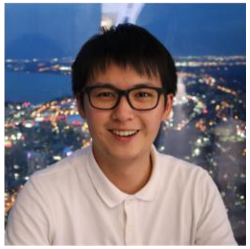
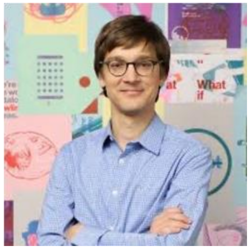
Open Graph Benchmark: Large-Scale Challenge

Weihua Hu

Joint work with
Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, Jure Leskovec

Benchmarks are Important

- Historically…
  - Computer vision

  - Natural Language Processing

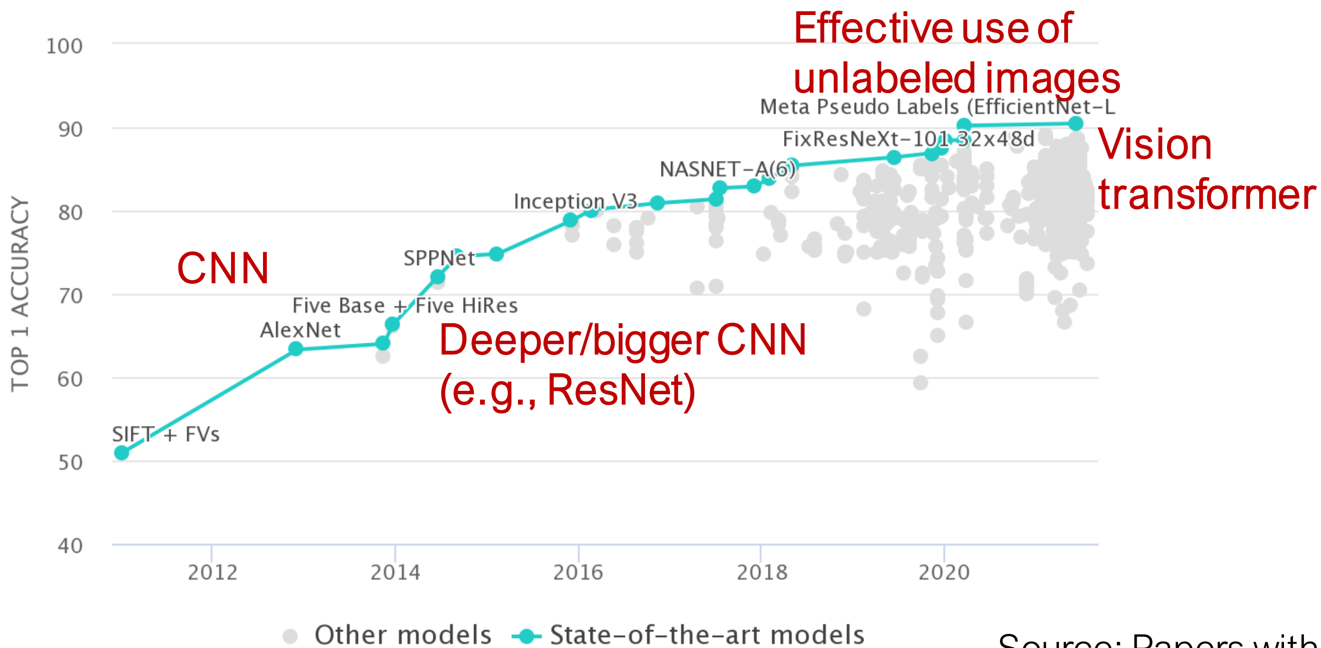
SQuAD
The Stanford Question Answering Dataset

GLUE

Benchmarks are Important

- Challenging and realistic benchmark has driven methodological innovation.

ImageNet performance



Source: Papers with code

Evaluating Graph ML

To advance research in graph ML, it is critical for our community to

develop diverse, challenging, and realistic benchmark datasets for machine learning on graphs

Open Graph Benchmark

- In May 2020, we introduced **OGB**: Realistic and diverse benchmark datasets for graph ML

OPEN GRAPH BENCHMARK

Webpage: https://ogb.stanford.edu/
Paper: https://arxiv.org/abs/2005.00687
Github: https://github.com/snap-stanford/ogb

Open Graph Benchmark

- OGB includes 15 datasets from diverse domains and tasks.

Task		Node property prediction ogbn-	
Domain	Nature	Society	Information
Small		arxiv	
Medium	proteins	products	mag
Large		papers100M	

Task		Link property prediction ogbl-	
Domain	Nature	Society	Information
Small	ddi	collab	biokg
Medium	ppa	citation2	wikikg2
Large			

Task		Graph property prediction ogbg-	
Domain	Nature	Society	Information
Small	molhiv		
Medium	molpcba / ppa		code2
Large			

OGB Python Package

- **Installation**

```
pip install ogb
```

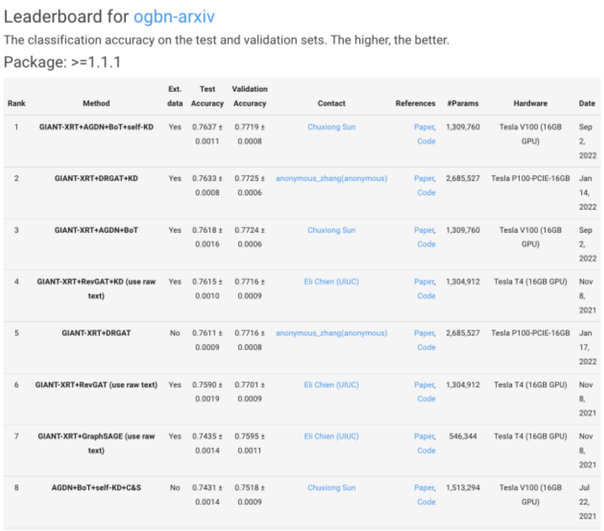
- **Data loading + splitting**

```python
from ogb.graphproppred import PygGraphPropPredDataset

dataset = PygGraphPropPredDataset(name='ogbg-molhiv')
split_idx = dataset.get_idx_split()
train_idx = split_idx['train']
```

- **Evaluation**

```python
from ogb.graphproppred import Evaluator

evaluator = Evaluator(name = 'ogbg-molhiv')
input_dict = {'y_true': y_true, 'y_pred': y_pred}
result_dict = evaluator.eval(input_dict)
```

Leaderboard


Leaderboard for ogbn-arxiv

The classification accuracy on the test and validation sets. The higher, the better.

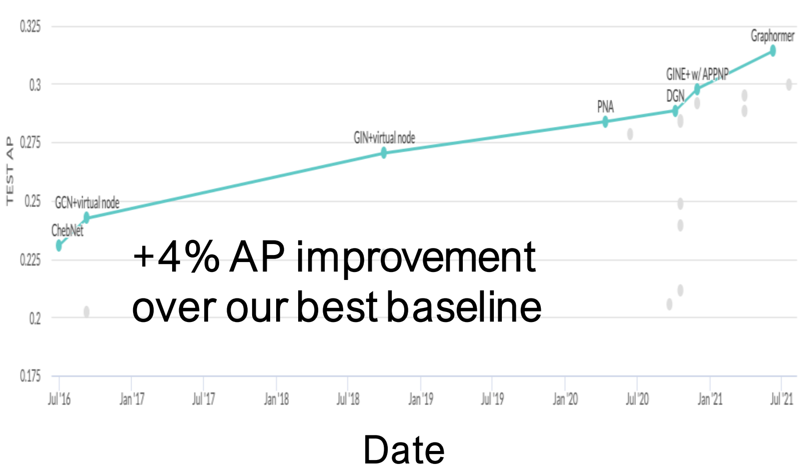
Package: >=1.1.1

Rank	Method	Ext. data	Test Accuracy	Validation Accuracy	Contact	References	#Params	Hardware	Date
1	GIANT-XRT+AGDN+BoT+self-KD	Yes	0.7637 ± 0.0011	0.7719 ± 0.0008	Chuxiong Sun	Paper, Code	1,309,760	Tesla V100 (16GB GPU)	Sep 2, 2022
2	GIANT-XRT+DRGAT+KD	Yes	0.7633 ± 0.0008	0.7725 ± 0.0006	anonymous_zhang(anonymous)	Paper, Code	2,685,527	Tesla P100-PCIE-16GB	Jan 14, 2022
3	GIANT-XRT+AGDN+BoT	Yes	0.7618 ± 0.0016	0.7724 ± 0.0006	Chuxiong Sun	Paper, Code	1,309,760	Tesla V100 (16GB GPU)	Sep 2, 2022
4	GIANT-XRT+RevGAT+KD (use raw text)	Yes	0.7615 ± 0.0010	0.7716 ± 0.0009	Eli Chien (UIUC)	Paper, Code	1,304,912	Tesla T4 (16GB GPU)	Nov 8, 2021
5	GIANT-XRT+DRGAT	No	0.7611 ± 0.0009	0.7716 ± 0.0008	anonymous_zhang(anonymous)	Paper, Code	2,685,527	Tesla P100-PCIE-16GB	Jan 17, 2022
6	GIANT-XRT+RevGAT (use raw text)	Yes	0.7590 ± 0.0019	0.7701 ± 0.0009	Eli Chien (UIUC)	Paper, Code	1,304,912	Tesla T4 (16GB GPU)	Nov 8, 2021
7	GIANT-XRT+GraphSAGE (use raw text)	Yes	0.7435 ± 0.0014	0.7595 ± 0.0011	Eli Chien (UIUC)	Paper, Code	546,344	Tesla T4 (16GB GPU)	Nov 8, 2021
8	AGDN+BoT+self-KD+C&S	No	0.7431 ± 0.0014	0.7518 ± 0.0009	Chuxiong Sun	Paper, Code	1,513,294	Tesla V100 (16GB GPU)	Jul 22, 2021

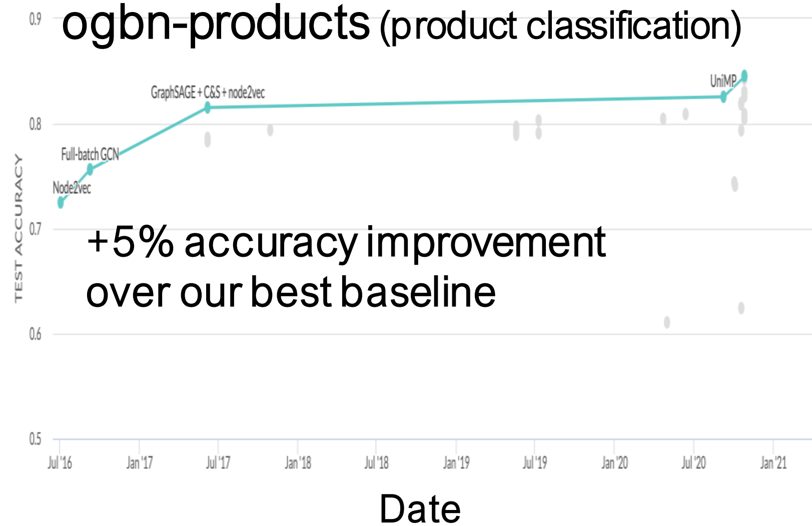
Open Graph Benchmark

- Many methods have been developed.
  - **Over 450 leaderboard submissions**
  - **Drastic accuracy improvement** on many datasets

ogbg-molpcba (molecule classification)



ogbn-products (product classification)



Source: Papers with code

Impact of OGB

- As of Sep 28th, 2022
  - 300K+ total dataset downloads
  - 350K+ total python package download
  - 1.5K Github stars
  - 800+ research papers use OGB

snap-stanford / ogb Public

Benchmark datasets, data loaders, and evaluators for graph machine learning

ogb.stanford.edu

MIT license

1.5k stars 331 forks

Open graph benchmark: Datasets for machine learning on graphs

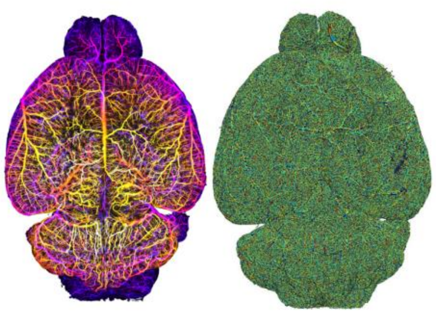
W Hu, M Fey, M Zitnik, Y Dong, H Ren… - Advances in neural …, 2020 - proceedings.neurips.cc

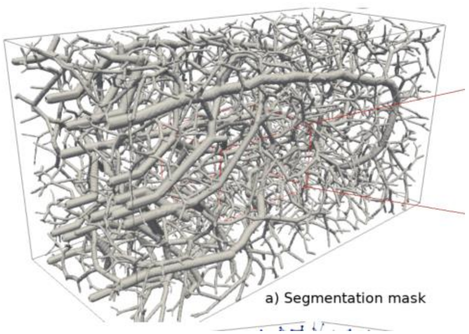
Abstract We present the Open Graph Benchmark (OGB), a diverse set of challenging and realistic benchmark datasets to facilitate scalable, robust, and reproducible graph machine learning (ML) research. OGB datasets are large-scale, encompass multiple important graph ML tasks, and cover a diverse range of domains, ranging from social and information networks to biological networks, molecular graphs, source code ASTs, and knowledge graphs. For each dataset, we provide a unified evaluation protocol using meaningful …

Save Cite Cited by 819 Related articles All 10 versions

OGB is Expanding

- **We actively look for the external dataset contributions**
- Recent contribution: ogbl-vessel by Paetzold et al. NeurIPS 2021
  - Link prediction over the incomplete whole brain vessel graph of a mouse.





Pushing Large-Scale Graph ML

- Large-scale graphs are ubiquitous
  - Billions of nodes and edges.
- But they are hard to handle
  - Training GNNs requires sophisticated mini-batching methods.
  - Embedding parameters can be huge.
  - Expensive IO, distributed training.
- We need an ML challenge to push the frontier!

OGB Large-Scale Challenge

- For the ACM KDD Cup 2021, we provided a set of three challenging large-scale graph datasets

OGB-LSC
— Large - Scale Challenge —

Webpage: https://ogb.stanford.edu/docs/lsc
Paper: https://arxiv.org/abs/2103.09430
Github: https://github.com/snap-stanford/ogb

OGB-LSC Stats

- Attracted huge attention from the community
  - 500+ registrations across the globe.
  - 123 teams submitted to the final test submission.
- Institutions:
  - Academia: 60%, Industry: 40%

OGB-LSC Datasets

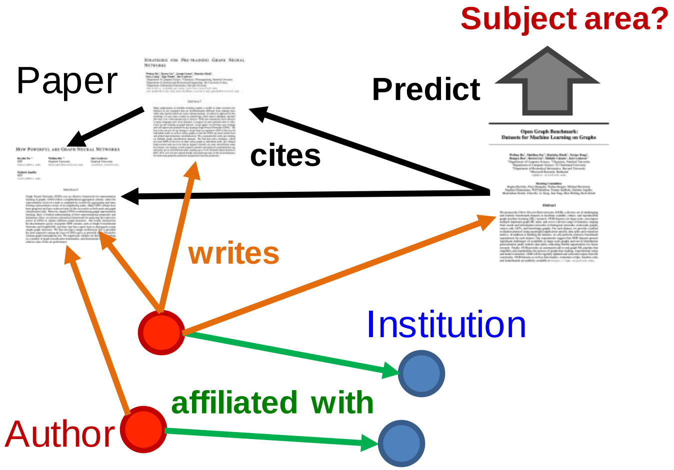
- LSC datasets are orders-of-magnitude larger than any exiting datasets

Task type	Dataset	Statistics	
Node-level	MAG240M-LSC	#nodes:	244,160,499
		#edges:	1,728,364,232
Link-level	WikiKG90M-LSC	#nodes:	87,143,637
		#edges:	504,220,369
Graph-level	PCQM4M-LSC	#graphs:	3,803,453
		#edges (total):	55,399,880

- Each dataset is practically relevant

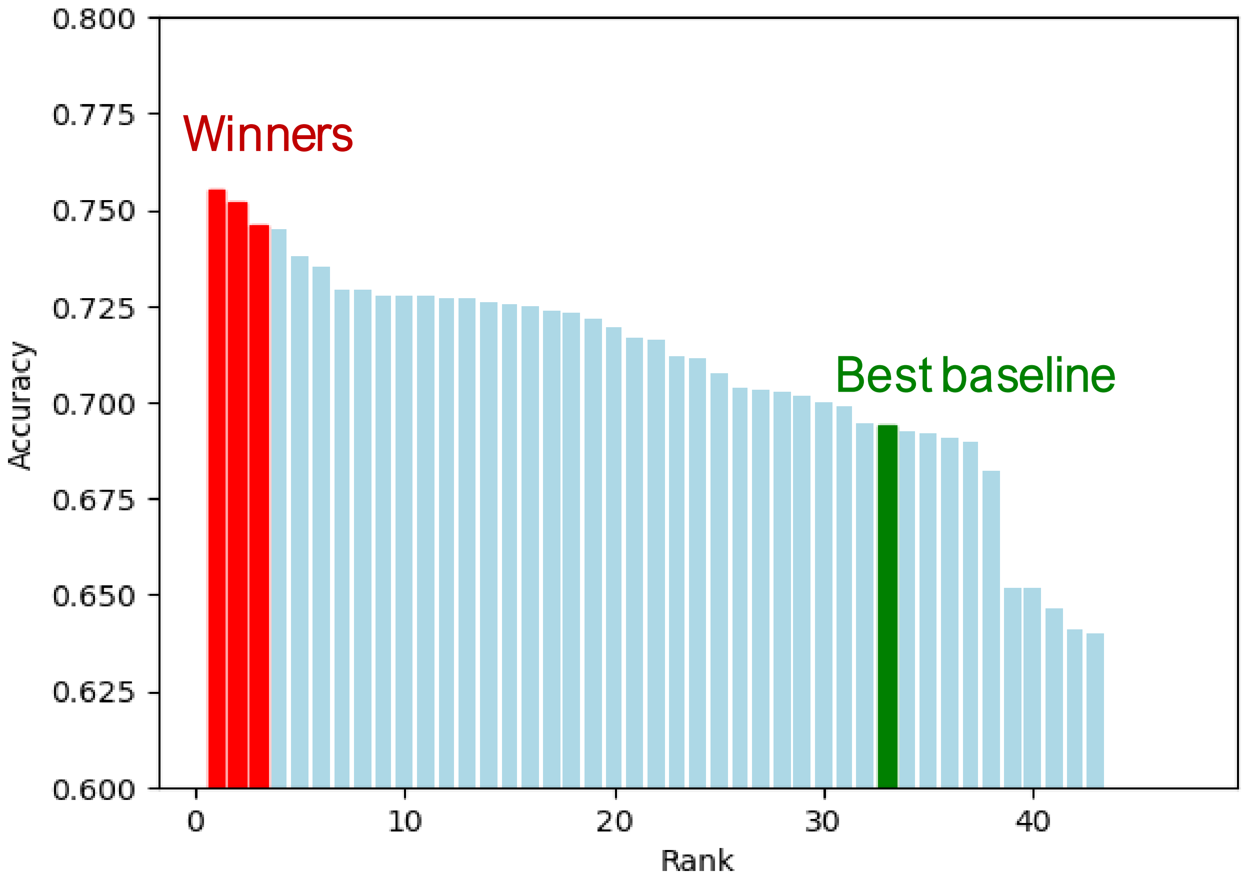
Node-Level: MAG240M

- Heterogeneous academic graph
- Task: Predict the subject areas of papers situated in the heterogeneous graph (node classification)



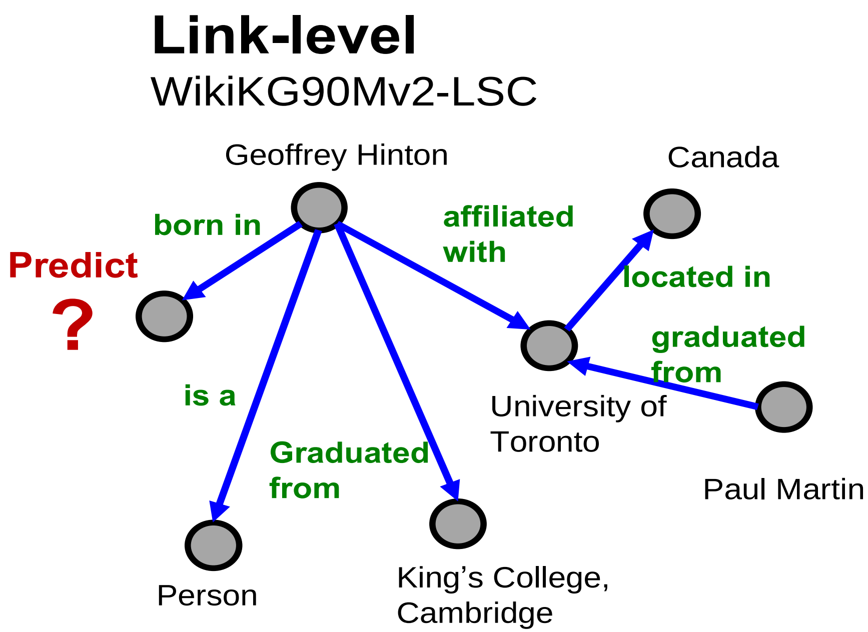
Performance Improvement

Accuracy: The higher, the better.



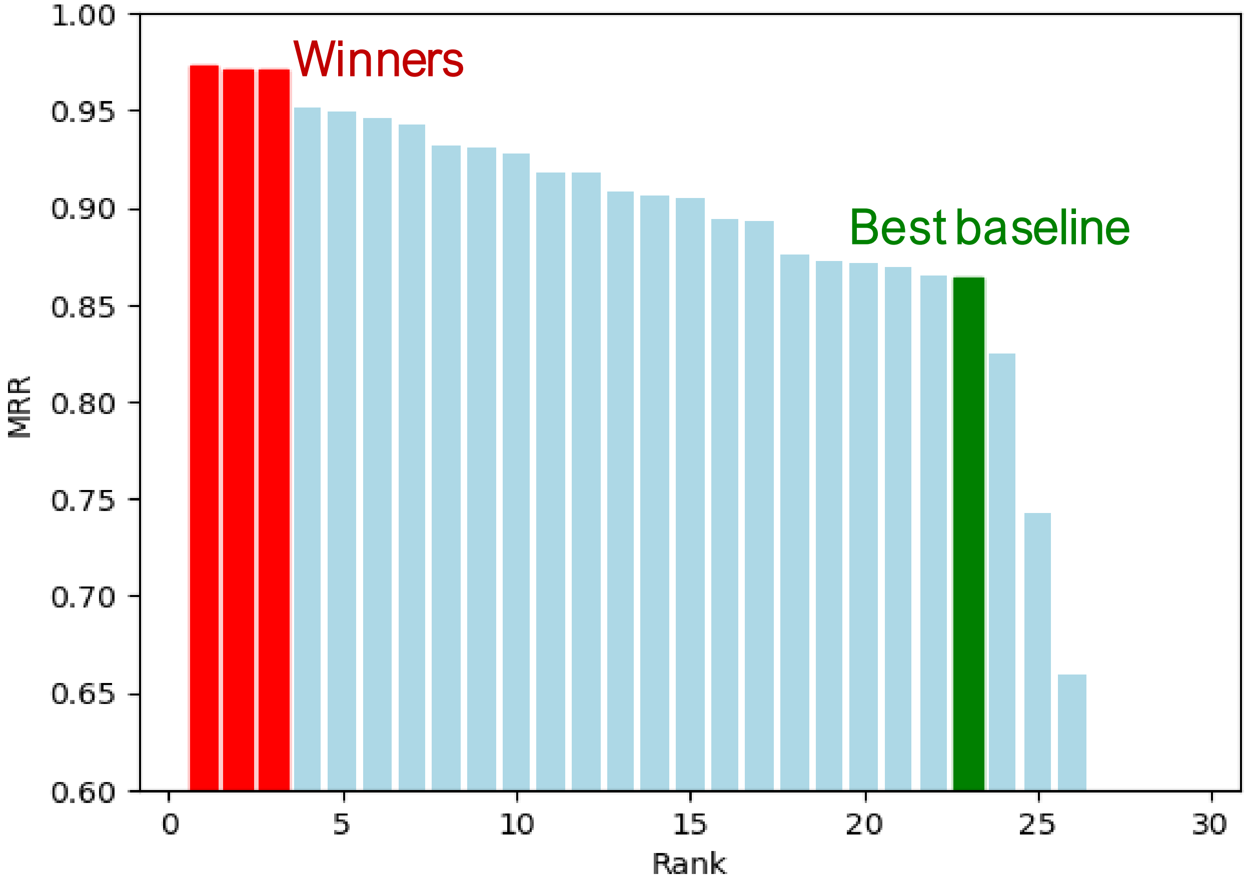
Link-Level: WikiKG90M

- Knowledge graph
- **Task**: Impute missing triplets (link prediction)



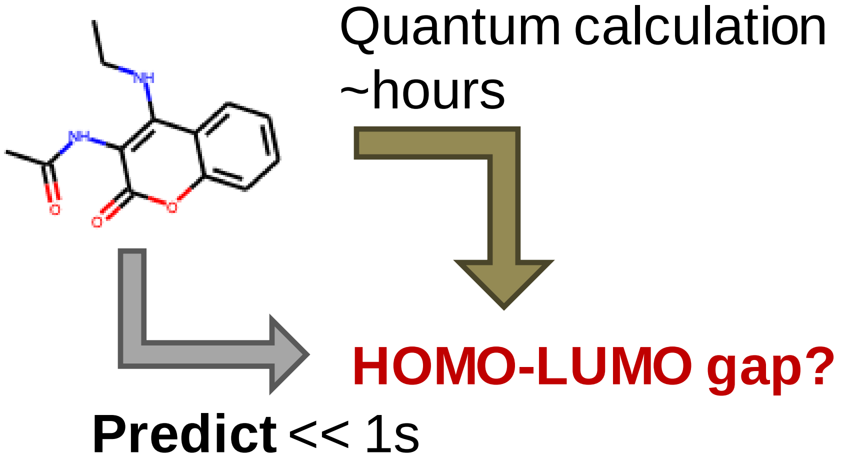
Performance Improvement

MRR: The higher, the better



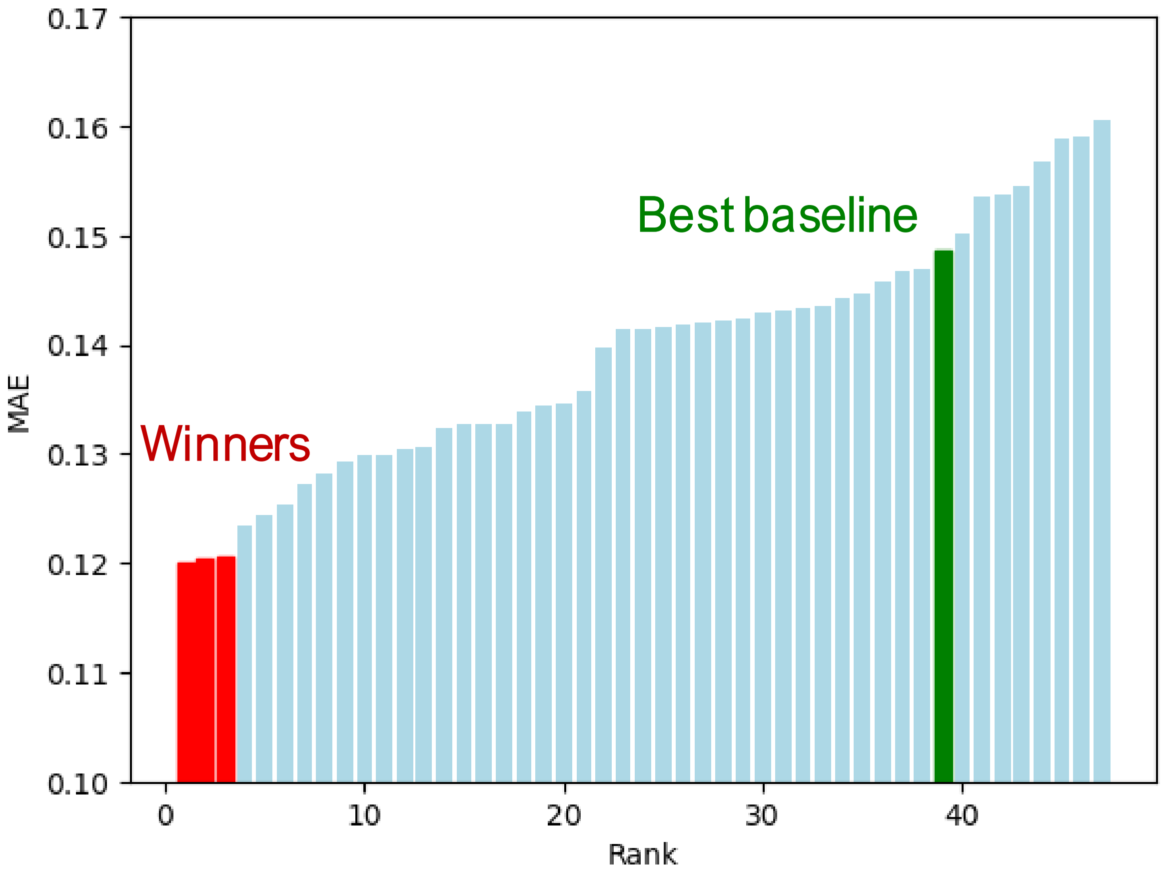
Graph-Level: PCQM4M

- Molecular graphs
- Task: Predict an important quantum chemistry property, the HOMO-LUMO gap, of a given molecule (graph regression).



Performance Improvement

MAE: The lower, the better.



Overall Observations

- Many novel techniques developed for the OGB-LSC large graphs
  - New mini-batch sampling techniques for hetero-generous graphs
  - New label propagation methods using GNNs.
  - New knowledge graph embedding models
  - New self-supervised learning methods for GNNs
  - Deeper, bigger, and more expressive GNNs
- More details are in our OGB-LSC paper: https://arxiv.org/abs/2103.09430

2^{nd} OGB-LSC

OGB-LSC @ NeurIPS 2022

Large - Scale Challenge

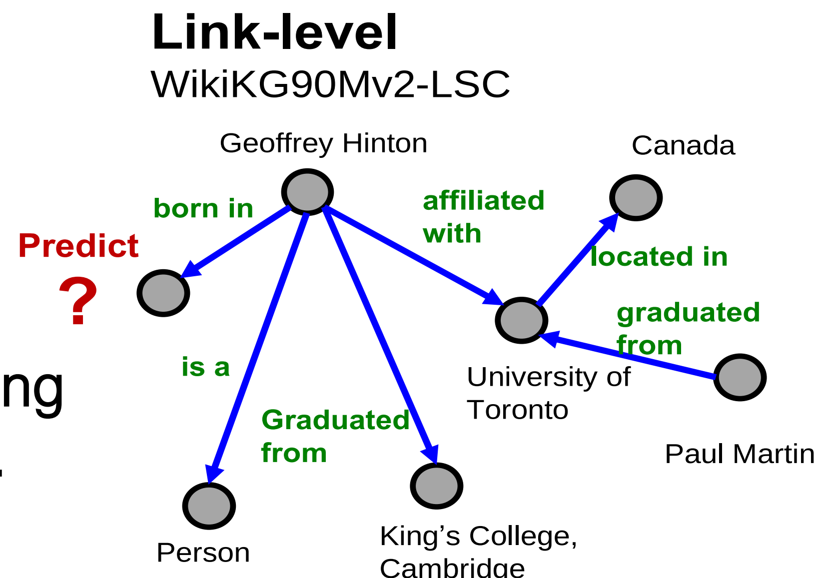
- The 2^{nd} iteration of OGB-LSC happening at the NeurIPS 2022.
- Webpage: https://ogb.stanford.edu/neurips2022/
- Winners and their solutions (code and technical report) will be announced on late November.

2nd OGB-LSC

- Uses the similar three datasets as the 1st OGB-LSC @ KDD Cup 2021.
  - Helps keep track of the progress every year
  - Similar to the annual ImageNet challenge.
- Some datasets have been updated to be more challenging and realistic.

Updates in 2nd OGB-LSC (1)

- No candidate sets for link prediction



1st OGB-LSC: Predict among the 1001 candidate nodes.

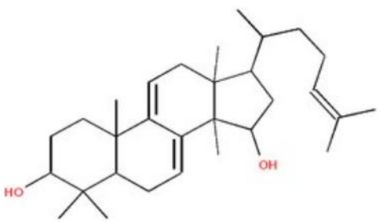
2nd OGB-LSC: Predict among all entities (~90M nodes)

Updates in 2^{nd} OGB-LSC (2)

- 3D molecular graph provided for training molecules

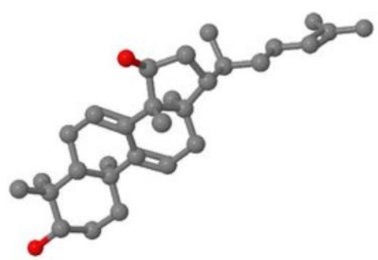
1^{st} OGB-LSC

2D molecular graph



Expensive Geometry optimization

3D molecular graph



Property prediction

HOMO-LUMO gap

2^{nd} OGB-LSC

Conclusions

- We presented **OGB and OGB-LSC** to accelerate research in graph ML.
- **OGB is expanding** with external dataset contributions.
- We are organizing **2nd OGB-LSC at NeurIPS 2022** to push large-scale graph ML!