Mixture of Scope Experts at Test: Generalizing Deeper Graph Neural Networks with Shallow Variants

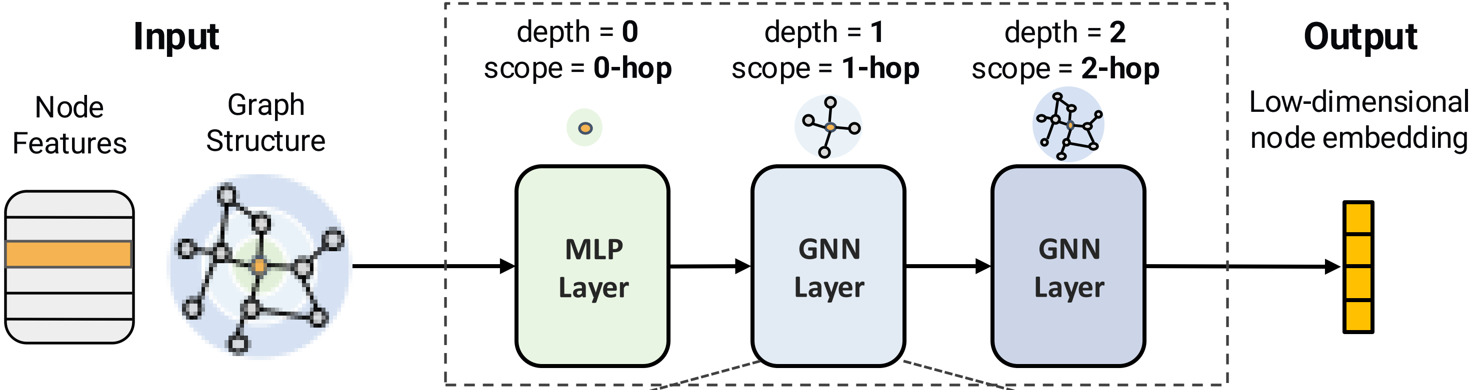
Gangda Deng[1], Hongkuan Zhou[1], Rajgopal Kannan[2], Viktor Prasanna[1]

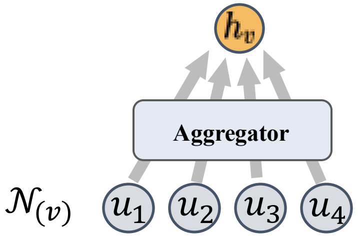
1: University of Southern California, Los Angeles, USA
2: DEVCOM ARL Army Research Office, Los Angeles, USA

gangdade@usc.edu, prasanna@usc.edu

Graph Neural Networks (GNNs)

The depth of a GNN model is coupled with its scope: Each GNN layer contains a neighbor aggregation function. L times of aggregation can perceive the entire L-hop neighborhood.





$$\mathbf{h}_v^{(\ell+1)} = \bigoplus_{u \in \mathcal{N}_{(v)}} \phi^{(\ell)}\left(\mathbf{h}_v^{(\ell)}, \mathbf{h}_u^{(\ell)}\right)$$

$\phi^{(\ell)}$: differentiable function such as **Linear Layer**

$\bigoplus_{u \in \mathcal{N}_{(v)}}$: neighbor aggregation function, e.g., **sum or mean**

The Long-lasting Depth Dilemma: Deeper GNNs struggle with generalization

(Step 1)

Deeper GNNs are desirable: increasing the model depth can exponentially incorporate more information.

Performance degradation is widely observed when going deep: number of layers > 3

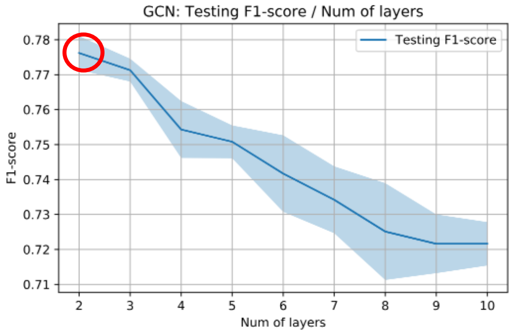
(Step 2)

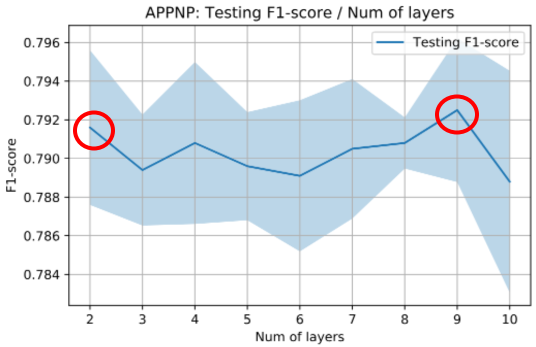
Existing solutions for deeper GNNs can **only alleviate the degradation** and achieve only **marginal gains** over their shallow variants.

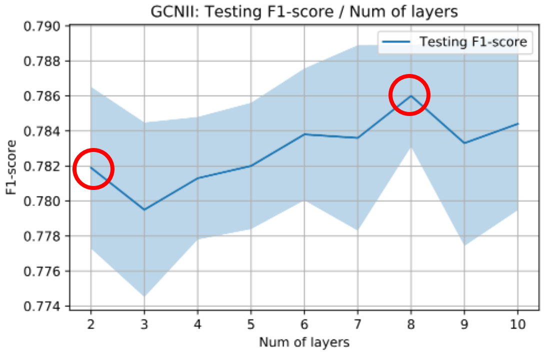
GCN, depth=2, acc=77.6%

APPNP, depth=9, acc=79.2%

GCNII, depth=8, acc=78.6%







Comparison of F1-score on the Citeseer dataset. Images from the paper "On Provable Benefits of Depth in Training Graph Convolutional Networks".

Error Decomposition for Deeper GNNs

Unlike previous studies that attribute the failure of deeper GNNs to a **single cause**, we argue that this failure stems from **multiple factors**, varying based on GNN architectures.

Current techniques face inherent trading-offs between these three types of errors.

True Risk

Minimized Empirical Risk

Empirical Risk

$$R(h_S) = \underbrace{\hat{R}_S(h_{S,ERM})}_{\text{representation error}} + \underbrace{\hat{R}_S(h_S) - \hat{R}_S(h_{S,ERM})}_{\text{optimization error}} + \underbrace{R(h_S) - \hat{R}_S(h_S)}_{\text{generalization error}}$$

Test Error Remain High for various deeper GNNs	Expressivity of Architecture	Training Difficulties	Generalization Gap
	Loss of Expressivity -> Over-smoothing Issue	Vanishing/Unstable Gradients -> Model Degradation	Powerful Models meet non-uniform patterns -> Overfitting Issue

A Subgroup Generalization View to Explain Deeper GNNs' Failure

Assumptions:
The graph is composed of non-overlapping **node subgroups**, with each subgroup containing nodes with the same homophily ratio.

Properties:
- Generalization Error for subgroup m depends on the **aggregated feature distance** and **homophily ratio difference**
- The minimum generalization error occurs at **different depths L** for subgroups i and j where $p_i > p_S$ and $p_j > p_S$
- Varying L yields a **larger generalization disparity on heterophilous** graphs than on homophilous graphs

Theorem 3.3 (GNN Subgroup generalization bound). *Assume the aggregated features $g^L(\mathbf{X}, \mathcal{G})$ share the same variance $\sigma^2\mathbf{I}$. Let θ be any classifier in the parameter set $\{\mathbf{W}^{(l)}\}_{l=1}^{L'}$ and S denote the training set. For any test subgroup $m \in \{1, \cdots, M\}$ and large enough number of the training nodes $N_S = |\mathcal{V}_S|$, with probability at least $1-\delta$ over the sample $\{y_v\}_{v \in V_S}$, there exists $0 < \alpha < \frac{1}{4}$ we have:*

$$\mathcal{L}_m^0(\theta) - \widehat{\mathcal{L}}_S^\gamma(\theta) \leq \mathcal{O}\left(\frac{\rho}{\sigma^2}\left(\boxed{\epsilon_m} + \rho\boxed{(p_S - p_m)}\boxed{\Gamma_{L-1}}\right)\right)$$

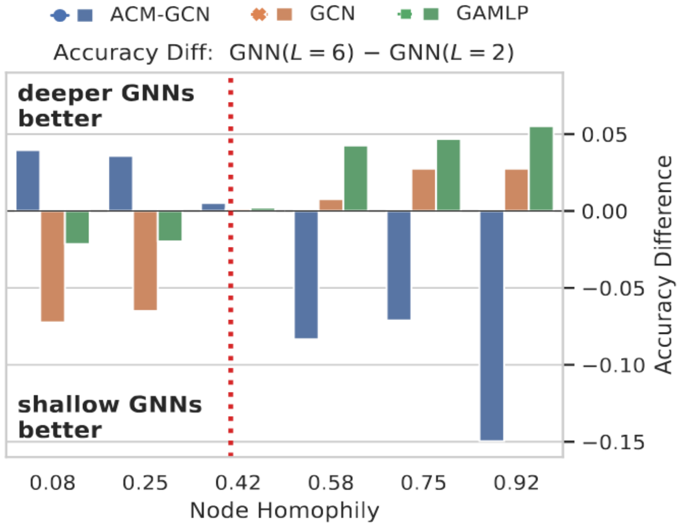
$$+\mathcal{O}\left(\frac{\|W\|^2 (\epsilon_m)^{2/L'}}{N_S^\alpha}\right) + \mathcal{O}\left(\frac{\ln(1/\delta)}{N_S^{2\alpha}}\right), \quad (2)$$

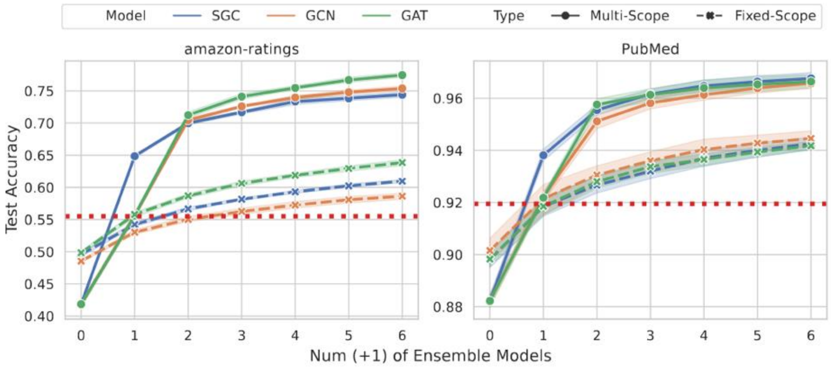
where $\rho := \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ is feature distribution separability, $\epsilon_m := \max_{u \in V_m} \min_{v \in V_S} \left\|g^L(\mathbf{X}, \mathcal{G})_u - g^L(\mathbf{X}, \mathcal{G})_v\right\|_2$ is the bound of the aggregated feature distance. $\Gamma_{L-1} := \mathbb{E}_{o \sim \Pr(o), o \in \{1, \ldots, M\}}\left[(p_o - q_o)^{L-1}\right]$ represents L-hop homophily coefficient, and $\|W\|^2 := \sum_{l=1}^{L'} \|\widetilde{W}_l\|_F^2$.

USC

Deeper GNNs Exhibit Generalization Preference Shift

- Increasing GNN depth enhances generalization for specific subgroups but inevitably **compromises generalization for others**
- Different depths of GNN can correctly predict a unique subset of nodes
- The generalization disparity of <u>models with different depths</u> is **significantly larger** than <u>models with the same depth</u> but different random seeds.

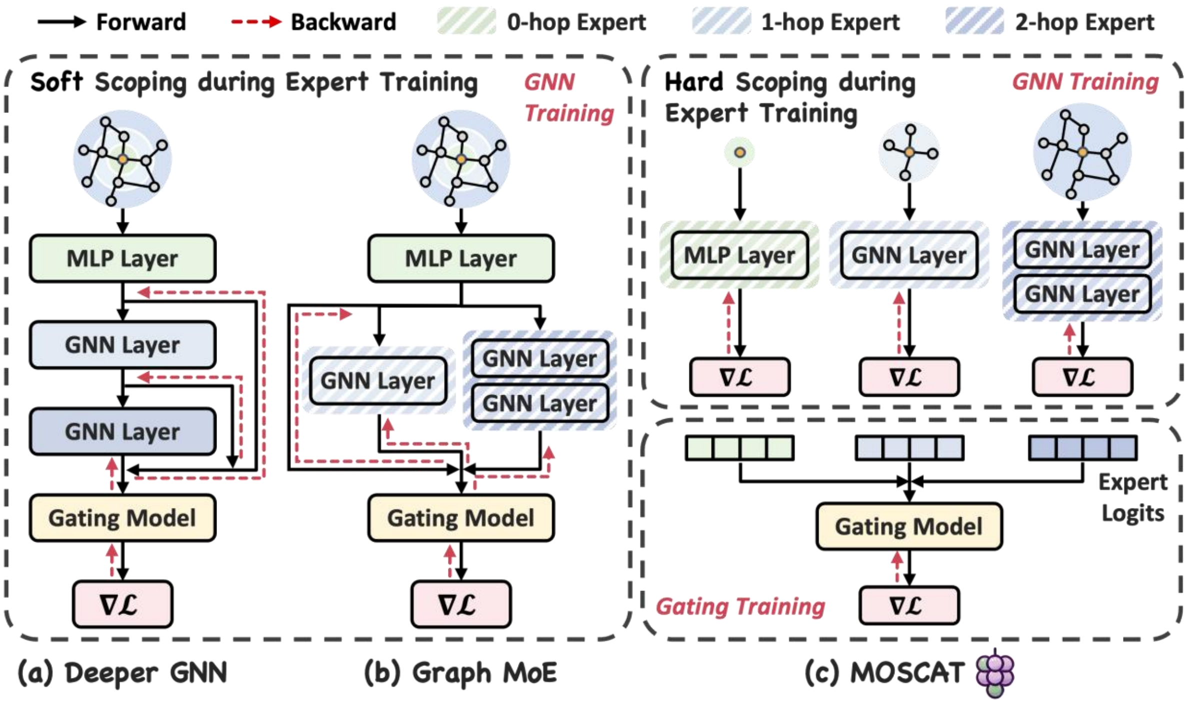




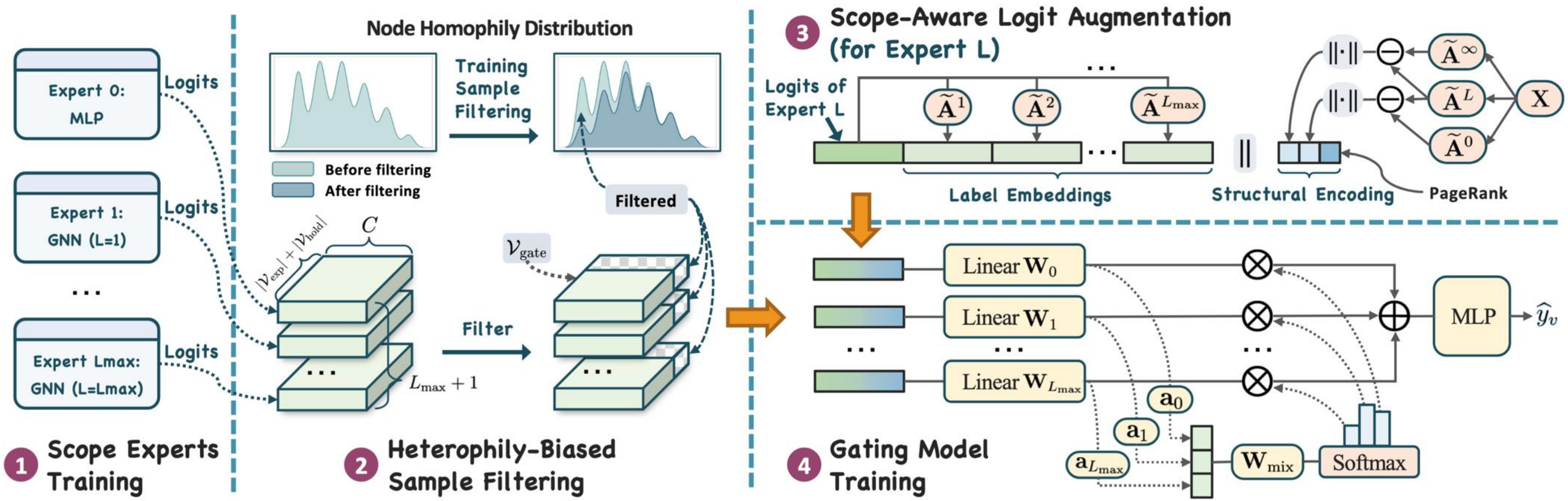
Test accuracy under *Oracle* model ensemble.

The overlapping ratio on Penn94 dataset.

Effective Remedy Paradigm: Mixing Deeper GNNs with their Shallow Variants to Improve Generalization



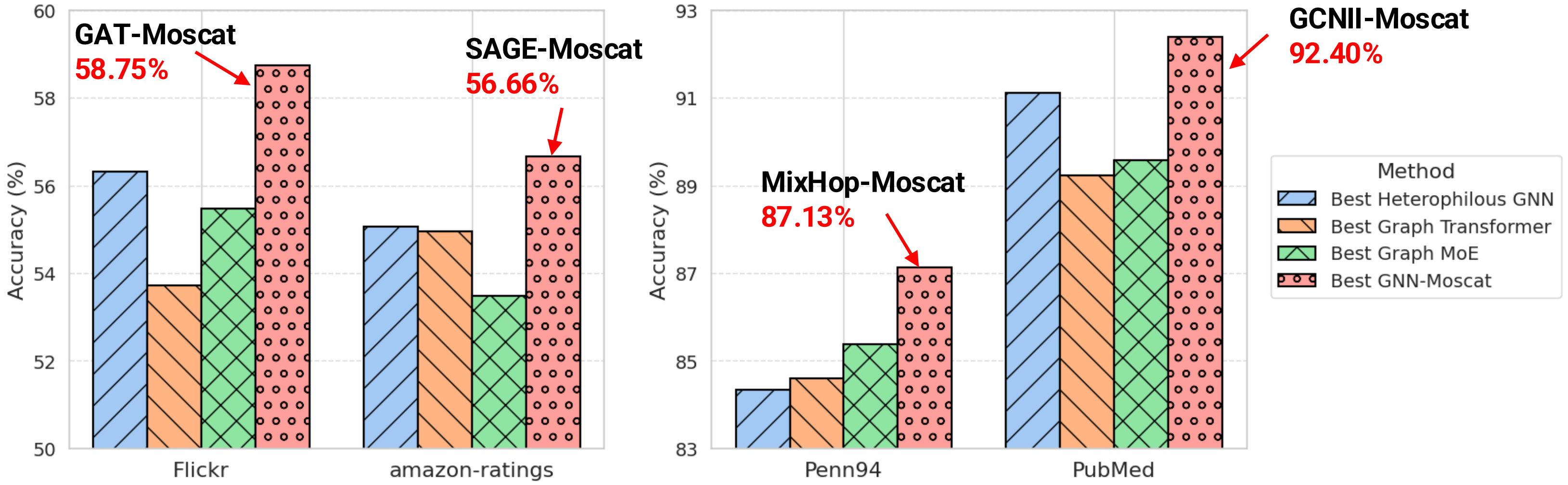
Proposed Method – Moscat: Mixture of Scope Experts at Test



Moscat Achieves New SOTA Performance

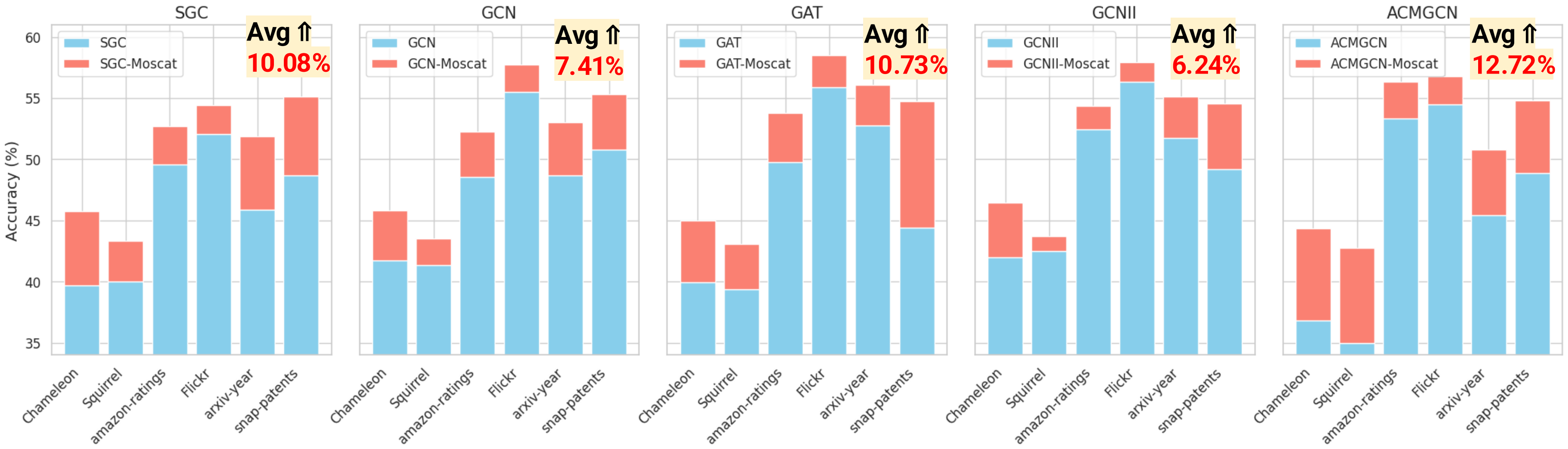
Baselines

- Heterophilous GNNs: GNNs designed for heterophilous graphs
- Graph Transformers: Applies global attention to enable a global scope
- Graph MoE: Mixture of multiple GNN experts including various of depth



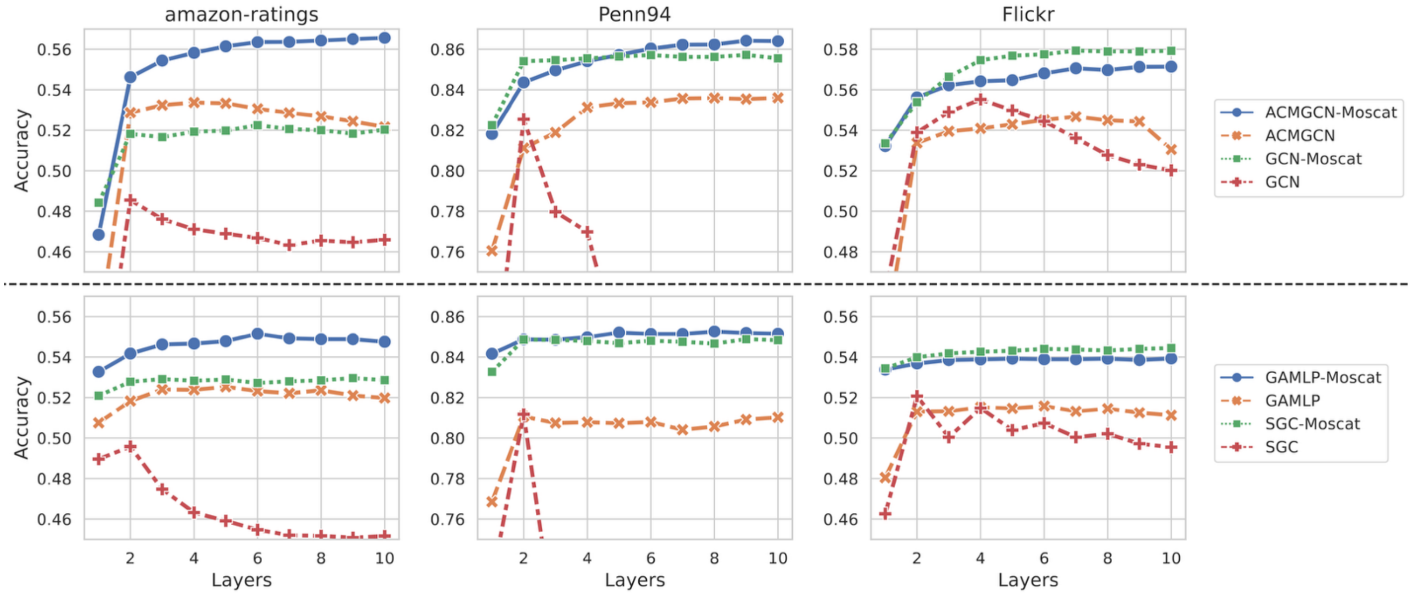
Moscat Strongly Improves GNNs with Diverse Architectures

- Significant Improvement: 6.24% ~ 12.72% improvements averaging on 6 datasets
- GCNII-Moscat achieves the **smallest gain**: GCNII aims to **avoid overfitting**, limiting Moscat's impact
- ACMGCN-Moscat achieves the **largest gain**: ACM-GCN is **more expressive** and is prone to overfitting



Moscat for Deeper GNNs

- **With a large depth**, GNN-Moscat outperforms GNN + other techniques designed (e.g., skip connections) for deeper GNNs
- Moscat can also apply on GNNs with skip connections to better leverage the depth and achieve further accuracy improvements



THANKS!

Code	Paper	Personal Website

USC