RETRIEVE, REASON, RECOVER: EVIDENCE-GROUNDED ICD CLINICAL CODING UNDER LONG TAILS AND MISSING LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated International Classification of Diseases (ICD) coding from clinical narratives faces three interrelated challenges; severe long-tail code distributions with sparse training examples, systematically incomplete labels from clinical undercoding, and a lack of interpretable evidence linking predictions to documentation. We propose HERA-ICD, a two-stage framework that reframes ICD coding as retrieval-augmented classification. Stage 1 performs hierarchical retrieval at document, code, and span granularities, enabling prototype-based few-shot and description-based zero-shot prediction for rare codes. In Stage 2, we construct per-instance graphs, whose topologies dynamically adapt to the retrieved evidence. A Graph Attention Network then reranks the candidates using document-specific reasoning. To handle missing labels, we integrate positive-unlabeled learning with graph-based label completion, propagating soft positives through the ICD hierarchy. We theoretically prove three key properties of our framework: the non-negative PU objective is unbiased and consistent, the generalization gap of the retrieval-then-rerank architecture is independent of vocabulary size, and prototype scoring generalizes from fewer than ten examples. Consequently, experiments on MIMIC-III and MIMIC-IV demonstrate consistent improvements over state-of-the-art methods, particularly on rare codes, while providing traceable evidence linking predictions to specific clinical spans.

1 INTRODUCTION

Automated International Classification of Diseases (ICD) coding uses artificial intelligence and natural language processing to transform unstructured clinical notes into standardized alphanumeric disease codes for applications such as billing, electronic health record management, and epidemiological research. Automated ICD coding systems achieve strong micro-F1 scores on frequent diagnoses. For example, PLM-ICD achieves 59.8% on the MIMIC-III full code set (Huang et al., 2022); yet, macro-F1 across these systems remains strikingly low, exposing a fundamental failure on the long tail of rare codes that constitute over half the ICD vocabulary (Edin et al., 2023). This disparity matters clinically, because rare codes often correspond to uncommon but serious conditions where coding errors have outsized consequences for patient care, reimbursement accuracy, and epidemiological tracking. We argue that the root cause is not insufficient model capacity, but rather the absence of three capabilities that human coders rely on: retrieving relevant evidence from documentation, reasoning over structured medical knowledge, and recognizing when a diagnosis is plausible but undocumented.

Figure 1 illustrates the three gaps. Current approaches address these needs only partially. Transformer models (Huang et al., 2022; Yuan et al., 2022; Li et al., 2023) rely on attention-based attribution rather than explicit evidence retrieval; graph methods (Cao et al., 2020; Zhou et al., 2021; Michalopoulos et al., 2022) use static code graphs shared by every document; knowledge-enhanced methods (Ren et al., 2025; Ge et al., 2024; Zhang et al., 2025) do not treat retrieval as a measurable IR problem; and although clinical documentation is systematically incomplete (Jin et al., 2023), only one prior work applies PU learning to ICD coding, without retrieval or label completion. No existing framework addresses all of these challenges jointly, and none offers finite-sample guarantees.

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

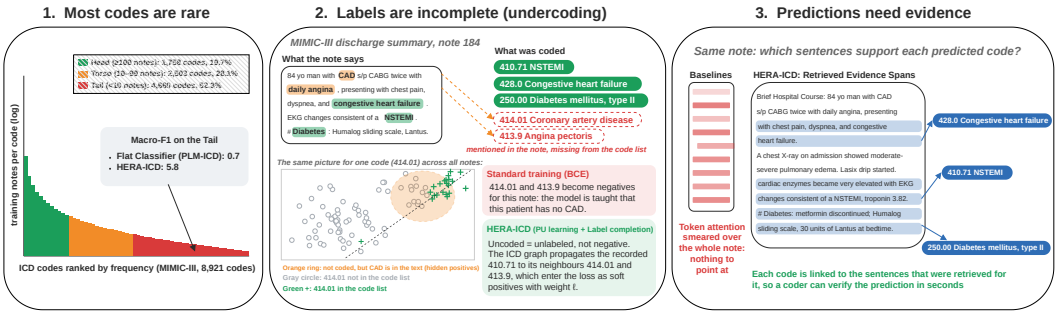

Figure 1: The three gaps in automated ICD coding. In MIMIC-III note 184, coronary artery disease and angina are documented but never coded, so standard training treats them as negatives.

We propose HERA-ICD (Hierarchical Evidence Retrieval and Adaptive Reranking for ICD Coding), a two-stage framework that reframes ICD coding as a retrieval-augmented classification problem. The key insight is that ICD coding, much like open-domain question answering, benefits from an explicit pipeline that *retrieves* evidence at multiple granularities, *reasons* over the retrieved context to rank codes, and *recovers* the codes that clinical documentation left unrecorded. Unlike end-to-end classification, this pipeline offers two key advantages: it yields traceable evidence, and it decouples evidence gathering from label assignment. As a result, each stage can be independently evaluated and improved.

Concretely, HERA-ICD operates across three levels of retrieval: document-level retrieval of similar notes via a FAISS index, code-level retrieval of code-specific examples for few-shot and zero-shot scoring, and span-level retrieval of text spans to ground each prediction using a bi-encoder and cross-encoder architecture. The retrieved evidence then conditions a per-instance graph whose nodes are candidate codes, evidence spans, and context notes, and a GAT reranks the candidates through document-specific reasoning.

To address missing labels, HERA-ICD integrates positive-unlabeled (PU) learning with graph-based label completion. Rather than treating unobserved codes as negatives, we propagate label information through the ICD hierarchy and co-occurrence graphs to assign soft positive weights to plausible missing codes, mitigating the bias against rare codes that standard cross-entropy training introduces. Our contributions are:

1. **Hierarchical multi-granularity retrieval** at the document, code, and span levels, enabling few-shot and zero-shot prediction with measurable retrieval quality.
2. **Evidence-conditioned dynamic graph reranking** that builds a per-instance graph from the retrieved evidence, so relational reasoning adapts to the document rather than applying a static code graph.
3. **Non-negative PU learning with graph-based label completion** that propagates labels through the ICD hierarchy and co-occurrence structure to handle incomplete clinical documentation.
4. **Comprehensive experiments** on MIMIC-III and MIMIC-IV improve macro-F1 by 3.3 points overall and 3.8 on rare codes, with ablations confirming each component, retrieval metrics quantifying each granularity, and a faithfulness evaluation showing that retrieved evidence aligns with expert-annotated clinical rationales.

We prove this design is sound in §4, where the PU objective is unbiased and consistent under a missing-completely-at-random model of undercoding, the generalization gap of the reranker is set by the reranker's own size and not by $|\mathcal{C}|$, prototype scoring for rare codes has excess risk $O(1/\sqrt{m})$, and label propagation converges geometrically.

2 Related Work

Automated ICD coding is dominated by three families of methods, and each leaves open a gap this paper addresses. Attention and transformer models, from CAML (Mullenbach et al., 2018) to PLM-ICD (Huang et al., 2022), encode the note well but attribute predictions with attention weights rather than retrieved evidence, and collapse on the long tail (Edin et al., 2023). Graph

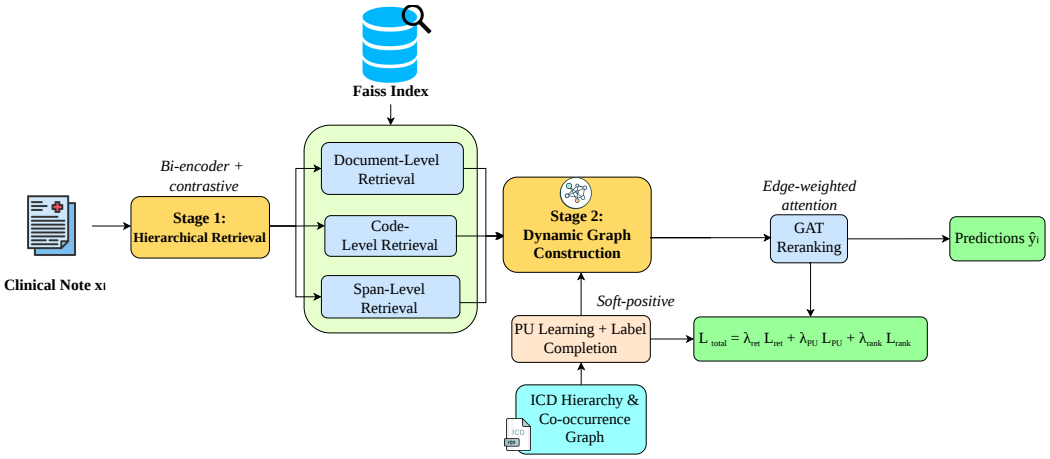

Figure 2: Overview of HERA-ICD. Stage 1 retrieves evidence at the document, code, and span level from FAISS indices. Stage 2 builds a per-instance graph and reranks the candidates with a GAT. PU learning with label completion supplies soft positives during training.

methods such as HyperCore (Cao et al., 2020) apply one *static* code graph to every document, so relational reasoning cannot adapt to the content of a particular note. Knowledge-enhanced and retrieval methods, among them TraceCoder (Ren et al., 2025), the retrieve-and-rerank model of Niu et al. (2023b) and the concurrent agentic system RAG-Coding (Gan et al., 2026), inject external knowledge but do not treat retrieval as a first-class information retrieval problem with measurable quality at several granularities, and none corrects for missing labels. On that last gap, Jin et al. (2023) is the only prior use of positive-unlabeled learning for ICD coding, and it reweights the loss without retrieval or label completion; the single positive label setting of Cole et al. (2021) is the special case of ours in which exactly one code per note is observed, while Xu et al. (2017) and Lee et al. (2025) give generalization and sample-complexity bounds for the positive-unlabeled setting that we extend to the multi-label case. Guo (2025) bound the risk of retrieval-augmented *generation*, whereas our Theorem 1 bounds a retrieval-augmented *classifier* in terms of the size of the retrieved graph. HERA-ICD is the first framework that closes all three gaps together, and the first to give unbiasedness and consistency guarantees for ICD coding under undercoding. Appendix D gives a full treatment of each family and covers few-shot and zero-shot ICD coding.

3 Methodology

Figure 2 overviews HERA-ICD. Stage 1 performs hierarchical retrieval (§3.2). Stage 2 reranks candidates via an evidence-conditioned graph (§3.3). PU learning with graph-based label completion handles missing labels during training (§3.4). Pseudocode for both stages is given in Appendix A.

3.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ be a dataset of clinical notes where $\mathbf{y}_i \in \{0,1\}^{|\mathcal{C}|}$ indicates ICD codes from vocabulary $\mathcal{C}$. Each input $\mathbf{x}_i$ is the free-text portion of a hospital discharge summary (the standard MIMIC input). Critically, $\mathbf{y}_i$ is *incomplete*, and systematic undercoding results in a significant number of reportable clinical conditions remaining undocumented (Jin et al., 2023). We partition $\mathcal{C}$ into frequent codes $\mathcal{C}_f$ ($\geq\tau$=10 examples), rare codes $\mathcal{C}_r$ ($< \tau$ examples), and unseen codes $\mathcal{C}_u$ (zero examples). On MIMIC-III, $|\mathcal{C}_r \cup \mathcal{C}_u|$ exceeds 50% of the vocabulary (Edin et al., 2023). HERA-ICD maps $\mathbf{x}_i$ to $\hat{\mathbf{y}}_i \in [0,1]^{|\mathcal{C}|}$ in two stages. Stage 1 retrieves candidate codes $\mathcal{C}_i^{\text{cand}} \subseteq \mathcal{C}$ and evidence $\mathcal{E}_i$ at multiple granularities. Stage 2 constructs an evidence-conditioned graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{A}_i)$ and reranks the candidates with a GAT.

3.2 Hierarchical Multi-Granularity Retrieval

All retrieval levels share a bi-encoder architecture with query encoder f_q and candidate encoder f_c, both initialized from Clinical-Longformer (Li et al., 2023). The query encoder f_q takes the clinical-note text, while the candidate encoder f_c takes the textual code description $\text{desc}(c)$ (the official

ICD title concatenated with its UMLS definition); both output 768-dimensional embeddings. Bi-encoders are trained using the InfoNCE contrastive loss $\mathcal{L}_{\text{ret}}$ (van den Oord et al., 2018) with cosine similarity, a learnable temperature τ, and a mix of in-batch and hard-mined negatives. $s_{\text{ret}}(\cdot)$ denotes the resulting score for a code, a note, or a span–code pair. To enable efficient retrieval, candidate embeddings are pre-computed and indexed using FAISS (Johnson et al., 2019) for sub-linear search.

Retrieval operates at three complementary granularities. At the *document level*, the K_d most similar training notes are retrieved, providing contextual prototypes whose label vectors supply an empirical prior over plausible codes. At the *code level*, for each candidate $c \in \mathcal{C}_i^{\text{cand}}$, we retrieve K_c code-specific examples (training notes labeled with c) from a per-code index, enabling prototype-based few-shot scoring. The prototype $\mathbf{p}_c = \frac{1}{K_c}\sum_{j=1}^{K_c} f_q(\mathbf{x}_j)$ is the mean encoder embedding of these retrieved notes, and the few-shot score is $s_{\text{fs}}(c) = \text{sim}(\mathbf{q}_i, \mathbf{p}_c)$ following Snell et al. (2017). For unseen codes (which have no training notes, so no prototype can be formed), we use the encoded code description directly as a zero-shot proxy, $s_{\text{zs}}(c) = \text{sim}(\mathbf{q}_i, f_c(\text{desc}(c)))$ (Rios & Kavuluru, 2018). At the *span level*, the note is segmented into overlapping spans, a bi-encoder retrieves the top-$2K_s$ spans per candidate code, and a cross-encoder reranker keeps the K_s most relevant, balancing efficiency with precision (Algorithm 1, Appendix A). The evidence set $\mathcal{E}_i = \mathcal{D}_i \cup (\bigcup_c \mathcal{E}_i^c) \cup \mathcal{S}_i$ aggregates all retrieved material for traceability. The three subsets are kept distinct because in Stage 2 each instantiates a different node type (context, code-example, and span nodes), so spans are not processed twice but rather *typed* separately when the graph is built.

3.3 Evidence-Conditioned Dynamic Graph Reranking

Unlike static graph methods (Cao et al., 2020; Michalopoulos et al., 2022) that apply fixed code relationships to all documents, we construct a graph *per instance* reflecting retrieved evidence.

Graph construction. For note $\mathbf{x}_i$, graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{A}_i)$ contains three node types. *Code nodes* carry features $\mathbf{h}_c = f_c(\text{desc}(c))$, *span nodes* carry features $\mathbf{h}_s = f_s(s)$, and *context nodes* come from retrieved notes. All features are projected to a shared d'-dimensional space via type-specific linear layers. Three edge types connect nodes. Code–code edges are weighted by the maximum of a hierarchy indicator, PMI co-occurrence, and semantic similarity,

$$w_{ab} = \max\big(\mathbf{1}[\text{parent}(c_a){=}c_b],\ \text{PMI}(c_a, c_b),\ \text{sim}(\mathbf{h}_{c_a}, \mathbf{h}_{c_b})\big), \tag{1}$$

pruned below threshold θ_{edge}; the maximum connects two codes whenever *any* of the three relations is strong, keeping the graph sparse while preserving the dominant relational signal. Span–code edges are weighted by $w_{sc} = s_{\text{ret}}(s, c) \cdot \text{sim}(\mathbf{h}_s, \mathbf{h}_c)$, and context–code edges connect retrieved notes to the codes in their label vectors (Algorithm 2). Every graph has at most $N = K + K_s + K_d + 1$ nodes, about 70 in our setting, a quantity that is independent of $|\mathcal{C}|$ and that enters the generalization bound in §4.

GAT reranking and score fusion. An L-layer GAT (Veličković et al., 2018) processes $\mathcal{G}_i$. We modify the attention proposed by Veličković et al. (2018) in one place, so that the retrieval-derived edge weight w_{vu} scales the attention coefficient of head m at layer l before it is normalized over the neighbourhood,

$$\alpha_{vu}^{(l,m)} = \frac{w_{vu}\exp(e_{vu}^{(l,m)})}{\sum_{u' \in \mathcal{N}(v)} w_{vu'}\exp(e_{vu'}^{(l,m)})}, \tag{2}$$

where $e_{vu}^{(l,m)}$ is the usual LeakyReLU attention logit; everything else, including multi-head concatenation, is unchanged. Evidence that the retriever scored highly therefore carries more weight in the aggregation, and this is the layer whose complexity Theorem 1 bounds. After L layers, the code-node representations $\mathbf{h}_c^{(L)}$ aggregate evidence from connected spans and related codes. A linear readout gives graph scores $s_{\text{graph}}(c) = \mathbf{w}_o^\top \mathbf{h}_c^{(L)} + b_o$. We fuse retrieval and graph scores through a learned, instance-specific gate:

$$\hat{y}_{i,c} = \sigma\big(g_c \cdot s_{\text{ret}}(c) + (1 - g_c) \cdot s_{\text{graph}}(c)\big), \quad g_c = \sigma\big(\mathbf{w}_g^\top [\mathbf{q}_i \,\|\, \mathbf{h}_c^{(L)}] + b_g\big), \tag{3}$$

where $g_c \in (0, 1)$ adapts per code and per instance, allowing the model to rely more on retrieval for well-represented codes and on graph reasoning for rare codes.

3.4 Positive-Unlabeled Learning with Graph Label Completion

Standard cross-entropy treats unobserved codes as negatives, yet clinical records are systematically undercoded (Jin et al., 2023). We address this through graph-based label propagation and a non-negative PU loss.

Label propagation. The global code graph $\mathcal{G}_{\text{code}}$ is built once over the full vocabulary, and its edge weights $w_{cc'}$ combine ICD parent–child links with the normalized PMI of code co-occurrence in the training labels, using the same $\max$ form as Eq. (1), with weights below θ_{edge} pruned. Let $\mathbf{W}$ be the row-normalized weight matrix of $\mathcal{G}_{\text{code}}$. For note i we initialize soft labels $\ell_{i,c}^{(0)} = \mathbb{1}[c \in \mathcal{P}_i]$ on the observed positives $\mathcal{P}_i$ and run T steps of the label-propagation iteration of Zhou et al. (2004),

$$\ell_{i,c}^{(t+1)} = \begin{cases} 1 & \text{if } c \in \mathcal{P}_i, \\ \alpha \sum_{c'} \mathbf{W}_{cc'}\, \ell_{i,c'}^{(t)} + (1-\alpha)\, \ell_{i,c}^{(0)} & \text{otherwise,} \end{cases} \tag{4}$$

where α is the propagation strength and observed positives stay clamped to one. Theorem 4 shows that this iteration converges geometrically to a unique closed-form fixed point, so the truncation at T is controlled. Codes with $\ell_{i,c}^{(T)} > \theta_{\text{soft}}$ become *soft positives* $\mathcal{S}_i^+$, capturing plausible missing labels (e.g., "coronary artery disease" when "acute myocardial infarction" is documented; Figure 5 in Appendix B).

Non-negative PU loss. In multi-label ICD coding, each code c defines a binary PU problem over the whole training set. Let $\mathcal{P}_c = \{i : c \in \mathcal{P}_i\}$ be the notes where c is observed, $\mathcal{S}_c^+ = \{i : c \in \mathcal{S}_i^+\}$ the notes where c is a soft positive, and $\mathcal{U}_c' = [n] \setminus (\mathcal{P}_c \cup \mathcal{S}_c^+)$ the remaining unlabeled notes. Let $\pi_c = \Pr(y_{i,c} = 1)$ be the class prior of c, and write $\ell^+(p) = -\log p$ and $\ell^-(p) = -\log(1-p)$. Following the unbiased PU risk of du Plessis et al. (2015) with the non-negative correction of Kiryo et al. (2017), we use the per-code estimator

$$\hat{R}_c = \frac{\pi_c}{|\mathcal{P}_c|} \sum_{i \in \mathcal{P}_c} \ell^+(\hat{y}_{i,c}) + \frac{1}{|\mathcal{S}_c^+|} \sum_{i \in \mathcal{S}_c^+} \ell_{i,c}^{(T)}\, \ell^+(\hat{y}_{i,c}) + \max\Big\{0,\ \frac{1}{|\mathcal{U}_c'|} \sum_{i \in \mathcal{U}_c'} \ell^-(\hat{y}_{i,c}) - \frac{\pi_c}{|\mathcal{P}_c|} \sum_{i \in \mathcal{P}_c} \ell^-(\hat{y}_{i,c})\Big\}, \tag{5}$$

and set $\mathcal{L}_{\text{PU}} = \sum_{c \in \mathcal{C}} \hat{R}_c$. The first term is the prior-weighted positive risk, the second is a weighted positive risk on propagated soft positives, and the third is the unbiased negative-class risk, estimated from unlabeled data by subtracting the positive contribution to the negative loss and thresholded at zero *at the aggregate level across the training set for each code*. Applying the correction per code and per dataset, rather than per instance, is what preserves the unbiasedness and consistency of the estimator (Proposition 1 and Theorem 2); in practice the sums are evaluated over mini-batches, as in Kiryo et al. (2017).

3.5 Training and Inference

The complete loss combines retrieval, PU, and ranking terms:

$$\mathcal{L} = \lambda_{\text{ret}}\, \mathcal{L}_{\text{ret}} + \lambda_{\text{PU}}\, \mathcal{L}_{\text{PU}} + \lambda_{\text{rank}}\, \mathcal{L}_{\text{rank}}, \tag{6}$$

where $\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}_i|\,|\mathcal{N}_i|} \sum_{c^+ \in \mathcal{P}_i} \sum_{c^- \in \mathcal{N}_i} \max(0,\, m - \hat{y}_{i,c^+} + \hat{y}_{i,c^-})$ is a margin ranking loss that pushes observed positives above negatives $\mathcal{N}_i$ sampled from candidate codes not in $\mathcal{P}_i$. Training proceeds in two phases. In Phase 1, the bi-encoders are trained with $\mathcal{L}_{\text{ret}}$ and the FAISS indices are built from the resulting embeddings. In Phase 2, the cached embeddings condition the reranker, and the GAT, fusion gate, and label-propagation module are trained jointly by minimizing $\lambda_{\text{PU}}\mathcal{L}_{\text{PU}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}$, while $\lambda_{\text{ret}}\mathcal{L}_{\text{ret}}$ is retained as a light fine-tuning term on the query encoder; label propagation is recomputed each epoch on the fixed $\mathcal{G}_{\text{code}}$. Candidate retrieval reduces the code space from $|\mathcal{C}|$ to K (typically 50), keeping graphs at $O(K + K_s + K_d)$ nodes; FAISS enables sub-linear retrieval, and graph construction is batched. At test time, each prediction is linked to retrieved spans and graph attention weights, providing traceable explanations.

4 Theoretical Analysis

We give five guarantees for HERA-ICD. Two are stated here. The corrected PU objective of Eq. (5) is unbiased (Proposition 1), and the generalization gap of the reranker is independent of the vocabulary

size (Theorem 1). Three further results are summarized at the end of this section and stated in full, with all proofs, in Appendix B. They cover consistency of the PU estimator, a few-shot bound for rare codes, and convergence of label propagation.

Unbiasedness of the multi-label non-negative PU loss. For code c let $R_c(f) = \pi_c \mathbb{E}_{x\sim P_c}[\ell^+(f(x))] + (1-\pi_c)\mathbb{E}_{x\sim N_c}[\ell^-(f(x))]$ be the true risk, where P_c and N_c are the positive and negative class-conditional distributions.

Proposition 1 (Non-negative PU loss for multi-label ICD coding)**.** *For each code c, the estimator $\hat{R}_c$ of Eq.* (5) *is, before thresholding, an unbiased estimator of $R_c(f)$ plus the weighted soft-positive risk; after thresholding it is the multi-label extension of the non-negative estimator of Kiryo et al. (2017), which is consistent and has lower variance than the unbiased estimator. Applying $\max(0,\cdot)$ per instance and per code, rather than to the aggregate negative-risk estimate of code c, breaks the unbiasedness property.*

Generalization of dynamic graph reranking. Let $\mathcal{H}$ be the class of L-layer GAT rerankers with H heads and per-head width d_h acting on graphs with $|\mathcal{V}_i| \le N$, under bounded parameters and inputs, edge weights in $[0,1]$, and 1-Lipschitz nonlinearities (Assumptions (A1)–(A4), Appendix B.3). Let $R(f)$ be the multi-label risk with average binary cross-entropy over the K candidates.

Theorem 1 (Generalization of dynamic graph reranking)**.** *With probability at least $1-\delta$ over a training set of n i.i.d. instances,*

$$\sup_{f\in\mathcal{H}} \big(R(f) - \hat{R}_n(f)\big) \;\le\; \tilde{O}\left(\sqrt{\frac{L\,H\,d_h^2 + \log(1/\delta)}{n}}\right), \tag{7}$$

where $\tilde{O}(\cdot)$ hides poly-logarithmic factors in N, B_W, B_a, B_x, and n.

The bound does not involve $|\mathcal{C}|$ at all, and it involves the graph size N only through poly-logarithmic factors. The two-stage design therefore replaces a flat classifier over $|\mathcal{C}| \approx 10^4$ labels by a reranker whose complexity is set by its own parameter count LHd_h^2 over a graph of about 70 nodes. This explains why HERA-ICD scales to the ICD-10 vocabulary of MIMIC-IV, where flat classifiers such as PLM-ICD lose most on macro-F1 (§5.2).

Further guarantees (Appendix B). Undercoding is modelled as missing completely at random, so a true positive of code c is recorded with probability $e_c \in (0,1]$ independently of the note (Assumption 1). Under it the estimator of Eq. (5) is consistent, with a finite-sample gap of order $1/\sqrt{e_c n_c}$ (Theorem 2), so the benefit of PU learning grows as undercoding worsens, which is the trend in Figure 4. This is the first finite-sample justification of PU learning for multi-label ICD coding. For a rare code with m examples and Δ-separable embeddings, prototype scoring has excess risk $O(1/\sqrt{m})$ even for $m < 10$ (Theorem 3), which is why code-level retrieval helps in the tail (Figure 3). Label propagation is an ℓ_∞ contraction with factor α, so it converges geometrically with truncation error at most α^T (Theorem 4), which at $\alpha = 0.3$ and $T = 20$ is $\approx 3.5 \times 10^{-11}$.

5 Experiments

We evaluate HERA-ICD on standard ICD coding benchmarks, comparing against baselines spanning transformer, graph, and knowledge-enhanced approaches. We conduct ablations to isolate each component's contribution, analyze performance across code frequency buckets, evaluate robustness to missing labels, and assess interpretability through evidence faithfulness tests.

5.1 Experimental Setup

Datasets. We evaluate the proposed algorithm on the MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) datasets, with statistics in Table 4 (Appendix C). MIMIC-III-Full contains all 8,921 ICD-9 codes and MIMIC-III-50 the 50 most frequent (Mullenbach et al., 2018); for MIMIC-IV we use both the ICD-9 (7,184 codes) and ICD-10 (11,246 codes) subsets.

Evaluation protocol. We follow the patient-level splits of Mullenbach et al. (2018), which prevent leakage across train, validation and test, and adopt the corrected macro-F1 of Edin et al. (2023),

Table 1: Results on MIMIC-III-Full. Best in **bold**, second-best underlined.

Category	Method	Mi-AUC	Ma-AUC	Mi-F1	Ma-F1	P@8	P@15
Attention-based	LAAT	98.8	91.9	57.5	9.9	73.8	59.1
	PLM-ICD	98.9	92.6	59.8	10.4	77.1	61.3
Knowledge-based	MSMN	99.2	95.0	58.4	10.3	75.2	59.9
	FLASH	99.0	–	58.4	–	75.8	60.2
	MRR	**99.5**	94.9	60.3	11.4	77.5	62.3
Hybrid	KEPT	–	–	59.9	11.8	77.1	61.5
	LCDL	98.9	**95.9**	59.0	11.1	75.9	60.1
	TraceCoder	98.8	92.3	60.5	12.5	77.5	62.4
	HERA-ICD	99.3	95.4	**62.8**	**15.8**	**78.8**	**63.7**

which averages only over codes present in the test set, consistent with recent work (Ren et al., 2025). We report micro-F1, macro-F1, micro-AUC, macro-AUC, and precision at K (P@5 for MIMIC-III-50; P@8 and P@15 for full-code settings).

Baselines and knowledge bases. Baselines span attention-based (LAAT, PLM-ICD), knowledge-based (MSMN, FLASH, BRSK, MRR) and hybrid (KEPT-Longformer, LCDL, MSAM, CoRelation, TraceCoder) methods (Vu et al., 2020; Huang et al., 2022; Yuan et al., 2022; Niu et al., 2023b; Zhao et al., 2024; Wang et al., 2024; Yang et al., 2022; 2025; Gomes et al., 2024; Luo et al., 2024; Ren et al., 2025), all using official implementations where available. The retrieval indices draw on three external sources, namely UMLS Metathesaurus concepts and synonyms ($\sim$500K concept–code pairs) (Bodenreider, 2004), medical Wikipedia snippets, and GPT-4–generated summaries of symptoms, causes and treatments for each code.

Implementation. Both encoders are initialized from Clinical-Longformer (Li et al., 2023), and the reranker is a 2-layer GAT with 8 heads over the $K{=}50$ retrieved candidates and $K_s{=}10$ spans, roughly 70 nodes per graph. Label propagation uses $T{=}20$, $\alpha{=}0.3$ and $\theta_{\text{soft}}{=}0.3$. The loss weights $\lambda_{\text{ret}}{=}1.0$, $\lambda_{\text{PU}}{=}1.0$ and $\lambda_{\text{rank}}{=}0.5$ were tuned on validation macro-F1; every other hyperparameter was fixed a priori and never selected on test data. All runs use a single NVIDIA A100, with remaining settings in Appendix C.1 and their sensitivity in Appendix C.4.

5.2 Main Results

Results on the frequent-code benchmark MIMIC-III-50 are in Appendix C (Table 5); with all codes frequent, margins are small and HERA-ICD still leads on every metric.

MIMIC-III-Full. Table 1 presents results on the full code set, as our PU learning and hierarchical retrieval components target rare codes. PLM-ICD achieves 59.8% micro-F1 but only 10.4% macro-F1, reflecting near-zero performance on codes outside the head. TraceCoder improves to 60.5% micro-F1 and 12.5% macro-F1 through knowledge enrichment. HERA-ICD achieves 62.8% micro-F1 and 15.8% macro-F1; the substantial macro-F1 improvement of +3.3 points over TraceCoder confirms that explicit retrieval and PU learning improve rare-code prediction beyond what knowledge integration alone provides.

MIMIC-IV. Table 2 reports results on MIMIC-IV. HERA-ICD maintains its advantage on both ICD-9 and ICD-10 subsets. The macro-F1 gap over baselines is larger on ICD-10 (+4.6% vs. +3.7% on ICD-9), where the vocabulary is 56% larger (11,246 vs. 7,184 codes) and the long-tail problem more severe, confirming that our framework's benefits scale with vocabulary size and label sparsity.

5.3 Retrieval Quality

Because the reranking stage depends entirely on Stage 1, we evaluate retrieval in isolation against the gold code set on MIMIC-III-Full. Candidate recall@50 is 94.1%, so gold codes are almost always retained in the $K{=}50$ candidate set and reranking, not retrieval miss, accounts for the bulk of remaining errors. Precision@K, recall@K and NDCG@K for all three granularities are in Table 8 (Appendix C).

Table 2: Results on MIMIC-IV. ◇Results using Nguyen et al. (2023) splits. Best in **bold**, second-best underlined.

MIMIC-IV-ICD9							
	Method	Mi-A	Ma-A	Mi-F1	Ma-F1	P@8	P@15
Attention-based	LAAT	99.3	96.0	61.7	26.4	68.9	52.7
	PLM-ICD	99.4	97.2	62.6	29.8	70.0	53.5
Knowledge-based	MSMN◇	**99.6**	96.8	61.2	27.8	68.9	–
Hybrid	CoRel.◇	99.5	96.8	62.4	30.0	70.1	–
	TraceCoder	99.5	97.5	63.7	33.7	71.1	54.4
	HERA-ICD	**99.6**	**97.9**	**65.5**	**37.4**	**72.6**	**55.9**
MIMIC-IV-ICD10							
	Method	Mi-A	Ma-A	Mi-F1	Ma-F1	P@8	P@15
Attention-based	LAAT	99.0	95.4	57.9	20.3	68.9	54.3
	PLM-ICD	99.2	96.6	58.5	21.1	69.9	55.0
Knowledge-based	MSMN◇	**99.6**	97.1	55.9	10.8	67.7	–
Hybrid	CoRel.◇	**99.6**	**97.2**	57.8	12.6	70.0	–
	TraceCoder	99.2	96.7	60.0	27.7	71.2	56.3
	HERA-ICD	99.4	97.1	**62.4**	**32.3**	**73.1**	**58.0**

5.4 Rare Code and Zero-Shot Performance

Figure 3 analyzes performance by code frequency on MIMIC-III-Full, where over 4,600 codes (52.3%) appear in fewer than ten training examples. In the Head bucket (≥100 examples), where data is abundant, the margin over TraceCoder is 1.8 points. In the Torso (10–99 examples) HERA-ICD reaches 14.1 against 10.5 for TraceCoder, a gain of 3.6, as hierarchical retrieval supplies complementary evidence at the code and span levels. The difference is starkest in the Tail (<10 examples), where PLM-ICD reaches only 0.7 and TraceCoder 2.0 while HERA-ICD reaches 5.8, through the combined effect of prototype-based few-shot scoring and PU learning correcting the bias from treating undocumented diagnoses as negatives. Removing few-shot and zero-shot scoring alone costs 1.2 macro-F1 (Table 3). This monotonically increasing advantage from Head to Tail is exactly where hierarchical retrieval and PU learning should help, since training signal is scarcest there.

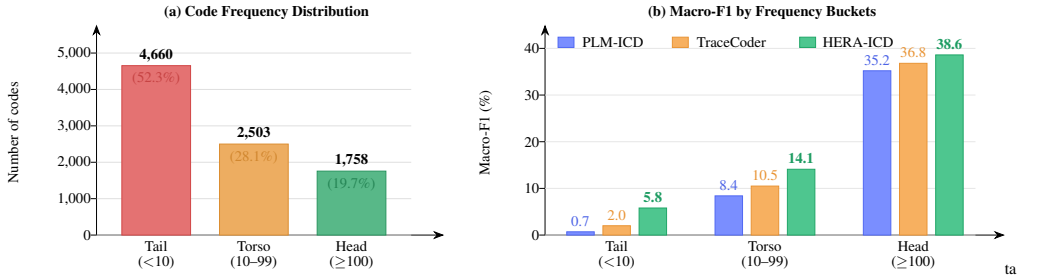

Figure 3: (a) Code frequency distribution on MIMIC-III-Full, where over 52% of codes have <10 training examples (Tail). (b) Macro-F1 by frequency buckets.

5.5 Ablation Study

We remove components from the full model and retrain on MIMIC-III-Full, reporting macro-F1 because it is most sensitive to the rare codes our components target (Table 3). Replacing PU learning with standard binary cross-entropy produces the largest drop, 3.2 points, confirming that missing-label bias is the dominant source of error for rare codes, and graph-based label completion contributes a further 1.5 on top of it, so propagating labels through the ICD hierarchy finds plausible positives that the PU loss alone does not. Falling back to document-level retrieval costs 2.3 points, showing that code-level and span-level evidence are complementary. Replacing the per-instance graph with a static HyperCore-style graph costs 1.6, which supports our core hypothesis that

Table 3: Ablation study on MIMIC-III-Full. Δ is the change in macro-F1 relative to the full model.

Variant	Ma-F1 (%)	Δ
HERA-ICD (full)	15.8	–
w/o PU learning (standard BCE)	12.6	−3.2
w/o graph label completion	14.3	−1.5
w/o hierarchical retrieval	13.5	−2.3
w/o graph reranking	14.1	−1.7
w/o few-shot / zero-shot scoring	14.6	−1.2
w/o cross-encoder reranking	15.1	−0.7
w/o external knowledge	14.4	−1.4
Static graph (HyperCore-style)	14.2	−1.6
Fixed fusion (g=0.5) vs. learned gate	15.2	−0.6

evidence-conditioned topology reasons more effectively than a fixed relational structure. Learned gating adds 0.6 over fixed interpolation, helpful but not essential.

5.6 Robustness to Missing Labels

To quantify robustness to undercoding, we synthetically drop $p\%$ of positive labels from training ($p \in \{10, 20, 30\}$) and compare HERA-ICD against a variant trained with standard BCE. Figure 4 shows macro-F1 as a function of p. HERA-ICD with PU learning degrades more gracefully than the BCE variant as positive labels are synthetically dropped from training data. The gap widens from 3.2% at p=0 to 5.4% at p=30%. The widening gap confirms that PU learning provides increasing benefit as documentation becomes more incomplete, precisely the regime encountered in real clinical data where undercoding rates of 20–40% are common (Jin et al., 2023).

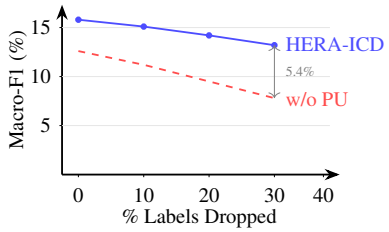

Figure 4: Macro-F1 on MIMIC-III-Full as positive labels are dropped.

5.7 Evidence Faithfulness

We assess whether the spans HERA-ICD attributes to each prediction are *faithful*, meaning actually responsible for the decision, and *aligned with clinician rationale*, using three measures. (i) *Comprehensiveness* is the drop in predicted probability when the attributed spans are removed from the note (higher is better). (ii) *Sufficiency* is the probability retained when only the attributed spans are kept (lower is better). (iii) *Rationale overlap* is the token-level F1 between attributed spans and a held-out set of clinician-annotated rationales. Against PLM-ICD attention attributions, the retrieved spans reach 0.41 comprehensiveness versus 0.18, 0.12 sufficiency versus 0.31, and 52.7% rationale overlap versus 34.2% (Table 9, Appendix C). Higher comprehensiveness and lower sufficiency indicate that HERA-ICD's explanations are more faithful to the prediction than attention heatmaps, and the overlap confirms alignment with expert annotations.

6 Conclusion

We presented HERA-ICD, which reframes ICD coding as retrieval-augmented classification. It retrieves evidence at the document, code and span level, reranks the retrieved candidates on a per-instance evidence graph, and recovers undocumented codes with a non-negative PU loss over graph-propagated soft positives. On MIMIC-III and MIMIC-IV it improves macro-F1 by 3.3 points overall and 3.8 points on rare codes over the strongest prior method, and every component contributes in ablation. The accompanying analysis shows that the PU objective is unbiased and consistent, and that the generalization gap of the reranker is independent of the vocabulary size, which is why the method scales to the ICD-10 label space. Two limitations remain. Per-instance graph construction adds inference overhead, and the consistency guarantee assumes that undercoding is missing completely at random. Our evaluation also covers diagnosis coding only. Natural extensions are a differentiable relaxation of edge pruning that learns the graph topology end to end, expectation-maximization refinement of the class priors π_c under undercoding (Bekker & Davis, 2020), and coding of procedures and radiology reports.

AI USE STATEMENT

The authors used generative AI tools while preparing this work. Their use covered drafting and revising manuscript text, restructuring sections, preparing figures. The research contributions, namely the method, the theoretical results and their proofs, the experimental design and the reported results, are the authors' own. The authors carefully reviewed, revised, and validated all AI-generated or AI-assisted content. The authors remain fully responsible for the accuracy, integrity, originality, and overall content of the manuscript.

ETHICS STATEMENT

This work is a secondary analysis of MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023), de-identified electronic health record datasets released under the PhysioNet credentialed access data use agreement. We collected no new data and had no contact with patients. Both datasets were de-identified by their providers and approved for research use with a waiver of informed consent, our use complies with that agreement, and no protected health information is redistributed. Figure 1 reproduces short, already de-identified excerpts from one discharge summary, with the surrogate identifiers left exactly as released.

Automated ICD coding has clinical and financial consequences. Undercoding distorts reimbursement and epidemiological statistics, and overcoding can constitute billing fraud. We therefore position HERA-ICD as decision support for professional coders rather than an autonomous coding system, and the retrieved evidence spans are intended to make every prediction auditable. The faithfulness evaluation in §5.7 tests whether those spans are actually responsible for the prediction, but it does not certify clinical safety. Deployment would require prospective validation and human review of every assigned code.

REPRODUCIBILITY STATEMENT

The method is specified in full in §3, with pseudocode for both stages in Appendix A and complete proofs of every stated result in Appendix B. Appendix C.1 lists all remaining hyperparameters and optimizer settings and Appendix C.4 reports their sensitivity. We use the standard patient-level splits of Mullenbach et al. (2018) and the corrected macro-F1 of Edin et al. (2023), so our numbers are directly comparable with published results.

Anonymized source code is available at `https://anonymous.4open.science/r/hera-icd-coding-F508`. It covers preprocessing, index construction, training, evaluation, and every ablation variant in Table 3, together with the configuration files that reproduce each experiment and a map from each equation, table and figure in the paper to the file that implements it. MIMIC-III and MIMIC-IV are credentialed datasets and cannot be redistributed, so the repository ships the download and preprocessing scripts needed to rebuild them from PhysioNet.

REFERENCES

Krishanu Das Baksi, Elijah Soba, John J Higgins, et al. MedCodER: A generative AI assistant for medical coding. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Industry Track (NAACL)*, pp. 449–459, 2025. doi: 10.18653/v1/2025.naacl-industry.37.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020. doi: 10.1007/s10994-020-05877-5.

Olivier Bodenreider. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004. doi: 10.1093/nar/gkh061.

Joseph S Boyle, Antanas Kascenas, Nicola Sahar, Maria Liakata, and Alison Q O'Neil. Automated clinical coding using off-the-shelf large language models. In *Deep Generative Models for Health Workshop, NeurIPS 2023*, 2023. arXiv:2310.06552.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3105–3114, 2020. doi: 10.18653/v1/2020.acl-main.282.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 933–942, 2021.

Marthinus du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, pp. 1386–1394, 2015.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2572–2582, 2023. doi: 10.1145/3539618.3591918.

Yidong Gan, David D. Nguyen, Yang Lin, Peter Zhong, Thanh Vu, Long Duong, and Yuan-Fang Li. RAG-Coding: Enhancing LLM medical coding with structured external knowledge. *arXiv preprint arXiv:2605.27377*, 2026.

Xueren Ge, Abhishek Satpathy, Ronald Dean Williams, John Stankovic, and Homa Alemzadeh. DKEC: Domain knowledge enhanced multi-label classification for diagnosis prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12798–12813, 2024. doi: 10.18653/v1/2024.emnlp-main.712.

Gonçalo Gomes, Inês Coutinho, and Bruno Martins. Accurate and well-calibrated ICD code assignment through attention over diverse label embeddings. In *EACL*, pp. 2302–2315, 2024.

Yingbo Guo. Retrieval-augmented generation as noisy in-context learning: A unified theory and risk bounds. *arXiv preprint arXiv:2506.03100*, 2025.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop (ClinicalNLP) at NAACL*, pp. 10–20, 2022. doi: 10.18653/v1/2022.clinicalnlp-1.2.

Zhichao Huang, Hanbing Dong, Aditya Pal, and Beibin Zhang. Surpassing GPT-4 medical coding with a two-stage approach. *arXiv preprint arXiv:2311.13735*, 2023.

Yucheng Jin, Yun Xiong, Dan Shi, Yifei Lin, Lifang He, Yao Zhang, Joseph M Plasek, Li Zhou, David W Bates, and Chunlei Tang. Learning from undercoded clinical records for automated international classification of diseases (ICD) coding. *Journal of the American Medical Informatics Association*, 30(3):438–446, 2023. doi: 10.1093/jamia/ocac230.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. doi: 10.1109/TBDATA.2019.2921572.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1674–1684, 2017.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Tom Lee, Nicolò Cesa-Bianchi, et al. Learning from positive and unlabeled examples – finite size sample bounds. *arXiv preprint arXiv:2507.07354*, 2025.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347, 2023. doi: 10.1093/jamia/ocac225.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Junyu Luo, Xiaochen Wang, Jingyu Wang, Ao Chang, Yaqing Wang, and Fenglong Ma. CoRelation: Boosting automatic ICD coding through contextualized code relation learning. *arXiv preprint arXiv:2406.11132*, 2024.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. ICD-BigBird: A contextual embedding model for ICD code classification. In *Proceedings of the 5th Clinical Natural Language Processing Workshop (BioNLP) at ACL*, pp. 330–336, 2022. doi: 10.18653/v1/2022.bionlp-1.32.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1101–1111, 2018. doi: 10.18653/v1/N18-1100.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, Stefan Winkler, Shao-Syun Huang, Jing-Jing Liu, and Chih-Jen Lin. MIMIC-IV-ICD: A new benchmark for extreme multilabel classification. *arXiv preprint arXiv:2304.13998*, 2023.

Kunying Niu, Yifan Wu, Yaohang Li, and Min Li. Hyperbolic graph convolutional neural network with contrastive learning for automated ICD coding. *Computers in Biology and Medicine*, 2023a. doi: 10.1016/j.compbiomed.2023.107789.

Kunying Niu, Yifan Wu, Yaohang Li, and Min Li. Retrieve and rerank for automated ICD coding via contrastive learning. *Journal of Biomedical Informatics*, 143:104396, 2023b. doi: 10.1016/j.jbi.2023.104396.

Mucheng Ren, He Chen, Yuchen Yan, Danqing Hu, Jun Xu, and Xian Zeng. TraceCoder: Towards traceable ICD coding via multi-source knowledge integration. *arXiv preprint arXiv:2510.15267*, 2025. To appear at IEEE BIBM 2025.

Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3132–3142, 2018. doi: 10.18653/v1/D18-1352.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric Xing. Generalized zero-shot text classification for ICD coding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4018–4024, 2020. doi: 10.24963/ijcai.2020/556.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3335–3341, 2020. doi: 10.24963/ijcai.2020/461.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Xiang Wang, Robert Mercer, and Frank Rudzicz. Multi-stage retrieve and re-rank model for automatic medical coding recommendation. In *NAACL*, pp. 4881–4891, 2024.

Yu Xu, Yang Zhang, and Jianfeng Hu. Multi-positive and unlabeled learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3182–3188, 2017.

Yaowen Yang, Hongfei Lin, Zhihao Yang, Yijia Zhang, Di Zhao, and Ling Luo. LCDL: Classification of ICD codes based on disease label co-occurrence dependency and longformer with medical knowledge. *Artificial Intelligence in Medicine*, 160:103041, 2025.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1767–1781, 2022. arXiv:2210.03304.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. Multi-label few-shot ICD coding as autoregressive generation with prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5366–5374, 2023. doi: 10.1609/aaai.v37i4.25668.

Juyong Yoo and Junmo Kim. How to leverage large language models for automatic ICD coding. *Computers in Biology and Medicine*, 189:109971, 2025. doi: 10.1016/j.compbiomed.2025.109971.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*, pp. 808–814, 2022. doi: 10.18653/v1/2022.acl-short.91.

Xu Zhang, Kun Zhang, Wenxin Ma, Rongsheng Wang, Chenxu Wu, Yingtai Li, and S Kevin Zhou. A general knowledge injection framework for ICD coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7180–7189, 2025. arXiv:2505.18708.

Zheng Zhao, Wei Lu, Xiaozhong Peng, Linlin Xing, Wei Zhang, and Changzheng Zheng. Automated ICD coding via contrastive learning with back-reference and synonym knowledge. *IEEE Transactions on Consumer Electronics*, 2024.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, 2004.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. Automatic ICD coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pp. 5948–5957, 2021. doi: 10.18653/v1/2021.acl-long.463.

A ALGORITHMS

Algorithm 1 gives the pseudocode for the Stage 1 hierarchical retrieval of §3.2, and Algorithm 2 for the Stage 2 graph construction of §3.3.

Algorithm 1 Hierarchical Multi-Granularity Retrieval

Require: Clinical note $\mathbf{x}_i$; FAISS indices $\mathcal{I}_{\text{code}}, \mathcal{I}_{\text{doc}}, \{\mathcal{I}_c\}_{c\in\mathcal{C}}$; hyperparameters K, K_d, K_c, K_s
Ensure: Candidate codes $\mathcal{C}_i^{\text{cand}}$, evidence set $\mathcal{E}_i$

$\mathbf{q}_i \leftarrow f_q(\mathbf{x}_i)$ {Encode clinical note}
$\mathcal{C}_i^{\text{cand}} \leftarrow \text{Retrieve}(\mathbf{q}_i, \mathcal{I}_{\text{code}}, K)$ {Candidate codes}
$\mathcal{D}_i \leftarrow \text{Retrieve}(\mathbf{q}_i, \mathcal{I}_{\text{doc}}, K_d)$ {Similar notes}
for each $c \in \mathcal{C}_i^{\text{cand}}$ **do**
  $\mathcal{E}_i^c \leftarrow \text{Retrieve}(\mathbf{q}_i, \mathcal{I}_c, K_c)$ {Code-level examples}
end for
$\mathcal{S}_i^{\text{bi}} \leftarrow \text{SpanRetrieve}(\mathbf{x}_i, \mathcal{C}_i^{\text{cand}}, 2K_s)$ {Bi-encoder span retrieval}
$\mathcal{S}_i \leftarrow \text{CrossEncoderRerank}(\mathcal{S}_i^{\text{bi}}, K_s)$ {Top-K_s spans}
$\mathcal{E}_i \leftarrow \mathcal{D}_i \cup (\bigcup_c \mathcal{E}_i^c) \cup \mathcal{S}_i$

return $\mathcal{C}_i^{\text{cand}}, \mathcal{E}_i$

Algorithm 2 Evidence-Conditioned Dynamic Graph Construction

Require: Candidates $\mathcal{C}_i^{\text{cand}}$, evidence $\mathcal{E}_i$, hierarchy $\mathcal{H}$, threshold θ_{edge}
Ensure: Graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{A}_i)$

$\mathcal{V}_i \leftarrow \emptyset,\ \mathcal{A}_i \leftarrow \emptyset$
for $c \in \mathcal{C}_i^{\text{cand}}$ **do**
  $\mathcal{V}_i \leftarrow \mathcal{V}_i \cup \{(c,\ \mathbf{W}_c f_c(\text{desc}(c)))\}$ {Code nodes}
end for
for $s \in \mathcal{S}_i$ **do**
  $\mathcal{V}_i \leftarrow \mathcal{V}_i \cup \{(s,\ \mathbf{W}_s f_s(s))\}$ {Span nodes}
end for
for $d \in \mathcal{D}_i$ **do**
  $\mathcal{V}_i \leftarrow \mathcal{V}_i \cup \{(d,\ \mathbf{W}_d f_q(d))\}$ {Context nodes}
end for
for $(c_a, c_b) \in \mathcal{C}_i^{\text{cand}} \times \mathcal{C}_i^{\text{cand}}$ **do**
  Compute w_{ab} via Eq. (1); add to $\mathcal{A}_i$ if $w_{ab} > \theta_{\text{edge}}$
end for
for $s \in \mathcal{S}_i, c \in \mathcal{C}_i^{\text{cand}}$ **do**
  $\mathcal{A}_i \leftarrow \mathcal{A}_i \cup \{(s, c,\ s_{\text{ret}}(s,c) \cdot \text{sim}(\mathbf{h}_s, \mathbf{h}_c))\}$ {Span–code edges}
end for
for $d \in \mathcal{D}_i,\ c \in \mathcal{C}_i^{\text{cand}} : c \in \text{labels}(d)$ **do**
  $\mathcal{A}_i \leftarrow \mathcal{A}_i \cup \{(d, c,\ s_{\text{ret}}(d))\}$ {Context–code edges}
end for
return $\mathcal{G}_i$

B PROOFS

B.1 PRELIMINARIES

To convert uniform convergence of the empirical risk into a generalization bound we use Rademacher complexity, which controls the supremum of the empirical process and yields sharper bounds than VC-dimension for real-valued function classes. Fix a graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ with $|\mathcal{V}| \le N$ and define the parameter vector

$$\theta = \left\{\mathbf{W}^{(l,m)}, \mathbf{a}^{(l,m)}\right\}_{l\in[L],\, m\in[H]} \cup \{\mathbf{w}_o, b_o, \mathbf{w}_g, b_g\}. \tag{8}$$

The total number of scalar parameters is

$$P = \sum_{l=1}^{L} H\big(d^{(l)} d_h + 2d_h\big) + d^{(L)} + 1 + (d_{\text{in}} + d^{(L)}) + 1 = O(LHd_h^2), \tag{9}$$

because $d^{(l)} \le H d_h$ for every layer.

Lemma 1 (Rademacher generalization bound, Mohri et al., 2012, Thm. 3.1)**.** *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to [0,1]$ and ℓ the binary cross-entropy loss. With probability at least $1-\delta$, $\sup_{f\in\mathcal{F}}\{R(f) - \hat{R}_n(f)\} \le 2\,\mathfrak{R}_n(\ell \circ \mathcal{F}) + 3\sqrt{\log(2/\delta)/(2n)}$.*

Lemma 2 (Contraction, Ledoux & Talagrand, 1991, Thm. 4.12)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be L_ϕ-Lipschitz with $\phi(0) = 0$. Then $\mathfrak{R}_n(\phi \circ \mathcal{F}) \le 2L_\phi\,\mathfrak{R}_n(\mathcal{F})$.*

Lemma 3 (Dudley's entropy integral, Wainwright, 2019, Thm. 5.22)**.** *Let $\mathcal{F}$ satisfy $\sup_{f\in\mathcal{F}} \|f\|_\infty \le B$. Then $\mathfrak{R}_n(\mathcal{F}) \le \inf_{\alpha>0}\big\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^B \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)}\,d\epsilon\big\}$, where $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ is the ϵ-covering number of $\mathcal{F}$ in the supremum norm.*

Lemma 4 (Covering number of a Euclidean ball, Wainwright, 2019, Ex. 5.8)**.** *The ϵ-covering number of a ball of radius R in $\mathbb{R}^P$ satisfies $\mathcal{N}(\epsilon, R\mathbb{B}_2^P, \|\cdot\|_2) \le (1 + 2R/\epsilon)^P$.*

B.2 Proof of Proposition 1

Proof. Fix a code c and write $s = f(x)$ for the predicted probability. The true risk is

$$R_c = \pi_c\,\mathbb{E}_{P_c}[\ell^+(s)] + (1-\pi_c)\,\mathbb{E}_{N_c}[\ell^-(s)]. \tag{10}$$

Because the unlabeled distribution is the mixture $U_c = \pi_c P_c + (1-\pi_c)N_c$, the negative loss on unlabeled data satisfies

$$\mathbb{E}_{U_c}[\ell^-(s)] = \pi_c\,\mathbb{E}_{P_c}[\ell^-(s)] + (1-\pi_c)\,\mathbb{E}_{N_c}[\ell^-(s)]. \tag{11}$$

Isolating the negative-class risk, $(1-\pi_c)\mathbb{E}_{N_c}[\ell^-(s)] = \mathbb{E}_{U_c}[\ell^-(s)] - \pi_c\mathbb{E}_{P_c}[\ell^-(s)]$, and substituting into Eq. (10),

$$R_c = \pi_c\,\mathbb{E}_{P_c}[\ell^+(s)] + \mathbb{E}_{U_c}[\ell^-(s)] - \pi_c\,\mathbb{E}_{P_c}[\ell^-(s)]. \tag{12}$$

The plug-in estimator of Eq. (12),

$$\hat{R}_c^{\text{uPU}} = \frac{\pi_c}{n_P}\sum_{i\in\mathcal{P}_c}\ell^+(\hat{y}_{i,c}) + \frac{1}{n_U}\sum_{i\in\mathcal{U}_c'}\ell^-(\hat{y}_{i,c}) - \frac{\pi_c}{n_P}\sum_{i\in\mathcal{P}_c}\ell^-(\hat{y}_{i,c}), \tag{13}$$

is unbiased but can become negative under finite-sample fluctuations because the last two terms estimate $(1-\pi_c)\mathbb{E}_{N_c}[\ell^-] \ge 0$ by a difference. Note that the second sum runs over $\mathcal{U}_c'$ rather than the full unlabeled pool, because label completion has already moved the most plausible missing positives into $\mathcal{S}_c^+$. The residual prior of $\mathcal{U}_c'$ therefore satisfies $\pi_c' \le \pi_c$, so using π_c in the subtracted term removes slightly too much and under-estimates the negative risk by $(\pi_c - \pi_c')\,\mathbb{E}_{P_c}[\ell^-]$. This errs towards predicting positives, which is the safe direction under undercoding, and the gap closes as propagation recovers more of the missing positives. Kiryo et al. (2017) show that replacing this difference by its positive part,

$$\hat{R}_c^{\text{nnPU}} = \pi_c\hat{R}_P^+ + \max\big(0,\ \hat{R}_U^- - \pi_c\hat{R}_P^-\big), \tag{14}$$

where $\hat{R}_P^+$, $\hat{R}_U^-$, $\hat{R}_P^-$ are the empirical averages of ℓ^+ on the positives, ℓ^- on the unlabeled set, and ℓ^- on the positives, yields a consistent estimator with lower variance that remains unbiased in expectation up to a bias term that vanishes as $n_P, n_U \to \infty$. Adding the soft-positive term, a weighted positive risk on the propagated set $\mathcal{S}_c^+$, gives Eq. (5).

Finally, the identity in Eq. (11) holds only for averages over the data distribution of code c. A per-instance quantity $\max(0, \ell^-(\hat{y}_{i,c}) - \pi_c\ell^+(\hat{y}_{i,c}))$ is not an estimate of any term in Eq. (12), because its expectation over i is not $(1-\pi_c)\mathbb{E}_{N_c}[\ell^-]$, because the positive part does not commute with the average and because the subtracted loss is evaluated on unlabeled rather than positive instances. Hence per-instance thresholding breaks unbiasedness, whereas the aggregate correction of Eq. (14) does not. □

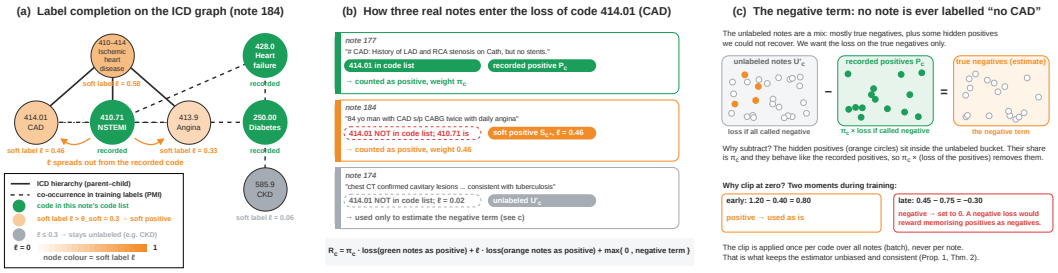

Figure 5: Learning from undercoded labels. (a) Soft labels spread from the recorded code 410.71 along hierarchy and co-occurrence edges of $\mathcal{G}_{\text{code}}$, and codes above θ_{soft} become soft positives. (b) Three real notes enter the loss of code 414.01 as a recorded positive, a soft positive and an unlabeled note. (c) The negative term subtracts the positive share π_c and is clipped at zero once per code, never per note.

B.3 Proof of Theorem 1

Assumptions. A retrieval-augmented classifier first restricts the label space from $|\mathcal{C}|$ to $K \ll |\mathcal{C}|$ candidates and then applies a GAT reranker. It is not obvious whether the empirical gains stem from overfitting to the retrieved subgraphs, or why the method scales to large vocabularies without a proportional increase in sample complexity. Let $\mathcal{H}$ be the hypothesis class of the L-layer GAT reranker with H heads and per-head hidden dimension d_h, and let every evidence-conditioned graph satisfy $|\mathcal{V}_i| \le N$. (A1) bounded parameters, $\|\mathbf{W}^{(l,m)}\|_{\text{op}} \le B_W$, $\|\mathbf{a}^{(l,m)}\|_2 \le B_a$, $\|\mathbf{w}_o\|_2 \le B_o$, $\|\mathbf{w}_g\|_2 \le B_g$, $|b| \le B_b$; (A2) bounded inputs, $\|\mathbf{h}_v^{(0)}\|_2 \le B_x$; (A3) edge weights in $[0,1]$ and neighbourhoods of size at most N; (A4) all nonlinearities (LeakyReLU, σ, softmax) are 1-Lipschitz with $\phi(0) = 0$. Consider the multi-label risk $R(f) = \mathbb{E}_{(x,y)}\big[\frac{1}{K}\sum_{c=1}^{K} \ell(y_c, f_c(x))\big]$ with binary cross-entropy ℓ.

We first establish three supporting propositions on the Lipschitzness of the GAT in its parameters, the covering number of the hypothesis class, and the resulting Rademacher complexity. The attention mechanism complicates the usual argument because the aggregation weights are themselves parameter-dependent.

Proposition 2 (Layer-wise Lipschitzness)**.** *Under (A1)–(A4), for any fixed graph $\mathcal{G}$ with $|\mathcal{V}| \le N$, the output of the GAT reranker satisfies $|f_\theta(\mathcal{G}) - f_{\theta'}(\mathcal{G})| \le \text{Lip}\,\|\theta - \theta'\|_2$ with $\log \text{Lip} = O\big(L \log(NHB_W B_a B_x)\big)$.*

Proof. We proceed by induction on the layer index l.

Base case ($l = 1$). For head m at layer 1 the attention logits are $e_{vu}^{(1,m)} = \text{LeakyReLU}\big((\mathbf{a}^{(1,m)})^\top[\mathbf{W}^{(1,m)}\mathbf{h}_v^{(0)} \| \mathbf{W}^{(1,m)}\mathbf{h}_u^{(0)}]\big)$. By (A4) LeakyReLU is 1-Lipschitz. The map $(\mathbf{W}, \mathbf{a}) \mapsto e_{vu}$ has partial derivatives bounded by $O(B_W B_x)$ with respect to $\mathbf{a}$ and $O(B_a B_x)$ with respect to $\mathbf{W}$, so the logits are $O(B_W B_a B_x)$-Lipschitz in the layer-1 parameters. The softmax over at most N neighbours is 1-Lipschitz in the ℓ_2 norm of the logit vector. The aggregation $\tilde{\mathbf{h}}_v = \sum_{u \in \mathcal{N}(v)} \alpha_{vu} \mathbf{W}^{(1,m)} \mathbf{h}_u^{(0)}$ is a convex combination, hence 1-Lipschitz in the vectors $\{\mathbf{W}\mathbf{h}_u\}$, so changing $\mathbf{W}$ by $\Delta\mathbf{W}$ changes the aggregate by at most $NB_x\|\Delta\mathbf{W}\|_{\text{op}}$. Combined with the Lipschitzness of the attention coefficients, the head output is $O(NB_W B_a B_x)$-Lipschitz in $(\mathbf{W}^{(1,m)}, \mathbf{a}^{(1,m)})$. After concatenating H heads and applying σ, the layer-1 features $\mathbf{H}^{(1)} \in \mathbb{R}^{N \times Hd_h}$ are $O(NHB_W B_a B_x)$-Lipschitz in the layer-1 parameters.

Inductive step. Assume $\mathbf{H}^{(l)}$ is $L^{(l)}$-Lipschitz in the parameters up to layer l and $\|\mathbf{h}_v^{(l)}\|_2 \le B^{(l)}$ for all v. Perturbing $(\Delta\mathbf{W}, \Delta\mathbf{a}, \Delta\mathbf{H}^{(l)})$ at layer $l+1$ and using the triangle inequality with the 1-Lipschitzness of LeakyReLU,

$$|e_{vu} - e'_{vu}| \le B_a \cdot O\big(B^{(l)}\|\Delta\mathbf{W}\|_{\text{op}} + B_W\|\Delta\mathbf{h}_v\|_2 + B_W\|\Delta\mathbf{h}_u\|_2\big) + O(B_W B^{(l)})\|\Delta\mathbf{a}\|_2. \quad (15)$$

By the inductive hypothesis $\|\Delta\mathbf{h}_v\|_2 \le L^{(l)}\|\Delta\theta_{\le l}\|_2$, so the logits are $O(B_a B^{(l)} + B_W L^{(l)})$-Lipschitz in the full parameter vector up to layer $l+1$. The softmax over at most N neighbours is 1-Lipschitz. In the aggregated message $\tilde{\mathbf{h}}_v^{(l+1,m)} = \sum_u \alpha_{vu}^{(l+1,m)} \mathbf{W}^{(l+1,m)} \mathbf{h}_u^{(l)}$, changing $\mathbf{W}^{(l+1,m)}$

by $\Delta\mathbf{W}$ changes the output by at most $NB^{(l)}\|\Delta\mathbf{W}\|_{\mathrm{op}}$, and changing the attention coefficients changes the convex combination by at most $NB_W B^{(l)}$ times the attention perturbation. Therefore

$$\|\tilde{\mathbf{h}}_v^{(l+1,m)} - \tilde{\mathbf{h}}_v'^{(l+1,m)}\|_2 \le O\big(NB_W B_a (B^{(l)})^2 + NB_W^2 B^{(l)} L^{(l)}\big)\, \|\Delta\theta_{\le l+1}\|_2, \tag{16}$$

and after concatenating H heads and applying σ,

$$L^{(l+1)} = O\big(HNB_W B_a (B^{(l)})^2 + HNB_W^2 B^{(l)} L^{(l)}\big). \tag{17}$$

Since $B^{(l)} \le (HNB_W)^l B_x$ (each layer multiplies features by at most HB_W and averages over at most N neighbours), induction on Eq. (17) gives $\log L^{(l)} = O\big(l \log(NHB_W B_a B_x)\big)$.

Readout. The readout $s_{\mathrm{graph}}(c) = \mathbf{w}_o^\top \mathbf{h}_c^{(L)} + b_o$ is affine, hence $O(\|\mathbf{w}_o\|_2 L^{(L)})$-Lipschitz. The gate $g_c = \sigma(\mathbf{w}_g^\top[\mathbf{q}_i \| \mathbf{h}_c^{(L)}] + b_g)$ composes a bounded linear map with the 1-Lipschitz sigmoid and is $O(\|\mathbf{w}_g\|_2 L^{(L)})$-Lipschitz. The fused score $\hat{y}_c = \sigma(g_c s_{\mathrm{ret}}(c) + (1-g_c) s_{\mathrm{graph}}(c))$ composes bounded bilinear operations with the sigmoid; since $s_{\mathrm{ret}}(c)$ is precomputed by the retrieval encoder and not learned by the reranker, the dependence on θ is only through g_c and $s_{\mathrm{graph}}(c)$, both $L^{(L)}$-Lipschitz up to polynomial factors in B_W, B_a, B_o, B_g. Hence $\log \mathrm{Lip} = O(L \log(NHB_W B_a B_x))$. □

The Lipschitz constant grows polynomially in N and exponentially in L; because only its logarithm enters the covering number, increasing the number of retrieved candidates does not catastrophically degrade the bound.

Proposition 3 (Covering number)**.** *Let $\mathcal{H}$ be the GAT class with parameters in a Euclidean ball of radius $B_\theta = \max\{B_W, B_a, B_o, B_g, B_b\}$ in $\mathbb{R}^P$, $P = O(LHd_h^2)$. Then for any $\epsilon > 0$, $\log \mathcal{N}(\epsilon, \mathcal{H}, \|\cdot\|_\infty) \le O\big(P \log(\mathrm{Lip}\, B_\theta/\epsilon)\big)$.*

Proof. By Proposition 2 every $f \in \mathcal{H}$ is Lip-Lipschitz in θ, so an ϵ/Lip-cover of the parameter ball induces an ϵ-cover of $\mathcal{H}$ in the supremum norm. By Lemma 4, $\mathcal{N}(\epsilon/\mathrm{Lip}, B_\theta \mathbb{B}_2^P, \|\cdot\|_2) \le (1 + 2\mathrm{Lip}\, B_\theta/\epsilon)^P$; taking logarithms yields the claim. □

Proposition 4 (Rademacher complexity of the GAT class)**.** *Under (A1)–(A4), $\mathfrak{R}_n(\mathcal{H}) \le \tilde{O}\big(\sqrt{LHd_h^2/n}\big) = \tilde{O}\big(\sqrt{P/n}\big)$, where $\tilde{O}$ hides poly-logarithmic factors in N, B_W, B_a, B_x, and n.*

Proof. The pre-sigmoid output is bounded because each layer performs weighted averages with coefficients in $[0,1]$ and multiplications by bounded operators, so by induction $\|\mathbf{h}_v^{(l)}\|_2 \le (HB_W)^l B_x$, so the readout is bounded by $B_{\mathrm{out}} := B_o (HB_W)^L B_x + B_b$. We apply Lemma 3 to the pre-sigmoid class (the sigmoid is a contraction) with $\alpha = 1/\sqrt{n}$ and Proposition 3:

$$\begin{aligned}\mathfrak{R}_n(\mathcal{H}) &\le \frac{4}{\sqrt{n}} + \frac{12}{\sqrt{n}} \int_{1/\sqrt{n}}^{B_{\mathrm{out}}} \sqrt{O\big(P \log(\mathrm{Lip}\, B_\theta/\epsilon)\big)}\, d\epsilon \\ &\le \frac{4}{\sqrt{n}} + \frac{12\sqrt{P}}{\sqrt{n}} \int_{1/\sqrt{n}}^{B_{\mathrm{out}}} \sqrt{\log(C/\epsilon)}\, d\epsilon, \end{aligned} \tag{18}$$

with $C = \mathrm{Lip}\, B_\theta$. The integrand is dominated by a poly-logarithmic factor, $\int_{1/\sqrt{n}}^{B_{\mathrm{out}}} \sqrt{\log(C/\epsilon)}\, d\epsilon \le B_{\mathrm{out}} \sqrt{\log(C\sqrt{n})} = \tilde{O}(1)$, since B_{out} is polynomially bounded in the problem parameters. Hence $\mathfrak{R}_n(\mathcal{H}) \le \tilde{O}(\sqrt{P/n}) = \tilde{O}(\sqrt{LHd_h^2/n})$. □

Proof of Theorem 1. Because the loss is an average over K codes and the GAT shares parameters across all code nodes, we apply the contraction lemma coordinate-wise. For each code c the map $s \mapsto \ell(y_c, \sigma(s))$ is 1-Lipschitz in the logit s, since $\partial_s \ell(y, \sigma(s)) = \sigma(s) - y$ has absolute value at most 1. Averaging over K coordinates introduces no $\sqrt{K}$ factor, so by Lemma 2, $\mathfrak{R}_n(\ell \circ \mathcal{H}) \le 2\,\mathfrak{R}_n(\mathcal{H})$. The Lipschitz constant Lip enters Proposition 3 only logarithmically, and Dudley's integral adds at

most a $\log N$ factor, so Proposition 4 gives $\mathfrak{R}_n(\mathcal{H}) \le \tilde{O}(\sqrt{LHd_h^2/n})$ with hidden factors in $\log N$. By Lemma 1, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{H}}\{R(f)-\hat{R}_n(f)\} \le 4\,\mathfrak{R}_n(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \le \tilde{O}\left(\sqrt{\frac{LHd_h^2}{n}}\right) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right) = \tilde{O}\left(\sqrt{\frac{LHd_h^2+\log(1/\delta)}{n}}\right), \tag{19}$$

which is Eq. (7). □

The bound involves neither $|\mathcal{C}|$ nor the graph size outside poly-logarithmic factors. Reducing the effective hypothesis class from a flat classifier over $|\mathcal{C}| \approx 10^4$ labels to a reranker over a graph of about 70 nodes decouples the complexity term from the vocabulary entirely, which is the fundamental statistical advantage of the retrieve-then-rerank design.

B.4 Consistency of the PU estimator

Without a consistency guarantee, a PU loss could remove the undercoding bias of the training set while introducing a new bias at test time. Finite-sample guarantees require a model of how labels go missing.

Assumption 1 (Missing completely at random)**.** *For each code c, the probability that a true positive label is observed does not depend on the instance, so that* $\Pr(c \in \mathcal{P}_i \mid y_{i,c}=1, \mathbf{x}_i) = e_c \in (0,1]$.

Under Assumption 1 the observed positives form a uniformly random subset of the true positives, the minimal condition under which unbiased PU risk estimation is possible (Bekker & Davis, 2020); extensions to missing-at-random undercoding are left to future work.

Theorem 2 (Consistency of multi-label non-negative PU)**.** *Under Assumption 1, let $\hat{R}_c^{\mathrm{nnPU}}$ be the estimator of Eq.* (5)*. Then $\mathbb{E}[\hat{R}_c^{\mathrm{nnPU}}] \to R_c$ as $n_P, n_U \to \infty$. Moreover, if the hypothesis class $\mathcal{F}$ has finite VC-dimension V, then with probability at least $1-\delta$,*

$$\left|R_c(\hat{f}) - \hat{R}_c^{\mathrm{nnPU}}(\hat{f})\right| \;\le\; O\left(\frac{1}{\sqrt{e_c\, n_c}} + \frac{1}{\sqrt{n_U}} + \sqrt{\frac{V\log n + \log(1/\delta)}{n}}\right), \tag{20}$$

where n_c is the number of true positives of c in the training set, n_U the number of unlabeled instances, and $n = n_c + n_U$.

Proof. Under Assumption 1 the observed positives $\mathcal{P}_c$ are a uniformly random subset of the true positives, so the empirical averages over $\mathcal{P}_c$ in Eq. (13) are unbiased for $\mathbb{E}_{P_c}[\cdot]$, and the remaining pool $\mathcal{U}_c'$ is a mixture with residual prior $\pi_c' \le \pi_c$, so the estimator is unbiased up to the conservative gap identified in the proof of Proposition 1; unbiasedness of $\hat{R}_c^{\mathrm{uPU}}$ then follows from Eq. (12), and the non-negative correction introduces a bias that vanishes as $n_P, n_U \to \infty$ (Kiryo et al., 2017), giving $\mathbb{E}[\hat{R}_c^{\mathrm{nnPU}}] \to R_c$. For the rate, the number of observed positives is $n_P \approx e_c n_c$, so Hoeffding's inequality applied to the positive averages contributes $O(1/\sqrt{e_c n_c})$ and applied to the unlabeled average contributes $O(1/\sqrt{n_U})$; a union bound over $\mathcal{F}$ via VC theory contributes $O(\sqrt{(V\log n + \log(1/\delta))/n})$. Summing the three terms yields Eq. (20). The factor $1/\sqrt{e_c n_c}$ shows that codes with small e_c (severe undercoding) suffer higher variance, quantifying the statistical cost of clinical undercoding. □

B.5 Few-shot generalization for rare codes

Rare codes often have fewer than ten training examples. Prototype-based scoring (§3.2) compensates, and the following bound shows that $m < 10$ examples suffice for non-trivial generalization.

Theorem 3 (Few-shot generalization for rare codes)**.** *Let c be a rare code with m training examples whose embeddings have mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{\Sigma}$, let $\mathbf{p}_c$ be the prototype (mean of the m embeddings), and assume the bi-encoder produces Δ-separable embeddings, $\|\mathbf{q}-\boldsymbol{\mu}_c\|_2 \le \|\mathbf{q}-\boldsymbol{\mu}_{c'}\|_2 - \Delta$*

for all $c' \neq c$ and all relevant queries $\mathbf{q}$*. Then the prototype classifier achieves excess risk*

$$\mathbb{E}[\mathcal{E}_{\text{excess}}] \;\leq\; \frac{4\,\mathrm{Tr}(\mathbf{\Sigma})}{\Delta^2\, m} + O\!\left(\frac{1}{\sqrt{m}}\right). \tag{21}$$

Proof. The prototype classifier is a nearest-centroid rule (Snell et al., 2017), so a query $\mathbf{q}$ satisfying the Δ-separability condition is misclassified only if the empirical prototype deviates from the population mean by more than $\Delta/2$. For m i.i.d. samples with covariance $\mathbf{\Sigma}$, $\mathbb{E}\|\mathbf{p}_c - \boldsymbol{\mu}_c\|_2^2 = \mathrm{Tr}(\mathbf{\Sigma})/m$, and Chebyshev's inequality gives $\Pr(\|\mathbf{p}_c - \boldsymbol{\mu}_c\|_2 > \Delta/2) \leq 4\,\mathrm{Tr}(\mathbf{\Sigma})/(\Delta^2 m)$. The $O(1/\sqrt{m})$ term accounts for the excess risk of the near-boundary queries whose margin is smaller than Δ, which by a standard margin argument decays at the parametric rate. □

B.6 Convergence of label propagation

Treating the output of a fixed-T propagation as a training signal is sound only if the iteration converges to a unique fixed point with a truncation error bounded independently of the instance. Since $\mathcal{G}_{\text{code}}$ mixes hierarchical and co-occurrence edges with widely varying degrees, it is neither regular nor symmetric, so convergence must be established for the row-stochastic matrix $\mathbf{W}$.

Theorem 4 (Label propagation convergence)**.** *Let $\mathbf{W}$ be the row-stochastic weight matrix of $\mathcal{G}_{\text{code}}$ and $\alpha \in (0,1)$. The iteration $\boldsymbol{\ell}^{(t+1)} = \alpha\mathbf{W}\boldsymbol{\ell}^{(t)} + (1-\alpha)\boldsymbol{\ell}^{(0)}$ converges to the unique fixed point $\boldsymbol{\ell}^* = (1-\alpha)(\mathbf{I} - \alpha\mathbf{W})^{-1}\boldsymbol{\ell}^{(0)}$, and $\|\boldsymbol{\ell}^{(T)} - \boldsymbol{\ell}^*\|_\infty \leq \alpha^T$.*

Proof. Since $\mathbf{W}$ is row-stochastic, $\|\mathbf{W}\|_\infty = 1$, so the map $\boldsymbol{\ell} \mapsto \alpha\mathbf{W}\boldsymbol{\ell} + (1-\alpha)\boldsymbol{\ell}^{(0)}$ is a contraction with factor $\alpha < 1$ in the ℓ_∞ norm. By the Banach fixed-point theorem the limit exists and is unique. The closed form follows from the geometric series $\sum_{t\geq 0}(\alpha\mathbf{W})^t = (\mathbf{I} - \alpha\mathbf{W})^{-1}$, which converges because $\|\alpha\mathbf{W}\|_\infty = \alpha < 1$. For the truncation error, $\boldsymbol{\ell}^{(T)} - \boldsymbol{\ell}^* = (\alpha\mathbf{W})^T(\boldsymbol{\ell}^{(0)} - \boldsymbol{\ell}^*)$, and since $\|\mathbf{W}^T\|_\infty = 1$ and both vectors lie in $[0,1]^{|\mathcal{C}|}$, $\|\boldsymbol{\ell}^{(T)} - \boldsymbol{\ell}^*\|_\infty \leq \alpha^T$. □

C Additional Experimental Details and Results

C.1 Implementation Details

This section completes the summary in §5.1. The query encoder f_q and the code encoder f_c are both initialized from Clinical-Longformer (Li et al., 2023) with a hidden size of 768. The GAT reranker uses L=2 layers, 8 attention heads, hidden dimension 512 and dropout 0.2. Besides the K=50 candidate codes and K_s=10 evidence spans reported in the main text, retrieval returns K_d=10 similar notes and K_c=5 code-level examples per candidate. The graph edge-pruning threshold is θ_{edge}=0.3, which yields an average node degree of 6.4, and the contrastive temperature τ is learned rather than tuned. Class priors π_c are estimated per code from training frequencies. Training runs for up to 20 epochs with early stopping at patience 3, using AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-5}, weight decay 0.01, linear warmup over 10% of the steps and gradient clipping at norm 1.0.

C.2 Dataset Statistics

The proposed algorithm is evaluated on MIMIC-III Johnson et al. (2016) and MIMIC-IV Johnson et al. (2023) datasets; key statistics of these datasets are summarized in Table 4.

Table 4: Dataset statistics. Rare codes appear in <10 training examples.

	Notes	Codes	% Rare	Avg. len.
MIMIC-III-Full	47,723	8,921	52.3%	1,456
MIMIC-III-50	11,368	50	0%	1,456
MIMIC-IV (ICD-9)	45,232	7,184	49.8%	1,621
MIMIC-IV (ICD-10)	61,438	11,246	58.1%	1,583

Table 5: Results on MIMIC-III-50. Best in **bold**, second-best underlined.

Category	Method	Mi-AUC	Ma-AUC	Mi-F1	Ma-F1	P@5
Attention-based	LAAT	94.6	92.5	71.5	66.6	67.5
	PLM-ICD	93.8	91.7	70.5	66.3	65.7
Knowledge-based	MSMN	94.4	92.5	72.5	68.3	68.0
	BRSK	94.9	92.9	72.8	69.1	67.7
	MRR	94.7	92.7	73.2	68.7	68.5
Hybrid	KEPT	94.8	92.6	72.9	68.9	67.3
	LCDL	95.1	93.1	73.0	69.0	68.6
	MSAM	95.0	93.2	73.2	69.0	68.2
	TraceCoder	95.1	93.2	73.4	69.1	68.6
	HERA-ICD	**95.4**	**93.7**	**74.2**	**69.5**	**69.3**

Table 6: Hyperparameter sensitivity on MIMIC-III-Full (macro-F1, %). Default setting in the central column.

Propagation α	0.1	0.2	**0.3**	0.4	0.5
Ma-F1	15.1	15.5	15.8	15.6	15.2
Soft thr. θ_{soft}	0.1	0.2	**0.3**	0.4	0.5
Ma-F1	14.9	15.4	15.8	15.5	15.0
Edge thr. θ_{edge}	0.1	0.2	**0.3**	0.4	0.5
Ma-F1	15.2	15.6	15.8	15.5	14.9

C.3 Results on MIMIC-III-50

Table 5 presents results on the 50 most frequent codes. Since all codes are frequent, the long-tail challenge is absent and differences between methods are smaller. TraceCoder leads existing methods with 73.4% micro-F1 and 69.1% macro-F1 through multi-source knowledge integration. HERA-ICD achieves 74.2% micro-F1 and 69.5% macro-F1; gains come primarily from evidence-conditioned graph reranking rather than PU learning, which has minimal effect when all codes are well-represented.

C.4 Hyperparameter Sensitivity

The loss weights λ were selected on validation macro-F1; all remaining hyperparameters were fixed a priori and never tuned on the test set. Table 6 sweeps the three most influential ones around their defaults on MIMIC-III-Full. Performance is stable across a wide range, indicating the model is not sensitive to their exact setting.

Varying θ_{edge} also controls graph density, and at the default value the per-instance graphs have an average node degree of 6.4, and accuracy is flat across the [0.2, 0.4] interval before degrading as the graph becomes either too sparse or too dense.

C.5 Computational Cost

Table 7 compares training and inference costs on MIMIC-III-Full. HERA-ICD's overhead comes from FAISS lookups (<1ms per query), per-instance graph construction, and the 2-layer GAT forward pass. Training is $\sim$2.6$\times$ slower than PLM-ICD but only 30% slower than TraceCoder, while inference adds $\sim$30% over TraceCoder due to parallelized retrieval and small graph sizes ($|\mathcal{V}_i| \approx 70$ nodes for K=50 candidates and K_s=10 spans). We consider this a favorable trade-off given the substantial macro-F1 gains and interpretability benefits.

Scaling and deployment. The candidate, document, and per-code FAISS indices are built once and amortized across all notes; embeddings are cached, so per-note cost is dominated by sub-millisecond lookups and a forward pass over a $\sim$70-node graph. Throughput therefore scales linearly and parallelizes trivially across GPUs, making batch-coding of a multi-million-record archive feasible as an offline job. We are nonetheless explicit about the trade-off, since the $\sim$30% inference overhead over TraceCoder and the per-instance graph construction make HERA-ICD better suited to

Table 7: Computational cost on MIMIC-III-Full (single A100 GPU, 40GB).

Method	Train (hr/epoch)	Infer (s/note)	Params (M)
PLM-ICD	∼0.5	0.067	125
TraceCoder	∼1.0	0.089	∼140
HERA-ICD	∼1.3	0.115	∼152

batch and back-office coding than to hard-real-time, point-of-care use; the slower training (Table 7) and memory for cached indices are further practical costs. Lowering per-instance graph-construction latency for low-latency clinical settings (e.g. via candidate caching or graph sparsification) is left to future work.

C.6 Retrieval Quality and Evidence Faithfulness

Table 8 reports retrieval quality at each granularity, summarized in §5.3. Table 9 reports the faithfulness measures of §5.7.

Table 8: Retrieval quality on MIMIC-III-Full (K=50).

Level	P@K	R@K	NDCG@K
Document (K_d=10)	41.2	68.5	55.3
Code (K=50)	27.4	94.1	72.6
Span (K_s=10)	63.8	71.2	68.4

Table 9: Evidence faithfulness on MIMIC-III-Full. Comp. (↑), Suff. (↓), Overlap = rationale token-F1 (↑).

Attribution	Comp. ↑	Suff. ↓	Overlap ↑
PLM-ICD (attention)	0.18	0.31	34.2
HERA-ICD (retrieved spans)	0.41	0.12	52.7

D Extended Related Work

This appendix expands §2. We organize prior work along two axes. Section D.1 covers encoding and classification approaches, and §D.2 covers methods that address the long-tail and missing-label challenges.

D.1 ICD Coding Models from Attention to Retrieval

Attention and transformer approaches. The modern line of automated ICD coding begins with CAML (Mullenbach et al., 2018), which introduced per-label attention over CNN-encoded text and established the MIMIC-III preprocessing pipeline used by virtually all subsequent work. LAAT (Vu et al., 2020) proposed a BiLSTM with label-specific attention, and PLM-ICD (Huang et al., 2022) demonstrated that a pretrained RoBERTa with segment pooling achieves the strongest results under standardized evaluation (Edin et al., 2023). Clinical-Longformer (Li et al., 2023) extended the context window to 4,096 tokens. More recently, large language models have been explored for ICD coding in zero-shot settings (Boyle et al., 2023; Huang et al., 2023; Yoo & Kim, 2025), improving rare-code coverage but without explicit retrieval mechanisms or missing-label handling.

Graph-based methods. Graph neural networks have been applied to model the ICD hierarchy and code co-occurrence. HyperCore (Cao et al., 2020) embeds codes in hyperbolic space with GCN layers, ICDBigBird (Michalopoulos et al., 2022) fuses BigBird with a GCN over co-occurrence edges, and HGCN-CL (Niu et al., 2023a) extends HyperCore with contrastive learning. A common limitation is that these methods construct static graphs applied identically to every document. HERA-ICD instead constructs per-instance evidence-conditioned graphs whose topology adapts to the retrieved evidence.

Knowledge-enhanced and retrieval approaches. Several methods integrate external knowledge. MSMN (Yuan et al., 2022) uses UMLS synonyms, KEPT (Yang et al., 2022) injects UMLS via

contrastive pretraining, DKEC (Ge et al., 2024) constructs heterogeneous knowledge graphs, and TraceCoder (Ren et al., 2025) integrates UMLS, Wikipedia, and LLM-generated knowledge through hybrid attention. GKI-ICD (Zhang et al., 2025) injects knowledge via retrieval-synthesized guidelines during training only. The closest retrieval-oriented work is that of Niu et al. (2023b), who proposed retrieve-and-rerank via contrastive learning, and MedCodER (Baksi et al., 2025), which combines LLM extraction with bi-encoder retrieval. RAG-Coding (Gan et al., 2026) is concurrent work that uses agentic retrieval-augmented LLM generation with structured external knowledge for ICD-10-CM; it targets LLM prompting rather than a trainable reranker, and does not model missing labels. None of these employs hierarchical multi-granularity retrieval, evidence-conditioned graph reranking, or PU learning for missing labels.

D.2 Long-Tail and Missing-Label Learning

Few-shot and zero-shot ICD coding. ZAGCNN (Rios & Kavuluru, 2018) provided the first evaluation of zero-shot ICD codes using GCN-propagated code descriptions. AGMC (Song et al., 2020) generates pseudo-features for zero-shot codes via adversarial models. KEPT (Yang et al., 2022) achieves 14.5% macro-F1 on rare codes through knowledge-enhanced prompting, and GP-soap (Yang et al., 2023) reframes ICD coding as autoregressive generation of code descriptions. Prototype-based methods (Snell et al., 2017) offer a natural framework but have not been combined with explicit evidence retrieval for ICD coding. Our hierarchical retrieval enables few-shot prediction through code-level prototypes and zero-shot prediction through description-to-label matching.

Positive-unlabeled learning. PU learning (du Plessis et al., 2015; Kiryo et al., 2017; Bekker & Davis, 2020) provides a principled framework when only positive and unlabeled examples are available; Lee et al. (2025) give tight finite-sample bounds for the single-label case, and Xu et al. (2017) extend the framework to multiple positive classes with generalization error of order $O(k^2/\sqrt{n})$. Within ICD coding, Jin et al. (2023) is the only prior work, proposing a reweighted PU loss for undercoded records without retrieval or graph-based label completion. In multi-label learning, the single-positive-label setting of Cole et al. (2021) is the special case of ours in which exactly one code per note is observed; our formulation allows any number of observed positives per note and adds propagated soft positives.

Theory of retrieval-augmented prediction. Guo (2025) give the first finite-sample risk bounds for retrieval-augmented generation, viewing retrieved context as noisy in-context examples. Our Theorem 1 is complementary, because it bounds the generalization gap of a retrieval-then-rerank *classifier* in terms of the size of the retrieved candidate graph, independently of the label vocabulary.