Selecting Top-*k* Data Science Models by Example Dataset

Mengying Wang[+], Sheng Guan[+], Hanchao Ma[+], Yiyang Bian[+], Haolai Che[+], Abhishek Daundkar[*], Alp Sehirlioglu[*], Yinghui Wu[+]

[+]Department of Computer and Data Science, [*]Department of Materials Science and Engineering, CWRU

CIKM '23, October 21–25, 2023, Birmingham, UK

CWRU

Birmingham 2023 CIKM

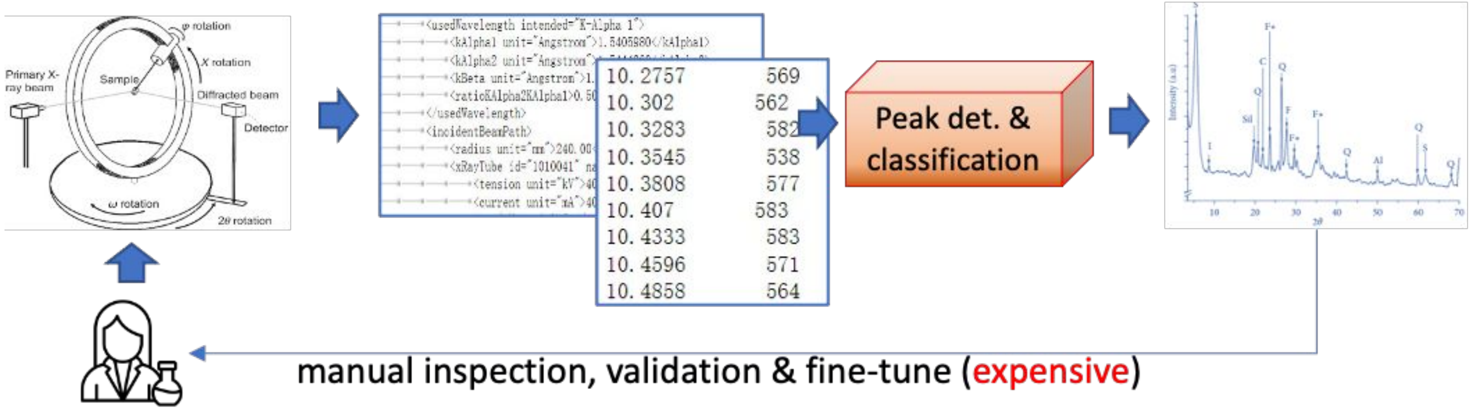
CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

Roadmap

- **Introduction**
  - Motivation
  - Model Selection Problem
- **ModsNet - Knowledge Graph-Based Model Search**
  - Extraction Module - Construct the Model-data Interaction Graph
  - Selection Module - Probe-and-Select Strategy
- **Prototype System**
- **Experiment**
  - Experiment Settings
  - Experiment Results
- **Conclusion & Future Work**

Motivation

In-Situ XRD analytics with ML models (e.g., regression)



🤩 Pre-trained models are invaluable resources:

- Machine Learning Models.
- Statistical/data analysis scripts.

⇨

🧐 How to make these models discoverable?

🤔 Search models by a 'query' dataset?"

Model Selection Problem

Problem:

Given a collection of models and associated metadata, recommend models with potentially high performance for a 'query' dataset.

- **Input:** a set of datasets and $\mathcal{D}$, pre-trained models $\mathcal{M}$, a (limited) amount of historical performance $\mathcal{H}$, a model performance measure $\boldsymbol{P}$, integer $\boldsymbol{k}$, and an example dataset d_q(a "query");
- **Output:** a set of $\boldsymbol{k}$ pre-trained models from $\mathcal{M}$ with expected good performance $\boldsymbol{P}$ over d_q.

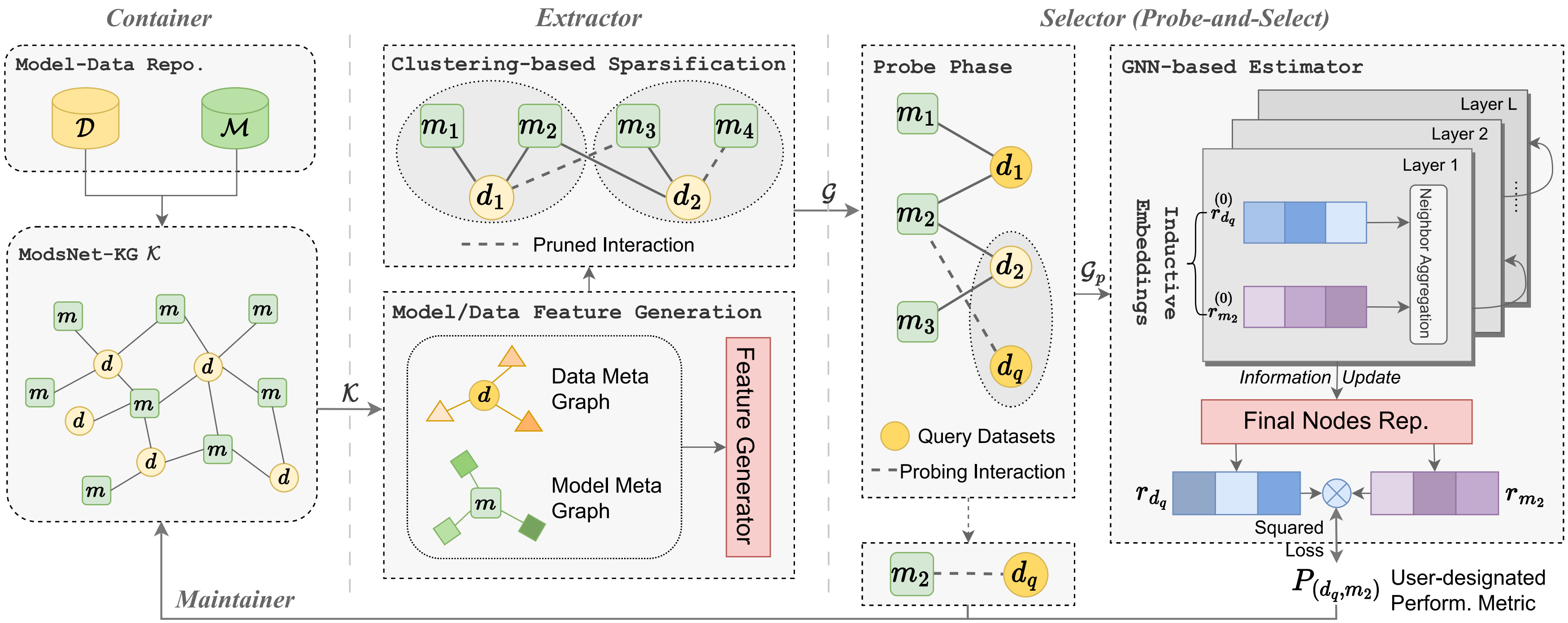
Challenges:

1. Modeling and incorporating knowledge. → Knowledge-enhanced.
2. Make recommendations for a new dataset without history records. → Probe-and-select.

Roadmap

- **Introduction**
  - Motivation
  - Model Selection Problem
- **ModsNet - Knowledge Graph-Based Model Search**
  - Extraction Module - Construct the Model-data Interaction Graph
  - Selection Module - Probe-and-Select Strategy
- **Prototype System**
- **Experiment**
  - Experiment Settings
  - Experiment Results
- **Conclusion & Future Work**

ModsNet - Knowledge Graph-Based Model Search



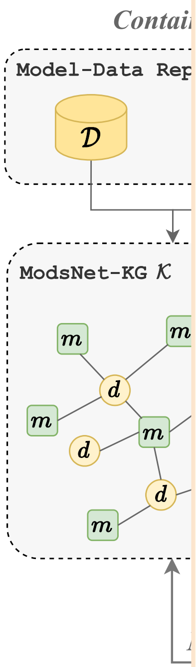
ModsNet - Knowledge Graph-Based Model Search

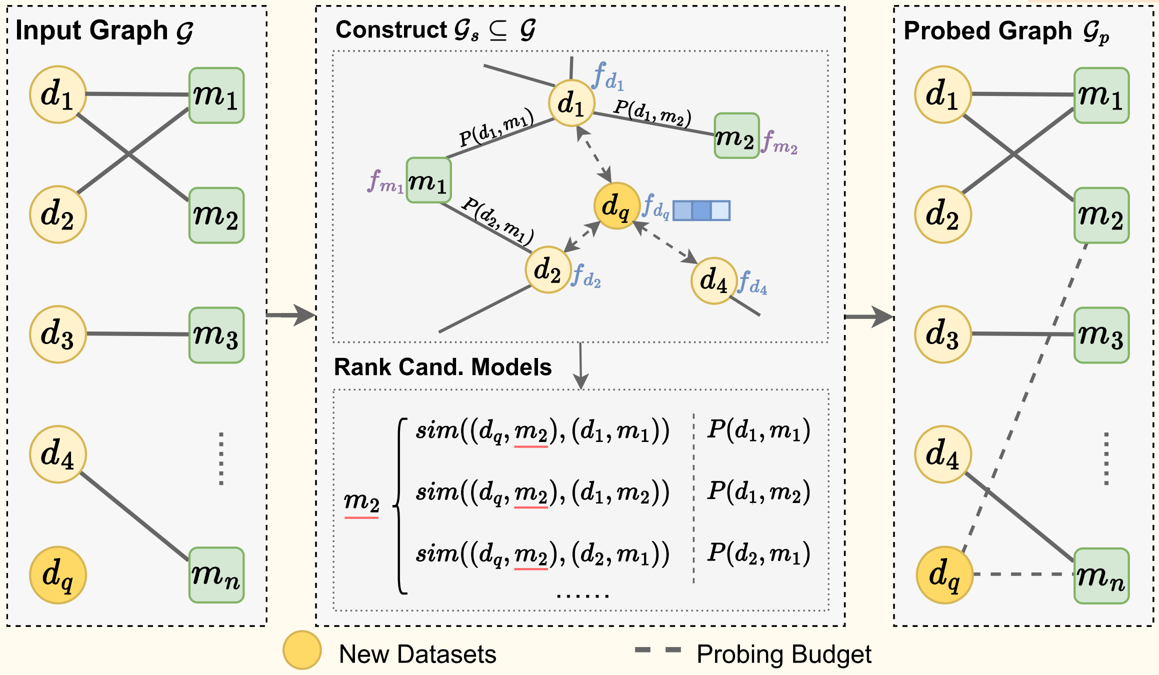
$$e_m = \frac{1}{(|\mathcal{D}_m|+1)^{\alpha}}\left(\sum_{d\in\mathcal{D}_m} e_d + e_{m_t}\right)$$

$$e_{d_q} = \frac{1}{(|\mathcal{M}_{d_q}|+1)^{\alpha}}\left(\sum_{m\in\mathcal{M}_{d_q}} e_m + e_{d_t}\right)$$

Selector

Model:





PU&GPU, o.s.), time

' features, model

e network

$m(\mathcal{F}(d), \mathcal{F}(d')))$

nsity

$$Score(m) = \frac{1}{|\mathcal{I}_s|}\sum_{e\in\mathcal{I}_s} lsim$$

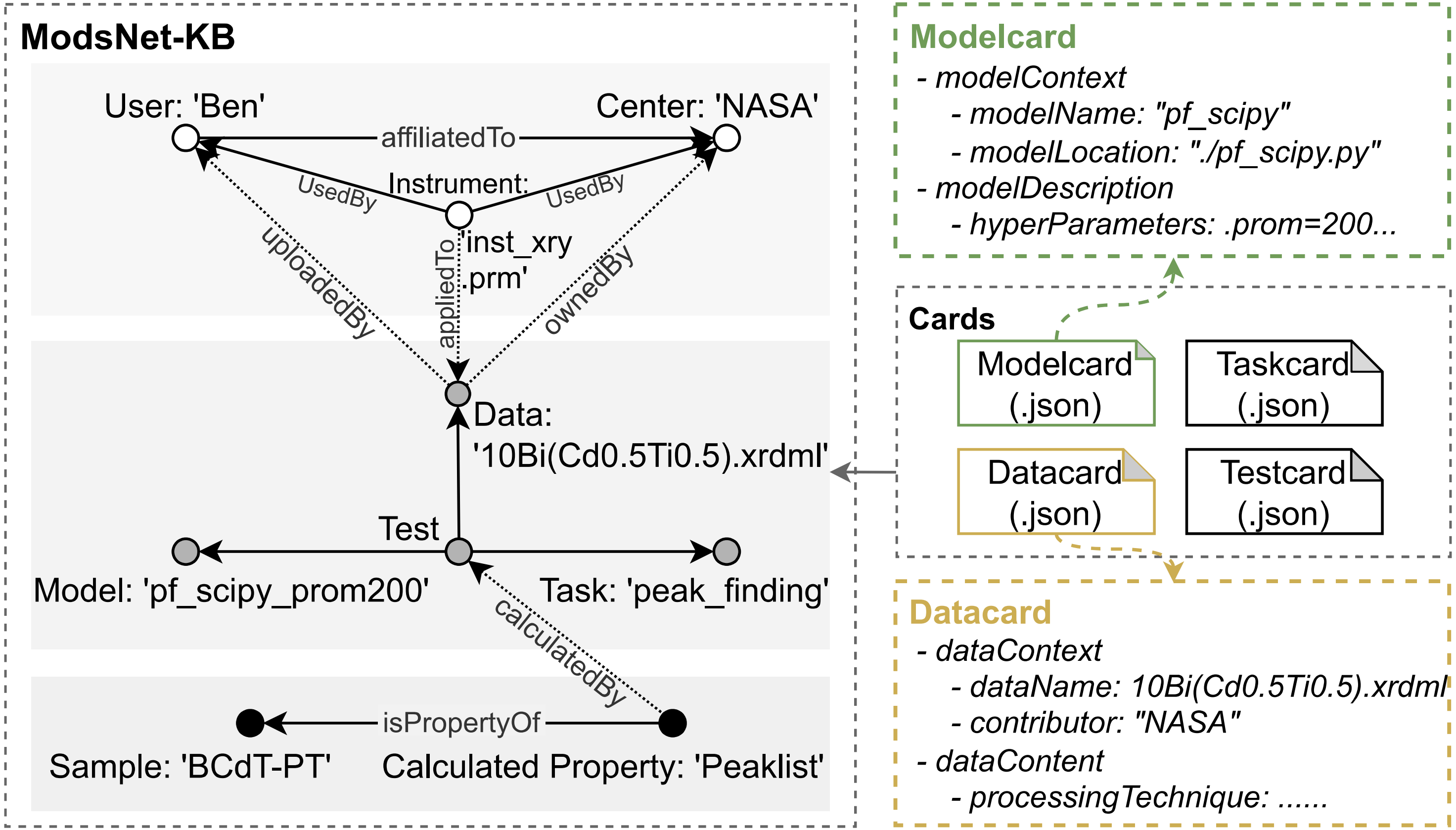
$$\mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{I}_{\text{train}}|}\sum_{(d_i,m_j)\in\mathcal{I}_{\text{train}}}\left(\hat{P}_{(d_i,m_j)} - P_{(d_i,m_j)}\right)^2 + \lambda\|\Theta\|_2^2$$

similarity, Jaccard similarity, har

Roadmap

- **Introduction**
  - Motivation
  - Model Selection Problem
- **ModsNet - Knowledge Graph-Based Model Search**
  - Extraction Module - Construct the Model-data Interaction Graph
  - Selection Module - Probe-and-Select Strategy
- **Prototype System**
- **Experiment**
  - Experiment Settings
  - Experiment Results
- **Conclusion & Future Work**

Prototype System



[1] *CRUX: Crowdsourced Materials Science Resource and Workflow Exploration. CIKM 22', Demo Track, Wang et al.*

Roadmap

- **Introduction**
  - Motivation
  - Model Selection Problem
- **ModsNet - Knowledge Graph-Based Model Search**
  - Extraction Module - Construct the Model-data Interaction Graph
  - Selection Module - Probe-and-Select Strategy
- **Prototype System**
- **Experiment**
  - Experiment Settings
  - Experiment Results
- **Conclusion & Future Work**

Experiment Settings - Datasets

Dataset	# Models	# Datasets	# Interactions	# Features	Density	Task
PKZoo	462	289	98257	21	0.73591	Peak Finding
KIZoo	1800	72	9304	41	0.07179	Image Classification
HFZoo	932	66	974	13	0.01583	Text Classification

PKZoo:

- Peak-finding models, XRD datasets.
- Crowdsourced from material science community, keep growing.
- Supported by material science experts.

KIZoo:

- Image datasets are collected from Kaggle.
- Self-curated, over 1,000 GPU hours, various CNN architectures.
- Recorded detailed training and testing information.

HFZoo:

- Text classifiers, text datasets.
- Crowdsourced from a fast-growing AI community.

Experiment Settings - Model Selection Methods

- **ModsNet and its three variants:**
  - ModsNet-C: optimized with clustering-based sparsification.
  - ModsNet-NoKG: operates without a knowledge graph.
  - ModsNet-RProb: utilizes random probes without filtering.
- **GNN-based methods:**
  - LightGCN
  - IDCF-GCN
  - INMO-GCN

  } Cope with the “cold-start” scenario by appending probe strategy of ModsNet.
- **CF-based methods:**
  - Collaborative Filtering: Cope with the “cold-start” scenario by dataset similarity.
  - Matchbox
- **Supervised learning methods:**
  - Linear Regression
  - Wide & Deep

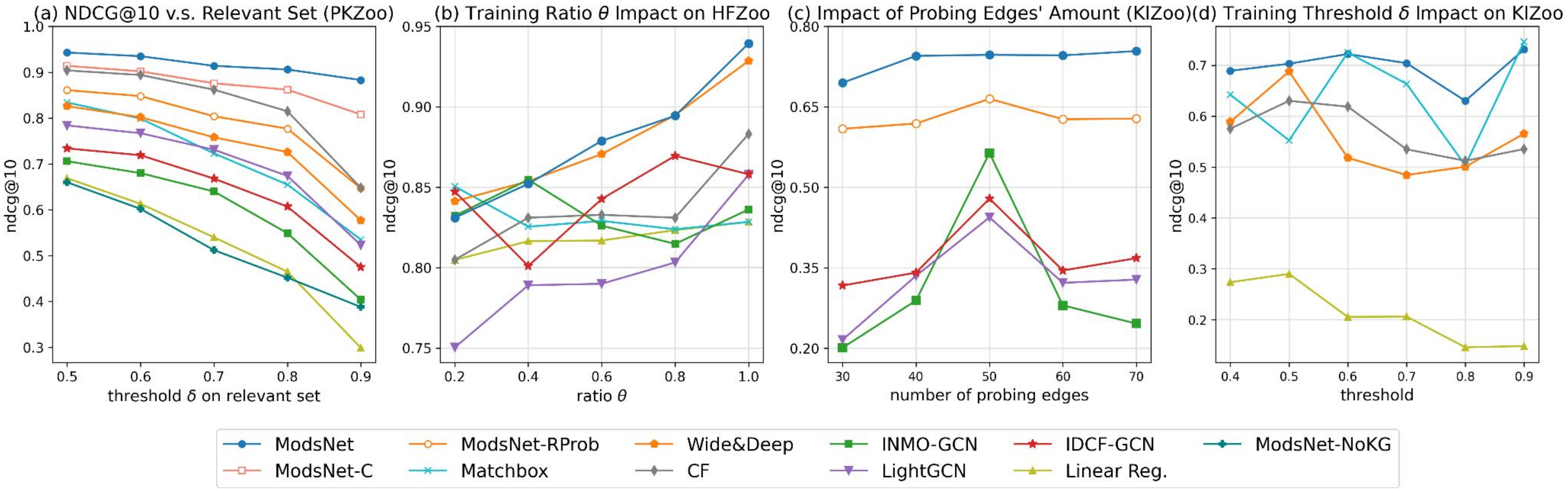
Experiment Results (Exp-1) - Effectiveness

Recommendation results over PKZoo:

metrics	Precision@5	Precision@10	Recall@5	Recall@10	NDCG@5	NDCG@10
ModsNet	**0.938**	**0.874**	**0.118**	<u>0.201</u>	**0.95**	**0.906**
ModsNet-C	<u>0.887</u>	<u>0.841</u>	0.108	**0.208**	<u>0.888</u>	<u>0.862</u>
ModsNet-RProb	0.867	0.733	<u>0.112</u>	0.186	0.859	0.777
ModsNet-NoKG	0.313	0.507	0.07	0.147	0.31	0.452
CF	0.882	0.797	0.092	0.168	0.875	0.815
Wide & Deep	0.79	0.687	0.084	0.14	0.808	0.726
lightGCN	0.759	0.654	0.07	0.122	0.749	0.674
Matchbox	0.641	0.61	0.093	0.162	0.701	0.655
IDCF-GCN	0.677	0.574	0.077	0.135	0.679	0.607
INMO-GCN	0.615	0.549	0.068	0.11	0.592	0.549
LinearRegression	0.528	0.482	0.033	0.058	0.484	0.465

- **ModsNet**: outperforms all methods.
- **ModsNet-C**: comparable with **ModsNet,** with 22.85% interactions pruned, speeded up 32.29%.
- Obvious gap between **ModsNet** and **ModsNet-RProb/ModsNet-NoKG**, with increases of 16% and 100.44% in NDCG@10, respectively.

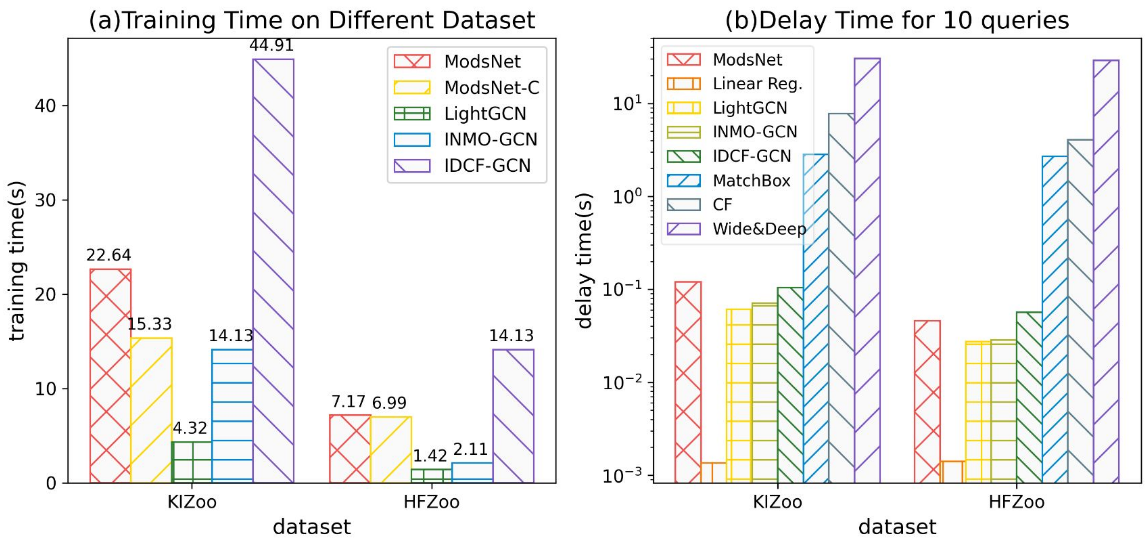
Experiment Results (Exp-2) - Impact of Factors



ModsNet performs stably in various settings:

- Fig(a) - varying the performance threshold δ on relevant set from 0.5 to 0.9.
- Fig(b) - varying interaction ratio θ in training set from 20% to 100%.
- Fig(c) - varying number of probe edges from 30 to 70.
- Fig(d) - varying the performance threshold δ on training set from 0.4 to 0.9.

Experiment Results (Exp-3) - Efficiency



- Fig(a) - **ModsNet-C** reduced the training time while keeping a relatively good performance.
- Fig(b) - **ModsNet** has proven to be significantly more efficient than other methods that have achieved comparable performance results, such as Wide & Deep, CF, and Matchbox.

Experiment Results (Exp-4) - Case Study

Query 1:
I have a dataset "tolgadincer/labeled-chest-xray-images", which model should I choose for classifying pneumonia? (k=1)

Response 1:
The selected models with estimated balanced accuracy

- Groundtruth{id: 1190, b_accuracy: 0.958}
- ModsNet-C{id: 1175, b_acccuracy: 0.925}
- LinearReg.{id: 1544, b_accuracy: 0.601}

Prediction Result 1:
Prediction result by selected models for the example image in the input dataset

- Groudtruth{pos: 0.9953, neg: 0.0047}
- ModsNet-C{pos: 0.9527, neg: 0.0473}
- LinearReg.{pos: 0.9099, neg: 0.0901}

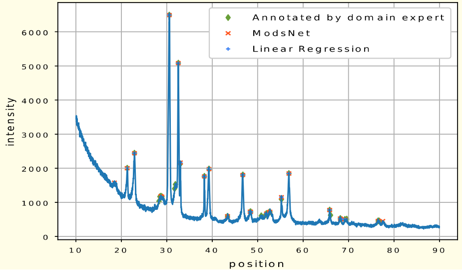
Query 2:
I have a XRDML file, which model should I choose for peak finding? (k=1)

```
<dataPoints>
  <positions axis="2Theta" unit="deg">
    <startPosition>10.000</startPosition>
    <endPosition>90.005</endPosition>
  </positions>
  <commonCountingTime unit="seconds">46.25<
  <intensities unit="counts">3464.000 3461.
</dataPoints>
```

Response 2:
The selected models with estimated f1_score

- Groundtruth{id: 311, f1_score: 0.85714}
- ModsNet{id: 291, f1_score: 0.80702}
- LinearReg.{id: 476, f1_score: 0.69388}

Prediction Result 2:
Visualization for results by selected models.



Conclusion & Future Work

- Investigated the problem of model selection given an example dataset.
- Proposed ***ModsNet***, supported by a prototype system:
  - A Knowledge Graph-Based framework.
  - Equipped with an inductive GNN-based regression model.
  - Optimized by a clustering-based sparsification strategy.
- Verified ModsNet's effectiveness and efficiency by three real-world datasets.

- Extend ModsNet for more domain-specific applications.
- Incorporate LLM to improve its explanbility.

crux-project.github.io

THANK YOU !

Email: mxw767@case.edu

Acknowledgment: This work is supported by NSF under
CNS-1932574, ECCS-1933279, CNS-2028748 and OAC-2104007.

Birmingham 2023 CIKM

Collected Features

Model:

Metadata: contributor, licenses, languages, task

Source code structure: AST topological features

Training Record: training dataset, base model, environment (CPU&GPU, o.s.), time cost, training performance

Model Info: model type, # parameters, hyperparameters, layers' features, model size, flops, inference time per step(CPU/GPU), topological depth of the network

Data:

Metadata: contributor, licenses, languages, organization, material sample, equipment, experiment settings: temperature, pressure, statistics of angles (2θ), intensity ranges

Activity: usability rating, hotness (#views, #votes, #downloads)

Statistics: # classes, size categories

Description: tasks/classes, textual descriptions

Interaction:

Model-data Pair: model id, dataset id

Evaluation Record: environment(GPU), testing cost

Metrics: accuracy, balanced accuracy, AUC, f1_score, precision, recall, Cosine similarity, Jaccard similarity, hamming loss, log loss

Comparison of Potential Methods

Approach	Method	External KG	Cold Start	Learning Cost	Query Time	Performance
KG-Based, Regression	Our Method	Yes	Yes	Low	Low	Always excellent
Supervised Learning Regression	Linear Regression	Yes	Yes	Low	Low	Not accurate enough
	Wide & Deep	Yes	Yes	High	High	Relatively excellent
Collaborative Filtering Regression	CF	No	No	Medium	Medium	Great for dense graphs, not for sparse
	Matchbox	Yes	Yes	High	High	Less sensitive than CF, relatively good
Graph Neural Network Link Prediction	LightGCN	No	No	Low	Low	Relatively good
	IDCF-GCN	No	No	Medium	Medium	Relatively good, inductive setting
	INMO-GCN	No	No	Low	Low	Relatively good, inductive setting

* This table outlines the initial methods. To ensure a fair comparison, baselines in the experimental study are adapted versions.