Research and Development of GNNs @ Amazon

Vassilis N. Ioannidis

Senior Applied Scientist, AWS AI Research & Education, AWS Graph ML, DGL
Ph.D. University of Minnesota

amazon | science

Our mission : AWS Graph ML (AGML)

Simplify research and development of graph machine learning (GML) (DGL)

Maximize the impact of GML across Amazon

Maximize the impact of GML across business (Neptune ML)

Overview of AGML research

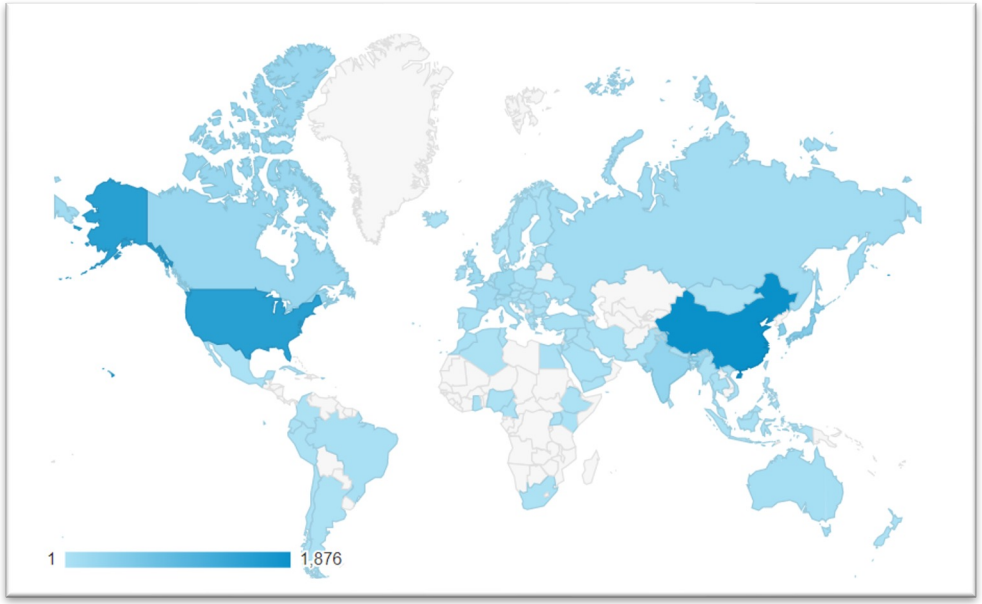
- GNN research in models, training methods and systems.
  - Efficient and scalable training
  - Better model performance

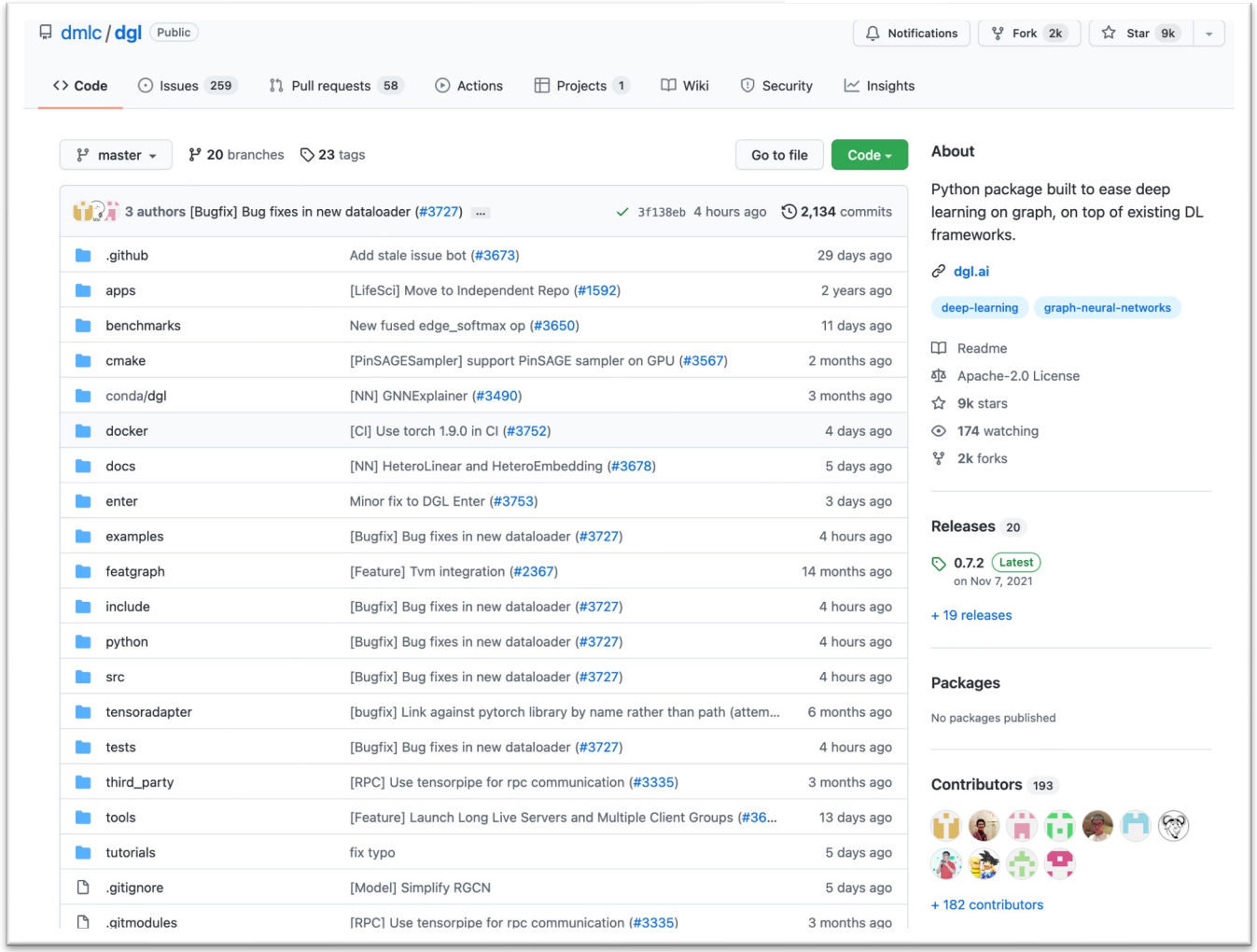
Scalable training:
- DGL-KE: Training Knowledge Graph Embeddings at Scale
- DistDGL: distributed GNN training at scale.
- Global neighbor sampling
- Compress embeddings for GNN with Tensor Decomposition
- TGL: temporal graph library
- Scalable temporal GNN training
- Scalable link prediction
- Adaptive GNN sampling for heterogeneous graphs
- Scalable GNN training with external memory
- Real-time GNN online inference system

Model performance:
- PanRep: universal representation learning
- GNN pre-training & fine-tuning
- Active learning for GNN
- Advanced task-aware sampling methods
- GNN ensembles
- LM-GNN for graphs with rich text features
- GNN distillation
- GNN explanability
- KG Reasoning
- KG for Health Science: DRKG, CTKG, Covid-19 KG
- Conditional equivariance for protein molecules
- GNNs for tabular data
- Improving drug design using GRNs

How to develop GNNs

DGL is a vibrant open-source effort



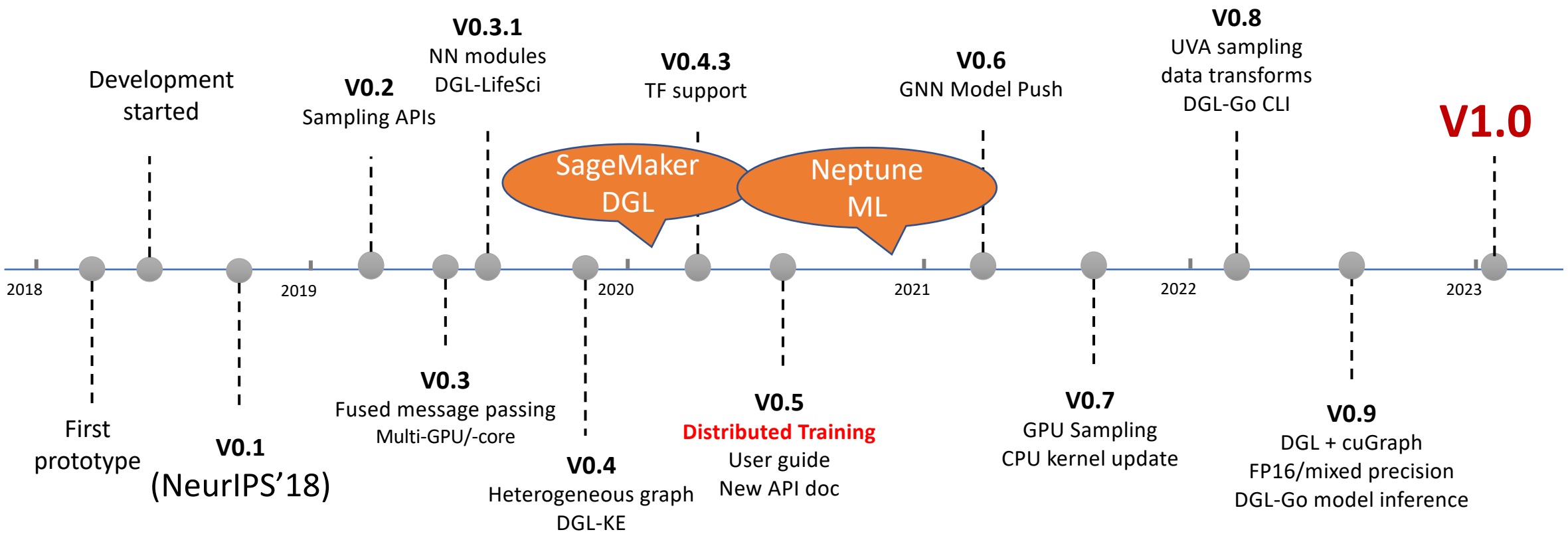


Vital stats

- Github stars: 10.1K
- Forks: 2.4K
- Contributors: 229
- DGL paper citations: 780
- CPU, GPU and x of them
- Scale: 1B nodes, 10B edges
- Model zoo: ~90

https://www.dgl.ai

DGL Release History



Active learning with GNNs

- Given Graph $G = (V, E)$, with features $v \in V$ $x_v \in R^d$ and labels $y_v \in \{1,2, \ldots, p\}$
- Annotating labels can be expensive in financial or medical data.
- *Can we select which nodes to annotate to improve GNN learning performance?*

Diversify the uncertain nodes in GNNs

- Uncertainty sampling selects nodes with high prediction uncertainty

- Maximizing the uncertainty may not explore well the classification boundary
  - Selected nodes may be clustered in one region of the graph – under exploration of the graph

- *How to diversify the uncertain samples for GNN training?*
  - Propagate the input features for *K* steps for the *M* most uncertain nodes
  - Cluster the propagate features in *C* clusters
  - Query the *C* nodes for their labels that are closest to the cluster centers

- Theoretical results confirming the advantage for diversifying the samples

Z. Dai, V. Ioannidis, S. Adeshina, Z. Jost, C. Faloutsos, G. Karypis "*ScatterSample: Diversified Label Sampling for Data Efficient Graph Neural Network Learning*" published in Learning on Graphs Conference 2022

ScatterSample: Diversified Label Sampling

Algorithm 1 *ScatterSample* Algorithm

Input: $\mathcal{V}_{train}$, GNN model M, number of propagation layers k, number of sampling round T, sampling redundancy r, initial sampling budget B_0 and total sampling budget B.
Initialize $S = \emptyset$
Compute $\mathbf{x}_n^{(k)}$ $\forall n \in \mathcal{V}_{train}$ as in (3).

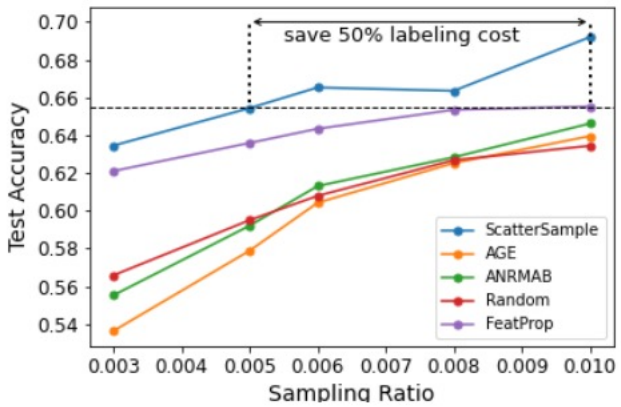
Initial Sampling:
Use k-means++ to cluster $\{\mathbf{x}_n^{(k)}\}$ into B_0 clusters.
Add a node closest to the cluster center per cluster to S.
Query the labels of nodes $v_n \in S$, denoted by $\mathbf{y}_S$.
Train model M using $(\mathbf{y}_S, \mathbf{X}, \mathcal{G})$.

Dynamic Sampling:
Initialize sampling round $t = 1$
while $t < T$ **do**
    Let $B_t = \min(B - |S|, (B - B_0)/T)$
    Use the **DiverseUncertainty** algorithm to select S_t
    Query the labels of S_t, and update $S = S \cup S_t$.
    Train model M over $(\mathbf{y}_S, \mathbf{X}, \mathcal{G})$. Update $t = t + 1$.
end while

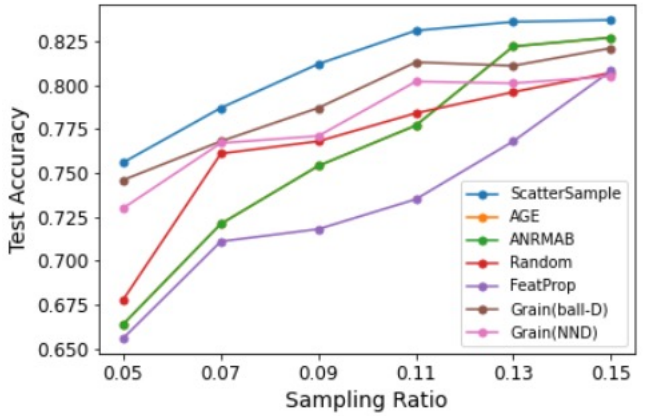
Algorithm 2 DiverseUncertainty Algorithm

Input: $\mathcal{V}_{train}$, $\{\mathbf{x}_n^{(k)} \ \forall n \in \mathcal{C}_t\}$, r, B_t
Compute $\phi_{entropy}(v)$ $\forall v \in \mathcal{V}_{train}$; see 2).
$\mathcal{C}_t \leftarrow \{rB_t$ nodes with largest $\phi_{entropy}(v)\}$.
Use k-means++ to cluster the $\mathbf{x}_n^{(k)}$ (for all $n \in \mathcal{C}_t$) into B_t clusters.
$S_t \leftarrow \emptyset$
for $j = 1, 2, \ldots, B_t$ **do**
    Compute the cluster center $\mathbf{v}_j$ of cluster j
    Pick node $x \leftarrow \arg\min_{n \in \mathcal{C}_t} \left\| \mathbf{x}_n^{(k)} - \mathbf{v}_j \right\|$
    $S_t \leftarrow S_t \cup \{x\}$
end for
Return S_t

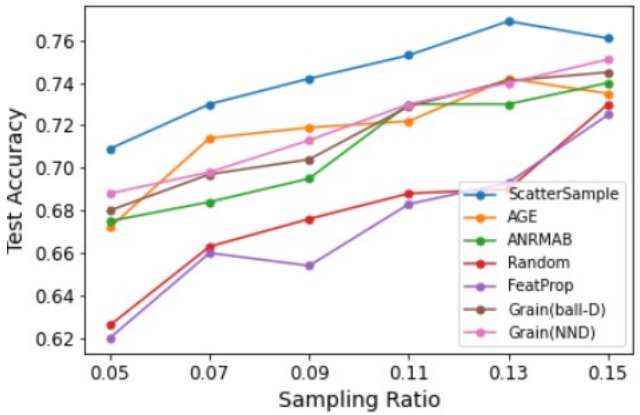
ScatterSample is data efficient



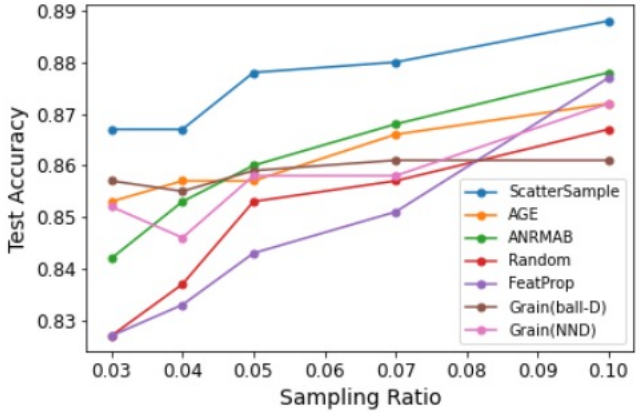
Ogbn-products



Cora



Citeseer



Pubmed

Complementing tabular methods with graphs

- Tree based methods competitive for iid tabular data
- Challenged in graph data that break iid assumption
- *Can we enhance the powerful Gradient Boosted Decision Trees (GBDT) with graph information?*

Complementing tabular methods with graphs

GBDT

Weak learner

$$h^{(t)} = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{m} \left\| -\frac{\partial \ell(\boldsymbol{y}_i, f^{(t)}(\boldsymbol{x}_i))}{\partial f^{(t)}(\boldsymbol{x}_i)} - h(\boldsymbol{x}_i) \right\|_2^2$$

$$f^{(t+1)}(\boldsymbol{x}) = f^{(t)}(\boldsymbol{x}) + \eta^{(t)} h^{(t)}(\boldsymbol{x})$$

- Smoothen the predictions, Laplacian regularizer

$$\ell_Z(\boldsymbol{Z}) \triangleq \|\boldsymbol{Z} - f(\boldsymbol{X}; \boldsymbol{\theta})\|_{\mathcal{F}}^2 + \lambda \mathrm{tr}\left[\boldsymbol{Z}^\top \boldsymbol{L} \boldsymbol{Z}\right]$$

- Extension to Laplacian kernel and robustness promoting regularizers

Graph-Aware Propagation as Bilevel Optimization

- Bilevel boosting solves the joint optimization problem
- Approximate smoothed solution by *k* gradient steps

- Compute *t* weak learners as follows
  - Perform *k* steps of gradient descent (graph propagation)
  - Express the function space gradient of the loss with respect to the *t* graph smoothed predictor
  - Learn a weak learner by approximating the gradient over all nodes

- Convergence guarantees attained
  - As *k* and *t* grow we have convergence to the loss

J. Chen, J. Mueller, V. N. Ioannidis, S. Adeshina, Y. Wang, T. Goldstein, and D. Wipf "Convergent Boosted Smoothing for Modeling Graph Data with Tabular Node Features," ICLR, Int. Conf. in Learn. Rep., May 2022

Tabular datasets

Regression				
Method	**House**	**County**	**VK**	**Avazu**
CatBoost	0.49 ± 0.01	1.30 ± 0.08	7.06 ± 0.22	0.1102 ± 0.02
AutoGluon	0.48 ± 0.01	1.28 ± 0.08	7.04 ± 0.22	0.1106 ± 0.02
BGNN	0.50 ± 0.02	1.29 ± 0.07	6.94 ± 0.21	0.1090 ± 0.01
GAT	0.54 ± 0.01	1.45 ± 0.06	7.22 ± 0.29	0.1134 ± 0.01
GCN	0.63 ± 0.01	1.48 ± 0.08	7.25 ± 0.19	0.1141 ± 0.02
AGNN	0.59 ± 0.01	1.45 ± 0.08	7.26 ± 0.20	0.1134 ± 0.02
APPNP	0.69 ± 0.01	1.50 ± 0.11	13.23 ± 0.12	0.1127 ± 0.01
CatBoost+ (ours)	0.55 ± 0.01	1.31 ± 0.12	6.92 ± 0.22	0.1080 ± 0.01
AG+ (ours)	**0.45 ± 0.01**	1.14 ± 0.06	**6.90 ± 0.22**	0.1075 ± 0.01
EBBS (ours)	**0.45 ± 0.01**	**1.13 ± 0.09**	6.95 ± 0.22	**0.1058 ± 0.01**

Classification				
Method	**Slap**	**DBLP**	**Coauthor-CS**	**Coauthor-Phy**
CatBoost	**96.3 ± 0.0**	91.3 ± 1.0	91.9 ± 0.2	94.6 ± 0.2
AutoGluon	96.2 ± 0.0	91.5 ± 1.0	92.1 ± 0.8	95.0 ± 0.0
BGNN	95.0 ± 0.0	88.9 ± 1.0	92.5 ± 0.2	96.4 ± 0.0
GAT	80.1 ± 1.0	80.2 ± 1.0	91.55± 0.5	95.4 ± 0.1
GCN	87.8 ± 1.0	42.8 ± 4.0	92.32 ± 0.4	95.4 ± 0.1
AGNN	89.2 ± 1.0	79.4 ± 1.0	92.65 ± 0.4	**96.9** ± 0.1
APPNP	89.5 ± 1.0	83.0 ± 2.0	93.2 ± 0.1	96.6 ± 0.2
AG+ (ours)	96.1 ± 0.3	**92.0 ± 1.0**	93.3 ± 0.4	96.6 ± 0.1
EBBS (ours)	**96.3 ± 0.4**	91.3 ± 0.6	**94.9 ± 0.2**	**96.9 ± 0.2**
SOTA	-	-	94.8[1]	**96.9**[2]

- AG+ : AutoGluon + smoothing
- Can we use more powerful learning algorithms?
    - Ongoing efforts to integrate GNN models with AutoGluon

Conclusion

- Ongoing research efforts
  - Modeling multi-modal (text, image) data with GNNs
  - Answering complex and temporal questions over KGs
  - Designing GNN models for life-science applications (protein, RNA, gene modeling)
  - Joint training of large language models and GNNs
  - Distillation in GNNs
- Products
  - (AWS) Neptune ML : A graph ML platform powered by GNNs for the Neptune database
  - (Amazon) Large scale representation learning for Amazon entities and relations

Thank you