Towards Foundation Models for Graph Reasoning & AI4Science

Michael Galkin

Hesham Mostafa

Santiago Miret

Oct 11th 2023

👋 Hello

- 2019: PhD at the University of Bonn (Germany) in CS focusing on graph algorithms and KG / NLP applications
- 2020-2022: Postdoc at Mila (Montreal) Graph ML 👉 **all the way** 👈
- 2023 - now: Research Scientist @ Intel AI
- Sometimes I write about graphs:
  - @graphml in Telegram
  - @mgalkin on Medium

4 Michaels @ ICML'23

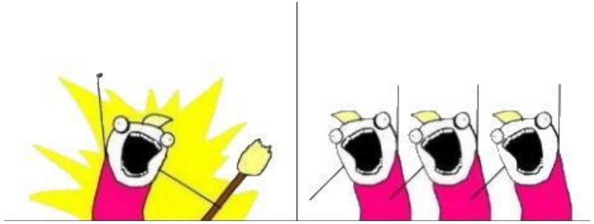
Foundation Models

A **single** model pre-trained (often) in the self-supervised fashion on **large amounts of data** that is applicable to **many downstream tasks**

- By in-context learning
- By fine-tuning

We Want Graph Foundation Models!

- ... Large!
  - Non strong signal that GNNs or Graph Transformers benefit from depth / increasing # params
  - Scaling laws for GNNs / GTs are non-existent
- ... Self-supervised pre-training!
  - No unified task
  - Limited signal that pre-training helps
- ... Uniform featurizing and Multi-modal!
  - But different 2D / 3D graphs, periodic structures, geometry



Foundation Models at Intel AI

intel labs

Knowledge Graph Reasoning

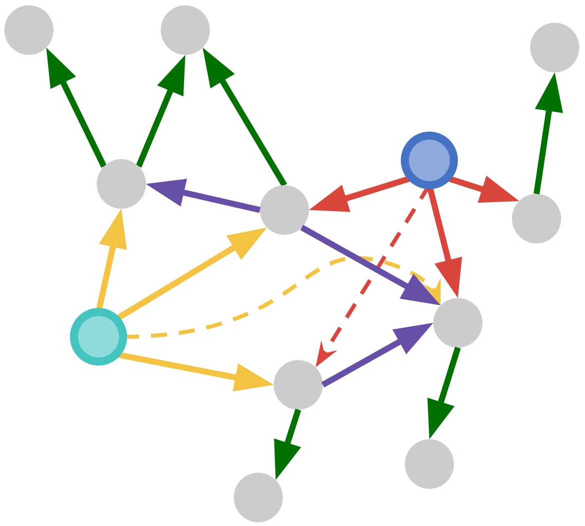
- At large-scale
- Inference on any domain
- All graph-level tasks (start from link prediction)

AI 4 Science

- Molecules, proteins, materials (crystals)
- Materials generation, eg, new catalysts

Foundation models: Graph Reasoning

- Simple link prediction
- Complex logical query answering
- … and beyond



Knowledge Graphs

Multi-relational graphs with **(subject, predicate, object)** triples.

Multi-domain graphs:

- **Encyclopedias** (Wikidata, Freebase)

  In search and retrieval-augmented LLMs

London (Google)

About

London, the capital of England and the United Kingdom, is a 21st-century city with history stretching back to Roman times. At its centre stand the imposing Houses of Parliament, the iconic 'Big Ben' clock tower and Westminster Abbey, site of British monarch coronations. Across the Thames River, the London Eye observation wheel provides panoramic views of the South Bank cultural complex, and the entire city. — Google

Weather: 57°F (14°C), Wind W at 7 mph (11 km/h), 78% Humidity More on weather.com

Local time: Thursday 7:29AM

Neighborhoods: Elephant and Castle, Chiswick, Brent Cross, MORE

Elevation: 36 ft (11 m)

Local government districts: 32 London boroughs; and the City of London

Region: London (Greater London)

Settled by Romans: AD 47; 1976 years ago; as Londinium

Feedback

London (Bing)

London is the capital and largest city of England and the United Kingdom, with a population of around 8.8 million. It stands on the River Thames in south-east England at the head of a 50-mile es… +
Wikipedia

gov.uk

Country England
Region London (Greater London)
Elevation 36 ft (11 m)
Sovereign state United Kingdom
See more

Knowledge Graphs

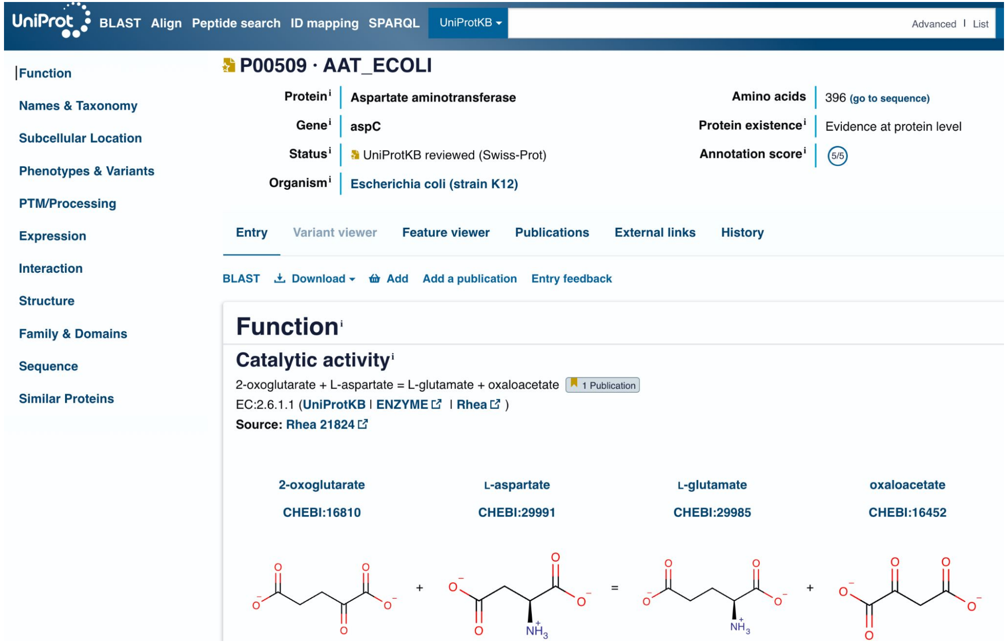
Multi-relational graphs with **(subject, predicate, object)** triples.

Multi-domain graphs:

- Encyclopedias (Wikidata, Freebase)
- **Sciences** (UniProt, DrugBank, Hetionet)

  eg, protein LMs are trained on UniProt

UniProt



Oct 11th 2023

Knowledge Graphs

Multi-relational graphs with **(subject, predicate, object)** triples.

Multi-domain graphs:

- Encyclopedias (Wikidata, Freebase)
- Sciences (UniProt, DrugBank, Hetionet)
- Thousands of **domain-specific KGs**

Spatiotemporal Urban KG

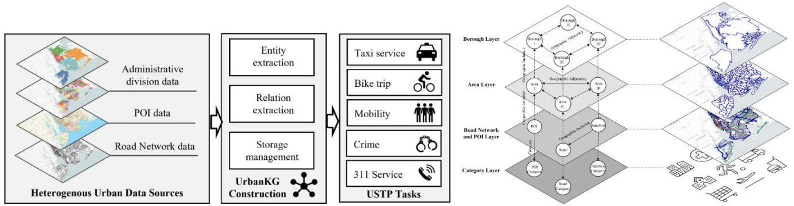
UUKG

The Unified Urban Knowledge Graph Dataset for Urban Spatiotemporal Prediction. PDF

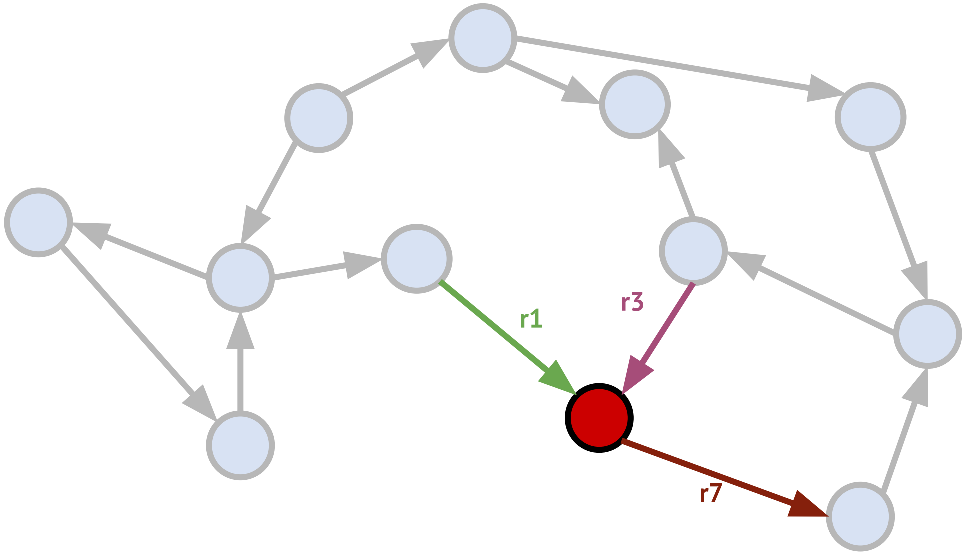
Overview • Installation • Dataset • How to Run • Directory Structure • Citation

Official repository of NeurIPS 2023 Dataset and Benchmark Track paper "UUKG: The Unified Urban Knowledge Graph Dataset for Urban Spatiotemporal Prediction". Please star, watch and fork our repo for the active updates!

1. Overview

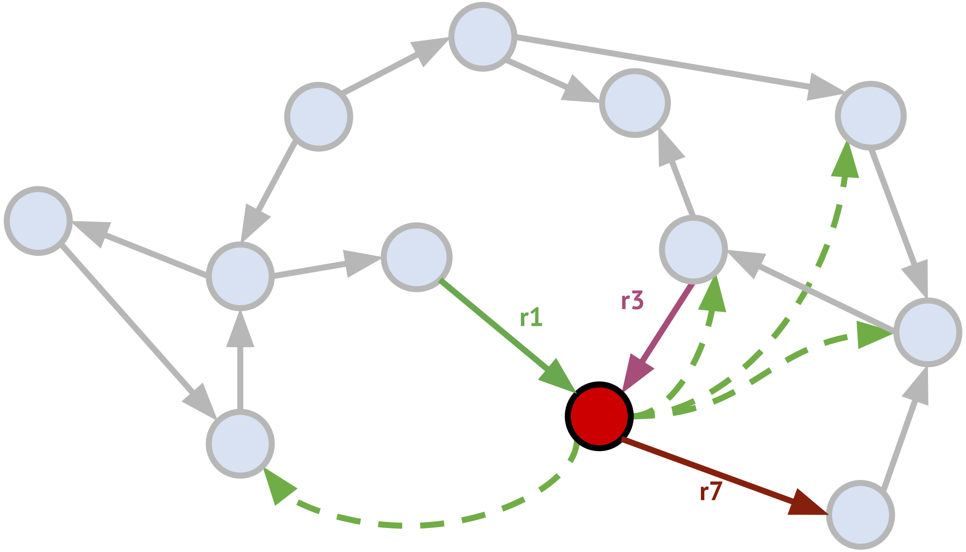


Knowledge Graphs: Setup



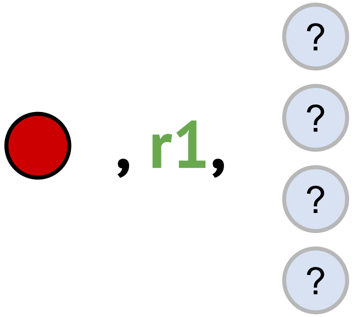
- Directed graphs (V, E)
- Explicit relation types (R)
- Input node features are **not** given
- **Transductive**: the same graph at inference
- **Inductive**: different graph at inference

Knowledge Graph Reasoning

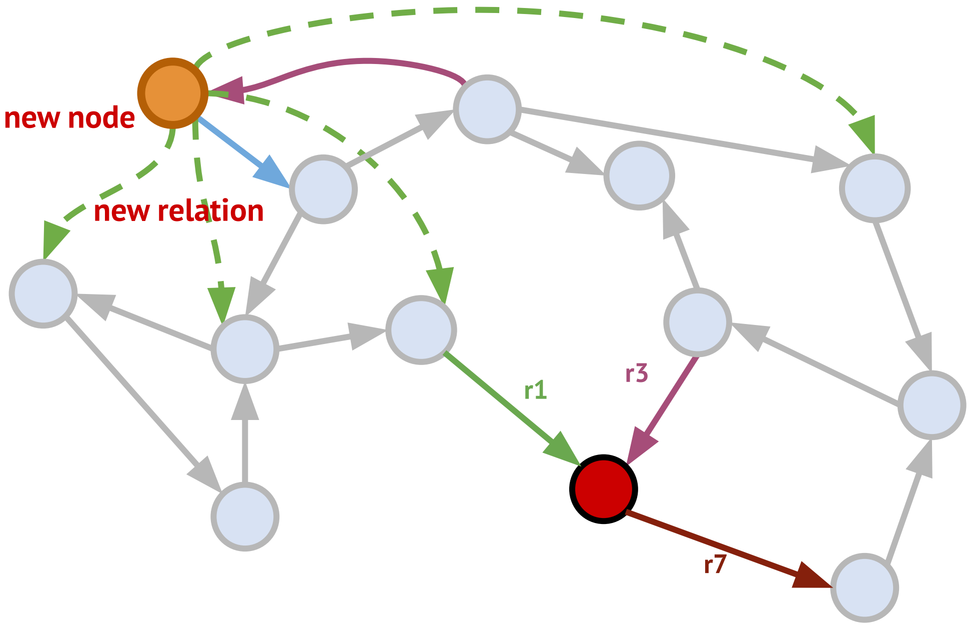


- Query: (head, relation, ?)

- Rank **all** entities as possible tails

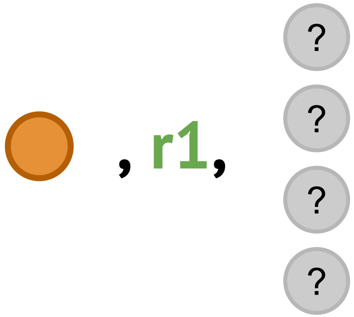


Inductive Graph Reasoning



- New nodes and relation types at inference time

- We still want to reason over new entities and relations

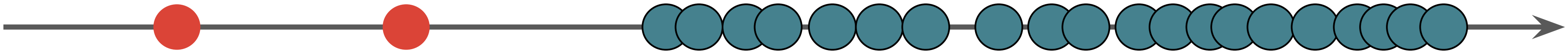


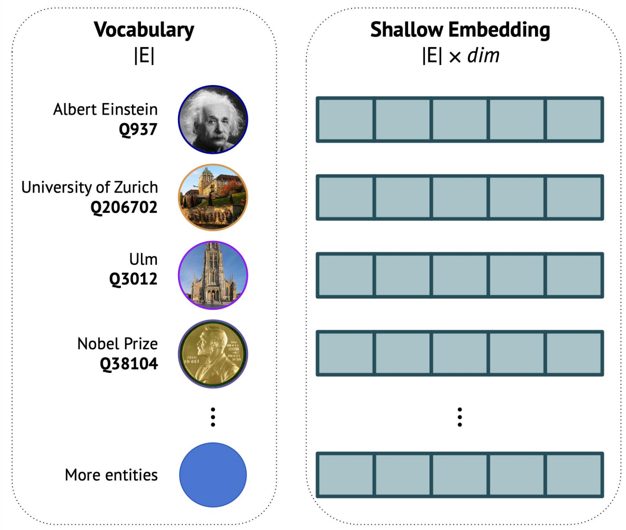
Brief History: 2011 -

RESCAL
[Nickel et al, ICML 2011]

TransE
[Bordes et al, NeurIPS 2013]

100+ KG embedding models since then 😱





Transductive models only: they learn graph-specific

- Entity embeddings (|V| x d)
- Relation embeddings (|R| x d)

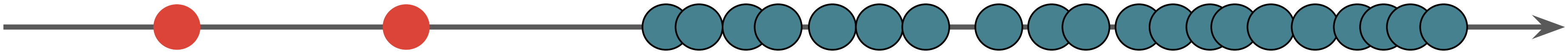
Brief History: 2011 -

Transductive | Triples | Supervised

RESCAL
[Nickel et al, ICML 2011]

TransE
[Bordes et al, NeurIPS 2013]

100+ KG embedding models since then 😱



Link Prediction on FB15k-237

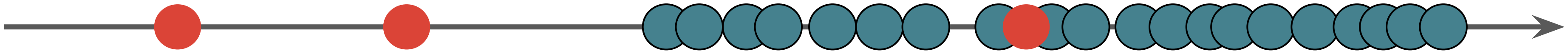
No substantial progress since 2018

Brief History: 2011 -

Transductive | **Triples** | **Supervised**

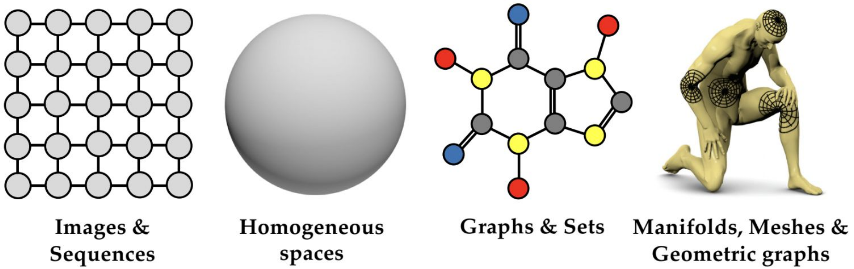
RESCAL
[Nickel et al, ICML 2011]

TransE
[Bordes et al, NeurIPS 2013]



Geometric DL
2018

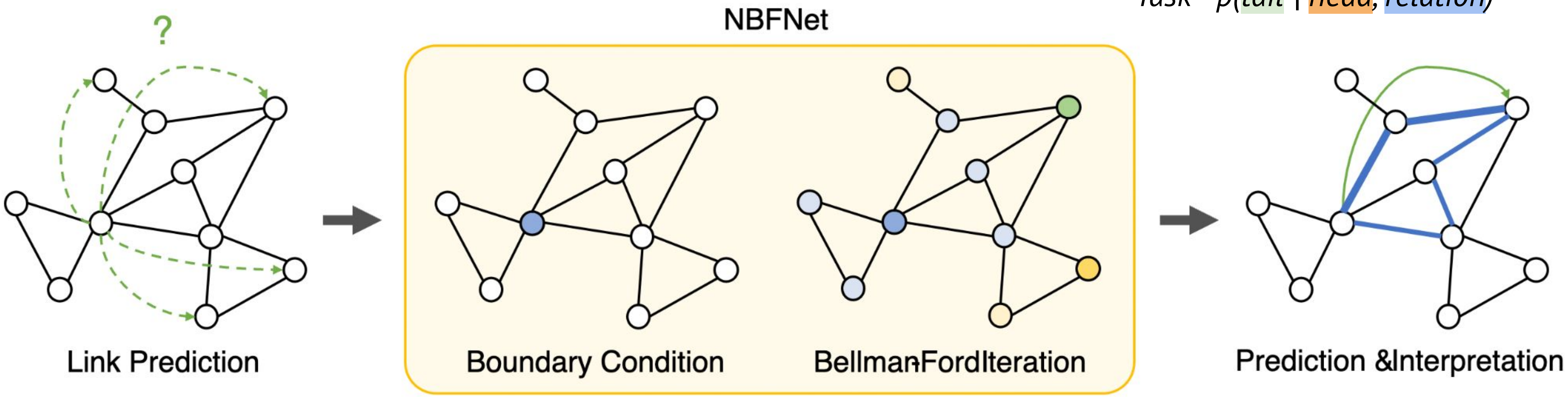
The "5G" of Geometric Deep Learning



https://geometricdeeplearning.com/

Breakthrough: Neural Bellman-Ford (2021)

Task - p(tail | head, relation)



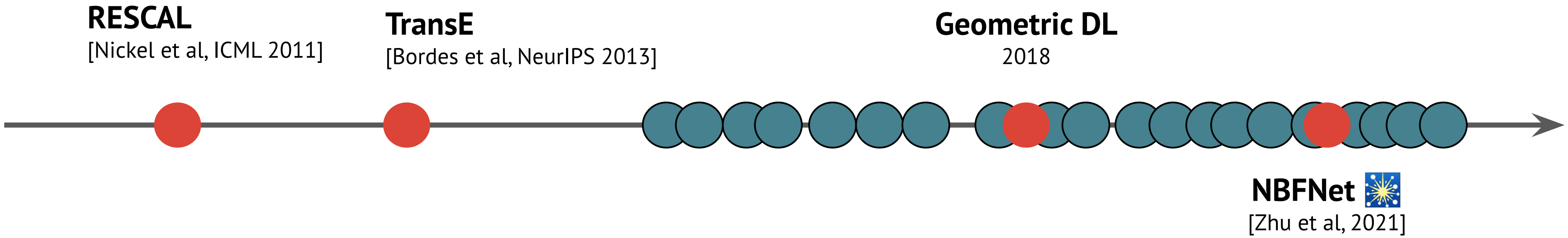
Idea:

1. Relations do not change at inference -> we can learn relation (edge type) embeddings
2. Initialize head node feature with the learnable relation vector (query)
3. Propagate for L layers, take final representations as final node features

Zhu et al. Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction. NeurIPS 2021

Brief History: 2011 -

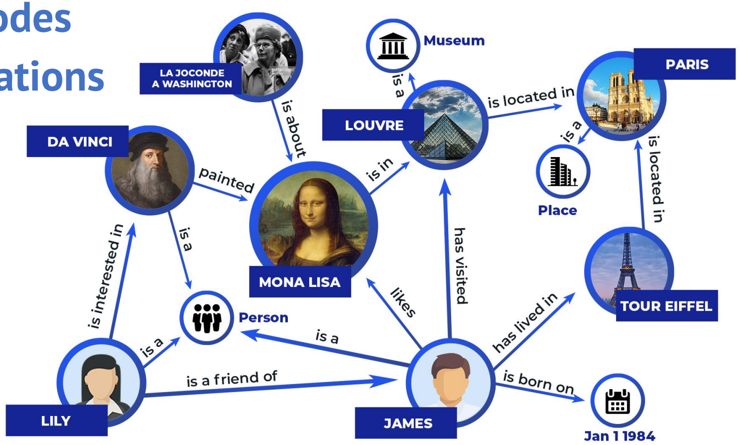
Inductive (ent) | **Triples** | **Supervised**

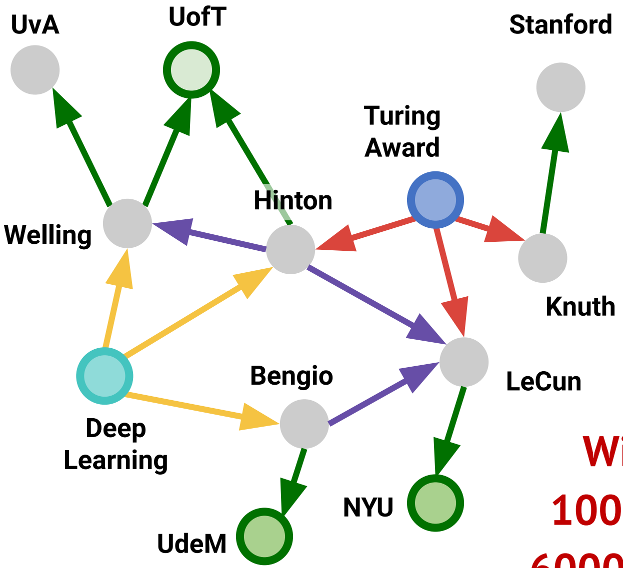


- **NBFNet** and Labeling Trick GNNs generalize to new nodes given **fixed relation types**:
- Is is possible to generalize to **both new nodes and new relation types?**

Foundation Models for Graph Reasoning

Freebase
86M nodes
1500 relations



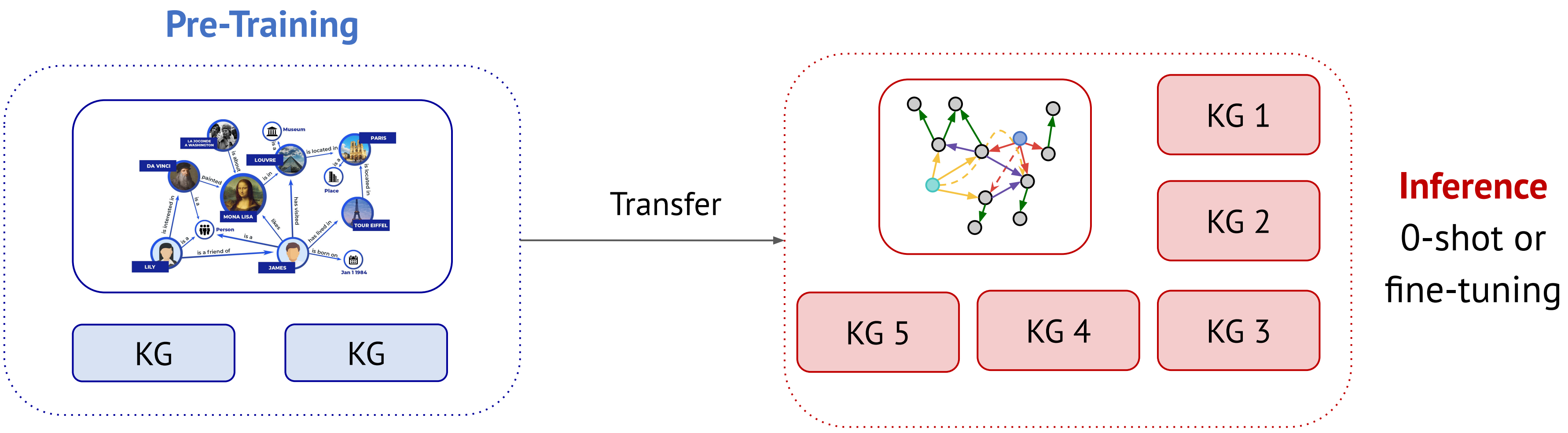


Wikidata
100M nodes
6000 relations

- ❑ Graph-specific embedding model
- ❑ Node embeddings: [86M x dim]
- ❑ Relation embeddings: [1500 x dim]
- ❑ Unique entity/relation vocabulary

- ❑ Graph-specific embedding model
- ❑ Node embeddings: [100M x dim]
- ❑ Relation embeddings: [6000 x dim]
- ❑ Unique entity/relation vocabulary

Foundation Models for Graph Reasoning

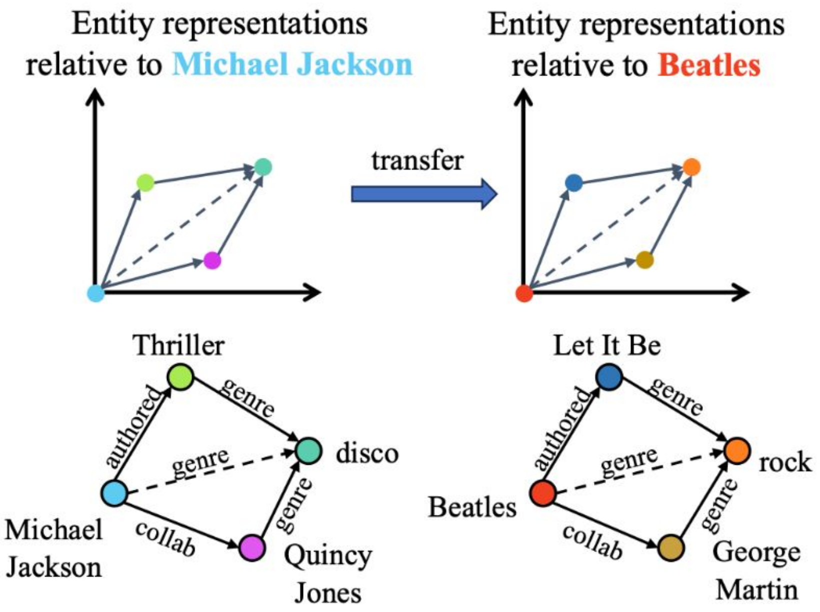


- ➔ We want to train a **single** model on one (or many) graph and run inference on **any other** possible KG
- ➔ Main problem: different entity and relation vocabularies
- ➔ For that, what is the transferable <u>invariance</u>?

Existing Inductive (entity) Models

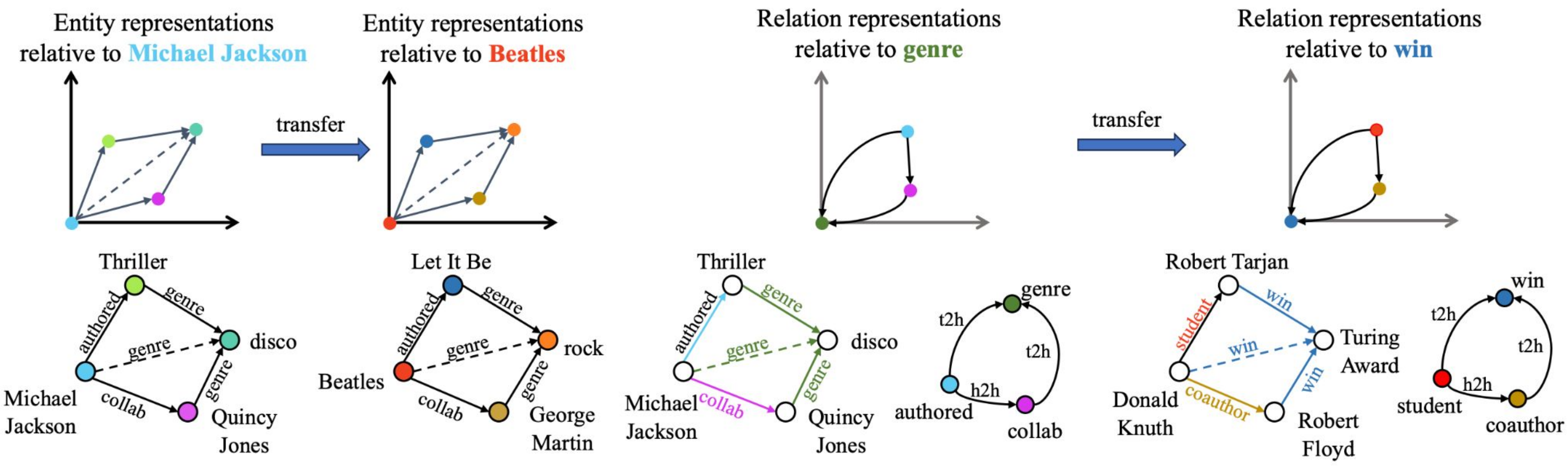
Most of existing models after NBFNet:

- **learn** relation embeddings
- build **relative** entity representations (using a labeling trick)
  - Initialize the head node with a learnable query vector ***q***
  - Other nodes <- 0
  - Message passing GNN
- Transfer to graphs with the **same relation types**



(a) Relative **entity** representations transfer to new entities (NBFNet, RED-GNN)

ULTRA: Unified, Learnable, Transferable



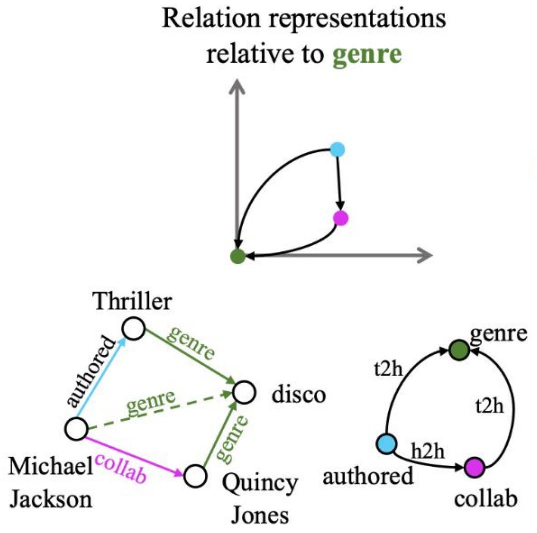
(a) Relative **entity** representations transfer to new entities (NBFNet, RED-GNN)

(b) Relative **relation** representations transfer to new relations (ULTRA)

ULTRA: Unified, Learnable, Transferable

- Let's try building a graph of relations

- 4 fundamental interactions:
  - Head-to-head (*h2h*)
  - Tail-to-head (*t2h*)
  - Tail-to-tail (*t2t*)
  - Head-to-tail (*h2t*)



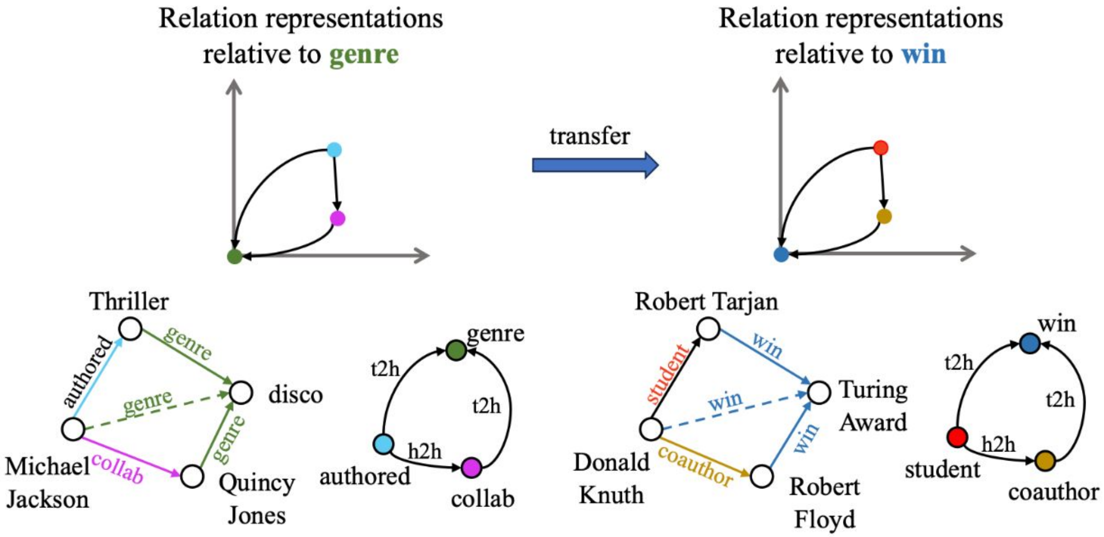
Observation: fundamental relations between relations remain the same!

ULTRA: Unified, Learnable, Transferable

- Let's try building a graph of relations

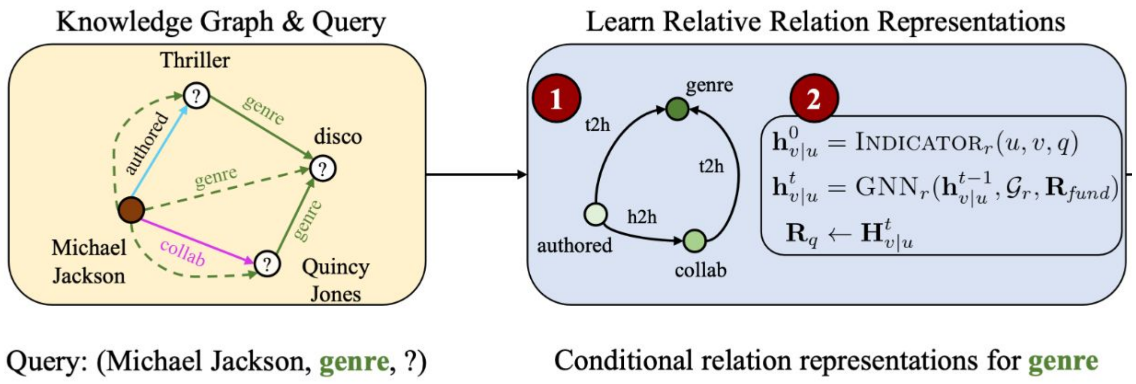
- 4 fundamental interactions:
  - Head-to-head (*h2h*)
  - Tail-to-head (*t2h*)
  - Tail-to-tail (*t2t*)
  - Head-to-tail (*h2t*)

- Can be used to infer **relative relation representations** of **new** relations



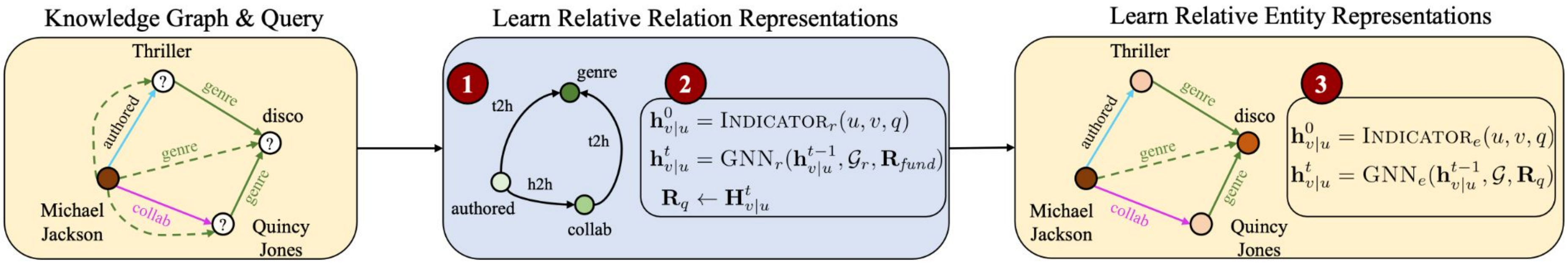
(b) Relative **relation** representations transfer to new relations (ULTRA)

Steps 1+2 : graph of relations + labeling trick



- ➔ Nodes = unique relations, edge types = 4 fundamental interactions
- ➔ Initialize the query relation node with $\mathbf{1}^d$
- ➔ Initialize the rest nodes with $\mathbf{0}^d$
- ➔ Message passing yields relative relation representations
- ➔ **Each relation = Unique relation representations |R| x d**

Step 3: run any inductive GNN



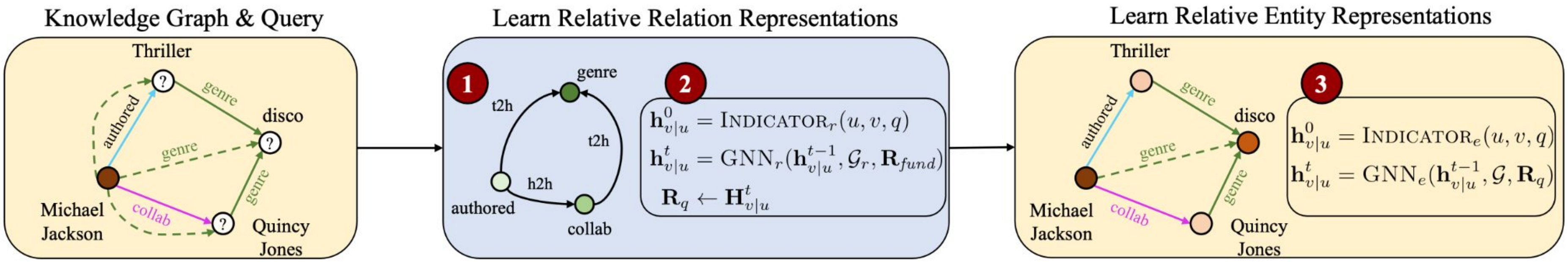
Query: (Michael Jackson, **genre**, ?)

Conditional relation representations for **genre**

Inductive link prediction using relation representations conditioned on **genre**

- ➔ **Each relation = Unique relation representations |R| x d**
- ➔ Use those relational representations for any inductive GNN (like NBFNet)

ULTRA: Foundation Model for KG Reasoning



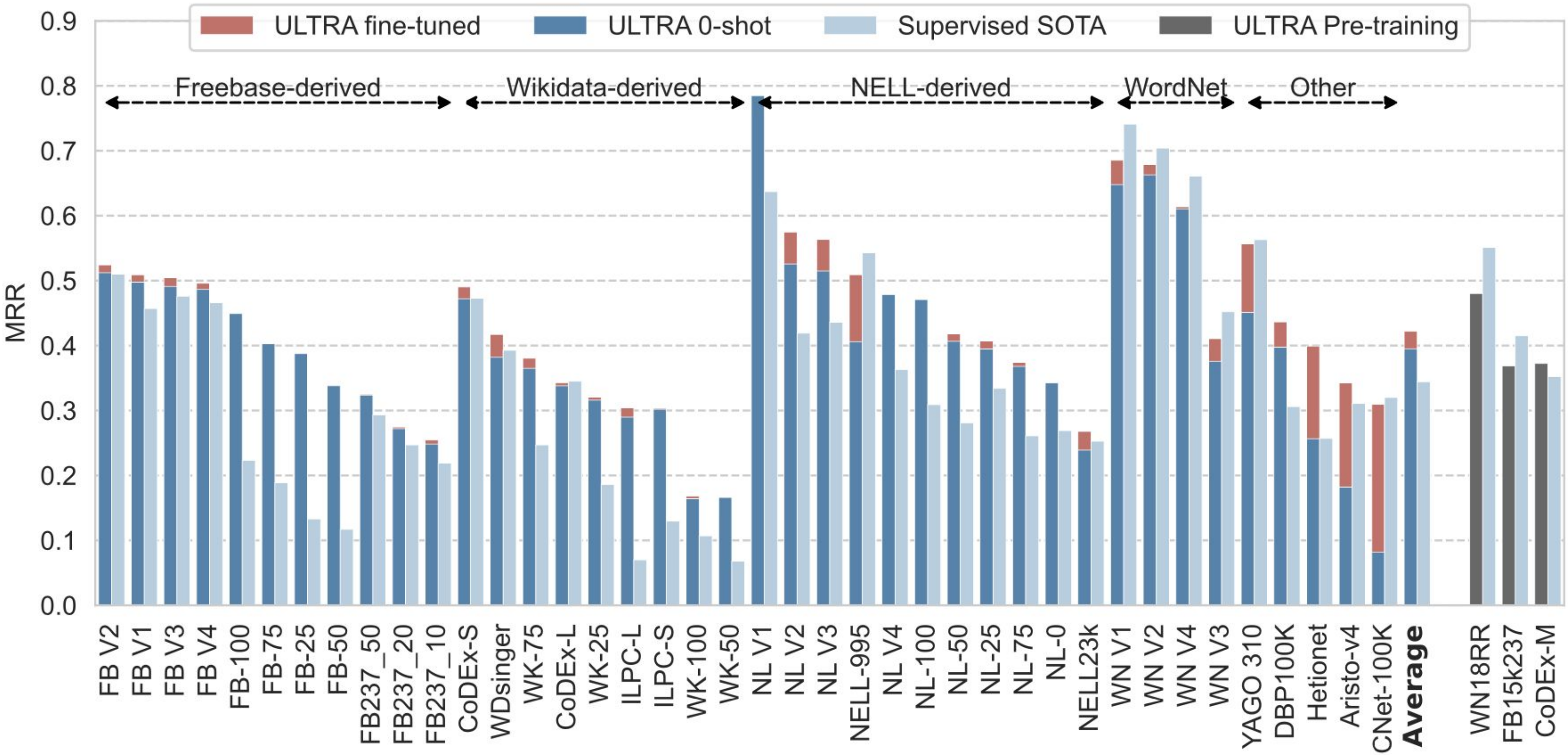
Query: (Michael Jackson, **genre**, ?)

Conditional relation representations for **genre**

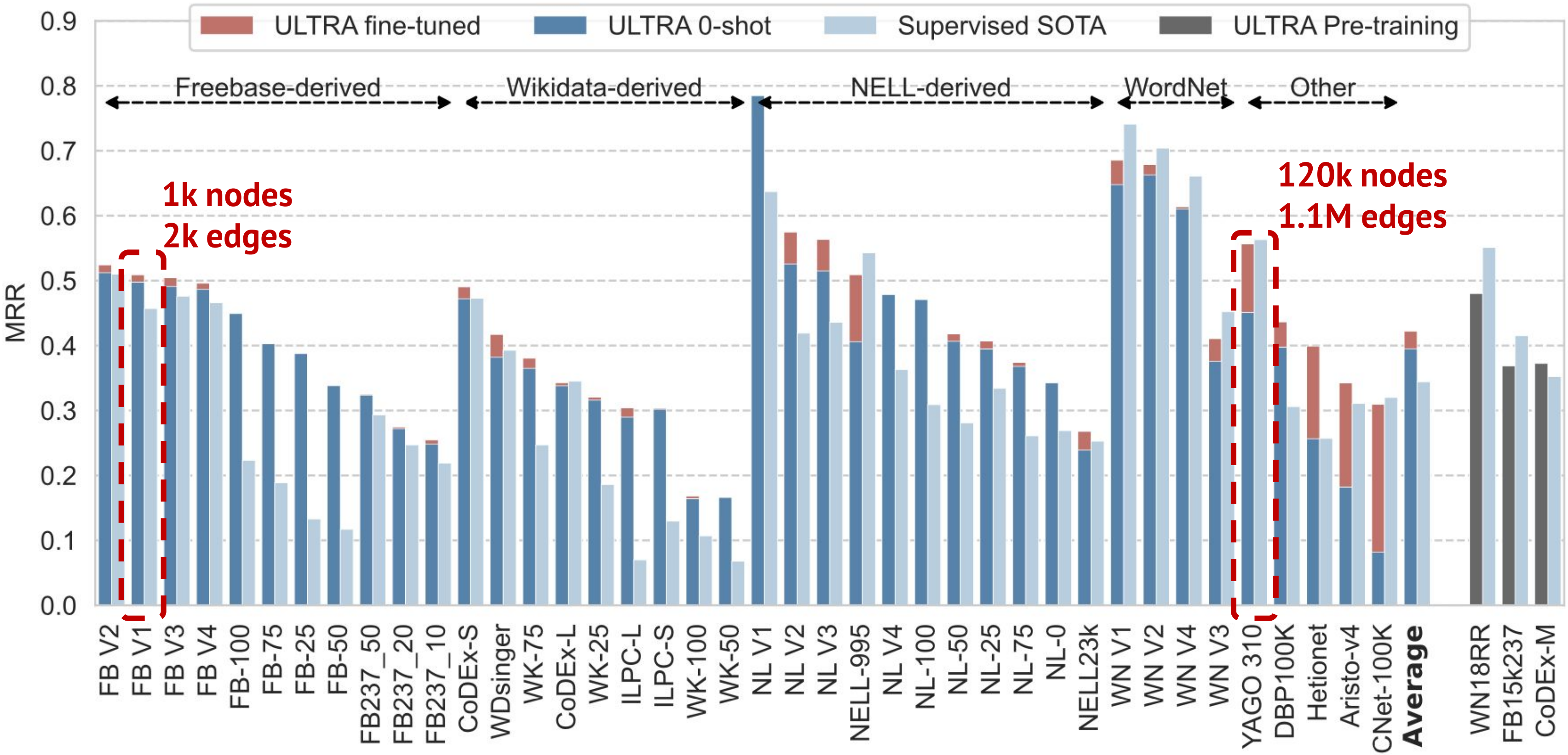
Inductive link prediction using relation representations conditioned on **genre**

- ✓ Doesn't need any input entity/relation features
- ✓ Learnable parameters: 4 fundamental relations (*h2t, t2t, t2h, h2h*) + GNN weights
- ✓ Generalizes to any graph of any size with any relation vocabulary
- ✓ Allows 0-shot inference and fine-tuning on any graph

Pre-trained ULTRA beats supervised SOTA in 0-shot inference on 50+ KGs



Generalization to different graph sizes



Oct 11th 2023

Pre-trained ULTRA beats supervised SOTA in 0-shot inference on 50+ KGs

Table 1: Zero-shot and fine-tuned performance of ULTRA compared to the published supervised SOTA on 51 datasets (as in Fig. 1 and Fig. 4). The zero-shot ULTRA outperforms supervised baselines on average and on inductive datasets. Fine-tuning improves the performance even further. We report pre-training performance to the fine-tuned version. More detailed results are in Appendix D.

Model	**Inductive** $(e)+(e,r)$ (27 graphs)		**Transductive** e (13 graphs)		**Total Avg** (40 graphs)		**Pretraining** (3 graphs)		**Inductive** $(e)+(e,r)$ (8 graphs)
	MRR	**H@10**	**MRR**	**H@10**	**MRR**	**H@10**	**MRR**	**H@10**	**Hits@10 (50 negs)**
Supervised SOTA	0.342	0.482	0.348	0.494	0.344	0.486	**0.439**	**0.585**	0.731
ULTRA 0-shot	0.435	0.603	0.312	0.458	0.395	0.556	-	-	0.859
ULTRA fine-tuned	**0.443**	**0.615**	**0.379**	**0.543**	**0.422**	**0.592**	0.407	0.568	**0.896**

- Fine-tuning is sample-efficient (2000 – 4000 batches at most)
- Fine-tuning boosts performance by further 10% relative to 0-shot

Generalization to New Unseen Domains

- ➢ Pre-trained on mostly general encyclopedia data (Freebase, Wikidata)

Graph	Domain	Supervised SOTA (MRR)	ULTRA (0-shot / ft) (MRR)
Hetionet	Biology, drugs	0.257	0.257 / **0.399**
ConceptNet	Commonsense reasoning	**0.320**	0.082 / <u>0.310</u>
Urban KG	Geography, location	0.552	0.556 / **0.618**

- ➢ Let us know more domain-specific KGs!

Pre-training + fine-tuning is better than training from scratch

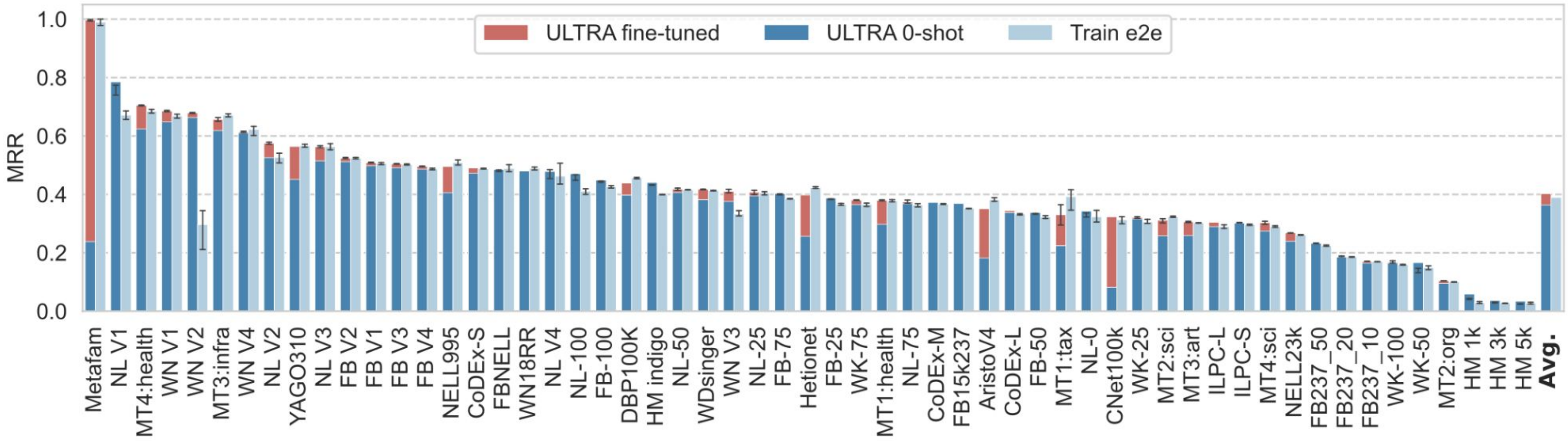


Figure 5: Comparison of zero-shot and fine-tuned ULTRA per-dataset performance against training a model from scratch on each dataset *(Train e2e)*. Zero-shot performance of a single pre-trained model is on par with training from scratch while fine-tuning yields overall best results.

\+ Save a ton of compute 😉

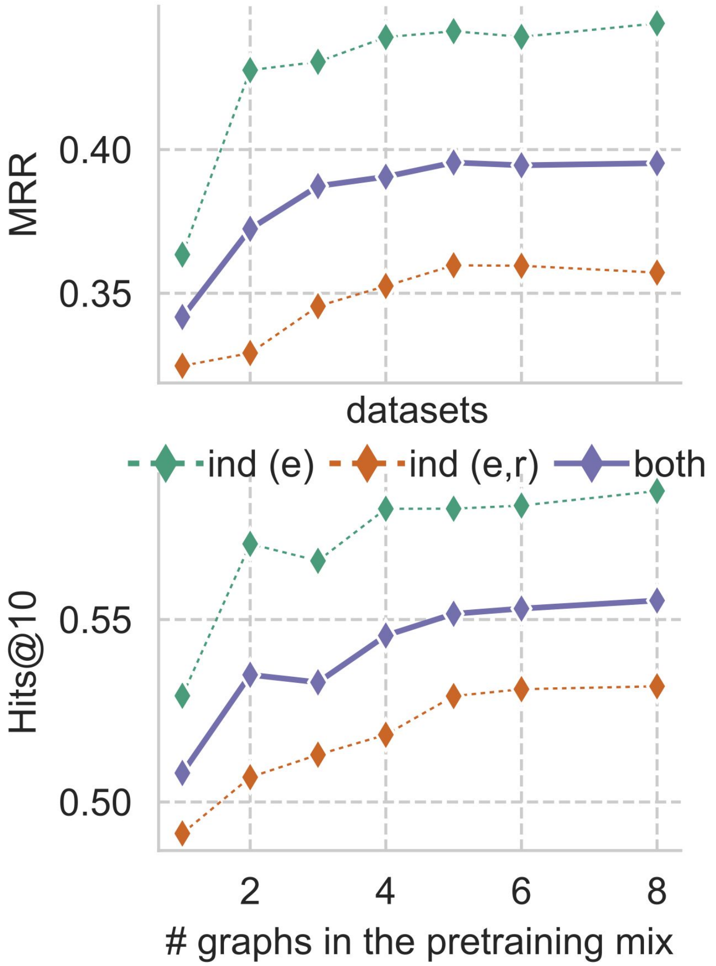
More data helps
0-shot inference

👀 Aggregated results over 40 KGs

👀 More diverse KGs in the pre-training data mix help
  - More relational graphs and their interactions

🤔 Saturation after training on 3-4 graphs

🤔 Scaling behavior to be investigated



Big Picture of KG Foundation Models

intel labs

Transductive | Triples | Supervised | Unimodal | Featurized

Inductive | Hyper-relational | Unsupervised | Multimodal | Non-featurized

SETTING

TASK

Link prediction

Node-level tasks

Graph-level tasks

Complex Query Answering

Theoretical Understanding

Scaling Laws

Foundation Model

Knowledge Graph

Open Challenges (internship projects)

1. Derive scaling laws
   - So far, the model doesn't improve after 200k params
   - Scaling model size vs scaling pre-training data

2. Investigate theoretical properties
   - Hints on the 2nd order logic and relations-of-relations

3. Extend to even more complex tasks (logical query answering)

4. Scale to LARGE graphs of billions of nodes

📜 > arxiv < 📜

Fresh from Oct, 9th

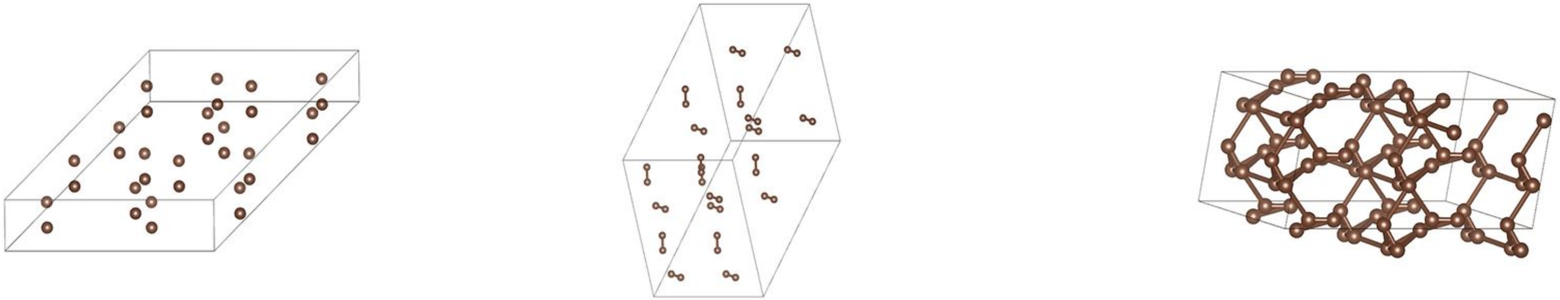
🚀 > Run the checkpoint on your own graph < 🚀

It's only 177k params

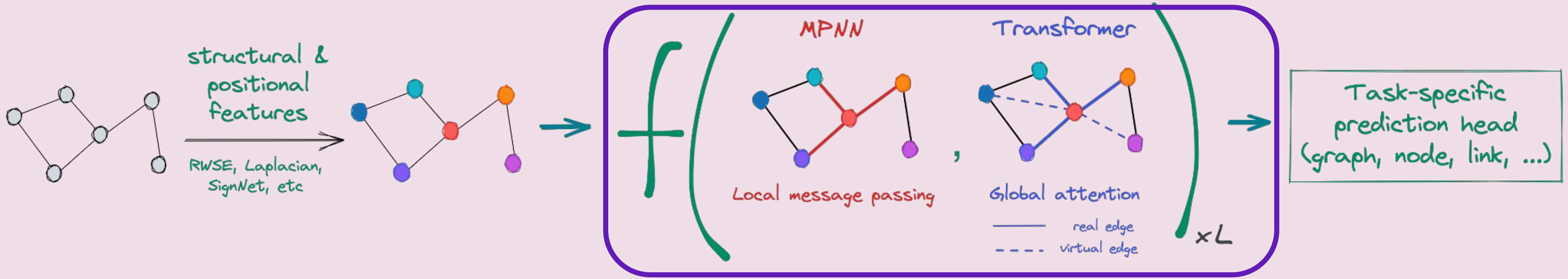
Galkin et al. Towards Foundation Models for Knowledge Graph Reasoning, 2023

Code & Data Coming soon

Foundation Models:
AI 4 Science



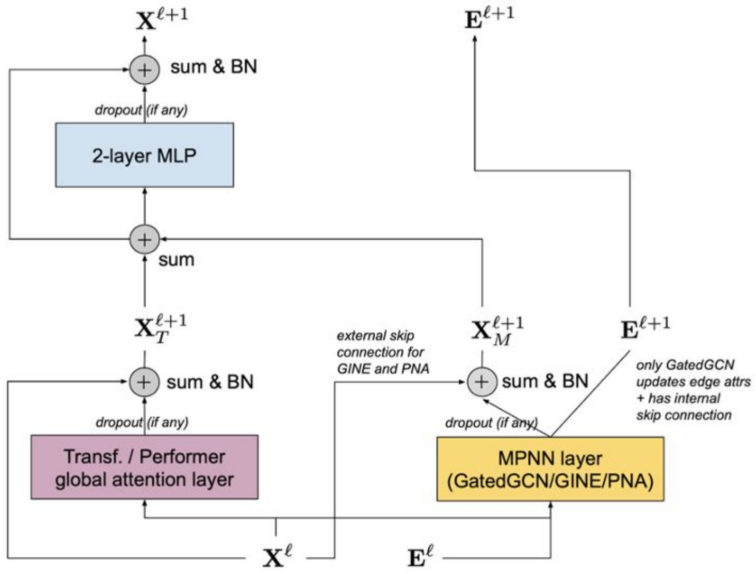
Bandgap-guided carbon structure generation
Source: https://distributionalgraphormer.github.io/



stack of L GPS layers

GraphGPS [Rampasek et al, 2022]

Entrance to the molecular ML



Combines Local MPNN and Transformer:

- Sum aggregation of the two representations
- Followed by a 2-layer MLP and skip-connections

Shameless plug: Best Graph Transformer of 2022

Recipe for a General, Powerful, Scalable Graph Transformer

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, A. Luu, Guy Wolf, D. Beaini · Computer Science · Neural Information Processing Systems · 25 May 2022

TLDR This paper proposes the first architecture with a complexity linear in the number of nodes and edges $O(N+E)$ by decoupled the local real-edge aggregation from the fully-connected Transformer, and argues that this decoupling does not negatively affect the expressivity, with the architecture being a universal function approximator on graphs. Expand

116 PDF · arXiv In Library Alert Cite

GraphGPS Public | Watch 9 | Fork 77 | Starred 455

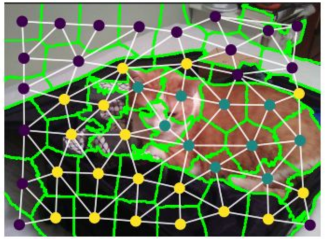
Model	PCQM4Mv2		
	Validation MAE ↓	Training MAE	# Param.
GCN-virtual	0.1153	n/a	4.9M
GIN-virtual	0.1083	n/a	6.7M
GRPE [48]	0.0890	n/a	46.2M
EGT [29]	0.0869	n/a	89.3M
Graphormer [51]	0.0864	0.0348	48.3M
GPS-small	0.0938	0.0653	6.2M
GPS-medium	0.0858	0.0726	19.4M

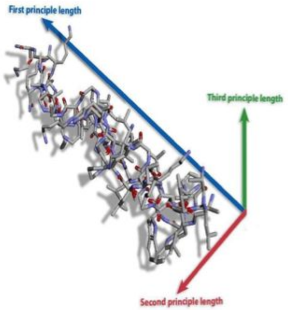
Model	ZINC
	MAE ↓
GCN [33]	0.367 ± 0.011
GIN [60]	0.526 ± 0.051
GatedGCN [7, 15]	0.282 ± 0.015
PNA [13]	0.188 ± 0.004
DGN [3]	0.168 ± 0.003
CIN [5]	0.079 ± 0.006
CRaWl [53]	0.085 ± 0.004
GIN-AK+ [67]	0.080 ± 0.001
SAN [36]	0.139 ± 0.006
Graphormer [62]	0.122 ± 0.006
K-Subgraph SAT [9]	0.094 ± 0.008
EGT [29]	0.108 ± 0.009
GPS (ours)	0.070 ± 0.004

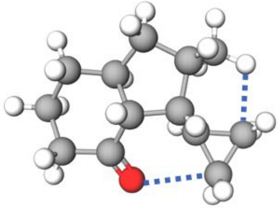
Long Range Graph Benchmark (LRGB) Results

- A new collection of datasets that require long range modeling for a network to perform well.

`reg-bound` graph







Model	PascalVOC-SP F1 score ↑	COCO-SP F1 score ↑	Peptides-func AP ↑	Peptides-struct MAE ↓	PCQM-Contact MRR ↑
GCN	0.1268 ± 0.0060	0.0841 ± 0.0010	0.5930 ± 0.0023	0.3496 ± 0.0013	0.3234 ± 0.0006
GINE	0.1265 ± 0.0076	0.1339 ± 0.0044	0.5498 ± 0.0079	0.3547 ± 0.0045	0.3180 ± 0.0027
GatedGCN	0.2873 ± 0.0219	**0.2641 ± 0.0045**	0.5864 ± 0.0077	0.3420 ± 0.0013	0.3218 ± 0.0011
GatedGCN+RWSE	0.2860 ± 0.0085	0.2574 ± 0.0034	0.6069 ± 0.0035	0.3357 ± 0.0006	0.3242 ± 0.0008
Transformer+LapPE	0.2694 ± 0.0098	**0.2618 ± 0.0031**	0.6326 ± 0.0126	**0.2529 ± 0.0016**	0.3174 ± 0.0020
SAN+LapPE	**0.3230 ± 0.0039**	0.2592 ± 0.0158*	**0.6384 ± 0.0121**	0.2683 ± 0.0043	**0.3350 ± 0.0003**
SAN+RWSE	**0.3216 ± 0.0027**	0.2434 ± 0.0156*	**0.6439 ± 0.0075**	**0.2545 ± 0.0012**	**0.3341 ± 0.0006**
GPS (ours)	**0.3748 ± 0.0109**	**0.3412 ± 0.0044**	**0.6535 ± 0.0041**	**0.2500 ± 0.0005**	**0.3337 ± 0.0006**

Dwivedi V.P., Rampášek L., Galkin M., Parviz A., Wolf G., Luu A.T. and Beaini D., *Long Range Graph Benchmark*. NeurIPS Datasets and Benchmarks 2022

GraphGPS++: ensembling 112 models

- **GraphGPS** hybrid architecture with Laplacian PEs and Random Walk SEs
- **Transformer-M** biased global attention with 2D/3D grouped input masking
- Denoising autoencoding auxiliary task (**Noisy Nodes**)

Table 4: Ensembled model performance on PCQM4Mv2 dataset. Models in the proxy set are trained on the `train+half_valid` data split whereas those in the full set are trained on all available data.

	Proxy Set			Main Set	
		Valid MAE			Ensembling
Case	**# Models**	**Avg.**	**Ensembled**	**# Models**	**Weight**
1: Baseline	10	0.0755	0.0725	35	1
2: No Atomic Number	4	0.0761	0.0734	16	0.5
3: FNN Dropout = 0.412	8	0.0759	0.0729	14	1
4: FNN Dropout = 0.412; No Atomic Number	5	0.0761	0.0736	7	0.5
5: Feature Set 2†	4	0.0755	0.0731	15	1
6: Feature Set 3†	4	0.0754	0.0731	14	1
7: Masking Weights = [1,2,2]	4	0.0754	0.0730	15	1
All	**39**	**0.0756**	**0.0722**	**112**	

† As defined in Table 2.

GPS++ is OGB LSC 2022 Winner in PCQM4M v2

Private Test Challenge

Leaderboard for PCQM4Mv2

Mean Absolute Error (MAE). The lower, the better.

Rank	Team	Test-challenge MAE
1	WeLoveGraphs	0.0719
2	ViSNet	0.0723
2	NVIDIA-PCQM4Mv2	0.0723

Public Test

Leaderboard for PCQM4Mv2

MAE on the test-dev and validation sets. The lower, the better.

Package: >=1.3.2

Rank	Method	Ensemble	Test-dev MAE	Validation MAE	Team	Contact	References	#Params	Hardware	Date
1	**GPS++**	Yes	0.0720	0.0778	GraphcoreValenceMILA	Dominic Masters (Graphcore/Valence/MILA)	Paper, Code	44,291,413	Graphcore BOW-POD16	Nov 18, 2022
2	**MolNet_Ensemble**	Yes	0.0753	0.0797	polixir.ai	zouxiaochuan (polixir.ai)	Paper, Code	32,047,874	8 RTX3090	Nov 1, 2022
3	**Global-ViSNet**	No	0.0766	0.0784	ViSNet	Tong Wang (Microsoft Research AI4Science)	Paper, Code	78,450,692	4 NVIDIA A100 GPUs	Oct 26, 2022

As a Foundation Model

Pre-training on PCQM4M v2 is a de-facto standard for other molecular tasks

Leaderboard for ogbg-molpcba

The Average Precision (AP) score on the test and validation sets. The higher, the better.

Note: The evaluation metric has been changed from PRC-AUC (Aug 11, 2020).

Package: >=1.2.2

Rank	Method	Ext. data	Test AP	Validation AP	Contact	References	#Params	Hardware	Date
1	**HIG(pre-trained on PCQM4M)**	Yes	0.3167 ± 0.0034	0.3252 ± 0.0043	Yan Wang (Tencent Youtu Lab)	Paper, Code	119,529,665	Tesla V100 (32GB)	Dec 28, 2021
2	**Graphormer (pre-trained on PCQM4M)**	Yes	0.3140 ± 0.0032	0.3227 ± 0.0024	Shuxin Zheng (Microsoft)	Paper, Code	119,529,664	NVIDIA Tesla V100 (16GB GPU)	Aug 2, 2021

How much molecular and scientific data is there?

Enormous LLM datasets vs scientific data

Beaini et al, *Towards Foundational Models for Molecular Learning on Large-Scale Multi-Task Datasets,* 2023.

How much data is there?

Fresh release: 100M molecules, 3000 tasks, 13B labels

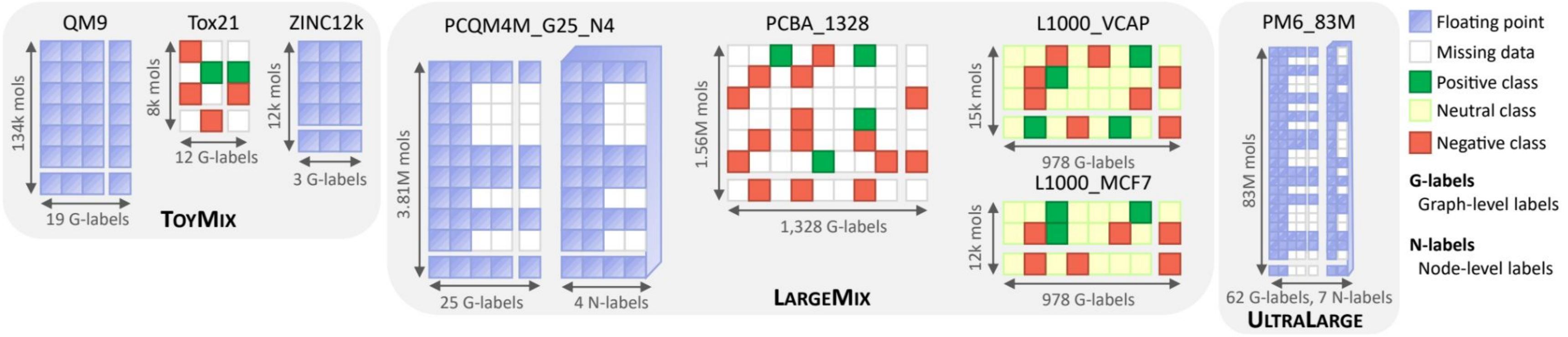


Figure 1: Visual summary of the proposed collections of molecular datasets. The "mixes" are meant to be predicted simultaneously in a multi-task fashion. They include quantum, chemical, and biological properties, categorical and continuous data points, graph-level and node-level tasks.

Beaini et al, *Towards Foundational Models for Molecular Learning on Large-Scale Multi-Task Datasets*, 2023.

What is the best pre-training objective?

Noisy Nodes [Godwin et al., 2022]
Input: 2D / 3D molecules
Output: Energy

- Aims to tackle the oversmoothing and overfitting problem in MPNNs
- Auxiliary denoising autoencoding
- Can be applied just to node and edge features, which is what we do
- 3D-based distance denoising didn't improve GPS++ performance :(

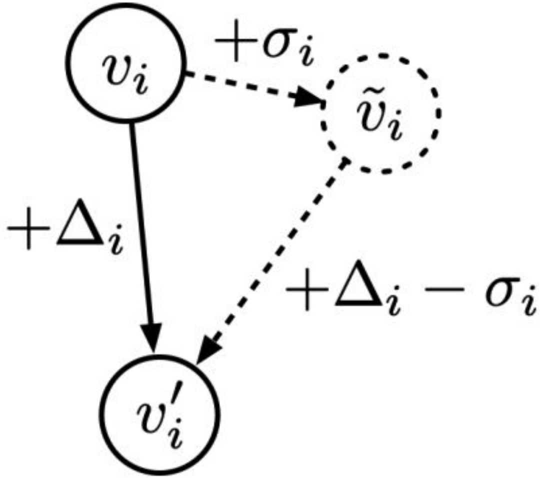


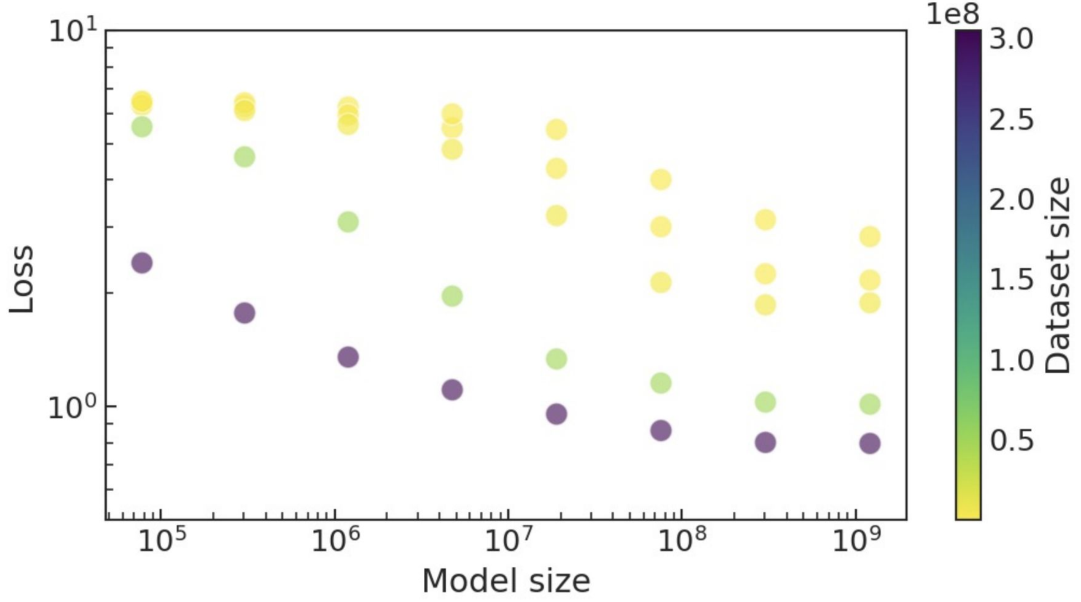
Figure 1: Noisy Node mechanics during training. Input positions are corrupted with noise σ, and the training objective is the node-level difference between target positions and the noisy inputs.

Godwin et al, *Simple gnn regularisation for 3d molecular property prediction & beyond,* ICLR 2022.

What is the best pre-training objective?

ChemGPT [Frey et al., 2022]
Input: SELFIES
Output: Next token

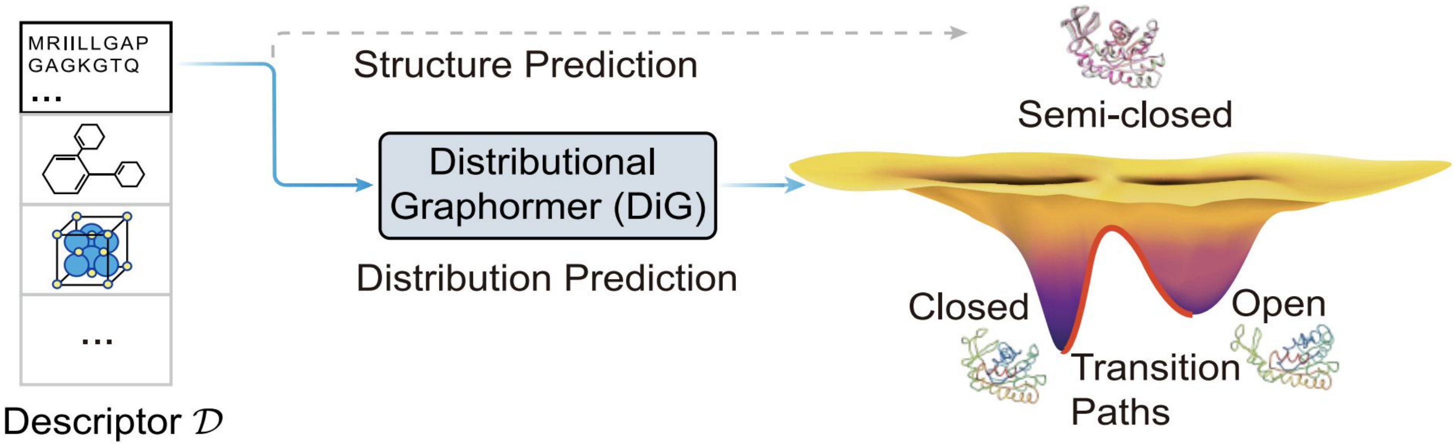
- Slap a transformer over string representations
- Some scaling laws can be derived



N Frey, R Soklaski, S Axelrod, S Samsi, R Gomez-Bombarelli, C Coley, V Gadepally, *Neural Scaling of Deep Chemical Models,* 2022.

What is the best pre-training objective?

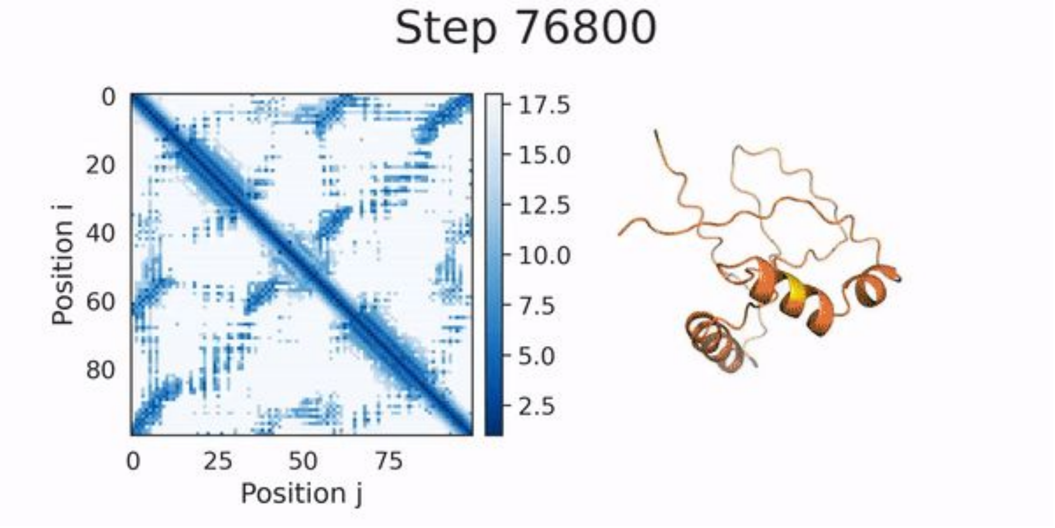
Distributional Graphormer [Frey et al., 2022]
Input: 3D structures (molecules, proteins, crystals)
Output: Equilibrium energy distribution + nice generative model



Zheng et al, *Distributional Graphormer: Towards Predicting Equilibrium Distributions for Molecular Systems with Deep Learning*, 2023.

Proteins: ESM-2 as a Foundation Model

ESM-2, ESMFold [Lin et al., 2022]
MLM on protein sequences
Bonus: 3D structure (folding) emerges from LM representations!



ESM Fold https://github.com/facebookresearch/esm

Lin, Akin, Rao, Hie et al, *Language models of protein sequences at the scale of evolution enable accurate structure prediction,* 2022.

Proteins: ESM-2 as a Foundation Model

ESM-2, ESMFold [Lin et al., 2022]
MLM on protein sequences
Bonus: 3D structure (folding) emerges from LM representations!

ESM-2 embeddings are used in a variety of protein models:

- **DiffDock** [Corso et al, ICLR 2023] - a diffusion model for protein-ligand docking

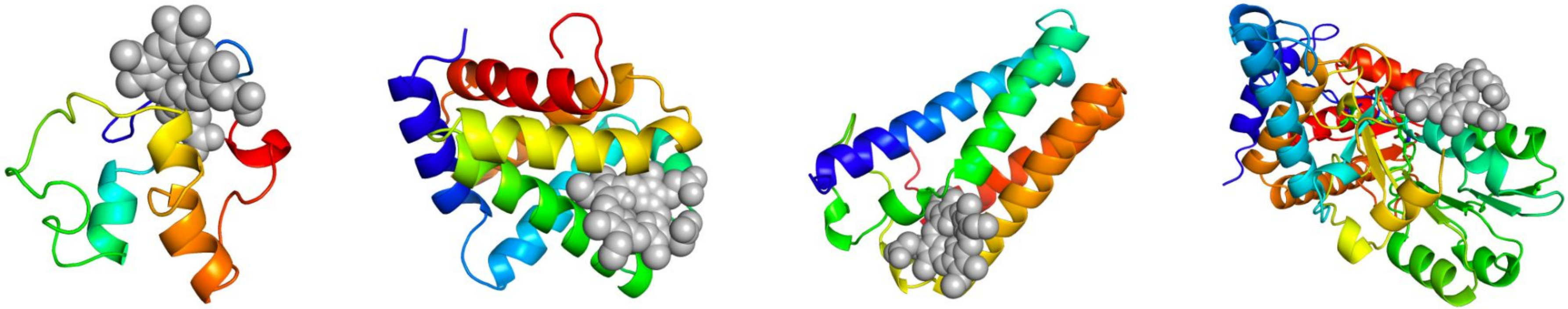
- **ProtST** [Xu, Yuan, et al, ICML 2023 Oral] - text-to-protein retrieval

Corso et al, *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking,* ICLR 2023
Xu, Yuan, et al, *ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts,* ICML 2023.

Shameless plug: ProtST

Joint pre-training on biomedical texts and protein sequences
Enables text-to-protein retrieval

Prompt **- FUNCTION: Binding to a heme, a compound composed of iron complexed in a porphyrin (tetrapyrrole) ring.**



(*1st*) 2N91-A:
- Affinity: **-7.3 (kcal/mol)**
- GO-MF label: **Bind**

(*2nd*) 1YHU-A:
- Affinity: **-7.9 (kcal/mol)**
- GO-MF label: **Bind**

(*3rd*) 5B3I-A:
- Affinity: **-8.1 (kcal/mol)**
- GO-MF label: **Bind**

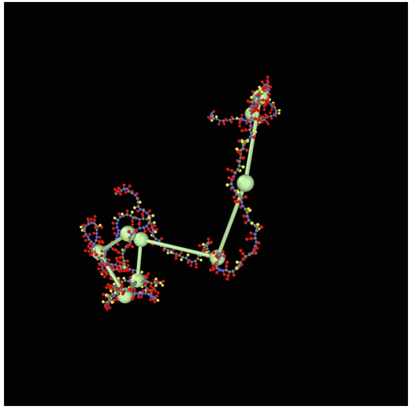
(*4th*) 5VPR-A:
- Affinity: **-7.4 (kcal/mol)**
- GO-MF label: **Non-bind**

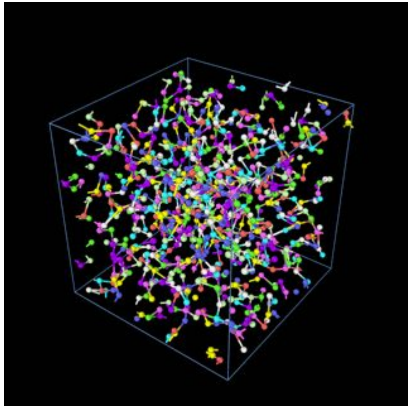
Figure 4: Zero-shot text-to-protein retrieval of heme binders based on ProtST-ESM-1b.

Xu, Yuan, Miret, Tang, *ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts,* ICML 2023 **Oral**.

Molecular Dynamics Simulations (MD)

- *aka* ML potentials, ML force fields
- Predict how a structure changes over time
  - eg, atoms 3D coordinates
  - you'd need to obtain energy, forces, acceleration, and integrate over the desired time period
- Can be applied to molecules, proteins, crystals, and materials in general
- Classic models: slow
  ML models: fast but no silver bullet





Fu et al. *Simulate Time-integrated Coarse-grained Molecular Dynamics with Multi-scale Graph Networks*. TMLR 2023

Back to Materials and Crystals

Open MatSci ML Toolkit : A Broad, Multi-Task Benchmark for Solid-State Materials Modeling

TMLR Open MatSciML Toolkit | OpenReview AI4Mat 2022 HPO | Lightning v1.8.6+ | PyTorch v1.12+ | DGL v0.9+ | PyG 2.3.1 | License MIT

https://github.com/IntelLabs/matsciml

Announcement Blog Post (Oct 9th)

- 6 datasets (1.5M materials)
- 3 baseline models
- Many training tasks incl. generative pipeline

Miret, Lee, Gonzales, Nassar, Spellings. *The Open MatSci ML Toolkit: A Flexible Framework for Machine Learning in Materials Science*. TMLR, 2023.
Lee, Gonzales, Nassar, Spellings, Galkin, Miret. *MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling*. 2023

MatSciML Toolkit & Benchmark

Task	Task Category	Data Source	#Train	#Validation	#Test	Metric
Energy Prediction Tasks						
S2EF	Property Reg.	OpenCatalyst Project [5]	2,000,000	1,000,000	-	MSE
IS2RE	Property Reg.	OpenCatalyst Project [5]	500,000	25,000	-	MSE
Formation Energy	Property Reg.	Materials Project [25]	108,159	30,904	15,456	MSE
LiPS	Property Reg.	LiPS [2]	17,500	5,000	2,500	MSE
OQMD	Property Reg.	OQMD [28]	818,076	204,519	-	MSE
NOMAD	Property Reg.	NOMAD [11]	111,056	27,764	-	MSE
CMD	Property Reg.	Carolina Materials Database [55]	171,548	42,887	-	MSE
Force Prediction Tasks						
S2EF	Property Reg.	OpenCatalyst Project [5]	2,000,000[1]	1,000,000	-	MAE
LiPS	Property Reg.	LiPS [2]	17,500	5,000	2,500	MAE
Property Prediction Tasks						
Material Bandgap	Property Reg.	Materials Project [25]	108,159	30,904	15,456	MSE
Fermi Energy	Property Reg.	Materials Project [25]	108,159	30,904	15,456	MSE
Stability	Property Class.	Materials Project [25]	108,159	30,904	15,456	ACC
Space Group	Property Class.	Materials Project [25]	108,159	30,904	15,456	ACC

Miret, Lee, Gonzales, Nassar, Spellings. *The Open MatSci ML Toolkit: A Flexible Framework for Machine Learning in Materials Science*. TMLR, 2023.
Lee, Gonzales, Nassar, Spellings, Galkin, Miret. *MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling*. 2023

Open Challenges (internship projects)

1. Designing a backbone model able to capture all the variety of 1.5M materials

2. Explore pre-training strategies

3. Improve physics-informed generative models for crystal structures

4. Run GNN-informed physical simulations (MD, DFT) for diverse materials systems at large scale

intel labs

Michael Galkin

Hesham Mostafa

Santiago Miret

Contact

mikhail.galkin@intel.com
hesham.mostafa@intel.com
santiago.miret@intel.com

Socials

@michael_galkin