Retrieval Over Training
Similarity search-based Model Selection for Time Series Anomaly Detection

Christos Panourgias *
Université Paris Cité, LIPADE
panourgiaschris@gmail.com

Roberto Stanzione *
École Normale Supérieure, Inria
roberto.stanzione@inria.fr

Adrien Petralia
Université Paris Cité, LIPADE
adrien.petralia@gmail.com

Themis Palpanas
Université Paris Cité, LIPADE
themis@mi.parisdescartes.fr

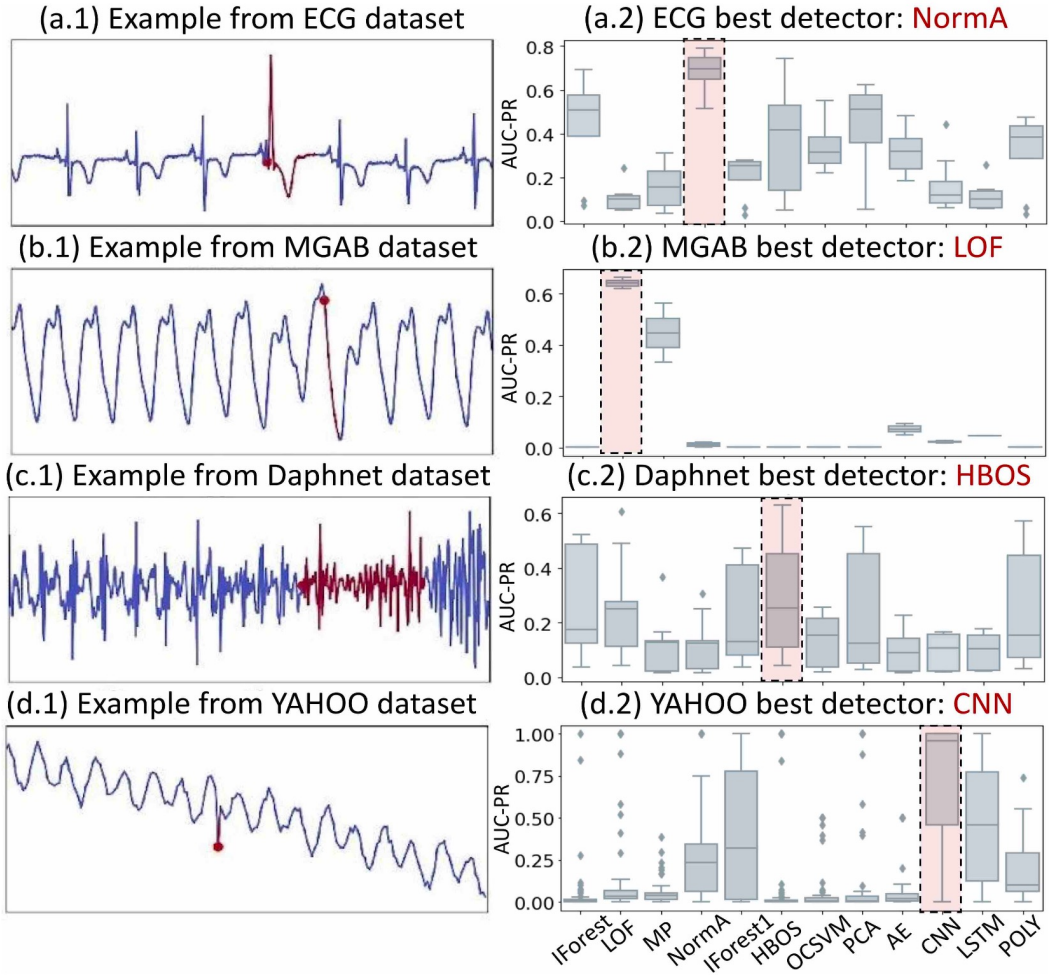
Paul Boniol
École Normale Supérieure, Inria
paul.boniol@inria.fr

*Equal Contribution.

NeurIPS 2026

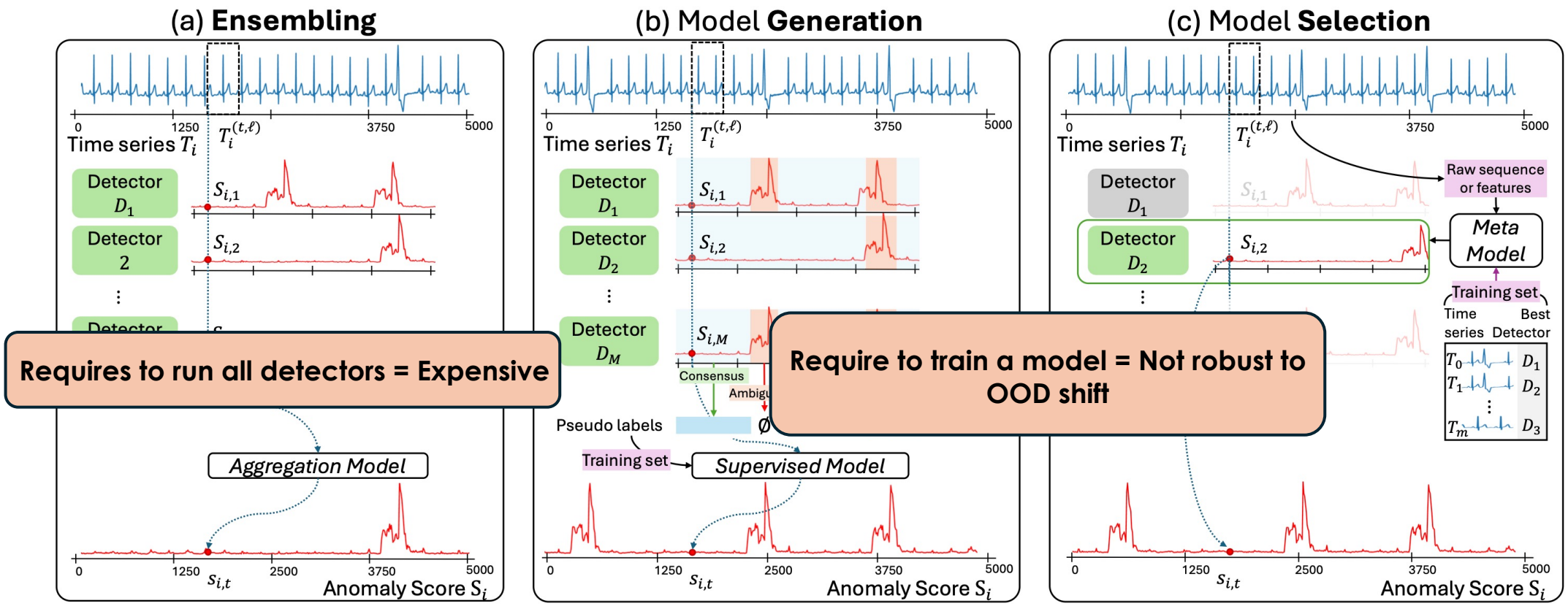
Automatic Solutions for TSAD

- TSAD appears across many domains
- Time series vary in structure, noise, seasonality and anomaly type
- **No single detector is best across all datasets**

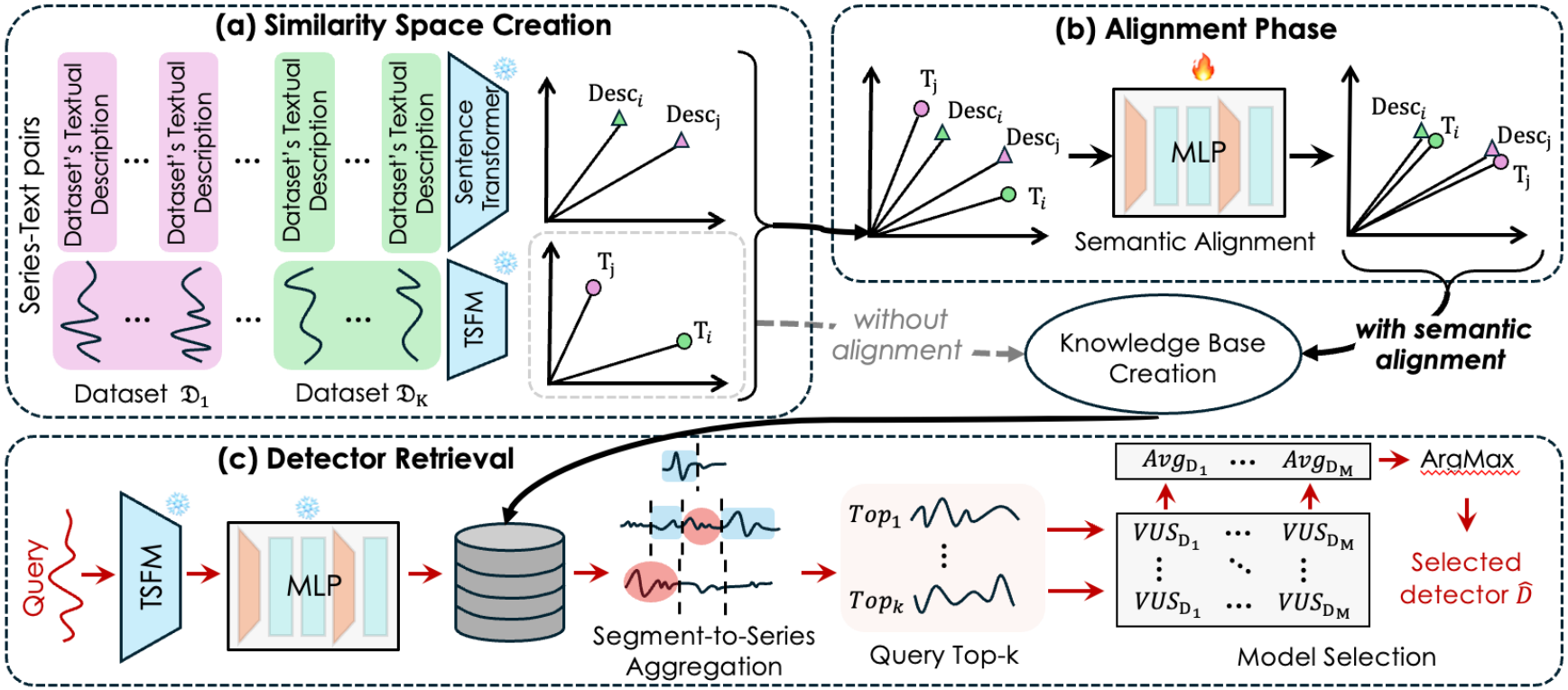


Sylligardos et al., "Choose Wisely: An Extensive Evaluation of Model Selection for Anomaly Detection in Time Series," PVLDB 2023.

Automatic Solutions for TSAD



RAMSAD: Retrieval-Augmented Model Selection for Anomaly Detection



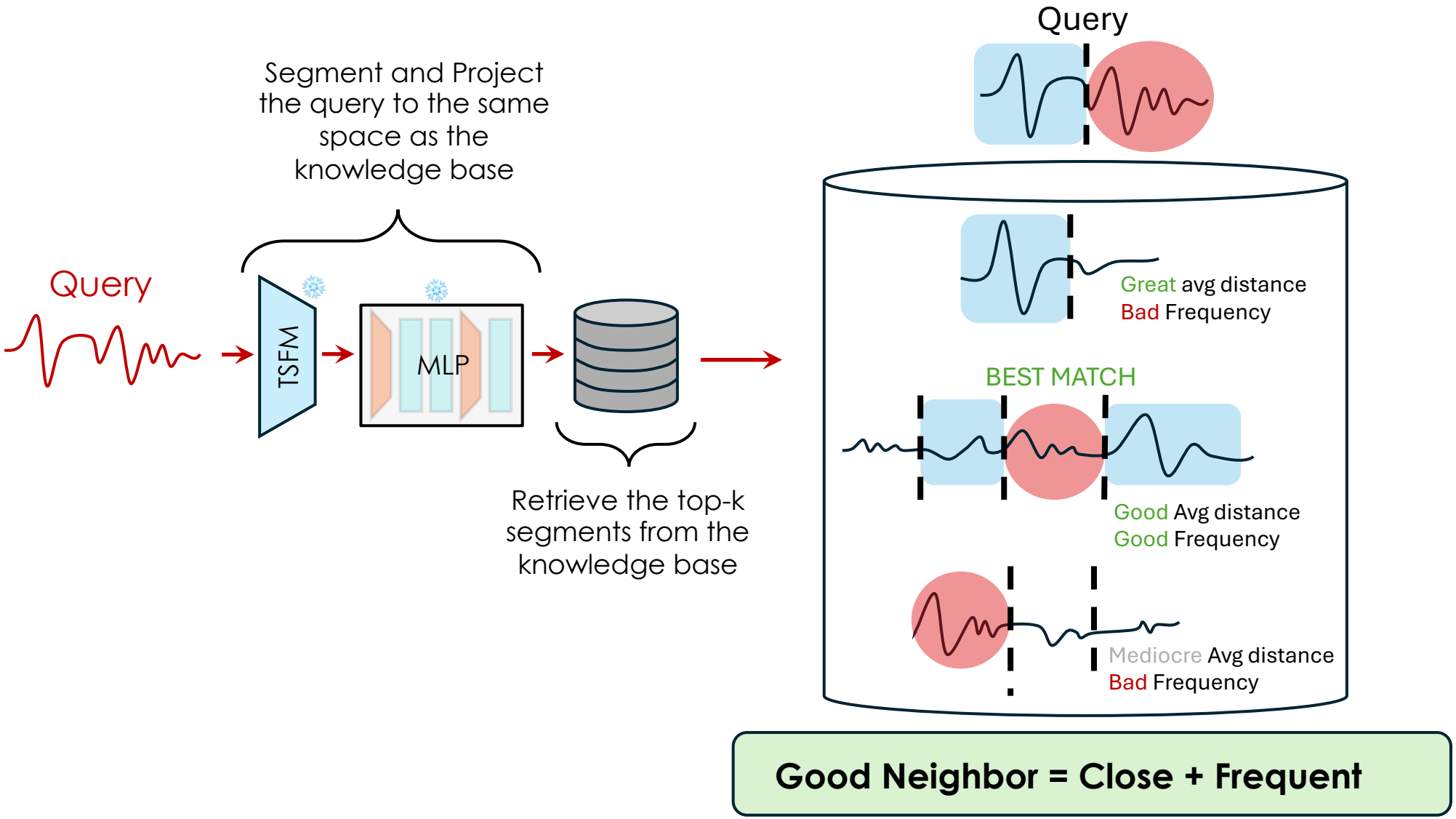
- **Create a similarity space** using TSFM embeddings and optional semantic alignment
- Build a **knowledge base** of past time series and detector scores in this space
- Given a new query **Retrieve top-k**
- Select the detector with the **best average performance** over retrieved neighbors

Building the Similarity Space



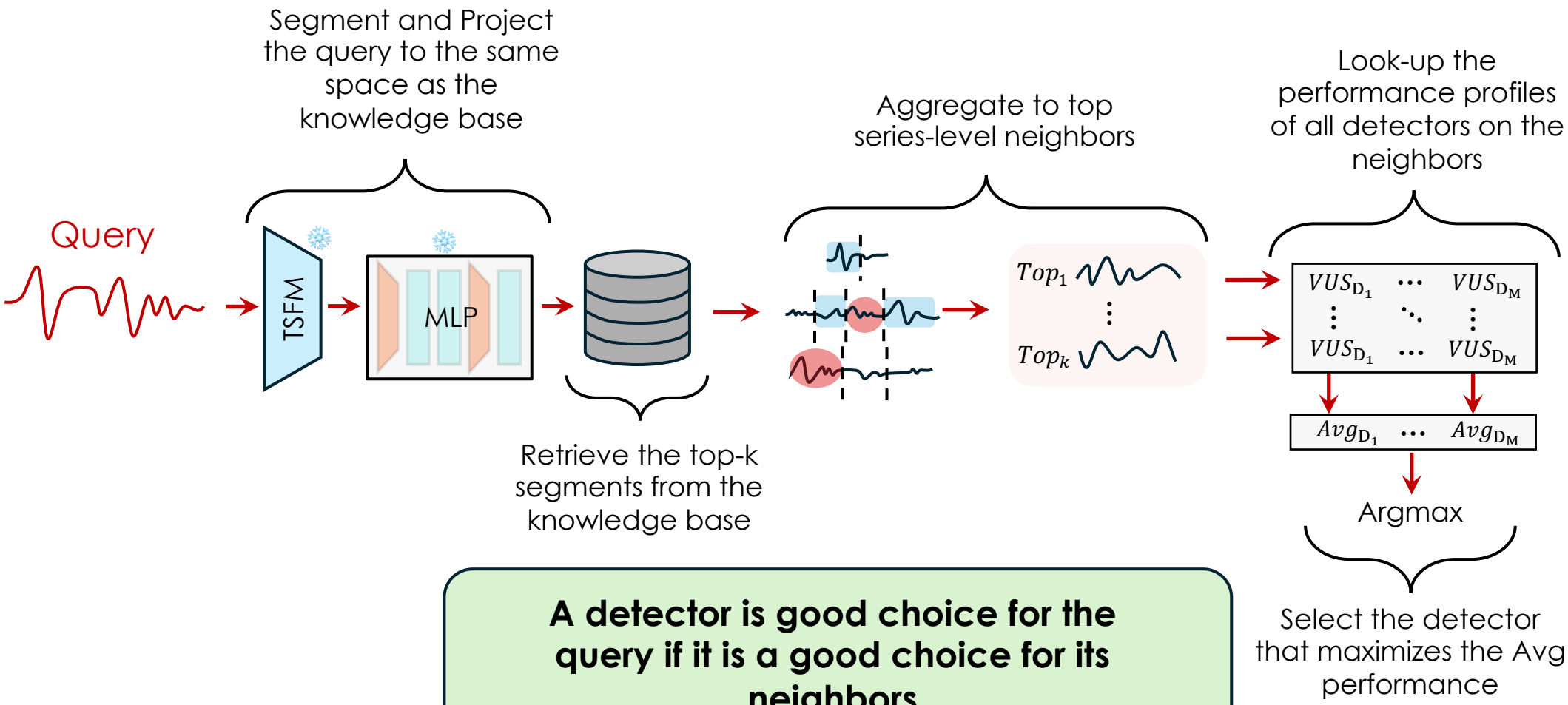
- Dataset descriptions provide **domain context**
- Alignment moves related series and descriptions **closer in the embedding space**
- In OOD, retrieved neighbors have **more similar anomaly meaning**

Retrieval and Detector Selection



Good Neighbor = Close + Frequent

Retrieval and Detector Selection



A detector is good choice for the query if it is a good choice for its neighbors

Experimental Setup

Datasets

- **870** univariate time series (TSB-AD)
- **9** Domains
- **ID** + **OOD** evaluation splits

Detector Pool

32 TSAD Detectors

Distance | Density | Prediction | Reconstruction

Baselines

25 automatic methods

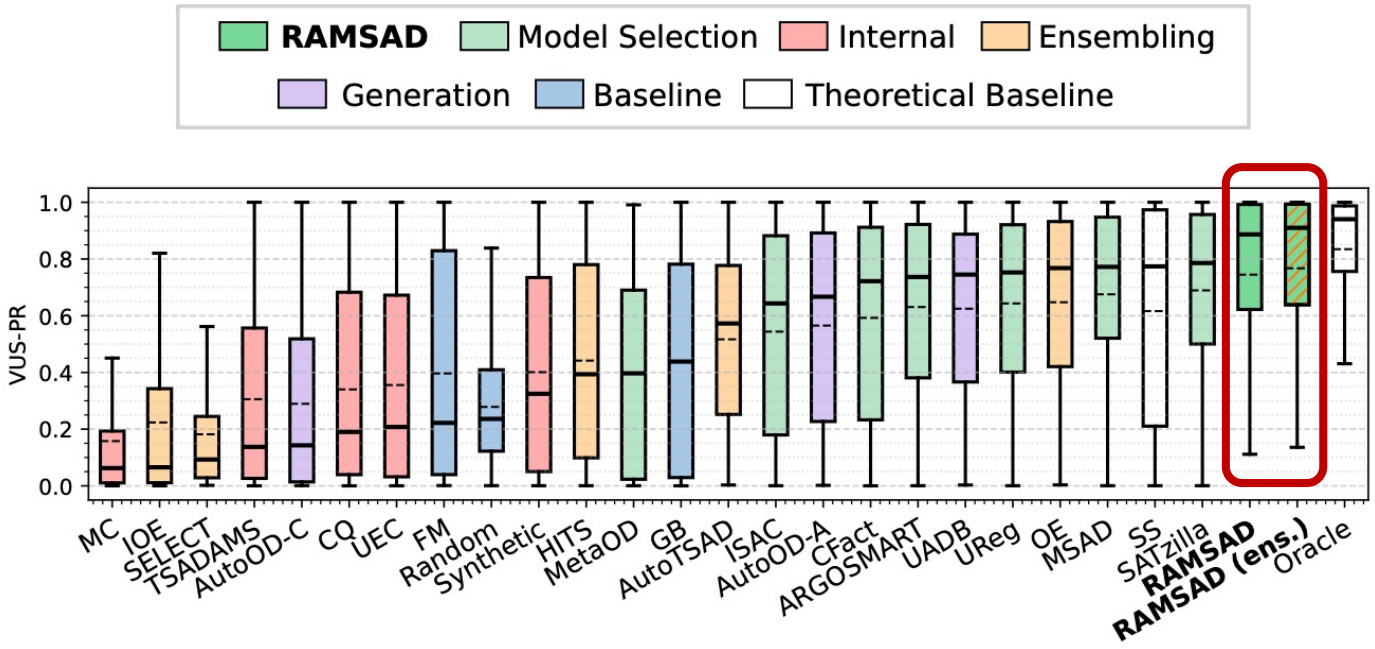
Selection | Ensembling | Generation

Evaluation Measure

VUS-PR

- Precision-recall-based
- Higher is Better

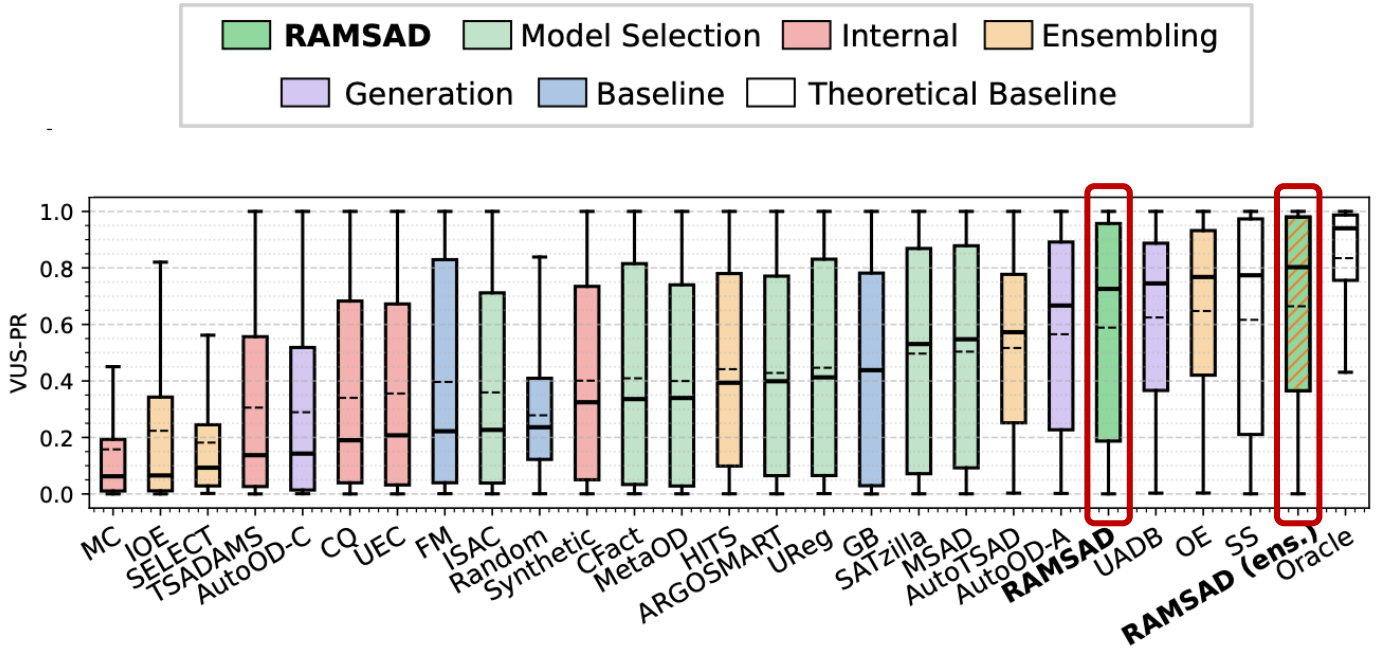
RAMSAD Leads in the In-Distribution Setting



In ID, retrieval is best: RAMSAD is the strongest single-model selector

- RAMSAD achieves the **best single-model performance**
- RAMSAD ensemble reaches **near-oracle performance**

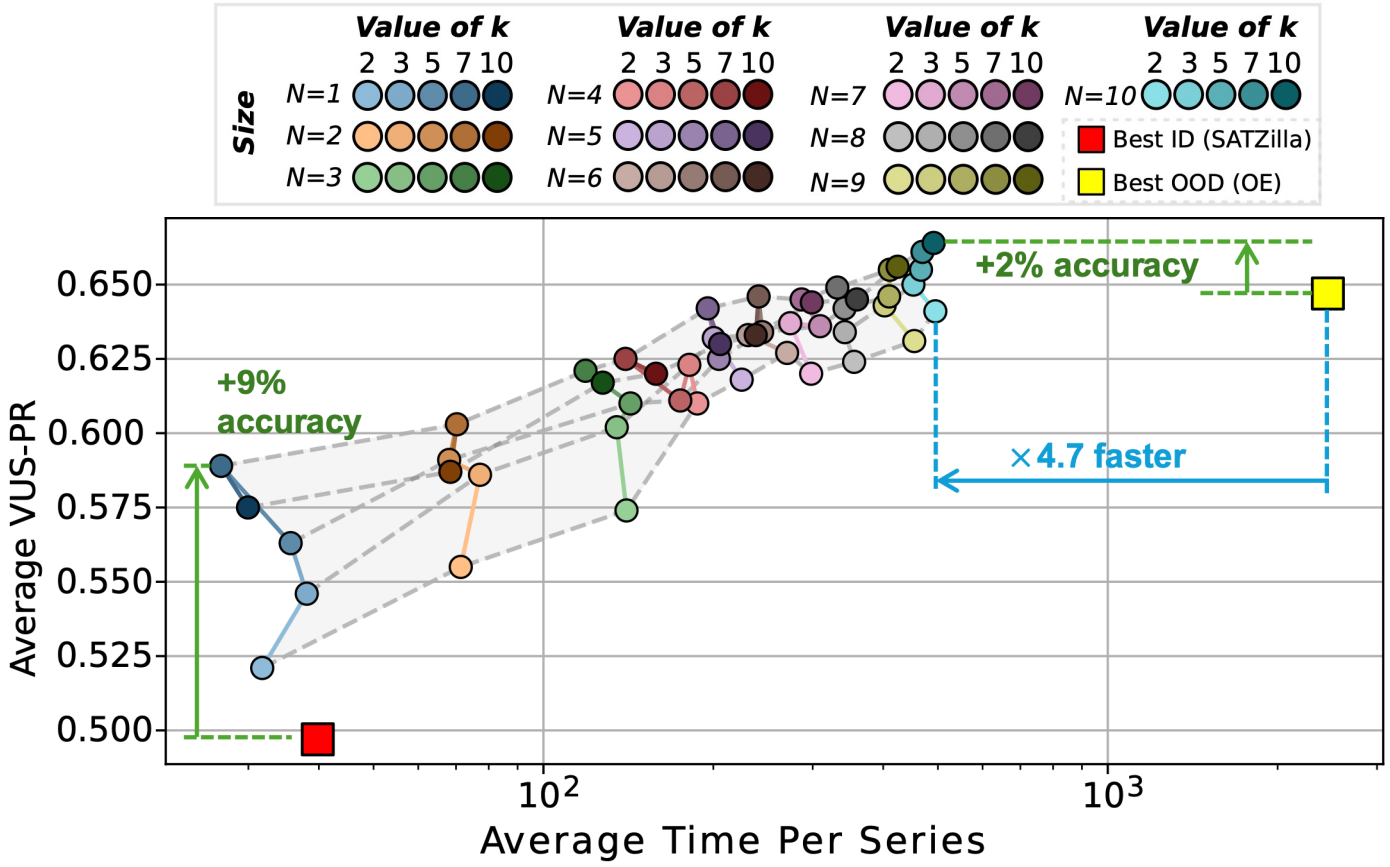
RAMSAD Is Most Robust in the OOD Setting



In OOD, RAMSAD is the most robust selector

- Selectors drop ~19–20% in OOD; RAMSAD drops only **16%**
- Single-model RAMSAD **matches OE and UADB, which run all 32 detectors**
- RAMSAD ensemble achieves the **best OOD rank**

RAMSAD Improves Accuracy without Full-Ensemble Cost



Higher accuracy, lower cost: RAMSAD runs only the useful detectors

- Single-model RAMSAD **is competitive with ensemble methods**
- RAMSAD ensemble beats the best ensemble baseline **using only 1/3 of the detectors**
- RAMSAD wins while staying **4.7x faster**

Conclusions

- **Retrieval instead of training:** RAMSAD picks a detector by looking at similar past series, so there's no selector to train.
- **Best model selector:** it beats all model-selection baselines.
- **Strong ensembles at lower cost: i**t reaches near-oracle performance and beats the best ensemble while running only 1/3 of the detectors, 4.7× faster.
- **Robust out-of-distribution:** semantically aligned TSFM spaces give the best retrieval, especially OOD.
- **Explainable by design:** the retrieved neighbors show why a detector was chosen.

Thank you



Code: github.com/Hendrix8/RAMSAD

Université Paris Cité

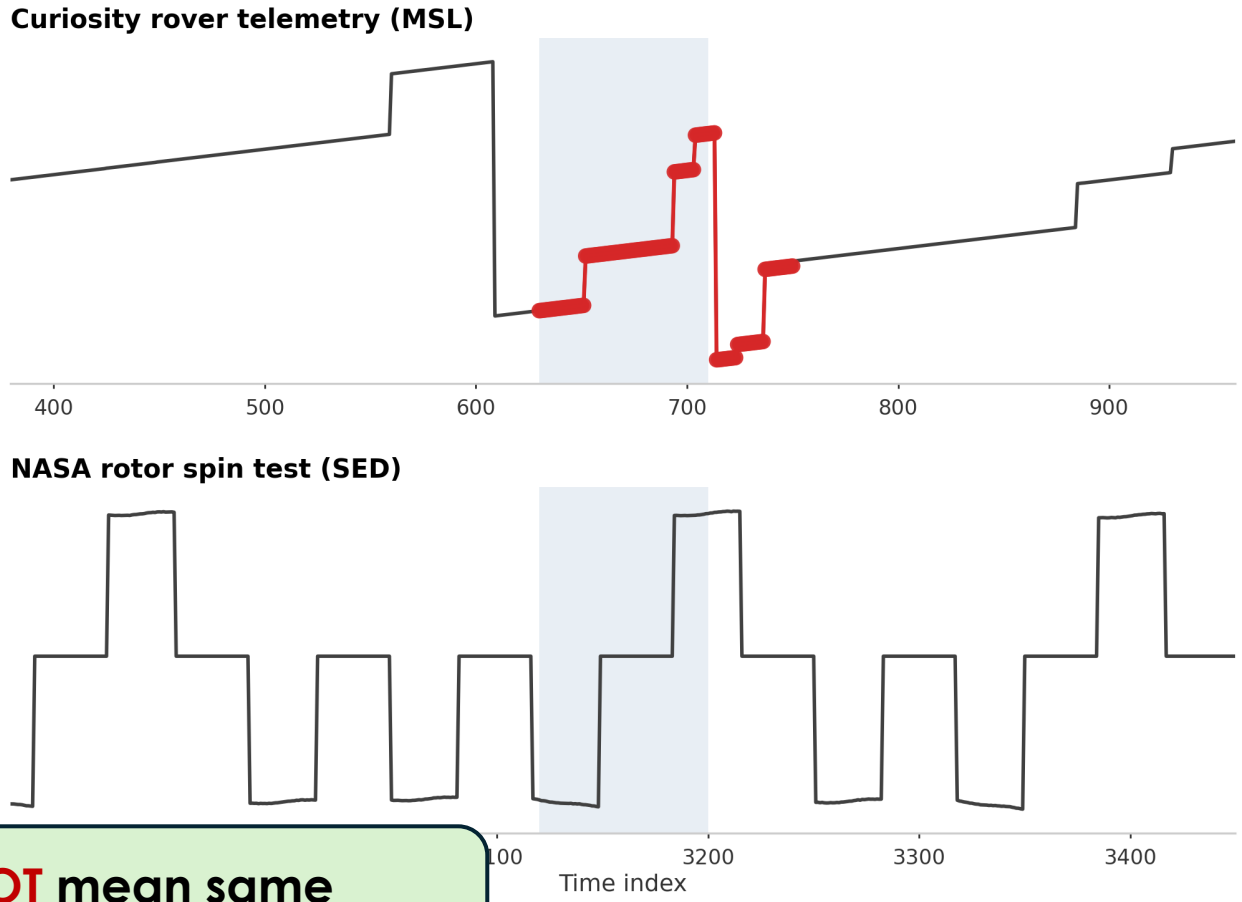
TwinODIS

Appendix

Ablations and design choices

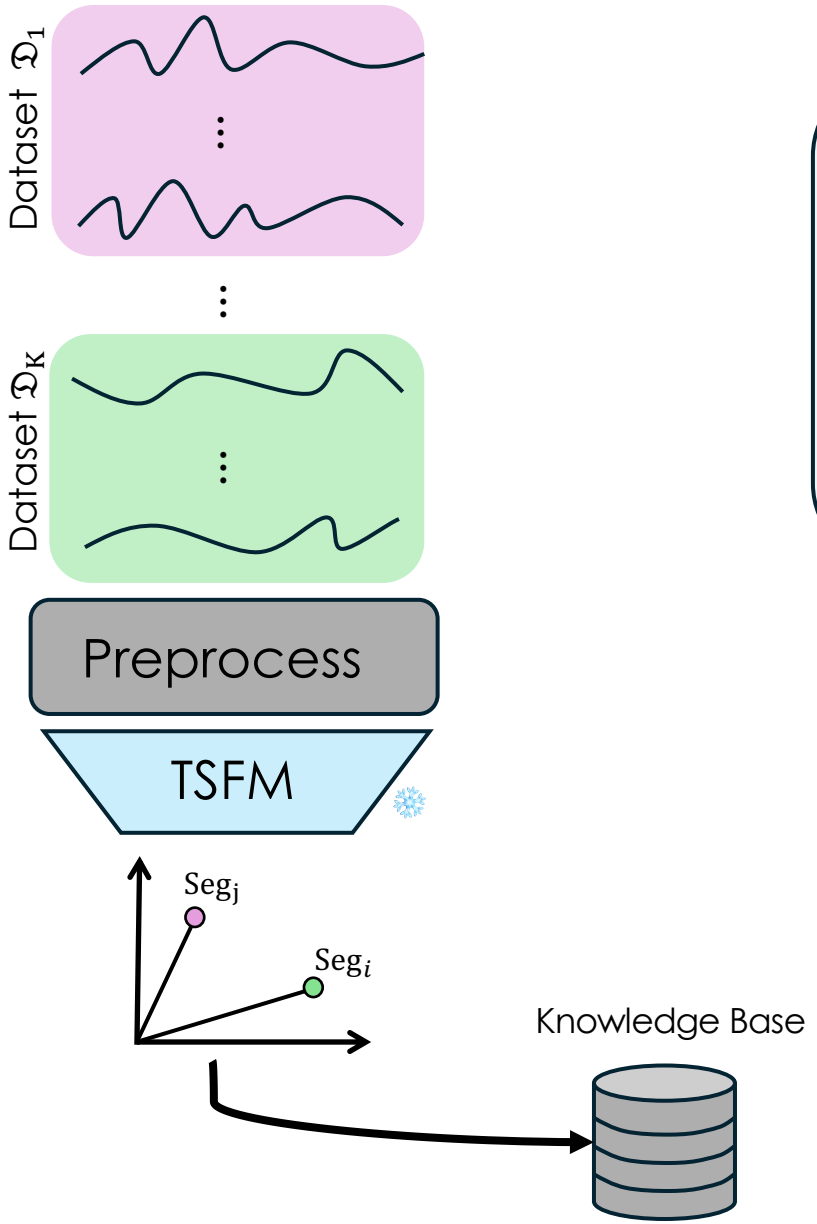
Why Not Compare Raw Time Series Directly?

- Raw similarity is shape-based and costly
- Similar patterns can mean different things across domains
- Time series also vary in length, making direct comparison harder



Same shape does NOT mean same anomaly meaning

Building the Retrieval Space



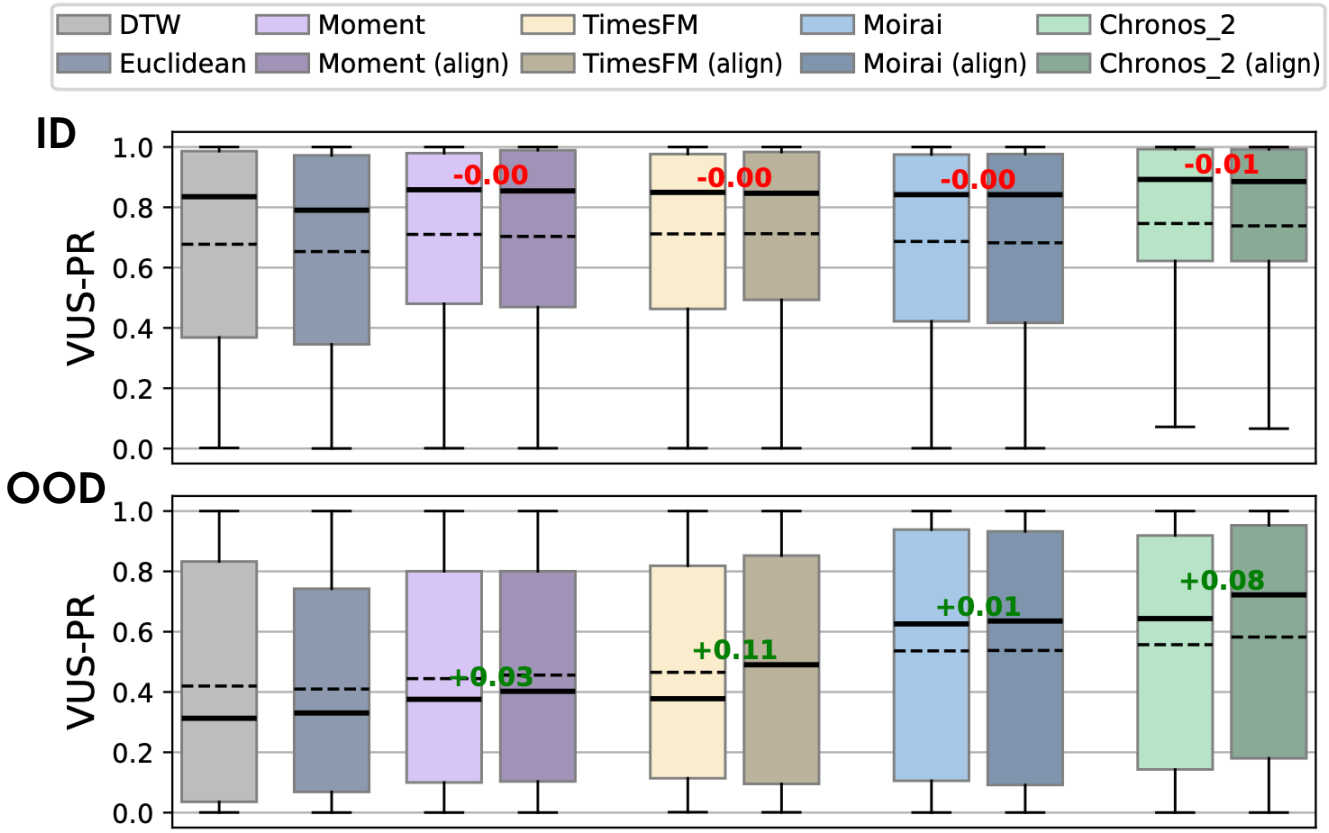
Why TSFM embeddings?

- TSFM embeddings create a **learned temporal feature space**
- Vector similarity becomes **efficient and more expressive**
- Fixed-size embeddings make retrieval easier to standardize

TSFM embeddings make similarity efficient and more expressive

Which Retrieval Space Works Best?

- Original spaces perform worse
- TSFM spaces improve retrieval
- Semantic alignment helps most in OOD
- Best results come from aligned TSFM spaces



Better retrieval space → Better selection