SUPPLEMENTARY MATERIAL

GNN models We follow the best hyperparameter and training settings given by the corresponding graph neural network papers, using the PyTorch-Geometric implementation[1] of GCN [**?**], GAT [**?**], SGC [**?**], and APPNP [**?**]. For SGC, we set the learning rate to 0.2, train for 100 epochs without early stopping, and tune weight decay for 60 iterations using Hyperopt[2], according to the original paper. We implement gfNN [**?**] by ourselves and follow the setting in the original paper. In order to capture the variance across different training runs, each model is run for 100 times, and we report the averaged results with standard deviations.

Width We conduct experiments on the influence of width on the models GCN and GAT, where we use the best hyperparameter settings and vary the hidden dimensions per layer in the range given by $\{2^i \mid 3 \leq i \leq 10\}$. Dropout layers are removed, and the number of epochs is set to 200 with early stopping after 10 epochs without improvement of the validation loss. Each model is run for 10 times.

Depth We conduct experiments on the influence of depth on the models GCN and GAT, where we follow the experimental setting in Appendix B by Kipf and Welling [**?**]. The number of layers is in the range $\{1, 2, \ldots, 10\}$. Each model is run for 10 times.

Graph density We conduct experiments on the influence of graph density on the models GCN and GAT. Different proportions of edges are removed randomly from 0% (original dataset) to 100% (no graph structure at all). Each model is run for 10 times.

New loss function We follow the setting by Tomani and Buettner [**?**], who also introduced an ECE-inspired loss function. An annealing coefficient is specified for the calibration error term since the early epochs are usually used for reaching the cross entropy minimum. More precisely, we define

$$\text{anneal\_coef} = \lambda \cdot \min\left\{1, \frac{\text{epoch}}{\text{EPOCHS} \cdot \text{anneal\_max}}\right\}, \quad (1)$$

$$\tilde{L}_{\text{cal}} = \text{anneal\_coef} \cdot L_{\text{cal}}\,, \quad \text{and} \quad (2)$$

$$L = \alpha \cdot L_{\text{ce}} + (1 - \alpha) \cdot \tilde{L}_{\text{cal}}\,, \quad (3)$$

where epoch and EPOCHS are the current training epoch and the total number of epochs, respectively, and λ, anneal_max, and α are hyperparameters. We set $\text{anneal\_max} = 1$, $\lambda = 1$, and tune α on the validation set in the range $\{0.95, 0.96, 0.97, 0.98, 0.99\}$. Each model (fixed hyperparameter setting) is run for 10 times.

[1]https://github.com/pyg-team/pytorch_geometric/tree/master/benchmark/citation

[2]https://github.com/hyperopt/hyperopt

TABLE I

CALIBRATED ACCURACY (MEAN±SD OVER 100 INDEPENDENT RUNS). TEMPERATURE SCALING AND RBS DO NOT CHANGE THE ACCURACY.

Dataset	Model	Uncal.	His	Iso	BBQ	Meta
Cora	GCN	81.43±0.60	80.38±0.82	81.80±0.57	81.34±0.67	79.23±1.61
	GAT	83.14±0.39	81.39±0.48	84.05±0.51	83.52±0.59	79.99±1.70
	SGC	81.19±0.05	79.91±0.13	81.16±0.11	79.83±0.24	78.77±1.88
	gfNN	78.73±5.04	79.06±1.16	80.21±1.28	79.96±1.31	76.30±5.51
	APPNP	83.68±0.36	82.52±0.46	83.45±0.45	83.20±0.53	81.33±1.79
Citeseer	GCN	71.32±0.70	71.93±0.84	72.39±0.66	71.79±0.99	68.22±4.13
	GAT	70.99±0.60	71.78±0.56	72.21±0.52	71.81±0.70	68.28±2.63
	SGC	72.46±0.15	74.13±0.05	73.81±0.12	73.32±0.13	69.19±2.08
	gfNN	67.33±6.58	71.74±1.22	71.98±1.15	71.10±1.22	64.63±6.61
	APPNP	72.10±0.38	72.94±0.48	72.90±0.48	72.63±0.83	69.64±2.29
Pubmed	GCN	79.23±0.43	79.01±0.55	79.03±0.46	78.85±0.67	76.99±4.62
	GAT	79.05±0.38	78.50±0.61	78.85±0.38	78.19±0.56	78.00±1.43
	SGC	78.72±0.04	79.05±0.08	79.30±0.03	79.88±0.19	77.83±1.57
	gfNN	77.94±2.32	77.92±1.11	78.16±0.98	77.76±1.33	75.66±3.05
	APPNP	80.09±0.25	80.12±0.44	80.15±0.30	79.50±0.47	78.37±1.64

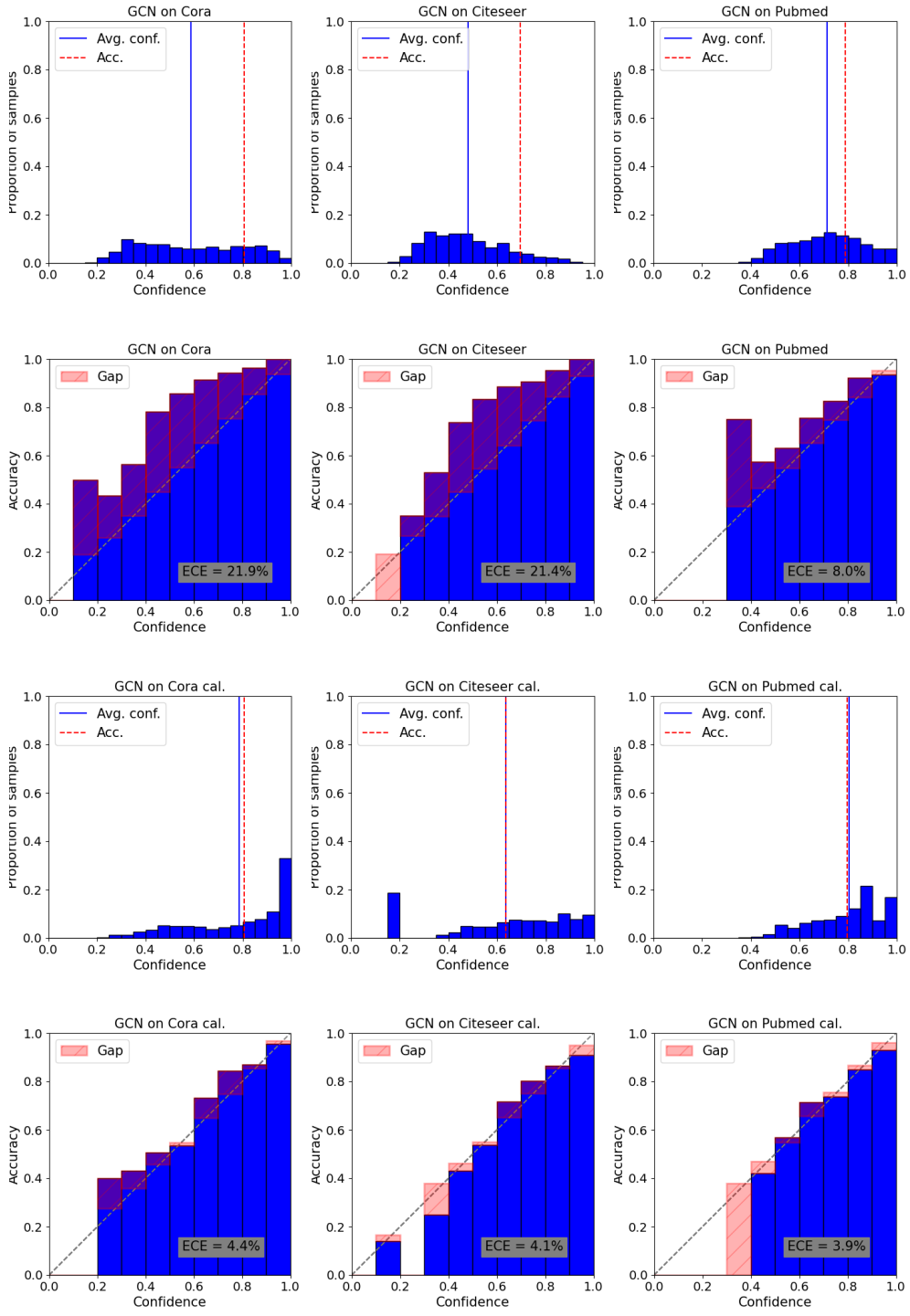

Fig. 1. Histograms and reliability diagrams for GCN.

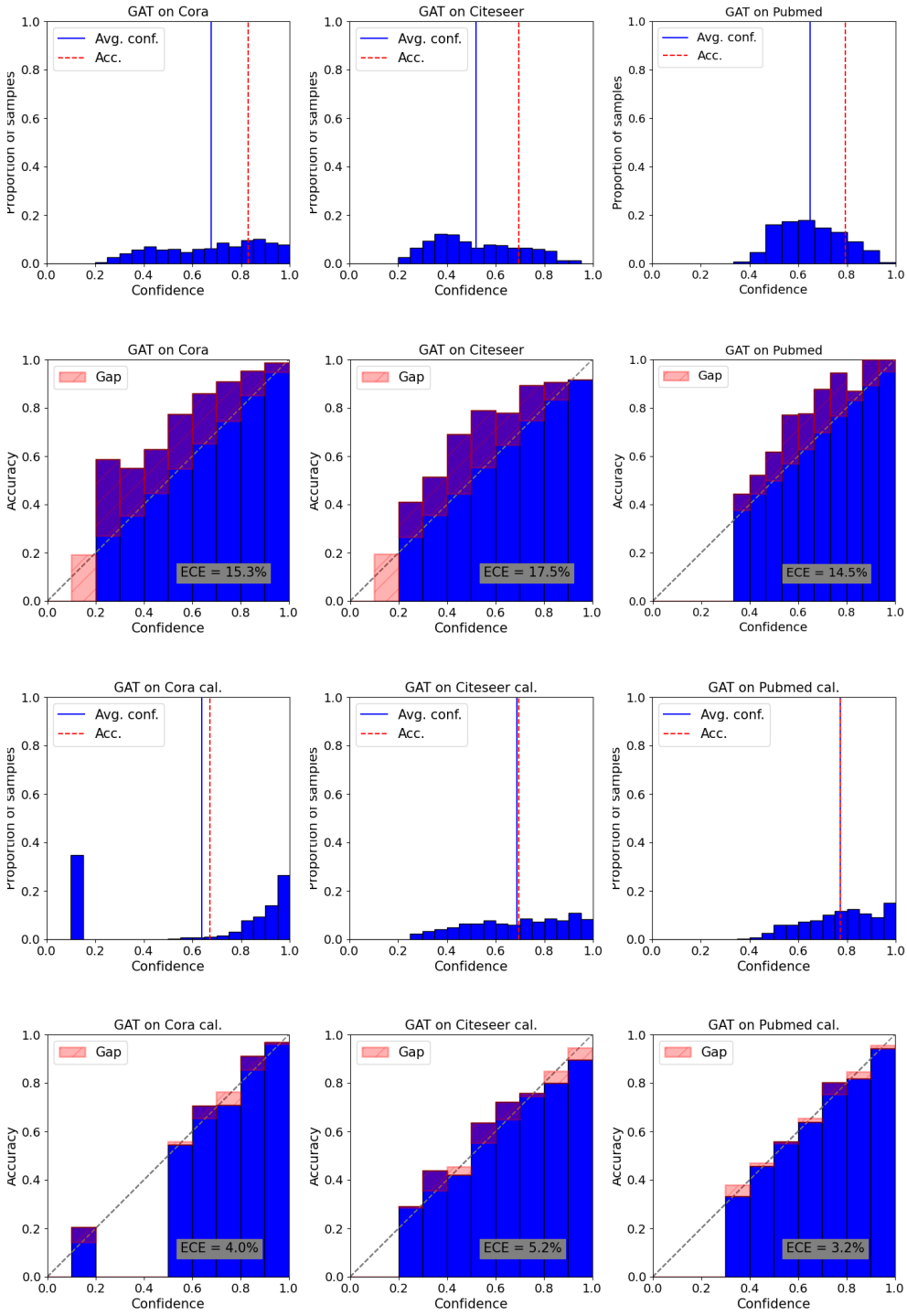

Fig. 2. Histograms and reliability diagrams for GAT.

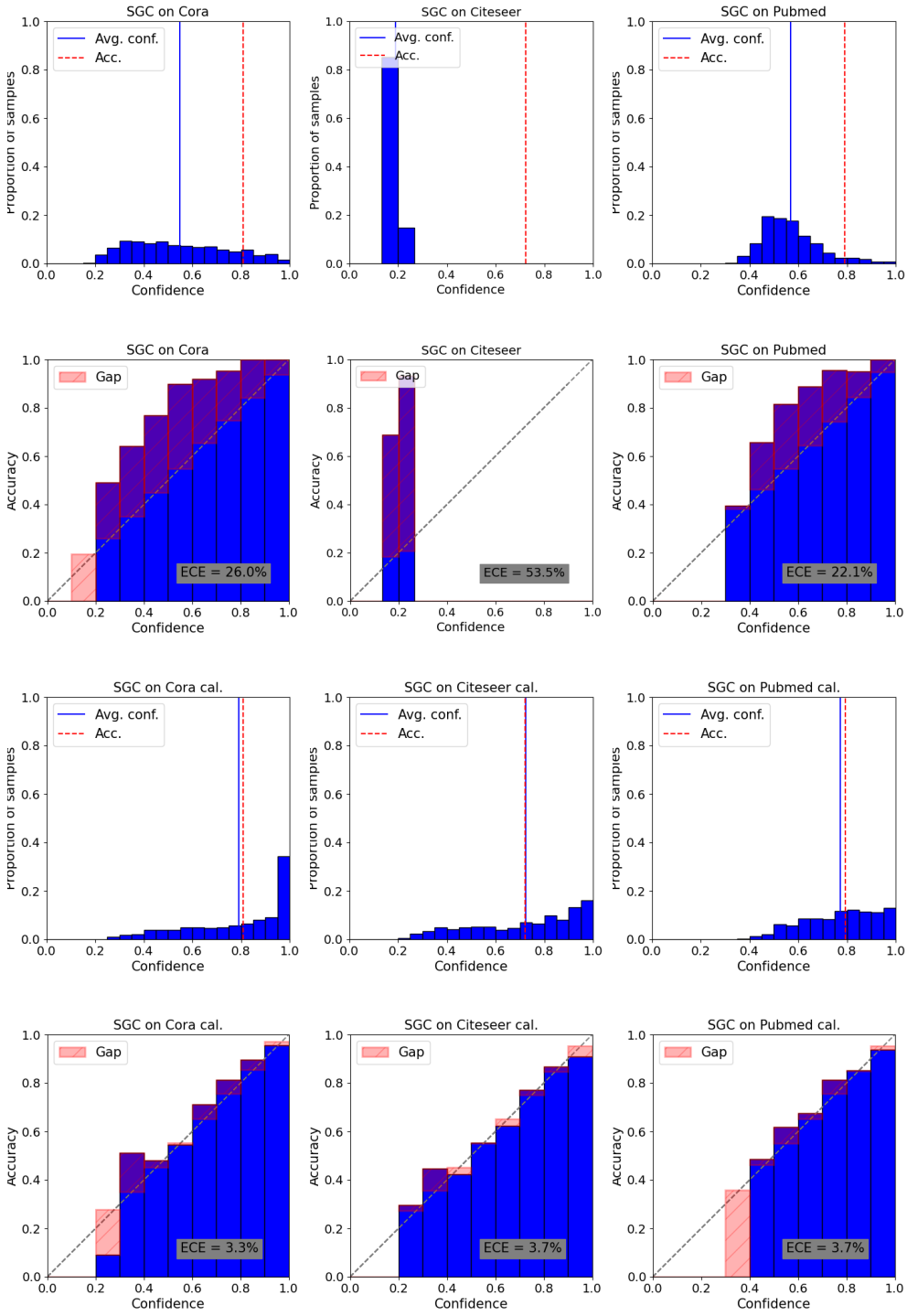

Fig. 3. Histograms and reliability diagrams for SGC.

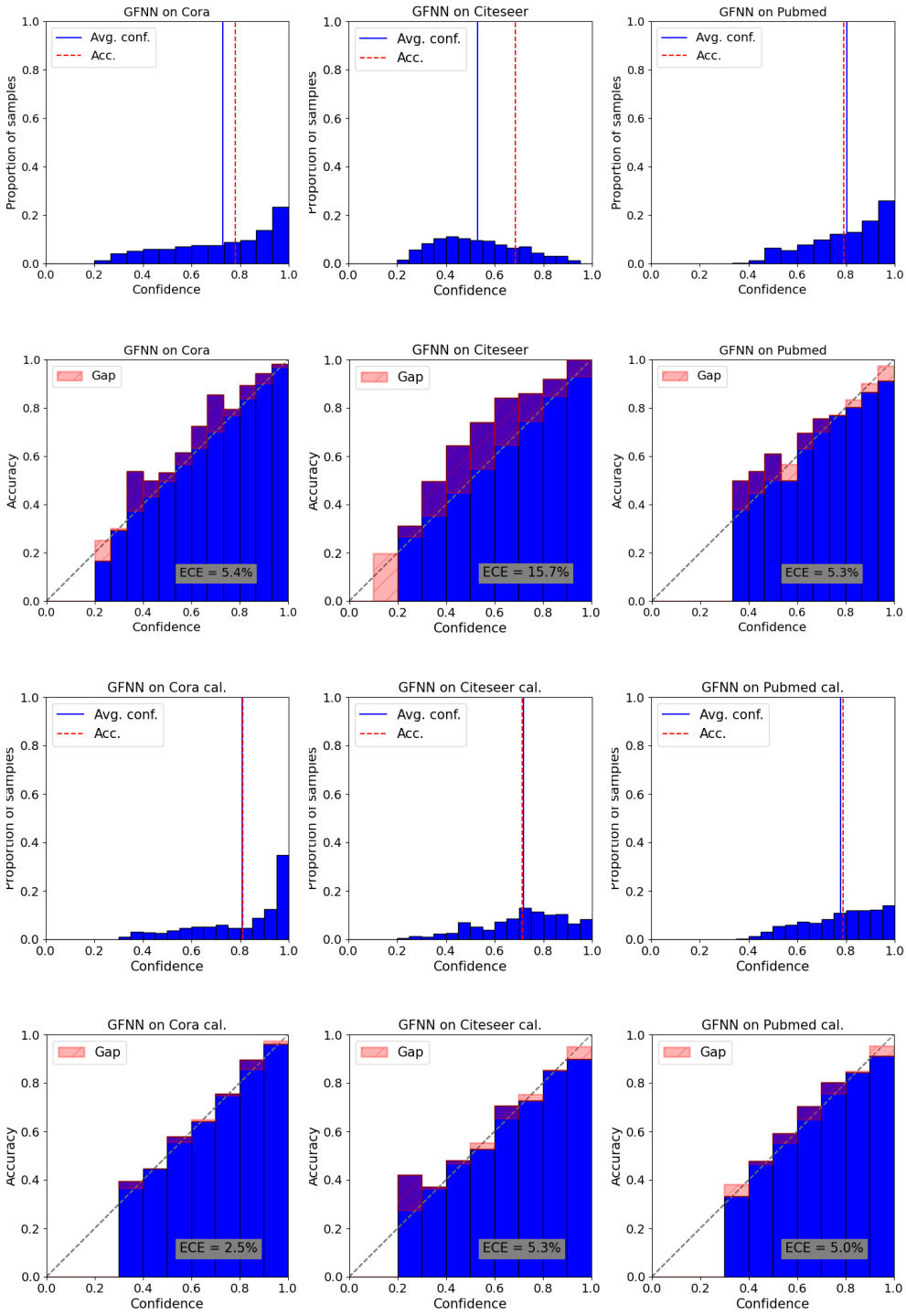

Fig. 4. Histograms and reliability diagrams for gfNN.

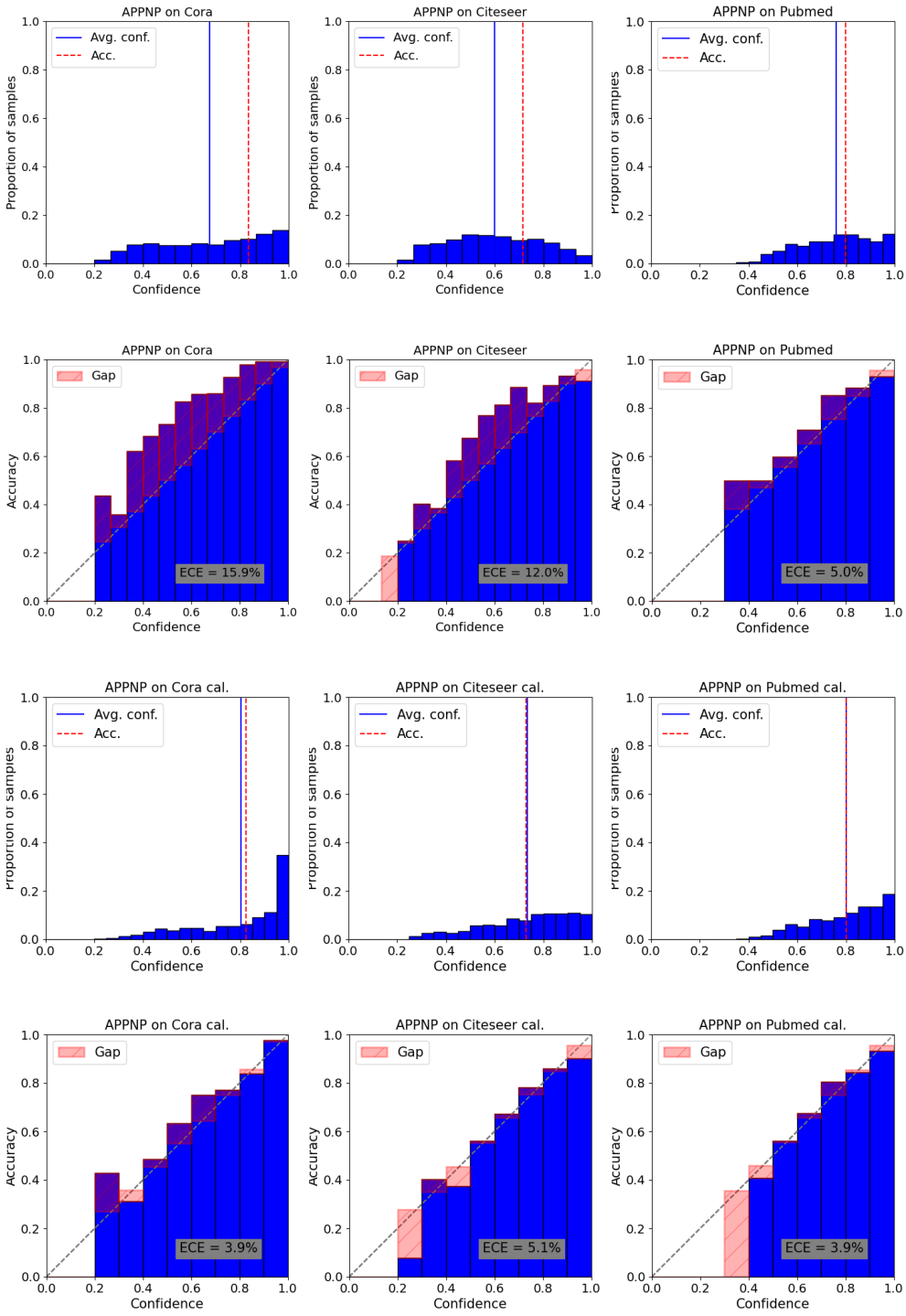

Fig. 5. Histograms and reliability diagrams for APPNP.

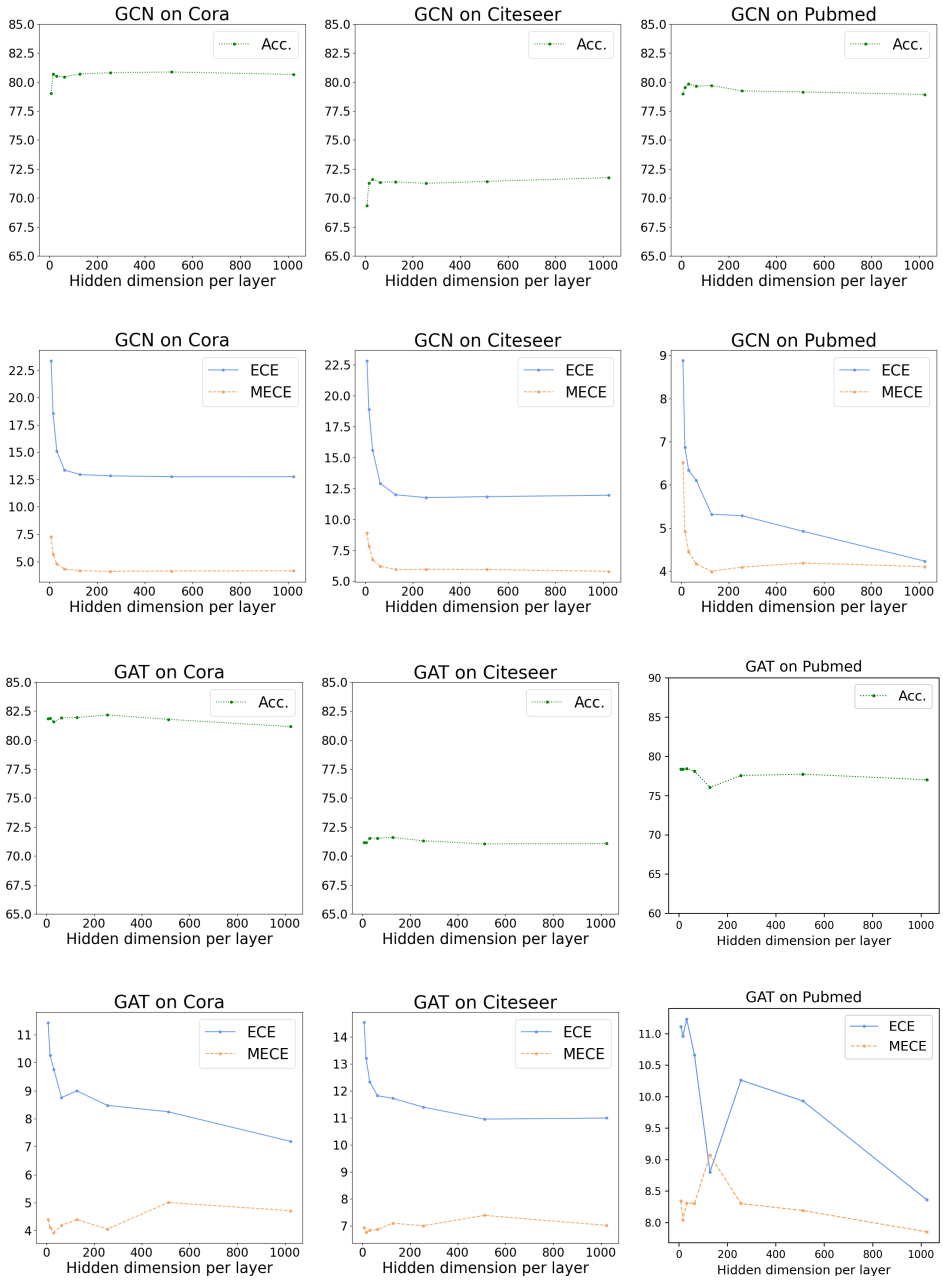

Fig. 6. Influence of width.

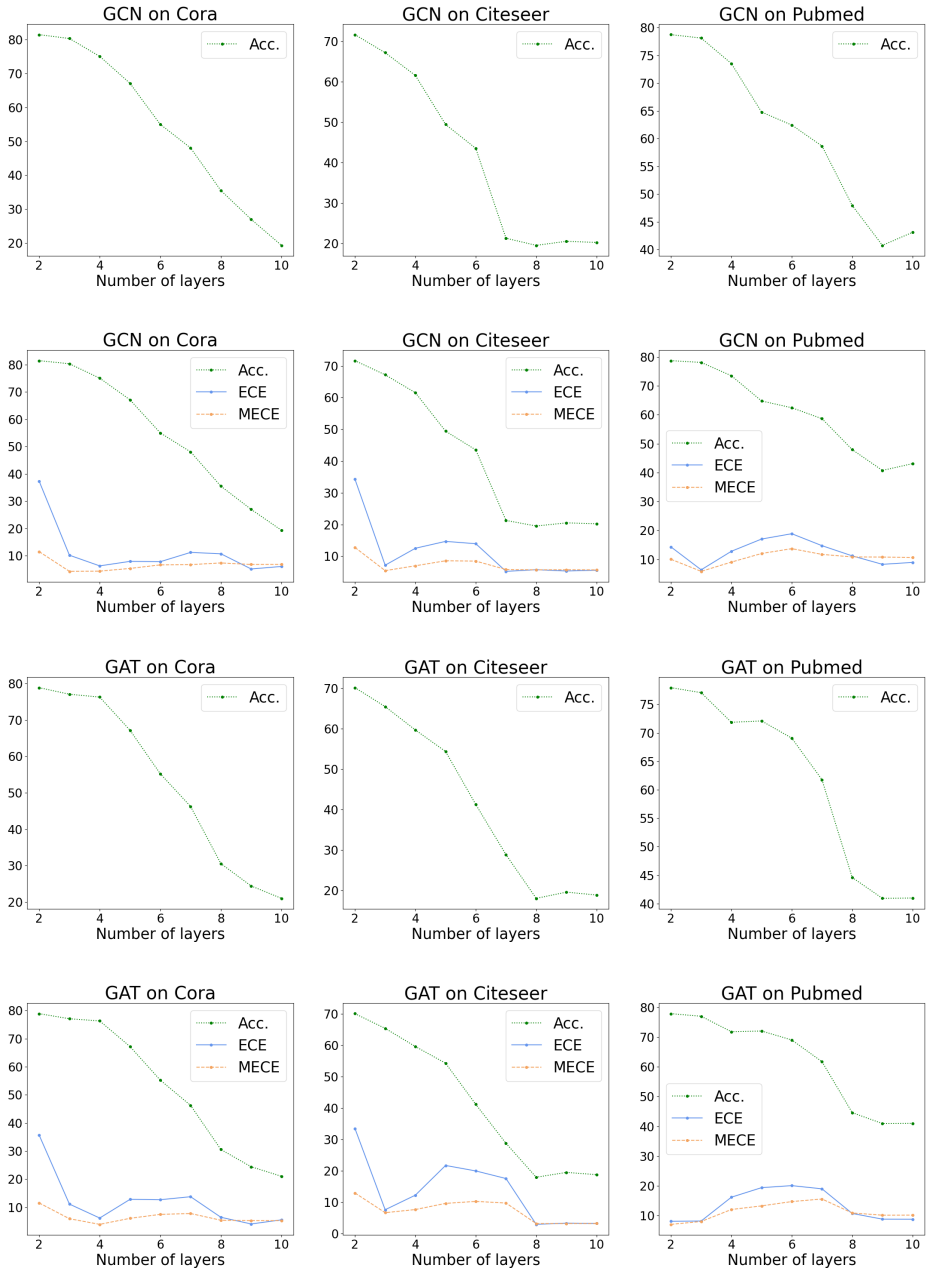

Fig. 7. Influence of depth.

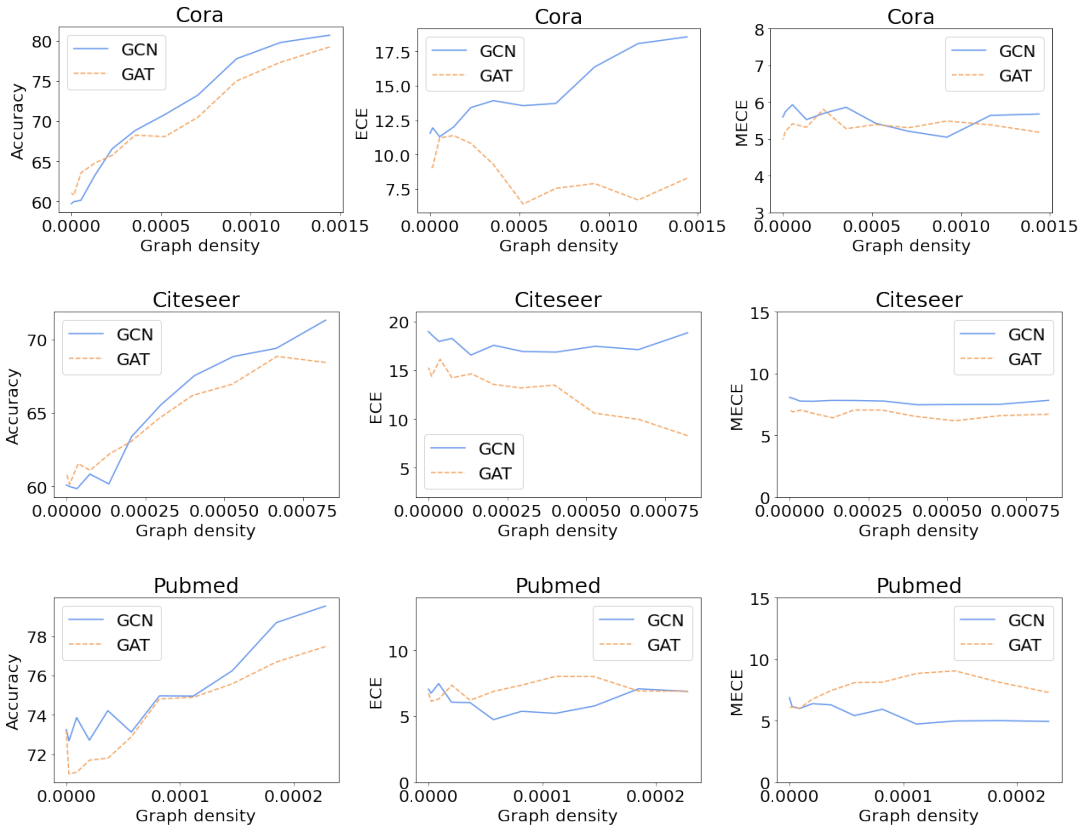

Fig. 8. Influence of graph density.