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

Applied Data Science Track Paper, KDD 2021

—

Junheng Hao[1,2], Chuan Lei[2], Vasilis Efthymiou[3], Abdul Quamar[2],
Fatma Özcan[4], Yizhou Sun[1], Wei Wang[1]

[1] University of California, Los Angeles (UCLA)
[2] IBM Research AI, Almaden
[3] FORTH-ICS
[4] Google

August, 2021 | Virtual Conference

IBM

Outline

UCLA IBM

Motivation: Data-to-Ontology Matching

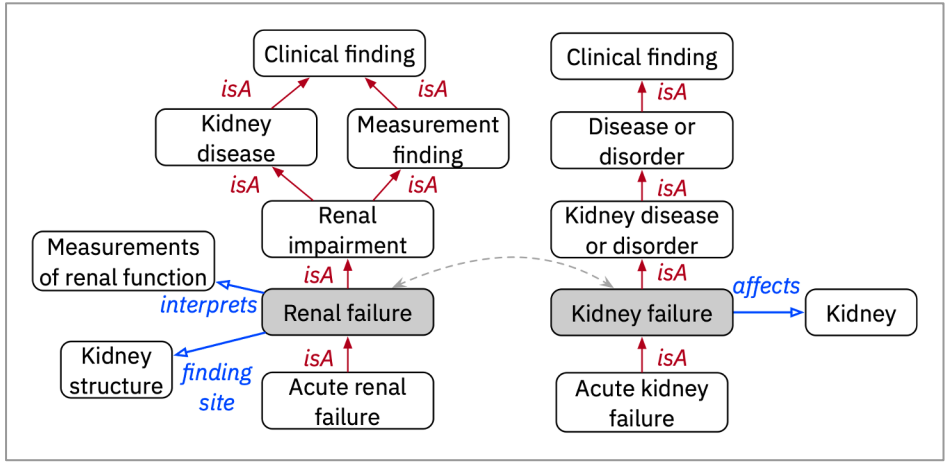
- MEDTO System Architecture Overview
- Model: Ontology Bootstrapping and Matching
- Experiments & Case Study
- Summary & Future Directions

World of AI-assisted Healthcare

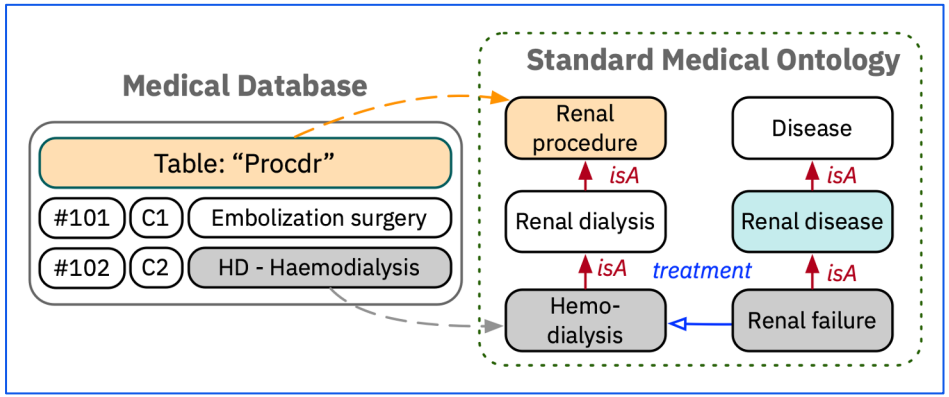
- Medical ontologies, many developed by experts, help define, standardize and organize concepts in the medical domain, which are foundational to support healthcare applications (such as clinic documentation, medical conversational system, Q&A).

Data-to-Ontology Matching

Ontology-Ontology Matching

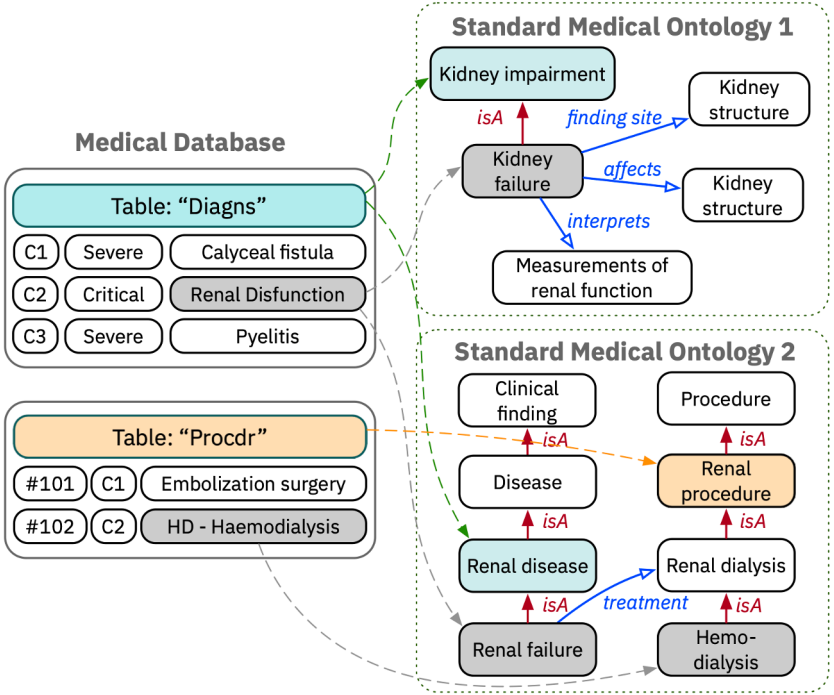


Data-to-ontology Matching



Data-to-Ontology Matching

- Large-scale clinical documents and medical record in databases
- Map database schema/tables to standard ontologies
  - Unifying and standardizing concepts in data
  - Enhancing downstream question answering and conversational systems
- Existing approaches are limited
  - Mappings between well-established ontologies cannot be directly applied on original data
  - Rule-based methods are hard to adapt to different domains → Low accuracy and robustness
- Challenges
  - Create a semantically rich ontology from databases
  - Effective matching techniques using various semantic features in the ontologies



Problem Statement

- Definitions:
  - Medical database D, represented by a relational schema S and its instances I
  - Medical ontology $O = (C, R, T)$, where C is the set of concepts, R is the set of relations, and $T = C \times R \times C$ is the set of triplets
- Problem Formulation:
  - Given a medical database D and a standard medical ontology O, the data to ontology matching problem is to find a set of matches M that map the schema S of D to the concepts in O, such that $\{(p, q) \in S \times O \mid p \equiv q\}$.

Outline

- Motivation: Data-to-Ontology Matching

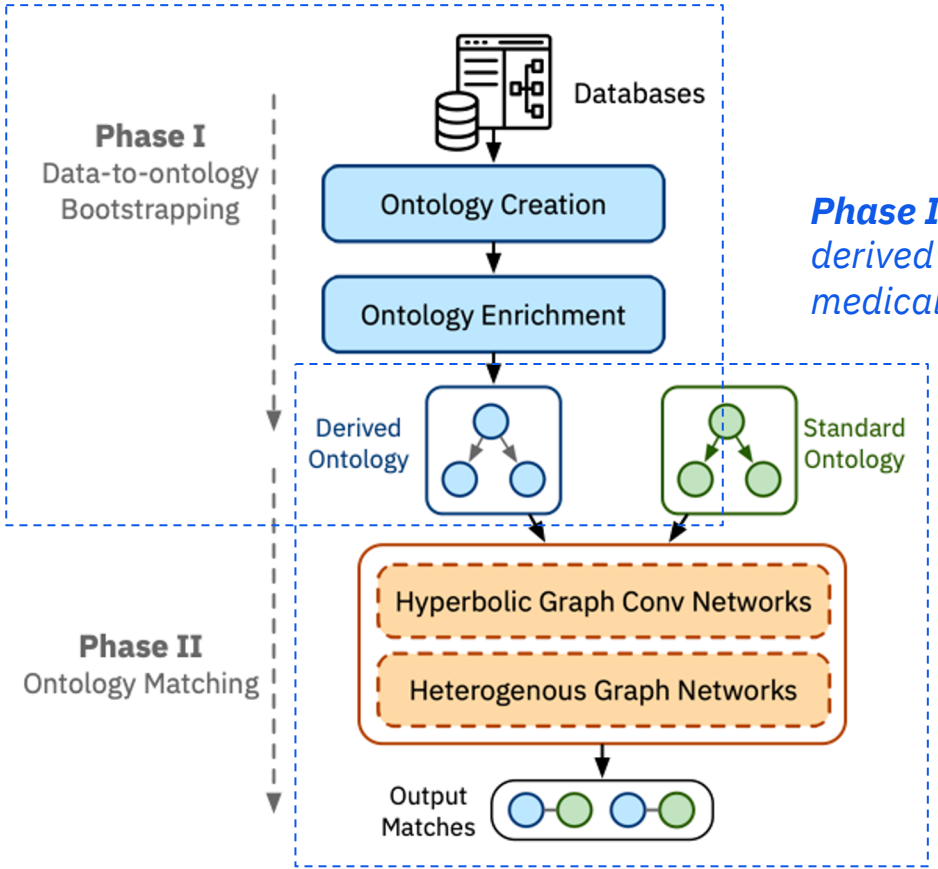
MEDTO System Architecture Overview

- Model: Ontology Bootstrapping and Matching
- Experiments & Case Study
- Summary & Future Directions

MEDTO System Architecture

Phase I: *A "cold-start" problem: How to create an ontology from a medical database?*

Ontology bootstrapping from medical databases



Phase II: *How to match the derived ontologies and standard medical ontologies?*

Ontology matching

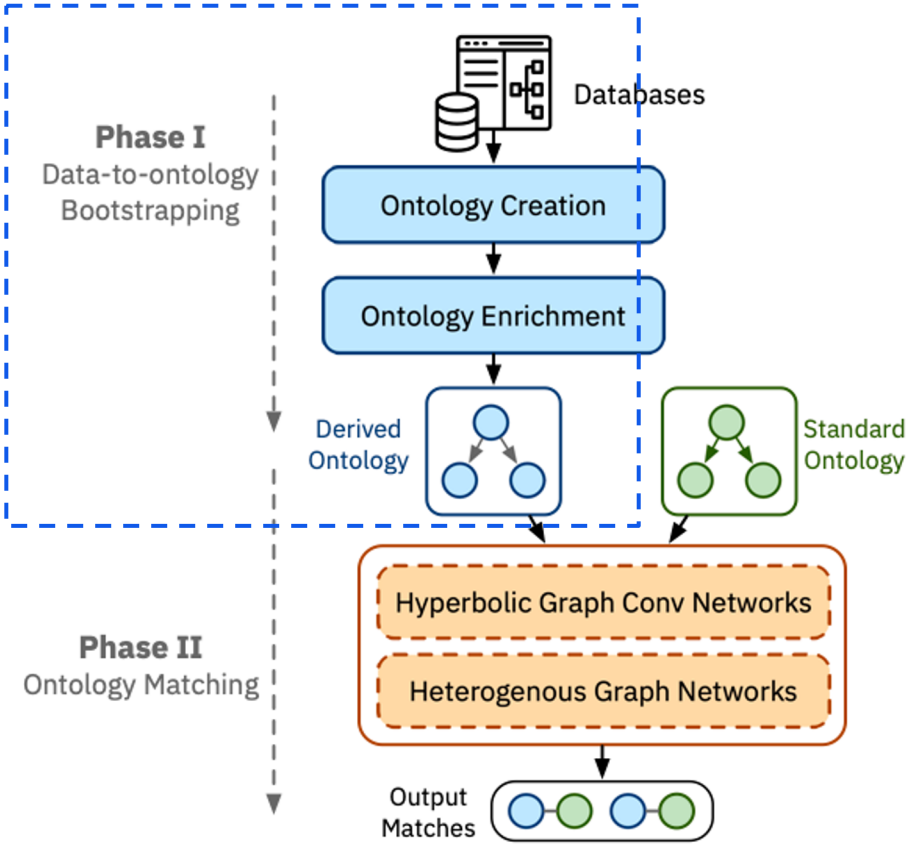
Outline

- Motivation: Data-to-Ontology Matching
- MEDTO System Architecture Overview
- **Model: Ontology Bootstrapping and Matching**
- Experiments & Case Study
- Summary & Future Directions

Phase I: Ontology bootstrapping

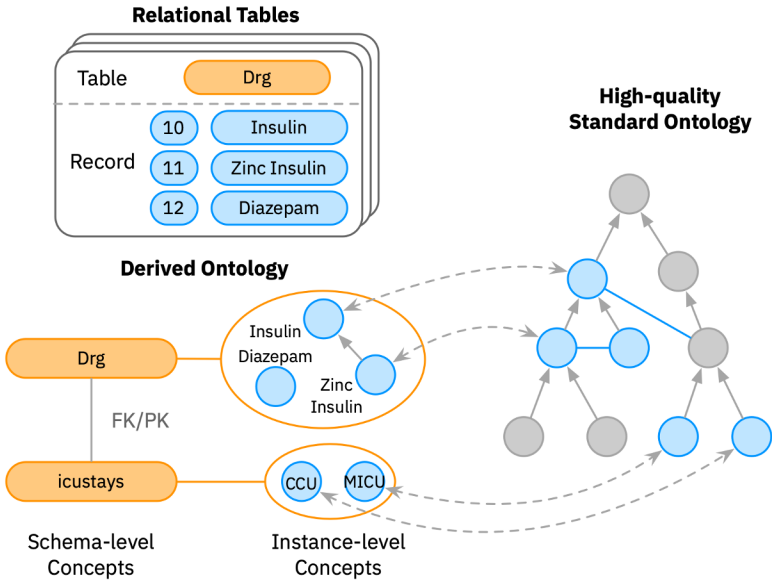
Goal: *As a "cold-start" problem: How to create an ontology from a medical database?*

Solution: *Ontology bootstrapping from medical databases*



Ontology Bootstrapping

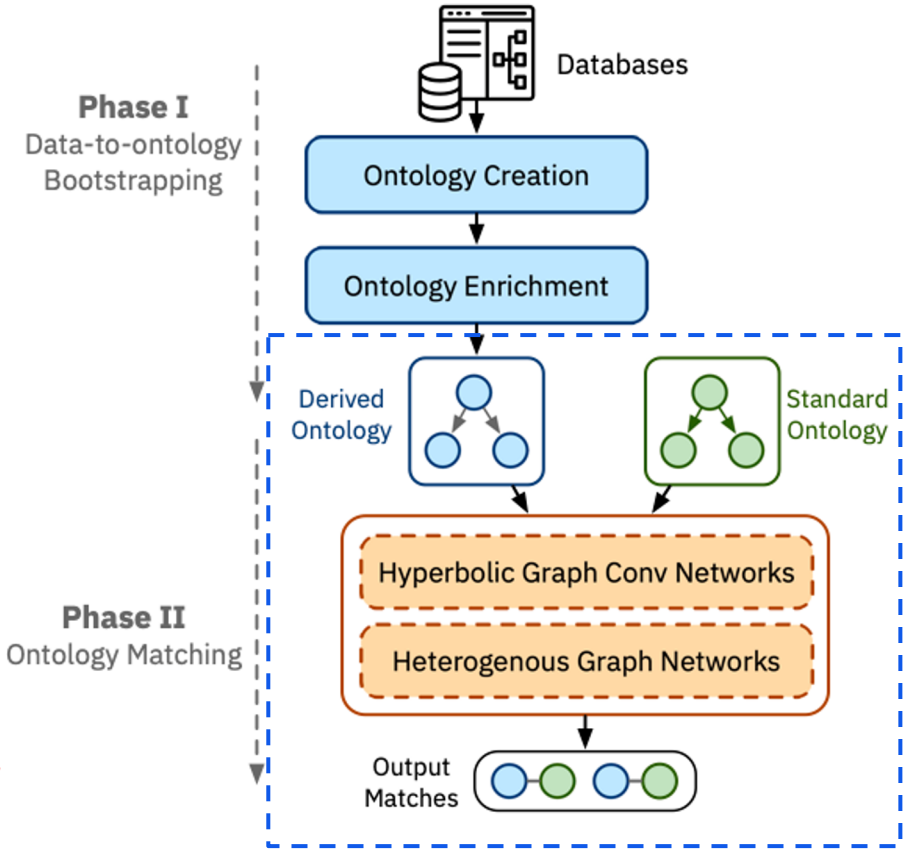
- **Goal:** Derive an ontology from a medical dataset stored in a relational database
- **Steps:** (1) Ontology creation; (2) Ontology Enrichment
- **Creation step:**
  - Concepts: Create a concept for each table with its representative columns as data properties
  - Relations: Add a relation between two concepts based on primary key-foreign key relationships between tables
- **Enrichment step:**
  - Concept augmentation: Add instance-level concepts (entries in table) to the created ontology, if instance-level concepts have their matches in the standard ontology
  - Neighborhood augmentation: Populating edges from standard ontology via pre-aligned seed concepts



Phase II: Ontology Matching

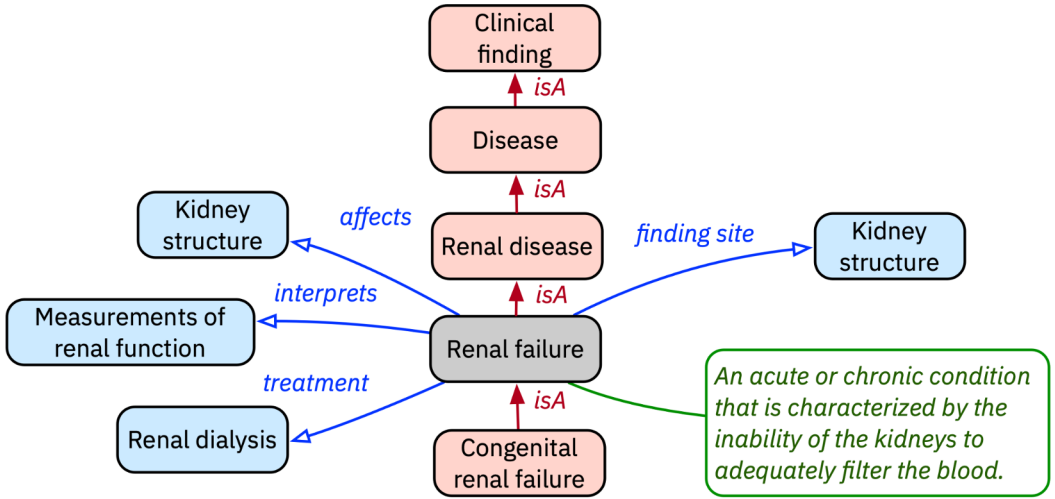
Goal: *How to match the derived ontologies and standard medical ontologies?*

Solution: *Hybrid graph neural network for ontology matching*



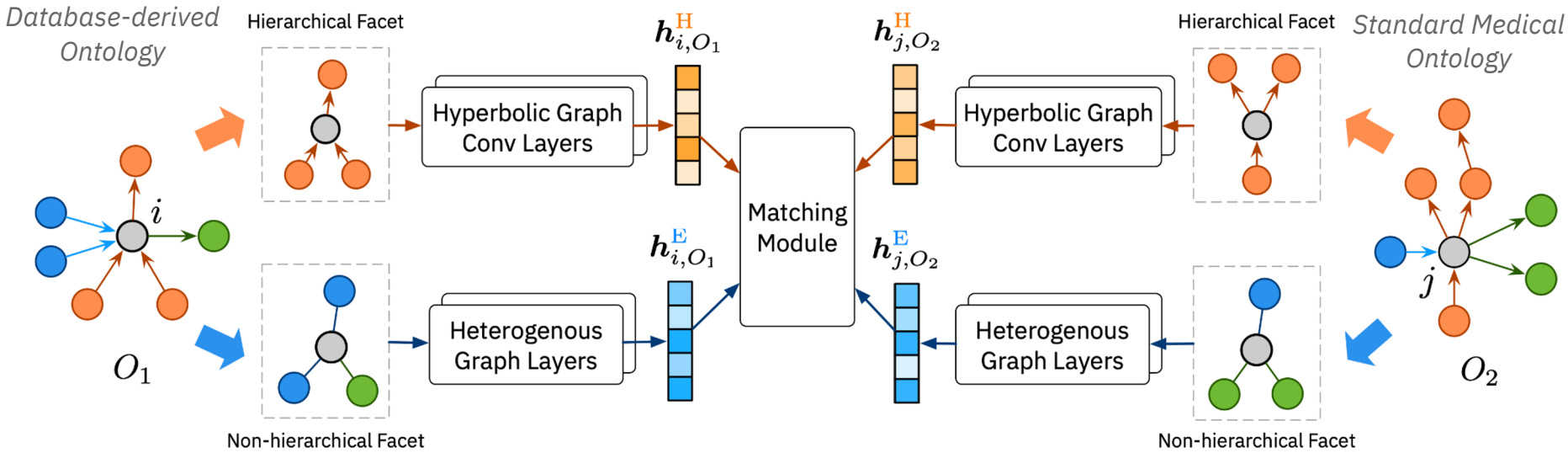
Matching Ontologies: Challenges

- After data have been transformed to one ontology, the next step is to match such ontology O_1 to high-quality standard ontologies O_2 .
- Challenge: Learn comprehensive representations from the *descriptive text features*, *hierarchical taxonomy features* (normally defined in "*Is-A*" relation)and *semantic relational facts* between concepts in the ontology, which are important to identify the match between two ontologies.



Ontology Matching

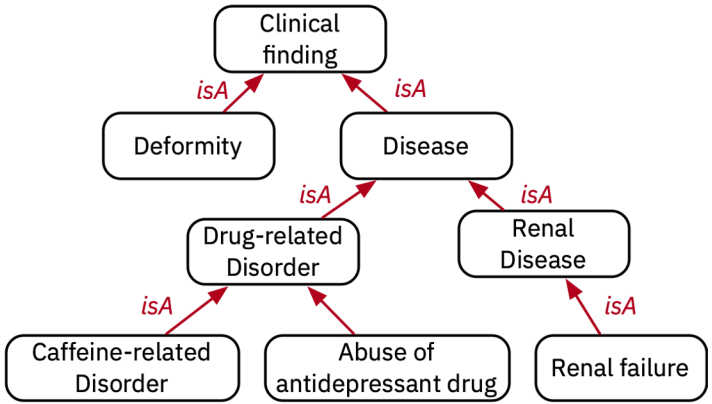
- Our solution: Hyperbolic Graph Convolution Module + Heterogeneous Graph Module
  - Focus on ontology hierarchical structures and relational structures



Architecture of ontology matching between two given ontologies

Hyperbolic Graph Convolution Layer

- Goal: Better capture concept hierarchies in medical ontologies by embeddings in the hyperbolic space
- Adopted from Hyperbolic Graph Convolutional Neural Network (HGCN) [1]



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

$$w_{ij} = \mathrm{SOFTMAX}\left(\mathrm{MLP}\left(\log_{\mathrm{o}}^K\left(\mathbf{h}_i^H\right) \,||\, \log_{\mathrm{o}}^K\left(\mathbf{h}_j^H\right)\right)\right)$$

$$\mathrm{AGG}^K(\mathbf{h}^H)_i = \exp_{\mathbf{h}_i^H}^K\left(\sum_{j\in\mathcal{N}(i)} w_{ij}\log_{\mathbf{h}_i^H}^K\left(\mathbf{h}_j^H\right)\right)$$

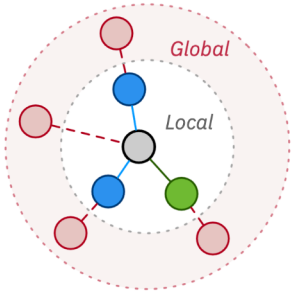
$$\mathbf{h}_i^{l,H} = \sigma^{\oplus^{K_{l-1},K_l}}\left(\mathrm{AGG}^{K_{l-1}}\left(\mathbf{h}^{l,H}\right)_i\right)$$

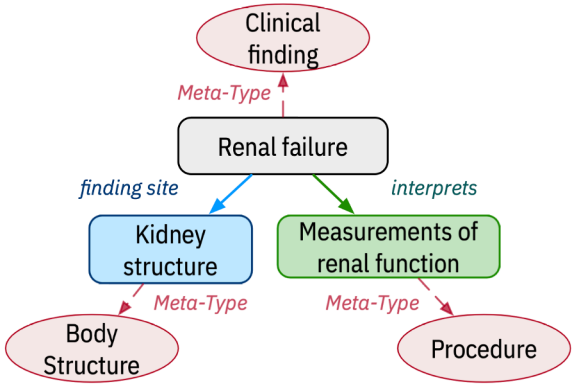
$$\mathcal{L}^{\mathcal{H}} = p((c_i,c_j)\in\mathcal{C}) = \left\{\exp\left[\frac{1}{t}\left(d^K\left(\mathbf{h}_i^H,\mathbf{h}_j^H\right)^2 - r\right)\right] + 1\right\}^{-1}$$

[1] I. Chami, Z. Ying, C. Ré, and J. Leskovec. Hyperbolic graph convolutional neural networks. In NeurIPS, pages 4869–4880, 2019.

Heterogeneous Graph Layer

- Goal: Model the multi-relational non-hierarchical relationships in the ontologies
- Enhance R-GCN [2] by using neighbor's top-level ancestor concepts (meta-type) in the ontology (e.g., *"kidney" → "body structure"*) as "global features"
- Both local and global context information are encoded by neighborhood aggregation





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

$\mathbf{h}_i^{l-1,E} \| \mathbf{g}_i^{l-1,E}$ $\qquad$ $\mathbf{h}_j^{l-1,E} \| \mathbf{g}_j^{l-1,E}$

$$\mathcal{L}^{\mathcal{E}} = \sum_{r\in\mathcal{R}}\sum_{i,j\in\mathcal{C}} w_{ij}^r \log \frac{\exp\left((\mathbf{h}_i^E)^\top A_r \mathbf{h}_j^E\right)}{\sum_{i'\in\mathcal{C}} \exp\left((\mathbf{h}_{i'}^E)^\top A_r \mathbf{h}_j^E\right)}$$

[2] M. Schlichtkrull, T. N. Kipf, P. Bloem, et. al. Modeling Relational Data with Graph Convolutional Networks. CoRR abs/1703.06103, 2017.

Matching and Training

- The final matching module takes pairs of concept embeddings and outputs prediction score, implemented by MLP (or Transformer encoder).

$$M\left(\mathbf{h}_i^U, \mathbf{h}_j^U\right) = \sigma\left(\mathbf{W}_2 \cdot \gamma\left(\mathbf{W}_1\left(\mathbf{h}_i^U || \mathbf{h}_j^U\right) + \mathbf{b}_1\right) + \mathbf{b}_2\right)$$

- Contrastive matching loss function

$$\mathcal{L}^M = \sum_{(i,j)\in\mathcal{M}^+} M(\mathbf{h}_i, \mathbf{h}_j) + \sum_{(i',j')\in\mathcal{M}^-} \omega\left[\lambda - M\left(\mathbf{h}_{i'}, \mathbf{h}_{j'}\right)\right]_+$$

- Joint training of all modules

$$\mathcal{L} = \mathcal{L}^M + \alpha_1 \cdot \left(\mathcal{L}^{\mathcal{H}}_{\mathcal{O}_1} + \mathcal{L}^{\mathcal{H}}_{\mathcal{O}_2}\right) + \alpha_2 \cdot \left(\mathcal{L}^{\mathcal{E}}_{\mathcal{O}_1} + \mathcal{L}^{\mathcal{E}}_{\mathcal{O}_2}\right)$$

Matching Loss

Loss of Hyperbolic GCN

Loss of Heterogeneous GNN

Outline

UCLA IBM

- Motivation: Data-to-Ontology Matching
- MEDTO System Architecture Overview
- Model: Ontology Bootstrapping and Matching

Experiments & Case Study

- Summary & Future Directions

Datasets

- Two medical databases: MIMIC-III [3] and MDX (IBM Micromedex)
  - **MIMIC-III:** Anonymized health-related record of 4000+ patients and their stays in ICU, including 21 tables on patient tracking, ICU data and hospitalization procedure.
  - **MDX:** A medical database of IBM Micromedex that contains 59 tables on drugs, adverse effects, indications, findings, etc.
- Three standard medical ontologies provided in OAEI Large BioMed Track. Stats:
  - **FMA** [4]: Declarative knowledge of human anatomy. → 78.9k concepts and "is-A" relations.
  - **NCI** [5]: Terminologies for clinical care and other basic research. → 56.9k concepts, 85.3k relations of 80 types (59.7k are "is-A").
  - **SNOMED CT** [6]: A collection of medical terms providing synonyms and definitions used in clinical reporting. → 76.7k concepts, 109.9k relations of 5 types (105.6k are "is-A")

[3] A. E. Johnson, T. J. Pollard, L. Shen, et al. Mimic-iii, a freely accessible critical care database. Scientific data, 3:160035, 2016.
[4] C. Rosse and J. L. V. M. Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. J. Biomed. Informatics, 36(6):478–500, 2003.
[5] S. de Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, and L. W. Wright. NCI thesaurus: Using science-based terminology to integrate cancer research results. In MEDINFO, volume 107, pages 33–37, 2004.
[6] K. Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. In Stud Health Technol Inform, volume 121, pages 279–290, 2006.

Data-to-Ontology Matching

UCLA IBM

- Medical databases: MIMIC-III and MDX
- Baselines: AML[7], LogMap[8], RDGCN[9] (SOTA from OpenEA)

Table: Matching MIMIC-III and MDX to SNOMED CT

Dataset	MIMIC-III ⇔ SNOMED		MDX ⇔ SNOMED	
Metric	Hits@10	Hits@30	Hits@10	Hits@30
AML	0.06 (1/15)	0.13 (2/15)	0.16 (3/19)	0.26 (5/19)
LogMap	0.20 (3/15)	0.20 (3/15)	0.21 (4/19)	0.37 (7/19)
MTransE	0.00 (0/15)	0.00 (0/15)	0.05 (1/19)	0.05 (1/19)
GCN-Align	0.20 (3/15)	0.33 (5/15)	0.32 (6/19)	0.42 (1/19)
RDGCN	0.27 (4/15)	0.40 (6/15)	0.32 (6/19)	0.58 (11/19)
MEDTO	**0.47 (7/15)**	**0.60 (9/15)**	**0.42 (8/19)**	**0.79 (15/19)**

*Significant performance improvement (**>50%** on MIMIC-III & **>25%** on MDX) compared to all baselines.*

[7] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto. The agreementmakerlight ontology matching system. In OTM, pages 527–541, 2013.
[8] E. Jiménez-Ruiz and B. C. Grau. Logmap: Logic-based and scalable ontology matching. In ISWC, pages 273–288, 2011.
[9] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, and D. Zhao. Relation-aware entity alignment for heterogeneous knowledge graphs. In IJCAI, pages 5278–5284, 2019.

Ontology-to-Ontology Matching

- Datasets: FMA, NCI and SNOMED from OAEI Challenge 2020 (all are standard medical ontologies) → *evaluate the ontology matching component of MEDTO*
- Baseline: Rule-based matchers (AML, LogMap), GNN-based KG entity alignment (OpenEA benchmark: MTransE, GCN-Align, RDGCN, etc.)

Better results over KG alignment and comparative performance over well-developed AML/LogMap

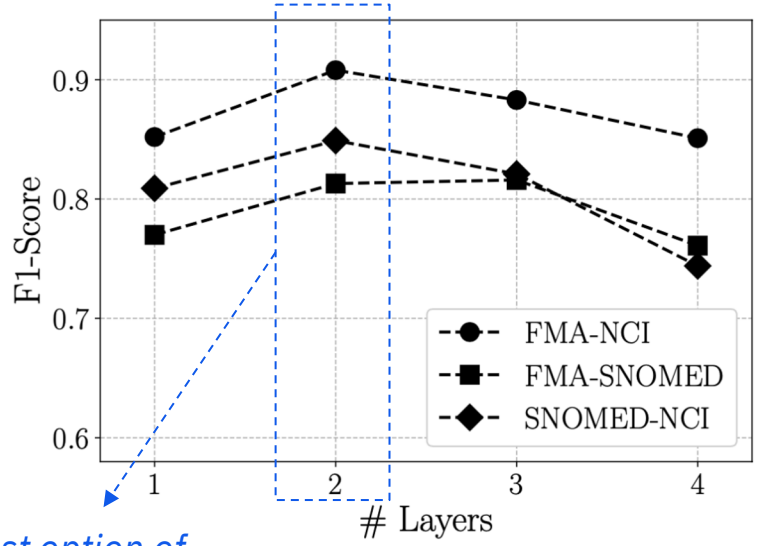
Table: Ontology matching on OAEI dataset

Datasets	FMA-NCI				FMA-SNOMED				NCI-SNOMED			
Metrics	P	R	F1	MRR	P	R	F1	MRR	P	R	F1	MRR
AML	0.942	**0.899**	**0.920**	–	**0.902**	0.729	0.806	–	0.890	0.744	0.810	–
LogMap	0.916	0.895	0.905	–	0.791	**0.850**	**0.819**	–	0.897	0.732	0.805	–
MTransE	0.627	0.640	0.633	0.416	0.505	0.475	0.490	0.372	0.254	0.378	0.304	0.349
GCN-Align	0.813	0.783	0.798	0.561	0.763	0.729	0.746	0.526	0.745	0.775	0.760	0.467
RDGCN	0.855	0.843	0.849	0.761	0.824	0.752	0.786	0.683	0.852	0.782	0.816	0.679
MEDTO	**0.944**	0.874	0.908	**0.783**	0.871	0.762	0.813	**0.690**	**0.901**	0.802	**0.849**	**0.704**
MEDTO (w/o HYP)	0.867	0.775	0.818	0.724	0.787	0.653	0.714	0.540	0.835	0.759	0.795	0.595
MEDTO (w/o HET)	0.927	0.851	0.887	0.763	0.863	0.747	0.801	0.676	0.881	**0.807**	0.842	0.688

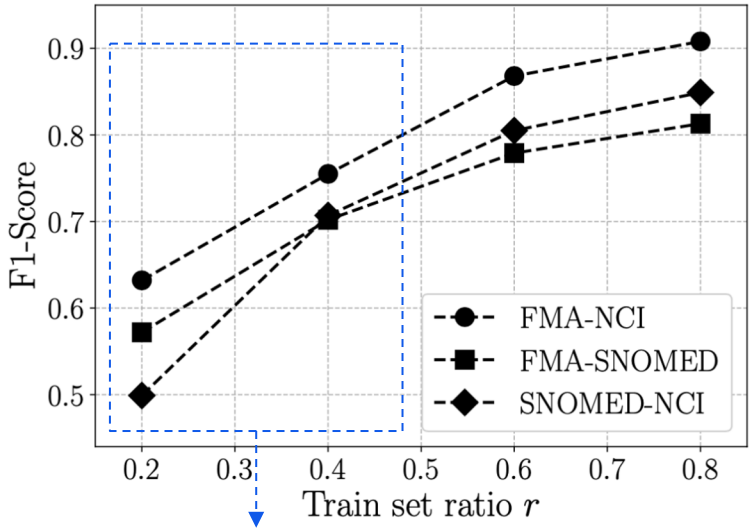
Both hyperbolic graph layers and heterogeneous graph layers contributes to the performance gain.

Hyperparameter Study

- Hyperparameters: (1) number of GNN layers in MEDTO matching; (2) training ratio of seed matches.



The best option of number of GNN layers in MEDTO is 2.



MEDTO can still perform fairly well when using a small set of train data (i.e., seed matches).

Case Study: MIMIC-III

- MEDTO finds more matches over MIMIC-III Tables compared to AML/LogMap.
- Ambiguous terms are challenging
  - Example: "***outputevents***", which specifically refers ***fluid output*** in most cases, which is captured by MEDTO. However, it mismatches with ***process output*** or ***output measurement*** in other models.
- MEDTO may sometimes fails
  - Lack of instance-level concepts during ontology bootstrapping
  - Sets of introduced instance-level concepts do not correctly reflect the content of table.

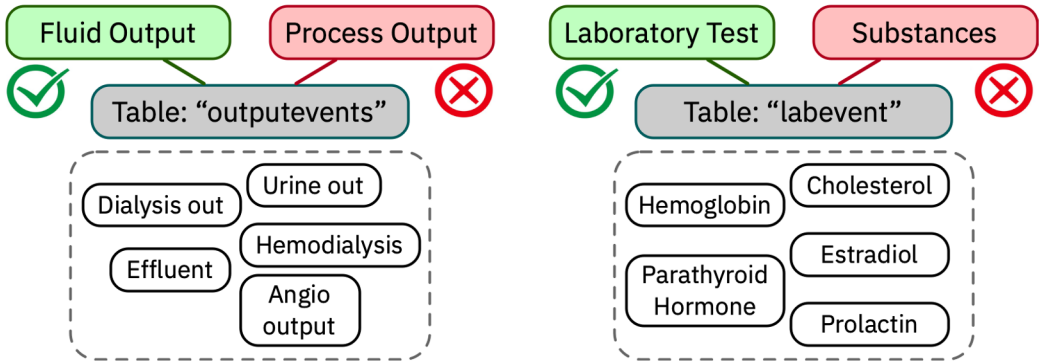


Table: Examples of MIMIC-III and MDX matching results

MIMIC-III Tables	AML	LogMap	RDGCN	MEDTO
patient	✓	✓	✓	✓
prescriptions	✓	✗	✓	✓
caregivers	✗	✓	✗	✓
services	✗	✓	✓	✓
outputevents	✗	✗	✓	✓
icustays	✗	✗	✗	✓
chartevents	✗	✗	✗	✗
labevents	✗	✗	✗	✗

MDX Tables	AML	LogMap	RDGCN	MEDTO
AdverseEffect	✓	✓	✓	✓
Dosage	✓	✓	✓	✓
DrugFoodInteraction	✗	✓	✓	✓
ContraIndication	✗	✗	✓	✓
DoseAdjustment	✗	✗	✗	✓
DrugRoute	✗	✗	✗	✗

Outline

UCLA IBM

- Motivation: Data-to-Ontology Matching
- MEDTO System Architecture Overview
- Model: Ontology Bootstrapping and Matching
- Experiments & Case Study

Summary & Future Directions

Summary & Future Directions

- **Summary**
  - End-to-end framework MEDTO for medical data to ontology matching
  - MEDTO creates a semantically enriched ontology from a given medical database and matches the derived ontology to standard ontologies
  - GNN-based ontology matching module capturing two facets of an ontology
  - Effectiveness shown on real-world medical databases
- **Future Directions**
  - Support more relations in an ontology (e.g., disjoint, equivalence statements, etc.)
  - Learn representations with ontological constraints applied to improve match predictions

Thank you!

Contact: jhao@cs.ucla.edu, chuan.lei@ibm.com

Scan the QR-code for more paper details!