OntoBook: Ontology-Grounded Synthetic Textbooks for Medical Encoder Pretraining

Rian Touchent, Éric de La Clergerie
Inria, Sorbonne Université
48 rue Barrault 75013 Paris, 21 rue de l'école de médecine 75006 Paris
{rian.touchent,eric.de_la_clergerie}@inria.fr

Abstract

We present OntoBook, a method that converts medical ontology structure into pretraining signal for encoder language models. Our approach has three stages: random walks through ontology graphs capture hierarchical and causal relations between medical codes, a large language model reformulates these walks into fluent textbook-style prose, and the resulting text is used to train ModernCamemBERT, a 149M-parameter French encoder, with two objectives on the same data: masked language modeling and relation prediction between code pairs. On three French medical coding benchmarks (FRACCO, Cantemist-FR, Distemist-FR), OntoBook achieves significant improvements over MLM-only pretraining, with +2.5 micro-F1 on FRACCO and +8.0 micro-F1 on Distemist. We find that alignment between objectives is necessary: misaligned training, where each task uses different data, causes a 30-point degradation. We release 1.3 million LLM-reformulated medical textbooks across three French ontologies (CIM-10, CCAM, ATC) and pretrained model checkpoints.

Keywords: knowledge graphs, ontology, medical coding, language model pretraining, multi-task learning

1. Introduction

Medical coding is the task of assigning standardized codes from medical ontologies to clinical documents. It is central to hospital billing, epidemiological surveillance, and clinical research. In France, over 30 million hospital stays per year require accurate coding using the CIM-10 classification (the French adaptation of ICD-10), CCAM procedure codes, and ATC medication codes. This task requires understanding not only medical vocabulary but also the relational structure of medical ontologies: hierarchical links between codes, causal associations, differential diagnoses, and exclusion rules. These relationships are explicitly encoded in ontology graphs but are absent from clinical text corpora.

Pretrained biomedical language models such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2022), and CamemBERT-bio (Touchent et al., 2023) learn medical vocabulary from large text corpora. However, they do not capture the relational structure of medical ontologies, since this structure is not expressed in running text. Knowledge graph approaches such as RDF2Vec (Ristoski and Paulheim, 2016) and Snomed2Vec (Agarwal et al., 2019) encode ontology structure through random walks, but produce static, non-contextualized embeddings that cannot benefit from transformer pretraining. Knowledge-enhanced transformers such as DRAGON (Yasunaga et al., 2022) and KEPLER (Wang et al., 2021) integrate knowledge graph structure into language model pretraining, but they rely on existing text-graph alignment rather than generating new training data from ontology structure alone.

We propose OntoBook, a pretraining method that converts medical ontology structure into training signal for encoder language models. Our approach proceeds in three stages. First, we generate random walks through ontology graphs to produce sequences that capture hierarchical, causal, and dif-

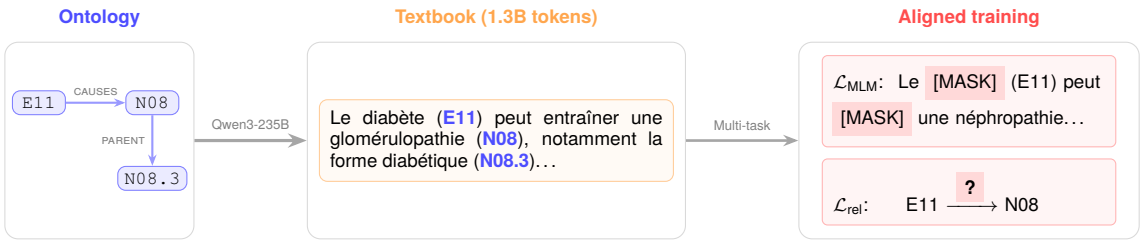

Figure 1: OntoBook pipeline. An ontology subgraph is converted into a random walk, then reformulated into textbook prose by Qwen3-235B. The encoder trains on this text with two aligned objectives: masked language modeling and relation prediction between code pairs.

Proceedings of the Knowledge Graphs and Large Language Models Workshop (KG-LLM) @ LREC26, pages 11–19
May 16, 2026.

ferential relationships between medical codes. Second, we reformulate these walks into fluent medical prose using a large language model, creating synthetic textbooks grounded in ontology structure. Third, we train an encoder with two objectives on the same textbook data: masked language modeling and relation prediction between code pairs. We refer to this co-training on shared data as *alignment*, and our experiments show that it is a necessary condition for the method to work.

We evaluate OntoBook on three French medical coding benchmarks and present ablation studies on the contribution of each component. We release 1.3 million LLM-reformulated medical textbooks across three French ontologies (CIM-10, CCAM, ATC) and all pretrained model checkpoints.

2. Related Work

This section reviews three research directions: biomedical language model pretraining, knowledge-enhanced language models, and automatic medical coding.

2.1. Biomedical Pretrained Language Models

Domain-specific pretraining has proven essential for biomedical NLP. BioBERT (Lee et al., 2020) further pretrained BERT on PubMed abstracts and PMC full-text articles, improving named entity recognition and relation extraction. PubMedBERT (Gu et al., 2022) demonstrated that pretraining from scratch on in-domain text outperforms mixed-domain initialization. ClinicalBERT (Huang et al., 2019) targeted clinical notes from MIMIC-III. For French, CamemBERT-bio (Touchent et al., 2023) adapted CamemBERT (Martin et al., 2020) to biomedical text, while DrBERT (Labrak et al., 2023) was trained on French biomedical and clinical text from the NACHOS corpus. These models learn from flat text corpora and do not capture the hierarchical structure of medical ontologies.

2.2. Knowledge-Enhanced Language Models

Random walks on knowledge graphs generate sequences capturing relational structure. RDF2Vec (Ristoski and Paulheim, 2016) adapted the walk-then-embed paradigm to RDF graphs. In the biomedical domain, Snomed2Vec (Agarwal et al., 2019) applied random walk and Poincaré embeddings to SNOMED-CT for clinical prediction tasks. OWL2Vec* (Chen et al., 2021) combined random walk sequences with lexical and logical features from OWL ontologies. These methods produce static embeddings (Word2Vec-style) and cannot be used for transformer pretraining.

Several approaches inject knowledge graph structure into transformers. ERNIE (Zhang et al., 2019) aligns entity embeddings with text representations. K-BERT (Liu et al., 2020) injects knowledge triples into the input sequence using soft-position indices and a visibility mask. For multi-task pretraining, KEPLER (Wang et al., 2021) combines MLM with TransE-style knowledge embedding on Wikidata. DRAGON (Yasunaga et al., 2022) jointly trains MLM and knowledge graph link prediction, with a biomedical variant on UMLS achieving +3% on MedQA. CODER (Yuan et al., 2022) uses UMLS relation triplets for contrastive pretraining, and SAPBERT (Liu et al., 2021) uses contrastive learning on UMLS synonyms for biomedical entity linking. Closest to the present work, BioOntoBERT (Shashikumar et al., 2023) generates sentences from biomedical ontologies using Onto2Sen templates and pretrains BERT with MLM on the resulting corpus. However, the template-based generation produces formulaic sentences (e.g., “X is a type of Y”) rather than fluent prose, and the method uses only MLM without multi-task relation prediction.

Recent work has explored synthetic data generation from knowledge sources. The Phi series (Gunasekar et al., 2023) demonstrated that LLM-generated textbook-quality data enables strong small-model performance. EntiGraph (Yang et al., 2025b) generates entity-centric synthetic data from knowledge sources but targets decoder language models rather than encoder pretraining.

2.3. Automatic Medical Coding

Medical coding assigns ICD, procedure, or medication codes to clinical text. CAML (Mullenbach et al., 2018) introduced label-wise attention for explainable predictions. PLM-ICD (Huang et al., 2022) showed that fine-tuning pretrained encoders with label attention directly improves coding performance. Kirchler et al. (2026) recently demonstrated that LLM embeddings improve transferability of EHR-based predictions across countries and coding systems, highlighting the importance of encoding medical knowledge structure for cross-system generalization. These methods improve the classification head or the input representations but rely on generic encoders pretrained on flat text corpora.

Table 1 summarizes how OntoBook combines components not jointly present in prior work.

3. Method

OntoBook converts medical ontology structure into pretraining signal through three stages: (1) generating random walks through ontology graphs, (2)

Method	Walks	LLM	MLM+Rel
Snomed2Vec	✓		
BioOntoBERT	✓		
DRAGON			✓
KEPLER			✓
Phi-1		✓	
OntoBook	✓	✓	✓

Table 1: Comparison with related approaches. OntoBook uniquely combines all three components for biomedical encoder pretraining.

reformulating these walks into fluent textbook prose using a large language model, and (3) training an encoder with multi-task learning combining masked language modeling and relation prediction. We first describe the ontologies used, then detail each stage. Figure 1 illustrates the full pipeline.

3.1. Medical Ontologies

We use three French medical ontologies published by the Agence du Numérique en Santé (ANS) through its terminology server (SMT) in RDF/OWL format. CIM-10 FR PMSI is the French adaptation of the WHO's ICD-10, enriched by the Agence Technique de l'Information sur l'Hospitalisation (ATIH) for the French hospital information system (Programme de Médicalisation des Systèmes d'Information, PMSI). It contains 19,161 diagnostic codes organized in a 5-level hierarchy (chapters, blocks, categories, subcategories), with textual annotations including 23,282 synonyms, 8,171 inclusion notes, and 381 definitions. CIM-10 is the only ontology with semantic edges beyond the hierarchy: 1,317 causal edges (e.g., E11 type 2 diabetes *causes* N08.3 diabetic glomerulopathy), 5,739 exclusion edges, and 341 manifestation edges, yielding a mean node degree of 2.77 compared to 2.00 for the purely hierarchical ontologies.

CCAM (Classification Commune des Actes Médicaux) is the French procedure classification maintained by the Caisse Nationale d'Assurance Maladie (CNAM), containing 38,191 procedure codes in a 4-level hierarchy with 8,991 synonyms and 1,188 disjointness constraints between mutually exclusive procedures. ATC (Anatomical Therapeutic Chemical) is the WHO medication classification with 6,950 substance codes in a strict 5-level hierarchy from anatomical group to chemical substance, with no semantic edges, no synonyms, and no definitions beyond labels. This contrast in relational richness directly shapes the walk generation strategy: CIM-10 walks can follow causal and differential diagnosis paths, whereas CCAM and ATC walks are limited to hierarchical and sibling traversals. Table 2 summarizes the structural properties of the three ontologies.

	CIM-10	CCAM	ATC
Graph structure			
Codes	19,161	38,191	6,950
Hierarchy depth	5	4	5
Hierarchical edges	19,139	38,191	6,950
Causal edges	1,317	—	—
Manifestation edges	341	—	—
Exclusion edges	5,739	—	—
Disjoint edges	—	1,188	—
Mean degree	2.77	2.06	2.00
Textual attributes			
Synonyms	23,282	8,991	0
Inclusion notes	8,171	—	—
Definitions	381	151	0

Table 2: Structural properties of the three French medical ontologies used for walk generation. CIM-10 is the only ontology with semantic edges beyond the hierarchy.

3.2. Ontology Walk Generation

We generate random walks through ontology graphs to capture relational structure in a sequential format suitable for language model pretraining. We sample walks starting from each code in the ontology. At each step, the next node is selected via weighted sampling where edge type weights depend on the walk type. This biased sampling is necessary because semantic edges are rare (Table 2), and a uniform walk would rarely traverse them. Each walk records the traversed codes along with their labels and the relation types connecting them. Walk length is sampled uniformly between 8 and 20 steps. We track visited nodes to avoid cycles, with a fallback that allows revisiting nodes when all neighbors have been explored.

For CIM-10, we define five walk types, each emphasizing different aspects of medical knowledge: *Etiologie* (causal chains), *Diagnostic Différentiel* (codes to distinguish clinically), *Codage Double* (mandatory code pairs linking etiology to manifestation), *Syndrome* (multi-system manifestations), and *Cross-Chapter* (connections between different ICD chapters, e.g., diabetes E11 to its renal complication N08). We generate 402,328 walks for CIM-10 (average 4,830 characters), 762,958 for CCAM, and 138,980 for ATC, totaling 1,304,266 walks.

3.3. Textbook Generation

Raw walks contain structured markers (e.g., "» `À distinguer de:`") that are useful for parsing but not natural for language model pretraining. We use

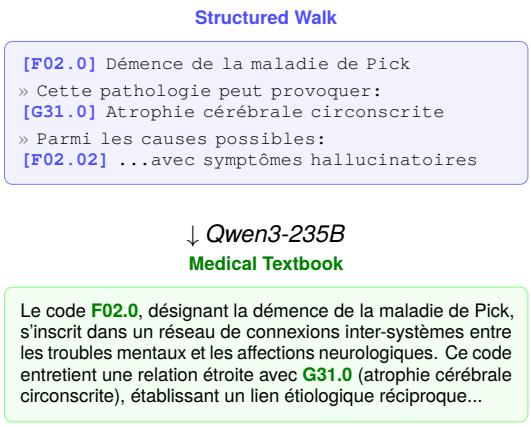

Figure 2: Walk-to-textbook transformation. The structured walk is reformulated into fluent medical prose while preserving all codes and relations.

a large language model to reformulate walks into fluent medical prose resembling textbook paragraphs.

We prompt Qwen3-235B-A22B-Instruct (Yang et al., 2025a) with FP8 quantization to transform each walk into a coherent paragraph. The prompt instructs the model to: (1) preserve all medical codes and their relationships exactly as stated, (2) use natural medical language without lists or bullet points, (3) not add any information beyond what appears in the walk. This ensures the textbook content remains grounded in the ontology structure. We apply two filters: we discard reformulations shorter than 50 characters or that fail to mention the source codes. The reformulation runs on 4 NVIDIA H100 GPUs using vLLM (Kwon et al., 2023) with guided decoding to ensure valid JSON output, taking approximately 20 hours for all 1.3 million walks. The resulting textbooks contain approximately 1.3 billion tokens across all three ontologies. Figure 2 shows an example transformation.

Figure 3 shows the system prompt used for CIM-10 walks (translated from French for readability; prompts for CCAM and ATC follow the same structure with ontology-specific terminology).

3.4. OntoBook Training

We train a ModernCamemBERT-base encoder (Antoun et al., 2025), initialized from a French checkpoint trained on 1 trillion tokens.

The model optimizes two objectives jointly. The first, $\mathcal{L}_{\text{MLM}}$, follows standard masked language modeling (Devlin et al., 2019): we mask 15% of tokens from textbook paragraphs and train to predict them. The second, $\mathcal{L}_{\text{rel}}$, classifies the relationship between pairs of medical codes. Our main model uses CIM-10; we explore transfer from CCAM and ATC in Section 4.5. We extract 282,907 code pairs from the CIM-10 ontology across 6 relation types: PARENT, CHILD, SIBLING, DIFFERENTIAL

Reformulation Prompt (CIM-10, translated)

You are an expert medical writer specializing in high-quality educational material.
Reformulate the provided CIM-10 walk into fluent, textbook-quality medical prose.
Strict rules: (1) **No invention**: never add information not in the source. (2) **Total fidelity**: preserve all definitions, codes, clinical notes, exclusions. (3) **Completeness**: do not omit any source information. (4) **Ignore metadata**: never mention walk types or structural headers.
Style: Continuous narrative prose, no bullet points or markdown. Integrate CIM-10 codes naturally. Preserve exact medical terminology.
Goal: Transform the walk into text suitable for a medical nosology textbook chapter while remaining 100% faithful to the original content.

Figure 3: System prompt for LLM-based walk reformulation. Temperature is set to 0 for deterministic, faithful output.

DIAGNOSIS, CAUSES, and CAUSED_BY. For each pair, we retrieve the corresponding textbook descriptions generated in Section 3.3. The input format is "[CLS] text$_A$ [SEP] text$_B$ [SEP]", where text$_A$ and text$_B$ are the textbook descriptions of the two codes, and a separate linear classification head on the [CLS] token predicts the relation type, trained with cross-entropy loss over the 6 classes.

The combined loss is:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \lambda \mathcal{L}_{\text{rel}} \tag{1}$$

where $\lambda = 1.0$. We train for 4 epochs with learning rate 2×10^{-5}, batch size 128, and AdamW optimizer with weight decay 0.01 and 1,000 warmup steps. Training takes approximately 4 hours per epoch on one NVIDIA H100 GPU. The key design choice is that both objectives operate on the *same textbook data*. We call this property *alignment*, and we show in Section 4.3 that it is essential for performance.

4. Experiments

4.1. Experimental Setup

Evaluation Benchmarks. We evaluate on the standard French medical coding evaluation suite:

- **FRACCO** (Pignat et al., 2025): a native French oncology corpus with 1,301 clinical cases annotated with ICD-O-3.1 codes, framed as multi-label classification over the top 100 codes.
- **Cantemist-FR** (Zaghir et al., 2024): the French adaptation of the Spanish tumor coding task (Miranda-Escalada et al., 2020), with 2,051 clinical documents.
- **Distemist-FR** (Zaghir et al., 2024): the French adaptation of the disease mention coding task (Miranda-Escalada et al., 2022).

Model	FRACCO	Cant.	Dist.	Avg.
External baselines				
CamemBERT-bio	20.2±0.2	12.1±0.4	9.0±0.2	13.8
DrBERT	36.3±0.7	37.7±1.0	22.5±0.7	32.1
ModernCamemBERT	56.4±1.0	63.5±1.3	23.4±1.7	47.7
Our models				
CodeInfill+MLM	56.5±0.2	66.4±1.9	20.0±0.9	47.6
MLM-only	55.8±0.4	66.0±1.1	24.2±5.8	48.7
OntoBook	**58.3**±0.3	**67.1**±1.1	**32.2**±1.1	**52.5**

Table 3: Main results on French medical coding (micro-F1 %). External baselines averaged over 9 seeds; our models over 3 seeds. Best results in bold.

Cantemist-FR and Distemist-FR are cross-lingual projections from Spanish. FRACCO is the only native French benchmark. All results are micro-F1 scores averaged over 3 random seeds.

Baselines. We compare against French biomedical language models: CamemBERT-bio (Touchent et al., 2023), DrBERT (Labrak et al., 2023), and ModernCamemBERT (Antoun et al., 2025) (which serves as our initialization checkpoint). We also report results for three ablations of our method: MLM-only (trained with $\mathcal{L}_{\text{MLM}}$ only on our textbooks), Rel-only (trained with $\mathcal{L}_{\text{rel}}$ only), and CodeInfill+MLM, which replaces relation prediction with a code infilling objective that masks medical code tokens (e.g., "E11", "N08.3") in textbook text and trains the model to predict them. We do not compare against knowledge-enhanced models such as DRAGON (Yasunaga et al., 2022) or SAPBERT (Liu et al., 2021) as they are English encoders that cannot be applied to French text.

Fine-tuning. For downstream evaluation, each pretrained encoder receives a clinical document as input and predicts a set of medical codes (multi-label classification over the top 100 codes per benchmark). We fine-tune with a linear classification head for 10 epochs, learning rate 2×10^{-5}, and batch size 16. External baselines (CamemBERT-bio, DrBERT, ModernCamemBERT) are averaged over 9 seeds; our models over 3 seeds due to the large number of configurations evaluated.

4.2. Main Results

Table 3 compares OntoBook against French biomedical baselines.

OntoBook outperforms all baselines on all three benchmarks. The improvement over MLM-only pretraining is significant on FRACCO (+2.5 micro-F1, $p < 0.001$) and Distemist (+8.0 micro-F1, $p < 0.001$), with consistent but not individually significant gains on Cantemist ($p = 0.11$). The gain is largest on Distemist, suggesting that ontology-aware pretraining particularly helps with fine-grained disease coding. OntoBook also reduces variance on Distemist from ±5.8 (MLM-only) to ±1.1.

Configuration	FRACCO	Cant.	Dist.	Avg.	Δ
OntoBook (aligned)	58.33	67.06	32.24	52.54	—
− $\mathcal{L}_{\text{rel}}$ (MLM-only)	55.81	66.01	24.23	48.68	−3.86
− $\mathcal{L}_{\text{MLM}}$ (Rel-only)	48.24	51.15	20.18	39.86	−12.68
− Alignment (misaligned)	33.39	20.11	12.63	22.04	−30.50
MLM-only (raw walks)	55.51	66.00	22.76	48.09	−4.45

Table 4: Ablation study (micro-F1 %). Misalignment degrades all benchmarks, with the largest drop on Cantemist (−46.95). All variants use the same base model and training budget.

4.3. Ablation Study

Table 4 ablates the key components of OntoBook.

Alignment is the most critical factor. Training $\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{rel}}$ on different data causes catastrophic degradation across all benchmarks, with Cantemist dropping from 67.06 to 20.11 (−46.95). Relation prediction alone also fails (−12.68 average F1), confirming that $\mathcal{L}_{\text{rel}}$ cannot serve as a standalone pretraining objective. The last row trains MLM-only on raw structured walks, removing both reformulation and relation prediction. The raw walk model (48.09) is close to the textbook MLM-only model (48.68), indicating that reformulation alone contributes little to MLM pretraining (+0.59 F1). However, the gap to full OntoBook is 4.45 points, reflecting the combined effect of adding both reformulation and relation prediction. This suggests that reformulation primarily benefits the relation prediction objective: on raw walks with structural markers (e.g., "» `À distinguer de:`"), the classifier can exploit formatting shortcuts rather than learning medical semantics, whereas fluent textbook prose forces the model to learn from content rather than format.

4.4. Training Dynamics and Hyperparameter Sensitivity

Figure 4 shows training dynamics and hyperparameter sensitivity. Epoch 1 (51.98) and epoch 2 (52.54) both outperform MLM-only pretraining (dashed line), but epoch 3 (49.43) degrades below it, possibly due to overfitting on the relation prediction task. All results in this paper use the epoch 2 checkpoint. The loss weight λ is robust across $\{0.1, 0.5, 1.0, 2.0, 5.0\}$: all values achieve 57.7–59.1% F1 on FRACCO. We use $\lambda = 1.0$ as it achieves the lowest variance (±0.52%). This robustness suggests that the exact balance between objectives matters less than ensuring both are present and aligned.

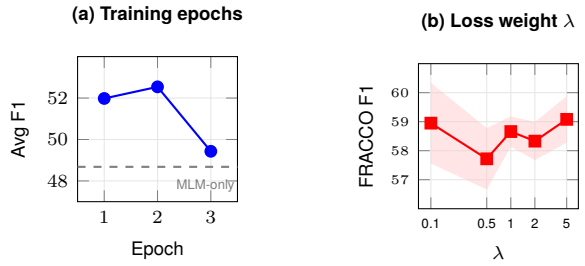

Figure 4: Training dynamics. (a) Average F1 peaks at epoch 2 then degrades. (b) Performance is stable across λ values (shaded: ± 1 std).

Ontology source	FRACCO	Cant.	Dist.	Avg.
CIM-10 (OntoBook)	58.33	67.06	32.24	52.54
All (CIM-10+CCAM+ATC)	59.31	**67.64**	31.44	52.80
CCAM only	58.59	66.54	34.26	**53.13**
ATC only	**59.66**	63.03	**36.05**	52.91

Table 5: Effect of ontology source on downstream performance (micro-F1 %). Each row trains a separate model using walks and relations from only that ontology. The main OntoBook model uses CIM-10 only. All models use the same training procedure and base checkpoint.

4.5. Multi-Ontology Transfer

The main OntoBook model uses CIM-10 walks (Table 3). To test whether other ontologies also provide useful pretraining signal, we train separate models on walks from each ontology individually and on all three combined.

All three individual ontologies improve over the MLM-only baseline, indicating that the pipeline generalizes across different medical knowledge structures. ATC (medications) yields the best Distemist score (+3.81 over CIM-10), despite Distemist being a disease coding task, suggesting that drug-disease relationships transfer well to disease understanding. CCAM (procedures) achieves the best average micro-F1 (53.13). Combining all three ontologies does not outperform the best individual ontologies. We hypothesize that this is because the relation prediction head must accommodate heterogeneous edge types across ontologies.

4.6. Probing Analysis

To understand how OntoBook organizes code representations, we evaluate with three probing tasks on CIM-10 code embeddings (mean-pooled over code descriptions): chapter classification (26 classes), hierarchy depth prediction (4 levels), and distance correlation (Spearman ρ between cosine distances and ontology tree distances). We compare against a code infilling variant (CodeInfill+MLM) that masks and predicts code tokens during pretraining.

Model	Chapter	Depth	Dist. ρ
Base	97.5%	99.3%	0.195
MLM-only	98.9%	99.4%	0.287
CodeInfill+MLM	**99.0%**	**99.8%**	**0.397**
Rel-only	98.0%	98.2%	0.203
OntoBook	98.7%	99.7%	0.073

Table 6: Probing analysis on CIM-10 code embeddings. CodeInfill+MLM achieves the best geometric organization but OntoBook achieves the best downstream performance (Table 3).

CodeInfill+MLM achieves the best probing metrics ($\rho = 0.397$) but the worst downstream performance among our models (47.6 average micro-F1, Table 3), while OntoBook has lower distance correlation ($\rho = 0.073$) despite achieving the best downstream performance (52.5 micro-F1). This inverse relationship suggests that for encoder pretraining, representational flexibility is more valuable than strict geometric preservation of ontology structure.

5. Discussion

The 30.50-point gap between aligned and misaligned training highlights the central role of objective alignment. When $\mathcal{L}_{\text{MLM}}$ trains on textbooks while $\mathcal{L}_{\text{rel}}$ trains on unrelated data, the model receives conflicting gradient signals: the MLM objective pushes representations toward textbook language patterns, while the relation objective pushes them toward a different data distribution. Aligned training avoids this conflict by ensuring both objectives reinforce the same semantic structure. Relation prediction guides attention toward ontologically meaningful patterns in the textbook text, while MLM ensures the model builds coherent representations of that text. This explains why relation prediction alone fails (-12.68 F1): it is a discriminative task that can exploit surface-level shortcuts, as demonstrated by the small gap between raw-walk and textbook MLM-only models (+0.59 F1). Combined with aligned MLM, however, it provides a complementary signal that improves downstream coding performance.

The cross-ontology transfer results suggest that our approach generalizes beyond CIM-10: training on CCAM or ATC alone also improves over the baseline despite targeting different medical domains. However, combining all ontologies does not outperform individual ones, possibly because the relation prediction head must accommodate heterogeneous edge types.

The probing analysis reveals a tension between geometric organization and downstream performance. CodeInfill+MLM achieves the best prob-

ing metrics ($\rho = 0.397$) but the worst downstream scores among our models (47.6 micro-F1), while OntoBook achieves the opposite ($\rho = 0.073$, 52.5 micro-F1). This suggests that for downstream coding tasks, flexible representations adapt better than rigid geometric structure that must be partially unlearned during fine-tuning.

LLM reformulation plays a dual role. For MLM alone, the effect is modest ($+0.59$ F1), but reformulation is essential for relation prediction: on raw walks, structural markers provide formatting shortcuts that the classifier can exploit without learning medical semantics, whereas fluent prose forces the model to learn from content rather than format. The full gap between OntoBook and MLM-only on raw walks is 4.45 points.

Limitations. All experiments use French medical coding. Extending to English ontologies such as ICD-11 and SNOMED-CT is left for future work. The reformulation step requires significant compute for LLM inference, and sensitivity to the choice of LLM has not been evaluated. Cantemist-FR and Distemist-FR are cross-lingual projections from Spanish (Zaghir et al., 2024), which may introduce translation artifacts; FRACCO is the only native French benchmark. All experiments use ModernCamemBERT; generalization to other architectures remains untested.

Future Work. Future work will extend OntoBook to English medical coding using UMLS and SNOMED-CT ontologies. A promising direction is cross-ontology walks that traverse edges between different ontologies (e.g., from a CIM-10 diagnosis to its ATC treatment to the CCAM procedure), generating richer walks that capture inter-ontology relationships in a single sequence.

6. Conclusion

We presented OntoBook, a method that converts medical ontology structure into pretraining signal for encoder language models through random walks, LLM-based textbook reformulation, and multi-task learning combining masked language modeling and relation prediction. Our key finding is that alignment between objectives is essential: both tasks must train on the same data, as misalignment causes a 30-point degradation. On French medical coding benchmarks, OntoBook improves over MLM-only pretraining by +2.5 micro-F1 on FRACCO and +8.0 on Distemist. We release 1.3 million LLM-reformulated medical textbooks across three French ontologies (CIM-10, CCAM, ATC) and pretrained model checkpoints.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011014393R2 made by GENCI.

7. Bibliographical References

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne Tamang, and Robert Rallo. 2019. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics. In *Proceedings of the KDD Workshop on Applied Data Science for Healthcare (DSHealth)*.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2025. ModernBERT or DeBERTaV3? Examining architecture and data influence on transformer encoder models performance. In *Proceedings of the International Joint Conference on Natural Language Processing and the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)*.

Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. 2021. OWL2Vec*: Embedding of OWL ontologies. *Machine Learning*, 110(7):1813–1845.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Suriya Gunasekar, Yi Zhang, Jyoti Anber, Sébastien Bubeck, Ronen Eldan, Neel Ganapathy, Yin Tat Lee, Yuanzhi Li, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding

with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20. ACL.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Matthias Kirchler, Marcelo Ferro, Valentina Lorenzini, et al. 2026. Large language models improve transferability of electronic health record-based predictions across countries and coding systems. *npj Digital Medicine*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*, pages 611–626. ACM.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238. ACL.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. ACL.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, and Martin Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the Cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*, volume 2664, pages 303–323.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: Results, methods, evaluation and multilingual resources. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2022), CEUR Workshop Proceedings*, volume 3180, pages 179–203.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. 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*, pages 1101–1111. ACL.

Johann Pignat, Milena Vucetic, Christophe Gaudet-Blavignac, Jamil Zaghir, Amandine Stettler, Fanny Amrein, Jonatan Bonjour, Jean-Philippe Goldman, Olivier Michielin, Christian Lovis, and Mina Bjelogrlic. 2025. FRACCO: A gold-standard annotated corpus of oncological entities with ICD-O-3.1 normalisation.

Petar Ristoski and Heiko Paulheim. 2016. RDF2Vec: RDF graph embeddings for data mining. In *International Semantic Web Conference*, pages 498–514. Springer.

Supreeth P Shashikumar, Fatemeh Amrollahi, and Shamim Nemati. 2023. Leveraging biomedical ontologies to boost pre-training for language models. In *Proceedings of the International Conference on Biomedical Ontologies (ICBO)*, volume 3603 of *CEUR Workshop Proceedings*.

Rian Touchent, Laurent Romary, and Éric De La Clergerie. 2023. CamemBERT-bio: a tasty French language model better for your health. In *Actes de la 30ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 323–334.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Wang, Bowen Zheng, Bowen Yu, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. 2025b. Synthetic continued pretraining. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pages 37309–37323.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Jamil Zaghir, Mina Bjelogrlic, Jean-Philippe Goldman, Soukaïna Aananou, Christophe Gaudet-Blavignac, and Christian Lovis. 2024. FRASIMED: A clinical French annotated resource produced through crosslingual BERT-based annotation projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7450–7460, Torino, Italia. ELRA and ICCL.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451. ACL.