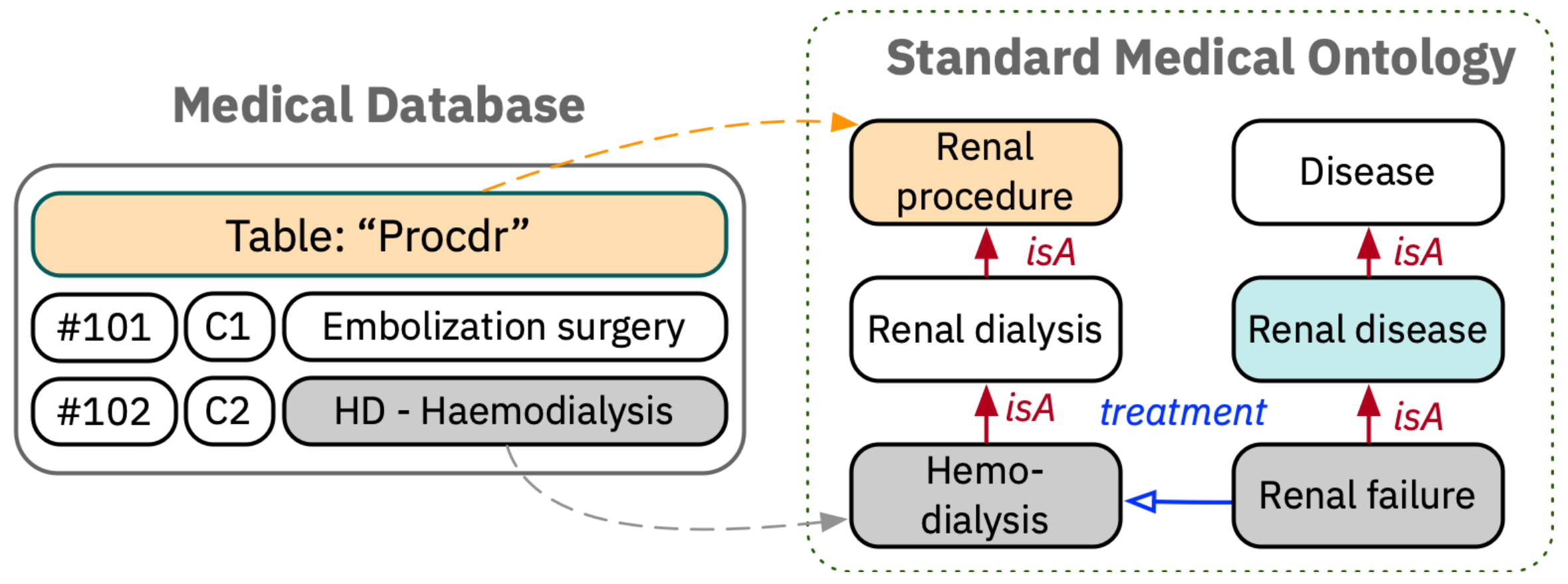
MEDTO: Medical Data to Ontology Matching Using Hybrid Graph Neural Networks

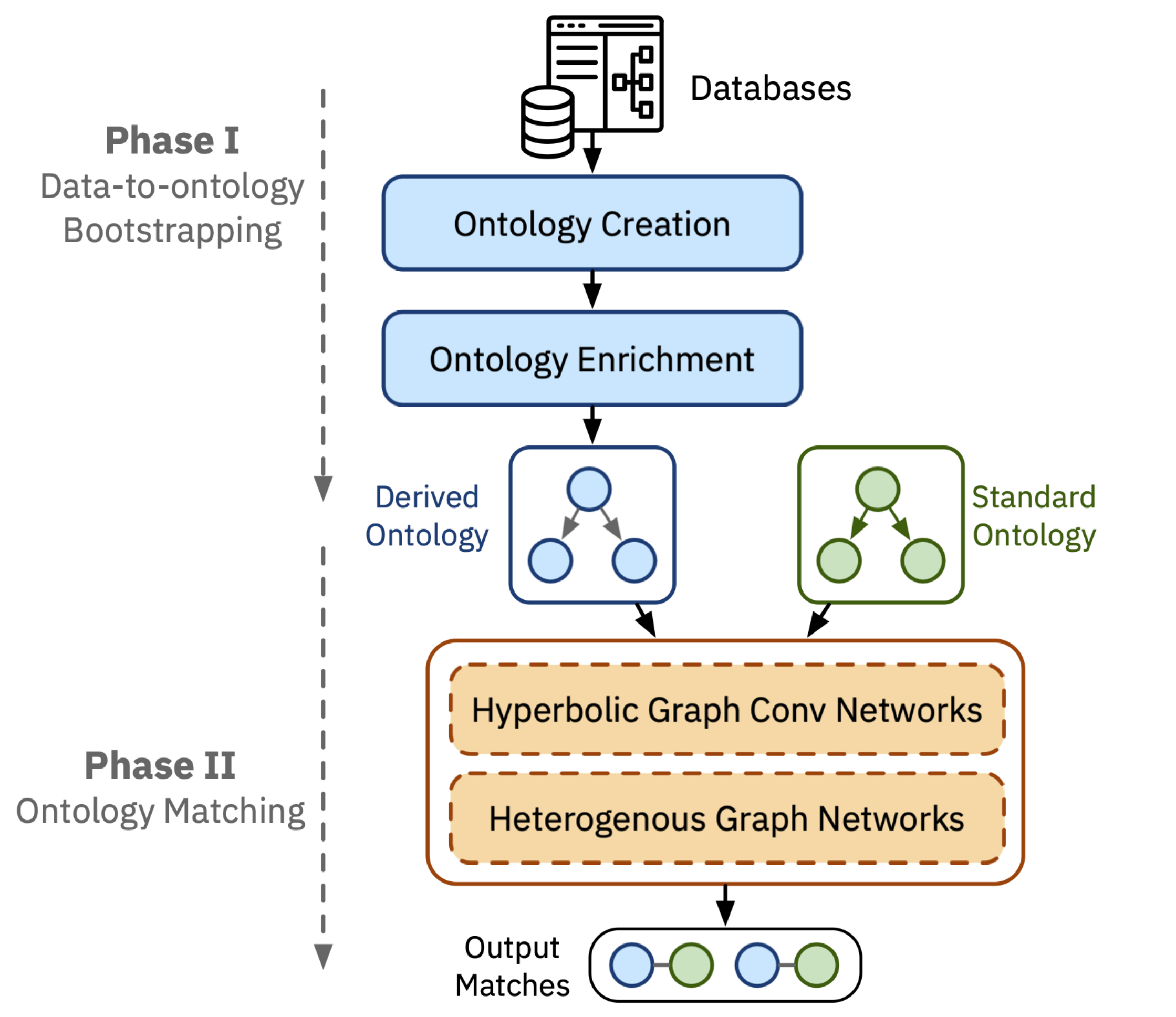
Junheng Hao, **Chuan Lei**, Vasilis Efthymiou, Abdul Quamar, Fatma Özcan, Yizhou Sun, Wei Wang
University of California Los Angeles, IBM Research - Almaden
Email: jhao@cs.ucla.edu | Website: https://www.haojunheng.com/project/medto

Data-to-Ontology Matching In Medical Domain

- Increasingly large-scale medical databases, in need of automatic AI-assisted analysis.
- Core task: Mapping database schema/tables to standard ontologies (for standardization)
- Existing methods focus on ontology matching, assuming ontologies are available for matching
- Effective data-to-ontology matching techniques

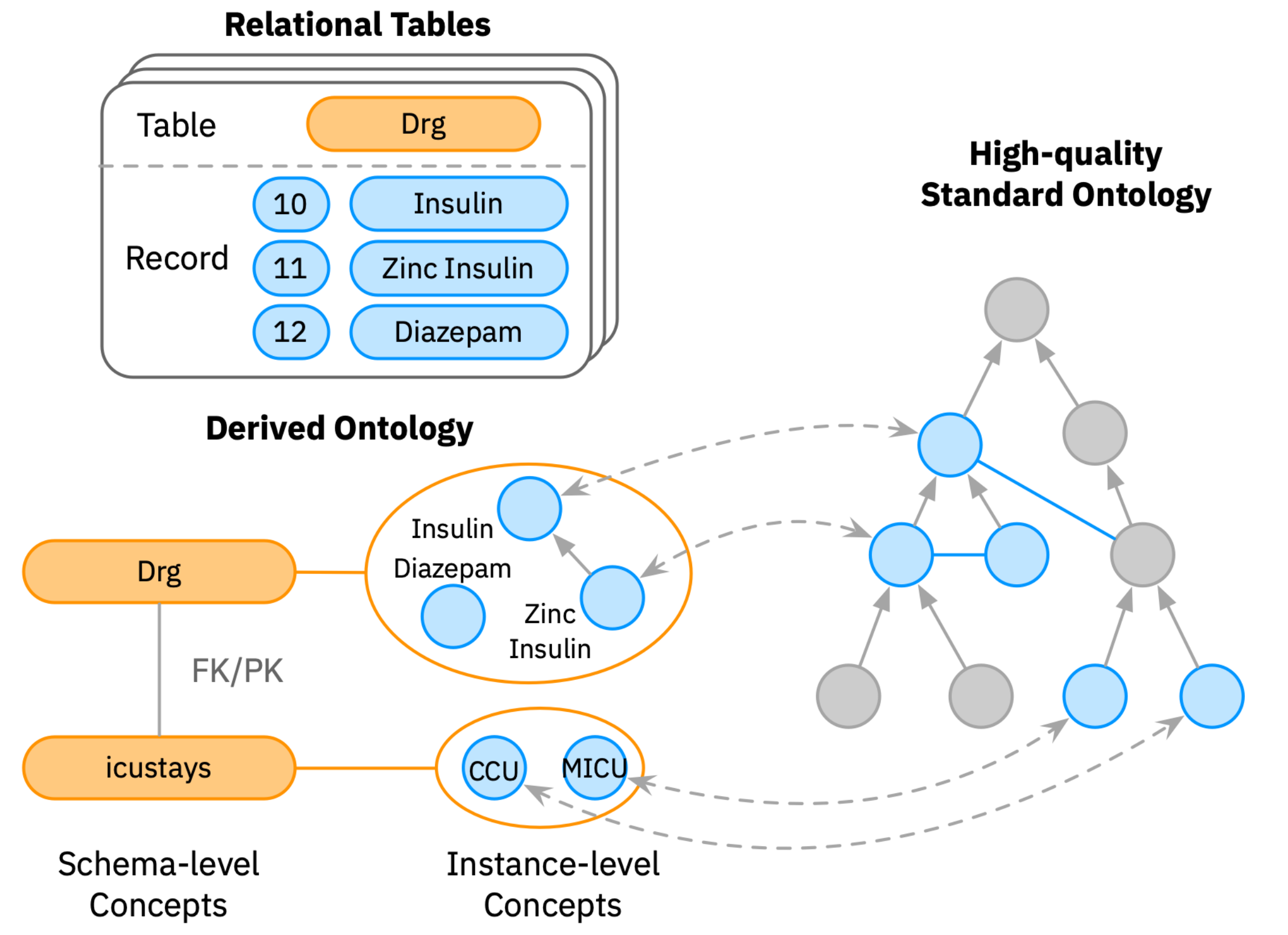


MEDTO: System Architecture

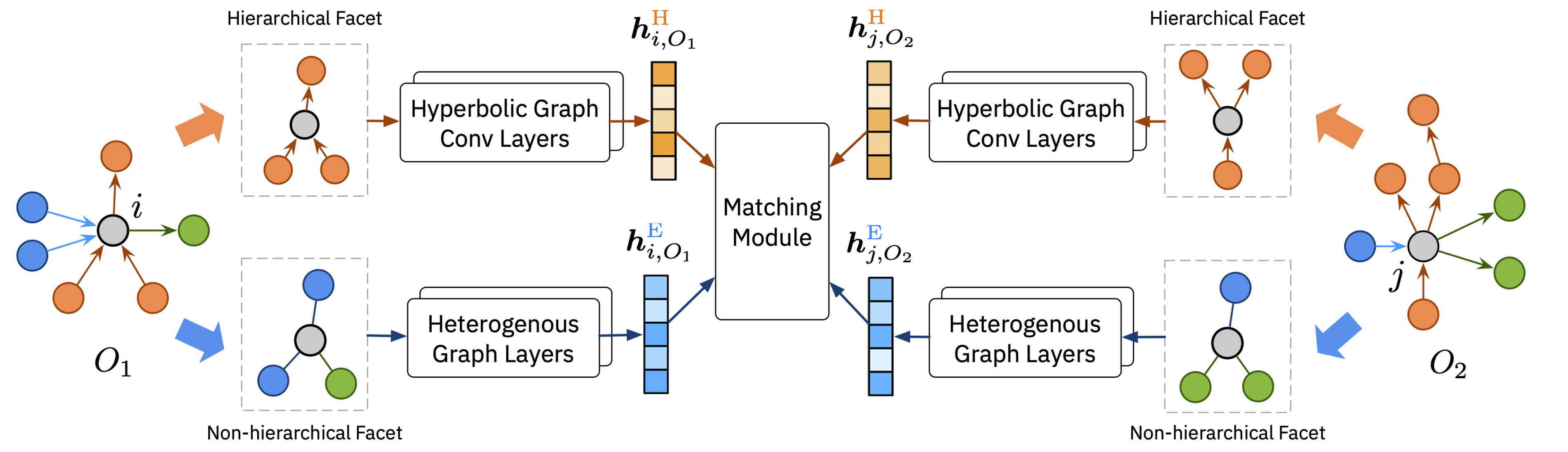


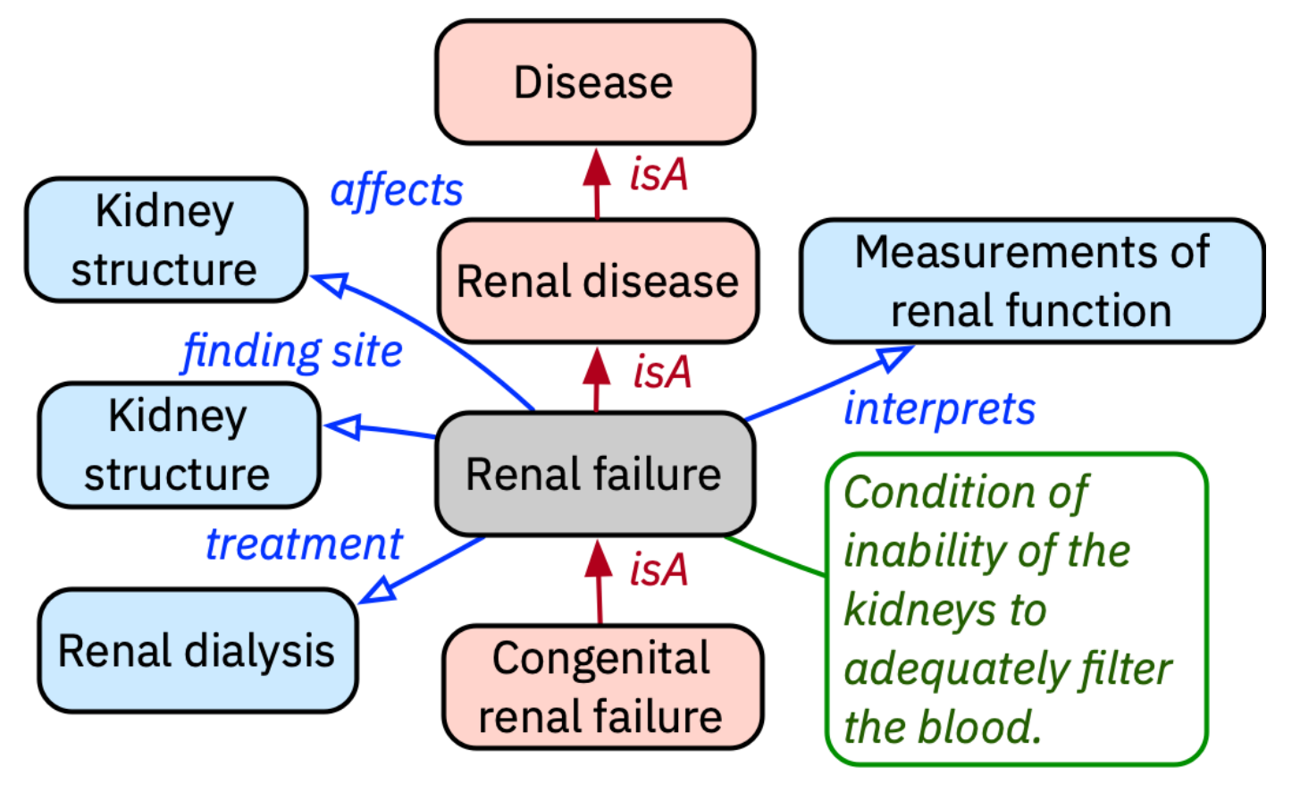
Phase 1: Bootstrapping

A two-step process from database tables: (1) Ontology creation; (2) Ontology Enrichment.



Phase 2: Ontology Matching (Graph Encoders)





Hyperbolic Graph Module (HYP): Better capture concept hierarchies or taxonomies in the medical ontologies in hyperbolic space (HGCN).

$$\mathbf{h}_i^{l,H} = \left(\mathbf{W}^l \otimes^{K_{l-1}} \mathbf{h}_i^{l-1,H}\right) \oplus^{K_{l-1}} \mathbf{b}^l,\ \mathbf{h}_i^{l,H} = \sigma^{\oplus^{K_{l-1},K_l}} \left(\mathrm{AGG}^{K_{l-1}}\left(\mathbf{h}_i^{l,H}\right)\right)$$

Heterogeneous Graph Module (HET): Model non-hierarchical relational facts with other concepts of multiple types in the ontologies

$$\mathbf{h}_i^{l,E} = \sigma(\mathbf{W}_0^l \cdot [\mathbf{h}_i^{l-1,E} || \mathbf{g}_i^{l-1,E}] + \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \frac{1}{c_{i,r}} \mathbf{W}_r^l \cdot [\mathbf{h}_j^{l-1,E} || \mathbf{g}_j^{l-1,E}])$$

Training

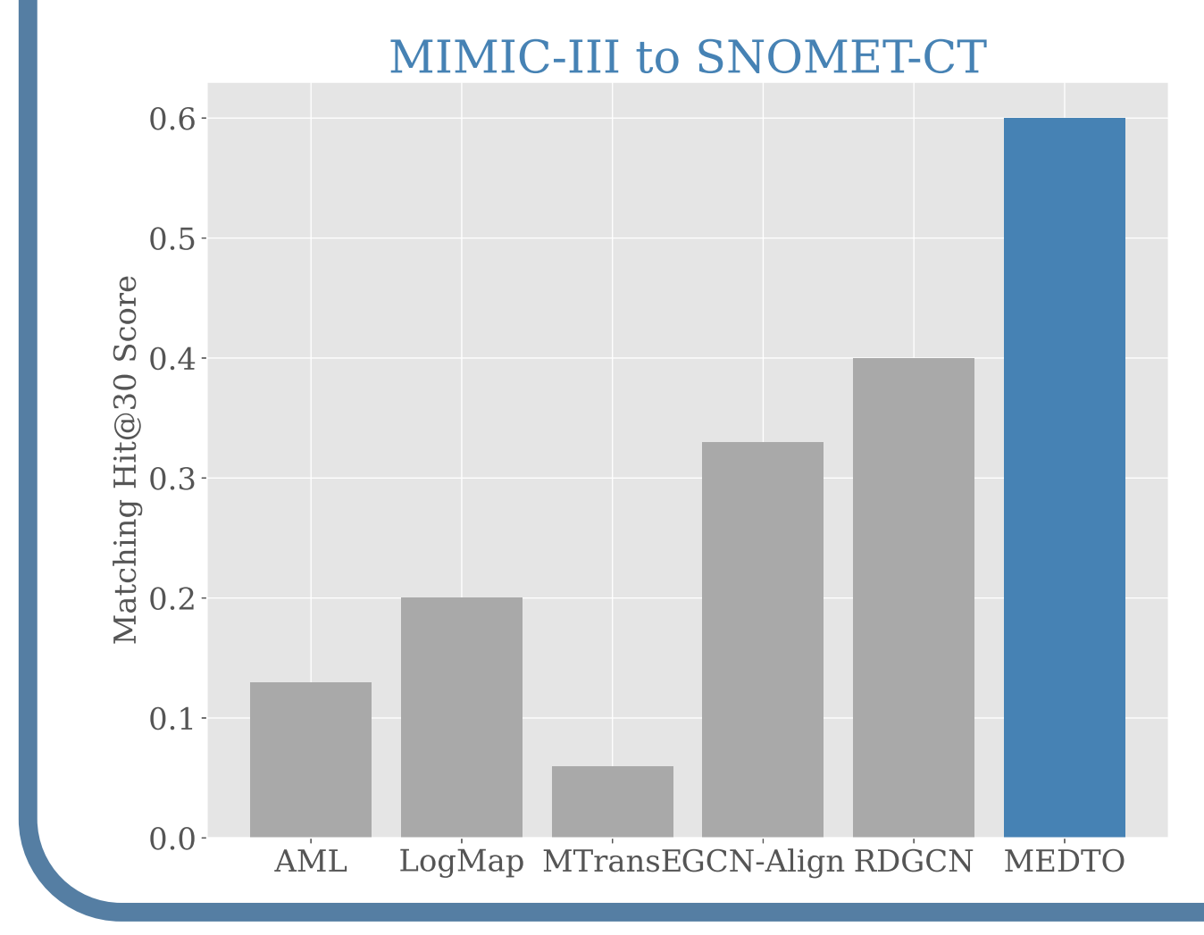
Matching: MLP, input as pairs of concept embeddings from O_1 and O_2

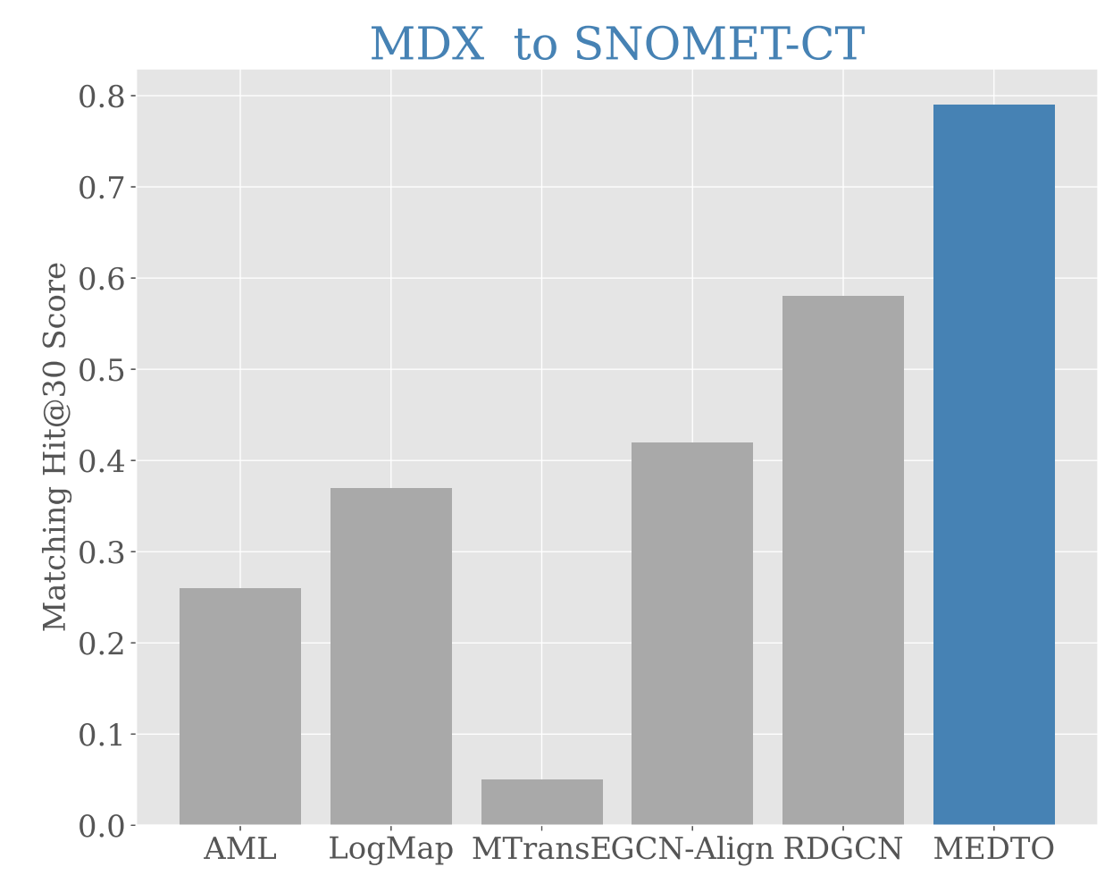
Training Loss: Matching loss + Graph decoders
$\mathcal{L} = \mathcal{L}^M + \alpha_1 \cdot (\mathcal{L}^{\mathrm{HYP}}_{\mathcal{O}_1} + \mathcal{L}^{\mathrm{HYP}}_{\mathcal{O}_2}) + \alpha_2 \cdot (\mathcal{L}^{\mathrm{HET}}_{\mathcal{O}_1} + \mathcal{L}^{\mathrm{HET}}_{\mathcal{O}_2})$

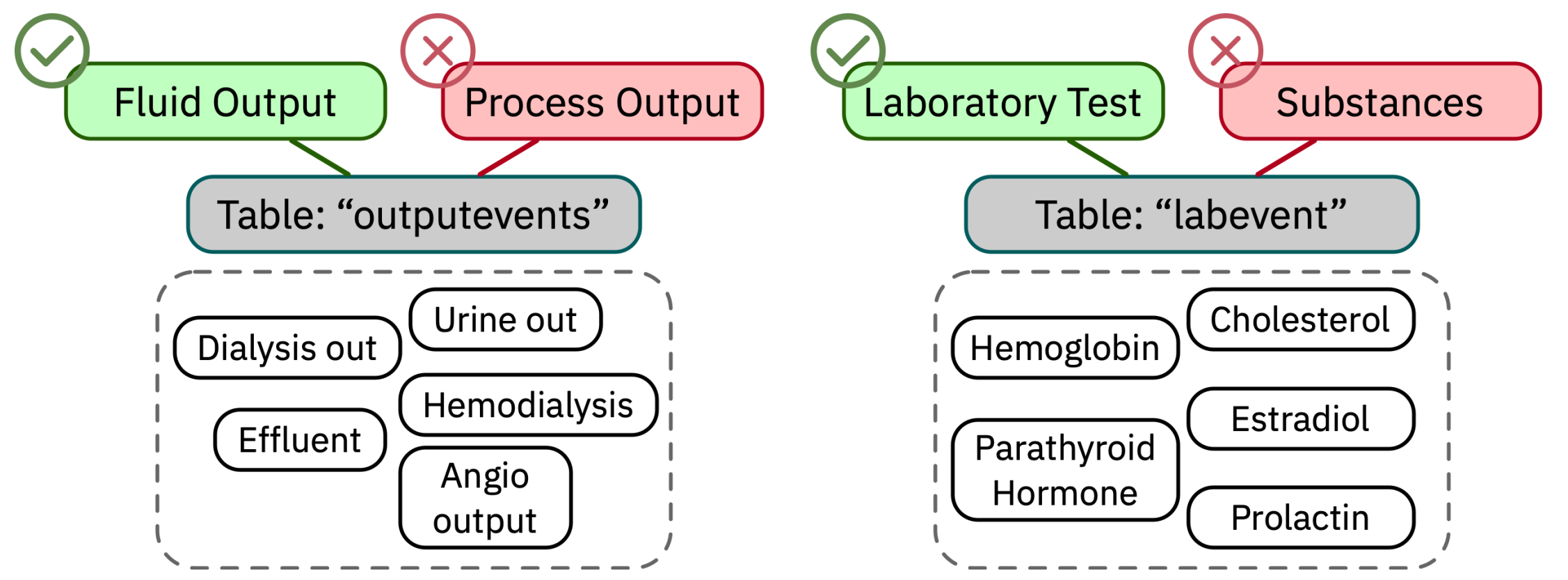
Dataset & Baselines

Medical Databases: MIMIC-III, IBM Micromedex (MDX)
Medical Ontologies: FMA, NCI, SNOMED-CT
Baselines: AML, LogMap, RDGCN, etc.

Experiments and Case Study on Medical Database



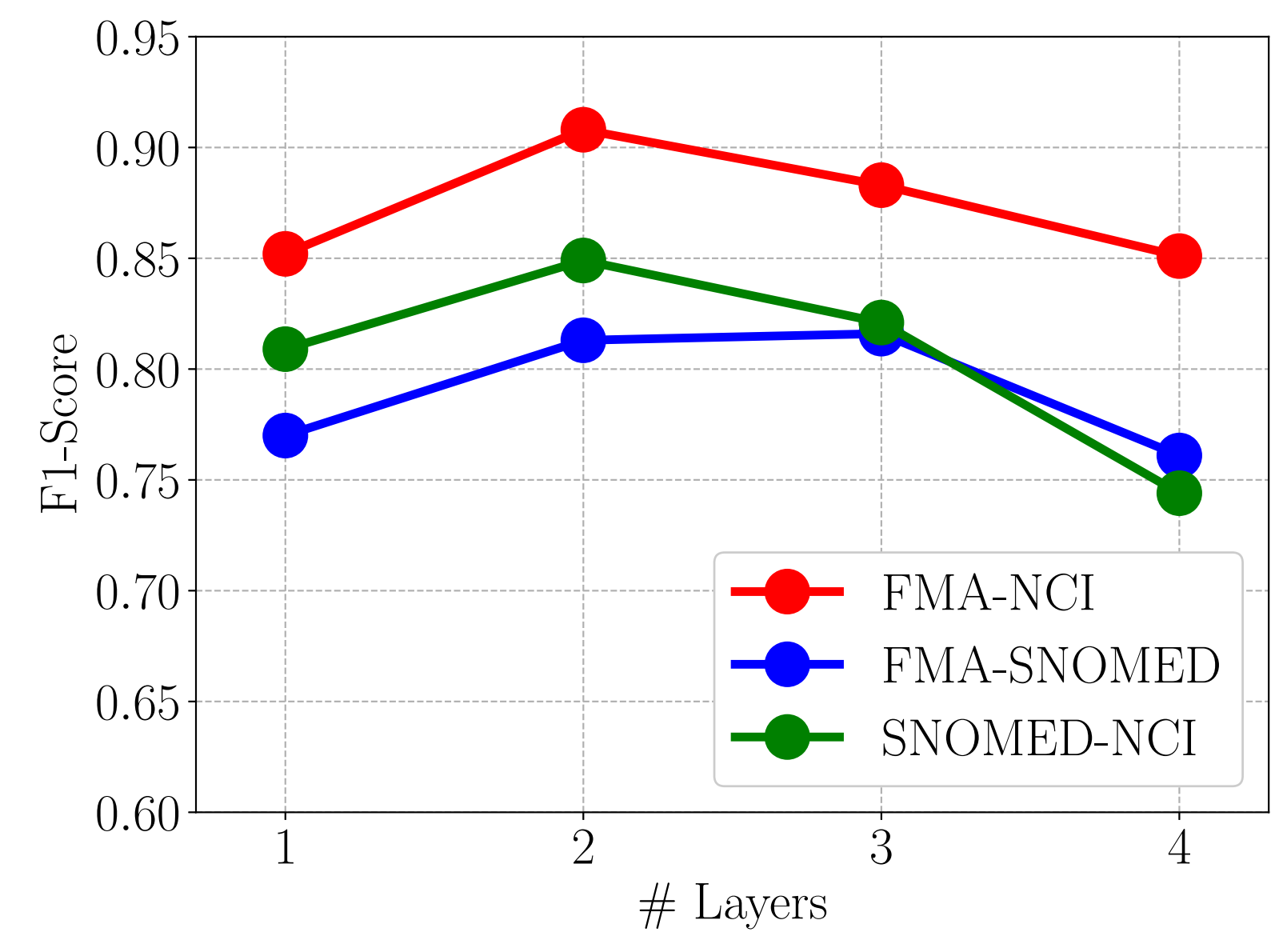


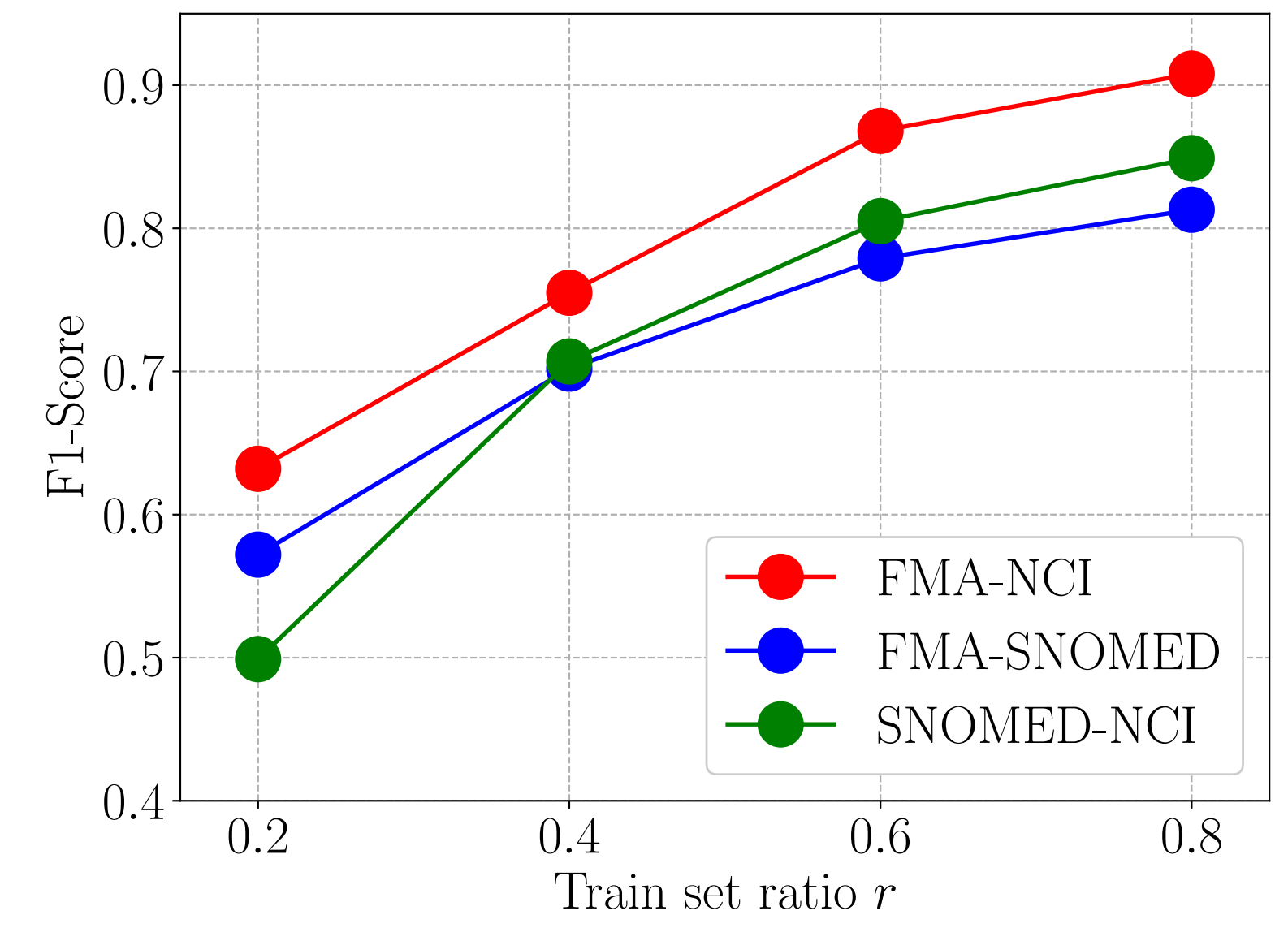


Experiments on Ontology Matching

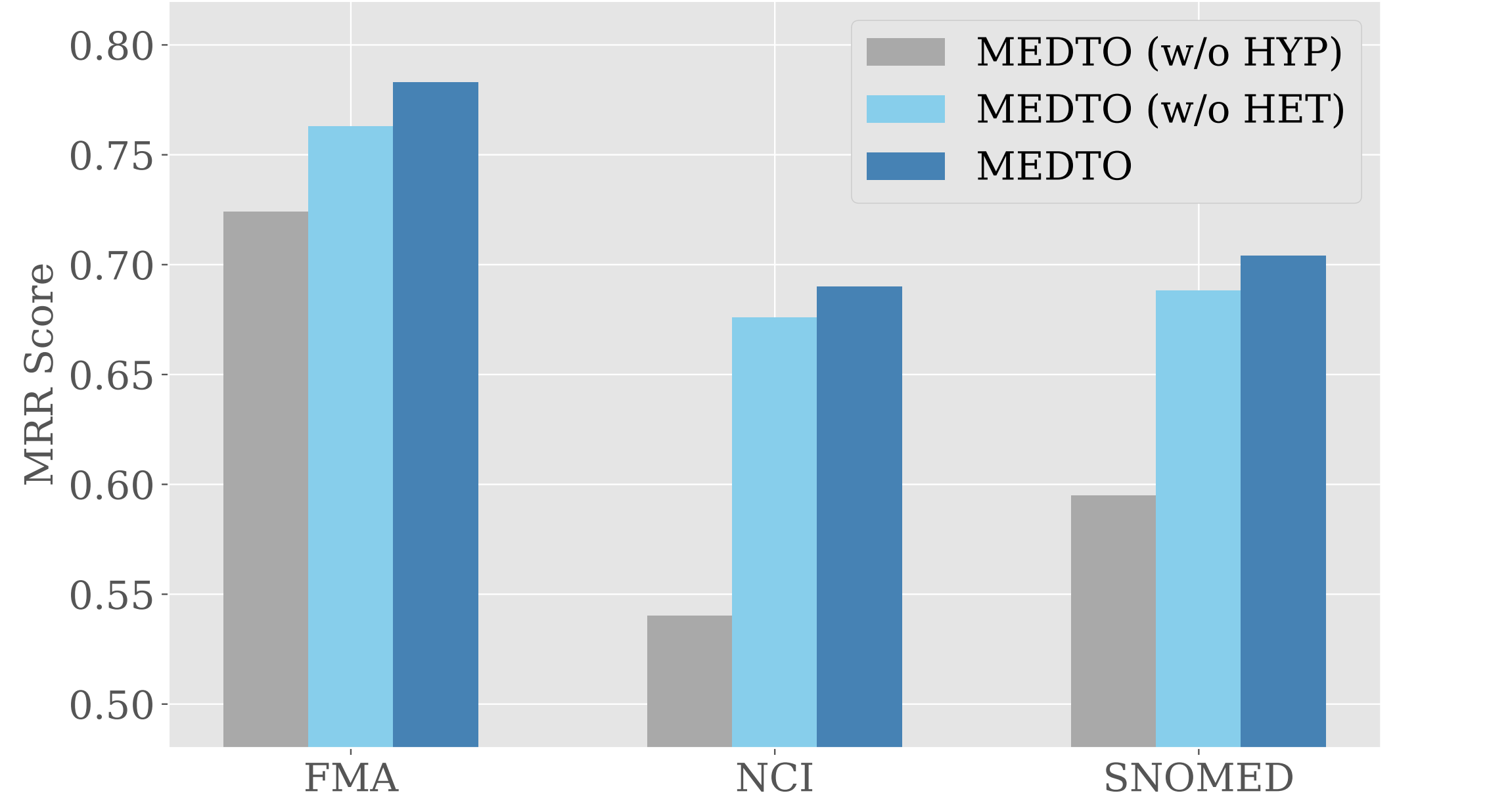
Model Groups	Datasets Metrics	FMA-NCI F1	 MRR	FMA-SNOMED F1	 MRR	NCI-SNOMED F1	 MRR
Rule-Based	AML	**0.920**	–	0.806	–	0.810	–
	LogMap	0.905	–	**0.819**	–	0.805	–
GNN-based Entity Alignment	MTransE	0.633	0.416	0.490	0.372	0.304	0.349
	GCN-Align	0.798	0.561	0.746	0.526	0.760	0.467
	RDGCN	0.849	0.761	0.786	0.683	0.816	0.679
Ours	MEDTO	0.908	**0.783**	0.813	**0.690**	**0.849**	**0.704**

Hyperparameters





MEDTO Variants



Acknowledgement

This work was done during the internship at IBM Research - Almaden. Special thanks to the members at IBM Research (Berthold Reinwald, Xiao Qin, Nasrullah Sheikh, Lingfei Wu, Yunyao Li, etc).