university of groningen / faculty of science and engineering / bernoulli institute

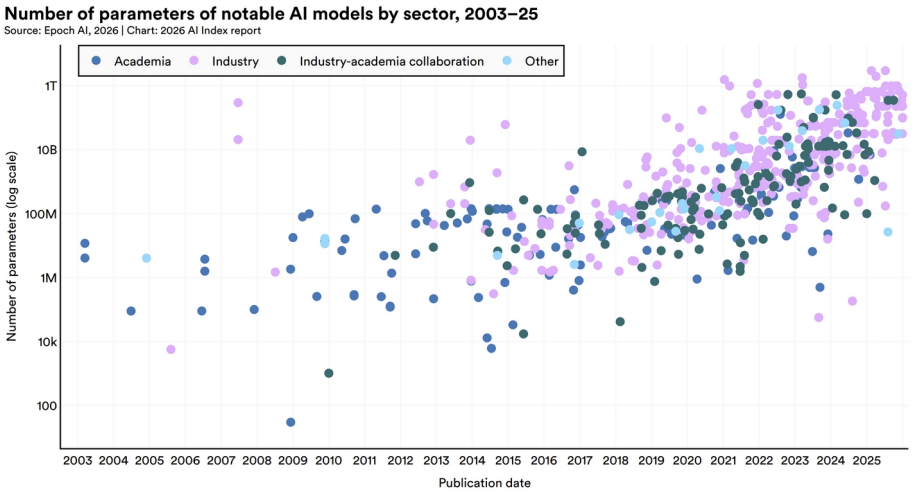
Foundation Models for Scalable and Sustainable AI

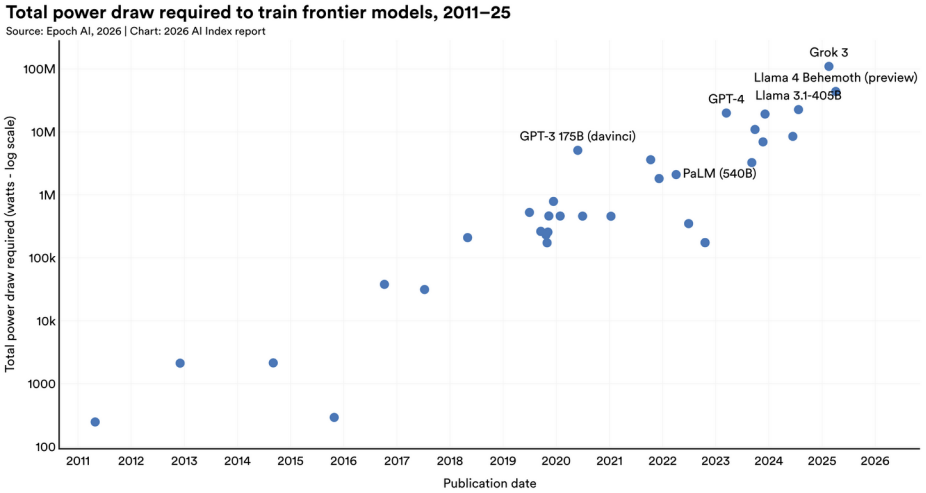
Andrés Tello

SummerSoC
Service Oriented Computing

June 29th - July 4th, 2026
Crete, Greece

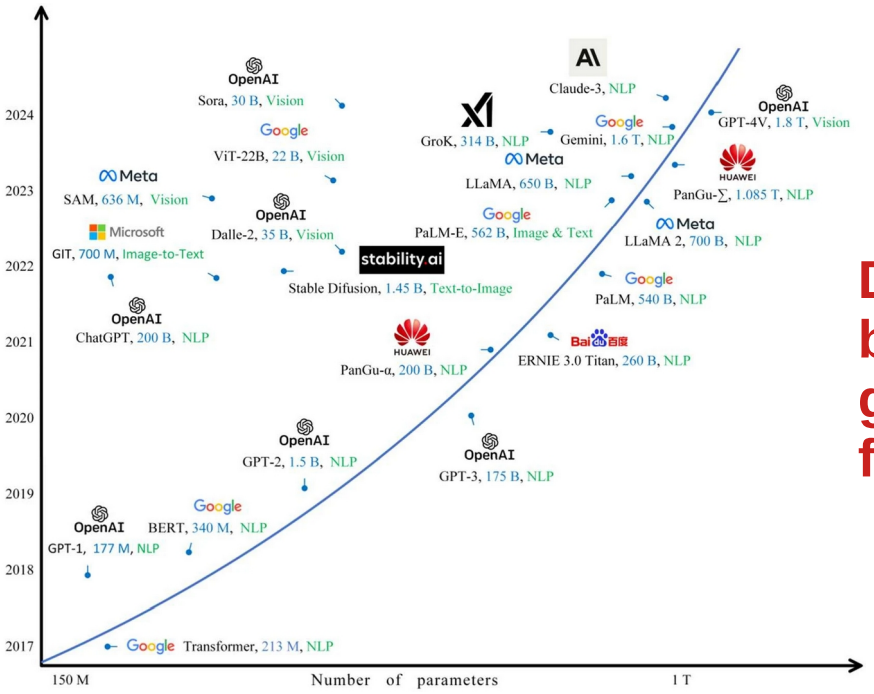
Foundation Models





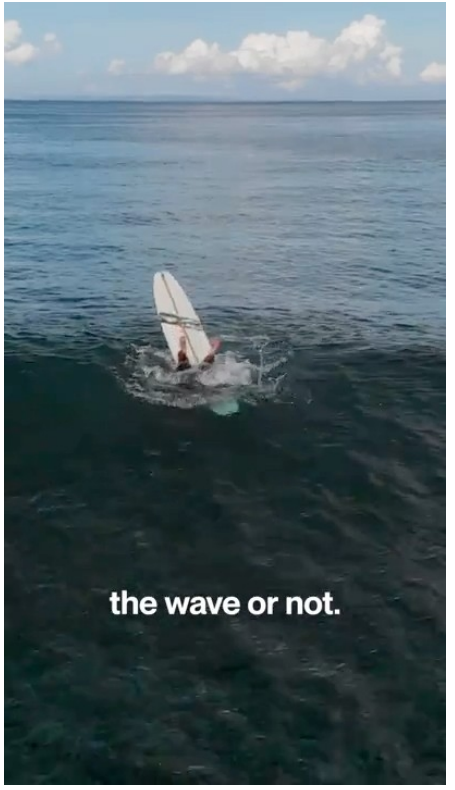
Sajadieh et al., “The AI Index 2026 Annual Report”

Foundation Models



Tu, et al., An overview of large AI models and their applications, 2024

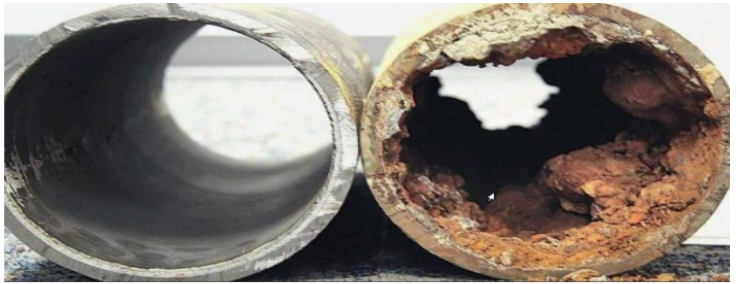
Do we really need billion-parameter general-purpose models for every domain?



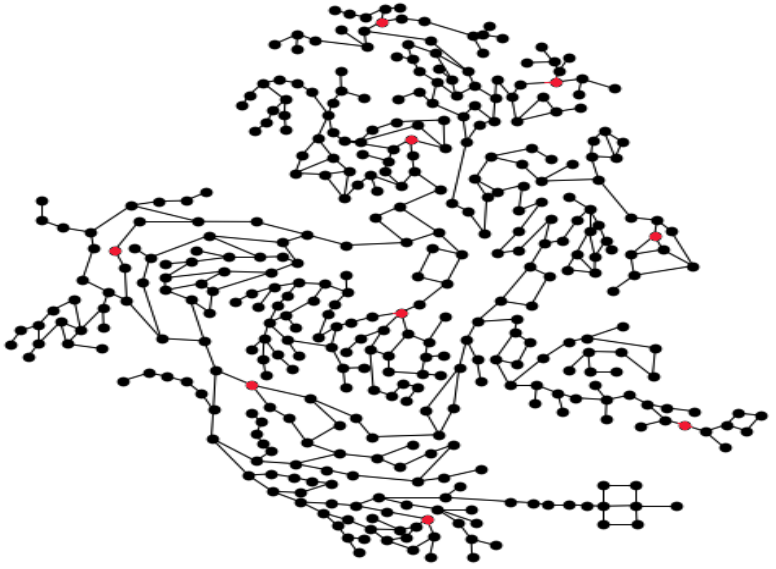
FMs design/use according the constraints of each domain.

Dutch drinking water network

~ 121,500 km of pipes



Foundation Models for WDNs



≠

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since 1966, when designers at Letraset and James Mosley, the librarian at St Bride Printing Library in London, took a 1914 Cicero translation and scrambled it to make dummy text for Letraset's Body Type sheets. It has survived not only many decades, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised thanks to these sheets and more recently with desktop publishing software like Aldus PageMaker and Microsoft Word including versions of Lorem Ipsum.

(a) ImageNet Synset: One sample image from each category (b) Corel-1000 Dataset: Sample images from each category

(c) Caltech-256 Dataset: One sample image from each category (d) Caltech-101 Dataset: One sample image from each category

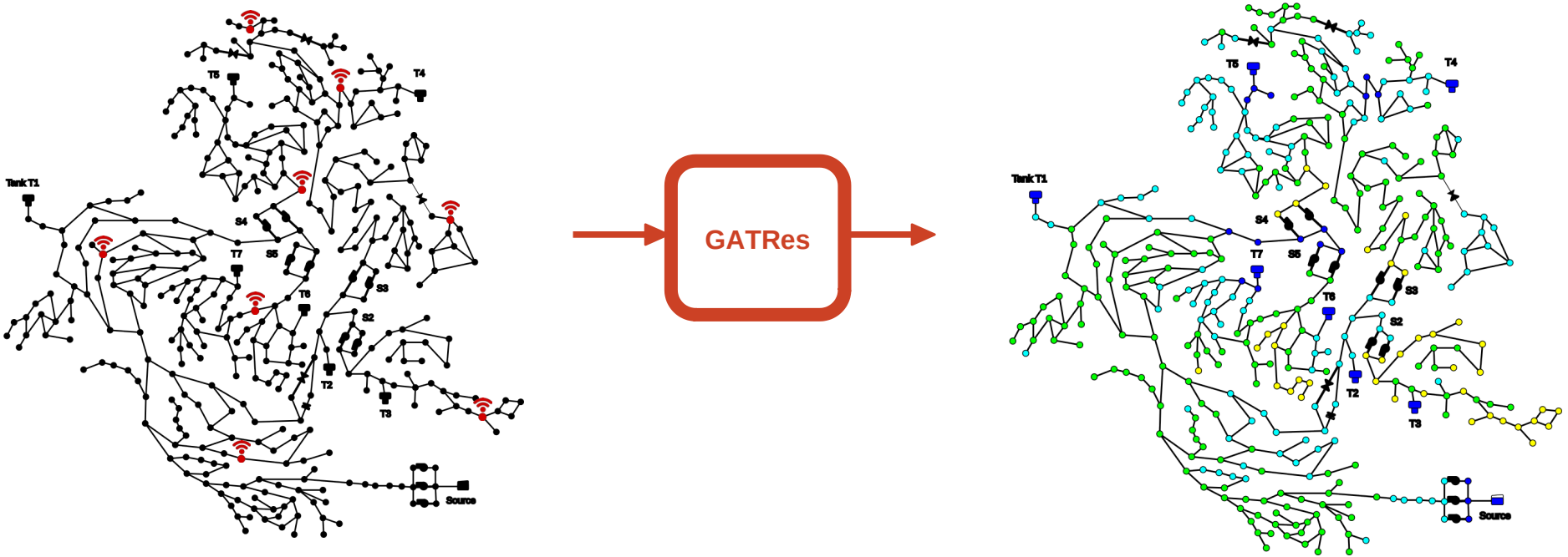
Table 5

Model Comparison in the Noisy Test Performed on 24-Hour Oosterbeek Water Distribution Network at 95% Masking Rate

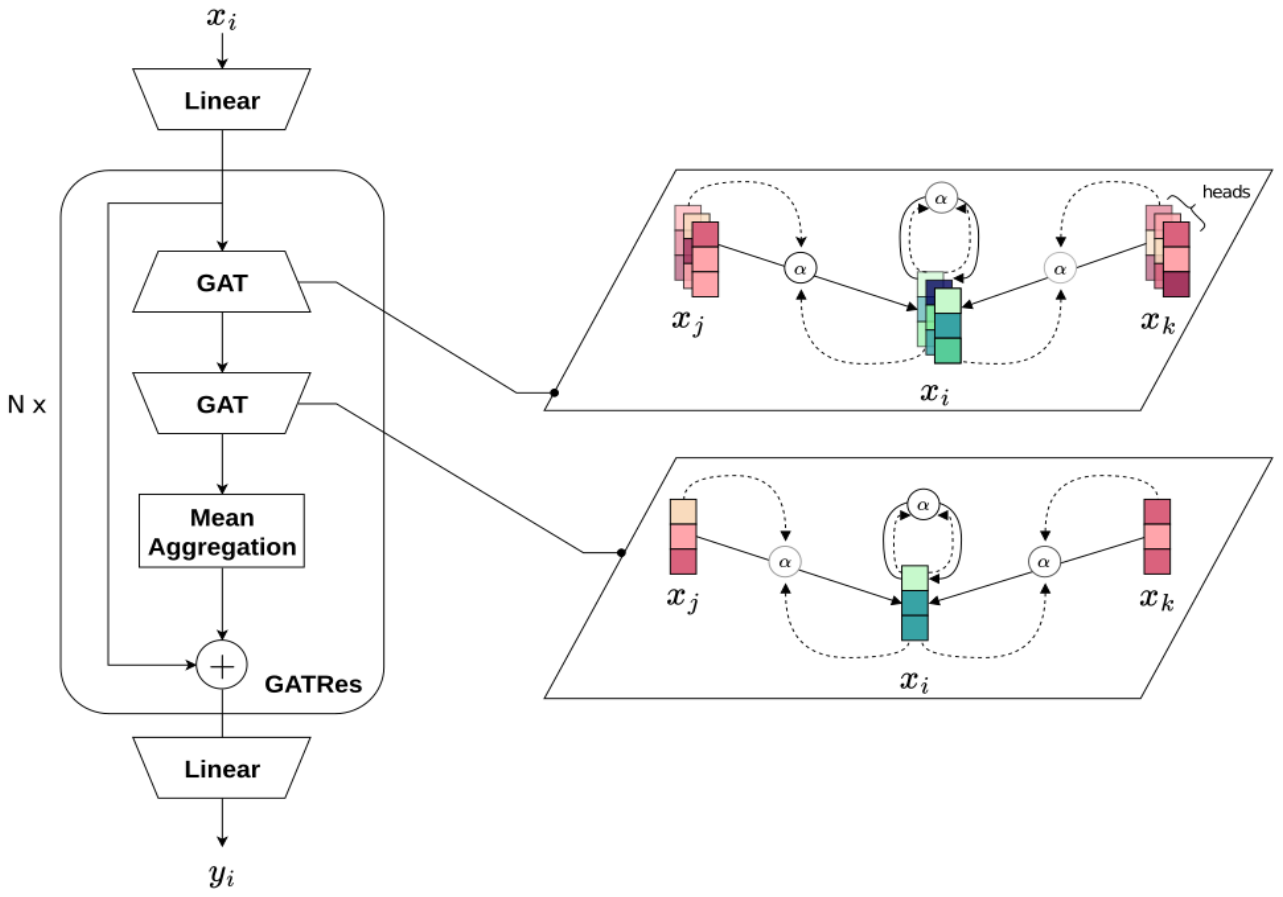
Model	#Million params (↓)	MAE (↓)	MAPE (↓)	NSE (↑)	Acc(@0.1) (↑)
GCNii (M. Chen et al., 2020)	0.63	6.696 ± 0.0838	0.2484 ± 0.0552	−0.1064 ± 0.0266	36.02 ± 0.4684
GAT (Veličković et al., 2018)	0.35	4.397 ± 0.3052	0.2112 ± 0.0767	0.1490 ± 0.1153	66.98 ± 1.6290
GraphConvWat (Hajgató et al., 2021a, 2021b)	0.92	3.611 ± 0.1234	0.1551 ± 0.0376	0.5877 ± 0.0370	62.99 ± 1.1600
GraphConvWat-tuned	**0.23**	2.347 ± 0.0252	0.0963 ± 0.0363	0.749 ± 0.0086	81.09 ± 0.3877
mGCN (Ashraf et al., 2023)	2.48	2.188 ± 0.0558	0.0948 ± 0.0155	0.6993 ± 0.0213	82.83 ± 0.4199
GATRes-small (ours)	0.66	**1.964** ± 0.0301	0.0802 ± 0.0458	**0.778** ± 0.0115	**86.56** ± 0.2826
GATRes-large (ours)	1.67	2.115 ± 0.0503	**0.0799** ± 0.0207	0.7417 ± 0.0140	83.43 ± 0.5044

Problem

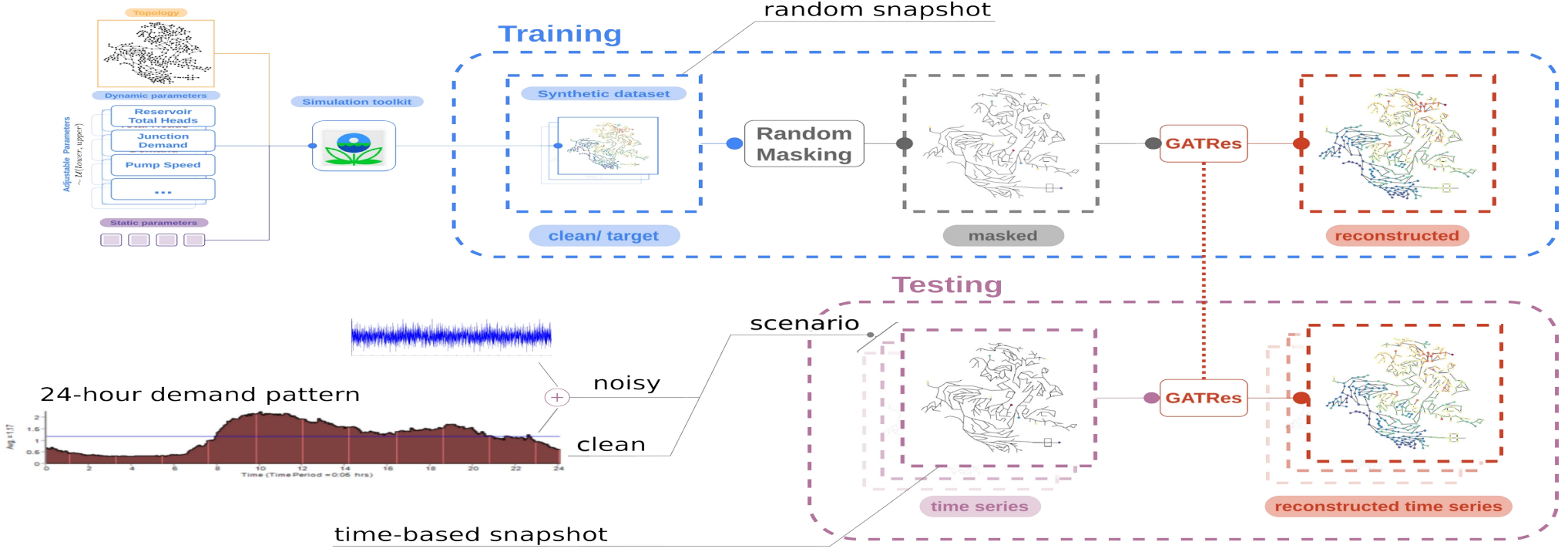
Reconstruct the pressures of the entire water distribution network using few existing sensors



GATRes Model

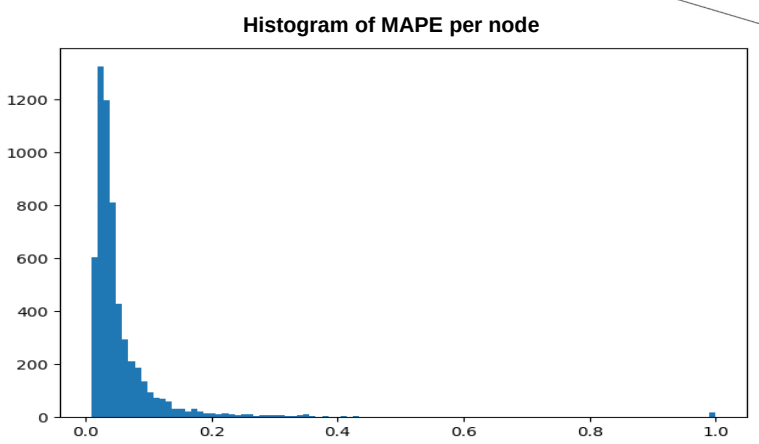


Model Training & Evaluation

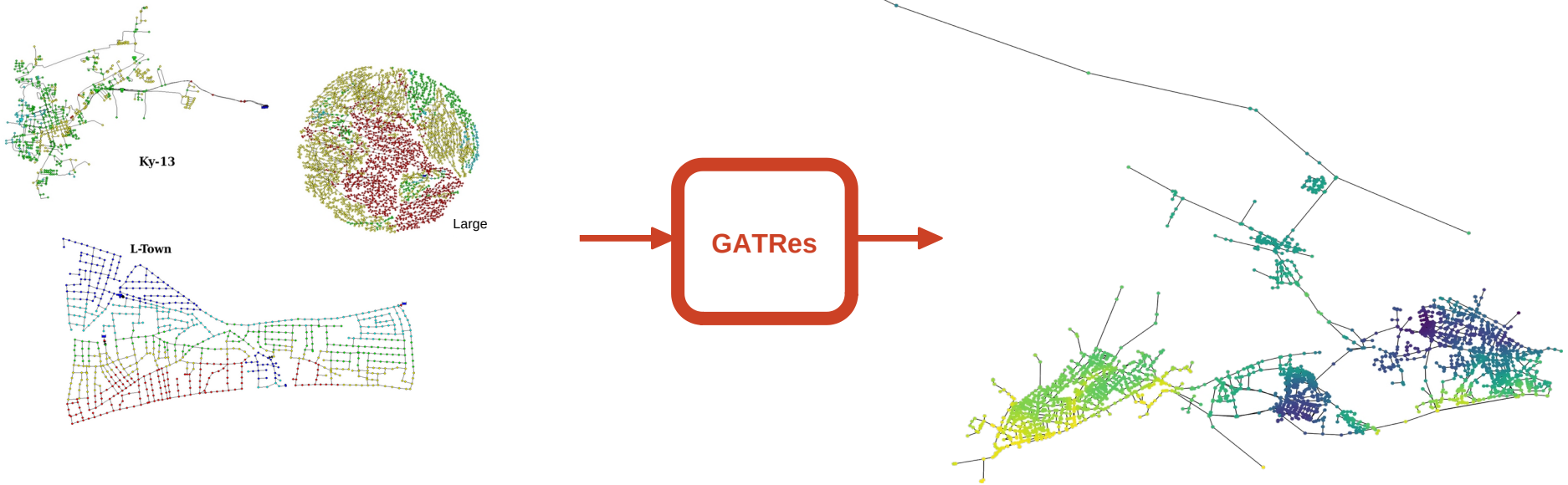


GATRes Performance

	Clean	Noisy
MAE	1.9370 ±0.0074	1.9640 ±0.0301
MAPE	0.0703 ±0.0005	0.0802 ±0.0458



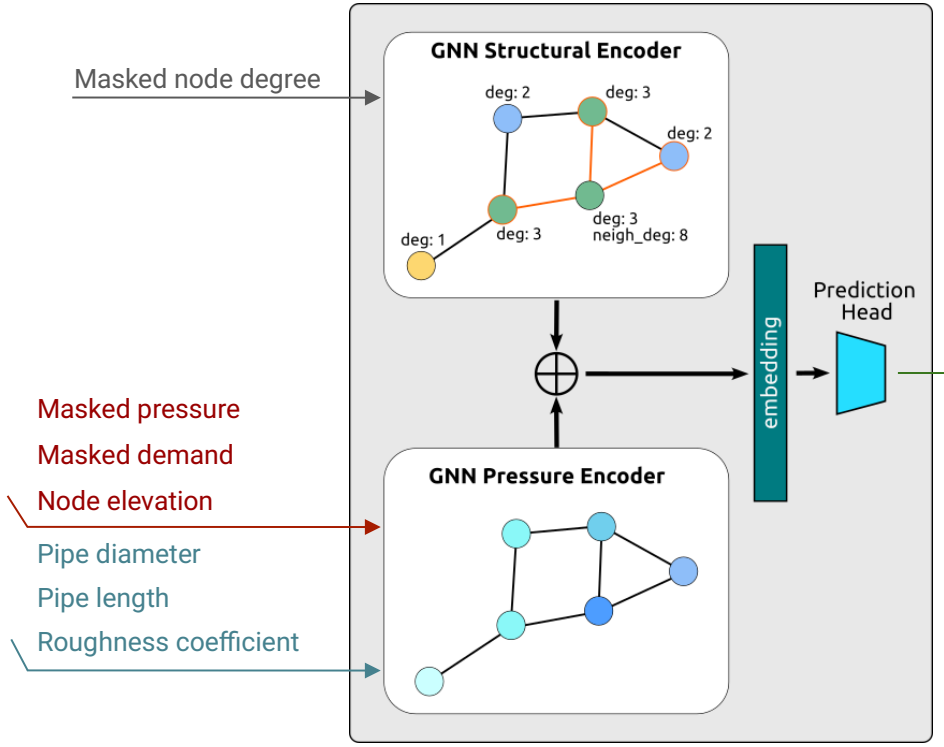
GATRes Generalization



Models	MAE (↓)	MAPE (↓)
Single-Graph	1.9370 ±0.0074	0.0703 ±0.0005
Multi-Graph Zero-Shot	3.0597 ±0.0074	0.0998 ±0.0005
Fine-tuned	**1.9097** ±0.0076	**0.0695** ±0.0005

How to improve Generalization and Transferability in WDNs domain?

Dual-Encoder GFM



$$\hat{p} = \sigma(W_1 X_{\text{pressure}} + W_2 X_{\text{degree}})$$

$$loss = \lambda \times degree\_loss \; + \; (1 - \lambda) \times pressure\_loss$$

$$\lambda = \lambda_{min} + 0.5 \times (\lambda_{max} - \lambda_{min}) \times \left(1 + cos\left(\pi \frac{epoch}{total\_epochs}\right)\right)$$

DiTEC-WDN Dataset

- **36 WDNs Baselines**
  - small (19 - ~500 nodes),
  - medium (500 - ~1000 nodes)
  - large (> 1000 nodes)

- **HuggingFace largest dataset for Graph Machine Learning**
  - 3.65 Terabytes of data
  - 1000 scenarios of WDN operation per network
  - More than 200 Million WDN states while operating under normal conditions.

Downloads last month 15,782

Total downloads

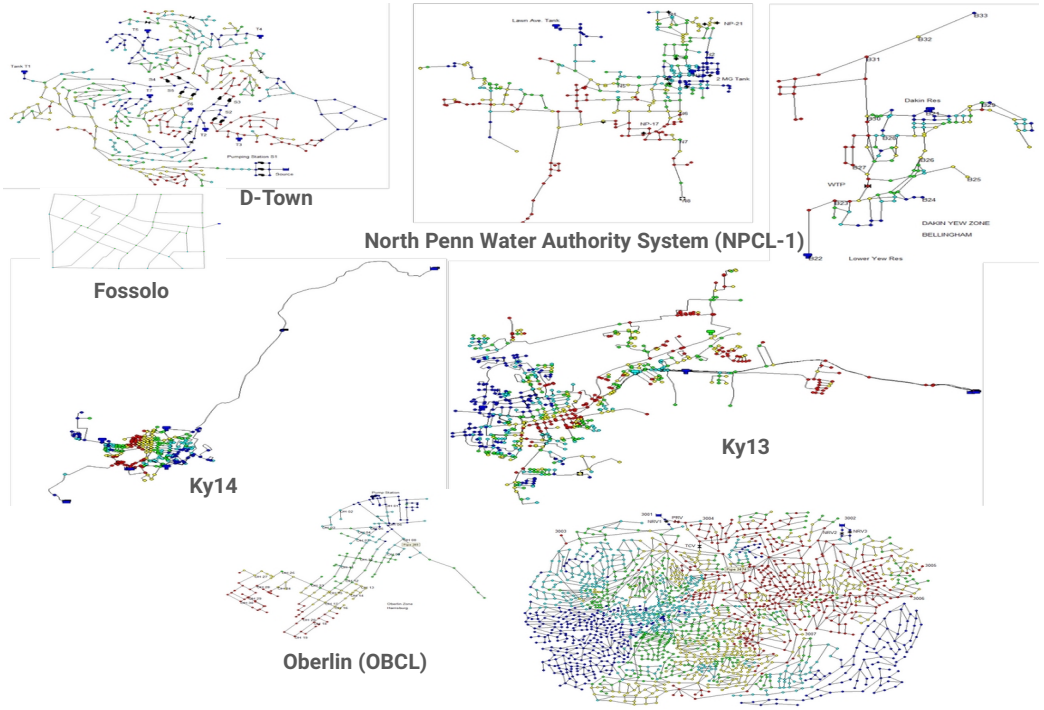
101,649

Dual-Encoder GFM

20 WDNs

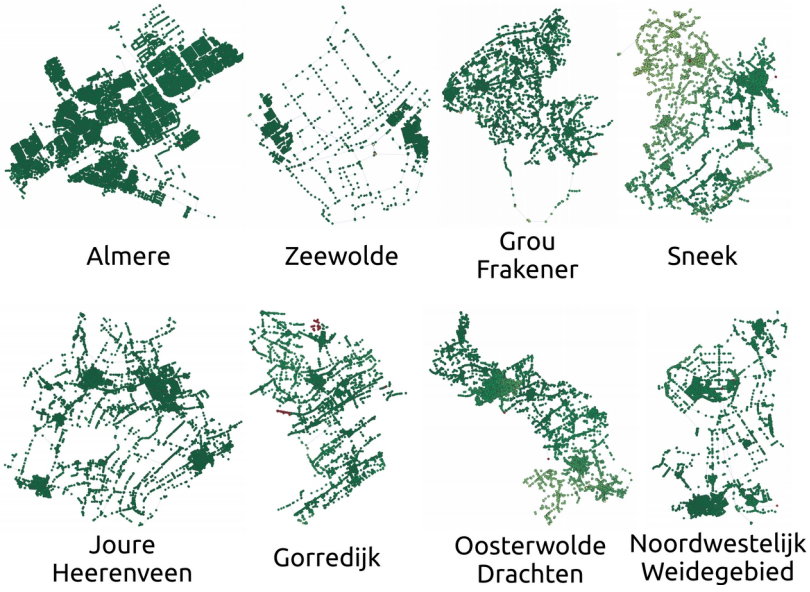
19 – 1900 nodes

TRAINING



8 WDNs Validated at Vitens

20,000+ nodes

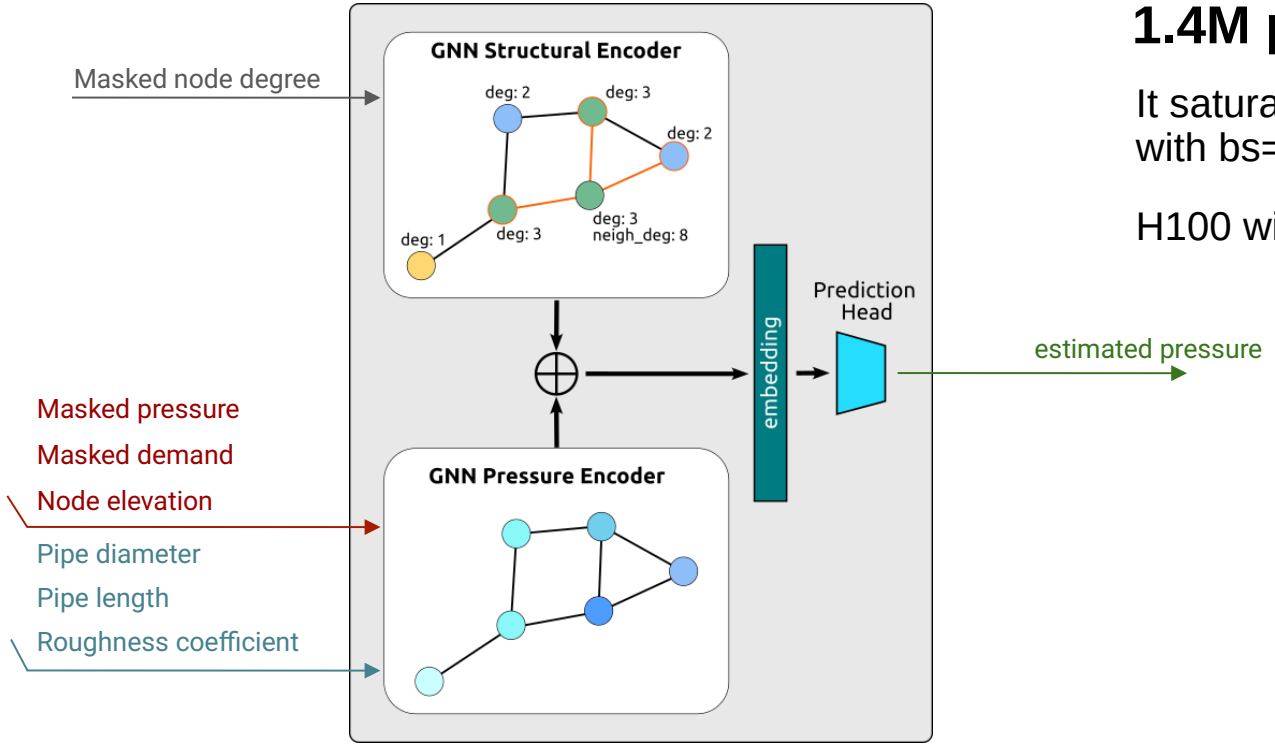


5min: 288 snapshots 1tr / 1val / **286 test**

15min: 96 snapshots 1tr / 1val / **94 test**

Now that we have a generalizable model, **can we make it more efficient?**

Small parameter count does not mean cheap training

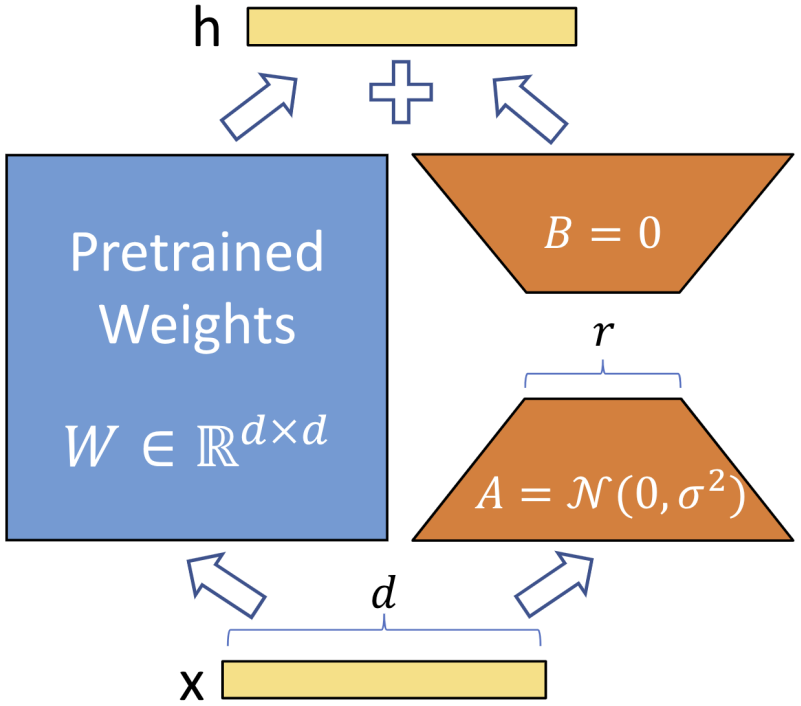


1.4M parameters

It saturated 24GB VRAM of an NVIDIA A30 with bs=64

H100 with 80GB max bs=192

Efficient Model Adaptation



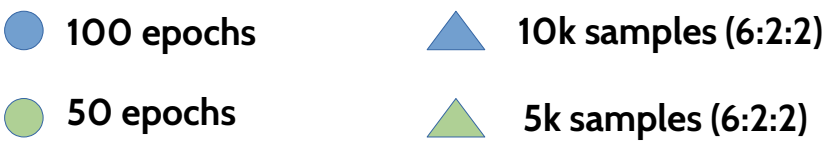
Low-Rank Adaptation (LoRA)

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

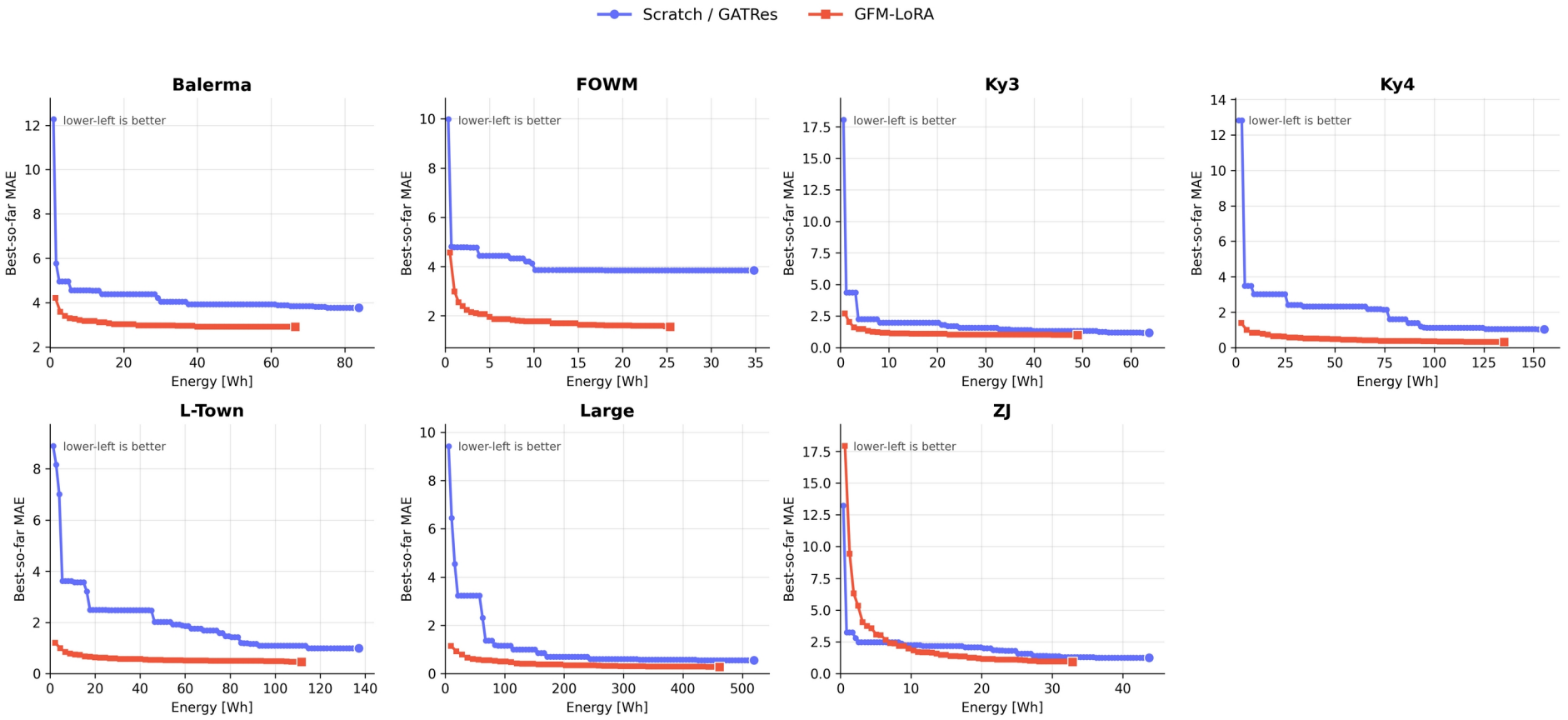
Hu et al., Lora: Low-rank adaptation of large language models., 2022

Experiments with LoRA

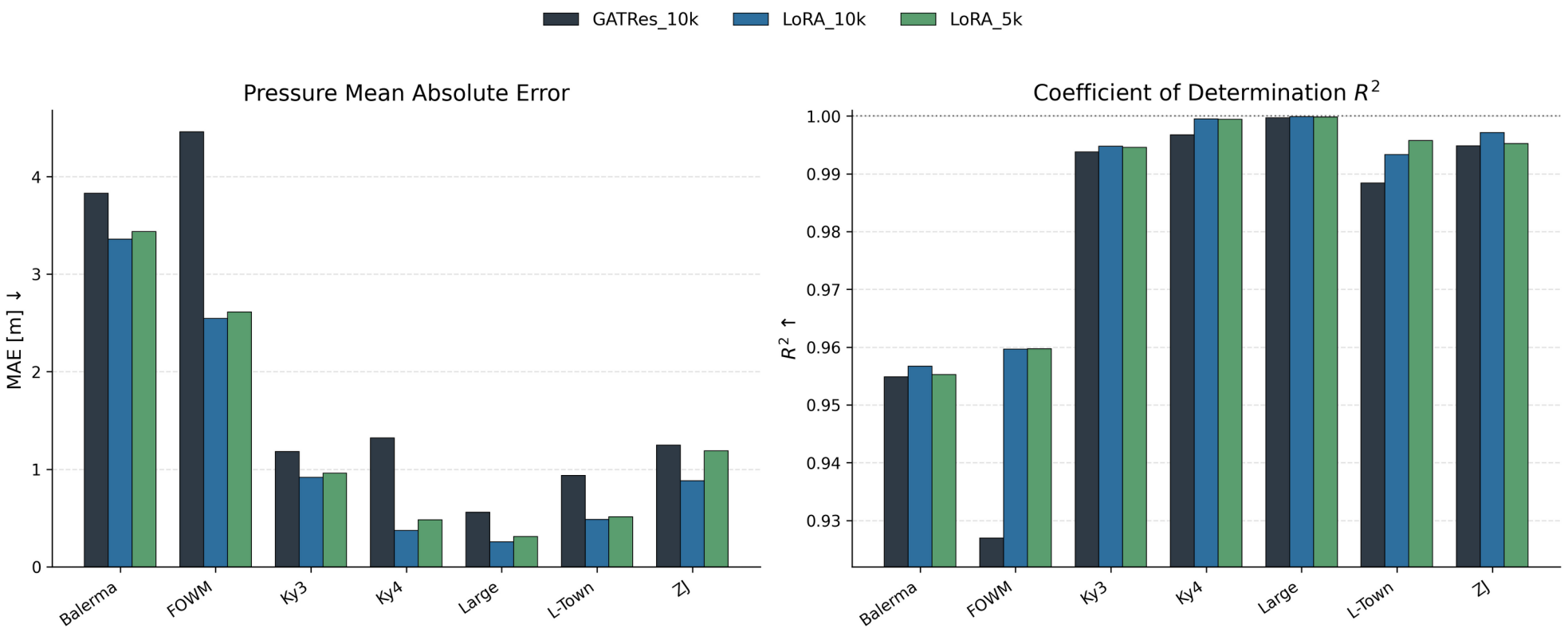
Training Strategy	Description
Single model baseline	Train one GNN from scratch for each target WDN
Dual Encoder GFM	Pretrain on 20 WDNs
GFM-LoRA	Freeze the GFM and adapt low-rank modules to each target WDN
GFM-LoRA 5k	Reduced target adaptation data budget



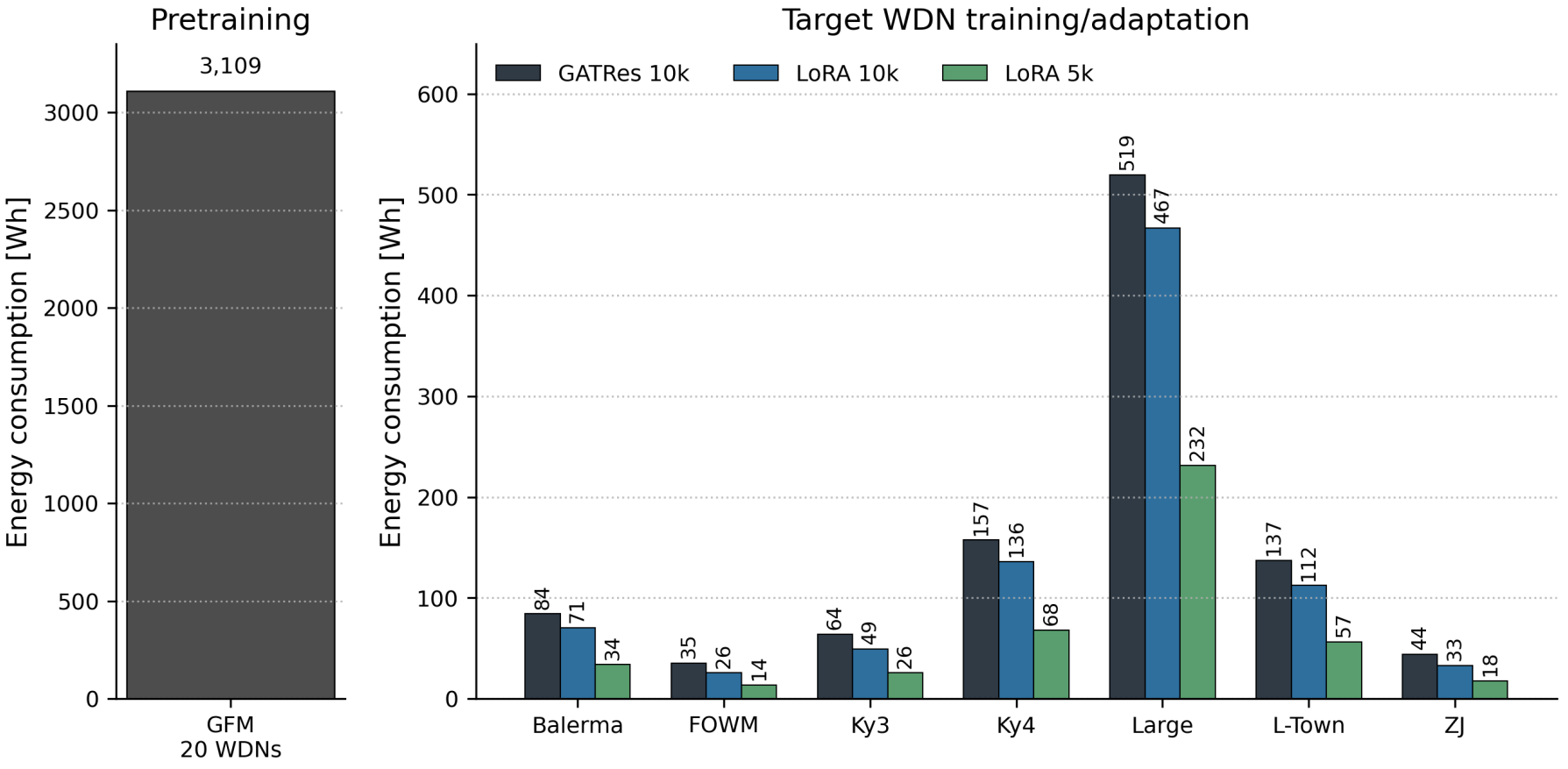
Energy/Performance across WDNs



Performance Comparison



Energy Consumption



Take aways

- Sustainable AI is not only about compressing huge models after over-scaling. It is about designing reusable, structurally appropriate, and cheap to adapt from the beginning.
- Small parameter count does not mean cheap training
- Energy-Performance trade-off was not observed in all our experiments.

Thank you