

University of Groningen

## Distorted or high-fidelity dataset for data augmentation in Graph Neural Networks?

Truong, Huy; Tello, Andrés; Degeler, Victoria; Lazovik, Alexander; Koellermeier, Julian

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*

2023

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Truong, H., Tello, A., Degeler, V., Lazovik, A., & Koellermeier, J. (2023). *Distorted or high-fidelity dataset for data augmentation in Graph Neural Networks? Improving pressure estimation in Water Distribution Systems*. Poster session presented at European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Turin (Torino), Italy.

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

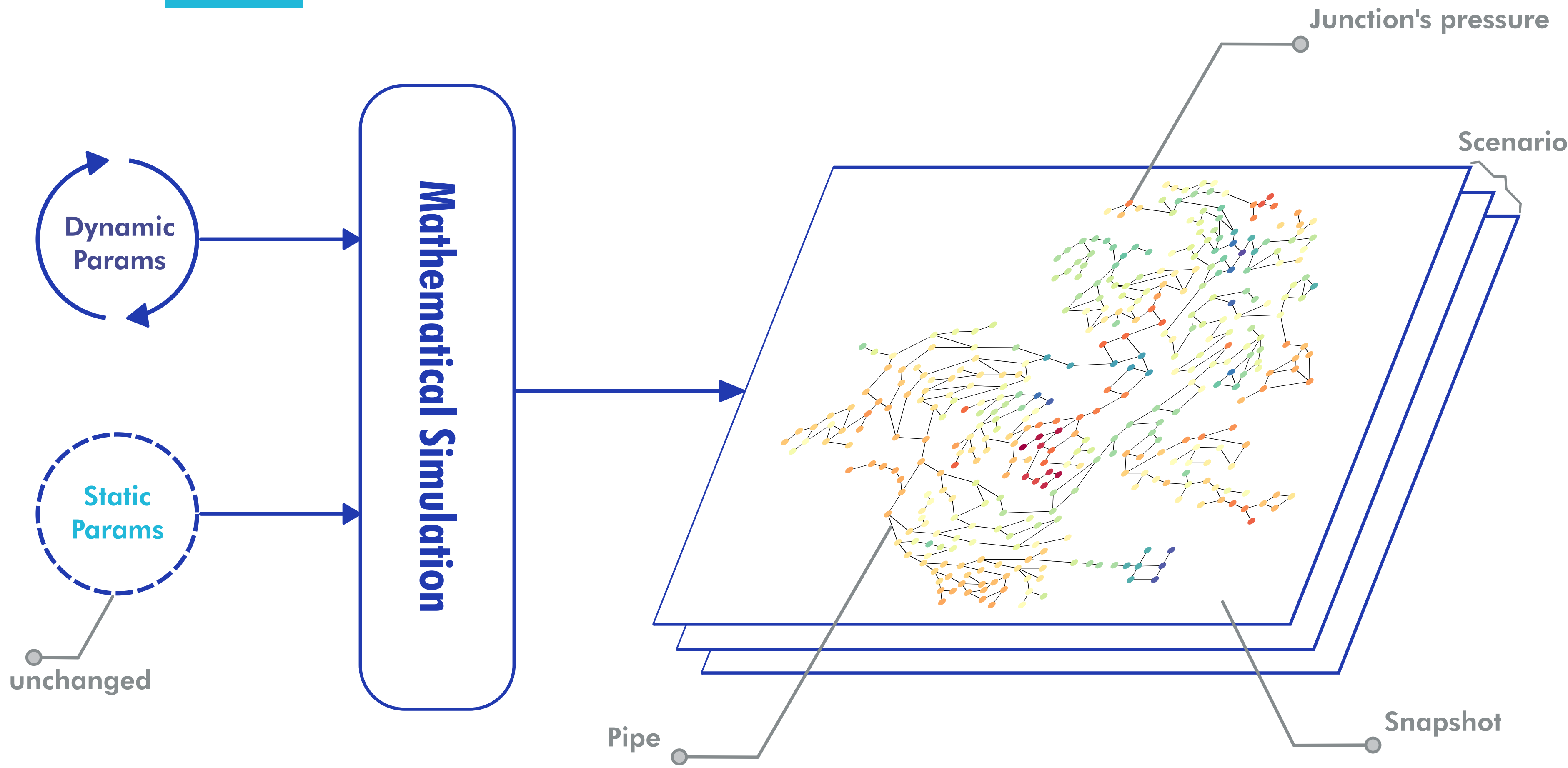
# Distorted or high-fidelity dataset for data augmentation in Graph Neural Networks?

## Improving pressure estimation in Water Distribution Systems

Huy Truong<sup>1\*</sup>, Andres Tello<sup>1\*</sup>, Victoria Degeler<sup>2</sup>, Alexander Lazovik<sup>1</sup>, Julian Koellermeier<sup>1</sup>

<sup>1</sup>University of Groningen, <sup>2</sup>University of Amsterdam, \* Equal contribution

### SIMPLE SYNTHETIC DATA GENERATION



#### Background

The objective is to estimate pressure values at unknown junctions given existing sensors.

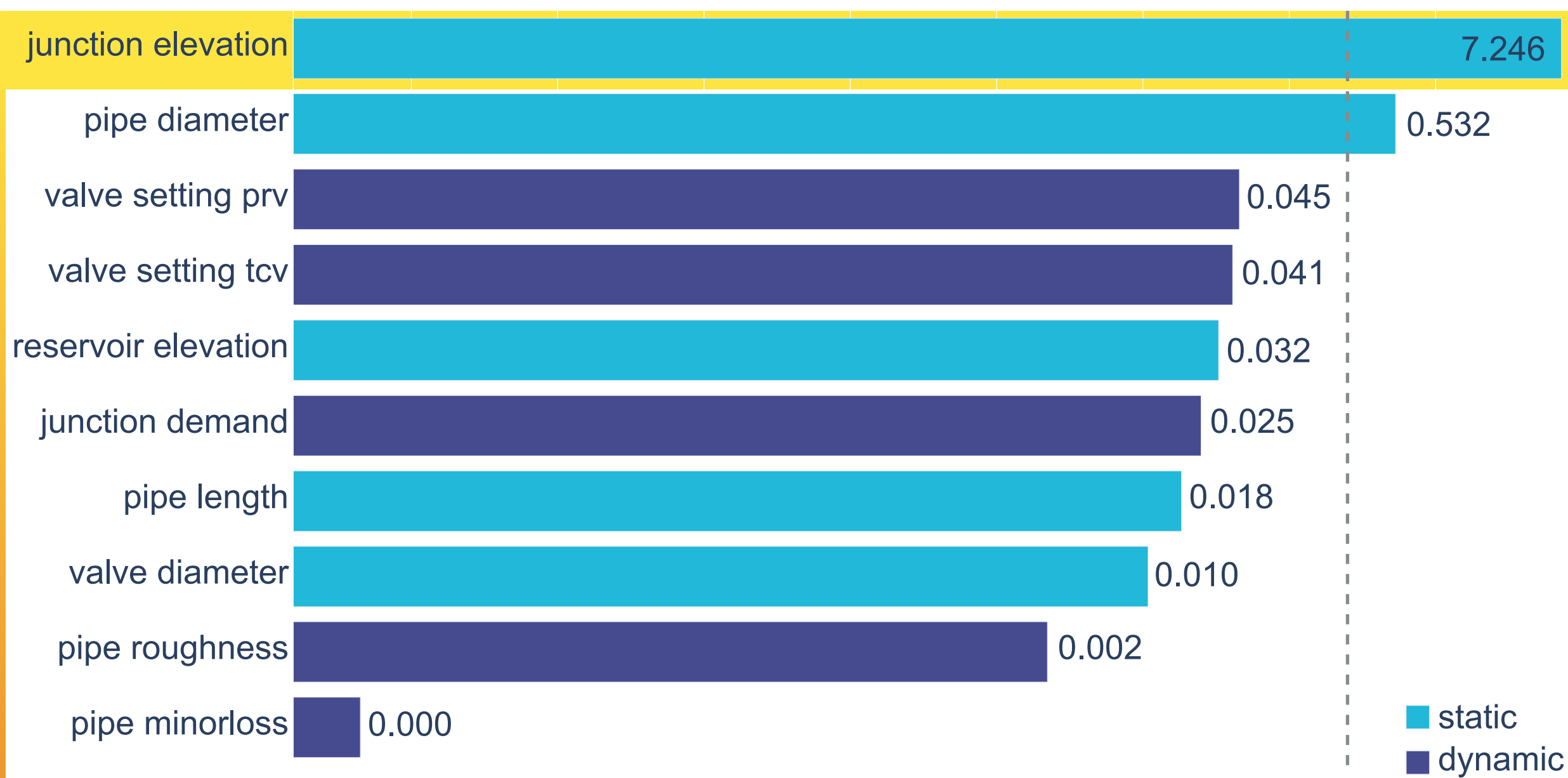
A **Simple** dataset is created using only dynamic parameters that can be changed in reality.

#### Issues

- Limited number of sensors
- Few In-Distribution (ID) scenes
- Test on Out-Of-Distribution (OOD) scenes

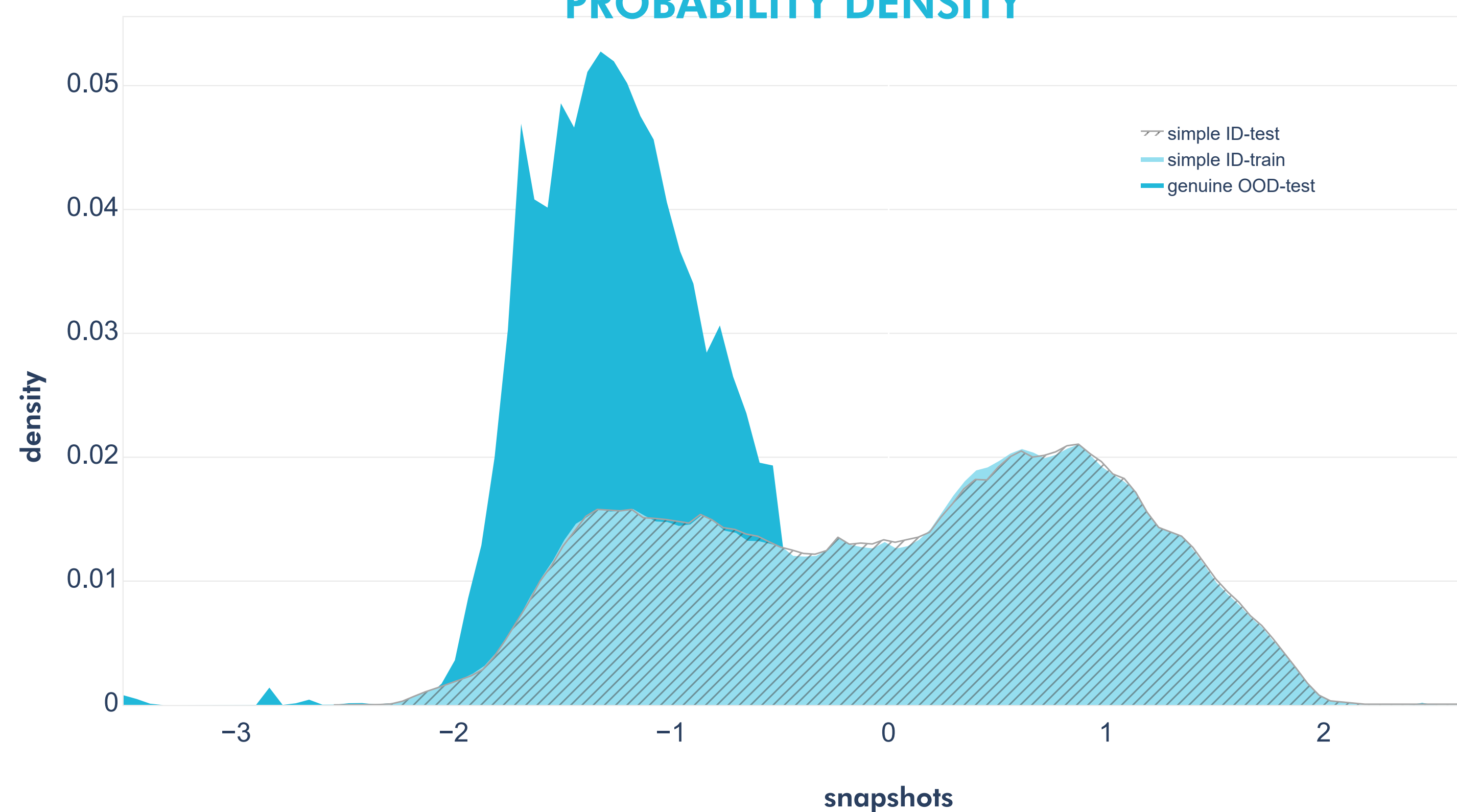
### SENSITIVITY ANALYSIS

Sensitivity score:  $|\text{Neighbor.Var}_{\text{factor}} - \text{Neighbor.Var}_{\text{unchange}}|$

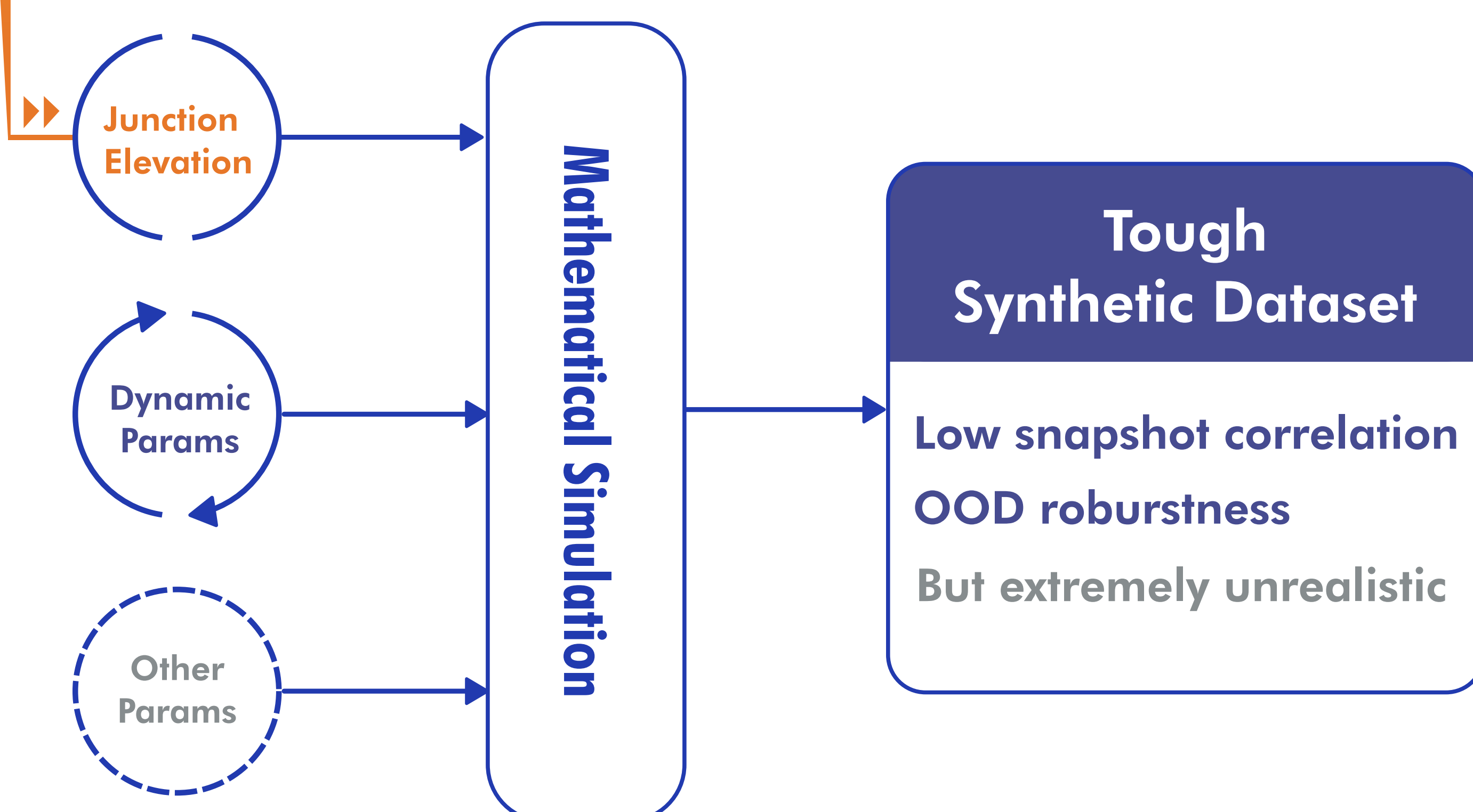


Idea: Distorting the most sensitivity parameter in the data generation phase

### PROBABILITY DENSITY



### ELEVATION-DISTORTED( **TOUGH** ) SYNTHETIC DATA GENERATION



#### Methodology

Apply an algorithm to strategically distort the **elevation** parameters at junctions to provide a distorted **Tough** dataset.

Use the dataset to train a Graph Neural Network to reconstruct pressure values at unknown positions in an unsupervised setting by masking out 95% of nodes and using visible nodes as input.

Test on an OOD data that the model has not encountered during training.

#### Results on OOD data

The model trained on the **Tough** dataset produces similar results on virtual sensors but shows a noticeable discrepancy on practical sensors compared to the model trained on **Simple** dataset.

Trained on	Virtual Sensors			Real Sensors		
	MAE ↓	MAPE ↓	NSE ↑	MAE ↓	MAPE ↓	NSE ↑
<b>Simple</b>	1.937	0.070	0.773	1.427	0.038	-1.294
<b>Tough</b>	2.017	0.075	0.790	0.713	0.026	0.818

#### Discussion

Relationship to Generative Model

Relationship to Active Learning