CANDIDATE LABEL RECONSTRUCTION WITH CARDINALITY PRESERVATION FOR MULTI-LABEL ECG

Anonymous authors
Paper under double-blind review

ABSTRACT

ECG recordings often contain multiple co-occurring diagnoses, while clinical annotations may also be ambiguous or incomplete. Partial Label Learning (PLL) provides a natural framework for this setting by learning from candidate label sets that contain the underlying true labels. However, existing PLL disambiguation strategies are not well-suited to multi-label ECG diagnosis for two reasons: uniform disambiguation can unnecessarily alter reliable candidate sets, while confidence-based approaches may suppress valid but less confident co-occurring diagnoses. We propose a novel disambiguation method, ReCal, which learns sample-specific reliability to determine how strongly each candidate set should be disambiguated and estimates label cardinality to preserve multiple valid diagnoses. Using several controlled ambiguity settings on PTB-XL, Chapman, and AMIGOS datasets, ReCal consistently improves both label disambiguation and downstream classification over existing methods.

1 INTRODUCTION

Electrocardiogram (ECG) analysis is a fundamental component of modern cardiovascular care, with non-invasive recordings of the heart's electrical activity. Recent advances in deep learning have improved the accuracy of automated ECG interpretation systems, achieving expert-level performance on diagnostic tasks (Johnson et al., 2025; Hannun et al., 2019). However, ECG diagnosis can remain ambiguous in clinical practice, as a recording may exhibit overlapping characteristics of multiple cardiac conditions and different clinicians may assign different diagnoses to the same recording (Rahmani et al., 2025b). Such ambiguity makes it challenging to associate some ECGs with a definitive label set, as some recordings may be associated with a set of plausible diagnostic labels (Ribeiro et al., 2020). To learn from such samples, we need to distinguish plausible labels from the underlying diagnoses while preserving the multi-label structure of the recording.

Partial Label Learning (PLL) frameworks enable learning from samples with a plausible set of labels, referred to as a *candidate label set*, including true and false-positive labels. While some methods treat all candidate labels as valid (Durand et al., 2019; Wu et al., 2022), false-positive candidates can mislead learning. Identification-based PLL methods therefore iteratively *disambiguate* candidate sets, typically using model predictions to assign confidence to likely correct labels (Wang et al., 2022a; Zhang et al., 2021; Wang et al., 2022b). Beyond this confidence-based refinement, several approaches incorporate additional structural information into disambiguation, such as instance- and label-level graph similarities (Wang et al., 2023), probabilistic graphical models that treat the ground-truth label set as a latent variable (Hang & Zhang, 2023), and label propagation based on nearest-neighbour reconstruction (Zhang & Fang, 2020). Despite these efforts, existing studies have primarily focused on improving the overall predictive performance of PLL models, while the behavior of disambiguation during training remains under-explored. In particular, existing disambiguation mechanisms generally refine candidate labels based on an estimated confidence, without explicitly considering *how much* each sample should be disambiguated or *how many* labels should be retained. These questions are particularly important in multi-label ECG diagnosis, where the amount of ambiguity can vary between samples and samples can have valid co-occurring diagnoses.

To address these questions, we first study how confidence-based disambiguation affects ECG samples with different levels of ambiguity during training. Fig. 1(a) shows the evolution of disambiguation deviation during training. While ambiguous samples reduce their deviation, reliable samples

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

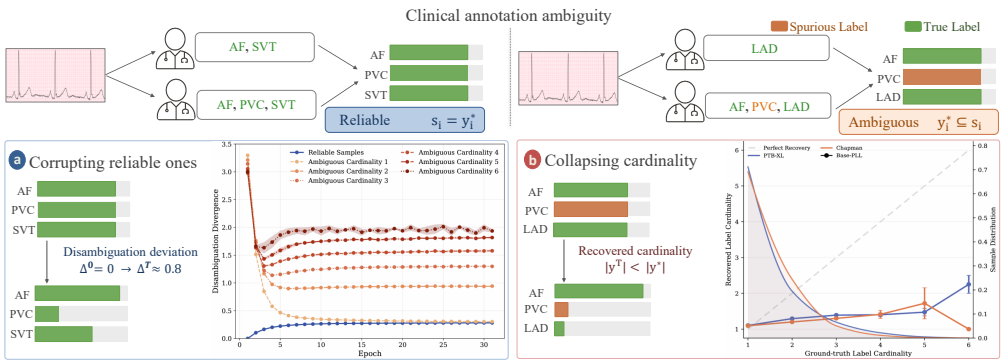


Figure 1: **Ambiguity and disambiguation challenges in multi-label ECG diagnosis.** (a) Disambiguation deviation on PTB-XL (Wagner et al., 2020), measured as the L2 distance between the disambiguated and ground-truth label set for reliable and ambiguous samples, grouped by ground-truth cardinality. Deviation of ambiguous samples generally decreases, whereas reliable samples deviate from their initially correct supervision. (b) Recovered label cardinality as a function of ground-truth label cardinality on PTB-XL and Chapman (Zheng et al., 2020). The recovered cardinality underestimates the ground-truth cardinality, indicating under-recovery of multi-label samples.

progressively deviate from their initially correct labels, indicating that disambiguation can benefit ambiguous samples while unnecessarily modifying reliable ones. Since real-world ECG datasets can contain both reliable and ambiguous candidate-label supervision (Ribeiro et al., 2020), this observation motivates adapting the degree of disambiguation to each sample. Interestingly, Fig. 1(a) also shows that higher-cardinality samples experience more disambiguation deviation. Accordingly, we next examine whether confidence-based disambiguation preserves the multi-label structure of ECG by comparing recovered and ground-truth label cardinality in Fig. 1(b). The recovered cardinality increasingly underestimates the ground-truth cardinality as the number of diagnoses increases, showing that samples with multiple diagnoses are disambiguated toward fewer labels. Although instance-dependent prediction cardinality has been recognized as an important consideration in multi-label classification (Cortes et al., 2024), existing PLL disambiguation mechanisms generally select correct candidate labels according to their relative confidence without explicitly modeling how many labels should be retained. These observations indicate that effective disambiguation for multi-label ECG diagnosis must account for both the degree of refinement and the number of diagnoses to preserve.

Motivated by these observations, we propose Candidate Label **Re**construction with **Ca**rdinality Preservation for Multi-**l**abel ECG (ReCal), a framework that adapts candidate-label disambiguation to each ECG sample. ReCal addresses two questions: *how much* the candidate labels be refined, and *how many* labels should be retained. First, ReCal learns a sample-specific weight that controls the degree of candidate-label refinement while scaling the training loss. Second, ReCal estimates the number of labels to retain and uses it to preserve valid co-occurring diagnoses during disambiguation. Accordingly, ReCal jointly adapts the *extent* of disambiguation and the *cardinality* of the recovered label set. We evaluate ReCal on multi-label ECG datasets under controlled ambiguity and demonstrate improved disambiguation and classification performance over baselines.

Our contributions are summarized as follows. ❶ We identify and explain two failures of PLL disambiguation for multi-label ECG diagnosis: uniform disambiguation of training samples and under-recovery of multi-label samples by suppressing weaker diagnoses. ❷ We propose ReCal, which controls the degree of disambiguation and estimates the cardinality of valid co-occurring diagnoses. ❸ Extensive evaluations across multiple ECG datasets and ambiguity-generation strategies show that our method consistently matches or outperforms prior state-of-the-art approaches. To enable fast reproducibility and contribute to the field, we make our code public.

2 Related Work

PLL addresses learning from ambiguously annotated data, where each instance is associated with a candidate label set containing the underlying unknown ground-truth label(s) together with false candidates. Existing methods can be broadly categorized according to how they use candidate-label

information. Average-based approaches treat all the candidate labels as valid supervision during training (Durand et al., 2019; Wu et al., 2022), without explicitly identifying the underlying true labels. In contrast, identification-based approaches focus on identifying the ground-truth labels by progressively refining the candidate set, *i.e.*, disambiguation (Lv et al., 2020; Zhang et al., 2021; Wang et al., 2022a). For example, PRODEN (Lv et al., 2020) progressively updates candidate-label weights according to model prediction confidence, while CAVL (Zhang et al., 2021) uses class activation values to estimate candidate-label importance. Furthermore, some approaches focus on log-tail distribution in the context of PLL using Multi-Focal Modifier which is a class frequency aware loss (Zhang et al., 2023). Semantic information has also been exploited to capture relationships between labels and instances by modeling class relationships, as in SST (Chen et al., 2022). Some approaches combine curriculum learning with PLL to gradually train on difficult samples by normalizing the loss based on candidate label size (Durand et al., 2019). Other PLL methods use additional information to improve disambiguation, including feature and label relationships through graph-based propagation (Wang et al., 2022b), class prototypes and contrastive learning (Wang et al., 2022a), and consistency regularization across augmented samples (Wu et al., 2022).

Partial multi-label learning has further applied PLL to multi-label learning, where multiple underlying ground-truths exist within the candidate label set. Most partial multi-label learning approaches focus on disambiguation. Some approaches rely on instance relationships to recover ground-truths shared between similar instances by clustering them (Xie & Huang, 2018). Similarly, feature and label correlations are used to recover the ground-truth labels (Wang et al., 2023). Some approaches treat ground-truth labels as latent variables and recover them with direct graphical models (Hang & Zhang, 2023). However, less attention has been given to how the disambiguation process itself affects samples with different supervision reliability and, in the multi-label setting, how it affects the number of valid labels recovered from a candidate set. Our work instead examines and incorporates both sample reliability and label cardinality into the disambiguation process.

3 METHODOLOGY

3.1 PROBLEM SETUP AND PRELIMINARIES

Let $\mathcal{Y} = \{1, \dots, C\}$ denote a set of diagnostic classes. Each ECG recording $\boldsymbol{x}_i \in \mathbb{R}^d$ is associated with an unknown ground-truth vector $\boldsymbol{y}_i^* \in \{0,1\}^C$, $\boldsymbol{y}_i^* \neq \mathbf{0}$, and an observed candidate-set vector $\boldsymbol{s}_i \in \{0,1\}^C$ satisfying $\boldsymbol{y}_i^* \preceq \boldsymbol{s}_i$, where $\preceq$ denotes the element-wise inequality. During training, we maintain a supervision vector $\boldsymbol{y}_i^{(t)} \in [0,1]^C$ at epoch t, supported on $\boldsymbol{s}_i$ and initialized as $\boldsymbol{y}_i^{(0)} = \boldsymbol{s}_i$, where $t \in \{0, \dots, T\}$. Let $\mathcal{S}_i = \operatorname{supp}(\boldsymbol{s}_i)$ for the candidate set, $\mathcal{Y}_i^* = \operatorname{supp}(\boldsymbol{y}_i^*) \subseteq \mathcal{S}_i$ for the true label set, and $k_i^* = |\mathcal{Y}_i^*|$ for the ground-truth cardinality. Since k_i^* may exceed one, this setting corresponds to partial multi-label learning. We call a sample *reliable* if $\boldsymbol{s}_i = \boldsymbol{y}_i^*$ and *ambiguous* otherwise. Given a training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{s}_i)\}_{i=1}^n$, we train a multi-label classifier $f(\cdot;\theta) : \mathbb{R}^d \to (0,1)^C$ with sigmoid outputs. Since the ground-truth vector $\boldsymbol{y}_i^*$ is unavailable during training, the model minimizes the empirical risk using the current supervision:

$$\mathcal{L}^{(t)}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(\boldsymbol{x}_i;\theta), \boldsymbol{y}_i^{(t)}\big), \tag{1}$$

where ℓ is the binary cross-entropy loss evaluated using the current supervision $\boldsymbol{y}_i^{(t)}$, $\theta^{(t)}$ denote the parameters, and $\boldsymbol{p}_i^{(t)} = f(\boldsymbol{x}_i;\theta^{(t)})$ denote the prediction. The supervision is refined by a disambiguation operator $\Phi_{\boldsymbol{s}_i}$ that uses the model prediction while restricting the supervision to the candidate set:

$$\boldsymbol{y}_i^{(t+1)} = \Phi_{\boldsymbol{s}_i}\left(\boldsymbol{y}_i^{(t)}, \boldsymbol{p}_i^{(t)}\right), \qquad \operatorname{supp}\big(\boldsymbol{y}_i^{(t+1)}\big) \subseteq \mathcal{S}_i. \tag{2}$$

An effective disambiguation process should suppress spurious candidate labels while preserving true diagnoses. In particular, it should avoid modifying candidate sets that are already reliable and should preserve the multi-label structure of samples with multiple underlying diagnoses. These requirements motivate our analysis of two complementary aspects of disambiguation: *how much* each sample should be disambiguated and *how many* labels should be retained during disambiguation.

Base-PLL. As a reference for our analysis, we consider a simple confidence-based scheme, referred to as *Base-PLL*, which trains with binary cross-entropy loss and uses the operator

$$\boldsymbol{\rho}_i^{(t)} = \Phi_{\boldsymbol{s}_i}\left(\boldsymbol{y}_i^{(t)}, \boldsymbol{p}_i^{(t)}\right) = \frac{\boldsymbol{s}_i \odot \boldsymbol{p}_i^{(t)}}{\max_{c\in\mathcal{S}_i} p_{ic}^{(t)}}, \tag{3}$$

i.e., the prediction is restricted to the candidates and normalized by its maximum. The component $\rho_{ic}^{(t)}$ is the confidence of candidate c relative to the most confident candidate $c_i^* \in \arg\max_{c\in\mathcal{S}_i} p_{ic}^{(t)}$, which we call the *dominant* candidate. [1]

3.2 Reliability-Adaptive Disambiguation

Existing identification-based PLL methods uniformly disambiguate every training sample, regardless of whether its candidate label set is ambiguous or already agrees with the underlying ground-truth labels. Disambiguation is useful for removing spurious candidate labels, but unnecessary when the candidate set is reliable.

Motivation: confidence sharpening can corrupt reliable samples. To understand this behavior, we analyze the Base-PLL operator in Eq. 3, which assigns each candidate its confidence relative to the dominant candidate. The following lemma shows that this operator does not account for whether a candidate set is already correct.

Lemma 1. *The Base-PLL disambiguation operator in Eq. 3 satisfies the following properties.*

- *(i)* $\rho_c \in (0,1]$ *for* $c \in \mathcal{S}$, $\rho_c = 0$ *for* $c \notin \mathcal{S}$, *and* $\rho_{c^*} = 1$. *Hence, every output of* $\Phi_{\boldsymbol{s}}$ *contains at least one saturated coordinate, regardless of the ground-truth cardinality* k^*.
- *(ii)* $\Phi_{\boldsymbol{s}}(\boldsymbol{y}, \boldsymbol{p})$ *does not depend on its first argument. The supervision history* $\boldsymbol{y}^{(0)}, \ldots, \boldsymbol{y}^{(t)}$ *therefore affects* $\boldsymbol{y}^{(t+1)}$ *only through the trained parameters* $\theta^{(t)}$.
- *(iii) For fixed* $(\boldsymbol{s}, \boldsymbol{p})$, *the output is identical for every ground-truth vector* $\boldsymbol{y}^* \preceq \boldsymbol{s}$. *In particular, disambiguation does not depend on whether the sample is reliable or on its cardinality* k^*.

The proof is given in Appendix A.2.2. These properties imply that confidence-based disambiguation continuously sharpens the relative confidence of candidate labels without knowing whether such sharpening is beneficial for a particular sample. To quantify the effect of disambiguation on ground-truth labels, we define the *disambiguation deviation* as follows:

$$\Delta_i^{(t)} = \left\|\boldsymbol{y}_i^* - \boldsymbol{y}_i^{(t)}\right\|_2^2. \tag{4}$$

Since $\boldsymbol{y}_i^*$ is unavailable during training, $\Delta_i^{(t)}$ is used only for analysis in the synthetic-ambiguity setting, where candidate sets are generated from known ground-truth labels. For a reliable sample, $\Delta_i^{(0)} = 0$. We decompose the disambiguation deviation into contributions from true and spurious candidate labels in the following proposition:

Proposition 1 (Deviation decomposition)**.** *Let* $\boldsymbol{y}^{(t+1)} = \boldsymbol{\rho}$ *be produced by Base-PLL. Then*

$$\Delta^{(t+1)} = \underbrace{\sum_{c\in\mathcal{Y}^*} (1-\rho_c)^2}_{A(\boldsymbol{\rho})} + \underbrace{\sum_{c\in\mathcal{S}\setminus\mathcal{Y}^*} \rho_c^2}_{B(\boldsymbol{\rho})}, \tag{5}$$

with $0 \le B(\boldsymbol{\rho}) \le |\mathcal{S}\setminus\mathcal{Y}^*|$. *Moreover,* $0 \le A(\boldsymbol{\rho}) \le k^* - 1$ *if the dominant candidate is a true label* ($c^* \in \mathcal{Y}^*$), *and* $0 < A(\boldsymbol{\rho}) < k^*$ *otherwise.*

The term $A(\boldsymbol{\rho})$ measures the *suppression* of true labels relative to the dominant candidate, and $B(\boldsymbol{\rho})$ measures the residual confidence on spurious candidates. To describe how these terms evolve, we formalize the notion that the model becomes more confident in one candidate relative to the others.

Definition 1 (Sharpening)**.** For a sample with candidate set $\mathcal{S}$, a prediction $\boldsymbol{p}'$ is a *sharpening* of $\boldsymbol{p}$ on $\mathcal{S}$ if the following two conditions hold:

[1] When analyzing a single sample, we omit the indices i and t and write $\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{\rho}, \mathcal{S}, \mathcal{Y}^*, k^*$ and c^*.

(i) The dominant candidate is unchanged, *i.e.*, $c^* \in \arg\max_{c \in \mathcal{S}} p_c$ and $c^* \in \arg\max_{c \in \mathcal{S}} p'_c$.

(ii) The relative confidence of every non-dominant candidate does not increase, *i.e.*, $\rho'_c \leq \rho_c$ for all $c \in \mathcal{S} \setminus \{c^*\}$, with strict inequality for at least one c.

The confidence sharpening therefore has two opposing effects: decreasing the relative confidence of non-dominant labels increases $A(\boldsymbol{\rho})$ when those labels are genuine, while decreasing the confidence of spurious candidates reduces $B(\boldsymbol{\rho})$. Therefore, sharpening can simultaneously improve candidate-label purification and suppress valid but less confident diagnoses. This trade-off is particularly important for reliable samples. When the candidate set is already correct, *i.e.*, $\mathcal{S} = \mathcal{Y}^*$, there are no spurious labels to purify and therefore $B(\boldsymbol{\rho}) = 0$.

Corollary 1 (Corruption of reliable multi-label samples)**.** *Let the sample be reliable ($\mathcal{S} = \mathcal{Y}^*$). Then $B \equiv 0$ and $\Delta^{(t+1)} = \sum_{c \in \mathcal{S}} (1 - \rho_c)^2$. Furthermore:*

(i) If $k^* = 1$, then $\Delta^{(t)} = 0$ for all t, i.e., reliable single-label samples are never corrupted.

(ii) If $k^* \geq 2$ and $\{p_c\}_{c \in \mathcal{S}}$ are not all equal, then $\Delta^{(1)} \geq (1 - \rho_{\min})^2 > 0 = \Delta^{(0)}$, where $\rho_{\min} = \min_{c \in \mathcal{S}} \rho_c$. Hence, the first disambiguation step strictly moves the supervision away from the ground-truth.

(iii) If $k^* \geq 2$, Δ increases under sharpening and approaches $k^* - 1$ as $\rho_c \to 0$ for all $c \neq c^*$.

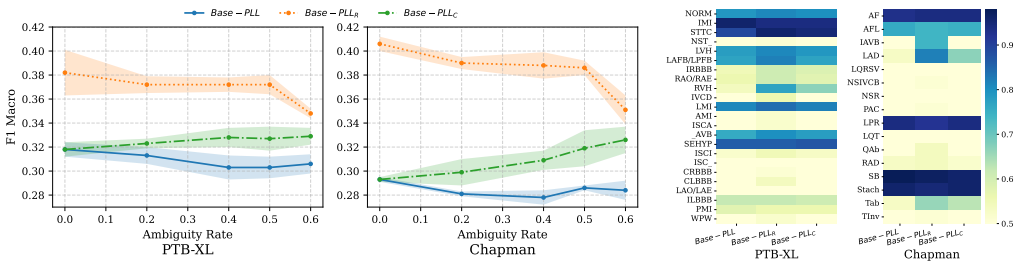
The proof is provided in Appendix A.2.5. By part (i), reliable single-label samples are never corrupted. By part (ii), for reliable multi-label samples, the first disambiguation step already moves the supervision away from the ground-truth. Part (ii) holds for any single application of the operator, whereas part (iii) is conditional on sharpening, *i.e.*, on how training shapes the predictions. Unlike ambiguous samples for which sharpening can remove spurious labels, sharpening a reliable multi-label sample moves its supervision away from the ground-truth. This is confirmed in Fig. 1, where reliable samples increase deviation from their initially correct supervision while ambiguous samples progressively reduce their deviation during training.

Controlled analysis. To empirically study the effect of this corruption on optimization, we perform an oracle study, denoted by Base-PLL$_{\text{R}}$, in which reliable samples are excluded from disambiguation during training. Fig. 2a reports the performance with and without reliable sample separation from disambiguation across different ambiguity levels and classes for PTB-XL and Chapman datasets. Separation improves classification performance on both datasets and for all ambiguity levels. This indicates that unnecessary updates to reliable samples degrade learning in PLL frameworks. To further investigate how this effect varies across classes, Fig. 2b presents the class-specific performance before and after excluding reliable samples from disambiguation. Certain conditions, *e.g.*, nonspecific intraventricular conduction delay (IVCD), include a broad set of conduction abnormalities, resulting in higher intra-class variability. Additionally, explainability analysis has shown that DNNs are more capable of learning representations that are consistent with cardiologists' reasoning for classes such as CLBBB or ILBBB (Bender et al., 2023). Since CLBBB is also among the classes with low sample count in the PTB-XL dataset ($\sim$100 samples), this could explain why separating reliable samples within disambiguation has been more beneficial for this class.

Sample-specific reliability learning. These observations suggest that disambiguation should be applied only to samples that need it. Inspired by noisy-sample detection (Wu et al., 2023), we assign each sample a learnable scalar $w_i \in \mathbb{R}$ to measure how reliable a sample is, where larger w_i indicate less reliable supervision. We jointly optimize the model parameters θ and the sample-specific weights w using the objective

$$\mathcal{J}^{(t)}(\theta, \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(\boldsymbol{x}_i; \theta), \boldsymbol{y}_i^{(t)}\big)\, e^{-w_i} + \frac{\lambda}{n} \sum_{i=1}^{n} w_i, \tag{6}$$

where $\lambda > 0$. The first term down-weights samples whose current supervision is hard to fit, and the second term prevents the trivial solution $w_i \to \infty$ that would suppress all samples. While this self-paced learning behavior prioritizes reliable samples over ambiguous ones, samples that are reliably disambiguated can regain influence throughout training. Consequently, we can use w_i to



(a) Performance over different ambiguity proportions. (b) Class-specific performance.

Figure 2: Impact of reliability- and cardinality-aware disambiguation on performance.

supervise disambiguation by estimating the reliability of a candidate label set and suppressing further disambiguation to avoid corrupting reliable label sets. To control disambiguation, we normalize the weights to $[0, 1]$ as w'_i and map them through the gate

$$\tilde{w}_i = \frac{1}{1 + \exp\big(-\tau(w'_i - 0.5)\big)}, \tag{7}$$

where $\tau > 0$ controls the sharpness. We perform a reliability-weighted disambiguation, as

$$\boldsymbol{y}_i^{(t+1)} = (1 - \tilde{w}_i)\, \boldsymbol{y}_i^{(t)} + \tilde{w}_i\, \tilde{\boldsymbol{y}}_i^{(t)}, \qquad \tilde{\boldsymbol{y}}_i^{(t)} = \Phi_{\boldsymbol{s}_i}\left(\boldsymbol{y}_i^{(t)}, \boldsymbol{p}_i^{(t)}\right), \tag{8}$$

where $\tilde{\boldsymbol{y}}_i^{(t)}$ is the disambiguated estimate. Equation 8 replaces the uniform disambiguation update with a sample-adaptive one, where reliable samples ($\tilde{w}_i \to 0$) keep the original candidate set, while ambiguous samples ($\tilde{w}_i \to 1$) are disambiguated. Thus, the learned reliability determines *how much* each sample should be disambiguated, addressing the unnecessary sharpening of reliable samples.

3.3 Cardinality-Aware Disambiguation

Confidence-based disambiguation determines *which* candidates are likely to be correct, but it does not model *how many* labels a sample should retain. Given a threshold $\gamma \in (0, 1)$, we define the labels recovered from the supervision vector and their number as

$$\hat{\boldsymbol{y}}_i^{(t)} = \mathbb{I}\left[\boldsymbol{y}_i^{(t)} \geq \gamma\right], \qquad k_i^{(t)} = \left\|\hat{\boldsymbol{y}}_i^{(t)}\right\|_1, \tag{9}$$

where $\mathbb{I}[\cdot]$ is applied element-wise. Existing confidence-based disambiguation does not include an objective that encourages the number of retained labels to match the number of true diagnoses. This can be problematic for multi-label samples, where valid diagnoses may have different confidence levels and weaker diagnoses can be suppressed by more dominant labels.

Motivation: confidence sharpening can suppress valid labels. Under Base-PLL, $\boldsymbol{y}_i^{(t)} = \boldsymbol{\rho}_i$, so a true label $c \in \mathcal{Y}_i^*$ is recovered if and only if $p_{ic} \geq \gamma \max_{c' \in \mathcal{S}_i} p_{ic'}$. The recovery of a true label therefore depends on its confidence *relative to the dominant candidate* rather than on its own confidence. By Lemma 1(i), the dominant candidate is always recovered, *i.e.*, $k_i^{(t)} \geq 1$. However, no explicit mechanism in the training objective or in the operator encourages $k_i^{(t)} \approx k_i^*$. As established in Lemma 1, for a fixed candidate set and model prediction, its output is invariant to the underlying ground-truth cardinality. From Eq. 5, decreasing the relative confidence of non-dominant true labels increases the suppression term $A(\boldsymbol{\rho})$. Hence, when several diagnoses are present, sharpening can progressively reduce the confidence of valid but non-dominant labels, causing the recovered cardinality to become smaller than the underlying ground-truth cardinality. This mechanism is consistent with Fig. 1(a): for ambiguous samples with high ground-truth cardinality, the deviation first decreases as $B(\boldsymbol{\rho})$ is reduced and then grows as $A(\boldsymbol{\rho})$ increases. As observed in Fig. 1(b), the skewed cardinality distribution of multi-label ECG datasets may further exacerbate this effect, since the model receives comparatively little supervision for predicting co-occurring labels. As a result, the recovered cardinality increasingly underestimates the ground-truth cardinality as the number of true diagnoses grows.

Table 1: Comparison with baselines on the PTB-XL, Chapman, and AMIGOS datasets under controlled candidate-label ambiguity. Macro-F1 is reported for overall performance as mean ± standard deviation. The best and second-best results are respectively shown in **bold** and underline.

	PTB-XL				Chapman			AMIGOS	
Method	Cardio-Feat	Deep	Random	Cardio-Class	Random	Deep	Treatment	Random	Circumplex
PRODEN	0.277 ±0.012	0.280 ±0.005	0.275 ±0.007	0.286 ±0.004	0.275 ±0.008	0.274 ±0.007	0.272 ±0.003	0.322 ±0.004	0.323 ±0.011
PARD	0.302 ±0.016	0.247 ±0.011	0.303 ±0.008	0.302 ±0.010	0.264 ±0.032	0.261 ±0.026	0.330 ±0.017	0.317 ±0.078	0.373 ±0.053
HST	0.164 ±0.006	0.253 ±0.017	0.173 ±0.014	0.209 ±0.020	0.297 ±0.003	0.281 ±0.014	0.288 ±0.011	0.574 ±0.013	0.610 ±0.035
SST	0.089 ±0.042	0.260 ±0.003	0.130 ±0.021	0.104 ±0.034	0.231 ±0.038	0.292 ±0.019	0.277 ±0.010	0.534 ±0.016	0.544 ±0.089
CAVL	0.094 ±0.010	0.210 ±0.009	0.189 ±0.010	0.183 ±0.006	0.241 ±0.006	0.215 ±0.015	0.240 ±0.016	0.239 ±0.019	0.255 ±0.006
CR	0.316 ±0.002	0.333 ±0.011	0.310 ±0.008	0.322 ±0.003	0.327 ±0.004	0.347 ±0.007	0.324 ±0.004	0.587 ±0.018	0.553 ±0.024
Base-PLL	0.312 ±0.004	0.332 ±0.006	0.303 ±0.009	0.306 ±0.008	0.286 ±0.002	0.304 ±0.004	0.280 ±0.003	0.304 ±0.004	0.517 ±0.013
ReCal	**0.347** ±0.004	**0.389** ±0.007	**0.343** ±0.009	**0.338** ±0.003	**0.388** ±0.002	**0.372** ±0.001	**0.381** ±0.003	**0.620 ±0.029**	**0.687 ±0.004**

Controlled analysis. To determine whether the observed cardinality collapse affects optimization, we perform a controlled experiment by modifying the confidence-based disambiguation step to prevent the recovered cardinality from falling below the ground-truth cardinality (*i.e.*, $|\hat{\boldsymbol{y}}_i^{(t)}| \geq |\boldsymbol{y}_i^*|$). This experiment uses the ground-truth cardinality only for analysis and is therefore unavailable in practical training. Nevertheless, it provides an upper bound on the benefit that can be obtained from preserving cardinality. As shown in Fig. 2a, preventing cardinality collapse improves performance over the basic confidence-based disambiguation. Interestingly, preserving cardinality seems to improve robustness of a PLL model against higher levels of ambiguity. This confirms that disambiguation should preserve the multi-label structure of each sample and avoid systematically collapsing samples with multiple true diagnoses toward low-cardinality disambiguation.

Instance-dependent cardinality estimation. Since the true cardinality k_i^* is unavailable during training, we estimate it from the samples that the reliability module currently identifies as reliable $\mathcal{R}^{(t)} = \{\, i \in \{1, \ldots, n\} \;:\; w_i' \leq \delta \,\}$, where $\delta \in (0, 1)$ is a threshold. For these samples, the recovered cardinality $k_i^{(t)}$ in Eq. 9 serves as a regression target. We train a lightweight auxiliary network $g_\phi : (0,1)^C \to \mathbb{R}$ that predicts the cardinality from the main model's prediction by minimizing

$$\mathcal{H}_{\text{card}}^{(t)}(\phi) = \frac{1}{|\mathcal{R}^{(t)}|} \sum_{i \in \mathcal{R}^{(t)}} \text{Huber}\Big(g_\phi\big(\boldsymbol{p}_i^{(t)}\big),\, k_i^{(t)} \Big). \tag{10}$$

Because the cardinality targets in $\mathcal{R}^{(t)}$ follow a long-tailed distribution, we sample training examples such that the number of examples drawn with cardinality k is proportional to n_k^α, where n_k is the number of samples in $\mathcal{R}^{(t)}$ with $k_i^{(t)} = k$ and $\alpha \in (0, 1)$. Setting $\alpha = 1$ recovers the original distribution and $\alpha = 0$ balances all cardinalities. Hence, intermediate values partially rebalance the distribution without aggressively oversampling rare cardinalities. The predicted cardinality is obtained by rounding and clipping to the valid range,

$$\hat{k}_i^{(t)} = \min\Big(\max\big(\text{round}\,\big(g_\phi(\boldsymbol{p}_i^{(t)})\big),\, 1\big),\, |\mathcal{S}_i| \Big). \tag{11}$$

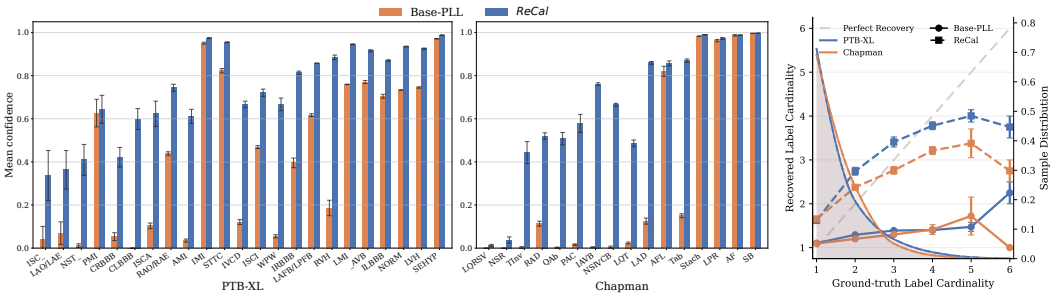
Restricting the cardinality supervision to the reliable subset couples the two components: the reliability mechanism identifies samples whose current supervision can provide a trustworthy estimate of cardinality. Once the auxiliary model has been trained, we obtain a predicted cardinality $\hat{k}_i^{(t)}$ for each instance and modify the disambiguation vector to preserve the estimated cardinality. Samples whose confidence vector contains fewer than $\hat{k}_i$ labels above the threshold are rescaled to map the top-$\hat{k}_i$ predicted classes into the range $[\gamma, 1]$. This transformation preserves the relative ordering of the selected candidates while preventing the confidence threshold from suppressing the $\hat{k}_i$ most plausible candidate labels. When the selected confidence values have negligible spread, all selected labels are instead assigned confidence one. The resulting disambiguation uses the model confidence to determine *which* candidate labels are likely to be correct, while the predicted cardinality determines *how many* labels should be retained, preserving the multi-label structure in disambiguation.

4 EXPERIMENTS

Datasets and ambiguity generation. We evaluate our method on the PTB-XL (Wagner et al., 2020) and Chapman-Shaoxing (Zheng et al., 2020) datasets for ECG diagnosis, and on the AMIGOS dataset (Miranda-Correa et al., 2018) for emotion recognition from ECG. We construct candidate label sets from the original multi-label annotations for a proportion $p \in [0, 0.6]$ of samples, with a probability $\epsilon \in [0, 1]$ of including a plausible label, following the setup of (Tian et al., 2023). Following the PLL setup for ECG diagnosis introduced in (Rahmani et al., 2025a), we consider multiple ambiguity-generation strategies: *random*, *instance-level cardiologist-driven*, *instance-level model-driven*, *class-level cardiologist-driven*, and *class-level treatment-driven* ambiguity. For emotion recognition, we follow (Zhang & Etemad, 2025) and consider two ambiguity-generation setups: *random* and *class-level Circumplex-driven* ambiguity. In particular, *instance-level* strategies introduce ambiguity according to individual ECG samples, whereas *class-level* strategies introduce ambiguity based on relationships between classes. Unless otherwise specified, the proportion of partially labelled training samples is set to $p = 0.5$.

Evaluation metrics. We evaluate the baselines in terms of both final classification performance and disambiguation performance. To assess classification performance, we use macro-F1 and AUC-ROC scores. To evaluate disambiguation performance, we define two additional metrics: (I) *Reliable sample Deviation* (RD), which measures the mean disambiguation deviation for reliable samples, and (II) *Ambiguous sample Deviation* (AD), which measures the average of the mean disambiguation deviation across groups of samples categorized by ground-truth label cardinality. We additionally report the average predicted cardinality and confidence assigned to positive labels for each class.

Overall performance. Table 1 reports the classification performance of ReCal and the baseline PLL methods under different ambiguity scenarios across three datasets. ReCal consistently improves over the baseline methods across the evaluated settings. These results indicate that adapting candidate-label disambiguation to the reliability and cardinality of individual samples is beneficial across different datasets and ambiguity-generation strategies. We further examine the robustness of these gains to changes in the amount of ambiguous supervision in Appendix A.4. The performance improvement remains consistent as the proportion of partially labelled samples increases and across different ambiguity-generation strategies, as shown in Fig. 4. Additional class-wise AUC-ROC results are provided in Fig. 5.



(a) Recovered class confidences.

(b) Recovered cardinality.

Figure 3: Analyzing disambiguation for PTB-XL and Chapman datasets.

Analyzing Candidate Label Disambiguation. To examine the effect of ReCal disambiguation, we assess how well different diagnostic classes are preserved during disambiguation. We measure the average confidence assigned during disambiguation to each class over samples for which that class belongs to the ground-truth label set. Accordingly, we compare ReCal with Base-PLL across diagnostic classes in Fig. 3a for both PTB-XL and Chapman datasets. To examine the relationship between class frequency and recovered confidence, we sort the classes according to their frequency in the training set, with the leftmost class having the lowest frequency. Interestingly, higher-frequency classes tend to exhibit higher confidence after disambiguation, whereas most minor classes receive lower confidence. While ReCal maintains high confidence for frequently occurring classes, it relatively increases the recovered confidence of minor classes, potentially preserving

them in disambiguation. We also study the disambiguation deviation within training of baselines in Fig. 8. We next examine whether confidence-based disambiguation preserves the number of co-occurring diagnoses. Fig. 3b reports the recovered label cardinality as a function of the ground-truth label cardinality. Base-PLL increasingly underestimates the number of labels as the ground-truth cardinality increases. ReCal instead uses an instance-dependent cardinality estimate learned from reliable samples to guide the number of labels retained during disambiguation. As shown in Fig. 3b, ReCal reduces the gap between the recovered and ground-truth cardinality, particularly for samples with higher ground-truth cardinality. These results support the effectiveness of explicitly accounting for instance-dependent cardinality during candidate-label disambiguation. We also compare additional PLL baselines across all datasets in Fig. 7, which show a similar tendency toward cardinality under-recovery, with ReCal preserving more labels as ground-truth cardinality increases.

Table 2: Ablation study of the ReCal framework and its components on PTB-XL, Chapman, and AMIGOS dataset under random ambiguity. Macro-F1 is reported for overall performance on classification, AD, and RD as disambiguation performance reported as mean ± standard deviation. The best and second-best results are respectively shown in **bold** and underline.

Method	PTB-XL			Chapman			Amigos		
	F1↑	AD↓	RD↓	F1↑	AD↓	RD↓	F1↑	AD↓	RD↓
Base-PLL	0.303 ± 0.009	1.312 ± 0.558	0.281 ± 0.486	0.286 ± 0.002	1.406 ± 0.706	0.320 ± 0.499	0.304 ± 0.004	1.002 ± 0.305	0.305 ± 0.485
ReCal$_{\text{w/o C}}$	0.335 ± 0.006	1.138 ± 0.417	**0.004 ± 0.042**	0.359 ± 0.007	1.235 ± 0.598	0.031 ± 0.179	0.615 ± 0.008	0.857 ± 0.081	0.138 ± 0.247
ReCal$_{\text{w/o R}}$	0.329 ± 0.007	1.228 ± 0.423	0.229 ± 0.417	0.317 ± 0.007	1.314 ± 0.615	0.268 ± 0.436	0.593 ± 0.009	**0.807 ± 0.038**	0.394 ± 0.409
ReCal	**0.343 ± 0.009**	**1.131 ± 0.338**	**0.004 ± 0.046**	**0.388 ± 0.002**	**1.191 ± 0.478**	**0.026 ± 0.152**	**0.620 ± 0.029**	0.843 ± 0.145	**0.136 ± 0.243**

Ablation Studies. ReCal combines two complementary mechanisms: sample-specific reliability adaptation and cardinality-aware disambiguation. We compare Base-PLL and ReCal with two variants that remove one component at a time: (I) ReCal without cardinality-aware disambiguation (ReCal$_{\text{w/o C}}$), and (II) ReCal without reliability-adaptive disambiguation (ReCal$_{\text{w/o R}}$). Since reliable samples are required to train the auxiliary MLP for cardinality estimation, ReCal$_{\text{w/o R}}$ retains the reliability weight in the training loss but removes it from subsequent disambiguation. Table 2 reports the results under random ambiguity across all three datasets using Macro-F1, AD, and RD. Removing the cardinality-aware disambiguation reduces Macro-F1 and increases AD relative to ReCal, highlighting the benefit of instance-dependent cardinality for preserving valid labels. In contrast, removing reliability adaptation substantially increases RD, particularly on PTB-XL and Chapman, demonstrating its role in preventing unnecessary disambiguation of reliable samples. Although the relative effect of each component varies across datasets and metrics, the consistent gains in classification performance and the complementary effects on AD and RD demonstrate the effectiveness of these mechanisms for disambiguation. To further examine the reliability weights, Fig. 6 shows their distributions for ambiguous and reliable samples during training, while Fig. 8 shows the resulting disambiguation deviation for reliable samples.

5 Conclusion

In this work, we first analyze confidence-based candidate-label disambiguation in multi-label ECG diagnosis and identify two systematic issues that can arise during disambiguation. Specifically, we show that applying the same degree of disambiguation across samples can unnecessarily modify candidate-label sets that already provide reliable supervision, while confidence-based refinement can suppress valid but less confident diagnoses and lead to under-recovery of multi-label samples. Motivated by these observations, we propose ReCal, a sample-adaptive candidate-label disambiguation framework that considers both the degree of disambiguation and the number of labels to retain. ReCal learns a sample-specific weight to adapt the strength of candidate-label refinement and estimates instance-dependent label cardinality to preserve valid co-occurring diagnoses during disambiguation. We evaluate ReCal on the PTB-XL and Chapman-Shaoxing datasets for ECG diagnosis, and on the AMIGOS dataset for emotion recognition from ECG, under multiple controlled ambiguity-generation strategies. The experimental results demonstrate that ReCal improves both candidate-label disambiguation and downstream multi-label ECG classification compared to prior PLL methods. Ablation studies further show that the two components provide complementary contributions to the overall performance. These results highlight the importance of adapting candidate-label disambiguation to the characteristics and multi-label structure of ECG samples.

AI USE STATEMENT

In this work, we used generative AI tools for grammatical check and readability of the paper text, feedback on experiment design, and suggesting titles. We have not used generative AI tools for developing framework, theoretical discussions, or mathematical proves, and generate synthetic data sets are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have ensured the reproducibility of our theoretical and empirical results. For the theoretical claims, all underlying assumptions and complete mathematical proofs are provided in Appendix A.1 and A.2. For the empirical evaluations, the description of the dataset processing steps, model architectures, and hyperparameter configurations are explained in Section A.3. Furthermore, the source code to reproduce the experiments are available in the anonymous supplementary material and will be available on GitHub upon publication.

REFERENCES

Theresa Bender, Jacqueline M. Beinecke, Dagmar Krefting, Carolin Müller, Henning Dathe, Tim Seidler, Nicolai Spicher, and Anne-Christin Hauschild. Analysis of a deep learning model for 12-lead ecg classification reveals learned features similar to diagnostic criteria. *IEEE journal of biomedical and health informatics*, 28(4):1848–1859, 2023.

T. Chen, T. Pu, H. Wu, Y. Xie, and L. Lin. Structured semantic transfer for multi-label recognition with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 339–346, 2022.

T. Chen, T. Pu, L. Liu, Y. Shi, Z. Yang, and L. Lin. Heterogeneous semantic transfer for multi-label recognition with partial labels. *International Journal of Computer Vision*, 132(12):6091–6106, 2024.

Corinna Cortes, Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Cardinality-aware set prediction and top-k classification. *Advances in neural information processing systems*, 37:18265–18309, 2024.

T. Durand, N. Mehrasa, and G. Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 647–657, 2019.

Jun-Yi Hang and Min-Ling Zhang. Partial multi-label learning with probabilistic graphical disambiguation. *Advances in Neural Information Processing Systems*, 36:1339–1351, 2023.

A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

L. S. Johnson, P. Zadrozniak, G. Jasina, A. Grotek-Cuprjak, J. G. Andrade, E. Svennberg, S. Z. Diederichsen, W. F. McIntyre, S. Stavrakis, and J. Benezet-Mazuecos. Artificial intelligence for direct-to-physician reporting of ambulatory electrocardiography. *Nature Medicine*, pp. 1–7, 2025.

J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning*, pp. 6500–6510. PMLR, 2020.

Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE transactions on affective computing*, 12(2):479–493, 2018.

S. Rahmani, R. Chatterjee, A. Etemad, and J. Hashemi. Dynamic prototype rehearsal for continual ecg arrhythmia detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2025a.

Sana Rahmani, Javad Hashemi, and Ali Etemad. Investigating ecg diagnosis with ambiguous labels using partial label learning. *arXiv preprint arXiv:2512.11095*, 2025b.

Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, and Wagner Meira Jr. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

Y. Tian, X. Yu, and S. Fu. Partial label learning: Taxonomy, analysis and outlook. *Neural Networks*, 161:708–734, 2023.

P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 2020.

H. Wang, M. Xia, Y. Li, Y. Mao, L. Feng, G. Chen, and J. Zhao. Solar: Sinkhorn label refinery for imbalanced partial-label learning. *Advances in Neural Information Processing Systems*, 35: 8104–8117, 2022a.

H. Wang, R. Xiao, Y. Li, L. Feng, G. Niu, G. Chen, and J. Zhao. Pico: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*, 2022b.

Haobo Wang, Shisong Yang, Gengyu Lyu, Weiwei Liu, Tianlei Hu, Ke Chen, Songhe Feng, and Gang Chen. Deep partial multi-label learning with graph disambiguation. *Proceedings of International Joint Conference on Artificial Intelligence, IJCAI*, 2023.

D.-D. Wu, D.-B. Wang, and M.-L. Zhang. Revisiting consistency regularization for deep partial label learning. In *International Conference on Machine Learning*, pp. 24212–24225. PMLR, 2022.

Tingting Wu, Xiao Ding, Hao Zhang, Jinglong Gao, Minji Tang, Li Du, Bing Qin, and Ting Liu. Discrimloss: A universal loss for hard samples and incorrect samples discrimination. *IEEE Transactions on Multimedia*, 26:1957–1968, 2023.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

F. Zhang, L. Feng, B. Han, T. Liu, G. Niu, T. Qin, and M. Sugiyama. Exploiting class activation value for partial-label learning. In *International Conference on Learning Representations*, 2021.

G. Zhang and A. Etemad. Partial label learning for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 2025.

Min-Ling Zhang and Jun-Peng Fang. Partial multi-label learning via credible label elicitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3587–3599, 2020.

W. Zhang, C. Liu, L. Zeng, B. Ooi, S. Tang, and Y. Zhuang. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1423–1432, 2023.

J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, 2020.

A Appendix

A.1 Why confidence-based disambiguation fails on multi-label ECG data

In addition to presenting the empirical results, we establish that the two identified failure modes arise as structural consequences of the max-normalized disambiguation operator employed by Base-PLL. In particular, these phenomena do not depend on a specific network architecture, optimization algorithm, or training trajectory.

A.2 Analysis of Confidence-Based Disambiguation

This section provides the proofs of the results in Sec. 3.2 and additional results on the behavior of the Base-PLL operator. The properties of a single disambiguation step (Lemma 1, Proposition 1, and Corollary 1(i)–(ii)) follow from the structure of the operator and hold for any network architecture and optimizer. The results on the evolution of the supervision over training are conditional on sharpening (Definition 1), *i.e.*, on how training shapes the predictions.

A.2.1 Setup and notation

We consider a single training sample and omit the indices i and t where no confusion arises. Let $\mathcal{Y} = \{1, \dots, C\}$ denote the label space, $\boldsymbol{s} \in \{0,1\}^C$ the candidate vector, and $\boldsymbol{y}^* \in \{0,1\}^C$ the unknown ground-truth vector with $\boldsymbol{y}^* \preceq \boldsymbol{s}$ and $\boldsymbol{y}^* \neq \mathbf{0}$. We write $\mathcal{S} = \operatorname{supp}(\boldsymbol{s})$, $\mathcal{Y}^* = \operatorname{supp}(\boldsymbol{y}^*) \subseteq \mathcal{S}$, $k^* = |\mathcal{Y}^*|$, and

$$a = |\mathcal{S} \setminus \mathcal{Y}^*| \tag{12}$$

for the number of spurious candidates, *i.e.*, the ambiguity level of the sample. A sample is reliable if and only if $a = 0$. Let $\boldsymbol{p} = f(\boldsymbol{x}; \theta) \in (0,1)^C$ denote the sigmoid output of the model. The Base-PLL operator in Eq. 3 can be written equivalently as

$$\Phi_{\boldsymbol{s}}(\boldsymbol{y}, \boldsymbol{p}) = \boldsymbol{\rho} = \frac{\boldsymbol{s} \odot \boldsymbol{p}}{\|\boldsymbol{s} \odot \boldsymbol{p}\|_\infty}, \qquad \|\boldsymbol{s} \odot \boldsymbol{p}\|_\infty = \max_{c \in \mathcal{S}} p_c, \tag{13}$$

where $\|\cdot\|_\infty$ denotes the max-norm, and the equality holds because $p_c > 0$ for all c. The dominant candidate is denoted by $c^* \in \arg\max_{c \in \mathcal{S}} p_c$. The disambiguation deviation is $\Delta^{(t)} = \|\boldsymbol{y}^* - \boldsymbol{y}^{(t)}\|_2^2$, as defined in Eq. 4.

A.2.2 Proof of Lemma 1

Proof. (i) Since $\boldsymbol{p} \in (0,1)^C$ and $\mathcal{S} \neq \emptyset$, we have $\max_{c \in \mathcal{S}} p_c > 0$, so the normalization is well defined. For $c \in \mathcal{S}$, $0 < p_c \leq \max_{c' \in \mathcal{S}} p_{c'}$, which gives $\rho_c \in (0,1]$ and $\rho_{c^*} = 1$. For $c \notin \mathcal{S}$, $s_c = 0$ and therefore $\rho_c = 0$.

(ii) The right-hand side of Eq. 13 does not contain $\boldsymbol{y}$. Hence, the supervision vector at epoch t has no *direct* influence on the supervision vector at epoch $t+1$; the new supervision is recomputed from the current prediction.[2]

(iii) The ground-truth vector $\boldsymbol{y}^*$ is not an argument of the operator in Eq. 13. Hence, for fixed $(\boldsymbol{s}, \boldsymbol{p})$, the output is the same for every $\boldsymbol{y}^* \preceq \boldsymbol{s}$, and in particular it does not depend on whether $\boldsymbol{s} = \boldsymbol{y}^*$ or on k^*.

□

Property (iii) implies that neither the training objective nor the disambiguation step contains a mechanism that aligns the number of retained labels with the true cardinality k^*: the operator is not only missing such a term, but is invariant to k^* by construction.

[2] Although $\boldsymbol{y}^{(t)}$ does not enter the operator, it influences $\boldsymbol{y}^{(t+1)}$ *indirectly*: $\boldsymbol{y}^{(t)}$ is the training target at epoch t, so it affects the gradients, the parameters $\theta^{(t)}$, and therefore the prediction $\boldsymbol{p}^{(t)}$ from which $\boldsymbol{y}^{(t+1)}$ is computed.

A.2.3 Proof of Proposition 1

Proof. Having $\boldsymbol{y}^{(t+1)} = \boldsymbol{\rho}$, we partition the label space into the disjoint sets $\mathcal{Y}^*$, $\mathcal{S} \setminus \mathcal{Y}^*$, and $\mathcal{Y} \setminus \mathcal{S}$. For $c \in \mathcal{Y}^*$, $y_c^* = 1$ and $y_c^{(t+1)} = \rho_c$, which contributes $(1-\rho_c)^2$. For $c \in \mathcal{S} \setminus \mathcal{Y}^*$, $y_c^* = 0$ and $y_c^{(t+1)} = \rho_c$, which contributes ρ_c^2. For $c \in \mathcal{Y} \setminus \mathcal{S}$, both y_c^* and $y_c^{(t+1)}$ are zero and contribute nothing. Summing over the three sets gives Eq. 5.

For the bounds, recall that $\rho_c \in (0, 1]$ for $c \in \mathcal{S}$. Each term of B satisfies $\rho_c^2 \in (0, 1]$, so $0 \leq B \leq |\mathcal{S} \setminus \mathcal{Y}^*| = a$. Each term of A satisfies $(1-\rho_c)^2 \in [0, 1)$, so $0 \leq A < k^*$. If $c^* \in \mathcal{Y}^*$, then $\rho_{c^*} = 1$ and its term vanishes, so A is a sum of $k^* - 1$ terms in $[0, 1)$ and $0 \leq A \leq k^* - 1$. □

The term A quantifies how much the true labels are reduced relative to the dominant candidate, which we refer to as *suppression*. The term B measures the residual relative confidence on spurious candidates, and its reduction corresponds to *purification*.

A.2.4 Properties of sharpening

Sharpening (Definition 1) is defined on the relative confidences ρ_c rather than on the raw probabilities p_c. Hence, all raw confidences may increase while the prediction is still a sharpening, provided that the non-dominant candidates increase more slowly than the dominant one. The relation is irreflexive, because it requires a strict decrease for at least one candidate, and transitive, because the dominant candidate is preserved and the inequalities compose. It is therefore a strict partial order, and not every pair of predictions is comparable; for example, two predictions are not comparable if one non-dominant candidate gains relative confidence while another loses it, or if the dominant candidate changes. Our results only require the relation to hold between the predictions being compared.

Lemma 2. *Let $\boldsymbol{p}'$ be a sharpening of $\boldsymbol{p}$ on $\mathcal{S}$, with relative confidences $\boldsymbol{\rho}'$ and $\boldsymbol{\rho}$. Then $A(\boldsymbol{\rho}') \geq A(\boldsymbol{\rho})$ and $B(\boldsymbol{\rho}') \leq B(\boldsymbol{\rho})$. Moreover, $A(\boldsymbol{\rho}') > A(\boldsymbol{\rho})$ if $\rho'_c < \rho_c$ for at least one $c \in \mathcal{Y}^* \setminus \{c^*\}$.*

Proof. The dominant candidate has $\rho'_{c^*} = \rho_{c^*} = 1$, so its term is unchanged in both A and B. For every other candidate $c \in \mathcal{S} \setminus \{c^*\}$, $0 < \rho'_c \leq \rho_c \leq 1$. For $c \in \mathcal{Y}^*$, this gives $1 - \rho'_c \geq 1 - \rho_c \geq 0$ and hence $(1-\rho'_c)^2 \geq (1-\rho_c)^2$, with strict inequality if $\rho'_c < \rho_c$. For $c \in \mathcal{S} \setminus \mathcal{Y}^*$, it gives $\rho'^2_c \leq \rho_c^2$. Summing the terms yields the claims. □

Sharpening therefore has two opposing effects: it increases the suppression of non-dominant true labels and decreases the residual confidence on spurious candidates. These opposing effects are the common mechanism behind the two failure modes.

A.2.5 Proof of Corollary 1

Proof. Since the sample is reliable, $\mathcal{S} = \mathcal{Y}^*$, so $\mathcal{S} \setminus \mathcal{Y}^* = \emptyset$ and $B \equiv 0$. Moreover, $c^* \in \mathcal{S} = \mathcal{Y}^*$. By Proposition 1, every supervision vector produced by the operator satisfies

$$\Delta = \sum_{c \in \mathcal{S}} (1 - \rho_c)^2. \tag{14}$$

In addition, $\boldsymbol{y}^{(0)} = \boldsymbol{s} = \boldsymbol{y}^*$, so $\Delta^{(0)} = 0$.

(i) If $k^* = 1$, then $\mathcal{S} = \{c^*\}$ and $\rho_{c^*} = 1$, so every output of the operator equals $\boldsymbol{s} = \boldsymbol{y}^*$. Hence, $\Delta^{(t)} = 0$ for all t.

(ii) If $k^* \geq 2$ and $\{p_c\}_{c \in \mathcal{S}}$ are not all equal, then some $c \in \mathcal{S}$ satisfies $p_c < \max_{c' \in \mathcal{S}} p_{c'}$, so $\rho_{\min} < 1$. Keeping only the term with minimal relative confidence in Eq. 14 gives $\Delta^{(1)} \geq (1 - \rho_{\min})^2 > 0 = \Delta^{(0)}$.

(iii) Let $\boldsymbol{p}'$ be a sharpening of $\boldsymbol{p}$. Since $\mathcal{Y}^* \setminus \{c^*\} = \mathcal{S} \setminus \{c^*\}$ and sharpening strictly decreases ρ_c for at least one c in this set, Lemma 2 gives $A(\boldsymbol{\rho}') > A(\boldsymbol{\rho})$, and therefore Δ strictly increases. If $\rho_c \to 0$ for all $c \in \mathcal{S} \setminus \{c^*\}$, each of these $k^* - 1$ terms in Eq. 14 converges to one, while the term of c^* remains zero. Hence, $\Delta \to k^* - 1$.

□

Part (i) shows that the increase in deviation of reliable samples in Fig. 1(a) cannot come from single-label samples, *i.e.*, Failure Mode I is a multi-label phenomenon. Part (ii) shows that, for reliable multi-label samples, already the first disambiguation step moves the supervision away from the ground-truth.

A.2.6 Deviation of ambiguous samples under sharpening

For an ambiguous sample, the initial deviation is $\Delta^{(0)} = \|\boldsymbol{y}^* - \boldsymbol{s}\|_2^2 = a$, since $\boldsymbol{s}$ and $\boldsymbol{y}^*$ differ exactly on the a spurious candidates.

Proposition 2. *Consider a sequence of predictions in which each prediction is a sharpening of the previous one, the dominant candidate is c^*, and $\rho_c \to 0$ for all $c \in \mathcal{S} \setminus \{c^*\}$. Along this sequence, A is non-decreasing and B is non-increasing, and the deviation of the Base-PLL supervision converges to*

$$\Delta \to \begin{cases} k^* - 1, & c^* \in \mathcal{Y}^*, \\ k^* + 1, & c^* \in \mathcal{S} \setminus \mathcal{Y}^*. \end{cases} \tag{15}$$

Proof. Monotonicity follows from Lemma 2. If $c^* \in \mathcal{Y}^*$, its term in A is zero, the remaining $k^* - 1$ terms of A converge to one, and all terms of B converge to zero. If $c^* \in \mathcal{S} \setminus \mathcal{Y}^*$, all k^* terms of A converge to one, the term of c^* in B equals one, and the remaining terms of B converge to zero. □

Comparing the limit with $\Delta^{(0)} = a$ explains the trajectories in Fig. 1(a) when the dominant candidate is a true label. If $a > k^* - 1$, purification outweighs suppression in the limit, and the supervision ends closer to the ground-truth than the candidate set. If $a < k^* - 1$, the limit exceeds the initial deviation, so the supervision eventually becomes worse than the unrefined candidate set, even if it improves at first while spurious candidates are purified. This is consistent with the decrease-then-increase trajectories of samples with high ground-truth cardinality. Note that Proposition 2 characterizes the limit and the direction of each term, but not the shape of the trajectory of Δ, which depends on the rates at which the relative confidences of true and spurious candidates decrease during training.

A.2.7 Recovered cardinality under sharpening

The following result supports the analysis of Failure Mode II in Sec. 3.3. Under Base-PLL, the supervision is $\boldsymbol{y}^{(t)} = \boldsymbol{\rho}$, so the recovered cardinality in Eq. 9 is $k = |\{c \in \mathcal{S} : \rho_c \geq \gamma\}|$ for a threshold $\gamma \in (0, 1)$.

Proposition 3. *For the Base-PLL supervision, the following statements hold:*

- *(i) $k \geq 1$ for every prediction, and a true label $c \in \mathcal{Y}^*$ is recovered if and only if $p_c \geq \gamma \max_{c' \in \mathcal{S}} p_{c'}$.*
- *(ii) If $\boldsymbol{p}'$ is a sharpening of $\boldsymbol{p}$, then $k(\boldsymbol{\rho}') \leq k(\boldsymbol{\rho})$.*
- *(iii) Along a sequence of sharpenings with $\rho_c \to 0$ for all $c \in \mathcal{S} \setminus \{c^*\}$, the recovered cardinality eventually equals one. In particular, for every sample with $k^* \geq 2$, at least $k^* - 1$ true labels are eventually lost.*

Proof. (i) Since $\rho_{c^*} = 1 \geq \gamma$, the dominant candidate is always recovered. By Eq. 13, $\rho_c \geq \gamma$ is equivalent to $p_c \geq \gamma \max_{c' \in \mathcal{S}} p_{c'}$. (ii) Under sharpening, ρ_{c^*} remains one and $\rho'_c \leq \rho_c$ for all other candidates, so $\mathbb{I}[\rho'_c \geq \gamma] \leq \mathbb{I}[\rho_c \geq \gamma]$ for every $c \in \mathcal{S}$. (iii) Since $\gamma > 0$, every non-dominant ρ_c eventually falls below γ, so only c^* remains recovered. At most one true label is therefore retained. □

Part (i) shows that whether a valid diagnosis is retained depends on its confidence relative to the dominant candidate rather than on its own confidence, and part (ii) shows that sharpening can only reduce the recovered cardinality. Together with Lemma 1(iii), this explains why nothing in Base-PLL counteracts the cardinality collapse observed in Fig. 1.

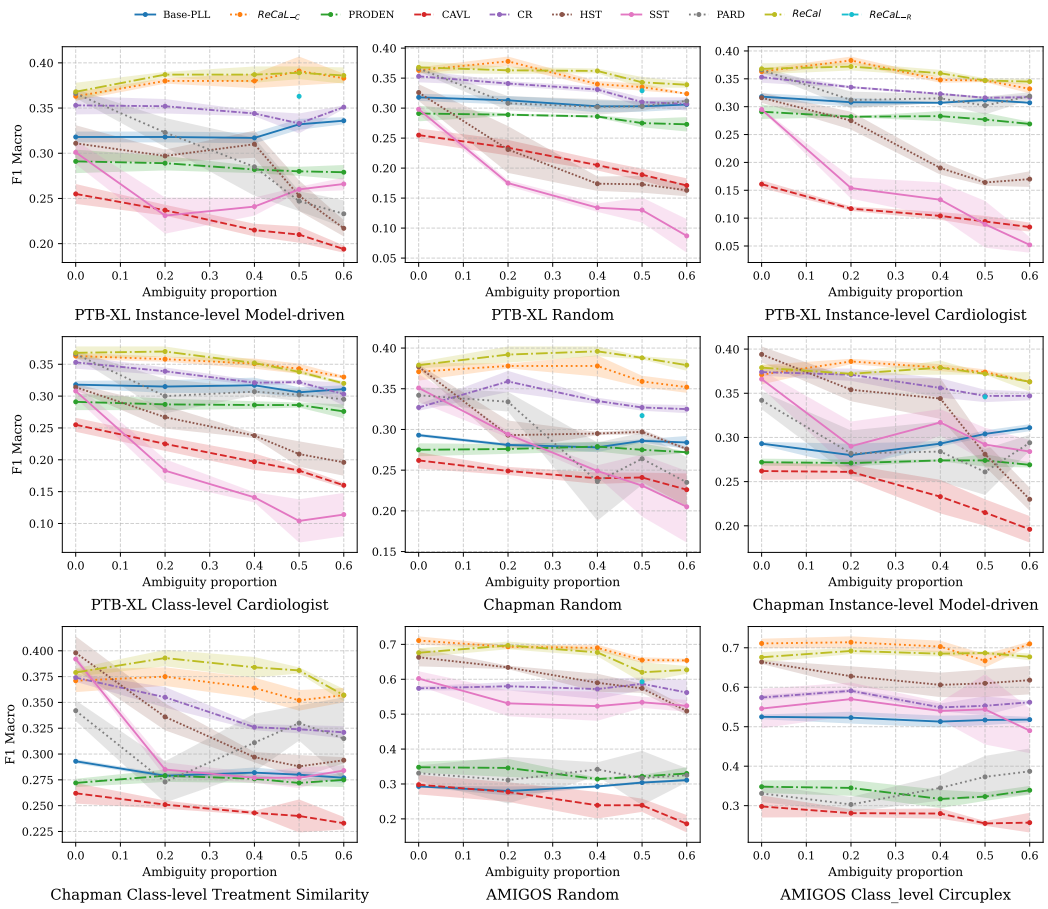


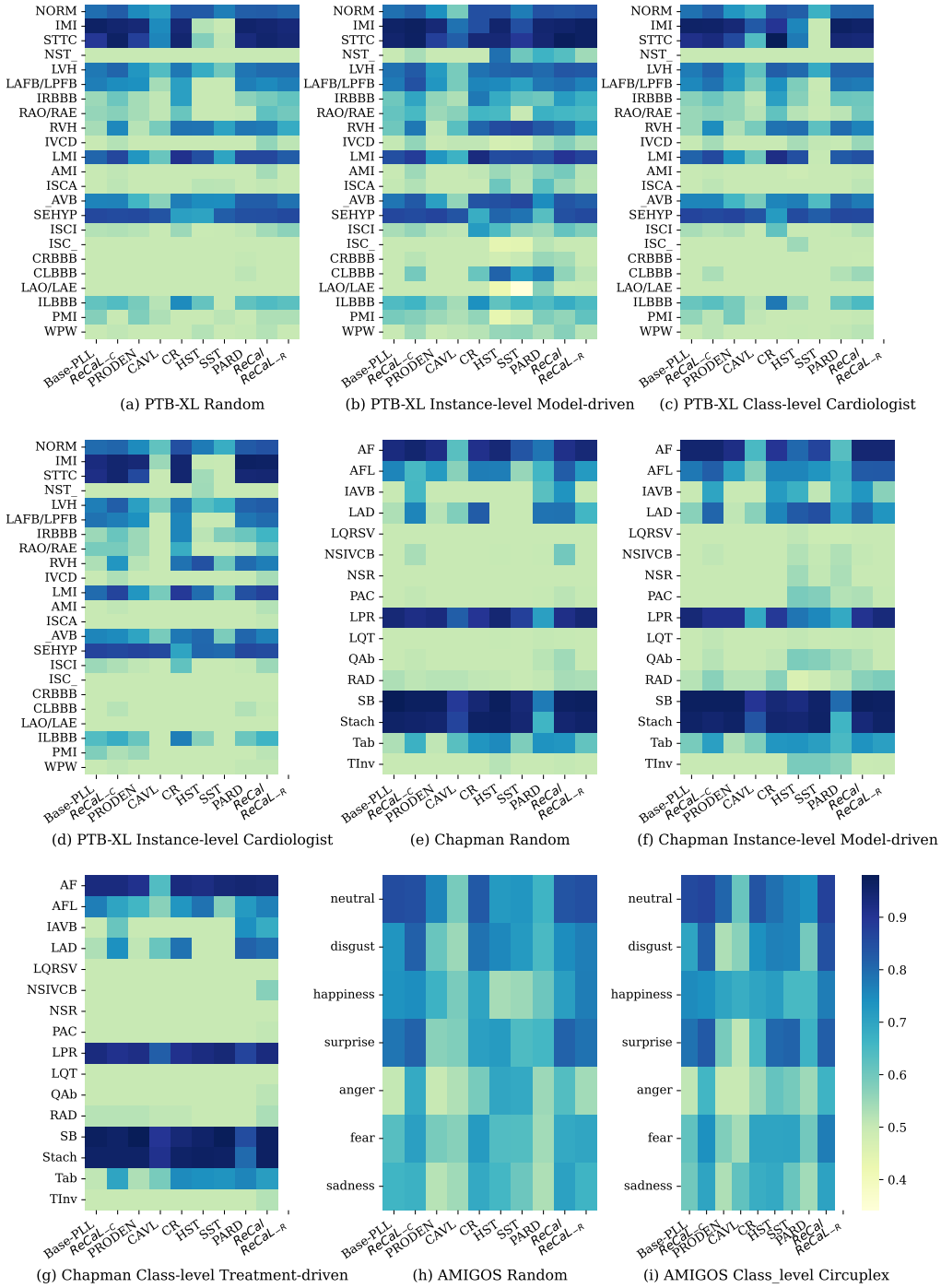
Figure 4: Sensitivity to the proportion of partially labelled training samples on PTB-XL and Chapman.

A.3 Baselines and Implementation Details

We compare ReCal with representative identification-based PLL methods, including PRODEN (Lv et al., 2020), CAVL (Zhang et al., 2021), CR (Wu et al., 2022), HST (Chen et al., 2024), SST (Chen et al., 2022) and PARD (Hang & Zhang, 2023). These methods cover different strategies for label disambiguation including model prediction confidence, and latent variable based estimation of the candidate label's confidence. We additionally include Base-PLL as a direct reference implementation of the confidence-based disambiguation analyzed in Sec. 3.2 and Sec. 3.3. We use the same data preprocessing, ECG representation, and backbone architecture for all baselines to ensure a fair comparison. We use a 1D ResNet18 as the shared backbone for all PLL methods, consisting of four stages with two bottleneck blocks per stage, followed by adaptive average pooling and a fully connected output layer. All methods are trained for 30 epochs using a batch size of 32. Each baseline follows its original network and parameter configuration as closely as possible. All experiments are implemented in PyTorch and conducted on an NVIDIA RTX 2080 GPU using three random seeds.

A.4 Additional Experiments

Sensitivity to partial-label proportion. In real-world applications, the proportion of samples associated with ambiguous candidate labels is unknown. We therefore vary the proportion of partially labelled samples while keeping the remaining samples reliable, and report the resulting macro-F1 scores in Fig. 4. As the proportion of partially labelled samples increases, the classification performance of the PLL methods generally degrades, reflecting the increasing difficulty of learning from less reliable supervision. Interestingly, the impact of increasing ambiguity varies across ambiguity-

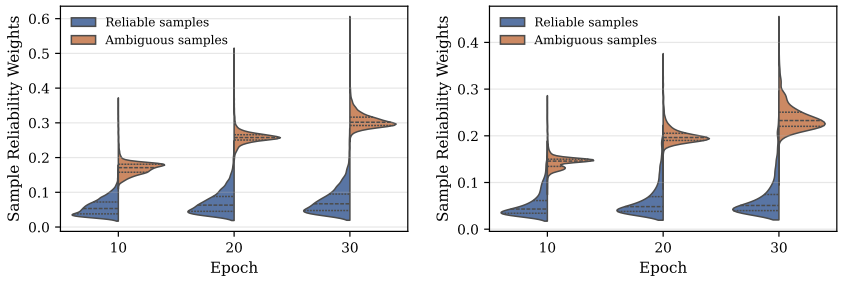


(a) PTB-XL Random (b) PTB-XL Instance-level Model-driven (c) PTB-XL Class-level Cardiologist

(d) PTB-XL Instance-level Cardiologist (e) Chapman Random (f) Chapman Instance-level Model-driven

(g) Chapman Class-level Treatment-driven (h) AMIGOS Random (i) AMIGOS Class_level Circuplex

Figure 5: Class-specific performance of baselines over AUCROC metric.



(a) PTB-XL dataset. (b) Chapman dataset

Figure 6: Distribution of the reliability weights.

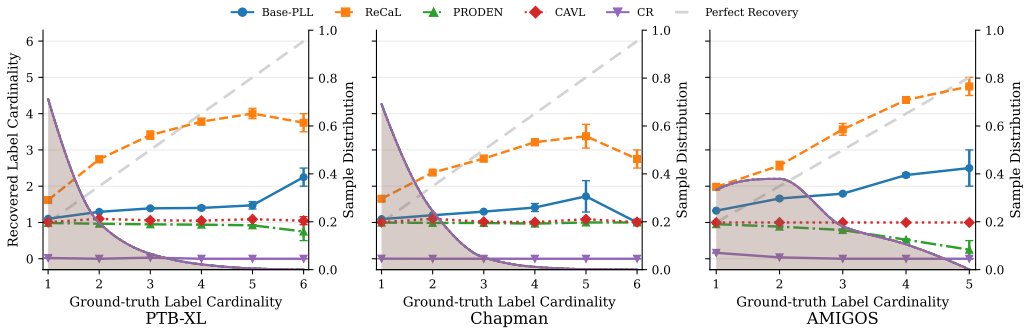


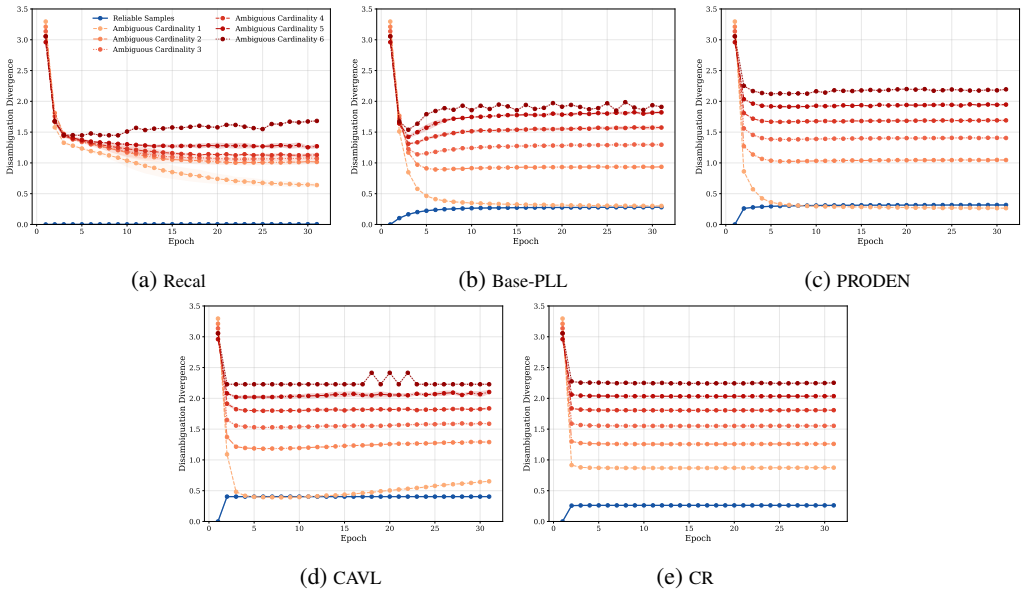
Figure 7: Recovered cardinality of the confidence-based baselines for PTB-XL, Chapman, and AMIGOS dataset.

generation strategies and methods. While some baselines maintain relatively stable performance under certain ambiguity types, they experience substantial performance degradation under others, indicating that different forms of ambiguity challenge disambiguation approaches differently. In contrast, ReCal maintains consistently higher performance across the evaluated proportions and ambiguity-generation strategies, suggesting that its sample-adaptive disambiguation is more robust to changes in both the amount and type of ambiguous supervision. This stability further supports the effectiveness and generalizability of ReCal across different ambiguity conditions.

Class specific performance Fig. 5 reports the class-specific AUCROC performance of the baselines for Chapman and PTB-XL datasets sorted by sample count on y-axis. As we see, different methods perform differently across different classes with most of them experience more performance degradation on minor classes. Based on the results, CR, and ReCal show an overall better performance across most of the classes compared to other baselines.

Visualization of reliability weights distribution. In this study, we visualize the distribution of the learned reliability weights for ambiguous and reliable sample during training on PTB-XL and Chapman datasets with random ambiguity. As we can observe in Fig. 6, the distribution of this learned weights are quite separable for ambiguous and reliable samples throughout training. This explains why the reliability adaptive weight can successfully separate these samples.

Recovered cardinality across PLL baselines. To examine if the cardinality under-recovery is specific to Base-PLL or is observed across other confidence-based disambiguation methods, we compare the recovered label cardinality of ReCal with several PLL baselines. Fig. 7 reports the recovered cardinality as a function of the ground-truth cardinality on PTB-XL, Chapman, and AMIGOS under random ambiguity. Most baselines increasingly underestimate the number of labels as the ground-truth cardinality increases, indicating a tendency to disambiguate multi-label samples toward fewer labels. In contrast, ReCal retains substantially more labels for samples with higher ground-truth cardinality, reducing the gap from the perfect-recovery line across the evaluated



(a) Recal (b) Base-PLL (c) PRODEN

(d) CAVL (e) CR

Figure 8: Disambiguation deviation of the confidence-based baselines on PTB-XL dataset over random ambiguity.

datasets. These results suggest that the under-recovery of multi-label samples is not limited to a particular implementation and further support the need to account for instance-dependent cardinality during disambiguation.

Disambiguation deviation across PLL baselines. To further examine the evolution of candidate-label disambiguation in training, we compare the disambiguation deviation of ReCal with several confidence-based PLL baselines on PTB-XL under random ambiguity. Fig. 8 reports the deviation for reliable samples and ambiguous samples grouped by ground-truth label cardinality. The methods exhibit different disambiguation trajectories across the training process, with the deviation of ambiguous samples depending on their ground-truth cardinality. ReCal maintains near-zero deviation for reliable samples while progressively refining ambiguous samples, whereas the other baselines generally introduce non-negligible deviation for reliable samples. The results provide further evidence that both the reliability of the candidate-label supervision and the number of underlying labels influence the behavior of confidence-based disambiguation.