each incoming batch is selected for expert audit, the cumulative annotation error (solid lines) remains below the target level, whereas the cumulative miscoverage (dashed lines) moves toward and remains near the target level ($\alpha = 0.1$) as the annotation streams progress. At the same time, panel (b) shows that the fraction of cells skipping expert review reaches as high as 70%, with its value varying across tissues and label spaces. Together, the two panels illustrate how randomized partial review can support online error control while allowing model-generated labels to replace expert annotation for part of the data stream.

1.2 Related work

Statistical guarantees for AI annotation. Our work adds to the recent literature on how model predictions can replace a portion of expert annotations while retaining explicit guarantees on label quality. PAC Labels combines AI predictions with selectively acquired expert labels within a fixed target collection and controls the overall labeling error of the completed dataset with high probability (Candès et al., 2025). Conformal Labeling instead uses a labeled calibration sample to select AI-labeled cases from an i.i.d. test sample, controlling the false discovery rate among the selected labels (Huang et al., 2025). These methods address settings where the labeled data for calibrating the annotation rule are representative of the unlabeled cases. CARED considers a different regime with potentially arbitrary workflow shift where expert labels are obtained through randomized audits. Its target is long-run error control over the full and non-audited streams, which in turn bounds the fraction of non-audited cases receiving erroneous automatic annotations.

Human-in-the-loop annotation and partial-review workflows. Our workflow is closely related to human-in-the-loop annotation and partial-review diagnostic pipelines, where model outputs reduce expert effort. In computational pathology, prior work has studied semi-automated annotation interfaces and practical protocols for reducing expert-labeling time and labor (Cazzaniga et al., 2025; Montezuma et al., 2023, 2025) especially pathologist-supervised workflows (Bodén et al., 2021; Rau et al., 2024; Gaffney, 2025). These works motivate the partial-review setting but focus mainly on workflow design and implementation, whereas our objective is to provide statistical guarantees through partial review. A complementary literature on learning to defer jointly learns when a model should predict and when a case should be deferred to an expert (Mozannar and Sontag, 2020; Charusaie et al., 2022). These methods optimize the prediction and deferral loss, whereas CARED aims for distribution-free long-run validity for the non-deferred cases. As CARED constructs prediction sets for deferred cases, the philosophy is close to Jin et al. (2026).

Conformal methods for cell-type annotation. Cell-type annotation is an important task in multiple areas. In single-cell genomics, representative procedures include reference mapping, consensus label transfer, atlas alignment, transfer learning, and, more recently, LLM-assisted annotation (Lotfollahi et al., 2024; Ergen et al., 2024; Korsunsky et al., 2019; Lotfollahi et al., 2022; Hou and Ji, 2024). In parallel, image-based pathology has begun moving from generic tissue-level modeling toward direct cell-level annotation and classification from histopathology or multiplex tissue images (Li et al., 2025b; Xu et al., 2025). More recently, conformal methods have also been developed for cell-type annotation (López-De-Castro et al., 2025). In particular, Corbetta et al. (2026) highlighted the distribution shift issue and modeled distribution shift as label shift (assuming the distribution of data within each cell type is invariant) or covariate shift (assuming the label distribution given features is invariant). However, these models do not resolve the adaptive, unpredictable changes prevalent in cell-type annotation settings (Figure 1). In addition, we focus on annotation settings whereas these methods focus on the construction of prediction sets.

Adaptive conformal inference. Our random review mechanism extends the adaptive conformal inference (ACI) framework (Gibbs and Candès, 2021) whose representative extensions concern general update rules (Angelopoulos et al., 2024; Gibbs and Candès, 2024) and time-series settings (Zaffran et al., 2022; Xu and Xie, 2023). Recent work starts to relax the full-feedback requirement, such as intermittently observed labels in imitation learning (Zhao et al., 2025) and semi-bandit feedback in retrieval (Ge et al., 2024). In contrast, we present a random review mechanism to overcome the expert review bottleneck and yield error control guarantees on the entire stream and the unreviewed cases.

Label-conditional and utility-aware prediction sets. Apart from validity, a growing literature studies the utility of prediction sets. This includes studies of training or optimization schemes that seek smaller or better-shaped conformal outputs (Stutz et al., 2021; Kiyani et al., 2024). A related direction incorporates structure in the label space so that the candidate labels remain coherent with domain knowledge (Goren et al., 2024; Mortier et al., 2025; Jin et al., 2026). Our utility-aware module is in spirit close to this literature: it incorporates an optimization module to improve the skipped ratio, reduce prediction set size, and encourage outputs that respect specified relations among cell types.

2 Background and setup

2.1 Formal setup

We now describe the formal setup of the AI annotation problem. For generality, we describe the setup for a generic annotation task, while referring to cell-type annotation at times as examples. We consider a streaming workflow where the cases arrive sequentially in batches. At time step t , the AI model receives an unlabeled batch $D_t = \{x_{t,1}, \dots, x_{t,B_t}\}$ with batch size $B_t := |D_t| \geq 1$.

The true labels $\{y_{t,i}\}_{i=1}^{B_t}$ belong to a finite label space $\mathcal{Y} = [K] = \{1, \dots, K\}$, where K is the total number of labels (cell types). The classification model used at time t produces for each case $x \in D_t$ a score vector $s_t(x) = (s_{1,t}(x), \dots, s_{K,t}(x)) \in [0, 1]^K$, where $s_{k,t}(x)$ is a model-produced score for class k , with larger values indicating stronger support for assigning x to that class. For each case $x_{t,i}$, based on the score $s_t(x_{t,i})$ and the information available at each time t , the task is to decide whether it can bypass the expert review that provides $y_{t,i}$ and, if so, to provide an automated annotation $\hat{y}_{t,i}$. This decision is denoted as $I_{t,i} \in \{0, 1\}$, so $I_{t,i} = 1$ means committing to the AI annotation.

We define the cumulative *annotation error* up to time t as

$$\text{AnnoErr}_t := \frac{\sum_{s=1}^t \sum_{i \in [B_s]: I_{s,i}=1} \mathbb{1}\{\hat{y}_{s,i} \neq y_{s,i}\}}{\sum_{s=1}^t B_s}.$$

Note that this quantity is fundamentally unobserved, since $y_{t,i}$ is not observed for the cases with $I_{t,i} = 1$. Such an error rate is analogous to the notion considered in Candès et al. (2025).

In addition, we assume that expert feedback is available throughout the process, so the procedure can actively request expert feedback, denoted by an indicator $E_{t,i} \in \{0, 1\}$ where $E_{t,i} = 1$ means the label $y_{t,i}$ is collected and observed before the next batch D_{t+1} . Together, cases in D_t fall into three categories:

- $E_{t,i} = 1$, whose true label $y_{t,i}$ is solicited from the expert and observed.
- $E_{t,i} = 0$ and $I_{t,i} = 1$, whose automated annotation $\hat{y}_{t,i}$ is considered as trustworthy so $y_{t,i}$ is never collected from the expert.
- $E_{t,i} = I_{t,i} = 0$, whose annotation cannot be automatically resolved by CARED but whose true label $y_{t,i}$ need not be observed before time $t + 1$; instead, it can be obtained later on using $\hat{C}_t(x_{t,i})$ as a candidate list, without interrupting our procedure.

We denote the audited set by $A_t := \{x_{t,i} : E_{t,i} = 1\}$. At each time point t , given the data stream $\{D_s\}_{s \leq t}$ and previously solicited feedback $\{y_{s,i} : E_{t,i} = 1\}_{s \leq t-1}$, we aim to produce the decisions $I_{t,i}$ and $\hat{y}_{t,i}$ while controlling the unobserved error rate AnnoErr_t below a user-specified level $\alpha \in (0, 1)$ in the long run.

As previewed, we accompany each case with a prediction set $\hat{C}_t(x) \subseteq [K]$. This set is said to cover the label if $y_{t,i} \in \hat{C}_t(x_{t,i})$. Following the literature, we define the *cumulative coverage error* up to time t as

$$\text{CovErr}_t := \frac{\sum_{s=1}^t \sum_{i=1}^{B_s} \mathbb{1}\{y_{s,i} \notin \hat{C}_s(x_{s,i})\}}{\sum_{s=1}^t B_s}.$$

As not all the $y_{t,i}$'s are observed, this coverage error is also unobserved for each time point t . Because every erroneous automated annotation necessarily corresponds to a prediction set that fails to contain the true label—either an empty set or a singleton set with the wrong label, the events counted by the numerator of AnnoErr_t form a subset of those for CovErr_t , and therefore we have $\text{AnnoErr}_t \leq \text{CovErr}_t$.

Additional notations for class-conditional errors. In cell-type annotation, it is important to maintain error rates within each *true* label class, because some classes can be rare but clinically important. We now introduce analogous notation for class-conditional errors. For each class $k \in [K]$, let $D_{k,t} \subseteq D_t$ denote the subset of batch- t cells whose ground-truth class is k . Then $\{D_{k,t}\}_{k=1}^K$ forms a partition of D_t .

Correspondingly, let $A_{k,t} := A_t \cap D_{k,t}$ be the reviewed cells from class k , and define the unreviewed subsets by $U_t := D_t \setminus A_t$, and $U_{k,t} := D_{k,t} \setminus A_{k,t}$. For a class- k cell $x \in D_{k,t}$ and a corresponding prediction set $\hat{C}_t(x)$, define the coverage error $\text{err}_{k,t}(x) := \mathbb{1}\{k \notin \hat{C}_t(x)\}$, which records whether the prediction set fails to cover the true label. For any horizon T , we write the group-wise annotation (within unreviewed cases) and coverage error as

$$\text{AnnoErr}_{k,T}^U = \frac{\sum_{t=1}^T \sum_{x_{t,i} \in U_{k,t} : I_{t,i}=1} \mathbb{1}\{\hat{y}_{t,i} \neq k\}}{\sum_{t=1}^T |U_{k,t}|}, \quad \text{CovErr}_{k,T}^D = \frac{\sum_{t=1}^T \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |D_{k,t}|}.$$

In addition, we define the cumulative coverage error rates over reviewed and unreviewed cells as

$$\text{CovErr}_{k,T}^A = \frac{\sum_{t=1}^T \sum_{x \in A_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |A_{k,t}|}, \quad \text{CovErr}_{k,T}^U = \frac{\sum_{t=1}^T \sum_{x \in U_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |U_{k,t}|}.$$

2.2 Preliminaries on conformal prediction

We next briefly review conformal prediction algorithms which will be useful in motivating our procedure. These algorithms may operate in different settings than ours, and we introduce them in their original setup.

Split conformal prediction (SCP) is a canonical variant of conformal prediction (Vovk et al., 2005; Lei et al., 2018). Assuming access to labeled data $\{(X_i, Y_i)\}_{i=1}^n$, the goal is to construct a prediction set $\hat{C}(X_{n+1}) \subseteq \mathcal{Y}$ for the unknown label Y_{n+1} of a newly observed test point $X_{n+1} \in \mathcal{X}$. The SCP method leverages any conformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, which is a function whose training process is independent of $\{(X_i, Y_i)\}_{i=1}^n \cup \{X_{n+1}\}$. Oftentimes the score s builds upon a pre-trained prediction model, and intuitively,

the score $s(x, y)$ measures how well the value y “conforms” to the predicted value based on x . The SCP prediction set takes the form

$$\hat{C}_\alpha^{\text{SCP}}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \geq \hat{q}_\alpha\},$$

where $\hat{q}_\alpha$ is the empirical α -quantile of $\{S_i\}_{i=1}^n \cup \{-\infty\}$, and $S_i = s(X_i, Y_i)$ for $i = 1, \dots, n$. Under the condition that $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, without any other assumptions on the data distribution or the score s , the coverage guarantee of SCP holds in finite sample: $\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha^{\text{SCP}}(X_{n+1})) \geq 1 - \alpha$.

The Mondrian conformal prediction (MCP) method extends SCP to achieve class-conditional coverage (Vovk et al., 2003), and we state this method here in accordance with SCP (assuming a pre-trained score function). Assuming access to the same data as above, and writing $\mathcal{Y} = \{1, \dots, K\}$, we denote $\mathcal{I}_k = \{i \leq n : Y_i = k\}$ as the calibration indices of class k . Then, let $\hat{q}_{k,\alpha}$ denote the empirical α -quantile of $\{s(X_i, k)\}_{i \in \mathcal{I}_k} \cup \{-\infty\}$. The MCP set for the new test point X_{n+1} is given by

$$\hat{C}_\alpha^{\text{MCP}}(X_{n+1}) = \{k \in \mathcal{Y} : s(X_{n+1}, k) \geq \hat{q}_{k,\alpha}\}.$$

Assuming the *within-class* exchangeability of $\{(X_i, Y_i) : Y_i = k\}_{i=1}^{n+1}$, these prediction sets satisfy

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha^{\text{MCP}}(X_{n+1}) \mid Y_{n+1} = k) \geq 1 - \alpha, \quad \forall k \in [K].$$

The essential advantage is that exchangeability is no longer imposed on the full distribution of (X, Y) . In an online process, the label proportion $\mathbb{P}(Y_t = k)$ may drift arbitrarily, and MCP maintains in-class coverage as long as the within-class distributions remain exchangeable. Moreover, the class-conditional coverage guarantee it delivers is stronger than the marginal coverage guarantee and is particularly suitable for cell-type annotation in which certain classes can be consequential.

While SCP and MCP still rely on exchangeability structures within data, adaptive conformal inference (ACI) is an online procedure that maintains coverage even when the data distribution drifts arbitrarily (Gibbs and Candes, 2021). The key strategy is to adjust the sets adaptively based on newly observed labels. Consider an online setting where data points $(X_t)_{t \geq 1}$ arrive one at a time, and the true label Y_t is revealed before time $t + 1$. The algorithm maintains a running level α_t , and at step t issues the prediction set $\hat{C}_{\alpha_t}(X_t)$ as in the SCP construction above. Once the true label Y_t is revealed, it updates the level by

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t), \quad \text{err}_t := \mathbb{1}\{Y_t \notin \hat{C}_{\alpha_t}(X_t)\},$$

with a fixed step size $\gamma > 0$ and any initialization $\alpha_1 \in \mathbb{R}$. For any sequence $(X_t, Y_t)_{t \geq 1}$, whether stationary, drifting, or adversarially generated, this procedure delivers the long-run guarantee $\frac{1}{T} \sum_{t=1}^T \text{err}_t \rightarrow \alpha$ as $T \rightarrow \infty$. However, the true label Y_t must be revealed before the next prediction can be issued.

3 Method and theoretical guarantees

In the following, we introduce the CARED framework in Section 3.1, the optional enhancement in Section 3.2, and the error control guarantees in Section 3.3.

3.1 Conformal annotation with randomized expert review

At the beginning, we assume access to an initial labeled block A_0 , among which the cells belonging to class $k \in [K]$ are denoted as $A_{k,0}$. This block provides the initial labeled pool used to construct class-wise prediction sets. Moreover, we initialize the class-wise levels by $\alpha_{k,1} = \alpha$ for every $k \in [K]$. We shall maintain the labeled sets $\{A_{k,t}\}_{k \in [K], t \geq 1}$ and the levels $\{\alpha_{k,t}\}_{k \in [K], t \geq 1}$ over time.

The procedure proceeds over time $t = 1, 2, \dots, T$. At each time point t , for each class k , a calibration set $D_{k,\text{cal}}(t)$ is constructed to be the class- k cells in the most recently reviewed $\min\{t, T_w\}$ batches, where $T_w \geq 1$ is a pre-specified parameter. That is, we define $D_{k,\text{cal}}(t) = \prod_{\tau=t-T_w}^{t-1} A_{k,\tau}$. This choice is computationally convenient and allows the calibration set to track local distributional changes. However, for theoretical

validity, any subset of all already-reviewed class- k cells, i.e., any subset of $\bigcup_{\tau=0}^{t-1} A_{k,\tau}$, are allowed. Also, a score $\tilde{s}_{k,t}(x) \in [0, 1]$ is computed based on a prediction model (in this paper, the NuClass (Xu et al., 2025) model), which might be updated over time, e.g., using the utility module described in Section 3.2. For any $u \in [0, 1]$, we denote the u -th empirical quantile of the calibration set as

$$\hat{q}_{k,t}(u) := \inf \left\{ q \in [0, 1] : \frac{1}{|D_{k,\text{cal}}(t)|} \sum_{x \in D_{k,\text{cal}}(t)} \mathbf{1}\{\tilde{s}_{k,t}(x) \leq q\} \geq u \right\}. \quad (1)$$

For completeness, we define $\hat{q}_{k,t}(u) = -\infty$ for $u < 0$ and $\hat{q}_{k,t}(u) = +\infty$ for $u > 1$. When $|D_{k,\text{cal}}(t)| = 0$, we set $\hat{q}_{k,t}(u) = -\infty$ for $u \leq 1$ and $\hat{q}_{k,t}(u) = +\infty$ for $u > 1$. Given the levels $\{\alpha_{k,t}\}_{k=1}^K$, we construct

$$\hat{C}_t(x) = \{k \in [K] : \tilde{s}_{k,t}(x) \geq \hat{q}_{k,t}(\alpha_{k,t})\}, \quad x \in D_t. \quad (2)$$

Our updating rule below following ACI may lead to $\alpha_{k,t} \notin [0, 1]$, yet the empirical quantiles are well-defined. To accommodate cases where $\hat{C}_t(x) = \emptyset$, we define the enlarged set

$$\hat{C}_t^+(x) = \begin{cases} \hat{C}_t(x), & \hat{C}_t(x) \neq \emptyset, \\ \{\arg \max_{k \in [K]} \tilde{s}_{k,t}(x)\}, & \hat{C}_t(x) = \emptyset, \end{cases} \quad (3)$$

where ties in the second case are broken randomly. Thus, $\hat{C}_t^+(x)$ is a singleton whenever $|\hat{C}_t(x)| \leq 1$. Since $\hat{C}_t(x) \subseteq \hat{C}_t^+(x)$, we know $\mathbf{1}\{k \notin \hat{C}_t^+(x)\} \leq \mathbf{1}\{k \notin \hat{C}_t(x)\}$. Our theory in Section 3.3 establishes the long-run coverage of $\hat{C}_t(x)$, which translates to the long-run coverage of $\hat{C}_t^+(x)$ and the annotation error bound.

The levels $\{\alpha_{k,t}\}$ are updated through an *unbiased randomized rounding scheme*. Fix a target reviewed fraction $p \in (0, 1)$. At time t , CARED chooses a realizable fraction p_t so that $p_t B_t$ is an integer and the cumulative reviewed fraction $(\sum_{t=1}^T p_t B_t) / (\sum_{t=1}^T B_t) \rightarrow p$ as $T \rightarrow \infty$; see Appendix B.1 for details. Then we sample uniformly without replacement from D_t the audit set of size $|A_t| = p_t B_t$. The annotation decision is

$$I_{t,i} := (1 - E_{t,i}) \mathbf{1}\{|\hat{C}_t(x_{t,i})| \leq 1\}.$$

Whenever $I_{t,i} = 1$, we define the automated annotation $\hat{y}_{t,i}$ as the unique element of the prediction set, $\hat{C}_t^+(x_{t,i}) = \{\hat{y}_{t,i}\}$. No automated annotation $\hat{y}_{t,i}$ is produced when $I_{t,i} = 0$. The audited labels are then used to update the online levels for subsequent steps similar to ACI:

$$\alpha_{k,t+1} = \alpha_{k,t} + \gamma_k \sum_{x \in A_{k,t}} (\alpha - \text{err}_{k,t}(x)), \quad k \in [K], \quad (4)$$

where γ_k is a pre-specified step size for class k , and $\text{err}_{k,t}(x) = \mathbf{1}\{k \notin \hat{C}_t(x)\}$ is the miscoverage indicator. In our experiments, we fix the class-wise step sizes according to cohort-specific class frequencies; the specific choices are given in Appendix C.1. Appendix D further discusses the performance of fixed step sizes compared with time-varying ones. The updated online levels are then used to compute the next batch's class-wise quantile cutoffs through (2). We summarize the procedure of CARED in Algorithm 1.

3.2 Optional model training with utility improvement in cell-type annotation

The procedure introduced in Section 3.1 works for any score functions $\tilde{s}_{k,t}(\cdot)$, but the choice of the score function determines the form of the prediction sets. In this part, we introduce an optional module that adapts the score functions over time. Specifically, we parameterize the transformed score as $\tilde{s}_t(x) = f_{\Theta_t}(s_t(x))$, where Θ_t is updated over time and $s_t(\cdot)$ is the original score given by the NuClass model. The construction of the score transformation f_{Θ_t} is provided in Algorithm A.2 in Appendix A.3.

Utility metrics and training objective. Recall that the t -th batch is $D_t = \{x_{t,1}, \dots, x_{t,B_t}\}$, where $B_t = |D_t|$. For any parameter Θ , let $\hat{C}_{t,i}(\Theta)$ denote the prediction set $\hat{C}_t(x_{t,i})$ obtained using the transformed scores $f_{\Theta}\{s_t(x_{t,i})\}$. The construction follows (1)–(2). Before updating Θ_t and deciding $\tilde{s}_t(\cdot)$, we first define

Algorithm 1 CARED

Input: Initial labeled block A_0 with labels in $[K]$; incoming batches $\{D_t\}_{t=1}^T$; target error level $\alpha \in (0, 1)$; target reviewed fraction $p \in (0, 1)$; calibration window $T_w \geq 1$; fixed step sizes $\gamma_k > 0$, $k \in [K]$.

- 1: Set $A_{k,0} \leftarrow \{x \in A_0 : y(x) = k\}$ and $\alpha_{k,1} \leftarrow \alpha$ for every $k \in [K]$.
- 2: **for** $t = 1, \dots, T$ **do**
- 3: Receive $D_t = \{x_{t,i}\}_{i=1}^{B_t}$ and score functions $\{\tilde{s}_{k,t} : \mathcal{X} \rightarrow [0, 1]\}_{k=1}^K$.
- 4: **for** $k = 1, \dots, K$ **do**
- 5: Set $D_{k,\text{cal}}(t) \leftarrow \bigcup_{\tau=\max\{0, t-T_w\}}^{t-1} A_{k,\tau}$.
- 6: Compute $\hat{q}_{k,t}(\alpha_{k,t})$ according to (1).
- 7: **end for**
- 8: Compute $\hat{C}_t(x_{t,i})$ and $\hat{C}_t^+(x_{t,i})$ according to (2) and (3) for every $i \in [B_t]$.
- 9: Draw $\xi_t \sim \text{Bernoulli}(pB_t - \lfloor pB_t \rfloor)$ and set $p_t \leftarrow (\lfloor pB_t \rfloor + \xi_t)/B_t$. ► See Appendix B.1
- 10: Sample $A_t \subseteq D_t$ uniformly without replacement with $|A_t| = p_t B_t$.
- 11: For every $i \in [B_t]$, set $E_{t,i} \leftarrow \mathbb{1}\{x_{t,i} \in A_t\}$ and $I_{t,i} \leftarrow (1 - E_{t,i}) \mathbb{1}\{|\hat{C}_t(x_{t,i})| \leq 1\}$.
- 12: Compute the unique $\hat{y}_{t,i} \in \hat{C}_t^+(x_{t,i})$ for each $i \in [B_t]$ with $I_{t,i} = 1$.
- 13: Obtain expert labels $y_{t,i}$ for all $E_{t,i} = 1$, and set $A_{k,t} \leftarrow \{x_{t,i} \in A_t : y_{t,i} = k\}$ for every $k \in [K]$.
- 14: For every $k \in [K]$, update $\alpha_{k,t+1} \leftarrow \alpha_{k,t} + \gamma_k \sum_{x \in A_{k,t}} (\alpha - \mathbb{1}\{k \notin \hat{C}_t(x)\})$.
- 15: **end for**

Output: Prediction sets $\hat{C}_t(x_{t,i})$, annotation decisions $I_{t,i}$ and annotation labels $\hat{y}_{t,i}$ for $I_{t,i} = 1$.

what properties are good for prediction sets. Below we quantify the utility of prediction sets for a batch D_t based on three criteria: (1) skipped ratio surrogate, (2) set size, and (3) structural coherence.

First, we encourage the score function to lead to a large fraction of automated annotations. This fraction (among cases not selected for expert audit) can be estimated using the entire batch: $\mathcal{J}_{\text{skip}}(\Theta) = \frac{1}{B_t} \sum_{i=1}^{B_t} \mathbb{1}\{|\hat{C}_{t,i}(\Theta)| \leq 1\}$. In addition, since non-singleton sets will be used as references for experts to resolve the ambiguity, it is sensible to encourage small set sizes. Smaller sets leave fewer candidate labels for the subsequent narrowing-down step and are therefore more useful in human-in-the-loop decision making (Zhang et al., 2024; Dhillon et al., 2024). We thus consider the second metric $\mathcal{J}_{\text{size}}(\Theta) = \frac{1}{B_t} \sum_{i=1}^{B_t} |\hat{C}_{t,i}(\Theta)|$. Finally, a non-singleton prediction set is more useful when the candidate labels respect the biological structures among the cell types (Kaymaz et al., 2021; Theunissen et al., 2024; Corbetta et al., 2026). We design a metric $\mathcal{J}_{\text{struct}}(\Theta)$ to reward structurally coherent prediction sets based on the subtype tree-structures in cell types; the detailed construction is given in Appendix A.1. The three components are combined into

$$\mathcal{J}(\Theta) = \lambda_{\text{size}} \mathcal{J}_{\text{size}}(\Theta) - \lambda_{\text{struct}} \mathcal{J}_{\text{struct}}(\Theta) - \lambda_{\text{skip}} \mathcal{J}_{\text{skip}}(\Theta), \quad (5)$$

where the weights are nonnegative; see Appendix C.1 for the choice in our applications. Thus, minimizing (5) penalizes large sets while rewarding structural coherence and a larger skipped-ratio.

Online score training. To train the score function, we first build a smooth surrogate for the objective $\mathcal{J}(\Theta)$. Specifically, since several terms involve indicators that are difficult for gradient descent updates, we replace these quantities via standard differentiable approximations. At each time t , we initialize $\Theta = 0$ and perform N gradient steps on the surrogate objective evaluated on the current unlabeled batch D_t . Before each update, we compute the class-wise conformal thresholds from the labeled calibration sets $\{D_{k,\text{cal}}(t)\}_{k=1}^K$ using the current transformation f_Θ . After each update, we recompute these thresholds and evaluate the original empirical objective $\mathcal{J}(\Theta)$ on D_t . Among the initialization and all gradient iterates, we select the parameter value with the lowest objective value as Θ_t . Notably, CARED allows Θ_t to depend on the current batch D_t , provided that Θ_t is constructed before the audit set A_t is sampled. The online training procedure is summarized in Algorithm A.1 and detailed in Appendix A.2.

3.3 Theoretical guarantees

In this part, we present the main theoretical guarantee, which establishes long-run class-wise validity in CARED based on the random audit alone, regardless of the distribution drift and the score models. At a high level, the batch-wise ACI-style recursion controls the error within the reviewed cells, and the uniform random audit transfers this control to the full stream, including the cells that are not reviewed. These theoretical results help contextualize the quality guarantee for the automated annotations from CARED.

Recall the notations from Section 2.1: $\text{CovErr}_{k,T}^A$, $\text{CovErr}_{k,T}^D$, and $\text{CovErr}_{k,T}^U$ denote the cumulative conformal errors over the audited cells, all cells, and unaudited cells that belong to class k , respectively, up to time T . The true class membership is unknown unless the case is reviewed by experts (i.e., $E_{t,i} \in \{0, 1\}$), and is used only for stating the guarantees. Let $\{\mathcal{G}_t\}_{t \geq 0}$ be the pre-review filtration, where $\mathcal{G}_t$ is the sigma-field generated by the information after the prediction sets for D_t are formed but before the audit set A_t is sampled. In particular, the class counts $|D_{k,t}|$ and the coverage indicators $\text{err}_{k,t}(x)$ for any $x \in D_{k,t}$ are $\mathcal{G}_t$ -measurable. Recall that $\{p_t\}_{t=1}^T$ are the fraction of cases sent for expert review in the batches.

We impose mild conditions on the data and the random audit, without imposing exchangeability or stationarity of the incoming batches or any parametric model on the data distribution.

Assumption 3.1. *Fixing a target reviewed fraction $p \in (0, 1)$, we assume that:*

- (i) *The batch size $B_t = |D_t|$ is uniformly controlled, obeying $B^- \leq B_t \leq B^+$ for some finite, positive integers $0 < B^- \leq B^+ < \infty$.*
- (ii) *For every $k \in [K]$, the cumulative number of class- k cells diverges, i.e., $\sum_{t=1}^\infty |D_{k,t}| = \infty$ as $T \rightarrow \infty$.*
- (iii) *Conditional on $\mathcal{G}_t$, the audit set $A_t \subseteq D_t$ is drawn uniformly without replacement with $|A_t| = p_t B_t$.*
- (iv) *The score function $\tilde{s}_t(\cdot)$ is adapted to $\mathcal{G}_t$.*

Assumption 3.1(i) and (iii) can be ensured through appropriate batch construction and audit sampling. Moreover, (ii) is a mild condition that every cell type continues to occur and can be updated throughout the workflow so the within-class asymptotics apply. Finally, (iv) requires the score function to be determined before drawing the random audit, but can depend on the batch D_t .

We begin with the guarantee on the expert-reviewed cells. The proof of Proposition 3.2 is in Appendix E.1.

Proposition 3.2 (Long-run coverage error control for reviewed cells). *Fix any class $k \in [K]$, suppose the accumulated class- k frequency is bounded away from zero, i.e., $\liminf_{T \rightarrow \infty} \frac{\sum_{t=1}^T |D_{k,t}|}{\sum_{t=1}^T B_t} \geq q_k$ for some constant $q_k > 0$. Under assumption 3.1, almost surely there exists some $T_0 < \infty$ such that for all $T \geq T_0$,*

$$|\text{CovErr}_{k,T}^A - \alpha| \leq \frac{\max\{\alpha, 1 - \alpha\}}{\gamma_k \sum_{t=1}^T |A_{k,t}|} + \frac{C}{T},$$

where $C = \frac{2B^+}{B^- - q_k} (1 + \frac{1}{pB^-})$. Consequently, $\text{CovErr}_{k,T}^A \rightarrow \alpha$ almost surely as $T \rightarrow \infty$.

Intuitively, the coverage guarantee on the reviewed cases inherit the guarantees from ACI. Our main result is that such guarantee translates to that on the entire stream, due to the unbiased random design. The proof of Theorem 3.3 is in Appendix E.2.

Theorem 3.3 (Long-run coverage error control for all cells). *Under Assumption 3.1, for every class k , as $T \rightarrow \infty$, the cumulative miscoverage over all cells and the cells not reviewed by the experts satisfies*

$$|\text{CovErr}_{k,T}^D - \alpha| \rightarrow 0, \quad |\text{CovErr}_{k,T}^U - \alpha| \rightarrow 0 \quad \text{a.s.}$$

These within-class coverage guarantees, together with the growing conditions on the group sizes, imply the overall coverage guarantee, i.e., $\text{CovErr}_T \rightarrow \alpha$ almost surely, and the coverage guarantee for all cases not reviewed by the experts by the lower-bounded audit probability, whose proofs are omitted for brevity. In a nutshell, it means that even though CARED only requests expert labels for a fraction of $p \in (0, 1)$ of

cases (in our experiments, the fraction can be as low as 10% or even lower), the prediction sets for all the incoming cases are still asymptotically well-calibrated and adapted to the arbitrarily drifting workflow.

More importantly for deployment, because $\widehat{C}_t(x_{t,i}) \subseteq \widehat{C}_t^+(x_{t,i})$, the annotation error is upper bounded by miscoverage. We present this result in Corollary 3.4.

Corollary 3.4 (Long-run automated-annotation error control). *Recall the overall cumulative annotation error AnnoErr_T and class-wise annotation error $\text{AnnoErr}_{k,T}$ defined in Section 2.1. Under Assumption 3.1,*

$$\limsup_{T \rightarrow \infty} \text{AnnoErr}_{k,T}^U \leq \alpha \quad \text{a.s.}, \quad k \in [K].$$

Like Theorem 3.3, these guarantees are distribution-free, requiring no exchangeability or stationarity conditions on the data stream. They also allow the AI models and scores to evolve arbitrarily over time. In addition, the per-class calibration ensures that the annotation error is rigorously controlled within every class, even though the true class memberships are unknown. This avoids neglecting clinically consequential but difficult-to-classify cell types. In practice, if the AI prediction quality is poor for one particular class, then CARED will accept fewer cases for automated annotation to maintain the error control.

For completeness, we establish an explicit convergence rate for the cumulative miscoverage error. Additional theory characterizing cumulative batch-wise deviations from the target level is in Appendix B.2.

Proposition 3.5 (Long-run error convergence rate). *Under the CARED procedure, suppose that the class-frequency condition in Proposition 3.2 holds for every class. Then, for every class k ,*

$$|\text{CovErr}_{k,T}^D - \alpha| \vee |\text{CovErr}_{k,T}^U - \alpha| = \mathcal{O}_p\left(\left(\sum_{t=1}^T |D_{k,t}|\right)^{-1/2}\right) = \mathcal{O}_p(T^{-1/2}).$$

4 Case study in cell-type annotation workflows

4.1 Data and annotation workflow

This case study evaluates CARED in the motivating cell-type annotation task where pathologists annotate cell types on whole-slide images (WSIs) in TissueLab (Li et al., 2025a). We describe the AI model, the tissue cohorts and their labels, and the construction of annotation streams.

We use the NuClass (Xu et al., 2025) model as the AI scores; it is a multi-scale histopathology model that detects nuclei and outputs a class-wise score vector for each detected cell. It provides three score sources, each used as the raw score $s_t(x)$: the *global* path, which incorporates the surrounding tissue context; the *local* path, which relies on nucleus-centered morphology; and the *gated* output, which adaptively fuses the two at the cell level. We use the three sources to show how CARED behaves under scores of different quality and sensitivity to tissue context. All cohorts below are fully held out from NuClass training and tuning.

The main experiments use three held-out cohorts from pancreas, lung, and ovary tissue, with label spaces of six, four, and three cell types, respectively (Table 1). The cohorts consist of one H&E-stained WSI per tissue, taken from three publicly available 10x Genomics Xenium datasets: FFPE pancreatic adenocarcinoma, non-small cell lung cancer, and ovarian serous carcinoma (10x Genomics, 2024c,a,b); see Appendix C for details on the H&E processing of the slide. We exclude cells on background or too close to the slide edge. Each cell was assigned a reference label from its spatial gene-expression profile: the 10x Genomics cell type with the highest normalized marker-gene score, which is viewed as the ground-truth cell type. Because these labels come from molecular measurements, they provide a reference independent of the H&E morphology. For audited cells, these reference labels are revealed to CARED to emulate expert review. For unaudited ones, they are used to compute the error rates only, and are never available to the procedure.

The label spaces reflect the cell populations. These cell types are critical because pathologists routinely distinguish these tissues based on cell’s spatial arrangements. Furthermore, they define tissue function and the tumor microenvironment: endocrine cells form the islets of Langerhans in the pancreas, while T cells,

Table 1: Data used in the case study. Class proportions are computed over all cells in each cohort.

Tissue cohort	# Cells	# Batches	Cell type (fractions %)
Pancreas	103,166	206	Acinar (12.8), Ductal (45.2), Endocrine (2.0), Endothelial (8.6), Fibroblast-like (27.6), T cell (3.8)
Lung	61,490	123	Endothelial (27.6), Epithelial malignant (61.8), T cell (5.9), Tumor-associated fibroblast (4.7)
Ovary	150,748	301	Endothelial (5.0), Epithelial malignant (85.3), Macrophage (9.7)

macrophages, and tumor-associated fibroblasts shape tumor progression and response to immunotherapy in lung and ovarian cancer. In this dataset, several clinically relevant cell types are especially rare: in the pancreas cohort, endocrine cells and T cells account for only 1.95% and 3.77% of all cells. Due to the class imbalance, a marginal error guarantee could be met while misannotating most of these rare cells, so we primarily target the class-conditional error control throughout.

Annotation data streams. To emulate how cells reach an annotator, we partition each cohort into spatially contiguous regions containing between $B^- = 490$ and $B^+ = 510$ cells. Each region plays the role of a field of view and forms one batch. One region, chosen at random, serves as the initial labeled block A_0 . The order in which the remaining regions are released is determined by annotation practice rather than by the tissue, so we consider three release schemes, each corresponding to a different practice:

- *Adjacent*: the next region is chosen among the unreleased neighbors of the current one, mimicking a pathologist panning across neighboring fields (Figure 1(a)).
- *Random*: regions are released in random order, mimicking the sampling of tiles or regions of interest when building annotated datasets (Qiu et al., 2023; Hägele et al., 2020).
- *Error-guided*: the next region is selected by identifying the class whose cumulative error on audited cells deviates most from α and releasing the region with the most cells predicted as that class. This mimics tools that steer annotators toward regions where the model currently struggles. Because each release depends on past audit outcomes, this scheme serves as an adversarial test with strong dependence; our guarantees still hold as long as the audit set is uniformly sampled given the past (Assumption 3.1(iii)).

The three schemes use the same cohorts, scoring model, and procedure, and differ only in the release order which determines the distribution drifts. Under the adjacent scheme, spatial autocorrelation yields gradual drift with abrupt changes at tissue-compartment boundaries (Figure 1); under the error-guided scheme, the drift is driven by the procedure’s past outcomes. Both of them lead to strong non-exchangeability. In the random scheme, due to random ordering, the batches are exchangeable. This exchangeability, however, holds at the level of regions only, since cells within a region share morphology and composition. With a limited number of regions (Table 1), such exchangeability does not suffice to justify error control on the cell level in each realized stream as the guarantees in Section 3.3.

Expert review protocol and interface. Figure 4 presents a demo interface for the annotation protocol. When a subsequent batch is processed, the interface displays all cells in the batch to preserve the spatial and morphological context needed for expert judgment, but only the uniformly sampled audit set A_t is flagged for review. Each flagged cell is outlined in the color of its highest-scoring class. The annotator inspects the flagged cells and corrects the displayed label whenever it is wrong. The resulting labels are then used in CARED to update the class-wise levels $\{\alpha_{k,t}\}$ before the next batch is released. For unreviewed cells, CARED determines which of them to annotate automatically; otherwise, the cell is deferred, and its prediction set is shown as a candidate list. In the three held-out cohorts, the annotator’s response is emulated by revealing the reference label of each flagged cell. Additional implementation details are reported in Appendix C.

Real expert annotation stream. To evaluate CARED on an annotation order produced by actual practice rather than by a release scheme, we also use expert annotations collected in TissueLab on a colon

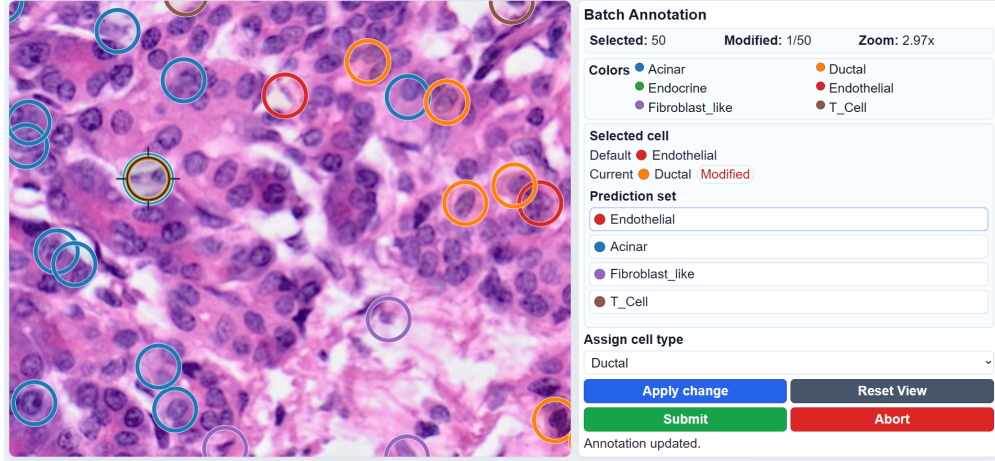


Figure 4: Interface for one pancreas batch in the partial-review workflow. The left panel shows a local histology field with reviewed cells highlighted by class-specific colored outlines. After clicking on a reviewed cell, the right panel displays its annotation information, including the default label, current label, and conformal prediction set. This example illustrates a reviewed cell whose label has been modified.

WSI containing 220,501 cells. One annotator (a research scientist trained in histopathology annotation under pathologist supervision) annotated 1,374 cells over 37 rounds, in the order they chose while navigating the slide. The annotated cells are grouped into three classes: tumor (20.5%), muscle (45.0%), and others (34.5%).¹ Both the stream order and the labels are determined by the expert behavior; results are reported in Section 4.4.

4.2 CARED controls cumulative annotation error

We present the cumulative coverage results using the global NuClass score; additional results for other score sources are reported in Appendix C. Unless otherwise stated, the experiments below use target error level $\alpha = 0.1$, target reviewed fraction $p = 0.1$, and a fixed random seed 0.

Figure 5 reports the whole-stream cumulative coverage error trajectories (which upper bounds the annotation error) under the three batch-release schemes. Within each true cell type, miscoverage over the entire stream approaches $\alpha = 0.1$ in most settings, even though true types are observed only for the 10% audited cells. The largest remaining gaps occur for rare types, such as tumor-associated fibroblasts in the adjacent lung stream, consistent with slower convergence for rare classes (Proposition 3.5). We also summarize the marginal errors per tissue, which also approaches the target level in each batch-release scheme.

Comparison with static calibration. A common alternative is to calibrate once on a few labeled regions, as when annotating selected ROIs or tiles (Qiu et al., 2023; Hägele et al., 2020). In this case, both covariate shift and label shift can be amplified as we show in the introduction. We emulate this setting on the pancreas cohort: in each of 200 repetitions, we randomly choose 10 batches as the calibration data in standard split conformal prediction and class-conditional conformal prediction in Section 2.2, which are then applied to all remaining cells. CARED starts from the same batches and runs under the error-guided scheme with $p = 0.1$, without the utility module. All of them target the error control of $\alpha = 0.1$. Across repetitions, the baselines’ final marginal errors spread widely around α , whereas CARED’s concentrate at α (Figure 6(a)). The gap widens within cell types (Figure 6(b)): class-wise calibration adjusts for class composition but not for morphological differences across regions, so its errors vary most for the rare endocrine cells. Because exchangeability holds only at the region level, such one-time calibration is valid on average over randomly sampled regions but not in a given run.

¹This category pools the original labels negative control, epithelial cells, lymphocytes, eosinophils, and plasma cells.

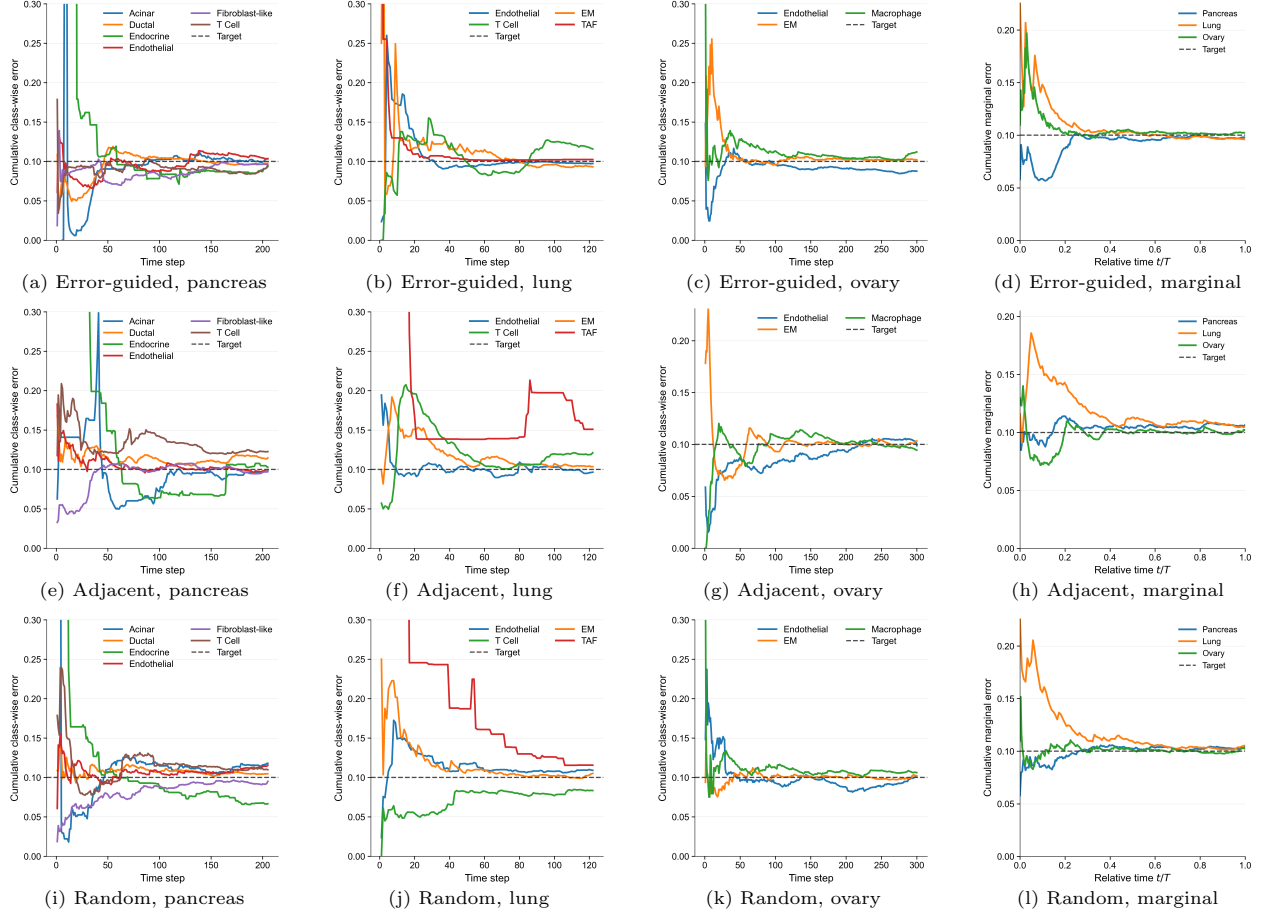


Figure 5: Cumulative error trajectories under the global NuClass score for three batch-release schemes. Rows correspond to the error-guided, adjacent, and random schemes, while columns correspond to pancreas, lung, ovary, and marginal results. The tissue-specific panels report class-wise cumulative error over time; the marginal panels report cumulative error by tissue against the relative time t/T . All experiments use target error level $\alpha = 0.1$ and reviewed fraction $p = 0.1$. Dashed horizontal lines mark the target error level.

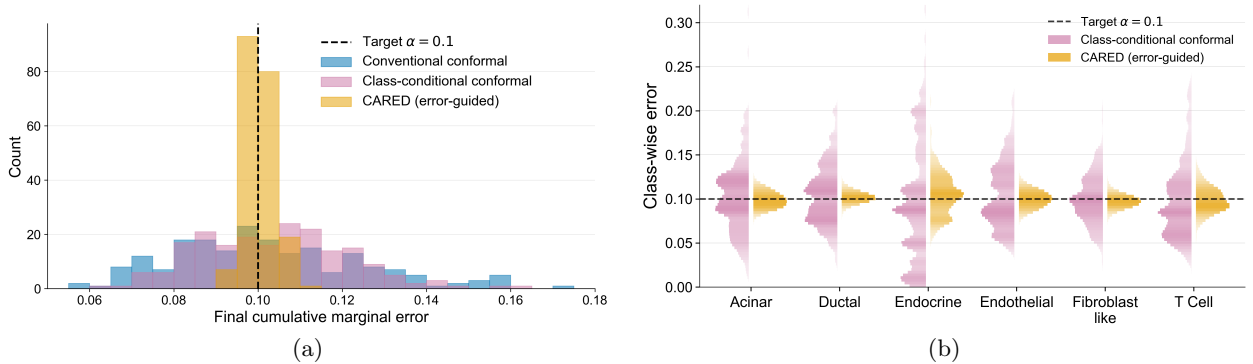


Figure 6: Comparison of final cumulative error on pancreas tissue using the global score source. Each experiment randomly selects 10 batches as calibration data for the baselines and as the initial batches for CARED. (a) Distribution of final cumulative marginal error over 200 repeated random experiments, where the dashed vertical line marks the target error level $\alpha = 0.1$. (b) Empirical distributions of final class-wise error over the same 200 experiments for class-conditional conformal prediction and our method, shown separately by cell type. The dashed horizontal line is the target level $\alpha = 0.1$.

4.3 Skipped ratio and prediction set utility

We measure the expert effort saved by the skipped ratio, the fraction of all cells that are not audited and receive an automated annotation, i.e., whose (restored) prediction set is a singleton (3). Figure 7 reports the skipped ratio at the final time T across three schemes and three tissue cohorts, both over all cells and stratified by the true cell types. The ratio computes the number of automatically annotated cells among all, including reviewed, cells. CARED allows a substantial fraction of cells to bypass expert labeling across the evaluated settings. In the four-class lung annotation task at $p = 0.1$ and $\alpha = 0.1$, CARED allows skipping as much as 70% of cells on average and over 80% of epithelial malignant cells. The skipped ratio also varies across cell types within a tissue: for instance, almost no endocrine cells in pancreas are annotated automatically; this is because the AI predictions are less reliable for this class so CARED defers more cells of this class.

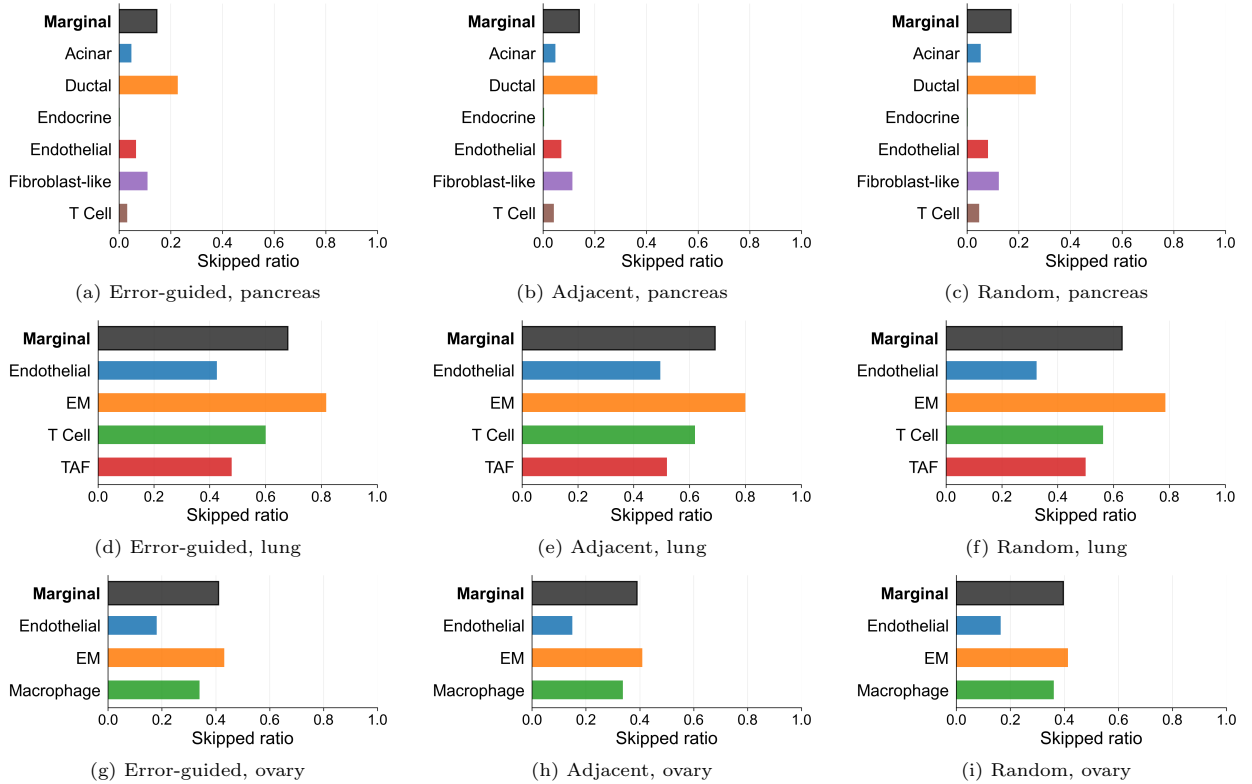


Figure 7: Final skipped ratios using the global NuClass score. Rows correspond to tissues, while columns correspond to batch-release schemes. Each panel reports the marginal skipped ratio (bold) together with ratios stratified by true label. Here, EM denotes Epithelial Malignant and TAF denotes Tumor Associated Fibroblast. All experiments use target level $\alpha = 0.1$ and reviewed fraction $p = 0.1$.

Cells that are not skipped receive a prediction set of candidate labels, whose value depends on how easily it can be resolved later. We evaluate the utility-aware score transformation introduced in Section 3.2 on the pancreas cohort where non-singleton candidate lists are common. Figure 8 compares the raw and CARED-transformed scores in terms of skipped ratio, average prediction set size, and structural coherence. Across all three release schemes, the utility-aware transformation consistently yields smaller candidate lists and improves structural coherence. It also increases the skipped ratio under error-guided release while yielding comparable skipped ratios under adjacent and random release. These results show that the utility module reduces human effort while producing more coherent candidate sets across different batch-release orders. Results for lung and ovary are reported in Figure 13 in Appendix C.2: the utility-aware transformation yields pronounced improvements for ovary, while its benefits are limited in lung, where the skipped ratio is already high and most candidate lists are already coherent. Sensitivity of the utility metrics to the utility

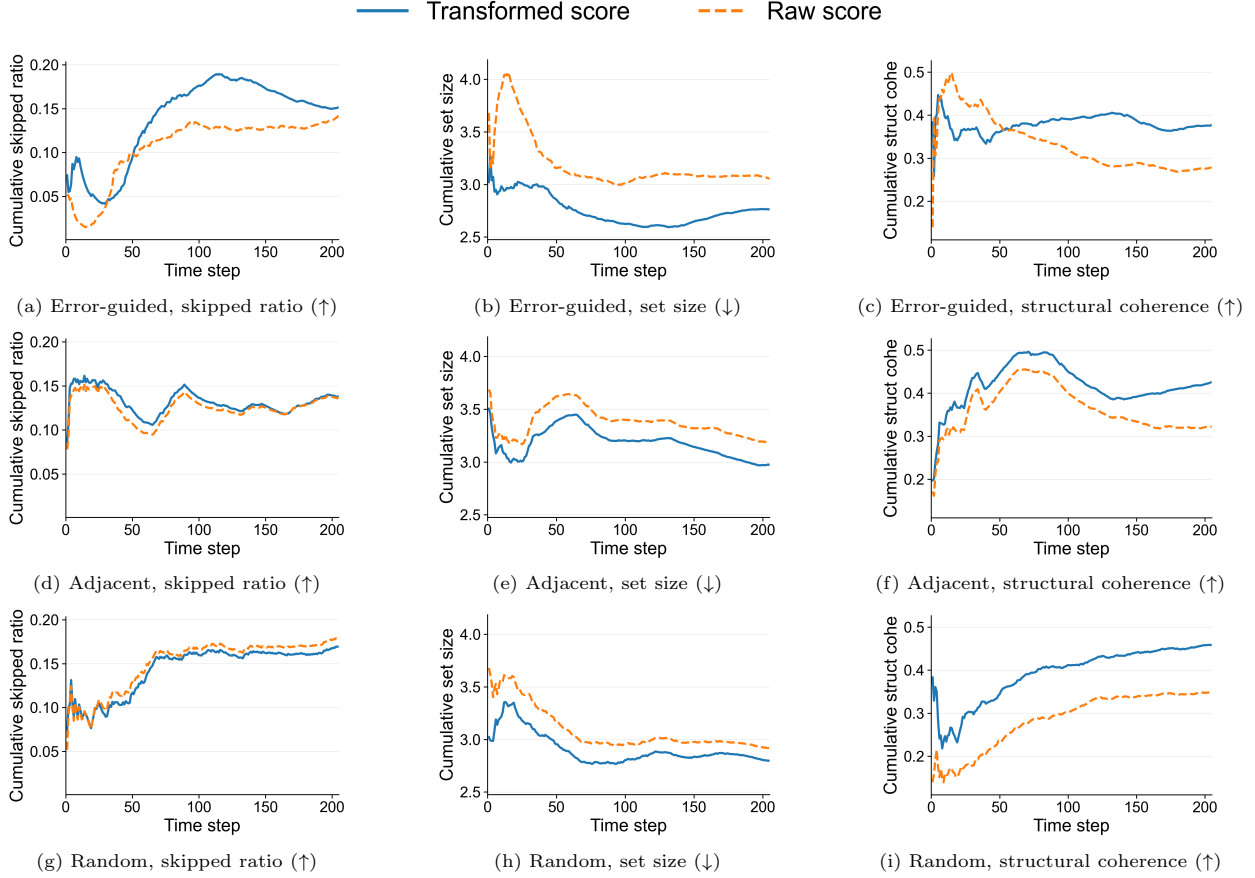


Figure 8: Utility of the raw and transformed scores over time on pancreas tissue. Rows correspond to the batch-release schemes, while columns correspond to the metrics. Higher skipped ratio and structural coherence and smaller prediction sets indicate better utility (indicated by the arrows in panel captions).

learning rate, random seed, and target error level is further discussed in Figure 15 in Appendix C.3.2.

4.4 Results with real expert annotations

We examine CARED on the colon data described in Section 4.1, where both the order of the 37 annotation rounds and the 1,374 labels come from an expert navigating the slide. The distribution drift is generated by real practice but the dataset is also small: with $p = 0.1$, CARED observes about 137 audited labels in total, about 4 per round. To avoid extremely scarce cell types which make error evaluation challenging, we group the cells are grouped into three classes: tumor, muscle, and others, with *others* combining several rare cell types. We run CARED with target error level $\alpha = 0.1$ using the global, local, and gated NuClass scores.

Figure 9(a) shows the cumulative annotation-error (solid) and miscoverage (dashed) trajectories compared with the expert annotations used in evaluation. Across the three NuClass scores, the cumulative annotation error and miscoverage remain close to the target level, showing the robust performance of CARED in real data drifts. Figure 9(b) reports the cumulative saved annotation cost, indicating that the proposed framework remains effective under distribution drifts created by real expert annotations.

4.5 Sensitivity analysis on cumulative error

We next assess the sensitivity of CARED to the choice of reviewed fraction p , random seed, and target error level α , using the pancreas cohort with the global NuClass score under the three batch-release schemes described in Section 4.1. Figure 10 summarizes the cumulative coverage error (which upper bounds the

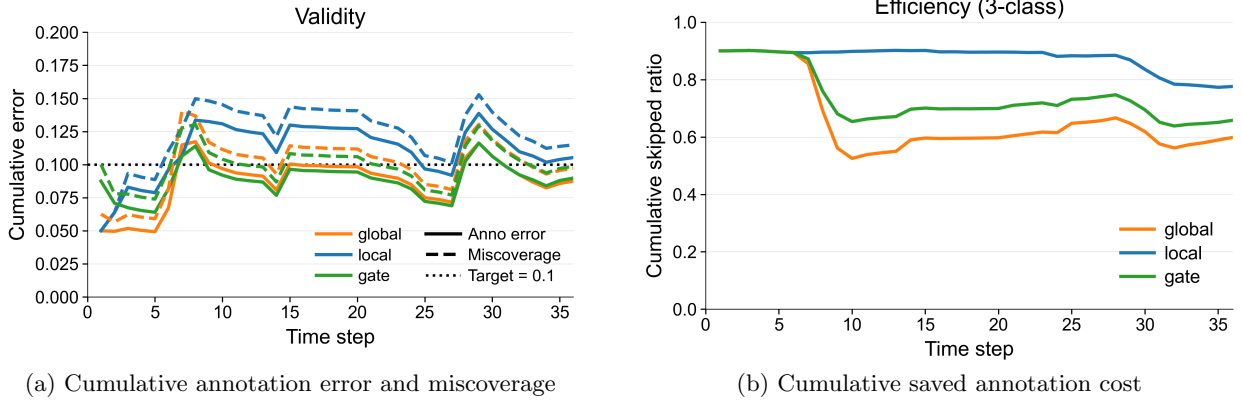


Figure 9: Results with real expert annotations on the colon dataset. (a) Cumulative annotation error (solid curves) and cumulative miscoverage (dashed curves) under the global, local, and gated NuClass score sources, with reviewed fraction $p = 0.1$ and target error level $\alpha = 0.1$; the black dotted horizontal line marks this target. (b) Cumulative skipped ratio under the same setting.

annotation error) trajectories as we vary the review fraction $p \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, random seeds, and the target level $\alpha \in \{0.01, 0.025, 0.05, 0.1, 0.2\}$.

Across all settings, the marginal coverage error moves towards the target level. Even with as few as a fraction of $p = 0.01$ of reviewed cases, the trajectories can still stabilize and approach the target level. Reducing the review rate does slow the convergence: the audit-sampling term in the proof of Proposition 3.5 scales as $\{(1-p)/(pN)\}^{1/2}$, where N is the total number of cells. The online-update term in Proposition 3.5 scales as $1/(\gamma_k p N_k)$, so a smaller p shrinks the effective step size and the levels adapt more slowly. The corresponding class-wise results are reported in Appendix C.3 (Figure 14). They show the same overall pattern, with larger finite-sample fluctuations only for rare cell types such as endocrine cells, consistent with the class-frequency dependence in Proposition 3.5; the marginal trajectories remain stable.

CARED converges as expected under multiple seeds, which determine the initial region and the audit samples. Across seeds (middle column in Figure 10), trajectories differ in the early stage and approach the target level from below or above depending on the starting region, but end close to α under every scheme.

Appendix D (Figure 17) further examines sensitivity to the scale of the fixed class-wise step sizes. On the pancreas task with the global NuClass score, scaling the default step sizes by factors from 0.1 to 10 consistently pulls the cumulative class-wise errors toward the target level $\alpha = 0.1$, demonstrating stable calibration over a broad range of step-size scales.

Discussion

We proposed CARED, which decides which AI-generated cell-type labels can be accepted without expert review while the annotation workflow drifts. Its validity is guaranteed by the random audit design alone, which yields long-run class-conditional control of miscoverage, and hence of annotation error, over both the full stream and the unreviewed cells, without exchangeability or stationarity conditions. On whole-slide images from pancreas, lung, and ovary tissues and on a real expert annotation stream, it tightly controls the error within cell types while allowing up to 70% of cells to bypass expert review. It builds prediction sets for difficult cases whose AI annotations are not accurate enough for later resolution. Its performance is robust to the choice of the review fraction and random seeds. Overall, CARED offers a perspective on how a small amount of human feedback can calibrate uncertainty quantification and yield effective and safe decisions.

Several directions remain open. First, our guarantee is asymptotic, and finite-sample confidence bounds for the realized error rate will be useful. Second, our error control is over the entire stream instead of the cases selected for automated annotation; the random audit design may be extended to calibrate the selective

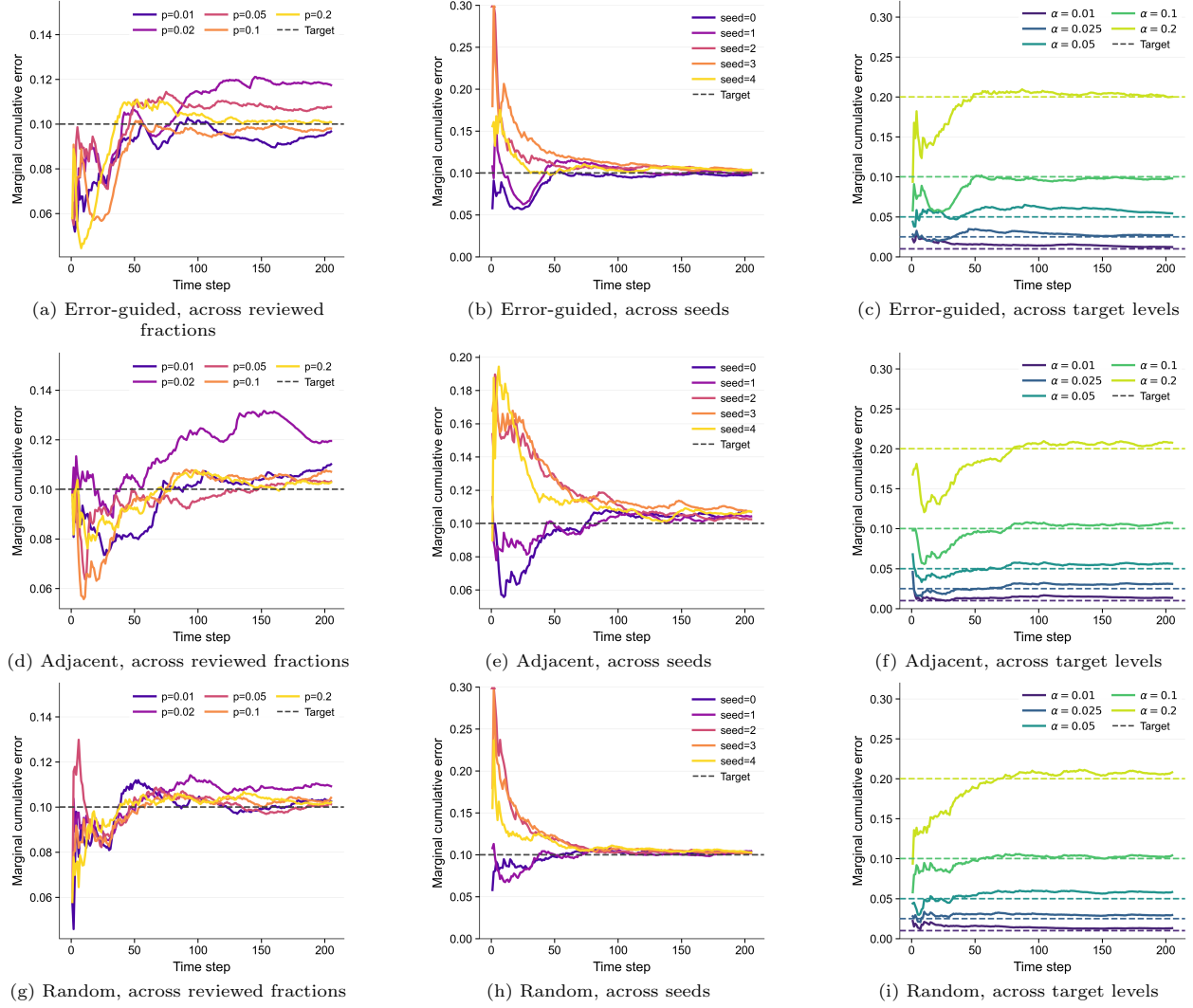


Figure 10: Sensitivity analysis of marginal cumulative error on pancreas tissue with the global NuClass score source. Rows correspond to the error-guided, adjacent, and random batch release schemes, while columns vary the reviewed fraction p , the random seed, and the target error level α . Dashed horizontal lines indicate the target error level used in each setting. In each column, only the parameter named in the column heading is varied; the other two retain their default values, $p = 0.1$, seed 0, and $\alpha = 0.1$, used in Sections 4.2 and 4.3.

errors. Finally, we focus primarily on retrospective evaluation, and a prospective deployment in real-time annotation tasks could further quantify the saved benefits and impact on downstream analysis and decisions.

References

- 10x Genomics (2024a). FFPE Human Lung Cancer Data with Human Immuno-Oncology Profiling Panel and Custom Add-on. Xenium In Situ Gene Expression Dataset by Xenium Onboard Analysis v2.0.0. May 14.
- 10x Genomics (2024b). FFPE Human Ovarian Cancer Data with Human Immuno-Oncology Profiling Panel and Custom Add-on. Xenium In Situ Gene Expression Dataset by Xenium Onboard Analysis v2.0.0. May 14.

- 10x Genomics (2024c). FFPE Human Pancreatic Ductal Adenocarcinoma Data with Human Immunology Profiling Panel. Xenium In Situ Gene Expression Dataset by Xenium Onboard Analysis v2.0.0. May 14.
- Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024). Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*.
- Angus, D. C., Khera, R., Lieu, T., Liu, V., Ahmad, F. S., Anderson, B., Bhavani, S. V., Bindman, A., Brennan, T., Celi, L. A., et al. (2025). Ai, health, and health care today and tomorrow: the JAMA summit report on artificial intelligence. *Jama*, 334(18):1650–1664.
- Aubreville, M., Stathonikos, N., Bertram, C. A., Klopffleisch, R., Ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T. A., Maier, A., et al. (2023). Mitosis domain generalization in histopathology images—the MIDOG challenge. *Medical Image Analysis*, 84:102699.
- Bernhardt, M., Castro, D. C., Tanno, R., Schwaighofer, A., Tezcan, K. C., Monteiro, M., Bannur, S., Lungren, M. P., Nori, A., Glocker, B., et al. (2022). Active label cleaning for improved dataset quality under resource constraints. *Nature communications*, 13(1):1161.
- Bodén, A. C., Molin, J., Garvin, S., West, R. A., Lundström, C., and Treanor, D. (2021). The human-in-the-loop: an evaluation of pathologists’ interaction with artificial intelligence in clinical practice. *Histopathology*, 79(2):210–218.
- Candès, E. J., Ilyas, A., and Zrnic, T. (2025). Probably approximately correct labels. *arXiv preprint arXiv:2506.10908*.
- Cazzaniga, G., Del Carro, F., Eccher, A., Becker, J. U., Gambaro, G., Rossi, M., Pieruzzi, F., Fraggetta, F., Pagni, F., and L’Imperio, V. (2025). Improving the annotation process in computational pathology: A pilot study with manual and semi-automated approaches on consumer and medical grade devices. *Journal of Imaging Informatics in Medicine*, 38(2):1112–1119.
- Charusaie, M.-A., Mozannar, H., Sontag, D., and Samadi, S. (2022). Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, pages 2972–3005. PMLR.
- Corbetta, D., Finos, L., Geistlinger, L., and Risso, D. (2026). Conformal inference for cell-type annotation with graph-structured constraints. *Journal of the Royal Statistical Society Series C: Applied Statistics*, page qlag037.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.
- Dhillon, G. S., Deligiannidis, G., and Rainforth, T. (2024). On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pages 1549–1557. PMLR.
- Ergen, C., Xing, G., Xu, C., Kim, M., Jayasuriya, M., McGeever, E., Oliveira Pisco, A., Streets, A., and Yosef, N. (2024). Consensus prediction of cell type labels in single-cell data with popV. *Nature Genetics*, 56(12):2731–2738.
- Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.
- Gaffney, M. H. (2025). Pathology in the artificial intelligence era: Guiding innovation and implementation to preserve human insight. *Academic Pathology*, 12(1):100166.
- Ge, H., Bastani, H., and Bastani, O. (2024). Stochastic online conformal prediction with semi-bandit feedback. *arXiv preprint arXiv:2405.13268*.

- Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.
- Gibbs, I. and Candès, E. J. (2024). Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36.
- Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., et al. (2023). LLMs accelerate annotation for medical information extraction. In *machine learning for health (ML4H)*, pages 82–100. PMLR.
- Goren, S., Galil, I., and El-Yaniv, R. (2024). Hierarchical selective classification. *Advances in Neural Information Processing Systems*, 37:111047–111073.
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410.
- Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.-R., and Binder, A. (2020). Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports*, 10(1):6423.
- Hall, P., Heyde, C. C., Birnbaum, Z., and Lukacs, E. (1980). *Martingale limit theory and its application*, volume 12. Elsevier.
- Hou, W. and Ji, Z. (2024). Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature methods*, 21(8):1462–1465.
- Howard, F. M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O. I., Kather, J. N., et al. (2021). The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423.
- Huang, H., Liao, W., Xi, H., Zeng, H., Zhao, M., and Wei, H. (2025). Model-agnostic selective labeling with provable statistical guarantees. *arXiv preprint arXiv:2510.14581*.
- Jin, Y. and Candès, E. J. (2023). Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41.
- Jin, Y., Moon, I., and Zitnik, M. (2026). Act or defer: Error-controlled decision policies for medical foundation models. *medRxiv*, pages 2026–02.
- Kaymaz, Y., Ganglberger, F., Tang, M., Haslinger, C., Fernandez-Albert, F., Lawless, N., and Sackton, T. B. (2021). HieRFIT: a hierarchical cell type classification tool for projections from complex single-cell atlas datasets. *Bioinformatics*, 37(23):4431–4436.
- Kim, C., Gadgil, S. U., and Lee, S.-I. (2026). Transparency of medical artificial intelligence systems. *Nature reviews bioengineering*, 4(1):11–29.
- Kiyani, S., Pappas, G., and Hassani, H. (2024). Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563.
- Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, 16(12):1289–1296.

- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.
- Li, S., Xu, J., Bao, T., Liu, Y., Liu, Y., Liu, Y., Wang, L., Lei, W., Wang, S., Xu, Y., et al. (2025a). A co-evolving agentic AI system for medical imaging analysis. *arXiv preprint arXiv:2509.20279*.
- Li, Z., Mirjahanmardi, S. H., Sali, R., Eweje, F., Gopaulchan, M., Kloker, L., Zhang, X., Li, G., Jiang, Y., and Li, R. (2025b). Automated cell annotation and classification on histopathology for spatial biomarker discovery. *Nature Communications*, 16(1):6240.
- Liang, J., Ren, Z., and Chen, Y. (2026). Optimal training-conditional regret for online conformal prediction. *arXiv preprint arXiv:2602.16537*.
- López-De-Castro, M., García-Galindo, A., González-Gomariz, J., and Armañanzas, R. (2025). Conformal inference for reliable single cell RNA-seq annotation. *Bioinformatics*, 41(10):btaf521.
- Lotfollahi, M., Hao, Y., Theis, F. J., and Satija, R. (2024). The future of rapid and automated single-cell data analysis using reference mapping. *Cell*, 187(10):2343–2358.
- Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology*, 40(1):121–130.
- Montezuma, D., Oliveira, S. P., Neto, P. C., Oliveira, D., Monteiro, A., Cardoso, J. S., and Macedo-Pinto, I. (2023). Annotating for artificial intelligence applications in digital pathology: a practical guide for pathologists and researchers. *Modern Pathology*, 36(4):100086.
- Montezuma, D., Oliveira, S. P., Tolkach, Y., Boor, P., Haragan, A., Carvalho, R., Della Mea, V., Kiehl, T.-R., Leh, S., Yousif, M., et al. (2025). Annotation practices in computational pathology: a european society of digital and integrative pathology (ESDIP) survey study. *Laboratory Investigation*, 105(3):102203.
- Mortier, T., Javanmardi, A., Sale, Y., Hüllermeier, E., and Waegeman, W. (2025). Conformal prediction in hierarchical classification. *arXiv preprint arXiv:2501.19038*.
- Mozannar, H. and Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR.
- Northcutt, C. G., Athalye, A., and Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.
- Qiu, J., Wilm, F., Öttl, M., Schlereth, M., Liu, C., Heimann, T., Aubreville, M., and Breininger, K. (2023). Adaptive region selection for active learning in whole slide image semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 90–100. Springer.
- Rau, T. T., Cross, W., Lastra, R. R., Lo, R. C.-L., Matoso, A., and Herrington, C. S. (2024). Closing the loop—the role of pathologists in digital and computational pathology research.
- Stacke, K., Eilertsen, G., Unger, J., and Lundström, C. (2020). Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336.
- Stutz, D., Cemgil, A. T., Doucet, A., et al. (2021). Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*.
- Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and Van Der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544.

- Theunissen, L., Mortier, T., Saeys, Y., and Waegeman, W. (2024). Uncertainty-aware single-cell annotation with a hierarchical reject option. *Bioinformatics*, 40(3):btae128.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
- Vovk, V., Lindsay, D., Nouretdinov, I., Gammerman, A., et al. (2003). Mondrian confidence machine. *Technical Report*.
- Xu, C. and Xie, Y. (2023). Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pages 38707–38727. PMLR.
- Xu, Y., Cui, Y., Li, M., and Huang, Z. (2025). Adaptive multi-scale integration unlocks robust cell annotation in histopathology images. *arXiv preprint arXiv:2511.13586*.
- Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *International conference on machine learning*, pages 25834–25866. PMLR.
- Zhang, D., Chatzimparmpas, A., Kamali, N., and Hullman, J. (2024). Evaluating the utility of conformal prediction sets for ai-advised image labeling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- Zhao, M., Admoni, H., Simmons, R., Ramdas, A., and Bajcsy, A. (2025). Conformalized interactive imitation learning: Handling expert shift and intermittent feedback. In *International Conference on Learning Representations*, volume 2025, pages 9238–9251.

A Details on utility module

This appendix gives further details on the optional utility module introduced in Section 3.2: the online selection algorithm, the structural-coherence metric, and the construction of the score transformation f_Θ .

A.1 Structural coherence and surrogate loss

Structural coherence. The structural-coherence metric captures coarse organization in the label space. Let $\{\mathcal{S}_1, \dots, \mathcal{S}_M\}$ be a partition of $[K]$ into M subtrees, where labels in the same subtree are structurally close (for example, by lineage, function, or morphology). Consider a map $g : [K] \rightarrow [M]$ that assigns each class to its subtree. For case i and subtree m , define

$$\text{Occ}_{i,m} = \max_{k:g(k)=m} \mathbb{1}\{k \in \hat{C}_{t,i}(\Theta)\}, \quad \text{Full}_{i,m} = \min_{k:g(k)=m} \mathbb{1}\{k \in \hat{C}_{t,i}(\Theta)\},$$

and

$$\text{Out}_{i,m} = \max_{m' \neq m} \text{Occ}_{i,m'}.$$

Here $\text{Occ}_{i,m}$ indicates whether subtree m occurs in the prediction set, $\text{Full}_{i,m}$ indicates whether the entire subtree is included, and $\text{Out}_{i,m}$ indicates whether the set also contains a label outside subtree m . We call a prediction set structurally coherent if

$$V_i(\Theta) := \sum_{m=1}^M \text{Occ}_{i,m} (1 - \text{Full}_{i,m}) \text{Out}_{i,m} = 0. \quad (\text{A.1})$$

This criterion penalizes a set that crosses subtree boundaries while including only a partial fragment of at least one touched subtree. It permits a small set contained in one subtree and a set formed by several complete subtrees, both of which represent uncertainty coherently. The batch-level score in (5) is

$$\mathcal{J}_{\text{struct}}(\Theta) = \frac{1}{B_t} \sum_{i=1}^{B_t} \mathbb{1}\{V_i(\Theta) = 0\}.$$

The rule in (A.1) is one admissible structural preference, not a required component of the conformal framework. Other domain-specific preferences, such as homogeneity, diversity, or a deeper hierarchy, can replace the structural component in the hard utility objective and its corresponding soft relaxation, while leaving the adaptive conformal framework in Section 3.1 unchanged.

Surrogate loss. The hard utility objective in (5) contains three operations that are not directly amenable to gradient-based updates: the membership indicators $\mathbb{1}\{k \in \hat{C}_{t,i}(\Theta)\}$, the set-size indicator $\mathbb{1}\{|\hat{C}_{t,i}(\Theta)| \leq 1\}$ in $J_{\text{skip}}(\Theta)$, and the structural-coherence indicator $\mathbb{1}\{V_i(\Theta) = 0\}$ in $J_{\text{struct}}(\Theta)$. We construct the surrogate loss $L_{\text{soft}}(\Theta)$ by replacing the operations above in (5) as follows.

- **Membership indicators.** Given the class-wise conformal thresholds $\{q_{k,t}\}_{k=1}^K$ computed under parameter Θ , $\mathbb{1}\{k \in \hat{C}_{t,i}(\Theta)\} = \mathbb{1}\{f_\Theta(s_t(x_{t,i}))_k \geq q_{k,t}\}$ is replaced by

$$z_{i,k}(\Theta) = \sigma\left(\frac{f_\Theta(s_t(x_{t,i}))_k - q_{k,t}}{T_{\text{temp}}}\right),$$

where T_{temp} controls the smoothness of the approximation.

- **Set-size indicator.** We interpret $z_{i,k}(\Theta)$ as the soft inclusion probability of class k . Then $\prod_{k=1}^K (1 - z_{i,k}(\Theta))$ corresponds to including no class, while $z_{i,k}(\Theta) \prod_{\ell \neq k} (1 - z_{i,\ell}(\Theta))$ corresponds to including class k and no other class. Summing the latter over k therefore corresponds to including exactly one class. Therefore, the indicator $\mathbb{1}\{|\hat{C}_{t,i}(\Theta)| \leq 1\}$ is replaced by

$$\prod_{k=1}^K (1 - z_{i,k}(\Theta)) + \sum_{k=1}^K z_{i,k}(\Theta) \prod_{\ell \neq k} (1 - z_{i,\ell}(\Theta)).$$

- **Structural-coherence indicator.** In (A.1), we replace the membership indicators in $\text{Occ}_{i,m}$, $\text{Full}_{i,m}$, and $\text{Out}_{i,m}$ by $z_{i,k}(\Theta)$, while retaining the max and min operations. Suppose this yields a surrogate $\hat{V}_i(\Theta)$, which reflects the extent of structural violation. We therefore replace $\mathbb{1}\{V_i(\Theta) = 0\}$ by one minus this expression, so that larger values continue to represent greater structural coherence. The retained max and min operations are differentiable almost everywhere and admit subgradients at ties, so they can be handled directly in the gradient updates.

A.2 Online selection algorithm

Algorithm A.1 gives the online selection procedure sketched in Section 3.2. It selects the score-transformation parameter Θ_t before the audit set is sampled.

Algorithm A.1 Online training of Θ_t at time step t

Input: Batch D_t with raw score vectors $\{s_t(x) : x \in D_t\}$; calibration sets $\{D_{k,\text{cal}}(t)\}_{k=1}^K$ (denoted by $D_{\text{cal}}(t)$); current online levels $\{\alpha_{k,t}\}_{k=1}^K$; candidate number N ; learning rate η .

Output: Selected parameter Θ_t and score transformation f_{Θ_t} .

```

1:  $\mathcal{J}^* \leftarrow \mathcal{J}(0)$ ;  $\Theta \leftarrow 0$ ;  $\Theta^* \leftarrow 0$ .
2: for  $n = 1, \dots, N$  do
3:   Compute class-wise conformal quantiles  $\{q_{k,t}\}_{k=1}^K$  on  $D_{\text{cal}}(t)$  using  $\Theta$ 
4:    $\Theta \leftarrow \Theta - \eta \nabla_{\Theta} \mathcal{L}_{\text{soft}}(D_t; \Theta, \{q_{k,t}\}_{k=1}^K)$  ► gradient step
5:   Using the updated  $\Theta$ , recompute  $\{q_{k,t}\}_{k=1}^K$  according to Equation (1), construct prediction sets on  $D_t$  according to Equation (2), and evaluate the utility objective  $\mathcal{J}(\Theta)$ 
6:   if  $\mathcal{J}(\Theta) < \mathcal{J}^*$  then
7:      $\mathcal{J}^* \leftarrow \mathcal{J}(\Theta)$ ;  $\Theta^* \leftarrow \Theta$ 
8:   end if
9: end for
10: Set  $\Theta_t \leftarrow \Theta^*$ .

```

A.3 Score transformation construction

Algorithm A.2 Score transformation f_{Θ}

Input: Score vector $s \in [0, 1]^K$; subtree partition $\{\mathcal{S}_1, \dots, \mathcal{S}_M\}$ of $[K]$; parameter matrices $\Theta^{\text{in}}, \Theta^{\text{out}} \in \mathbb{R}^{K \times M}$.

Output: Transformed score vector $\tilde{s} = f_{\Theta}(s) \in \mathbb{R}^K$.

```

1:  $s_{\max} \leftarrow \max_k s_k$ 
2:  $u_k \leftarrow \text{logit}(s_k / s_{\max})$  for all  $k$ 
3:  $\mathcal{U} \leftarrow [K]$  ► unprocessed classes
4: for  $r = 1, \dots, M$  do
5:    $\ell_r \leftarrow \arg \max_{k \in \mathcal{U}} u_k$  ► leader class
6:    $m \leftarrow$  subtree index such that  $\ell_r \in \mathcal{S}_m$ 
7:   for  $k \in (\mathcal{S}_m \cap \mathcal{U}) \setminus \{\ell_r\}$  do
8:      $u_k \leftarrow u_k + \Theta_{k,r}^{\text{in}}$  ► updates scores associated to the leader subtree
9:   end for
10:  for  $j \in \mathcal{U} \setminus \mathcal{S}_m$  do
11:     $u_j \leftarrow u_j - \Theta_{j,r}^{\text{out}}$  ► updates scores associated to other subtrees
12:  end for
13:   $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}_m$ 
14: end for
15:  $\tilde{s}_k \leftarrow s_{\max} \cdot \sigma(u_k)$  for all  $k$  ►  $\tilde{s}_k \in (0, s_{\max})$ 
16: return  $\tilde{s} = f_{\Theta}(s) \in \mathbb{R}^K$ 

```

Algorithm A.2 gives the construction of the score transformation f_Θ used in Section 3.2. Given a score vector s and a parameter Θ , it returns the transformed score $\tilde{s} = f_\Theta(s)$. The construction of f_Θ is guided by three principles. First, initializing Θ^{in} and Θ^{out} at zero should yield the identity map $f_0(s) = s$, so the unmodified score vector is always available as a baseline candidate. Second, every transformed entry should lie in $(0, s_{\max}) \subset (0, 1)$, preserving the score range required by the conformal construction. Finally, the transformation should respect the prescribed subtree structure, coordinating score changes within each subtree while allowing different subtrees to change differently. For numerical stability, in implementation we lower-bound $s_{\max}$ by 10^{-12} and clip each normalized score $s_k/s_{\max}$ to $[10^{-4}, 1 - 10^{-4}]$ before applying the logit transformation. These numerical safeguards avoid undefined values at the endpoints and are omitted from Algorithm A.2 for notational simplicity.

B Additional theoretical results

B.1 Unbiased randomized rounding

When the target reviewed fraction is fixed at p , the desired audit budget pB_t may not be an integer. The following randomized rounding scheme implements this budget while preventing the resulting rounding error from accumulating systematically over time, and yields the class-wise asymptotic relations required in the proofs of Lemma F.2 and Theorem 3.3; its proof is deferred to Appendix E.5.

Proposition B.1 (Unbiased randomized rounding). *Fix $p \in (0, 1)$. Let $\mathcal{G}'_t$ be the time- t sigma-field generated by all random quantities determined after D_t is realized but before choosing p_t . At time t , before A_t is sampled, set*

$$q_t := pB_t - \lfloor pB_t \rfloor$$

and draw $\xi_t \mid \mathcal{G}'_t \sim \text{Bernoulli}(q_t)$. Now sample the audit set with size

$$|A_t| := \lfloor pB_t \rfloor + \xi_t,$$

Under Assumption 3.1(i)–(iii), the realized reviewed fractions $p_t = |A_t|/B_t$ under this rule satisfy, for every class $k \in [K]$,

$$\frac{\sum_{t=1}^T (p_t - p) |D_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} \rightarrow 0 \quad a.s., \quad \frac{\sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |A_{k,t}|} \rightarrow 0 \quad a.s.$$

B.2 A cumulative batch-wise deviation guarantee

The guarantees in Section 3.3 control cumulative class-wise errors, the relevant notion for long-run validity. In practice, one also wants to understand how the procedure behaves locally across individual batches (Gibbs and Candès, 2024). A natural way to quantify such local behavior is to aggregate the absolute per-batch deviations from the target error level. The result below gives an upper bound for this cumulative local-deviation measure, and discusses its asymptotic properties as the time index T and batch size $|D_{k,t}|$ get large. Note that in this subsection, the incoming batch D_t may be selected using review feedback observed before time t (as in the error-guided scheme).

We first set up the problem. Fix a class k . Here we consider only nonempty class- k batches because we stipulate that there is no deviation on empty sets. As a simplification, we assume that $|D_{k,t}| > 0$ for all t .

More notation. Write the batch-wise class- k error rates as

$$\text{BatErr}_{k,t}^D := \frac{\sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{|D_{k,t}|}, \quad \text{BatErr}_{k,t}^A := \frac{\sum_{x \in A_{k,t}} \text{err}_{k,t}(x)}{|A_{k,t}|},$$

whenever $|A_{k,t}| > 0$. If $|A_{k,t}| = 0$, we define $\text{BatErr}_{k,t}^A := \text{BatErr}_{k,t}^D$. The quantity of interest is the cumulative batch-wise deviation of the all-cell error rate from the target level:

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \alpha|.$$

Define the effective batch step size and the clipped online level by

$$\eta_{k,t} := \gamma_k |A_{k,t}|, \quad \bar{\alpha}_{k,t} := \text{clip}_{[0,1]}(\alpha_{k,t}).$$

Assumption B.2. *Throughout this subsection, assume the following:*

(i) *The lower bound of the batch size B^- satisfies that*

$$B^- > \max\left\{\frac{1}{p}, \frac{1}{1-p}\right\}.$$

(ii) *For every class $k \in [K]$, the effective batch step size satisfies*

$$\eta_{k,t} \leq 1 \quad \text{for all } t.$$

(iii) *For every class $k \in [K]$ and time t , the scores*

$$\{\tilde{s}_{k,t}(x) : x \in D_{k,\text{cal}}(t) \cup D_{k,t}\}$$

are distinct.

The discussion on batch-wise deviation is sufficiently informative only when the batch sizes are not excessively small. Assumption B.2(i) can be interpreted as "at least one cell per batch is chosen for online review on average", which is a plausible condition for large-scale settings. This condition leads to intermediate consequences

$$B^- > 2 \quad \text{and} \quad 0 < p^- \leq p_t \leq p^+ < 1 \quad \text{for all } t,$$

for some deterministic constants p^-, p^+ . Specifically, the second consequence holds with $p^- = p - \frac{1}{B^-} > 0$ and $p^+ = p + \frac{1}{B^-} < 1$. Assumption B.2(ii) can be satisfied when the step size γ_k is chosen reasonably small; a sufficient condition is $\gamma_k < \frac{1}{B^+}$ for every class k .

Definition B.3 (Calibration discrepancy). *For each time t with $|D_{k,t}| > 0$, define the calibration discrepancy as*

$$\Delta_{k,t} := \sup_{u \in [0,1]} \left| \frac{\sum_{x \in D_{k,t}} \mathbf{1}\{\tilde{s}_{k,t}(x) \leq \hat{q}_{k,t}(u)\}}{|D_{k,t}|} - \frac{\sum_{x \in D_{k,\text{cal}}(t)} \mathbf{1}\{\tilde{s}_{k,t}(x) \leq \hat{q}_{k,t}(u)\}}{|D_{k,\text{cal}}(t)|} \right| + \frac{1}{|D_{k,\text{cal}}(t)|}.$$

For $\alpha_{k,t} \in [0,1]$, the clipped $\bar{\alpha}_{k,t} = \alpha_{k,t}$. By the class-wise conformal construction in Section 3.1, a class- k error occurs when $\tilde{s}_{k,t}(x) < \hat{q}_{k,t}(\bar{\alpha}_{k,t})$; Assumption B.2(iii) ensures that no score in $D_{k,t}$ equals this threshold, so

$$\text{BatErr}_{k,t}^D = \frac{\sum_{x \in D_{k,t}} \mathbf{1}\{\tilde{s}_{k,t}(x) \leq \hat{q}_{k,t}(\bar{\alpha}_{k,t})\}}{|D_{k,t}|}.$$

Here, $\hat{q}_{k,t}(\bar{\alpha}_{k,t})$ is the smallest threshold for which at least a $\bar{\alpha}_{k,t}$ fraction of the calibration scores lie below the threshold. Assumption B.2(iii) makes the calibration scores distinct, so this empirical fraction can exceed $\bar{\alpha}_{k,t}$ by at most one calibration observation, or $1/|D_{k,\text{cal}}(t)|$. Therefore,

$$|\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| \leq \Delta_{k,t}. \tag{B.2}$$

When $\alpha_{k,t} \notin [0,1]$, the same inequality follows directly from the quantile conventions in Section 3.1. The first term in $\Delta_{k,t}$ captures the calibration mismatch caused by distribution shift between the current class- k

batch and the corresponding calibration set; whereas the second term in $\Delta_{k,t}$ accounts for the finite-sample correction, which is negligible since $|D_{k,\text{cal}}(t)|$ is large.

Before establishing an informative upper bound on the cumulative absolute errors, we impose additional assumptions that may not hold in every practical setting. Nevertheless, they facilitate an interpretable characterization of the batchwise fluctuations of the errors produced by CARED. We consider two distinct sets of assumptions, each leading to a different guarantee: one constrains the range of the class- k batch sizes $|D_{k,t}|$, whereas the other assumes a nondecreasing sequence formed by the effective step sizes $\eta_{k,t}$. Subsections B.2.1 and B.2.2 present the corresponding results, respectively.

B.2.1 Cumulative deviation under constrained batch size

Assumption B.4 (Bounded batch size range). *For any class k , there exist deterministic constants $d_k > 0$ and $M_k \geq 1$ such that*

$$d_k \leq |D_{k,t}| \leq M_k d_k \quad \text{for every } t.$$

That is, the number of class- k cells remains controlled across batches.

Proposition B.5 (Cumulative batch-wise deviation guarantee). *Consider any class $k \in [K]$, suppose that Assumptions 3.1, B.2, and B.4 hold. Then, for any $\epsilon \in (0, 1)$, define*

$$\mathcal{R}_{k,T}(\epsilon) := \frac{1}{p^-} \left[\frac{T}{2} \sqrt{\frac{B^-}{B^- - 1} p^+ (1 - p^-)} \frac{M_k}{d_k} + \frac{5}{2} \sqrt{T p^+ (1 - p^-)} \frac{M_k}{d_k} \log \frac{2}{\epsilon} + \frac{5M_k}{12} \log \frac{2}{\epsilon} \right].$$

With probability at least $1 - \epsilon$, it holds that

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \alpha| \leq \left(1 + \frac{p^+}{p^-} M_k\right) \sum_{t=1}^T \Delta_{k,t} + \mathcal{R}_{k,T}(\epsilon).$$

Corollary B.6 (Asymptotic cumulative batch-wise deviation guarantee). *Under the setting of Proposition B.5, we consider a sequence of instances for which $T \rightarrow \infty$ and the batch size becomes as large as possible, as in $d_k \rightarrow \infty$. Then, for any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$,*

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \alpha| \leq (1 + M_k) \sum_{t=1}^T \Delta_{k,t} + \mathcal{O}\left(\frac{T}{\sqrt{d_k}}\right) + o_T(T).$$

Thus, in the combined regime where $T \rightarrow \infty$ and $d_k \rightarrow \infty$, all terms other than the calibration mismatch are asymptotically negligible relative to T . Therefore, the calibration mismatch is the only asymptotically non-negligible contribution that can accumulate linearly with the horizon, with coefficient $1 + M_k$, which approaches 2 as $M_k \rightarrow 1$.

B.2.2 Cumulative deviation under nondecreasing effective step size

Assumption B.7 (Nondecreasing effective step size). *Consider any class $k \in [K]$ and any horizon T , assume that almost surely*

$$0 < \eta_{k,1} \leq \eta_{k,2} \leq \dots \leq \eta_{k,T} < 1.$$

That is, the effective step sizes across batches form a nondecreasing sequence.

Proposition B.8 (Cumulative batch-wise deviation guarantee). *Fix a class $k \in [K]$. Suppose that Assumptions 3.1, B.2, and B.7 hold. Define*

$$\mu_{k,t} := \frac{1}{\sqrt{2p^- |D_{k,t}|}} + \exp\left(-\frac{p^- |D_{k,t}|}{12}\right), \quad (\text{B.3})$$

and

$$S_{k,T}(\epsilon) := \sum_{t=1}^T \mu_{k,t} + \sqrt{2T \log \frac{1}{\epsilon}} + \frac{1}{3} \log \frac{1}{\epsilon}. \quad (\text{B.4})$$

Then, for every $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$,

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \alpha| \leq 2 \sum_{t=1}^T \Delta_{k,t} + S_{k,T}(\epsilon).$$

Equations (B.3) and (B.4) are constructed in this way because $S_{k,T}(\epsilon)$ forms an upper bound for cumulative sampling fluctuations $\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \text{BatErr}_{k,t}^A|$; this is justified in Lemma F.3.

Corollary B.9 (Asymptotic cumulative batch-wise deviation guarantee). *Under the setting of Proposition B.8, we consider a sequence of instances for which $T \rightarrow \infty$ and the batch size becomes asymptotically large, that is, there exists d_k such that $d_k \leq |D_{k,t}|$ for any $1 \leq t \leq T$ and $d_k \rightarrow \infty$. Then, for every $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$,*

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \alpha| \leq 2 \sum_{t=1}^T \Delta_{k,t} + \mathcal{O}\left(\frac{T}{\sqrt{d_k}}\right) + o_T(T).$$

Thus, as the horizon T and the class- k batch size grow asymptotically large, the calibration mismatch is the only nonnegligible term that accumulates linearly with time, with coefficient 2. Unlike the coefficient in Corollary B.6, this value does not depend on the class-size range.

C Additional details and results of experiments

C.1 Experimental details

This subsection collects implementation details and auxiliary experimental settings omitted from Section 4. The details include the data processing, the selection rule for the class-wise step sizes γ_k , the structural subtree partition used by the utility-aware module, and the settings for the experiments in Section 4.

Dataset processing. In each dataset (cohort), the H&E stain was applied to the same tissue section after the Xenium run and imaged at $40\times$ ($0.274 \mu\text{m}/\text{pixel}$). For all experiments, the images are downsampled to approximately $0.415 \mu\text{m}/\text{pixel}$. Cell locations were registered to the H&E image using 10x Genomics Xenium Ranger, a post-processing pipeline for Xenium data, then filtered by 1,000 cells.

Selection rule for γ_k . In the experiments, the step sizes are selected before the online stream is run and then kept fixed over time. To satisfy the assumptions in Appendix B.2, we choose γ_k based on the class frequency in the test cohort, so that the effective batch step size $\gamma_k |A_{k,t}|$ remains below one with margin throughout the stream. Although this simulation uses oracle class-frequency information, in practice, the frequency can be roughly specified from domain knowledge. The assignment rule for γ_k is reported in Table 2.

Structural subtree. In our experiments, the structural subtree partition is specified by grouping cell types with closely related cellular characteristics into the same subtree. Specifically, the pancreas labels are grouped into three subtrees according to cellular function, while the ovary and lung labels are organized into two subtrees according to the benign-malignant distinction. The concrete subtree assignments are reported in Table 2.

Parameters used in experiments. Table 2 lists the experimental details and parameter values used in the main experiments.

Table 2: Details and parameters in the main experiments.

Name and notation	Choice used in the experiments
Held-out tissues	Pancreas, ovary, and lung
Label space for pancreas	Acinar, Ductal, Endocrine, Endothelial, Fibroblast-like, T Cell
Subtree partition for pancreas	$\mathcal{S}_1 = \{\text{Acinar, Ductal, Endocrine}\};$ $\mathcal{S}_2 = \{\text{Endothelial, Fibroblast-like}\}; \mathcal{S}_3 = \{\text{T Cell}\}$
Label space for ovary	Endothelial, Epithelial Malignant, Macrophage
Subtree partition for ovary	$\mathcal{S}_1 = \{\text{Endothelial, Macrophage}\}; \mathcal{S}_2 = \{\text{Epithelial Malignant}\}$
Label space for lung	Endothelial, Epithelial Malignant, T Cell, Tumor Associated Fibroblast
Subtree partition for lung	$\mathcal{S}_1 = \{\text{Endothelial, T Cell, Tumor Associated Fibroblast}\};$ $\mathcal{S}_2 = \{\text{Epithelial Malignant}\}$
Target error level α	0.1
Target reviewed fraction p	0.1
Step size of frequent classes ($> 25\%$)	$\gamma_k = 10^{-3}$
Step size of moderate classes ($5\%-25\%$)	$\gamma_k = 2 \times 10^{-3}$
Step size of scarce classes ($< 5\%$)	$\gamma_k = 10^{-2}$
Initial labeled block A_0	The first randomly selected batch
Batch-size bounds B^-, B^+	$B^- = 490, B^+ = 510$
Calibration window T_w	30 batches
Utility candidate number N	100
Utility learning rate η	0.1
Utility weights $(\lambda_{\text{skip}}, \lambda_{\text{size}}, \lambda_{\text{struct}})$	(20.0, 1.0, 5.0)
Soft-membership temperature T_{temp}	0.1
Random seed	0

C.2 More experimental results

For reference, we report the online level trajectories $\alpha_{k,t}$ under the *error-guided* batch-release scheme in the experiments reported in Section 4. These trajectories are shown in Figure 11. We then provide two additional sets of experimental results to show CARED’s stability. First, we report the cumulative conformal error trajectories on the three tissues for the remaining local and gated NuClass score sources, under the error-guided batch-release scheme. Second, we compare the utility metrics obtained from the utility-aware transformed scores and raw scores on the lung and ovary cohorts using the global NuClass score.

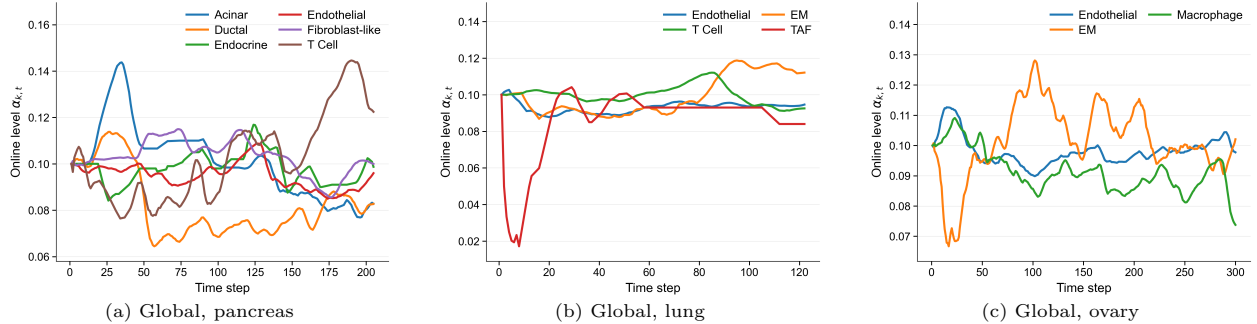


Figure 11: Online level trajectories $\alpha_{k,t}$ under the error-guided batch-release scheme using the global NuClass score. Panels (a)–(c) correspond to pancreas, lung, and ovary, respectively; each curve represents a cell type. The curves are smoothed with a window size of 7.

Together with the results under the global score source in Section 4, Figure 12 extends the same empirical

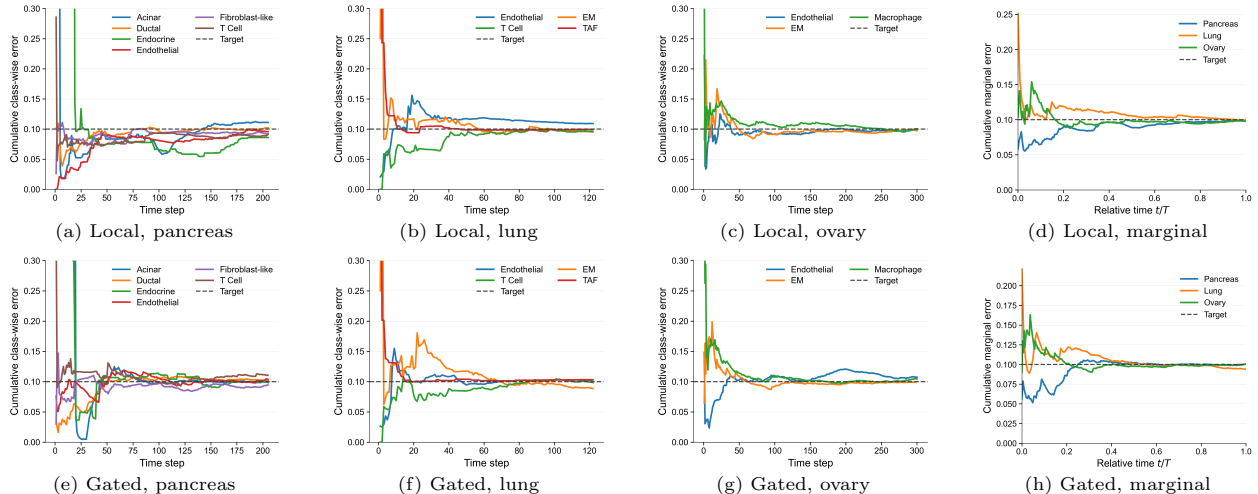


Figure 12: Additional cumulative conformal error trajectories under the error-guided batch-release scheme. Rows correspond to the local and gated NuClass scores, while columns show pancreas, lung, ovary, and marginal results. The first three columns report class-wise errors within each tissue; the marginal panels compare the marginal errors of the three tissues against the fraction of processed batches. Dashed horizontal lines mark the target error level $\alpha = 0.1$.

validity check to the remaining local and gated NuClass score sources, showing that the method continues to control class-wise errors across the considered tissues and score sources. We next examine the utility results on the lung and ovary tasks in Figure 13. For the lung results, the utility-aware score transformation shows a minor disadvantage relative to the raw-score baseline in the cumulative utility metrics. This does not contradict the intended scope of our framework, since the utility module is guaranteed to deliver prediction sets with no lower utility for a fixed batch of cells, while the resulting score transformation may still induce a less favorable trajectory of selected batches and the online levels $\alpha_{k,t}$ when cumulative utility is evaluated over the full stream. Moreover, the lung annotation task is not the primary target setting for the utility module: the skipped ratio is already high in this task, and the subtree structure has already been satisfied well. The more important evidence is the six-class pancreas annotation task, where the utility module shows clearer gains, as shown in Figure 8.

C.3 Additional sensitivity analyses

C.3.1 Class-wise cumulative error

Section 4.5 reports sensitivity analyses for the marginal cumulative error under variations of the reviewed fraction p , random seed, target error level α , and batch release scheme. Here, for each sensitivity setting, we report the final cumulative empirical error for each cell type after processing the full annotation stream, and compare these class-wise errors across the three batch release schemes.

Figure 14 can be further interpreted through Proposition 3.5, which implies that the deviation of the class-wise cumulative empirical error from the target level decreases at the inverse-square-root rate in the total number of cells from that class. Therefore, for a fixed finite stream, rarer cell types are expected to show greater sensitivity. This accounts for the relatively large fluctuations for endocrine cells, which constitute only 1.95% of the stream, and also for the milder instability of T cells and endothelial cells, which are two relatively scarce classes in this experiment. Note that the marginal cumulative-error trajectories in Section 4.5 remain stable, and the remaining class-wise curves in Figure 14 do not show comparable fluctuations. Thus, the class-wise fluctuations reflect finite-sample variability rather than a failure of the empirical error-control pattern.

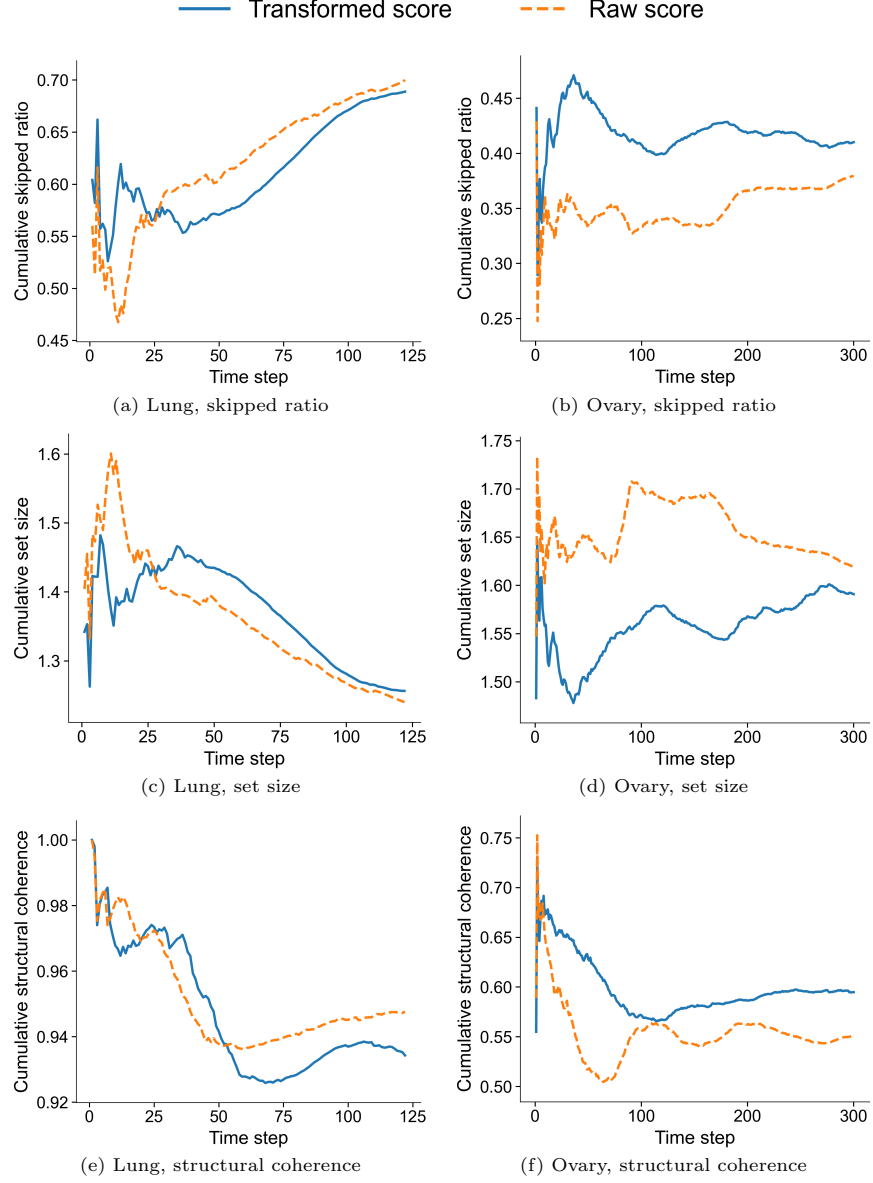


Figure 13: Additional cumulative utility comparison under the error-guided batch-release scheme using the global NuClass score. Rows correspond to the skipped ratio, average prediction set size, and structural coherence, while columns report lung and ovary. Solid blue curves use the utility-aware transformed score, while dashed orange curves use the raw score without the utility module. Higher skipped ratio and structural coherence and smaller prediction sets indicate better utility.

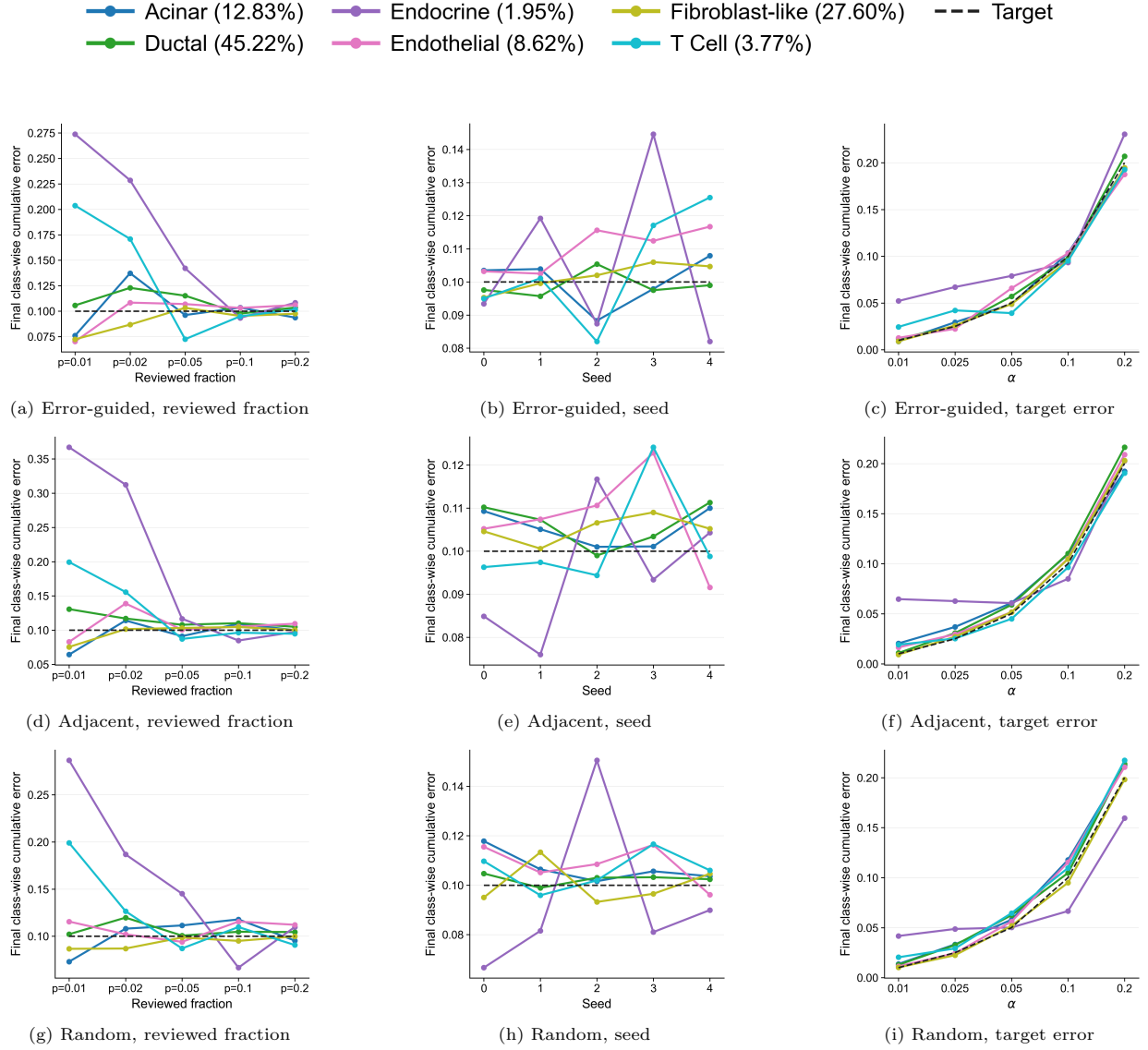


Figure 14: Class-wise sensitivity analysis of final cumulative empirical error on pancreas tissue with the global score source. Rows correspond to the error-guided, adjacent, and random batch release schemes, while columns vary the reviewed fraction p , the random seed, and the target error level α . Dashed horizontal lines indicate the target error level used in each setting, and percentages in the shared legend denote the empirical fraction of each cell type in the stream.

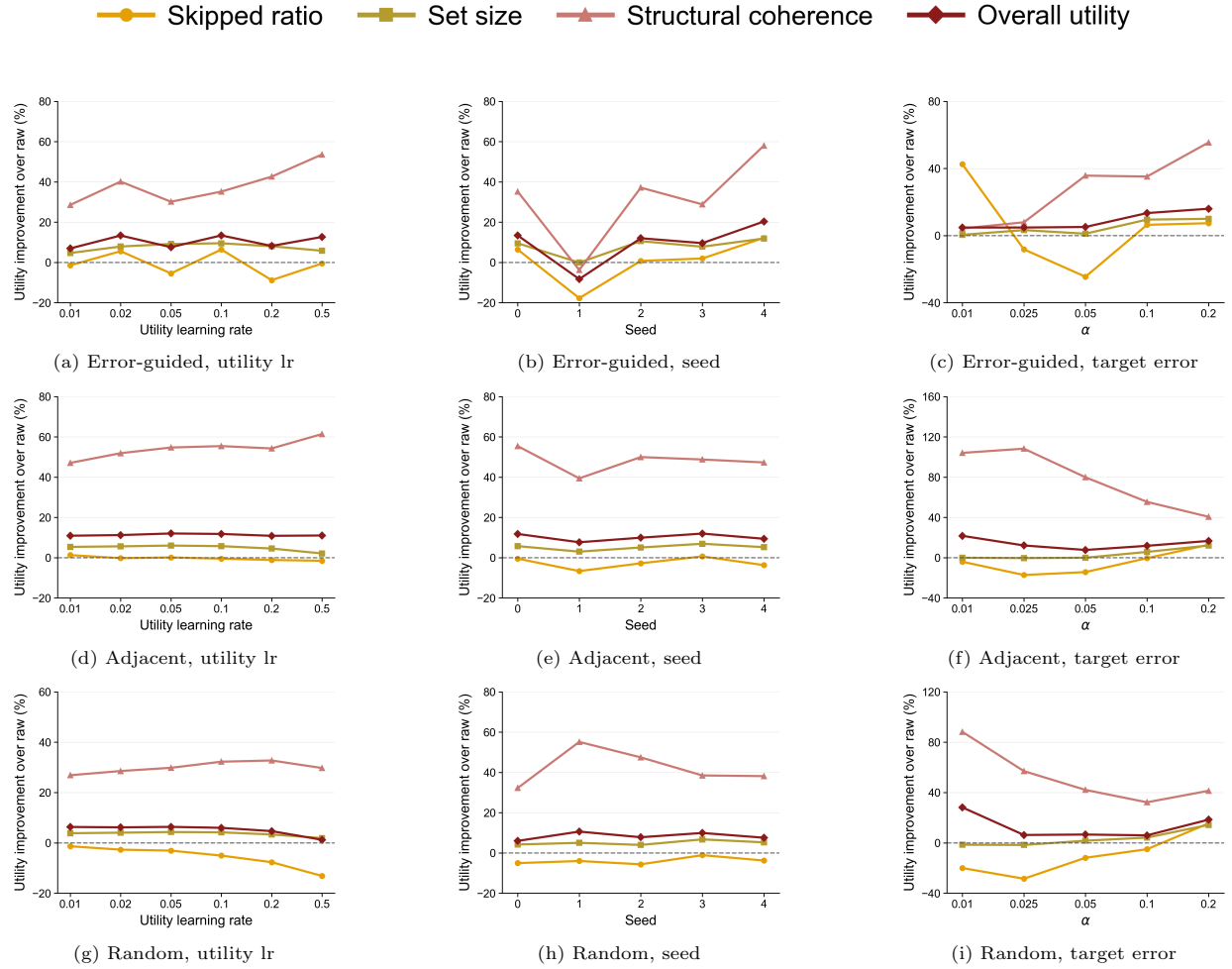


Figure 15: Sensitivity analysis of utility metrics on the six-class pancreas annotation task with the global score source. Rows correspond to the error-guided, adjacent, and random batch release schemes, while columns vary the utility learning rate η , the random seed, and the target error level α . Each panel compares the utility metrics obtained from the utility-aware transformed score and the raw score. The shared legend identifies the four utility metrics.

C.3.2 Utility metrics

We further examine the stability of the utility improvements on the six-class pancreas annotation task, as summarized in Figure 15.

For each batch release scheme and parameter setting, we compare the prediction sets produced by the utility-aware transformed score with those obtained from the raw scores. The three columns vary the utility learning rate η , the random seed, and the target error level α , and the three rows correspond to the error-guided, adjacent, and random batch release schemes. The plotted value is the percentage improvement over the raw-score baseline, computed from the cumulative stream-level utility metrics after processing the full annotation stream. Overall utility is the weighted aggregate of skipped ratio, set size, and structural coherence, using $(\lambda_{\text{skip}}, \lambda_{\text{size}}, \lambda_{\text{struct}}) = (20, 1, 5)$ as in Appendix C.1; its percentage improvement is normalized by the coordinate-wise worst-case six-class reference $(0, 6, 0)$, corresponding to zero skipped ratio, full set size, and zero structural coherence.

Figure 15 shows that the utility module consistently improves overall utility over the raw-score baseline in almost every considered case. The component-wise gains are more uneven: structural coherence improves substantially in most settings, whereas skipped ratio and set size can change more modestly or vary with the sensitivity parameter. This is expected because the aggregate utility balances three competing criteria, and the relative emphasis can be adjusted through the utility weights.

D Discussions on stepsize selection

The main experiments use a constant class-wise step size γ_k in the online update (4). This choice is aligned with the distribution-free guarantee established in Section 3.3: for each class k , the online level is updated as

$$\alpha_{k,t+1} = \alpha_{k,t} + \gamma_k \sum_{x \in A_{k,t}} (\alpha - \text{err}_{k,t}(x)),$$

where γ_k is fixed over time and selected according to the class frequency rule reported in Appendix C.1.

Decaying stepsize schedules provide a natural alternative. For the standard adaptive conformal prediction setting, Angelopoulos et al. (2024) show that step sizes of order $\gamma_t \propto t^{-1/2-\varepsilon}$, with $0 < \varepsilon < 1/2$, can retain long-run convergence of cumulative error while improving the stability of local error fluctuations. However, their local convergence theory is developed for the i.i.d. setting, and therefore should not be read as a general local-error guarantee under distribution shift. A complementary analysis of adaptive conformal prediction under distribution drift highlights the same trade-off from the non-i.i.d. perspective: smaller or decaying step sizes can stabilize the calibration-quantile trajectory within relatively stationary segments, but may adapt more slowly after distributional changes (Liang et al., 2026). Thus, decaying step sizes are a reasonable stability-oriented variant, but not a uniformly preferable replacement for the constant step sizes used in our main experiments.

To examine this issue, we rerun the pancreas experiment with the global NuClass score using the same initial class-wise scales γ_k as in the main experiment, but replace the constant update by

$$\gamma_{k,t} = \gamma_k t^{-2/3}.$$

Figure 16 compares the resulting per-batch all-cell deviations $|\text{BatErr}_{k,t}^D - \alpha|$ with those obtained from the constant stepsize update. The comparison shows a mixed pattern: the decaying schedule is closer to the target in some intervals and classes, while the constant schedule is closer in others. This is consistent with the stability-adaptation trade-off above, and explains why we keep the constant schedule as the default while reporting the decaying schedule as an empirical variant.

In practice, the numerical choice of γ_k is usually guided by empirical considerations rather than by an exact calculation. Although Theorem 3.3 shows that the cumulative class-wise error converges asymptotically for any fixed positive step size, the finite-stream behavior still depends on the scale of γ_k . To illustrate this

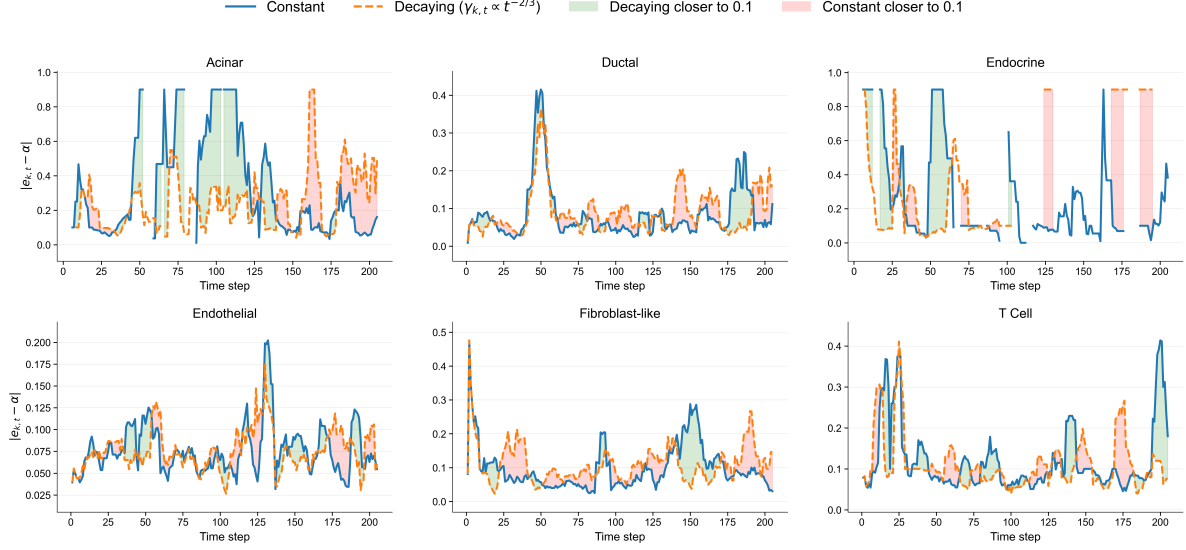


Figure 16: Comparison between constant and decaying class-wise step sizes on the pancreas experiment with the global NuClass score. The curves report per-batch all-cell deviations $|\text{BatErr}_{k,t}^D - \alpha|$ for each class under the constant schedule used in the main experiments and the decaying schedule $\gamma_{k,t} = \gamma_k t^{-2/3}$. Shaded intervals indicate which schedule is closer to the target error level $\alpha = 0.1$ at each time.

sensitivity, Figure 17 compares the constant stepsize update after multiplying the default class-wise step sizes by factors ranging from 0.1 to 10. Across this range, class-wise errors are generally pulled toward the target level $\alpha = 0.1$, although the finite-sample trajectories differ by class. The main exception is the endocrine class under large stepsize multipliers, where the finite-stream trajectory remains visibly above the target for longer. This behavior is consistent with the scarce-class instability discussed in Appendix C.3.1: endocrine cells account for only a small fraction of the stream, so their class-wise cumulative error is more sensitive in finite samples. This supports using the reported γ_k values as a pragmatic default rather than a deliberately optimal tuning.

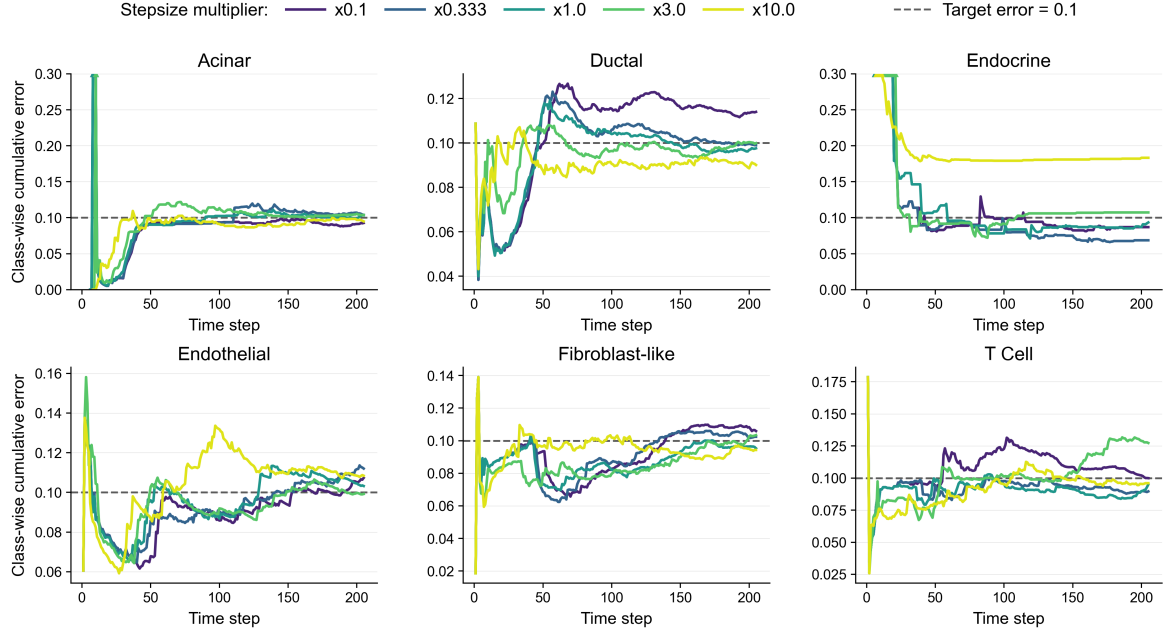


Figure 17: Sensitivity of cumulative class-wise error trajectories to the scale of the constant step sizes on the pancreas experiment with the global NuClass score. Each curve uses the same class-wise stepsize rule as in the main experiments, multiplied by the factor shown in the legend. The dashed line marks the target error level $\alpha = 0.1$.

E Technical proofs

E.1 Proof of Proposition 3.2

Telescoping the update rule (4) from the initial value $\alpha_{k,1} = \alpha$,

$$\alpha_{k,T+1} = \alpha + \gamma_k \sum_{t=1}^T \sum_{x \in A_{k,t}} (\alpha - \text{err}_{k,t}(x)).$$

Rearranging, $\sum_{t=1}^T \sum_{x \in A_{k,t}} \text{err}_{k,t}(x) - \alpha \sum_{t=1}^T |A_{k,t}| = (\alpha - \alpha_{k,T+1})/\gamma_k$. Dividing both sides by $\sum_{t=1}^T |A_{k,t}|$ and taking absolute values yields the exact identity

$$|\text{CovErr}_{k,T}^A - \alpha| = \frac{|\alpha_{k,T+1} - \alpha|}{\gamma_k \sum_{t=1}^T |A_{k,t}|}. \quad (\text{E.1})$$

It remains to bound $|\alpha_{k,T+1} - \alpha|$. Set $\bar{M}_k := \sup_{t \geq 1} |A_{k,t}|$. Since $\text{err}_{k,t}(x) \in \{0, 1\}$, the maximum single-step change satisfies

$$|\alpha_{k,t+1} - \alpha_{k,t}| = \gamma_k \left| \sum_{x \in A_{k,t}} (\alpha - \text{err}_{k,t}(x)) \right| \leq \gamma_k |A_{k,t}| \max\{\alpha, 1 - \alpha\} \leq \gamma_k \bar{M}_k \max\{\alpha, 1 - \alpha\}. \quad (\text{E.2})$$

We claim that for all $t \geq 1$,

$$\alpha_{k,t} \in [-\gamma_k \bar{M}_k (1 - \alpha), 1 + \gamma_k \bar{M}_k \alpha]. \quad (\text{E.3})$$

We establish the lower bound; the upper bound is symmetric. Fix $t \geq 1$. If $\alpha_{k,t} \geq 0$, the lower bound is immediate. Otherwise, let $t' < t$ be the last time with $\alpha_{k,t'} \geq 0$, which exists since $\alpha_{k,1} = \alpha \geq 0$. By (E.2),

$$\alpha_{k,t'+1} \geq \alpha_{k,t'} - \gamma_k \bar{M}_k (1 - \alpha) \geq -\gamma_k \bar{M}_k (1 - \alpha).$$

If $\alpha_{k,s} < 0$ for some s with $t' < s \leq t$, the quantile convention (1) sets $\hat{q}_{k,s}(\alpha_{k,s}) = -\infty$, so class k belongs to every prediction set and $\text{err}_{k,s}(x) = 0$ for all $x \in A_{k,s}$. The update then gives

$$\alpha_{k,s+1} - \alpha_{k,s} = \gamma_k |A_{k,s}| \alpha \geq 0,$$

so $\alpha_{k,s+1} \geq \alpha_{k,s}$; this holds regardless of whether $|A_{k,s}| = 0$. By the choice of t' , $\alpha_{k,s} < 0$ for every s with $t' < s \leq t$, so applying this monotonicity for $s = t' + 1, \dots, t - 1$ yields

$$\alpha_{k,t} \geq \alpha_{k,t'+1} \geq -\gamma_k \bar{M}_k (1 - \alpha).$$

Hence (E.3) holds, and consequently

$$|\alpha_{k,T+1} - \alpha| \leq \max\{\alpha, 1 - \alpha\} + \gamma_k \bar{M}_k \max\{\alpha, 1 - \alpha\}. \quad (\text{E.4})$$

Substituting (E.4) into (E.1):

$$|\text{CovErr}_{k,T}^A - \alpha| \leq \frac{\max\{\alpha, 1 - \alpha\}}{\gamma_k \sum_{t=1}^T |A_{k,t}|} + \frac{\bar{M}_k \max\{\alpha, 1 - \alpha\}}{\sum_{t=1}^T |A_{k,t}|}.$$

The first term matches the stated bound. For the second term, Lemma F.2 gives $\sum_{t=1}^T |A_{k,t}| / \sum_{t=1}^T |D_{k,t}| \rightarrow p$ a.s. By the assumed frequency lower bound and $B_t \geq B^-$, almost surely there exists some $T_0 < \infty$ such that for all $T \geq T_0$,

$$\sum_{t=1}^T |A_{k,t}| \geq \frac{p q_k T B^-}{2}.$$

Since $|A_{k,t}| \leq |A_t| \leq p_t B^+$, we have $\bar{M}_k \leq (\sup_t p_t) B^+$. Using unbiased randomized rounding, it follows

$$\sup_t p_t \leq \sup_t \frac{\lceil p B_t \rceil}{B_t} \leq p + \frac{1}{B^-}, \quad \frac{\sup_t p_t}{p} \leq 1 + \frac{1}{p B^-}.$$

Therefore, on the same probability-one event, for all $T \geq T_0$,

$$\frac{\bar{M}_k \max\{\alpha, 1 - \alpha\}}{\sum_{t=1}^T |A_{k,t}|} \leq \frac{2(\sup_t p_t) B^+}{p q_k T B^-} \leq \frac{1}{T} \cdot \frac{2B^+}{B^- q_k} \left(1 + \frac{1}{p B^-}\right) \quad \text{a.s.},$$

Since both terms vanish as $T \rightarrow \infty$, we further have

$$|\text{CovErr}_{k,T}^A - \alpha| \rightarrow 0 \quad \text{a.s.}$$

This proves the proposition. $\square$

E.2 Proof of Theorem 3.3

Fix a class $k \in [K]$. We first prove the convergence of accumulated error on all cells. Combining (E.1) and (E.4), and noting that $\bar{M}_k \leq B^+$, gives

$$|\text{CovErr}_{k,T}^A - \alpha| \leq \frac{\max\{\alpha, 1 - \alpha\}(1 + \gamma_k B^+)}{\gamma_k \sum_{t=1}^T |A_{k,t}|} \rightarrow 0 \quad \text{a.s.}$$

By Assumption 3.1(ii) and Lemma F.2,

$$\sum_{t=1}^T |D_{k,t}| \rightarrow \infty, \quad \frac{\sum_{t=1}^T |A_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} \rightarrow p > 0 \quad \text{a.s.},$$

and hence $\sum_{t=1}^T |A_{k,t}| \rightarrow \infty$ a.s., which justifies the above convergence. Thus, by the triangle inequality, it suffices to show

$$|\text{CovErr}_{k,T}^D - \text{CovErr}_{k,T}^A| \rightarrow 0 \quad \text{a.s.} \quad (\text{E.5})$$

For convenience, write

$$Z_{k,t} := \sum_{x \in A_{k,t}} \text{err}_{k,t}(x),$$

and the centered sampling residual as

$$W_{k,t} := Z_{k,t} - p_t \sum_{x \in D_{k,t}} \text{err}_{k,t}(x).$$

By the definitions above,

$$\text{CovErr}_{k,T}^A - \text{CovErr}_{k,T}^D = \frac{\sum_{t=1}^T W_{k,t} + \sum_{t=1}^T p_t \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |A_{k,t}|} - \frac{\sum_{t=1}^T \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |D_{k,t}|}.$$

We then use $\sum_{t=1}^T p_t \sum_{x \in D_{k,t}} \text{err}_{k,t}(x) = p \sum_{t=1}^T \sum_{x \in D_{k,t}} \text{err}_{k,t}(x) + \sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)$. Factoring the full-batch average from the resulting middle two terms gives

$$\text{CovErr}_{k,T}^A - \text{CovErr}_{k,T}^D = \underbrace{\frac{\sum_{t=1}^T W_{k,t}}{\sum_{t=1}^T |A_{k,t}|}}_{\text{I}} + \underbrace{\text{CovErr}_{k,T}^D \left(\frac{p \sum_{t=1}^T |D_{k,t}|}{\sum_{t=1}^T |A_{k,t}|} - 1 \right)}_{\text{II}} + \underbrace{\frac{\sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |A_{k,t}|}}_{\text{III}}. \quad (\text{E.6})$$

We now treat the three terms in (E.6) separately.

Term I. We first verify the martingale conditions for $W_{k,t}$. Using the audit indicators $E_{t,i} = \mathbf{1}\{x_{t,i} \in A_t\}$, we can write

$$W_{k,t} = \sum_{i: y_{t,i}=k} \text{err}_{k,t}(x_{t,i})(E_{t,i} - p_t).$$

Conditional on $\mathcal{G}_t$, the error indicators and p_t are fixed. Uniform sampling without replacement gives

$$\mathbb{E}[E_{t,i} \mid \mathcal{G}_t] = p_t, \quad \text{Var}(E_{t,i} \mid \mathcal{G}_t) = p_t(1 - p_t).$$

Consequently,

$$\mathbb{E}[W_{k,t} \mid \mathcal{G}_t] = 0. \quad (\text{E.7})$$

Moreover, when $B_t \geq 2$, for distinct i, j ,

$$\text{Cov}(E_{t,i}, E_{t,j} \mid \mathcal{G}_t) = -\frac{p_t(1 - p_t)}{B_t - 1} \leq 0.$$

Since the error indicators are nonnegative and binary, for $B_t \geq 2$ the nonpositive covariance terms yield

$$\begin{aligned} \mathbb{E}[W_{k,t}^2 \mid \mathcal{G}_t] &= \text{Var}(W_{k,t} \mid \mathcal{G}_t) \\ &\leq p_t(1 - p_t) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)^2 \\ &= p_t(1 - p_t) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x) \\ &\leq p_t(1 - p_t) |D_{k,t}| \leq \frac{1}{4} |D_{k,t}|. \end{aligned} \quad (\text{E.8})$$

When $B_t = 1$, the audit set is deterministic conditional on $\mathcal{G}_t$, so $W_{k,t} = 0$ and (E.8) holds as well. It also holds when $|D_{k,t}| = 0$, since the defining sum is empty. Thus, (E.7) and (E.8) hold for all t . Equations (E.7) and (E.8) verify the hypotheses of Lemma F.1 with $b_t = |D_{k,t}|$ and $C = 1/4$. This sequence is nonnegative, $\mathcal{G}_t$ -measurable, and satisfies $\sum_{t=1}^T b_t \rightarrow \infty$ a.s. by Assumption 3.1(ii). Therefore, $\sum_{t=1}^T W_{k,t} / \sum_{t=1}^T |D_{k,t}| \rightarrow 0$ a.s. Lemma F.2 gives $\sum_{t=1}^T |D_{k,t}| / \sum_{t=1}^T |A_{k,t}| \rightarrow 1/p$ a.s., so

$$\text{Term I} = \frac{\sum_{t=1}^T W_{k,t}}{\sum_{t=1}^T |D_{k,t}|} \cdot \frac{\sum_{t=1}^T |D_{k,t}|}{\sum_{t=1}^T |A_{k,t}|} \rightarrow 0 \cdot \frac{1}{p} = 0 \quad \text{a.s.}$$

Term II. By Lemma F.2, $\sum_{t=1}^T |A_{k,t}| / \sum_{t=1}^T |D_{k,t}| \rightarrow p$ a.s., so

$$p \sum_{t=1}^T |D_{k,t}| / \sum_{t=1}^T |A_{k,t}| - 1 \rightarrow 0 \quad \text{a.s.}$$

Since $0 \leq \text{CovErr}_{k,T}^D \leq 1$, Term II converges to zero almost surely.

Term III. By Proposition B.1,

$$\frac{\sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |A_{k,t}|} \rightarrow 0 \quad \text{a.s.},$$

which is precisely this term.

Hence all three terms in (E.6) vanish a.s., proving (E.5) and therefore

$$|\text{CovErr}_{k,T}^D - \alpha| \rightarrow 0 \quad \text{a.s.}$$

For the unreviewed cells, since $U_{k,t} = D_{k,t} \setminus A_{k,t}$, writing the total error on $D_{k,t}$ as the sum of those on $A_{k,t}$ and $U_{k,t}$ and rearranging:

$$\text{CovErr}_{k,T}^U - \text{CovErr}_{k,T}^D = \frac{\sum_{t=1}^T |A_{k,t}|}{\sum_{t=1}^T |U_{k,t}|} (\text{CovErr}_{k,T}^D - \text{CovErr}_{k,T}^A). \quad (\text{E.9})$$

By Lemma F.2, $\sum_{t=1}^T |A_{k,t}| / \sum_{t=1}^T |U_{k,t}| \rightarrow p/(1-p)$, which is finite. Since (E.5) holds, identity (E.9) yields

$$|\text{CovErr}_{k,T}^U - \text{CovErr}_{k,T}^D| \rightarrow 0 \quad \text{a.s.},$$

and the triangle inequality completes the proof. $\square$

E.3 Proof of Corollary 3.4

Proof of Corollary 3.4. Fix a class $k \in [K]$. For any $x_{t,i} \in D_{k,t}$ with $I_{t,i} = 1$, Section 3.1 gives $\widehat{C}_t^+(x_{t,i}) = \{\hat{y}_{t,i}\}$. Since $\widehat{C}_t(x_{t,i}) \subseteq \widehat{C}_t^+(x_{t,i})$, an incorrect automated annotation implies that the original prediction set does not contain the true class:

$$\mathbb{1}\{\hat{y}_{t,i} \neq k\} \leq \mathbb{1}\{k \notin \widehat{C}_t(x_{t,i})\}.$$

Summing this inequality over the automatically annotated class- k cells and then over all batches gives

$$\text{AnnoErr}_{k,T}^D \leq \text{CovErr}_{k,T}^D, \quad \text{AnnoErr}_{k,T}^U \leq \text{CovErr}_{k,T}^U.$$

Theorem 3.3 shows that both quantities on the right-hand side converge almost surely to α . Therefore,

$$\limsup_{T \rightarrow \infty} \text{AnnoErr}_{k,T}^D \leq \alpha, \quad \limsup_{T \rightarrow \infty} \text{AnnoErr}_{k,T}^U \leq \alpha \quad \text{a.s.}$$

$\square$

E.4 Proof of Proposition 3.5

Fix a class $k \in [K]$. We retain the notation from the proof of Theorem 3.3. In particular, (E.6) decomposes $\text{CovErr}_{k,T}^A - \text{CovErr}_{k,T}^D$ into the three underbraced terms denoted there by I, II, and III. The proof below gives rate bounds for these three terms. By the triangle inequality,

$$|\text{CovErr}_{k,T}^D - \alpha| \leq |\text{CovErr}_{k,T}^D - \text{CovErr}_{k,T}^A| + |\text{CovErr}_{k,T}^A - \alpha|.$$

Proposition 3.2 gives

$$|\text{CovErr}_{k,T}^A - \alpha| = \mathcal{O}\left(\left(\sum_{t=1}^T |A_{k,t}|\right)^{-1}\right) = \mathcal{O}(T^{-1}) \quad \text{a.s.}$$

Moreover, the class-frequency condition in Proposition 3.2 together with Assumption 3.1(i) implies

$$\sum_{t=1}^T |D_{k,t}| \asymp T \quad \text{a.s.},$$

while Lemma F.2 gives

$$\sum_{t=1}^T |A_{k,t}| \sim p \sum_{t=1}^T |D_{k,t}| \quad \text{a.s.}$$

It is therefore enough to prove that each of I, II, and III in (E.6) is $\mathcal{O}_p(T^{-1/2})$.

Term I in (E.6). By (E.7), $(W_{k,t})$ is a martingale difference sequence, and (E.8) gives $\mathbb{E}[W_{k,t}^2 | \mathcal{G}_t] \leq \frac{1}{4}|D_{k,t}|$. Hence, by the orthogonality of martingale differences and $|D_{k,t}| \leq B^+$,

$$\mathbb{E}\left[\left(\sum_{t=1}^T W_{k,t}\right)^2\right] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[W_{k,t}^2 | \mathcal{G}_t]\right] \leq \frac{1}{4}B^+T.$$

Chebyshev's inequality therefore gives $\sum_{t=1}^T W_{k,t} = \mathcal{O}_p(\sqrt{T})$. Using $\sum_{t=1}^T |A_{k,t}| \asymp T$ a.s.,

$$\text{Term I} = \frac{\sum_{t=1}^T W_{k,t}}{\sum_{t=1}^T |A_{k,t}|} = \mathcal{O}_p(T^{-1/2}).$$

For the two rounding terms in the remaining parts of the decomposition, let $\mathcal{G}'_t$ denote the time- t sigma-field generated by all random quantities determined after D_t is realized but before the randomized rounding variable that determines p_t is drawn. Thus $\mathcal{G}'_t$ precedes the pre-review sigma-field $\mathcal{G}_t$, in which p_t is already realized and the annotation subset A_t has not yet been sampled. Under the unbiased randomized rounding rule,

$$p_t - p = \frac{\xi_t - q_t}{B_t}, \quad q_t = pB_t - \lfloor pB_t \rfloor, \quad \xi_t | \mathcal{G}'_t \sim \text{Bernoulli}(q_t),$$

and therefore

$$\mathbb{E}[p_t - p | \mathcal{G}'_t] = 0, \quad \mathbb{E}[(p_t - p)^2 | \mathcal{G}'_t] = \frac{q_t(1 - q_t)}{B_t^2} \leq \frac{1}{4B_t^2}. \quad (\text{E.11})$$

Since $|D_{k,t}|$ and $\sum_{x \in D_{k,t}} \text{err}_{k,t}(x)$ are $\mathcal{G}'_t$ -measurable and $0 \leq \sum_{x \in D_{k,t}} \text{err}_{k,t}(x) \leq |D_{k,t}| \leq B_t$, (E.11) implies

$$\begin{aligned} \mathbb{E}[(p_t - p)|D_{k,t}| | \mathcal{G}'_t] &= 0, \\ \mathbb{E}[(p_t - p)^2 |D_{k,t}|^2 | \mathcal{G}'_t] &\leq \frac{1}{4} \mathbb{1}\{|D_{k,t}| > 0\}, \end{aligned} \quad (\text{E.12})$$

and

$$\begin{aligned} \mathbb{E}\left[(p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x) \middle| \mathcal{G}'_t\right] &= 0, \\ \mathbb{E}\left[\left((p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)\right)^2 \middle| \mathcal{G}'_t\right] &\leq \frac{1}{4} \mathbb{1}\{|D_{k,t}| > 0\}. \end{aligned} \quad (\text{E.13})$$

Term II in (E.6). Write the numerator of the bracketed factor as

$$p \sum_{t=1}^T |D_{k,t}| - \sum_{t=1}^T |A_{k,t}| = - \sum_{t=1}^T Y_{k,t} - \sum_{t=1}^T (p_t - p)|D_{k,t}|,$$

where $Y_{k,t} := |A_{k,t}| - p_t |D_{k,t}|$. The proof of Lemma F.2 establishes (F.1) and (F.2); hence $(Y_{k,t})$ is a martingale difference sequence and $\mathbb{E}[Y_{k,t}^2 | \mathcal{G}_t] \leq C |D_{k,t}|$. The same second-moment calculation as in Term I gives $\sum_{t=1}^T Y_{k,t} = \mathcal{O}_p(\sqrt{T})$. By (E.12), $((p_t - p)|D_{k,t}|)$ is also a martingale difference sequence with respect to the rounding filtration, and

$$\mathbb{E} \left[\left(\sum_{t=1}^T (p_t - p) |D_{k,t}| \right)^2 \right] \leq \frac{T}{4},$$

so $\sum_{t=1}^T (p_t - p) |D_{k,t}| = \mathcal{O}_p(\sqrt{T})$. Therefore, the bracketed factor in Term II is $\mathcal{O}_p(T^{-1/2})$ after dividing by $\sum_{t=1}^T |A_{k,t}| \asymp T$, and since $0 \leq \text{CovErr}_{k,T}^D \leq 1$, we conclude

$$\text{Term II} = \mathcal{O}_p(T^{-1/2}).$$

Term III in (E.6). By (E.13), $((p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x))$ is also a martingale difference sequence with respect to the rounding filtration. Its conditional second moments are bounded by 1/4, so the same fixed- T second-moment argument gives

$$\sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x) = \mathcal{O}_p(\sqrt{T}).$$

Dividing by $\sum_{t=1}^T |A_{k,t}| \asymp T$ therefore yields Term III = $\mathcal{O}_p(T^{-1/2})$. Combining the three terms via (E.6),

$$|\text{CovErr}_{k,T}^D - \text{CovErr}_{k,T}^A| = \mathcal{O}_p(T^{-1/2}).$$

Adding the $\mathcal{O}(T^{-1})$ contribution from Proposition 3.2 and using $\sum_{t=1}^T |D_{k,t}| \asymp T$, the triangle inequality gives

$$|\text{CovErr}_{k,T}^D - \alpha| = \mathcal{O}_p \left(\left(\sum_{t=1}^T |D_{k,t}| \right)^{-1/2} \right) = \mathcal{O}_p(T^{-1/2}).$$

For the unreviewed rate, identity (E.9) and Lemma F.2 give

$$|\text{CovErr}_{k,T}^U - \text{CovErr}_{k,T}^D| = \frac{\sum_{t=1}^T |A_{k,t}|}{\sum_{t=1}^T |U_{k,t}|} |\text{CovErr}_{k,T}^D - \text{CovErr}_{k,T}^A| = \mathcal{O}_p(1) \cdot \mathcal{O}_p(T^{-1/2}),$$

and the triangle inequality gives

$$|\text{CovErr}_{k,T}^U - \alpha| = \mathcal{O}_p \left(\left(\sum_{t=1}^T |D_{k,t}| \right)^{-1/2} \right) = \mathcal{O}_p(T^{-1/2}).$$

□

E.5 Proof of Proposition B.1

Fix a class $k \in [K]$. The rounding moment bounds in (E.12) and (E.13) show that $((p_t - p)|D_{k,t}|)$ and $((p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x))$ are martingale difference sequences with conditional second moments bounded by a constant multiple of $|D_{k,t}|$. Applying Lemma F.1 to these two sequences gives

$$\frac{\sum_{t=1}^T (p_t - p) |D_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} \rightarrow 0, \quad \frac{\sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |D_{k,t}|} \rightarrow 0 \quad \text{a.s.}$$

It remains only to replace the denominator in the second convergence by $\sum_{t=1}^T |A_{k,t}|$. Conditional on $\mathcal{G}_t$, where p_t has already been realized and the audit set A_t has not yet been sampled, $|A_{k,t}| = \sum_{i: y_{t,i} = k} E_{t,i}$ is

hypergeometric with mean $p_t|D_{k,t}|$ and variance bounded by $C|D_{k,t}|$ for a deterministic constant $C < \infty$. Therefore Lemma F.1, together with the first convergence above, gives

$$\frac{\sum_{t=1}^T |A_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} = \frac{\sum_{t=1}^T \{|A_{k,t}| - p_t|D_{k,t}|\}}{\sum_{t=1}^T |D_{k,t}|} + \frac{\sum_{t=1}^T (p_t - p)|D_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} + p \rightarrow p \quad \text{a.s.}$$

Since $p > 0$, it follows that

$$\frac{\sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x)}{\sum_{t=1}^T |A_{k,t}|} = \frac{\sum_{t=1}^T (p_t - p) \sum_{x \in D_{k,t}} \text{err}_{k,t}(x) / \sum_{t=1}^T |D_{k,t}|}{\sum_{t=1}^T |A_{k,t}| / \sum_{t=1}^T |D_{k,t}|} \rightarrow 0 \quad \text{a.s.}$$

Together with the first convergence, this establishes the two asymptotic relations claimed in the proposition. Since K is finite, the conclusions hold simultaneously for all classes after intersecting the corresponding probability-one events. $\square$

E.6 Proofs for Proposition B.5 and Corollary B.6

Proof of Proposition B.5. Fix a class $k \in [K]$ and write

$$x_{k,t} := \alpha_{k,t} - \alpha.$$

Recall that $\bar{\alpha}_{k,t}$ denotes $\alpha_{k,t}$ clipped to the interval $[0, 1]$. Hence, $x_{k,t}$ and $\bar{\alpha}_{k,t} - \alpha$ have the same sign and $|\bar{\alpha}_{k,t} - \alpha| \leq |x_{k,t}|$. Since $0 \leq \eta_{k,t} \leq 1$, we have

$$|x_{k,t} - \eta_{k,t}(\bar{\alpha}_{k,t} - \alpha)| = |x_{k,t}| - \eta_{k,t}|\bar{\alpha}_{k,t} - \alpha|.$$

The update rule and the triangle inequality therefore give

$$\eta_{k,t}|\bar{\alpha}_{k,t} - \alpha| \leq |x_{k,t}| - |x_{k,t+1}| + \eta_{k,t}(|\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| + |\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D|). \quad (\text{E.14})$$

Define

$$Z_{k,t} := |A_{k,t}| |\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D|, \quad Y_{k,t} := (|A_{k,t}| - p_t|D_{k,t}|) (|\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| - |\bar{\alpha}_{k,t} - \alpha|).$$

Summing (E.14), using $\eta_{k,t} = \gamma_k|A_{k,t}|$, and dividing by $\gamma_k > 0$ gives

$$\sum_{t=1}^T |A_{k,t}| |\bar{\alpha}_{k,t} - \alpha| \leq \frac{|x_{k,1}| - |x_{k,T+1}|}{\gamma_k} + \sum_{t=1}^T |A_{k,t}| |\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| + \sum_{t=1}^T Z_{k,t}.$$

The initialization $\alpha_{k,1} = \alpha$ implies $x_{k,1} = 0$, so the telescoping term is nonpositive and can be discarded. Moreover,

$$|A_{k,t}| (|\bar{\alpha}_{k,t} - \alpha| - |\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D|) = p_t|D_{k,t}| (|\bar{\alpha}_{k,t} - \alpha| - |\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D|) - Y_{k,t}.$$

Thus, adding and subtracting $p_t|D_{k,t}|$ in the preceding sum gives

$$\sum_{t=1}^T p_t|D_{k,t}| |\bar{\alpha}_{k,t} - \alpha| \leq \sum_{t=1}^T p_t|D_{k,t}| |\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| + \sum_{t=1}^T Z_{k,t} + \sum_{t=1}^T Y_{k,t}.$$

Using $p_t|D_{k,t}| \geq p^- d_k$ then yields

$$p^- d_k \sum_{t=1}^T |\bar{\alpha}_{k,t} - \alpha| \leq \sum_{t=1}^T p_t|D_{k,t}| |\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| + \sum_{t=1}^T Z_{k,t} + \sum_{t=1}^T Y_{k,t}. \quad (\text{E.15})$$

Indeed, by (B.2), $|\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| \leq \Delta_{k,t}$, while $p_t |D_{k,t}| \leq p^+ M_k d_k$. Hence,

$$\sum_{t=1}^T p_t |D_{k,t}| |\bar{\alpha}_{k,t} - \text{BatErr}_{k,t}^D| \leq p^+ M_k d_k \sum_{t=1}^T \Delta_{k,t}.$$

We next control the remaining two sums. Conditional on $\mathcal{G}_t$, the reviewed cells are sampled uniformly without replacement from D_t . The term $Z_{k,t}$ is the absolute centered sample sum of the class- k errors, and its conditional second moment is

$$\begin{aligned} \mathbb{E}[Z_{k,t}^2 | \mathcal{G}_t] &= p_t(1-p_t) \frac{B_t}{B_t-1} |D_{k,t}| \text{BatErr}_{k,t}^D (1 - \text{BatErr}_{k,t}^D) \\ &\leq \frac{1}{4} \frac{B^-}{B^- - 1} p^+(1-p^-) M_k d_k. \end{aligned}$$

Consequently, the conditional Cauchy–Schwarz inequality gives

$$\mathbb{E}[Z_{k,t} | \mathcal{G}_t] \leq \frac{1}{2} \sqrt{\frac{B^-}{B^- - 1} p^+(1-p^-) M_k d_k}.$$

Moreover, $0 \leq Z_{k,t} \leq |D_{k,t}|/4 \leq M_k d_k/4$. Since $B^- \geq 2$, the conditional variance of $Z_{k,t} - \mathbb{E}[Z_{k,t} | \mathcal{G}_t]$ is at most $p^+(1-p^-) M_k d_k/2$.

For $Y_{k,t}$, uniform sampling gives $\mathbb{E}[|A_{k,t}| | \mathcal{G}_t] = p_t |D_{k,t}|$. Since the second factor in the definition of $Y_{k,t}$ is determined before review, $(Y_{k,t})_{t=1}^T$ is a martingale difference sequence. This factor has absolute value at most one, so $|Y_{k,t}| \leq M_k d_k$, and its conditional variance is bounded by

$$\text{Var}(|A_{k,t}| | \mathcal{G}_t) = p_t(1-p_t) \frac{B_t - |D_{k,t}|}{B_t - 1} |D_{k,t}| \leq p^+(1-p^-) M_k d_k.$$

Applying the rescaled form of Lemma F.4 to $(Z_{k,t} - \mathbb{E}[Z_{k,t} | \mathcal{G}_t])_{t=1}^T$ and $(Y_{k,t})_{t=1}^T$, and taking a union bound, shows that with probability at least $1 - \epsilon$,

$$\begin{aligned} \sum_{t=1}^T Z_{k,t} &\leq \frac{T}{2} \sqrt{\frac{B^-}{B^- - 1} p^+(1-p^-) M_k d_k} + \sqrt{p^+(1-p^-) M_k d_k T \log \frac{2}{\epsilon}} + \frac{M_k d_k}{12} \log \frac{2}{\epsilon}, \\ \sum_{t=1}^T Y_{k,t} &\leq \sqrt{2p^+(1-p^-) M_k d_k T \log \frac{2}{\epsilon}} + \frac{M_k d_k}{3} \log \frac{2}{\epsilon}. \end{aligned} \tag{E.17}$$

The two square-root deviation terms in (E.17) have total coefficient $1 + \sqrt{2} < 5/2$, and the two logarithmic terms sum to $(5M_k d_k/12) \log(2/\epsilon)$. Substituting (E.17) and the calibration bound into (E.15), and then dividing by $p^- d_k$, gives

$$\sum_{t=1}^T |\bar{\alpha}_{k,t} - \alpha| \leq \frac{p^+}{p^-} M_k \sum_{t=1}^T \Delta_{k,t} + \mathcal{R}_{k,T}(\epsilon).$$

Finally, the triangle inequality and (B.2) imply

$$|\text{BatErr}_{k,t}^D - \alpha| \leq |\text{BatErr}_{k,t}^D - \bar{\alpha}_{k,t}| + |\bar{\alpha}_{k,t} - \alpha| \leq \Delta_{k,t} + |\bar{\alpha}_{k,t} - \alpha|.$$

Summing this inequality and applying the preceding bound yields

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \alpha| \leq \left(1 + \frac{p^+}{p^-} M_k\right) \sum_{t=1}^T \Delta_{k,t} + \mathcal{R}_{k,T}(\epsilon),$$

which proves the proposition. $\square$

Proof of Corollary B.6. Under randomized rounding, take $B^- := \min_{1 \leq t \leq T} B_t$. The bounded class-size condition implies $B^- \geq d_k$, so $B^- \rightarrow \infty$ as $d_k \rightarrow \infty$. Hence $p^- = p - 1/B^-$ and $p^+ = p + 1/B^-$ satisfy

$$\frac{p^+}{p^-} = 1 + \mathcal{O}\left(\frac{1}{d_k}\right), \quad \frac{1}{p^-} = \mathcal{O}(1).$$

Also, $\Delta_{k,t} \leq 2$, and therefore

$$\left(\frac{p^+}{p^-} - 1\right) M_k \sum_{t=1}^T \Delta_{k,t} = \mathcal{O}\left(\frac{T}{d_k}\right).$$

For fixed ϵ , the explicit expression for $\mathcal{R}_{k,T}(\epsilon)$ gives

$$\mathcal{R}_{k,T}(\epsilon) = \mathcal{O}\left(\frac{T}{\sqrt{d_k}}\right) + o_T(T).$$

Substituting these two bounds into Proposition B.5, and absorbing $\mathcal{O}(T/d_k)$ into $\mathcal{O}(T/\sqrt{d_k})$, proves the corollary. $\square$

E.7 Proofs for Proposition B.8 and Corollary B.9

Proof of Proposition B.8. Write $x_{k,t} := \alpha_{k,t} - \alpha$. Dividing (E.14) by $\eta_{k,t} > 0$, and then using (B.2) and the triangle inequality, gives

$$|\text{BatErr}_{k,t}^D - \alpha| \leq 2\Delta_{k,t} + |\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| + \frac{|x_{k,t}| - |x_{k,t+1}|}{\eta_{k,t}}. \quad (\text{E.18})$$

By summation by parts, $x_{k,1} = 0$, and the nondecreasing effective step sizes,

$$\sum_{t=1}^T \frac{|x_{k,t}| - |x_{k,t+1}|}{\eta_{k,t}} \leq 0. \quad (\text{E.19})$$

Summing (E.18) and applying (E.19) yields

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^D - \alpha| \leq 2 \sum_{t=1}^T \Delta_{k,t} + \sum_{t=1}^T |\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D|.$$

By Lemma F.3, the second term on the right-hand side is at most $S_{k,T}(\epsilon)$ with probability at least $1 - \epsilon$, which proves the proposition. $\square$

Proof of Corollary B.9. By assumption, $|D_{k,t}| \geq d_k$ for every $t \leq T$. For fixed $\epsilon \in (0, 1)$, the definition of $S_{k,T}(\epsilon)$ therefore gives

$$\begin{aligned} S_{k,T}(\epsilon) &\leq T \left\{ \frac{1}{\sqrt{2p^- d_k}} + \exp\left(-\frac{p^- d_k}{12}\right) \right\} + \sqrt{2T \log \frac{1}{\epsilon}} + \frac{1}{3} \log \frac{1}{\epsilon} \\ &= \mathcal{O}\left(\frac{T}{\sqrt{d_k}}\right) + o_T(T), \end{aligned}$$

where p^- is fixed and positive. Combining this bound with Proposition B.8 proves the corollary. $\square$

F Auxiliary lemmas

Lemma F.1 (Weighted strong law for martingale differences). *Let $(\mathcal{G}_t)_{t \geq 1}$ be a filtration, and let Y_t be $\mathcal{G}_{t+1}$ -measurable with*

$$\mathbb{E}[Y_t \mid \mathcal{G}_t] = 0.$$

Suppose there exists a nonnegative $\mathcal{G}_t$ -measurable sequence $(b_t)_{t \geq 1}$ and a constant $C < \infty$ such that

$$\mathbb{E}[Y_t^2 \mid \mathcal{G}_t] \leq C b_t \quad \text{a.s.}$$

If $\sum_{t=1}^T b_t \rightarrow \infty$ almost surely, then

$$\frac{\sum_{t=1}^T Y_t}{\sum_{t=1}^T b_t} \rightarrow 0 \quad \text{a.s.}$$

Proof. Set $S_0 = 1$ and $S_t = 1 + \sum_{i=1}^t b_i$. Then S_t is positive, nondecreasing, and $\mathcal{G}_t$ -measurable. For every $t \geq 1$,

$$\frac{b_t}{S_t^2} \leq \frac{b_t}{S_{t-1}S_t} = \frac{1}{S_{t-1}} - \frac{1}{S_t},$$

and consequently,

$$\sum_{t=1}^{\infty} \frac{b_t}{S_t^2} \leq 1.$$

Hence,

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[Y_t^2 \mid \mathcal{G}_t]}{S_t^2} \leq C \sum_{t=1}^{\infty} \frac{b_t}{S_t^2} \leq C < \infty \quad \text{a.s.}$$

Since $S_T \rightarrow \infty$ almost surely, the case $p = 2$ of Theorem 2.18 in [Hall et al. \(1980\)](#) yields

$$\frac{\sum_{t=1}^T Y_t}{S_T} \rightarrow 0 \quad \text{a.s.,}$$

Finally, $S_T / \sum_{t=1}^T b_t \rightarrow 1$ almost surely, which gives the stated conclusion. $\square$

Lemma F.2 (Class-wise reviewed fraction). *Under Assumption 3.1, for every class $k \in [K]$, the fraction of reviewed cells converges to the target fraction p with probability 1, i.e.,*

$$\frac{\sum_{t=1}^T |A_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} \rightarrow p \quad \text{a.s.}$$

Proof. Fix a class $k \in [K]$ and define

$$Y_{k,t} := |A_{k,t}| - p_t |D_{k,t}|.$$

It suffices to show that

$$\frac{\sum_{t=1}^T Y_{k,t}}{\sum_{t=1}^T |D_{k,t}|} \rightarrow 0 \quad \text{a.s.}$$

Indeed, by Proposition B.1,

$$\frac{\sum_{t=1}^T (p_t - p) |D_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} \rightarrow 0 \quad \text{a.s.,}$$

and

$$\frac{\sum_{t=1}^T |A_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} = \frac{\sum_{t=1}^T Y_{k,t}}{\sum_{t=1}^T |D_{k,t}|} + \frac{\sum_{t=1}^T (p_t - p) |D_{k,t}|}{\sum_{t=1}^T |D_{k,t}|} + p.$$

Now we verify the conditions of Lemma F.1. Recall that $\mathcal{G}_t$ denotes the pre-review sigma-field after the prediction sets for D_t have been formed and before the audit set A_t is sampled. Conditional on $\mathcal{G}_t$, the incoming batch D_t , the class count $|D_{k,t}|$, and the reviewed fraction p_t are fixed. By Assumption 3.1(iii), $|A_{k,t}|$ follows a hypergeometric distribution with population size B_t , number of class- k cells $|D_{k,t}|$, and draw size $p_t B_t$. Therefore

$$\mathbb{E}(|A_{k,t}| \mid \mathcal{G}_t) = \frac{|A_t|}{B_t} |D_{k,t}| = p_t |D_{k,t}|,$$

and consequently

$$\mathbb{E}(Y_{k,t} \mid \mathcal{G}_t) = 0. \quad (\text{F.1})$$

If $B_t = 1$, the conditional distribution of $|A_{k,t}|$ is degenerate and its variance is zero. If $B_t \geq 2$, the hypergeometric variance formula gives

$$\text{Var}(|A_{k,t}| \mid \mathcal{G}_t) = |A_t| \frac{|D_{k,t}|}{B_t} \left(1 - \frac{|D_{k,t}|}{B_t}\right) \frac{B_t - |A_t|}{B_t - 1}.$$

For $B_t \geq 2$, the factors after $|A_t|$ are bounded by one and $|A_t| \leq B_t$, whereas for $B_t = 1$ the variance is zero. Thus there exists a constant $C < \infty$ such that

$$\mathbb{E}(Y_{k,t}^2 \mid \mathcal{G}_t) = \text{Var}(|A_{k,t}| \mid \mathcal{G}_t) \leq C|D_{k,t}|. \quad (\text{F.2})$$

By (F.1) and (F.2), taking $b_t = |D_{k,t}|$, which is nonnegative and $\mathcal{G}_t$ -measurable, in Lemma F.1 yields

$$\frac{\sum_{t=1}^T Y_{k,t}}{\sum_{t=1}^T |D_{k,t}|} \rightarrow 0 \quad \text{a.s.},$$

where $\sum_{t=1}^T |D_{k,t}| \rightarrow \infty$ follows directly from Assumption 3.1(ii). Combining this with the preceding decomposition proves the result. $\square$

Lemma F.3 (Cumulative sampling fluctuation). *Fix a class $k \in [K]$. Suppose that Assumptions 3.1 and B.2 hold. Then, for any horizon $T \geq 1$ and any $\zeta \in (0, 1)$, with probability at least $1 - \zeta$,*

$$\sum_{t=1}^T |\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| \leq \sum_{t=1}^T \mu_{k,t} + \sqrt{2T \log \frac{1}{\zeta}} + \frac{1}{3} \log \frac{1}{\zeta} = S_{k,T}(\zeta),$$

where $\mu_{k,t}$ is defined in (B.3).

Proof. Fix t and define

$$\mathcal{H}_t := \mathcal{G}_t \vee \sigma(|A_{k,t}|),$$

where $\mathcal{G}_t$ is the previously defined pre-review sigma-field and $|A_{k,t}|$ denotes the number of reviewed class- k cells. Conditional on $\mathcal{H}_t$ and on the event $|A_{k,t}| = m > 0$, the number of erroneous reviewed class- k cells $\sum_{x \in A_{k,t}} \text{err}_{k,t}(x)$ has a hypergeometric distribution. The population size is $|D_{k,t}|$, the success count is $\sum_{x \in D_{k,t}} \text{err}_{k,t}(x)$, and the draw size is m . Since

$$\text{BatErr}_{k,t}^A = \frac{1}{m} \sum_{x \in A_{k,t}} \text{err}_{k,t}(x),$$

the conditional mean is $\text{BatErr}_{k,t}^D$. If $|D_{k,t}| = 1$, then the conditional distribution is degenerate and the variance is zero. If $|D_{k,t}| \geq 2$, the hypergeometric variance formula gives

$$\begin{aligned} \text{Var}(\text{BatErr}_{k,t}^A \mid \mathcal{H}_t, |A_{k,t}| = m) &= \frac{1}{m^2} m \text{BatErr}_{k,t}^D (1 - \text{BatErr}_{k,t}^D) \frac{|D_{k,t}| - m}{|D_{k,t}| - 1} \\ &\leq \frac{\text{BatErr}_{k,t}^D (1 - \text{BatErr}_{k,t}^D)}{m} \leq \frac{1}{4m}, \end{aligned}$$

where the last inequality uses $0 \leq \text{BatErr}_{k,t}^D \leq 1$. Thus the same upper bound holds for all $|D_{k,t}| \geq 1$. Together with the convention $\text{BatErr}_{k,t}^A = \text{BatErr}_{k,t}^D$ when $|A_{k,t}| = 0$, this gives

$$\mathbb{E}[(\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D)^2 \mid \mathcal{H}_t] \leq \frac{\mathbb{1}\{|A_{k,t}| > 0\}}{4|A_{k,t}|}. \quad (\text{F.3})$$

Since $\mathcal{G}_t \subseteq \mathcal{H}_t$, by the tower property and conditional Jensen's inequality, (F.3) gives

$$\begin{aligned} \mathbb{E}[|\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| \mid \mathcal{G}_t] &= \mathbb{E}[\mathbb{E}[|\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| \mid \mathcal{H}_t] \mid \mathcal{G}_t] \\ &\leq \mathbb{E}\left[\mathbb{E}\left\{\left[(\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D)^2 \mid \mathcal{H}_t\right]\right\}^{1/2} \mid \mathcal{G}_t\right] \\ &\leq \mathbb{E}\left[\frac{\mathbb{1}\{|A_{k,t}| > 0\}}{2\sqrt{|A_{k,t}|}} \mid \mathcal{G}_t\right]. \end{aligned} \quad (\text{F.4})$$

Define the high-count event

$$\mathcal{E}_{k,t}^{\text{high}} = \{|A_{k,t}| \geq p^- |D_{k,t}|/2\}.$$

Conditional on $\mathcal{G}_t$, the count $|A_{k,t}|$ follows a hypergeometric distribution with mean $p_t |D_{k,t}|$. Since $p_t \geq p^-$, the multiplicative Chernoff bound for sampling without replacement yields

$$\begin{aligned} \mathbb{P}((\mathcal{E}_{k,t}^{\text{high}})^c \mid \mathcal{G}_t) &= \mathbb{P}\left(|A_{k,t}| < \frac{1}{2}p^- |D_{k,t}| \mid \mathcal{G}_t\right) \\ &\leq \mathbb{P}\left(|A_{k,t}| < \frac{1}{2}p_t |D_{k,t}| \mid \mathcal{G}_t\right) \\ &\leq 2 \exp\left(-\frac{p_t |D_{k,t}|}{12}\right) \leq 2 \exp\left(-\frac{p^- |D_{k,t}|}{12}\right). \end{aligned} \quad (\text{F.5})$$

Now define the centered fluctuation

$$V_{k,t} := |\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| - \mathbb{E}[|\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| \mid \mathcal{G}_t].$$

Then $V_{k,t} \in \mathcal{G}_{t+1}$, $\mathbb{E}(V_{k,t} \mid \mathcal{G}_t) = 0$, and $|V_{k,t}| \leq 1$. It remains to bound the predictable part of $|\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D|$ and the conditional second moment of $V_{k,t}$. On the high-count event $\mathcal{E}_{k,t}^{\text{high}}$, the two factors $\mathbb{1}\{|A_{k,t}| > 0\}/(2\sqrt{|A_{k,t}|})$ and $\mathbb{1}\{|A_{k,t}| > 0\}/(4|A_{k,t}|)$ are bounded by $(2p^- |D_{k,t}|)^{-1/2}$ and $(2p^- |D_{k,t}|)^{-1}$, respectively. On $(\mathcal{E}_{k,t}^{\text{high}})^c$, they are bounded by $1/2$ and $1/4$. Combining these bounds with (F.4) and (F.5) gives

$$\begin{aligned} \mathbb{E}[|\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| \mid \mathcal{G}_t] &\leq \frac{1}{\sqrt{2p^- |D_{k,t}|}} + \frac{1}{2} \mathbb{P}((\mathcal{E}_{k,t}^{\text{high}})^c \mid \mathcal{G}_t) \\ &\leq \frac{1}{\sqrt{2p^- |D_{k,t}|}} + \exp\left(-\frac{p^- |D_{k,t}|}{12}\right) = \mu_{k,t} \end{aligned} \quad (\text{F.6})$$

For the centered fluctuation, since $|V_{k,t}| \leq 1$, the predictable quadratic variation satisfies

$$\sum_{t=1}^T \mathbb{E}[V_{k,t}^2 \mid \mathcal{G}_t] \leq T \quad \text{a.s.}$$

Applying Lemma F.4 to the martingale difference sequence $(V_{k,t})_{t=1}^T$ therefore gives, with probability at least $1 - \zeta$,

$$\sum_{t=1}^T V_{k,t} \leq \sqrt{2T \log \frac{1}{\zeta}} + \frac{1}{3} \log \frac{1}{\zeta}.$$

By the definition of $V_{k,t}$ and (F.6),

$$\begin{aligned} \sum_{t=1}^T |\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| &= \sum_{t=1}^T \mathbb{E}[|\text{BatErr}_{k,t}^A - \text{BatErr}_{k,t}^D| \mid \mathcal{G}_t] + \sum_{t=1}^T V_{k,t} \\ &\leq \sum_{t=1}^T \mu_{k,t} + \sqrt{2T \log \frac{1}{\zeta}} + \frac{1}{3} \log \frac{1}{\zeta} = S_{k,T}(\zeta), \end{aligned}$$

which proves the lemma. $\square$

Lemma F.4 (Corollary of Freedman's inequality). *Let $(\mathcal{G}_t)_{t \geq 0}$ be a filtration. Suppose V_t is $\mathcal{G}_{t+1}$ -measurable with $\mathbb{E}[V_t | \mathcal{G}_t] = 0$ and $|V_t| \leq 1$ a.s. Assume further that*

$$\sum_{t=1}^T \mathbb{E}(V_t^2 | \mathcal{G}_t) \leq B \quad \text{a.s.}$$

for some constant $B > 0$. Then for any $\zeta \in (0, 1)$, with probability at least $1 - \zeta$,

$$\sum_{t=1}^T V_t \leq \sqrt{2B \log \frac{1}{\zeta}} + \frac{1}{3} \log \frac{1}{\zeta}.$$

Proof. Theorem 4.1 of [Freedman \(1975\)](#) states: for a martingale difference sequence with increments bounded above by one, if $S_n = \sum_{t=1}^n V_t$ denotes its partial sum and $W_n = \sum_{t=1}^n \mathbb{E}(V_t^2 | \mathcal{G}_t)$ its predictable quadratic variation, then for all $a, b > 0$,

$$\mathbb{P}(\exists n : S_n \geq a \text{ and } W_n \leq b) \leq \left(\frac{b}{a+b} \right)^{a+b} e^a.$$

Since $W_T \leq B$ a.s. by assumption, the event $\{S_T \geq a\}$ implies the event in the display (at $n = T$). We therefore apply the theorem with $b = B$ and

$$a = \sqrt{2B \log \frac{1}{\zeta}} + \frac{1}{3} \log \frac{1}{\zeta}.$$

Set $L = \log(1/\zeta)$. Since $b = B$, the choice of a gives $a = \sqrt{2bL} + L/3$. We compute

$$\begin{aligned} \left(\frac{b}{a+b} \right)^{a+b} e^a &= \exp \left\{ -(a+b) \log \left(1 + \frac{a}{b} \right) + a \right\} \\ &= \exp \left\{ -b \left[\left(1 + \frac{a}{b} \right) \log \left(1 + \frac{a}{b} \right) - \frac{a}{b} \right] \right\}. \end{aligned}$$

Setting $r = L/b$, we have $a/b = \sqrt{2r} + r/3$. We first establish

$$\left(1 + \sqrt{2r} + \frac{r}{3} \right) \log \left(1 + \sqrt{2r} + \frac{r}{3} \right) - \sqrt{2r} - \frac{r}{3} \geq r, \quad r \geq 0, \quad (\text{F.7})$$

To prove (F.7), define

$$F(x) = \left(1 + x + \frac{x^2}{6} \right) \log \left(1 + x + \frac{x^2}{6} \right) - x - \frac{2x^2}{3}, \quad x \geq 0.$$

Direct differentiation gives $F(0) = F'(0) = 0$ and

$$F''(x) = \frac{1}{3} \left[\log \left(1 + x + \frac{x^2}{6} \right) - 1 + \frac{1}{1 + x + x^2/6} \right] \geq 0,$$

where the inequality follows from $\log z \geq 1 - z^{-1}$ for $z \geq 1$. Thus $F(x) \geq 0$ for all $x \geq 0$. Taking $x = \sqrt{2r}$ proves (F.7). Applying (F.7) with $r = L/b$ yields

$$b \left[\left(1 + \frac{a}{b} \right) \log \left(1 + \frac{a}{b} \right) - \frac{a}{b} \right] \geq br = L.$$

Consequently,

$$\left(\frac{b}{a+b} \right)^{a+b} e^a \leq e^{-L} = \zeta.$$

Therefore $\mathbb{P}(S_T \geq a) \leq \zeta$, which proves the stated bound. $\square$