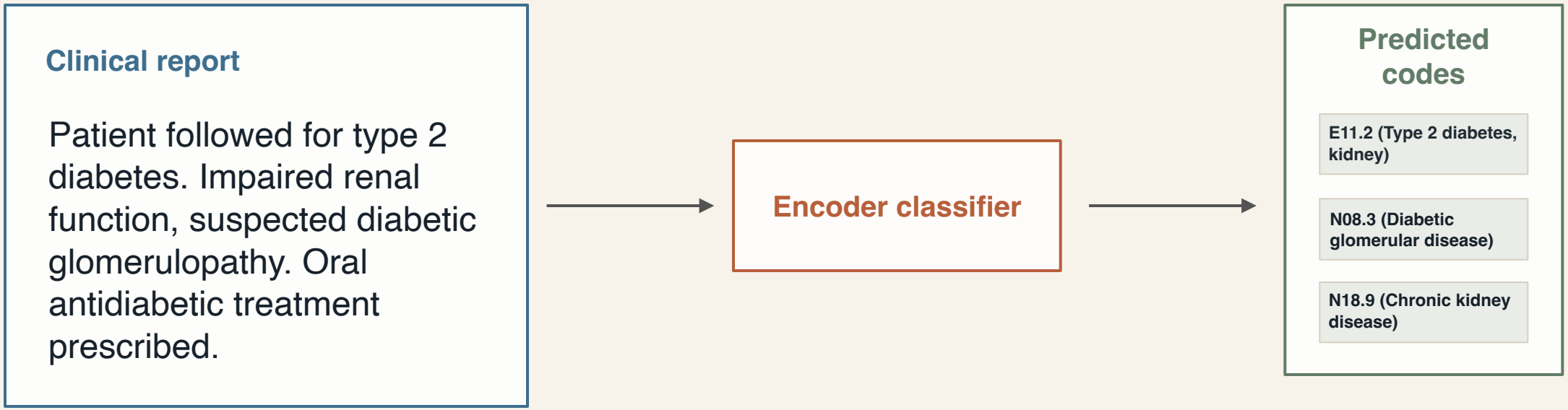
OntoBook: ontology-grounded synthetic textbooks for medical encoder pretraining

Turning medical code graphs into language modelling training signal

Rian Touchent, Éric de La Clergerie
Inria, Sorbonne Université

From clinical reports to standardized codes



Why it matters: in France, tens of millions of hospital stays per year require accurate coding for billing, epidemiological surveillance and clinical research.

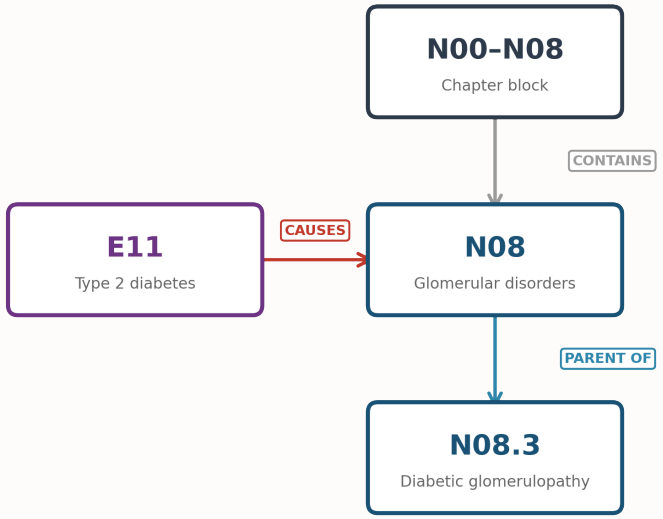
METHOD

The task needs ontology structure, but biomedical encoders mostly learn from flat text

What text pretraining sees

Mr. W., 26 years old, from New Caledonia, was referred to the West Paris allergy center (CAOP) one week after a shock following an infusion of Ambisome* (liposomal amphotericin B). This patient was first hospitalized over three years ago in Nouméa for a left apical thoracic mass eroding two ribs.

What coding also requires



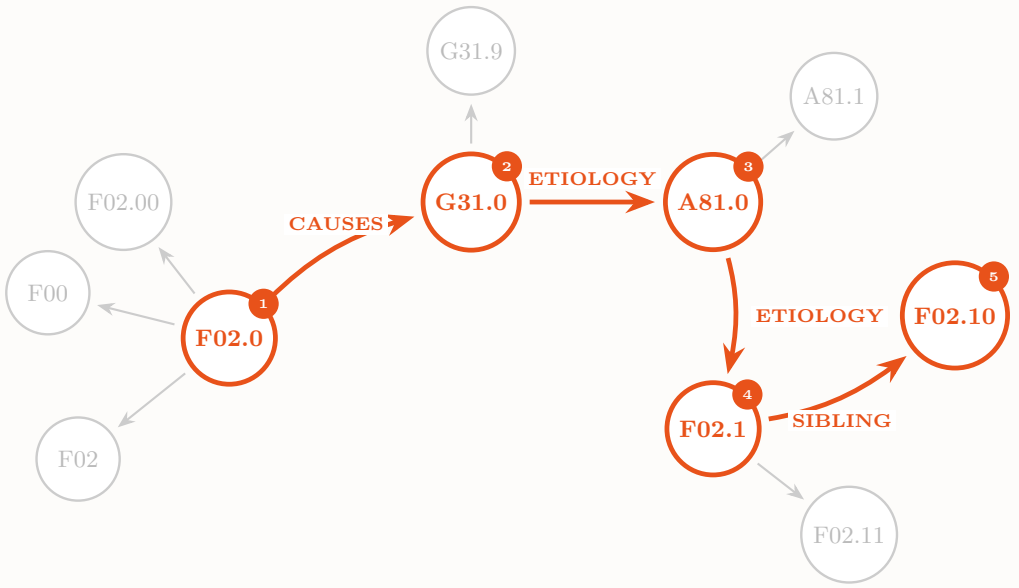
Research question: Can ontology structure become text that an encoder can pretrain on?

Linearizing ontology walks into training text

Relation types

1. **Etiology**
   what causes this disease?
2. **Differential**
   what is it not?
3. **Syndrome**
   what does it look like?
4. **Cross-chapter**
   how does it connect?
5. **Dual coding**
   what codes go together?

Random walk



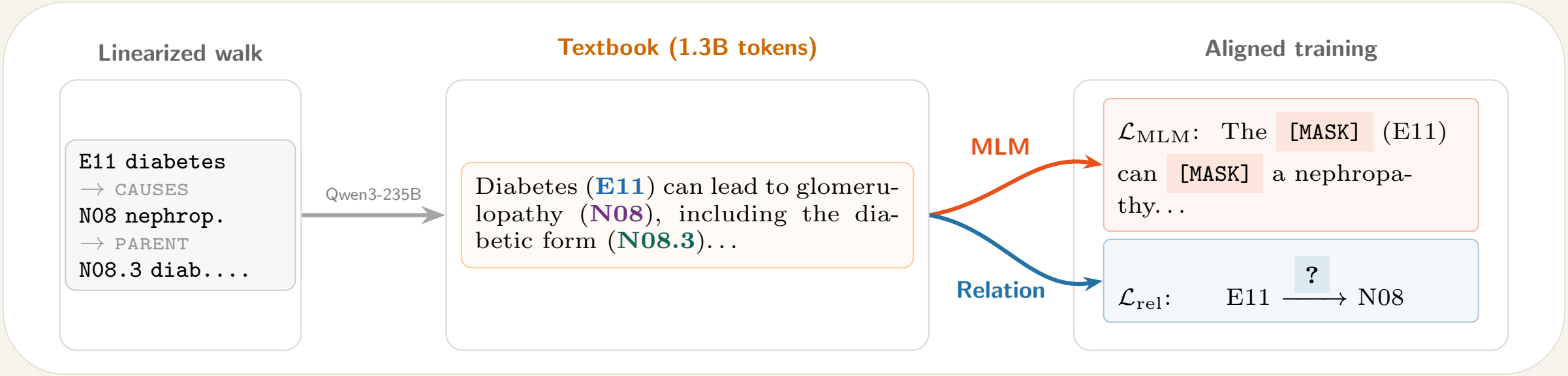
Linearize

```
> F02.0 -- Dementia in Pick's disease
-> This condition may be the cause of:
   G31.0 (Circumscribed cerebral atrophy)

> G31.0 -- Circumscribed cerebral atrophy
-> The etiology includes:
   A81.0 (Creutzfeldt-Jakob disease)

> A81.0 -- Creutzfeldt-Jakob disease
-> The etiology includes:
   F02.1 (Dementia in CJD)

> F02.1 -- Dementia in Creutzfeldt-Jakob
-> In the same nosological group:
   F02.10 (without additional symptoms)

> F02.10 -- Dementia in CJD [A81.0]
            - without additional symptoms
```

METHOD

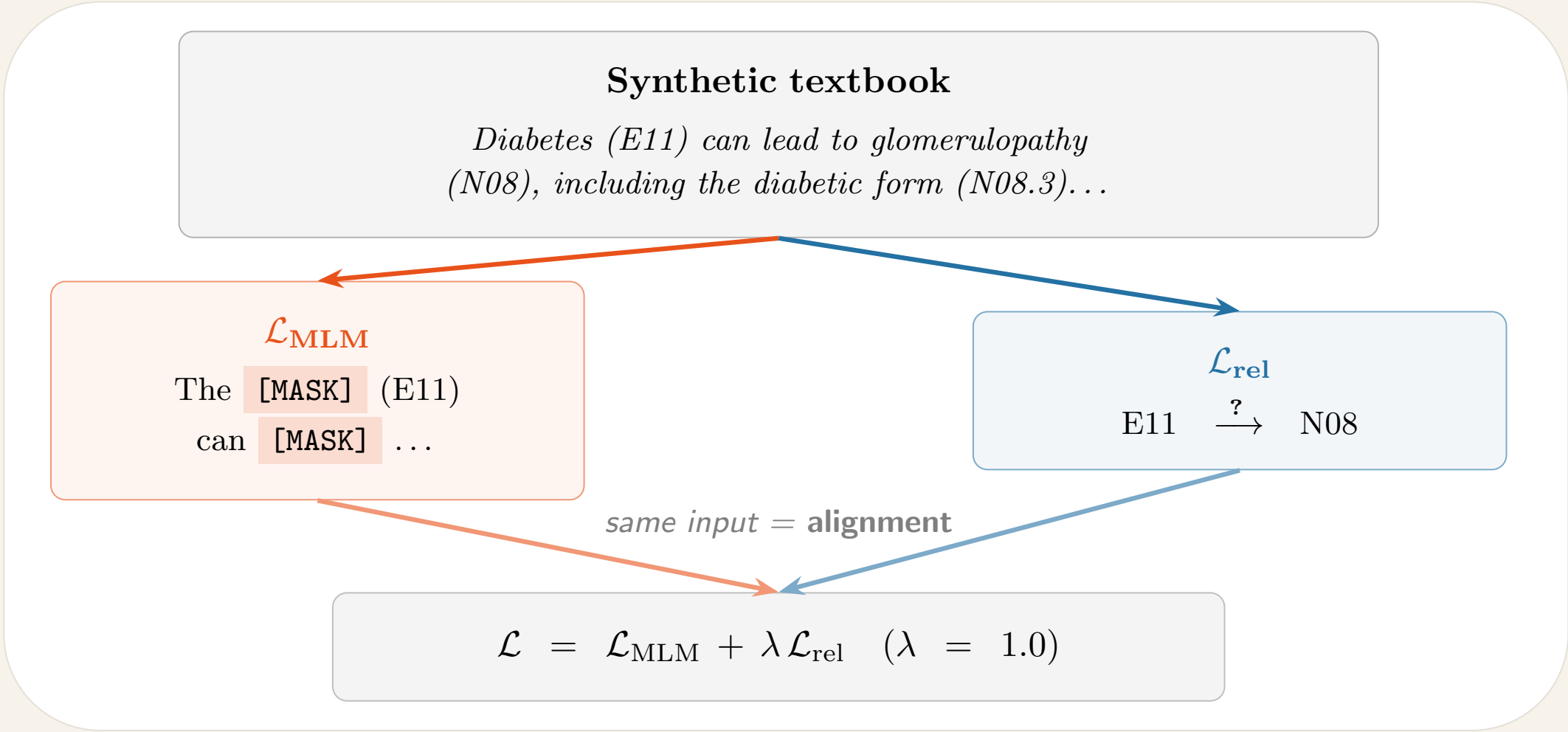
From graph to prose

Two objectives on the same generated prose



Relational structure becomes readable medical text, then MLM and relation prediction reinforce the same signal.

Both objectives read the same data



METHOD

Alternative objectives

MLM-only

$\mathcal{L}_{\text{MLM}}$

Diabetes `[MASK]` can `[MASK]` to `[MASK]` (N08)...

mask random tokens

CodeInfill

$\mathcal{L}_{\text{MLM}}$ (code tokens only)

Diabetes `[MASK]` can lead to glomerulopathy `[MASK]` ...

mask code tokens only

OntoBook

$\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{rel}}$

Diabetes (E11) `[MASK]` glomerulopathy `[MASK]` ...

$$\text{E11} \xrightarrow{?} \text{N08}$$

mask random + predict relation

RESULTS

OntoBook outperforms baselines and other training objectives

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**

PROBING

Better ontology geometry is not the same as better coding performance

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

CodeInfill probes best, but OntoBook performs best downstream: useful representations need flexibility, not only geometric fidelity.

ABLATION

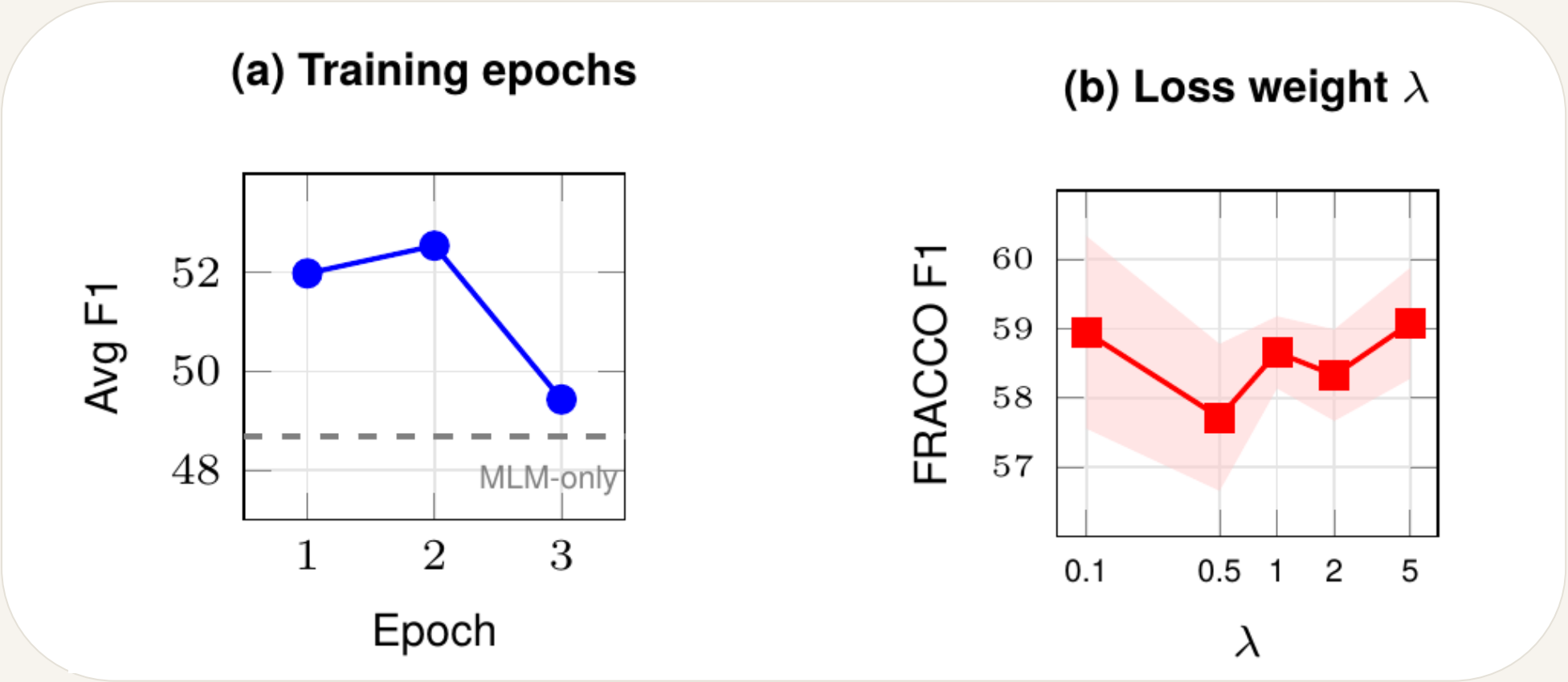
Misalignment costs 30 F1 points

We isolate the training objectives under the same base model and budget.

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

Misaligned training collapses by 30.5 average F1 points; relation-only also fails as a standalone objective.

Most of the benefit arrives early; the loss balance is robust



Transfer across ontologies

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

Conclusion

Ontology signal improves coding

- +4.8 avg F1 over ModernCamemBERT
- +8.8 F1 on Distemist, the hardest benchmark

Geometric fidelity hurts fine-tuning

- CodeInfill mirrors ontology best ($\rho = 0.397$)
- but scores worst downstream (47.6 vs 52.5)

Alignment is the key

- misalignment costs 30.5 F1 points
- MLM and $\mathcal{L}_{\text{rel}}$ must train on the same data

Next: cross-ontology walks

- CIM-10 diagnosis $\rightarrow$ ATC medication $\rightarrow$ CCAM procedure
- 1.3M textbooks + checkpoints released

Questions?

POSITIONING

OntoBook combines three ideas that prior approaches keep separate.

	Walks	LLM prose	MLM + relation
Snomed2Vec	✓		
BioOntoBERT	✓		
DRAGON / KEPLER			✓
Phi-style data		✓	
OntoBook	✓	✓	✓