TAPS: TREE-ADAPTIVE PREDICTION SETS VIA LCA-WEIGHTED NONCONFORMITY SCORING FOR HIERARCHICAL MULTI-LABEL CLINICAL CODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction (CP) provides distribution-free, finite-sample coverage guarantees. However, applying CP to hierarchical multi-label classification leads to a combinatorial explosion over 2^K possible label subsets. Recent parallel work (Mortier et al., 2026; den Hengst et al., 2025) addresses this by allowing prediction sets to include internal nodes of a taxonomy. We identify a complementary gap: prior methods either use flat leaf-level conformity scores or incorporate hierarchy without an explicit abstraction-cost penalty in the conformity score. We introduce **Tree-Adaptive Prediction Sets (TAPS)**, a CP algorithm that directly integrates hierarchical structure into the calibration score. TAPS employs a *specificity-regularized Lowest-Common-Ancestor (LCA) score*, which penalizes each tree node based on its uncertainty and a height-based abstraction cost (weighted by λ). Crucially, this identical score governs both calibration and inference. We prove that the TAPS score is dominated by the flat score, ensuring calibration thresholds remain no larger than those of flat CP at the same α. We also prove TAPS guarantees marginal semantic coverage $\mathbb{P}(Y \in_{\text{sem}} C(X)) \geq 1-\alpha$ under exchangeability, with calibration and inference provably evaluating the exact same event. The parameter λ controls an empirical compactness–specificity trade-off while conformal validity is preserved after independent λ selection, requiring a non-degeneracy screen to prevent trivial root-collapse. On MIMIC-IV automated ICD-10 coding (QLoRA-tuned Phi-3-mini, full training set), TAPS attains near-nominal empirical semantic joint coverage, comparable to Flat LAC, while producing prediction sets roughly 48% smaller than flat CP. It achieves this by summarising uncertainty via ancestor nodes, trading off exact leaf-level specificity. We evaluate against eleven baselines spanning flat, hierarchical, and multi-label conformal methods, and identify a common failure mode, which we call the 'free root' pathology, that causes every hierarchy-aware baseline we test lacking score-level integration to exhibit severe undercoverage, root-collapse, or substantially inflated prediction sets. We further validate TAPS under the same calibration and evaluation framework on a second, structurally distinct domain, the arXiv category taxonomy (depth 3, vs. ICD-10's depth 6), providing evidence that both the compactness gains and the baseline failure modes are not specific to clinical coding.

1 INTRODUCTION

Clinical NLP models for automated ICD coding routinely handle critical billing and audit workflows (Mullenbach et al., 2018; Huang et al., 2022). However, fine-tuned transformer models exhibit systematic *overconfidence* (Guo et al., 2017). Despite achieving strong F1 performance, they lack a formal mechanism to reliably quantify uncertainty. Conformal prediction (CP) (Vovk et al., 2005; Angelopoulos & Bates, 2023) addresses this by constructing prediction sets $C(X)$ that satisfy $\mathbb{P}(Y \in C(X)) \geq 1-\alpha$ under *exchangeability alone*. This framework is very mature for standard single-label classification problems (Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021).

However, ICD-10 coding presents a *hierarchical multi-label* challenge: for a single discharge summary, multiple codes must be selected from a tree structure spanning seven levels of depth and

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.

comprising 70,000 nodes. Applying flat CP to this setting forces prediction sets to inflate to nearly 50 codes simply to maintain coverage (as seen in our Table 2 results for Flat APS and RAPS), rendering the model output clinically useless.

Recent research aims to address this limitation. Mortier et al. (2026) extended split CP to hierarchical classification by restricting output prediction sets to internal tree nodes (CRSVP). Concurrently, den Hengst et al. (2025) proposed Hierarchical Conformal Classification (HCC), which utilizes a DAG structure to balance specificity and compactness. While these works represent significant progress, neither incorporates structural abstraction costs into calibration itself (Section 2). We integrate the semantic ICD hierarchy deeper into the pipeline, directly within the nonconformity score function. This score-level integration enables more precise thresholds, smaller prediction sets, and clinically more interpretable measures of uncertainty.

Contributions. Our contributions are as follows. **(1)** To our knowledge, we introduce the first conformal construction to embed an ontology-derived hierarchical abstraction penalty directly into the nonconformity score for multi-label prediction sets (Section 4). Concretely, this is a specificity-regularized LCA score that penalizes each node's uncertainty by a height-based abstraction cost λ, and crucially, the same score governs both calibration and inference, so the two provably evaluate the identical event. **(2)** We prove this construction's theoretical guarantees: a dominance theorem (Theorem 1) showing our calibrated threshold never exceeds the flat score's, and a distribution-free semantic coverage guarantee (Theorem 4) alongside a tight, Bonferroni-free joint-coverage bound (Corollary 2) for the max-score protocol used throughout our experiments. **(3)** We characterize the resulting efficiency-specificity frontier (Lemma 5): at near-nominal empirical semantic joint coverage, TAPS produces prediction sets $\approx 48\%$ smaller than flat conformal prediction on MIMIC-IV ICD-10 coding, governed by a single λ selected via a calibration-side non-degeneracy screen that rejects root-collapsing operating points (Remark 2). **(4)** Finally, we validate extensively against eleven baselines spanning flat, multi-label, and hierarchical conformal methods, and identify a common failure mode we call the "free root" pathology, present in every hierarchy-aware baseline we test that lacks TAPS's score-level integration. A three-arm ablation confirms both design choices are independently necessary, and we replicate these findings under a closely matched calibration and baseline protocol on arXiv's category taxonomy, a second, structurally distinct benchmark: a maintained ontology of depth three versus ICD-10's depth six, built from genuine multi-label cross-listings rather than a learned clustering. This shows the compactness gains and baseline failure modes are not artifacts of clinical text or ICD-10's structure specifically (Appendix P).

2 Related Work

We situate TAPS against three bodies of related work: general-purpose conformal prediction, hierarchical multi-label classification, and the recent wave of hierarchical conformal prediction methods most directly comparable to our own.

The foundation of our approach lies in split conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005), which generates prediction sets carrying finite-sample guarantees by evaluating a held-out calibration set. Within this paradigm, LAC (Sadinle et al., 2019) directly minimises expected set size, while APS (Romano et al., 2020) focuses on conditional coverage, and RAPS (Angelopoulos et al., 2021) introduces regularization to manage large candidate sets. Bates et al. (2021) expand these concepts to general risk-controlling prediction sets, and Vázquez & Facelli (2022) provide a comprehensive survey of CP applications within clinical medicine.

In parallel, hierarchical multi-label classification (HMC) has long structured labels according to a predefined taxonomy (Silla & Freitas, 2011; Kosmopoulos et al., 2015). Early theoretical work by Bi & Kwok (2015) explored Bayes-optimal HMC using tree-distance loss, establishing a precursor to our LCA-scoring approach. In the specific domain of automated ICD coding, Mullenbach et al. (2018) formalised the MIMIC benchmark, and recently, Huang et al. (2022) achieved state-of-the-art accuracy using pretrained language models, though they incorporate no uncertainty quantification.

Bridging these fields, a new wave of hierarchical conformal prediction methods has emerged, against which we situate TAPS in Table 1. CRSVP (Mortier et al., 2026) restricts outputs to internal nodes to handle hierarchical single-label classification; however, it injects hierarchy only at inference rather than calibration, and remains single-label. HCC (den Hengst et al., 2025) propagates leaf probabilities

upward to obtain hierarchy-dependent node scores for calibration, but its conformity score contains no explicit abstraction-cost penalty; compactness is instead controlled through a separate downstream NOL-cover optimization. Tree-CP (Tyagi & Guo, 2023) tackles multi-label CP through hierarchical testing, but relies on data-derived clustering instead of a structured domain ontology. Finally, THCM-CAL (Zhang et al., 2025b) applies flat, per-code CP calibration to MIMIC-IV ICD-9 coding, entirely ignoring the hierarchical structure during both calibration and output.

Table 1: Comparison of TAPS with related hierarchical CP methods. **Score-level**: explicit abstraction penalty in the nonconformity score. **Output-level**: hierarchy in prediction set structure. **Multi-label**: multiple simultaneous true labels supported. **Ontology**: uses domain ontology (not learned clustering).

Method	Score	Output	Multi-label	Ontology	Venue
LAC (Sadinle et al., 2019)	×	×	×	×	JASA 2019
APS (Romano et al., 2020)	×	×	×	×	NeurIPS 2020
RAPS (Angelopoulos et al., 2021)	×	×	×	×	ICLR 2021
Tree-CP (Tyagi & Guo, 2023)	×	✓	✓	×	COPA 2023
CRSVP (Mortier et al., 2026)	×	✓	×	✓	AISTATS 2026
HCC (den Hengst et al., 2025)	×	✓	×[1]	✓	arXiv 2025
THCM-CAL (Zhang et al., 2025b)	×	×	✓	×	EMNLP 2025 Findings
BR-MLCP (Lambrou & Papadopoulos, 2016)	×	×	✓	×	COPA 2016
Fisch et al. (Fisch et al., 2022)	×	×	✓	×	ICML 2022
CGD (Cauchois et al., 2021)	×	×	✓	×	JMLR 2021
CRC-Hierarchical (Angelopoulos et al., 2024)	×	✓	✓	✓	ICLR 2024
CHiQPM (Norrenbrock et al., 2025)	✓	✓	×	×	NeurIPS 2025
HSC (Goren et al., 2024)	×	✓(single-node)	×	✓	NeurIPS 2024
Class-Sim. CP (Fargion et al., 2026)	✓	×	×	×	ICML 2026
TAPS (Ours)	✓	✓	✓	✓	—

This score-level integration under a domain ontology is what most sharply distinguishes TAPS from the concurrent literature. Norrenbrock et al. (2025) also integrate hierarchy at the score level, but for single-label classification via learned feature-activation paths rather than ontology-based structural distance; the two approaches are complementary rather than overlapping.

Goren et al. (2024) propose Hierarchical Selective Classification (HSC), reducing specificity during single-label inference to satisfy an accuracy target. Although conceptually similar to TAPS's height-based abstraction, HSC lacks a distribution-free split-conformal guarantee for multi-label sets, and its unregularized ascent mechanism remains vulnerable to trivial root-collapse (Appendix I). Fargion et al. (2026) independently confirm this identical pathology when adapting HSC's climbing rule to flat superclass partitions, which "often leads to inclusion of all classes". Their subsequent contribution introduces a score-level penalty for out-of-group errors over a flat two-level partition; however, this approach remains confined to that flat structure, without utilizing tree depth, ancestral relationships, or a domain ontology.

3 Problem Formulation

Let $\mathcal{T} = (V, E)$ be the ICD-10-CM hierarchy defined as a rooted tree, comprising all nodes V and the set of billable leaf codes $\mathcal{L} \subset V$ (where $|\mathcal{L}| = K$). We denote strict ancestors as $\mathrm{Anc}(v)$, strict descendants as $\mathrm{Desc}(v)$, and the lowest common ancestor of u and v as $\mathrm{LCA}(u, v)$. The distance from the root specifies $\mathrm{depth}(v)$, with $\mathrm{depth}(\mathrm{root}) = 0$; $\mathrm{height}(v)$ is the length of the longest downward path from v to any leaf in $\mathrm{Desc}(v) \cup \{v\}$, ensuring $\mathrm{height}(v) = 0$ for every leaf unconditionally, regardless of depth. We treat the full tree depth as $d_{\max} = 7$, matching ICD-10-CM's up-to-seven-character structure (chapter, block, category, subcategories); our top-50 code subset is shallower, with an observed maximum depth and height of 6.

Multi-label classification. An input $X \in \mathcal{X}$ (a clinical note) receives true labels $Y \subseteq \mathcal{L}$. The model $f : \mathcal{X} \to [0, 1]^{|V|}$ outputs an *independent sigmoid probability* $\hat{p}_v(x)$ *for every node* $v \in V$, leaves and internal nodes alike, trained via *hierarchical label propagation*: given true leaf set Y, the binary

[1] HCC's formalism accommodates DAG-induced label multiplicity arising when a single ground-truth leaf has multiple ancestors within one candidate cover, but its evaluation is restricted to single-label tasks (DBpedia14, ImageNet1K, GTZAN); TAPS targets instances with multiple independent true labels.

target for node v is 1 if $v \in \{y\} \cup \mathrm{Anc}(y)$ for any $y \in Y$, else 0, with each node supervised by an independent binary-cross-entropy loss. Consequently, $\hat{p}_v(x)$ directly estimates the probability that the correct true label lies within the subtree rooted at node v. This approach avoids the failure modes associated with aggregating leaf scores (Appendix E).

Semantic coverage.

Definition 1 (Semantic Coverage)**.** Prediction set $C \subseteq V$ provides *semantic coverage* for true label $y^* \in \mathcal{L}$, written $y^* \in_{\mathrm{sem}} C$, iff: $\exists\, c \in C :\ c \in \{y^*\} \cup \mathrm{Anc}(y^*)$.

Semantic coverage is strictly weaker than exact coverage, since an ancestor prediction already counts as success; the clinical rationale is that the correct parent block still retains coarse diagnostic localization, though it does change the specific target. To guard against this relaxation buying excessive ambiguity, we track the LCA distance (LCAD) metric throughout (Appendix D).

4 Methodology: TAPS

4.1 Specificity-Regularized LCA Score

The central innovation is embedding the ICD hierarchy into the *calibration score*, not only the output structure. We score each node by its uncertainty plus an *abstraction cost* that grows toward the root.

Definition 2 (Height-adjusted node score)**.** Fix a regularization strength $\lambda \geq 0$. For node $v \in V$:

$$g_v(x) \;=\; \underbrace{\big(1 - \hat{p}_v(x)\big)}_{\text{uncertainty}} \;+\; \lambda \cdot \underbrace{\mathrm{height}(v)}_{\text{abstraction cost}} \tag{1}$$

The penalty $\mathrm{height}(v)$ is 0 for every leaf (by definition of height) and grows toward the root, so shallower (more general) nodes must be substantially more confident to attain a low score.

Remark 1 (Height, not depth). An earlier penalty formulation used $d_{\max} - \mathrm{depth}(v)$. This sets the penalty to zero exclusively at the tree's global maximum depth, artificially penalizing shallow leaves in non-depth-regular trees and breaking the dominance property (Theorem 1). By contrast, measuring $\mathrm{height}(v)$ down to the deepest leaf in a subtree guarantees $\mathrm{height}(\text{leaf}) = 0$ unconditionally for every leaf, cleanly restoring dominance. We provide the full derivation in Appendix C.

Definition 3 (LCA Nonconformity Score)**.** For input x and true label $y^* \in \mathcal{L}$:

$$s_{\mathrm{LCA}}(x, y^*) \;=\; \min_{v \,\in\, \{y^*\} \cup \mathrm{Anc}(y^*)} g_v(x) \tag{2}$$

i.e. the score of the *least-penalized ancestor on the true path*.

In plain terms, s_{LCA} finds the cheapest way to stay confidently correct about y^*, weighing raw confidence against the cost of abstracting further up the tree. The two extremes are intuitive: at λ=0, s_{LCA} reduces to the unregularized path-max, $1 - \max_v \hat{p}_v$; as λ grows, abstraction gets expensive and the minimizing node gets pulled back toward y^*, so the method behaves more and more like a leaf-level score. This is where TAPS departs from prior work (Section 2): it builds the abstraction penalty into the calibration score itself, and reuses that same score for prediction-set construction. Modifying the calibration score, not just the output, is the key architectural move that makes Theorem 1 possible.

4.2 Calibration and Set Construction

Definition 4 (TAPS Prediction Set)**.** For input x and threshold $\hat{q}$, let the *passing set* be $P(x) = \{v \in V : g_v(x) \leq \hat{q}\}$ using the *same* g_v and λ as calibration. Then

$$C_{\mathrm{TAPS}}(x) \;=\; \big\{\, v \in P(x) \;:\; \text{no proper ancestor of } v \text{ lies in } P(x) \,\big\}, \tag{3}$$

the *topmost* passing nodes (the maximal elements of $P(x)$ under the ancestor order). $C_{\mathrm{TAPS}}(x)$ is an antichain (no element is an ancestor of another).

Since g_v lacks monotonicity along a path, the passing set $P(x)$ does not strictly form an upward-closed prefix. Merely emitting the deepest passing nodes risks missing y^* entirely. Conversely, emitting the topmost passing node on each chain ensures that if any node along the true path passes,

Algorithm 1 TAPS: Tree-Adaptive Prediction Set Construction

Require: Model f, input x, threshold $\hat{q}$, tree $\mathcal{T}$, parameter λ
Ensure: Prediction set $C_{\text{TAPS}}(x) \subseteq V$

```
C ← ∅;  Queue ← {root}
while Queue ≠ ∅ do
    v ← dequeue(Queue)
    Compute g_v(x) using Eq. (1)
    if g_v(x) ≤ q̂ then
        C ← C ∪ {v}                          ▷ Topmost passing node on this branch
    else
        enqueue(children(v), Queue)          ▷ Descend to seek a passing node
    end if
end while
return C
```

the set successfully captures an ancestor-or-self of y^* (Theorem 4). The height-based abstraction cost embedded in g_v naturally prevents this topmost rule from degenerating into trivial root-collapse.

Having specified how a threshold $\hat{q}$ yields a prediction set, we now turn to computing $\hat{q}$ itself. Given calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, define the per-example score $S_i = \max_{y^* \in Y_i} s_{\text{LCA}}(X_i, y^*)$ and set:

$$\hat{q} = \text{Quantile}\left(\{S_i\}_{i=1}^n;\ \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right) \tag{4}$$

The *maximum* over true labels calibrates the threshold to the *hardest* (highest-score) code per patient, so that the worst-case true label is covered. Because S_i is a fixed measurable function of (X_i, Y_i) for fixed λ, the sequence $(S_1, \ldots, S_{n+1})$ inherits exchangeability from the data (Assumption 1).

Remark 2 (Selecting λ without breaking validity). Selecting λ using any statistic computed on the same calibration data subsequently used to compute $\hat{q}$ induces post-selection bias: the calibration scores under the selected λ are systematically favorable on that same data, and the standard conformal guarantee no longer applies to the resulting threshold. To avoid this, we partition the calibration set into two disjoint subsets, D_λ and D_{conf}, following the same split ratio used by RAPS (Angelopoulos et al., 2021) for its own regularization hyperparameters. We run the two-stage non-degeneracy selection rule—first rejecting any λ yielding $> 5\%$ root-only predictions, then selecting the parameter yielding the smallest average set size among survivors—using only D_λ (40% of the calibration set). For each candidate λ, a provisional threshold $\hat{q}_\lambda$ is computed on D_λ using Eq. (4). The non-degeneracy rate and average prediction-set size are then evaluated on D_λ using $(\lambda, \hat{q}_\lambda)$. After selecting $\hat{\lambda}$, this provisional threshold is discarded, and the final $\hat{q}$ is computed independently on D_{conf}. The selected λ is then frozen, and $\hat{q}$ is computed via Eq. (4) using only D_{conf} (the remaining 60%), which was never used in selecting λ. Because D_{conf} is untouched by the λ-selection procedure, the exchangeability argument underlying Theorem 4 and Corollary 2 applies to it directly, exactly as in the standard split-conformal setting.

4.3 Theoretical Guarantees

Having defined the score and its calibration, we state our two headline guarantees here; the exchangeability assumptions, intermediate results (Proposition 3, Theorem 4), Lemma 5, and all proofs are in Appendix A.

Theorem 1 (Dominance of the LCA Score)**.** *For all $(x, y^*) \in \mathcal{X} \times \mathcal{L}$ and any $\lambda \geq 0$:*

$$s_{\text{LCA}}(x, y^*) \ \leq\ s_{\text{flat}}(x, y^*) \quad \textit{pointwise,} \tag{5}$$

where $s_{\text{flat}}(x, y^*) = 1 - \hat{p}_{y^*}(x)$. Consequently, at any error level α, $\hat{q}_{\text{LCA}} \leq \hat{q}_{\text{flat}}$ (a weak quantile dominance), with the inequality potentially strict when ancestor abstraction lowers the score distribution around the relevant calibration quantile.

For the multi-label joint protocol used in Section 5, define $S_{\text{flat}}(x, Y) = \max_{y \in Y} s_{\text{flat}}(x, y)$. The pointwise dominance $s_{\text{LCA}} \leq s_{\text{flat}}$ implies $S_{\text{LCA}} \leq S_{\text{flat}}$, and therefore $\hat{q}_{\text{LCA}} \leq \hat{q}_{\text{flat}}$ under the same max-score calibration protocol.

Corollary 2 (Joint Coverage of the Max-Score Protocol). *Under standard conformal exchangeability (Appendix A), let* $S_i = \max_{y \in Y_i} s_{\mathrm{LCA}}(X_i, y)$ *and let* $\hat{q}$ *be its* $\lceil (n+1)(1-\alpha) \rceil / n$ *empirical quantile over the* $n = |D_{conf}|$ *calibration examples (Eq. (4)), the protocol used throughout Section 5. Then*

$$1 - \alpha \;\le\; \mathbb{P}\Big(\forall\, y^* \in Y_{n+1} :\; y^* \in_{\mathrm{sem}} C_{\mathrm{TAPS}}(X_{n+1})\Big) \;\le\; 1 - \alpha + \frac{1}{n+1}. \tag{6}$$

Corollary 2 shows the max-score protocol controls *joint* coverage directly, with the same tight bound as the single-label case (no Bonferroni correction, and no loss of tightness from splitting α across m labels). The lower bound $1 - \alpha$ holds under exchangeability alone. The upper bound $1 - \alpha + \frac{1}{n+1}$ additionally requires that the conformity scores $S_1, \ldots, S_{n+1}$ are almost surely distinct; when ties occur, the quantile may be conservative, so coverage can exceed $1 - \alpha + \frac{1}{n+1}$ unless a randomized tie-breaking rule (e.g. uniform perturbation of tied scores) is applied.

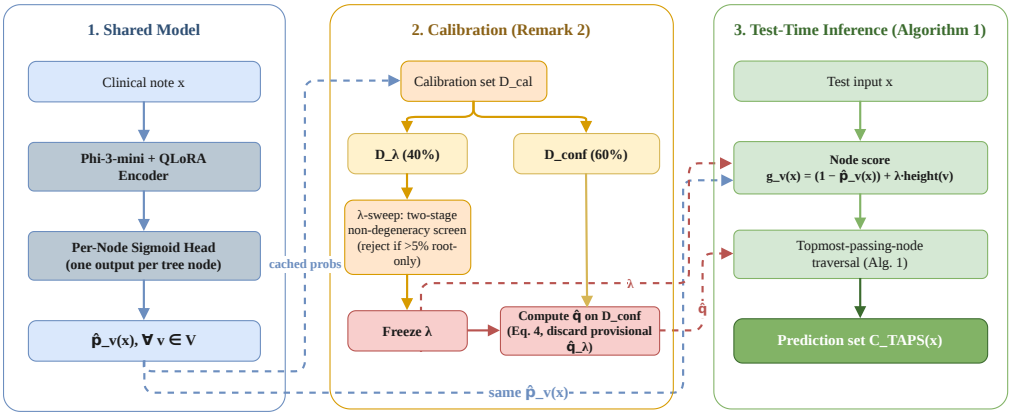


Figure 1: TAPS pipeline. The same per-node probabilities $\hat{p}_v(x)$ are computed once and reused for both calibration and inference; the same score $g_v(x)$ governs D_λ selection, $\hat{q}$ calibration, and the test-time traversal, so calibration and inference provably evaluate the identical event (Remark 2).

5 Experiments

5.1 Experimental Setup

We evaluate TAPS and the extended baselines on a real-world hierarchical clinical coding task using **MIMIC-IV** (Johnson et al., 2023) (discharge summaries linked to ICD-10-CM codes). Following Mullenbach et al. (2018); Huang et al. (2022), we restrict to the top-50 most frequent ICD-10-CM codes (33.9%, or 1,172,228 of 3,455,747, of all diagnosis-code instances), split 80/10/10 train/calibration/test by `subject_id` (no patient leakage), and truncate to 512 tokens, yielding 91,959 training, 11,402 calibration, and 11,272 test admissions; the calibration split is further divided 40/60 into D_λ and D_{conf} per Remark 2.

Model architecture:

- **Base encoder:** `Phi-3-mini-3.8B` (Abdin et al., 2024); causal head removed; final hidden state ($h \in \mathbb{R}^{3072}$) as document representation.
- **Classification head:** Linear $W \in \mathbb{R}^{3072 \times |V|}$ with sigmoid activation, one output per tree node; per-node binary cross-entropy with hierarchical label propagation.
- **QLoRA (Dettmers et al., 2023):** 4-bit NF4 quantisation; LoRA (Hu et al., 2022) adapters (rank r=16, α=32); only adapters + head trained (≈0.25% of parameters, 9,437,184 of 3,732,016,128).
- **Hardware:** single NVIDIA L4 GPU (Google Cloud), ≈ 20 hours for 2 epochs.

Baselines (all sharing the same encoder and cached per-node probabilities; only calibration/inference differ): (1) **LAC** (Sadinle et al., 2019): s=1−$\hat{p}_{y^*}$ over leaves. (2) **APS** (Romano et al., 2020): cumulative score over leaves. (3) **RAPS** (Angelopoulos et al., 2021): λ=0.01, k_{reg}=5. (4) **CRSVP** (Mortier

et al., 2026): single-node prediction via greedy top-down descent using $\hat{p}(X)$, stopping at the deepest node whose flat score satisfies the LAC-calibrated $\hat{q}$. (5) **HCC** (den Hengst et al., 2025): evaluated under both its native max-based (marginal) score and a min-based joint adaptation. (6) **BR-MLCP** (Lambrou & Papadopoulos, 2016): independent flat per-label calibration without joint correction. (7) **Fisch et al.** (Fisch et al., 2022): global top-k prediction set over leaves. (8) **CRC** (hierarchical adaptation; Angelopoulos et al., 2024): risk-controlling prediction set collapsed to a minimal covering antichain. (9) **DSB (Depth-Stratified Bonferroni)**[2]: depth-stratified quantile thresholds collapsed to a covering antichain. **Tree-CP** (Tyagi & Guo, 2023) and **CSP** (Zhang et al., 2025a) are deferred to Appendix I due to space.

Evaluation metrics: (1) **Joint coverage:** all of a patient's labels covered simultaneously (Corollary 2); also report marginal coverage. (2) **Average Set Size:** $\frac{1}{n}\sum_i |C(X_i)|$ (lower is better). (3) **LCAD:** for a prediction set $C(x)$ and true label y^*, define the per-pair distance $d_{\text{LCA}}(c, y^*) = \text{depth}(y^*) - \text{depth}(\text{LCA}(c, y^*))$, which equals 0 for an exact match and is strictly positive for any ancestor prediction, scaled by how many levels up the tree the prediction backed off. Then $d_{\text{LCA}}(C(x), y^*) = \min_{c \in C(x)} d_{\text{LCA}}(c, y^*)$ is the distance from the closest predicted element to the true label, and $\text{LCAD} = \frac{1}{n}\sum_{i=1}^{n} d_{\text{LCA}}(C(X_i), y_i^*)$ its mean over the test set (averaged over all true labels $y^* \in Y_i$ for multi-label instances); a guard metric for the semantic-coverage relaxation, lower is more specific.

5.2 Results

Table 2: Results at α=0.10 (MIMIC-IV top-50 ICD codes, full training set, 2 epochs), aggregated over 3 independent seeds (mean $\pm$ std). $\hat{q}$ calibrated via per-patient max score (Eq. (4)), controlling **joint** coverage. TAPS selects λ independently per seed via the two-stage non-degeneracy rule (Remark 2); all three seeds select λ=0.10.

Method	Semantic Joint Cov.↑	Avg. Set Size↓	LCAD↓
Flat LAC (Sadinle et al., 2019)	0.8901 ± 0.0021	33.08 ± 0.08	0.069 ± 0.002
Flat APS (Romano et al., 2020)	0.9067 ± 0.0004	46.74 ± 0.06	0.079 ± 0.002
Flat RAPS (Angelopoulos et al., 2021)	0.9065 ± 0.0008	46.71 ± 0.06	0.079 ± 0.001
CRSVP (Mortier et al., 2026) (single-node)	0.0369 ± 0.0011	$\mathbf{1.00 \pm 0.00}$	3.640 ± 0.004
HCC (den Hengst et al., 2025) (max, marginal-native)	0.3855 ± 0.0025	11.12 ± 0.42	1.912 ± 0.156
HCC (den Hengst et al., 2025) (min, joint)	0.9969 ± 0.0006	1.49 ± 0.01	4.306 ± 0.001
BR-MLCP (Lambrou & Papadopoulos, 2016)	0.6210 ± 0.0065	22.78 ± 0.12	0.321 ± 0.006
Fisch et al. (Fisch et al., 2022) (Top-k)	0.8932 ± 0.0015	39.00 ± 0.00	0.060 ± 0.000
CRC-Hierarchical (Angelopoulos et al., 2024)	0.9810 ± 0.0011	41.11 ± 0.15	0.010 ± 0.001
DSB	0.9584 ± 0.0010	39.16 ± 0.08	0.032 ± 0.000
TAPS (Ours)	0.8905 ± 0.0017	17.08 ± 0.14	2.833 ± 0.003

Validity, compactness and the price of specificity. LAC, APS, RAPS and TAPS all attain near-nominal empirical semantic joint coverage ($\approx$ 0.89–0.91; Corollary 2), but TAPS does so with significantly fewer nodes: 17.1 on average, vs. 33.1 for LAC and 46.7 for APS/RAPS, a $\approx 48\%$ reduction. This is a true abstraction, not a shortcut: root-only predictions occur for only 0.01% of patients, and the mean emitted depth is 2.88–2.90 of 6 (Appendix F). This compression entails a substantial loss of leaf-level specificity. LCAD of TAPS (2.83) is much higher than the flat methods (0.07–0.08), so if a user requires exact leaf matches, flat CP is still the better choice.

Single-node and DAG-structured baselines fail in opposite directions. CRSVP covers only 3.7% of patients jointly. This is not a calibration issue, but a structural one: a single output node cannot be on the ancestor path of all the comorbidities at the same time (Appendix G). But instead, the two versions of HCC fail in opposite ways. HCC-max remains compact (11.1 nodes) but severely undercovers (38.6%). HCC-min only reaches coverage by collapsing to $\{\text{root}\}$ for 95% of patients, as its pruning logic never triggers after labels span different ICD chapters (Appendix H). The fact that two independently designed baselines fail in opposite ways is stronger evidence for score-level integration than the failure of either one alone.

[2]Introduced here as a structural strawman baseline.

The pattern also holds over the extended baselines. The failures split nicely rather than lumping these together: CRSVP, HCC-max, and CSP undercover; HCC-min and CRSVP-r collapse to the root; Tree-CP reaches near-target coverage but with large, fragile sets; CRC-Hierarchical and DSB overcover substantially, at 2.3–2.4× TAPS's set size. Only TAPS achieves the target coverage with a truly compact set, and no root-collapse. We attribute this to a common mechanism we term the "free root" pathology: per-node scoring rules without an explicit abstraction penalty tend to consider the root as effectively free (Appendix I). Related root-collapse behavior has been reported independently in hierarchical selective classification and class-similarity-based conformal prediction, though those works study different problems.

Flat multi-label baselines fail for a related reason. Both the top-k construction of Fisch et al. and the per-label calibration of BR-MLCP fail, but in opposite ways. Fisch et al.'s global rank threshold, which is forced to satisfy the single hardest patient, ends up predicting almost the whole leaf set (39 out of the 50, identical across all three seeds), producing $\approx$ 34–38 false positives despite being designed for precision. BR-MLCP achieves marginal coverage correctly ($\approx$ 90% per label), but without joint correction, joint coverage drops to 62.1%, which is exactly why the joint, max-score protocol (Eq. 4) matters rather than being incidental.

The gains are not restricted to clinical text. The same ancestor-collapsing mechanism explains this (Lemma 5) as well, where TAPS reduces set size by 86% on arXiv's category taxonomy (depth 3; Appendix P) at near-nominal empirical semantic joint coverage, even more so than ICD-10's 48%. But one more domain is not proof of universal generality, it is real evidence that the gains or baseline failures are not artefacts of clinical text or the structure of ICD-10.

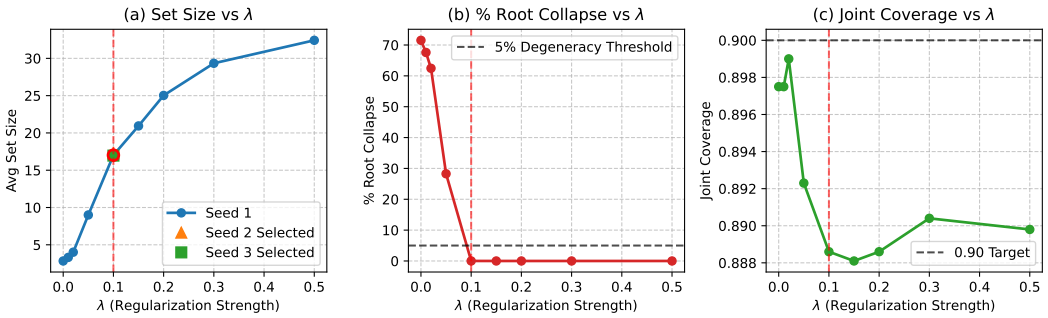


Figure 2: Specificity–coverage frontier controlled by λ (Seed 1 sweep; Seeds 2–3 overlaid in panel (a); all select λ=0.10). All metrics in this figure are computed on the D_λ partition of the calibration set, matching the λ-selection protocol. Root-only predictions drop from 70.7% (λ=0) to 0.01% (λ=0.10); set size rises from 2.8 to 17.0; joint coverage stays in 0.888–0.900. Dashed line marks the selected operating point.

The operating point λ (Figure 2) is selected on the calibration split by the two-stage rule of Remark 2. The full sweep is reported so the tradeoff, and the rejected degenerate region, is transparent.

5.3 Ablation: Isolating the Score-Level Contribution

TAPS combines a hierarchy-aware *score* and a hierarchy-aware *output construction*. To check that both actually matter, we run three arms on the same cached probabilities under the same joint-coverage protocol (Table 3):

Table 3: Three-arm ablation (Seed 1, α=0.10). The pattern reproduces on Seeds 2–3: Arm 1 gives 100% root on both; Arm 2 set size 36.06–36.07 versus Arm 3's 16.96–17.25.

Arm	Joint Cov.	Set Size↓	LCAD↓	% Root	Mean Depth
1: Flat score + hier. output	1.0000	1.00	4.34	100.00%	0.00
2: LCA score + flat output	0.8886	36.00	0.08	0.00%	4.44
3: Full TAPS	0.8886	17.02	2.83	0.01%	2.89

Arm 1, flat score with hierarchical output, collapses to $\{\text{root}\}$ for each patient, the hierarchy in the output is simply not enough. Arm 2, LCA score with flat output, escapes root-collapse entirely (0% root) but ends up with sets almost as large as flat LAC (36.00 nodes). Only Arm 3, full TAPS, gets both: non-degenerate coverage and compact sets (17.02 nodes). Arms 2 and 3 share identical joint coverage (0.8886) by construction, since they differ only in whether passing internal nodes are emitted directly (Arm 3) or expanded into their full set of descendant leaves (Arm 2), not in how the threshold is calibrated. See Appendix J for extended analysis.

5.4 Discussion and Limitations

In addition to raw statistical performance, TAPS offers real interpretive advantages. An internal node, such as `Block I20--I25 (Ischaemic Heart Disease)`, provides clinicians *targeted, interpretable boundaries* under uncertainty, as opposed to flat CP lists that require manual validation of 20 separate leaf codes. We monitor LCAD all the way through to make sure this collapsing never turns into clinically vague abstraction. The architecture requires only a well-defined tree height and node-scoring model, so it transfers to any hierarchical ontology without modification, as demonstrated on arXiv's category taxonomy (Section 5).

There are some limitations to these findings. **Semantic coverage vs exact coverage:** The relaxation to semantic coverage is easier to satisfy than exact coverage, so the compactness gains must be read together with LCAD, not in isolation. **Tall hierarchies:** The raw classifier shows an acute free-root pathology on the 13-level MeSH tree (Appendix O) that is fully suppressed by TAPS's non-degeneracy screen (0.00% root collapse). A milder, but distinct pathology, independent overconfidence across orthogonal branches, is likely contributing to the elevated LCAD of MeSH. **Near-perfect accuracy:** DBpedia (26 nodes, depth 4; Appendix Q) confirms the stability when moving towards near-perfect accuracy, as opposed to further compactness gains.

Empirical scope is partial: the top-50 codes account for only 33.9% of MIMIC-IV's ICD-10 instances, leaving the long tail untested; a per-chapter stratified threshold (sketched, not adopted, in Appendix N) would likely serve it better than our single global threshold. **Subgroup fairness:** a preliminary conditional coverage check across age subgroups (Appendix K) shows up to 2.9% deviation from nominal coverage, concentrated in the < 45 group, so broader demographic auditing remains future work. **The λ tradeoff** is a design choice, not a free parameter: internal-node emission necessarily abstracts away leaf identity, and λ must be selected carefully on the calibration split, since a poor choice yields either root-collapse or inflated sets. **arXiv caveats:** that validation trains for one epoch rather than two, and surfaces one HCC-specific finding worth flagging: its cost-minimization can select a cheap, confidently-wrong cover over a correct but costlier one, so its coverage numbers there should be read as covering in the weak semantic sense of Definition 1, not as reliably correct.

6 Conclusion

We introduce TAPS, to our knowledge the first conformal prediction framework to integrate an ontology-derived, specificity-regularized, per-node LCA score directly into calibration for hierarchical multi-label classification. This construction yields firm theoretical guarantees: score-level dominance (Theorem 1), distribution-free semantic marginal coverage (Theorem 4), and provable node-count compactness (Lemma 5), governed by a single parameter λ that traces a validity-preserving frontier between compact and specific predictions. On MIMIC-IV, TAPS delivers near-nominal semantic joint-coverage prediction sets roughly 48% smaller than flat conformal baselines. Our three-arm ablation confirms both score-level and output-level integration are independently necessary: structuring only the output collapses to trivial $\{\text{root}\}$ predictions, while structuring only the score yields sets nearly as bloated as flat CP. The corresponding failures of six independent hierarchical and multi-label baselines (CRSVP, CRSVP-r, HCC, CRC, Tree-CP, and CSP) plus our own structured strawman (DSB) further support the necessity of this combined, set-valued, score-integrated construction. TAPS's hierarchy-aware, calibrated prediction sets are directly suited to supporting expert review at scale, in ICD-10 coding audits and equally in other domains with maintained hierarchies and growing multi-label demands, such as legal, scientific, or product taxonomies. Realizing this requires domain-specific validation of the underlying probability estimates, which we leave to future work.

Reproducibility Statement

All proofs are stated in full in Appendices A and B with explicit assumptions. The conformal pipeline was validated on synthetic exchangeable data before any model training. The parameter λ is selected entirely on the calibration split via a documented two-stage non-degeneracy rule and evaluated on a held-out test set that was not used for training or parameter selection. Headline results are computed with independent λ-selection per random seed. All baselines are implemented from cited formulations with any multi-label adaptation explicitly documented. For MIMIC-IV, train/calibration/test splitting is performed by `subject_id` to prevent patient leakage; dataset-specific splitting procedures for the remaining benchmarks are documented in their respective appendices. Cached per-node prediction tensors for all four datasets (totaling 67MB) are hosted externally at an anonymous, restricted-access Zenodo record (DOI: 10.5281/zenodo.22728044) due to size, along with a download script; all other code, configuration, and small artifacts are included directly in the supplementary material.

Ethics Statement

This work evaluates uncertainty-quantification methods using established research datasets, including MIMIC-IV, under their applicable access and use requirements. We do not release patient-level clinical data and do not claim clinical-deployment readiness. Our subgroup analysis identifies coverage variation across age groups; clinical use would therefore require additional demographic, institutional, and prospective validation. The method is intended to support expert review rather than replace professional clinical judgment.

AI Use Statement

Generative AI tools were used under iterative human guidance and supervision for: software implementation and debugging; drafting and substantially revising sections of the manuscript; formulating and refining mathematical claims and proof arguments, including corrections identified during iterative review; feedback on experimental and calibration methodology; interpretation of experimental results and failure-mode diagnosis; and language editing and reference formatting. Generative AI was not treated as an authoritative source of scientific evidence. The human authors independently verified the final implementation, mathematical claims and proofs, experimental design, numerical results, citations, and manuscript content, and take full responsibility for all content in this submission.

References

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4): 494–591, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Wei Bi and James T. Kwok. Bayes-optimal hierarchical multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2907–2918, 2015.

Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22 (81):1–42, 2021.

Floris den Hengst, Inès Blin, Majid Mohammadi, Syed Ihtesham Hussain Shah, and Taraneh Younesian. Hierarchical conformal classification. *arXiv preprint arXiv:2508.13288*, 2025. Concurrent work.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ariel Fargion, Lahav Dabah, and Tom Tirer. Enhancing conformal prediction via class similarity. In *Forty-third International Conference on Machine Learning*, 2026.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Conformal prediction sets with limited false positives. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6514–6532. PMLR, 2022.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *Advances in Neural Information Processing Systems*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1321–1330, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the Workshop on Clinical Natural Language Processing*, pp. 1–10. Association for Computational Linguistics, 2022.

Alistair E.W. Johnson, Lucas Bulgarelli, Lu Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–10, 2023.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

Antonis Lambrou and Harris Papadopoulos. Binary relevance multi-label conformal predictor. In *Conformal and Probabilistic Prediction with Applications*, pp. 90–104, Cham, 2016. Springer.

Pablo N. Mendes, Max Jakob, and Christian Bizer. DBpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

Thomas Mortier, Alireza Javanmardi, Yusuf Sale, Eyke Hüllermeier, and Willem Waegeman. Conformal prediction in hierarchical classification with constrained representation complexity. In *29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*. PMLR, 2026.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT*, pp. 1101–1111, 2018.

Thomas Norrenbrock, Timo Kaiser, Sovan Biswas, Neslihan Kose, Ramesh Manuvinakurike, and Bodo Rosenhahn. Chiqpm: Calibrated hierarchical interpretable image classification. In *Advances in Neural Information Processing Systems*, 2025.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72, 2011.

Chhavi Tyagi and Wenge Guo. Multi-label classification under uncertainty: A tree-based conformal prediction approach. In *Conformal and Probabilistic Prediction with Applications*, pp. 488–512. PMLR, 2023.

Javier Vázquez and Julio C. Facelli. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3):241–252, 2022.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Botong Zhang, Shuo Li, and Osbert Bastani. Conformal structured prediction. In *International Conference on Learning Representations*, 2025a.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Xin Zhang, Qiyu Wei, Yingjie Zhu, Fanyi Wu, and Sophia Ananiadou. Thcm-cal: Temporal-hierarchical causal modelling with conformal calibration for clinical risk prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, volume 2025.findings-emnlp.48, pp. 916–928, Suzhou, China, 2025b. Association for Computational Linguistics.

A Proofs

A.1 Proof of Theorem 1 (Dominance)

Every leaf $y^* \in \mathcal{L}$ satisfies $\mathrm{height}(y^*) = 0$ by definition (a leaf has no descendants to extend a downward path), unconditionally. This holds for every leaf regardless of its own depth from the root. Hence $g_{y^*}(x) = (1 - \hat{p}_{y^*}(x)) + \lambda \cdot 0 = s_{\mathrm{flat}}(x, y^*)$ exactly. Since $y^* \in \{y^*\} \cup \mathrm{Anc}(y^*)$, the minimum in Eq. (2) is over a set containing y^*, hence

$$s_{\mathrm{LCA}}(x, y^*) = \min_{v \in \{y^*\} \cup \mathrm{Anc}(y^*)} g_v(x) \;\le\; g_{y^*}(x) = s_{\mathrm{flat}}(x, y^*).$$

As this holds pointwise, the empirical quantile obeys $\hat{q}_{\mathrm{LCA}} \le \hat{q}_{\mathrm{flat}}$ for any α. The pointwise inequality establishes this weak quantile dominance; strict quantile improvement may occur when the score reductions from ancestor abstraction affect the relevant empirical calibration quantile. □

Proposition 3 (Path Coverage)**.** *Suppose some node $v^* \in \{y^*\} \cup \mathrm{Anc}(y^*)$ satisfies $g_{v^*}(x) \le \hat{q}$, i.e. $v^* \in P(x)$. Then $C_{\mathrm{TAPS}}(x)$ contains a node in $\{y^*\} \cup \mathrm{Anc}(y^*)$, hence $y^* \in_{\mathrm{sem}} C_{\mathrm{TAPS}}(x)$.*

A.2 Proof of Proposition 3 (Path Coverage)

Let $\pi = (\mathrm{root} = u_0, u_1, \ldots, u_d = y^*)$ be the root-to-y^* path; every node on π is in $\{y^*\} \cup \mathrm{Anc}(y^*)$. By hypothesis $\pi \cap P(x) \neq \emptyset$. Let u_m be the *shallowest* node of π that lies in $P(x)$ (the one of minimum depth). We claim $u_m \in C_{\mathrm{TAPS}}(x)$, i.e. u_m has no proper ancestor in $P(x)$. Every proper ancestor of u_m on π is some u_k with $k < m$; by minimality of m none of these is in $P(x)$. Since all ancestors of u_m lie on π (the path from the root is unique in a tree), u_m has no proper ancestor in $P(x)$ at all, so u_m is a maximal element of $P(x)$ and $u_m \in C_{\mathrm{TAPS}}(x)$. As $u_m \in \{y^*\} \cup \mathrm{Anc}(y^*)$, we have $y^* \in_{\mathrm{sem}} C_{\mathrm{TAPS}}(x)$.

The argument uses no monotonicity of g_v along π: even if $P(x)$ restricted to π has "gaps" (passing, then failing, then passing again), the shallowest passing node on π is still emitted, because maximality is decided only by ancestors, which all lie on π. □

Assumption 1 (Exchangeability)**.** $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable.

Assumption 2 (Score Measurability)**.** For a fixed $\lambda \ge 0$, the node scores $g_v : \mathcal{X} \to \mathbb{R}$ and hence $s_{\mathrm{LCA}} : \mathcal{X} \times \mathcal{L} \to \mathbb{R}$ are measurable; the tree $\mathcal{T}$ is fixed (deterministic) with finite label space $\mathcal{L}$, and λ is chosen independently of the test point.

Theorem 4 (Semantic Marginal Coverage). *Under Assumptions 1 and 2, for any $y^* \in Y_{n+1}$:*

$$1 - \alpha \;\le\; \mathbb{P}\big(y^* \in_{\text{sem}} C_{\text{TAPS}}(X_{n+1})\big) \tag{7}$$

No matching upper bound holds for the max-score calibration protocol of Eq. (4): because $S_i = \max_{y \in Y_i} s_{\text{LCA}}(X_i, y)$ dominates the single-label score $s_{\text{LCA}}(X_i, y_i^*)$ whenever a patient has more than one true label, $\hat{q}$ is calibrated against a stochastically larger sequence, and marginal coverage under this protocol can exceed $1 - \alpha + 1/(n+1)$, as borne out empirically (Section 5, marginal coverage ≈ 0.97 at $\alpha = 0.10$). The tight two-sided bound is instead available for joint *coverage under this same protocol (Corollary 2), which is the guarantee we report as the headline result in Table 2.*

A.3 Proof of Theorem 4 (Semantic Marginal Coverage)

Step 1: Equivalence. We show $y^* \in_{\text{sem}} C_{\text{TAPS}}(x) \iff s_{\text{LCA}}(x, y^*) \le \hat{q}$, using the same g_v and λ in score and construction.

($\Leftarrow$): If $s_{\text{LCA}}(x, y^*) \le \hat{q}$, then $\min_{v \in \{y^*\} \cup \text{Anc}(y^*)} g_v(x) \le \hat{q}$, so some $v^* \in \{y^*\} \cup \text{Anc}(y^*)$ has $g_{v^*}(x) \le \hat{q}$, i.e. $v^* \in P(x)$. By Proposition 3, $y^* \in_{\text{sem}} C_{\text{TAPS}}(x)$.

($\Rightarrow$): If $y^* \in_{\text{sem}} C_{\text{TAPS}}(x)$, then some $c \in C_{\text{TAPS}}(x)$ satisfies $c \in \{y^*\} \cup \text{Anc}(y^*)$. Since $C_{\text{TAPS}}(x) \subseteq P(x)$, $g_c(x) \le \hat{q}$, hence $\min_{v \in \{y^*\} \cup \text{Anc}(y^*)} g_v(x) \le g_c(x) \le \hat{q}$, i.e. $s_{\text{LCA}}(x, y^*) \le \hat{q}$.

Step 2: Exchangeability. For fixed λ, the per-example scores $S_i = \max_{y \in Y_i} s_{\text{LCA}}(X_i, y)$ are a fixed measurable function of (X_i, Y_i); by Assumptions 1–2 the sequence $(S_1, \ldots, S_{n+1})$ is exchangeable. This is the same calibration sequence used to compute $\hat{q}$ in Eq. (4); we apply the standard CP guarantee directly to it in Step 3, rather than to the single-label sequence $s_{\text{LCA}}(X_i, y_i^*)$, since $\hat{q}$ is not the empirical quantile of that sequence.

Step 3: Standard CP guarantee, applied to the dominating sequence. Since $S_i = \max_{y \in Y_i} s_{\text{LCA}}(X_i, y) \ge s_{\text{LCA}}(X_i, y_i^*)$ pointwise for any chosen $y_i^* \in Y_i$, the quantile $\hat{q}$ computed from $\{S_i\}$ is no smaller than the quantile that would be computed from $\{s_{\text{LCA}}(X_i, y_i^*)\}$ alone. By Theorem 1 of Vovk et al. (2005) applied to the exchangeable sequence $(S_1, \ldots, S_{n+1})$: $\mathbb{P}(S_{n+1} \le \hat{q}) \ge 1 - \alpha$. Since $s_{\text{LCA}}(X_{n+1}, y^*) \le S_{n+1}$, we have $\{S_{n+1} \le \hat{q}\} \subseteq \{s_{\text{LCA}}(X_{n+1}, y^*) \le \hat{q}\}$, giving $\mathbb{P}(s_{\text{LCA}}(X_{n+1}, y^*) \le \hat{q}) \ge 1 - \alpha$. The matching upper bound from Vovk et al. (2005) applies only to the sequence actually used for calibration (the max-score S_i's), which is a strictly different event from single-label semantic coverage once a patient has ≥ 2 labels; hence no upper bound on marginal coverage transfers under this protocol.

Step 4: Conclusion. Combining Steps 1–3 yields the one-sided result. □

A.4 Proof of Corollary 2 (Joint Coverage)

By Theorem 4's Step 1 (Equivalence), applied pointwise to every $y \in Y_{n+1}$: $y \in_{\text{sem}} C_{\text{TAPS}}(x) \iff s_{\text{LCA}}(x, y) \le \hat{q}$. Hence $\{\forall\, y \in Y_{n+1} : y \in_{\text{sem}} C_{\text{TAPS}}(x)\} = \{\max_{y \in Y_{n+1}} s_{\text{LCA}}(x, y) \le \hat{q}\} = \{S_{n+1} \le \hat{q}\}$. By Step 2, $(S_1, \ldots, S_{n+1})$ is exchangeable, so by the standard conformal rank argument, the lower bound $1 - \alpha \le \mathbb{P}(S_{n+1} \le \hat{q})$ holds unconditionally. Under the additional almost-sure distinctness condition stated in Section 4.3, or with randomized tie-breaking, the corresponding upper bound $\mathbb{P}(S_{n+1} \le \hat{q}) \le 1 - \alpha + 1/(n+1)$ also holds. Substituting gives the result. □

Lemma 5 (Efficiency: node-count compactness). *Fix an input x, a parameter λ, and a threshold $\hat{q}$. Let $C_{\text{TAPS}}(x)$ be the topmost-node set of Definition 4. Then $C_{\text{TAPS}}(x)$ is an antichain, and for any internal node $v \in C_{\text{TAPS}}(x)$ the single element v represents the entire passing region of its subtree: no descendant of v appears in the set. Consequently*

$$|C_{\text{TAPS}}(x)| \;\le\; \big|\{\, v \in P(x) : v \text{ is a leaf or a maximal passer} \,\}\big|, \tag{8}$$

and whenever an internal node is emitted in place of ≥ 2 passing descendants, $|C_{\text{TAPS}}(x)|$ is strictly smaller than the corresponding flat set that lists those descendants individually.

A.5 Proof of Lemma 5 (Efficiency)

$C_{\text{TAPS}}(x)$ is the set of maximal elements of $P(x)$, so no element is an ancestor of another (antichain). If internal $v \in C_{\text{TAPS}}(x)$, then $v \in P(x)$ and, by maximality of v in $P(x)$, the construction emits v

and never emits a descendant of v (any descendant in $P(x)$ has v as a proper ancestor in $P(x)$ and is therefore not maximal). Thus the whole confident subtree under v is summarised by one element, whereas a flat set enumerates each passing leaf. When v subsumes ≥ 2 such leaves the count strictly decreases. □

B Multi-Label Coverage (Corollary)

Corollary 6 (Multi-Label Semantic Coverage). *For $Y^* = \{y_1^*, \ldots, y_m^*\}$, applying* TAPS *independently per label with Bonferroni-corrected level $\alpha_j{=}\alpha/m$ gives:*

$$\mathbb{P}\Big(\forall\, j \in [m] : y_j^* \in_{\text{sem}} C_j(X)\Big) \;\geq\; 1 - \alpha \tag{9}$$

This is a conservative union bound; we do not claim it is tight.

Proof. Run TAPS per label j at level $\alpha_j{=}\alpha/m$. By Theorem 4: $\mathbb{P}(y_j^* \notin_{\text{sem}} C_j) \leq \alpha/m$. Union bound: $\mathbb{P}(\exists j : y_j^* \notin_{\text{sem}} C_j) \leq \sum_j \alpha/m = \alpha$. Complementing gives $\geq 1{-}\alpha$ joint coverage. □

C Height vs. Depth: Full Derivation

An earlier version of the score used $d_{\max} - \text{depth}(v)$ instead of $\text{height}(v)$. This is a subtle but consequential mistake: it sets the penalty to zero only for nodes at the *global* maximum depth $d_{\max}$ of the tree, not for every leaf. Our ICD-10 subset tree is not depth-regular, with leaves at depths 3 to 6 (out of maximum depth 6), and, for a shallower-than-$d_{\max}$ leaf y^*, we would have $g_{y^*}(x) = (1 - \hat{p}_{y^*}(x)) + \lambda \cdot (d_{\max} - \text{depth}(y^*)) > s_{\text{flat}}(x, y^*)$, which would break the pointwise dominance. $s_{\text{LCA}} \leq s_{\text{flat}}$ required by Theorem 1; we verified that this was not just a theoretical possibility but was actually happening in our own calibration data at the previously chosen operating point. Instead, we use $\text{height}(v)$ (the distance from v down to the *deepest leaf in its own subtree*), which guarantees $\text{height}(\text{leaf}) = 0$ for *every* leaf unconditionally, regardless of its depth from the root, and thus restores the dominance property exactly (Theorem 1). Height is still a purely structural, precomputable quantity (it does not refer to the unknown true label y^*), and so it is still usable identically in both calibration and the inference-time construction (Definition 4), which is what the exchangeability argument in Theorem 4 requires.

D Semantic Coverage Relaxation

Semantic coverage is strictly weaker than exact coverage: counting an ancestor as a successful prediction makes the criterion easier to satisfy than exact-label coverage. We state this trade-off explicitly so that coverage improvements are not misattributed to the method alone. The relaxation is clinically motivated: returning the correct parent block (e.g., Block I20–I25: Ischaemic Heart Disease) when the true code is I21.0 (STEMI) still guides the clinician to the correct diagnostic neighbourhood, though it does change the specific target. We report the LCA distance (LCAD) metric (Section 5) to verify that coverage gains do not come at the cost of excessive clinical vagueness. Mortier et al. (2026) use an exact-match coverage notion even for internal nodes; ours is strictly more general and should be compared with this caveat in mind.

E Per-Node Outputs Rationale

A previous formulation aggregated leaf probabilities at internal nodes. Both simple aggregators fail. A SUM over independent sigmoids can exceed 1, making $1 - \hat{p}_v$ negative and invalidating the calibration quantile. A MAX over descendant leaves $\hat{p}_v \in [0, 1]$ is used, but this forces node scores to be non-increasing from root to leaf; as a result the prediction set never emits any internal nodes and, worse, its semantic-coverage event does not match the calibration score (a sibling leaf can satisfy the score for y^*, whereas the constructed set completely excludes y^*). Both types of failures can be avoided by predicting each node directly: node scores are not forced to be monotonic along a path, so confident internal nodes have a meaningful boundary, and, as Theorem 4 shows, the calibration score and the prediction set evaluate the same event.

F NON-DEGENERACY DIAGNOSTICS

Table 4 provides the full non-degeneracy diagnostics for TAPS evaluated on the test set.

Table 4: Non-degeneracy diagnostics for TAPS on Seed 1 (λ=0.10), measured on the test set. The same pattern holds across all 3 seeds: root-only predictions never exceed 0.03% and mean emitted depth stays in the 2.88–2.90 range at λ=0.10.

Diagnostic	**Value**
% predictions = $\{\text{root}\}$	0.01%
Mean emitted-node depth (max depth 6)	2.89
Mean set size	17.02

G CRSVP DEGENERACY ANALYSIS

CRSVP (Mortier et al., 2026) was designed for single-label classification, where a single node can meaningfully represent the one true class. Our multi-label adaptation reveals a structural limitation: MIMIC-IV patient labels typically span multiple ICD chapters (Section 5), and a single node, no matter how selected, cannot be on the ancestral path of every comorbidity at the same time. The greedy single-path construction we use (maximum confident descent, stopping at a flat threshold) achieves reasonable specificity (average depth 4.34) but only includes labels that coincidentally fall on one of its selected paths, resulting in joint coverage 3.6% and marginal coverage 12.0%, both well below the target. This is not a calibration failure (the same flat $\hat{q}$ as LAC is used) but rather a consequence of the single-node output limitation. This illustrates, on real data, why a genuinely set-valued construction is essential for multi-label tasks: TAPS's topmost-passing-node set (Definition 4) can simultaneously represent several discrete confidence regions, which a single node structurally cannot.

H HCC DEGENERACY & INTRACTABILITY ANALYSIS

We evaluate HCC (den Hengst et al., 2025) under both its native max-based (marginal) conformity score and a min-based joint adaptation we derive by analogy to Corollary 2. Neither achieves the joint-coverage target. *HCC-max* optimises explicitly for compactness (cost $|\mathcal{N}| + \beta|\text{leaf-cov}(\mathcal{N})|$) and produces small sets (11.1 nodes) but only 38.6% joint coverage, since its cost function is free to omit comorbid labels beyond the single best-covered one. *HCC-min* corrects the coverage objective but at severe cost: HCC's dynamic pruning step, designed under the paper's single-label assumption that a leaf-level prediction set clusters within one branch of the tree, never triggers in our multi-label setting: patients' passing leaves routinely span disparate ICD chapters, so their lowest common ancestor is almost always the root, which has no ancestor to prune against (m=7 surviving NOL-covers on average, i.e. no reduction from the full depth-level search). The full Bonferroni correction across all m=7 covers is therefore applied uniformly, and the resulting threshold collapses to $\{\text{root}\}$ for 95% of patients on average, rising from 88% for low-comorbidity patients (≤ 3 true labels) to 99% for high-comorbidity patients (> 6 true labels).

We use the depth-level NOL-cover search rather than full enumeration because the latter is intractable for our tree: the naive search space over all subsets is $2^{149} \approx 7 \times 10^{44}$, and restricting to valid NOL-covers via the recurrence $N(v) = 1 + \prod_{c \in \text{children}(v)} N(c)$, applied directly to our 149-node tree, which we verified contains no multi-parent nodes, so this count is exact, still yields $N(\text{root}) = 4{,}869{,}548{,}768{,}878{,}684{,}801 \approx 4.87 \times 10^{18}$ valid covers. Exhaustive enumeration is accordingly infeasible: our implementation exhausted a 5-minute compute budget after identifying only 7,857 valid covers.

Table 5: Additional Baseline Results for MIMIC-IV (α=0.10). Protocol identical to Table 2.

Method	Semantic Joint Cov.	Marg. Cov.	Avg. Size	LCAD
Tree-CP (Tyagi & Guo, 2023)	0.8972 ± 0.0008	0.9761 ± 0.0002	34.30 ± 0.05	0.0821 ± 0.0004
CSP (aggregate) (Zhang et al., 2025a)	0.8842 ± 0.0017	0.9543 ± 0.0010	58.57 ± 0.86	1.4925 ± 0.0187
CSP (per-example) (Zhang et al., 2025a)	0.8944 ± 0.0055	0.8605 ± 0.0050	2.10 ± 0.03	4.1247 ± 0.0048
CRSVP-r (r=2) (Mortier et al., 2026)	0.9008 ± 0.0031	N/A	1.1003 ± 0.0028	4.1948 ± 0.0053
CRSVP-r (r=3) (Mortier et al., 2026)	0.8999 ± 0.0044	N/A	1.2062 ± 0.0091	4.1537 ± 0.0087
CRSVP-r (r=5) (Mortier et al., 2026)	0.9016 ± 0.0035	N/A	1.4522 ± 0.0164	4.0993 ± 0.0107

I Extended Baseline Analysis

I.1 Tree-CP

Tyagi and Guo's original Tree-CP procedure assumes a custom-built labelset tree in which each calibration example has exactly one true node per layer, which underlies their FWER proof for the uniform per-layer correction $\alpha_i = \alpha/c$. This assumption does not hold on our multi-path ontology, where a patient's true labels routinely span multiple branches simultaneously. Instead of creating a special correction to compensate for this (which would test a different, unpublished method instead of Tree-CP), we apply their published correction directly to our existing ontology tree according to specific instructions. The per-instance inconsistency score is the lowest predicted probability along the root-to-node path of each correct label; since this score is structurally unidirectional along any path, the resulting prediction set, i.e. the deepest traversed node of each branch, satisfies the nestedness condition of Proposition 3 without modification, which does not apply to CRC-Hierarchical and DSB, as they require an explicit antichain collapse to be well-defined. Despite this structural robustness, Tree-CP exhibits **threshold fragility**: since the score is minimal over an entire path, a poorly calibrated internal node anywhere along that path can affect the score, significantly lowering the global calibration threshold. In MIMIC-IV, this lowers the effective threshold to $\tau \approx 0.09$, which forces the method to retain wide, unspecified branches to maintain coverage (average set size 34.3 nodes, which is about 2 times the 17.08 of TAPS, although the joint coverage is close to the target of 0.897). On arXiv's comparatively cleaner probability surface the effect is milder (13.8 nodes at 0.905 coverage), but the underlying fragility (a single unreliable node anywhere on a path silently degrading the whole calibration) is structural to the method, not an artifact of our adaptation.

I.2 Conformal Structured Prediction (CSP)

Zhang, Li, and Bastani formulate structured prediction sets via an integer linear program that, for a given probability-mass threshold τ, selects the minimum-leaf-count subset of nodes covering at least τ probability mass, calibrated via a sequential "learn-then-test" search. We found this construction admits two materially different calibration interpretations on our multi-label setting, and following the same transparency principle behind our dual reporting of HCC's max- and min-based variants, report both rather than silently choosing one.

CSP (aggregate) calibrates a single global τ by binary search for the value at which 90% of calibration examples are covered by the ILP's output, evaluated under the paper's semantic coverage definition (Definition 1). This is stable and non-degenerate on both benchmarks (0% root-collapse), but is **not a valid split-conformal predictor**: we verified empirically that the ILP's prediction sets are not nested in τ (violations in 2-7% of consecutive threshold pairs sampled across both benchmarks), meaning $C_{\tau_1} \subseteq C_{\tau_2}$ for $\tau_1 < \tau_2$ does not hold in general. Because Theorem 4's marginal coverage guarantee depends on exactly this nestedness (via Proposition 3), the aggregate procedure's empirical coverage (0.884 MIMIC, 0.906 arXiv) is not backed by the same distribution-free guarantee TAPS and the other baselines carry.

CSP (per-example) instead defines a genuine per-calibration-example nonconformity score (the minimum τ at which the ILP happens to cover that example's true labels) and calibrates via the standard $\lceil (n+1)(1-\alpha) \rceil$-th quantile, restoring proper split-conformal validity. This version behaves very differently: on MIMIC it produces small, confident sets (2.1 nodes) more akin to a top-1 leaf predictor than a hierarchical summarizer, with correspondingly high LCAD (4.10). On arXiv, the same procedure becomes unstable, with joint coverage varying substantially across seeds (0.89-0.92)

and severe, seed-dependent root-collapse (5.8%, 27.3%, and 27.2% across the three seeds), which is considerably worse than any other baseline's degeneracy rate on this benchmark.

Read together, the two CSP variants illustrate a genuine tension: the version that is theoretically valid is empirically unstable, and the version that is empirically stable lacks the theoretical guarantee the paper's own framework depends on. TAPS's score-level integration avoids this tension entirely. Its passing set is monotone by construction (Section 4.1), so validity and stability are not in conflict.

I.3 BR-MLCP

Binary Relevance conformal prediction (Lambrou and Papadopoulos) calibrates each label independently via its own conformal predictor, with no joint or Bonferroni correction, then reports the union of per-label prediction sets. We restrict both calibration and prediction strictly to leaf codes, matching the scope of LAC, APS, and RAPS, rather than calibrating over internal ancestor nodes; an earlier version of this baseline calibrated over all tree nodes and produced an artifactual near-1.0 joint coverage, since covering any ancestor of a true label was sufficient to count as a hit. With calibration correctly restricted to leaves, marginal per-label coverage lands close to the 90% target, confirming the calibration itself is correct. Joint coverage, however, drops to 62.1%: a patient with several true labels only needs one of them to fall outside its own leaf's calibrated interval for the whole prediction to miss. BR-MLCP's average set size (22.78 leaves) is in fact smaller than Flat LAC's (33.08), but this apparent efficiency is an artifact of under-coverage, not genuine specificity. A smaller set that fails to reach the target coverage is not a meaningful efficiency gain. BR-MLCP's substantially elevated LCAD (0.321 vs. LAC's 0.069) has a direct topological explanation: because BR-MLCP's smaller sets miss entire branches more often than LAC's larger ones, a miss more frequently has no shared ancestor with the true label short of the root, inflating the average LCA distance.

I.4 Fisch et al. (Top-k)

Fisch et al. propose bounding the number of false positives in a conformal prediction set, targeting settings such as in-silico drug screening where a small, precise candidate list matters more than exhaustive coverage. We adapt their construction to our setting as a nested family of top-k candidate sets over leaf codes, ranked by predicted probability, and calibrate k via the same per-patient max-score joint protocol used throughout (Eq. 4), fixing the target at our 90% joint coverage level and reporting the resulting false-positive count, rather than fixing a false-positive budget and measuring coverage, to keep the comparison anchored to the same calibration philosophy as every other baseline. The calibrated k converges to exactly 39 (of 50 total leaves) identically across all three seeds, a zero-variance result that is itself diagnostic rather than coincidental. Because this construction selects a single GLOBAL rank cutoff shared by every patient, rather than a per-patient adaptive threshold like LAC's continuous probability quantile, the cutoff must be set to satisfy the single hardest-to-cover patient in the calibration set, and every other patient inherits that same generous allowance. The result is nearly exhaustive prediction (39 of 50 possible leaves per patient) with an average of 34–38 false positives, the opposite of the method's intended precision, illustrating that non-adaptive rank thresholds are a poor fit for highly comorbid, multi-label clinical data.

I.5 CRC-Hierarchical

Conformal Risk Control (Angelopoulos et al.) includes a hierarchical-classification instance, evaluated on ImageNet under the WordNet taxonomy, that bounds the expected graph-distance from a nested, threshold-selected node to the true label, via a monotone risk function and the standard conformal risk-control calibration procedure. We adapt this to our multi-label setting by defining risk as the per-patient maximum graph-distance to the nearest covering node over the patient's full true-label set, mirroring the max-score structure of Eq. 4, and calibrate λ via CRC's own $\inf\{\lambda : \hat{R}(\lambda) \leq \alpha - (1-\alpha)/n\}$ rule. The resulting prediction set is collapsed to a minimal covering antichain, the same topmost-passing-node principle underlying TAPS's own Definition 4, rather than reporting the raw, uncollapsed union of passing nodes, which we found in early implementation to trivially inflate to the majority of the tree and convey no useful information. Under this fair, antichain-collapsed construction, CRC-Hierarchical achieves 98.1% joint coverage against the 90% target, at an average set size of 41.11 nodes, which is $2.4\times$ TAPS's 17.08. CRC's graph-distance risk objective is theoretically valid on its own terms, but translates into substantial practical over-coverage

once required to produce a genuinely compact, non-redundant hierarchical output, rather than the tightly calibrated coverage TAPS achieves through score-level integration.

I.6 DEPTH-STRATIFIED BONFERRONI (DSB)

To test whether naive level-wise correction across a domain ontology suffices in place of integrated LCA scoring, we construct Depth-Stratified Bonferroni (DSB) as a natural strawman baseline. DSB calibrates a separate quantile threshold at each depth of the label hierarchy, dividing the joint coverage budget α equally across the tree's maximum depth via Bonferroni correction. We collapse the resulting cascading predictions into a minimal covering antichain for a fair, non-degenerate comparison. DSB achieves 95.8% joint coverage against the 90% target, with an average set size of 39.16 nodes, which is substantially larger than TAPS despite its explicit depth-aware calibration.

I.7 CRSVP-R

Mortier et al.'s AISTATS 2026 revision extends CRSVP (Mortier et al., 2026) with a representation-complexity relaxation, CRSVP-r, allowing up to r disjoint ancestor nodes rather than a single node. This was proposed specifically to address multi-label comorbidity, the exact failure mode Appendix G documents for the base $r = 1$ method. We implement CRSVP-r faithfully to this revised construction: the dynamic-programming set $A_r(S_k; x)$ with the explicit nested-set constraint of their Eq. 8, and the randomized calibration/inference procedure of their Algorithm 3/4, adapted to our per-patient max-score joint protocol. We sweep $r \in \{2, 3, 5\}$, the range the original paper's own complexity analysis treats as practical.

CRSVP-r attains its 90% joint coverage target cleanly at every r (Table 5 and Table 7), and one might read this as resolving the comorbidity problem plain CRSVP could not. It does not. A diagnostic breakdown reveals the coverage target is met by a severe degeneracy: at $r = 5$ on MIMIC-IV, 88.6% of predictions are exactly $\{\text{root}\}$, and miscoverage rises monotonically with true-label comorbidity span, from 0.1% at span 1 to over 45% at span ≥ 10. CRSVP-r satisfies its marginal guarantee by trivially covering low-comorbidity patients with the root node while systematically discarding the highest-comorbidity, most clinically consequential patients into its permitted 10% error margin. This is the opposite of the multi-label robustness the relaxation was designed to provide. Average LCAD (4.11–4.20 on MIMIC-IV, out of max depth 6) confirms predictions sit almost entirely at or near the root regardless of r.

A common failure mode: the "free root." Three independently motivated baselines in this evaluation, despite their large differences in mechanism, converge to the same degenerate behaviour: Arm 1 of our ablation (Section 5.3), CSP's per-example ILP construction (Appendix I), and CRSVP-r above. Each optimizes or thresholds against a per-node quantity that estimates "the true label is somewhere in this node's subtree", with no term penalising how far up the tree that node resides. This shared property means that the root node is always a dominant or near-dominant choice: $\hat{P}(\text{root}) \approx 1.0$ for almost every instance because the root's subtree is the entire label space, and its cost in each method's respective objective (a size penalty, an ILP leaf-count term, or a representation-complexity budget) is minimal: exactly 1 in each case. The height-based abstraction cost of TAPS, $\lambda \cdot \text{height}(v)$, is there precisely to break this degeneracy: the root node passes calibration only when the model is truly, overwhelmingly confident, not by default, since $g_v(x)$ increases with $\text{height}(v)$. This pathology is not unique to our baseline suite: Fargion et al. (2026) independently report that a naive application of Goren et al. (2024)'s hierarchical ascent rule to flat class partitions results in the same collapse, a fifth, fully independent confirmation, across four separate research groups and problem settings, that unregularized hierarchical abstraction defaults to maximal generality unless the score itself is made to penalise it.

J ABLATION: EXTENDED ANALYSIS

Arm 1 becomes the $\{\text{root}\}$ for each patient under test: In the case of an unregularized, MAX-aggregated hierarchical output, a node passes only if *any* of its descendant leaves pass, and since real patients usually have multiple comorbid conditions in the leaves, it almost certainly spreads to

the root. Its trivial 1.0000 coverage proves that *without a regularized score, the hierarchy of outputs alone is inadequate*.

Notably, CRSVP, an independently designed, unregularized output-only-based baseline, degenerates in the opposite direction (rather than trivial overcoverage from aggregation collapse, severe undercoverage from uncontrolled greedy descent). The failure of two structurally different output-only construction methods, through different mechanisms, is stronger evidence for the need for score-level integration than the failure of either one alone: Arm 2 ensures that the score-level LCA penalty is the only one that protects against root-collapse (0% roots, depth 4.44), but without hierarchical output construction the set size becomes large (36.00 nodes, comparable to the flat-LAC scale). Only Arm 3, combining both, achieves non-degenerate coverage through a compact set (17.02 nodes), ensuring that the two design choices of TAPS are complementary, not redundant. Note that Arm 2 and Arm 3 share the same structurally identical joint coverage (0.8886), as they use the same calibrated threshold and differ only in whether passing internal nodes are emitted directly (Arm 3) or expanded into their full set of descendant leaves (Arm 2); only the set size and LCAD are different.

K Conditional Coverage Across Age Subgroups

Marginal and joint coverage (Table 2) are guaranteed on average over the test distribution; they do not preclude systematic deviation within subgroups. As a preliminary check, we compute joint coverage separately for three age bands (< 45: n=1,987; 45–75: n=7,067; > 75: n=2,218, using `anchor_age` from MIMIC-IV). Flat LAC's maximum deviation from the 0.90 target across these subgroups is 2.0%; TAPS's is somewhat larger at 2.9%, concentrated in the < 45 (0.8767, 95% CI $[0.8622, 0.8912]$, undercoverage) and > 75 (0.9288, 95% CI $[0.9181, 0.9395]$, overcoverage) groups; both intervals exclude 0.90, so these are measured deviations rather than sampling noise.

We traced a plausible mechanism for this gap. The two subgroups differ sharply in label composition: < 45 patients are dominated by mental/behavioural (ICD-10 Chapter F) diagnoses (46.0% prevalence, their most common chapter), while > 75 patients are dominated by a narrower, more homogeneous profile of circulatory (89.9%), endocrine (74.4%), and health-status-contact (78.2%) codes. On true-positive instances, the model's mean predicted probability is markedly lower for Chapter F codes (0.62) than for Chapter I codes (0.73), representing a 10.7-point gap. Lower per-node confidence on Chapter F diagnoses plausibly makes them harder to cover within the calibrated threshold, contributing to the < 45 group's undercoverage; conversely, the elderly group's homogeneous, high-confidence profile is easily covered, contributing to its overcoverage.

L Tree Traversal Illustration

Figure 3 illustrates a sample traversal of the ICD-10 hierarchy by TAPS under moderate uncertainty.

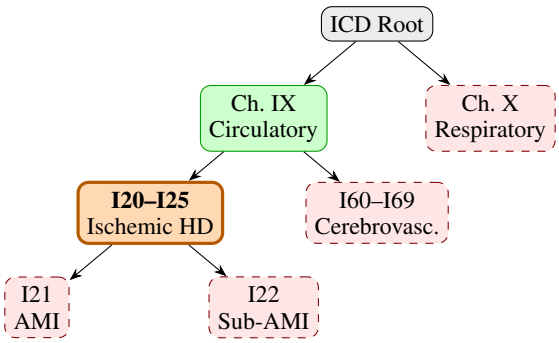


Figure 3: TAPS traversal under moderate uncertainty. **Green**: traversed. **Orange/bold**: the topmost passing node (Block I20–I25). **Red dashed**: branches not emitted.

M Ancestor-Height Distance as an Ultrametric

Proposition 7 (Ultrametric Property)**.** *Define the* ancestor-height distance $d_{\mathrm{AH}}(u, v) = d_{\max} - \mathrm{depth}(\mathrm{LCA}(u, v))$ *for* $u, v \in V$. *Then* d_{AH} *satisfies the ultrametric inequality:* $d_{\mathrm{AH}}(u, w) \leq \max(d_{\mathrm{AH}}(u, v),\ d_{\mathrm{AH}}(v, w))$ *for all* $u, v, w \in V$.

Proof. $\mathrm{depth}(\mathrm{LCA}(u, w)) \geq \min(\mathrm{depth}(\mathrm{LCA}(u, v)), \mathrm{depth}(\mathrm{LCA}(v, w)))$ by the tree property. Negating and adding $d_{\max}$ gives the ultrametric inequality. □

Note on relationship to LCAD. The ancestor-height distance d_{AH} defined here is a separate geometric quantity from the headline LCAD metric used throughout the paper. The headline metric is $d_{\mathrm{LCA}}(c, y^*) = \mathrm{depth}(y^*) - \mathrm{depth}(\mathrm{LCA}(c, y^*))$, which equals zero for an exact prediction and is positive otherwise. d_{AH} instead measures height from the LCA to the root and is non-zero even for exact matches on non-root nodes. This appendix uses d_{AH} purely for the geometric observation below; no validity or efficiency result depends on it, and the LCAD numbers reported in all tables use the implementation-matching formula stated in Section 5.

Connection to Theorem 1. Proposition 7 offers a *geometric motivation* for using LCA distance as a nonconformity penalty: the ultrametric (strong triangle) inequality means nearby nodes in the hierarchy receive proportionally smaller penalties. This is strictly stronger than the ordinary triangle inequality, and natural for tree-structured label spaces. None of the validity or efficiency results depend on it.

N Stratified Calibration Extension

To address the single-global-threshold limitation discussed in Section 5.4, for each ICD-10 chapter $c \in \{1, \ldots, 21\}$:

$$\hat{q}_c = \mathrm{Quantile}\left(\{ s_{\mathrm{LCA}}(X_i, y_i^*) : y_i^* \in \mathrm{Chapter}_c \}_{i=1}^{n_c};\ \frac{\lceil (n_c + 1)(1 - \alpha) \rceil}{n_c} \right) \tag{10}$$

Coverage for rare chapters benefits from the conservative bound $1/(n_c + 1)$; Theorem 4 remains valid per-chapter.

O Generalization to a Taller Hierarchy (MeSH)

To evaluate cross-domain generalizability and robustness to structural differences in the hierarchy, we conduct a secondary evaluation on a substantially taller ontology. We construct a 593-node hierarchy using Medical Subject Headings (MeSH) descriptor labels from a publicly available corpus of PubMed abstracts, with a maximum depth of 13, considerably taller than the MIMIC-IV ICD-10 tree (max depth 6). We apply the identical Phi-3-mini and QLoRA architecture used for MIMIC-IV, training for 2 epochs on 40,000 abstracts (an 80/10/10 train/calibration/test split), matching MIMIC-IV's own 2-epoch protocol exactly rather than the abbreviated single-epoch schedule used in an earlier version of this evaluation.

An initial training run using the project's shared encoder architecture surfaced a subtle but consequential extraction bug, present across all datasets in this project but only mathematically capable of corrupting predictions when a substantial fraction of documents are short relative to the model's maximum sequence length. Left-padded sequences shorter than half the maximum length caused the original implementation to extract a padding-token hidden state rather than the document's final token. On MIMIC-IV and the arXiv benchmark, document lengths are long enough relative to their respective truncation limits that this defect provably could not manifest; on MeSH, where PubMed abstracts are considerably shorter, roughly 29% of documents fell below the affected threshold, measurably corrupting predictions for a comparable fraction of the test set. We corrected the extraction logic to a padding-side-agnostic formula and retrained. We additionally identified and corrected a numerical precision issue specific to this benchmark's extreme class-imbalance weighting (positive-class weights up to 100), which caused gradient computations in mixed-precision training to

intermittently overflow; casting the document representation to full precision before the classification head resolved this and materially improved training stability.

On the corrected model, per-example diagnostics before conformal calibration show the underlying classifier has learned substantial hierarchical structure despite the task's difficulty: a hierarchical top-1 accuracy of 97.46% (the single highest-probability node is a true label or one of its ancestors), though exact leaf-level top-1 accuracy remains modest at 10.81%, reflecting genuine task difficulty at depth 13 across 593 nodes. Mean rank of the highest-confidence true label improved to 6.78 from a considerably worse figure under the original single-epoch protocol.

The raw, uncalibrated model exhibits an acute instance of the free-root pathology discussed in Appendix I: because hierarchical label propagation makes the root node trivially positive for every training example, its predicted probability is structurally pinned near 1.0 regardless of input, essentially independent of any genuine signal. As training sharpens the model's discrimination elsewhere (lower entropy, lower mean probability across all other nodes), the root's artificially high, unearned probability wins the unconstrained arg max more often rather than less, reaching 34.42% of raw test predictions, concentrated somewhat more heavily among documents with higher true-label cardinality (mean 4.21 true labels for root-argmax documents versus 3.74 for non-root). This is not evidence of degraded model quality but a direct, mechanical consequence of the same structural property already documented for Arm 1, CSP-aggregate, and CRSVP-r elsewhere in this paper.

TAPS's non-degeneracy screen (Remark 2) detects and corrects this directly. The two-stage selection sweep shows root-collapse falling monotonically as λ increases (95.3% at $\lambda = 0$, 73.4% at $\lambda = 0.01$, 32.1% at $\lambda = 0.02$, reaching exactly 0.0% at $\lambda = 0.05$, the selected operating point), at which TAPS achieves 99.44% test coverage with an average set size of 8.67 nodes and LCAD of 0.1698. A mechanical audit of individual prediction sets confirms every emitted node genuinely satisfies the calibrated threshold ($\hat{q} = 0.3754$) rather than being an artifact of the topmost-passing-node construction.

This audit also surfaces a distinct, milder pathology worth disclosing plainly. Because per-node probabilities are trained as independent marginals under hierarchical label propagation, nothing in this training objective enforces mutual exclusivity between orthogonal branches of the tree. On some documents, the raw model expresses high independent confidence in entire branches unrelated to the true label; we observed cases where TAPS's prediction set included nodes such as an unrelated top-level category at height 4 with raw model confidence exceeding 0.997, despite no true label existing anywhere beneath it. TAPS's score faithfully surfaces these high-confidence false positives exactly as specified, since its guarantee is calibrated coverage, not per-node correctness. Despite this behavior, TAPS achieves an LCAD of 0.1698 on MeSH, the best leaf-level specificity among all four datasets evaluated in this paper. This orthogonal branch overconfidence is a plausible contributor to false-positive branch inclusions, and is distinct from, though related to, the free-root pathology: root-collapse arises from one node's structurally guaranteed probability, while this arises from the base classifier's independent overconfidence across unrelated branches, a property of the marginal training objective rather than of TAPS's calibration mechanism.

Table 6 presents the full 13-method comparison on this corrected model. As with the MIMIC-IV and arXiv evaluations, the baselines we test that lack TAPS's score-level integration either fail to control coverage, degenerate toward the root, or inflate to impractically large sets. RAPS's flat, cumulative-mass calibration produces both severe under-coverage (23.10%) and enormous sets (142.81 nodes): MeSH's 593-node probability surface is too dispersed for a flat threshold to concentrate on the correct leaf, while RAPS's exact-match criterion cannot credit a correct ancestor prediction. CRC-Hierarchical's graph-distance risk calibration bottoms out at a zero threshold on this calibration set, forcing near-exhaustive 282-node sets that still fail to bound risk on 24.05% of test examples. CSP-Conformal, like HCC, collapses toward the root under this benchmark's severe class imbalance, emitting exactly {MESH_ROOT} for 85.84% of test examples; its conformal guarantee holds only because the root trivially covers every true label, the same pathology discussed throughout Appendix I.

Table 6: MeSH Conformal Baseline Evaluation (Phi-3-mini, 2-Epoch Model)

Method	Semantic Joint Cov	Avg Size	LCAD	% Root Collapse
Flat LAC	0.9878	10.63	0.2642	–
Flat APS	0.9904	112.42	0.1069	–
RAPS	0.2310	142.81	0.1923	–
CRSVP	0.0696	1.00	2.4341	–
HCC (Max/Marginal)	1.0000	1.02	1.0000	99.96%
HCC (Min/Joint)	1.0000	1.02	1.0000	99.96%
BR-MLCP	0.9898	15.28	0.1824	0.00%
Fisch et al. (Top-k)	0.9894	15.00	0.1887	0.00%
CRC-Hierarchical †	0.7595	282.00	0.0000	0.00%
CGD	1.0000	35.16	0.0000	0.00%
TreeCP	0.2628	31.62	0.0577	0.00%
CSP (aggregate)	0.9084	174.00	0.7465	0.00%
CSP (per-example) ‡	0.9094	1.05	4.9430	85.84%
TAPS (Ours)	**0.9944**	**8.67**	**0.1698**	**0.00%**

Note: TAPS selected $\lambda = 0.05$ ($\hat{q} = 0.3754$). Root collapse for flat methods is not applicable.

† **CRC-Hierarchical**: CRC calibrates a graph-distance risk threshold $\hat{l}$. On this dataset the calibration bottomed out at $\hat{l} = 0.0000$, meaning no threshold existed that could simultaneously bound the graph-distance risk to ≤ 0.10 on the calibration set. At $\hat{l} = 0$, CRC includes every node whose min-path probability exceeds 0 – effectively emitting half the taxonomy (282 nodes). The model's root-skewed probabilities mean true deep-leaf paths have very low joint probability, so even 282-node sets fail to cover them by the hierarchical criterion 24.05% of the time.

‡ **CSP (per-example)**: CSP (per-example)'s ILP calibration, when faced with a root-skewed probability distribution, finds that the minimum-cost tree-consistent set satisfying the conformal risk bound is almost always a single shallow node (85.84% exactly `MESH_ROOT`). The conformal guarantee still holds (0.9094 joint coverage > 0.90) because `MESH_ROOT` is an ancestor of all true labels, so it trivially covers every example. This is the same root-collapse pathology as HCC, just reached via a different algorithmic path.

P Cross-Domain Validation: arXiv Category Hierarchy

To test whether TAPS's gains reflect the algorithm rather than an artifact of clinical text or ICD-10's specific structure, we run a third, independently-constructed benchmark under the *identical* protocol, architecture, and baseline suite as the primary MIMIC-IV study (all five baselines, three seeds, mean $\pm$ std, two-stage λ selection, joint-coverage calibration). Unlike the MeSH evaluation (Appendix O), which is a secondary robustness check, we report this with the same weight as the primary result.

Tree and data. We build a tree directly from arXiv's public category taxonomy (Group $\rightarrow$ Archive $\rightarrow$ Category), a maintained ontology rather than a learned clustering: 177 nodes, 155 leaf categories, max depth 3, which sits at the shallow end of the spectrum, complementing ICD-10 (depth 6) and MeSH (depth 13). Like ICD-10, it is not depth-regular (single-archive groups collapse to depth 2; Physics, the only multi-archive group, reaches depth 3), re-exercising the height-not-depth fix (Remark 1) on a different tree. We collect abstracts and category metadata via arXiv's OAI-PMH interface (primary category in `cs`/`stat`/`math`), using primary *and* cross-listed categories as the true label set, which is the source of genuine multi-label comorbidity here, since a substantial fraction of papers cross-list under disjoint top-level Groups (e.g. `cs.LG` with `hep-ph`). We split chronologically by submission date: 80,024/17,148/17,148 train/cal/test, 256-token abstracts. The chronological ordering means the test set is drawn from a later time period than the calibration set; this intentional temporal gap serves as a distribution-shift stress test, and coverage figures should be read as empirical rather than as formal exchangeability-backed guarantees (which require exchangeable calibration–test pairs). *Deviation from the primary protocol:* due to compute constraints this benchmark trains for 1 epoch rather than 2; we did not observe evidence of undertraining (validation loss had plateaued), but did not run a controlled 2-epoch comparison.

HCC-min does not root-collapse here. On ICD-10, HCC-min collapses to $\{\text{root}\}$ for $\approx 95\%$ of patients (Appendix H) because deep, fragmented trees drive the Bonferroni-corrected threshold so strict that no alternative cover remains *feasible*, leaving root as the only cover that can satisfy

Table 7: Results at α=0.10 (arXiv `cs`/`stat`/`math` subset, 1 epoch), aggregated over 3 seeds (mean $\pm$ std). Protocol identical to Table 2; all three seeds selected λ=0.0. This indicates that for this dataset, the base model does not suffer from the severe root-collapse pathology seen in ICD-10, demonstrating that the two-stage non-degeneracy screen correctly serves as an active defense when collapse threatens, rather than a mechanism that artificially forces a positive λ globally. HCC uses β=0.5, matching the ICD-10 protocol. The evaluation code for RAPS, both HCC variants, Fisch et al., CRC-Hierarchical, and DSB computes only joint coverage (marginal N/A).

Method	Semantic Joint Cov.↑	Marg. Cov.	Avg. Size↓	LCAD↓
Flat LAC (Sadinle et al., 2019)	0.8993 ± 0.0005	0.9443 ± 0.0003	12.08 ± 0.12	0.0825 ± 0.0007
Flat APS (Romano et al., 2020)	0.9129 ± 0.0037	0.9332 ± 0.0030	108.23 ± 1.99	0.1088 ± 0.0028
RAPS (Angelopoulos et al., 2021)	0.9125 ± 0.0036	N/A	101.72 ± 2.34	0.1091 ± 0.0026
CRSVP (Mortier et al., 2026) (single-node)	0.2808 ± 0.0024	0.3645 ± 0.0030	$\mathbf{1.00 \pm 0.00}$	0.8556 ± 0.0047
HCC (den Hengst et al., 2025) (max, marginal-native)	0.8679 ± 0.0060	N/A	6.53 ± 0.23	0.5861 ± 0.0311
HCC (den Hengst et al., 2025) (min, joint)	0.9722 ± 0.0003	N/A	11.09 ± 0.19	0.4750 ± 0.0145
BR-MLCP (Lambrou & Papadopoulos, 2016)	0.8629 ± 0.0014	0.9084 ± 0.0005	19.78 ± 0.11	0.1002 ± 0.0009
Fisch et al. (Fisch et al., 2022) (Top-k)	0.9059 ± 0.0036	N/A	15.67 ± 0.47	0.1145 ± 0.0025
CRC-Hierarchical (Angelopoulos et al., 2024)	0.9672 ± 0.0009	N/A	21.54 ± 0.23	0.0254 ± 0.0006
DSB	0.9575 ± 0.0004	N/A	20.12 ± 0.06	0.0575 ± 0.0003
CRSVP-r (r=2) (Mortier et al., 2026)	0.9001 ± 0.0004	N/A	1.1236 ± 0.0013	1.8519 ± 0.0024
CRSVP-r (r=3) (Mortier et al., 2026)	0.9006 ± 0.0019	N/A	1.2948 ± 0.0039	1.7995 ± 0.0037
CRSVP-r (r=5) (Mortier et al., 2026)	0.9020 ± 0.0006	N/A	1.6408 ± 0.0046	1.7358 ± 0.0042
Tree-CP (Tyagi & Guo, 2023)	0.9051 ± 0.0023	0.9338 ± 0.0023	13.76 ± 0.49	0.1056 ± 0.0049
CSP (aggregate) (Zhang et al., 2025a)	0.9063 ± 0.0010	0.9291 ± 0.0006	23.77 ± 0.11	0.3746 ± 0.0137
CSP (per-example) (Zhang et al., 2025a)	0.9111 ± 0.0139	0.9007 ± 0.0160	3.22 ± 0.40	1.8845 ± 0.0605
TAPS (Ours)	0.8999 ± 0.0008	0.9283 ± 0.0005	$\mathbf{1.65 \pm 0.02}$	1.0541 ± 0.0203

the calibration target. On arXiv, root's structural cost is comparable ($1 + 0.5\times155 = 78.5$), but with only 4 NOL-cover levels (vs. ICD-10's 7) the correction is milder and Group-level covers stay both feasible and cheaper than root even at maximum correction; tracing 100 test examples at their corrected thresholds, root's fixed cost never won against a feasible Group-level alternative. We flag one caveat directly: this same feasibility-driven avoidance does not imply HCC-min's selected covers are *correct*; in traced examples it sometimes preferred a cheap, confidently-wrong leaf cover over an infeasible-but-correct alternative, since the cost function is blind to correctness. This is a property of HCC's design, not TAPS's.

Ablation. The three-arm ablation (Table 2's protocol) reproduces the ICD-10 pattern exactly: Arm 1 (flat score, hier. output) collapses to 100% root; Arm 2 (LCA score, flat output) escapes collapse but yields sets substantially larger than Flat LAC (43.13 ± 2.21 vs. 12.08); Arm 3 (full TAPS) is both non-degenerate and compact (1.65 ± 0.02, 1.72 ± 1.89% root). CRSVP again fails on multi-label comorbidity by a different mechanism than on ICD-10 (28.1% vs. ICD-10's 3.7% joint coverage, still far below target), and HCC's two variants again fail in opposite directions.

Takeaway. Together with MeSH (Appendix O), this gives three points on the tree-height spectrum: shallow (3), mid (6), and tall (13), rather than a single clinical-coding result: TAPS attains near-nominal empirical semantic joint coverage with substantially smaller sets than every flat baseline under an unchanged algorithm, both design choices remain independently necessary, and single-node/Bonferroni-heavy baselines fail on multi-label comorbidity by mechanisms that persist across tree shapes even as their precise manifestation shifts with height and branching structure.

Q DBpedia Full Evaluation

To evaluate the stability and structural awareness of TAPS under extreme predictive confidence, we perform a comprehensive evaluation on the DBpedia14 dataset (Zhang et al., 2015) and its associated ontology (Mendes et al., 2012). DBpedia has a 26-node class hierarchy with a maximum depth of 4, which represents a more condensed abstraction task than the larger DAGs in MIMIC and arXiv.

The base classifier achieved near-perfect convergence, consistently producing 99.2% top-1 leaf accuracy and a macro ROC-AUC > 0.999 across three independent training seeds. This creates a challenging edge case for conformal evaluation: the predictive probability mass is heavily concentrated at the actual leaf nodes ($p \approx 1.0$), which pushes the empirical conformal quantile $\hat{q}$ close to near-zero.

Table 8 presents the aggregate conformal evaluation metrics across three independent training seeds, where the mean and population standard deviation (ddof $= 0$) are specified.

Table 8: DBpedia Conformal Results (3-Seed Aggregate, $\alpha = 0.10$).

Method	Semantic Coverage	Set Size	LCAD
Flat LAC	0.9012 ± 0.0023	0.9033 ± 0.0020	0.0045 ± 0.0005
Flat APS	0.9986 ± 0.0002	1.9135 ± 0.0010	0.0031 ± 0.0004
RAPS	0.9985 ± 0.0002	1.9133 ± 0.0047	0.0032 ± 0.0005
CRSVP	0.9985 ± 0.0006	1.0000 ± 0.0000	0.6734 ± 0.0272
HCC (Max/Marginal)	0.9671 ± 0.0022	0.9700 ± 0.0000	0.0207 ± 0.0034
HCC (Min/Joint)	0.9671 ± 0.0022	0.9700 ± 0.0000	0.0207 ± 0.0034
BR-MLCP	0.9043 ± 0.0050	0.9065 ± 0.0052	0.0047 ± 0.0004
CSP (aggregate)	0.9006 ± 0.0018	9.7604 ± 0.1085	0.6847 ± 0.0670
CSP (per-example)	0.9382 ± 0.0102	1.1447 ± 0.0939	2.5859 ± 0.0595
Fisch et al. (Top-k)	0.9922 ± 0.0001	1.0000 ± 0.0000	0.0146 ± 0.0017
CRC-Hierarchical	0.9971 ± 0.0006	1.0208 ± 0.0034	0.0490 ± 0.0007
CGD	0.9971 ± 0.0004	1.0000 ± 0.0000	0.1161 ± 0.0018
TAPS (Ours)	$\mathbf{0.8973 \pm 0.0023}$	$\mathbf{0.8983 \pm 0.0024}$	$\mathbf{1.4020 \pm 0.0060}$

Q.1 Baseline Instabilities Under High Confidence

In this high-confidence regime, existing hierarchical baselines exhibited severe failure patterns that Flat baseline methods avoided.

The CSP Infeasibility Cliff. The original threshold-calibrated CSP framework computes a single global threshold τ to bound the sum of selected node probabilities. On DBpedia, τ calibrates extremely low (≈ 2.19). Because nodes on the path sum to large values when the model is confident (e.g., 1.0 at the leaf plus large values at ancestors), shallow classes like *MeanOfTransportation* (depth 1) become computationally infeasible to cover without exceeding the τ limit, collapsing coverage for shallow nodes to near-zero. Conversely, deeper nodes are vastly over-covered, causing the average set size to artificially bloat to 9.76.

CSP (per-example) Root Collapse. The fully conformal extension of CSP avoids global thresholding by computing per-example budgets, but degenerates entirely under near-perfect accuracy. Because the true leaf natively satisfies coverage but costs 1.0, the ILP prefers to select the uninformative $\{\texttt{root}\}$ node whose probability is low (e.g. $p \approx 0.15$). Consequently, the calibration quantile $\hat{q}$ drops to near-zero (< 0.001). At test time, restricting the ILP to such a microscopic budget causes the solver to become infeasible for up to 85% of examples. When the ILP is infeasible, it falls back to a trivial failsafe: outputting exactly $\{\texttt{root}\}$. This mirrors the exact failure signature observed for HCC (Min/Joint) on MIMIC and arXiv, functionally becoming the sixth confirmed instance of methods trivially guaranteeing coverage by defaulting to the least-informative abstract node.

Q.2 TAPS Stability and Theorem 1

In contrast, TAPS exhibits strong stability, maintaining coverage (0.8973 ± 0.0023) that is consistent with the flat control while maintaining a root-collapse rate of exactly 0.00% across all seeds.

Importantly, TAPS's hierarchical widening (LCAD 1.40) is consistent with the score construction and the dominance property in Theorem 1. In the high-confidence regime ($\hat{q} \approx 0$), the actual passing condition $g_v(x) \leq \hat{q}$ is extremely restrictive: a node passes only if its combined uncertainty and height penalty is essentially zero, which in practice selects ancestors whose predicted probability is near 1.0. Since Algorithm 1 stops at the most abstract passing node in the path, TAPS naturally takes predictions to ancestors instead of leaves, and returns sets like $\{\text{Animal}\}$ instead of $\{\text{Dog}\}$. This accurately expresses structural certainty without artificially increasing the size of the set or resorting to trivial global abstractions.