BTS: Building Timeseries Dataset Empowering Large-Scale Building Analytics

Arian Prabowo (UNSW), Xiachong Lin (UNSW, CSIRO), Imran Razzak (UNSW), Hao Xue (UNSW), Emily W. Yap (UOW), Matthew Amos (CSIRO), Flora D. Salim (UNSW)

(UNSW) University of New South Wales
(UOW) University of Wollongong
(CSIRO) Commonwealth Scientific and Industrial Research Organisation

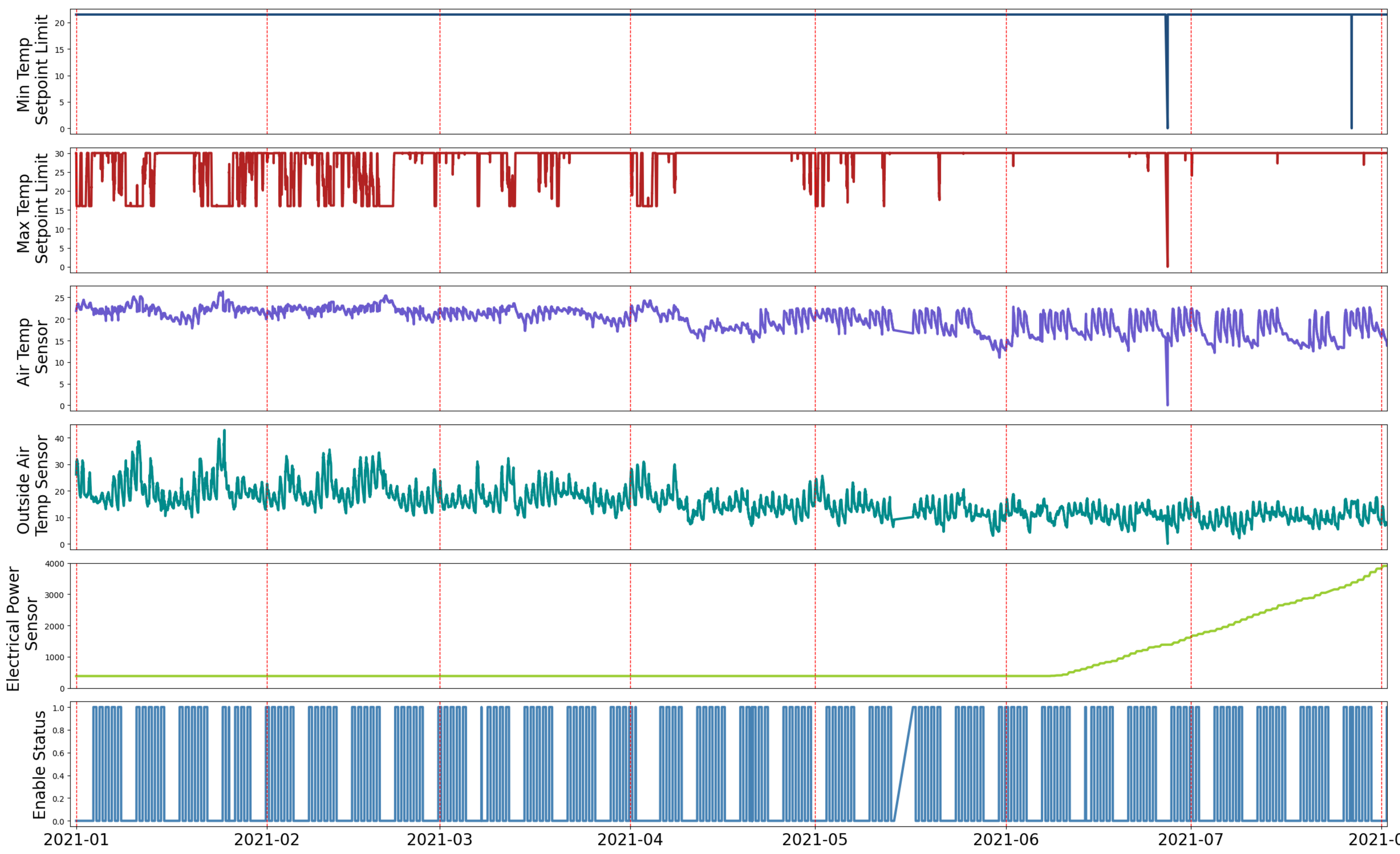
Dataset

The Building Time Series (BTS) dataset is a **multi-year time-series** dataset collected from **three anonymised buildings in Australia**. This dataset enables diverse research opportunities, including:

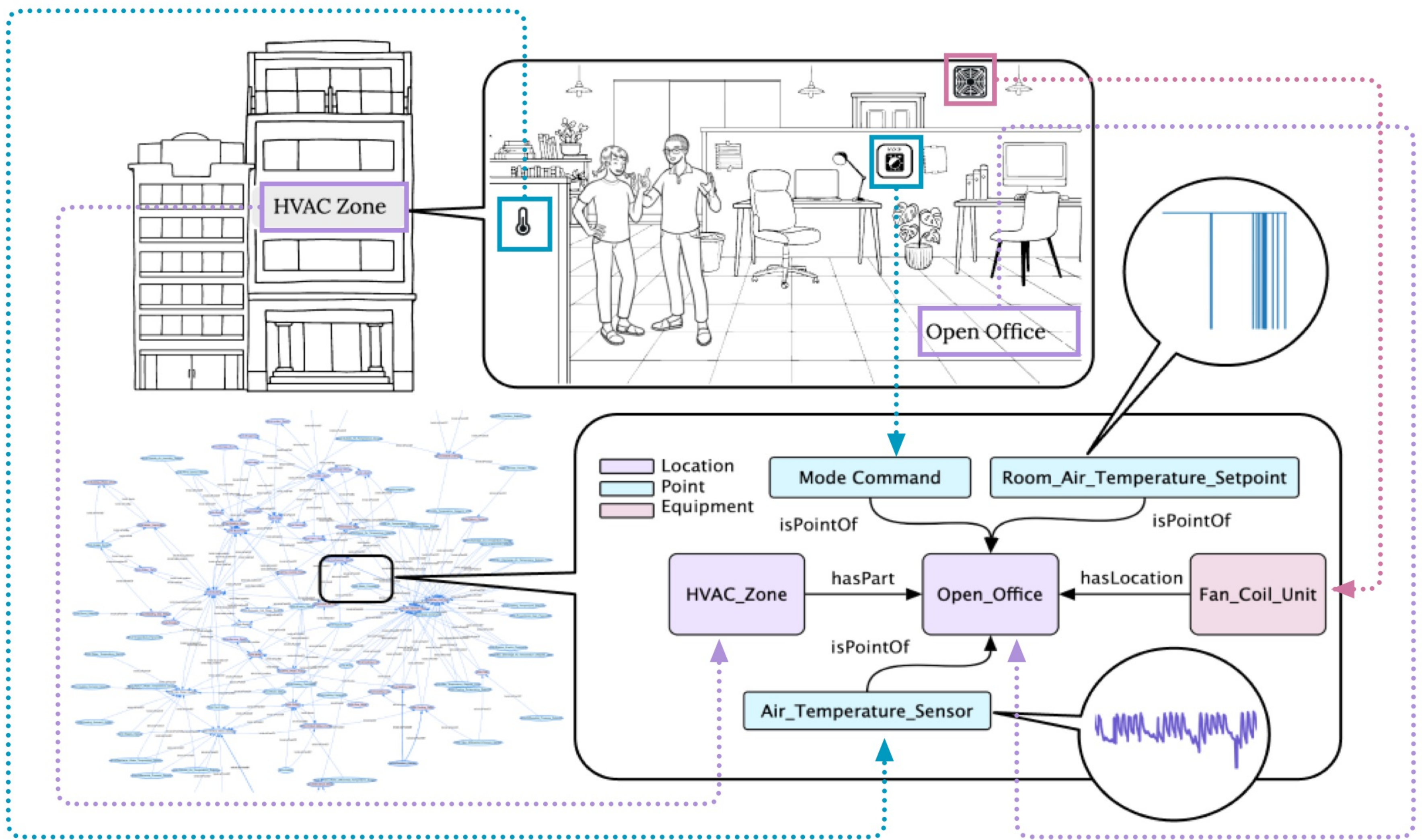
- Building analytics:
  - Optimizing energy, emissions, and occupant comfort.
  - Developing AI or LLM-powered chat systems and copilots for smart buildings.
- Fundamental machine learning research:
  - Addressing domain shift and domain adaptation.
  - Exploring multimodal learning with knowledge graphs.
  - Tackling unbalanced multivariate time series with long-tail distributions.

Spanning 2021 to 2024, it contains over **15,000 time series** across **300 unique classes**.

Visualisation of six timeseries from the snippet



In addition to the time-series data, BTS includes a **metadata** schema in the form of a **knowledge graph** that captures relationships between timeseries and their associated physical, logical, and virtual entities.



Summary statistics of our dataset, when compared with LBNL59, the only other comparable dataset in the existing literature.

(In brackets are the unique counts)

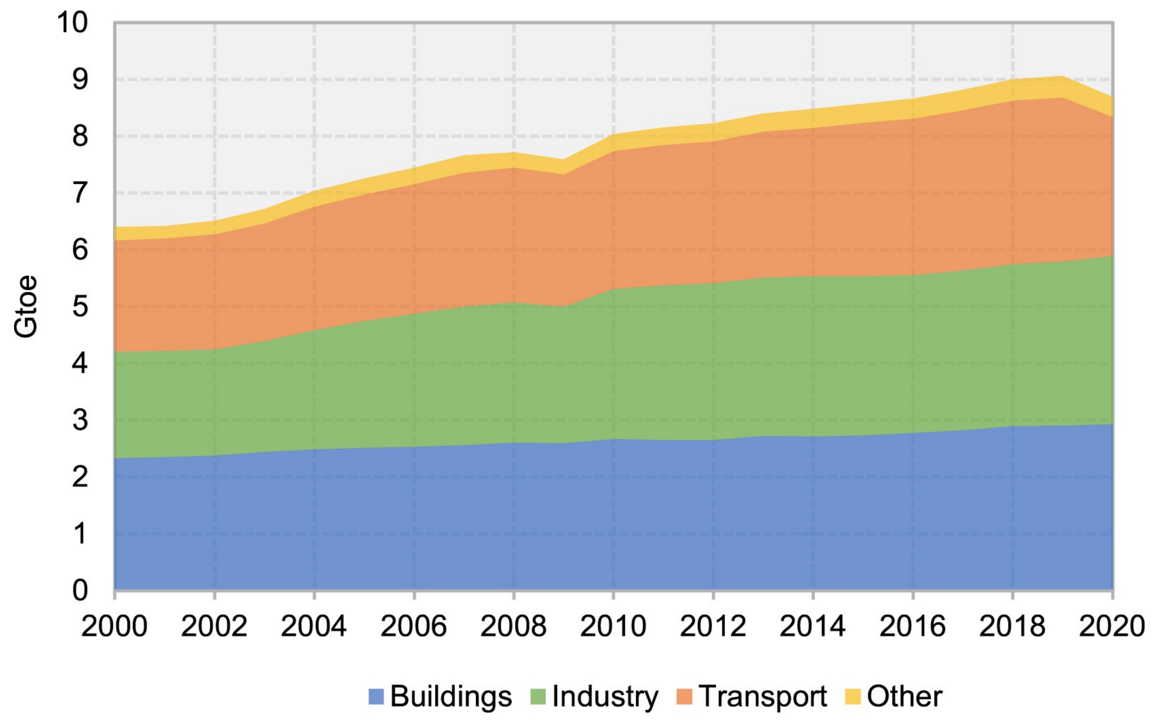
	Count (Unique)	LBNL59		BTS_A		BTS_B		BTS_C	
Top Level	Collection	0	(0)	4	(2)	2	(2)	8	(1)
	Equipment	59	(3)	547	(24)	159	(25)	963	(41)
	Location	73	(3)	481	(9)	68	(17)	381	(26)
	Point	230	(11)	8374	(126)	851	(57)	10440	(159)
	Timeseries	337		8349		851		5347	
Point Subclass	Alarm	0	(0)	798	(16)	5	(2)	109	(8)
	Command	0	(0)	363	(6)	97	(5)	785	(13)
	Parameter	0	(0)	79	(6)	36	(2)	935	(17)
	Sensor	144	(8)	4396	(56)	266	(25)	4062	(68)
	Setpoint	86	(3)	772	(26)	232	(16)	1629	(41)
	Status	0	(0)	1628	(17)	110	(6)	2187	(19)
	Location	Berkeley, USA		Undisclosed locations in Australia					
	Start Date	01-01-2018		01-01-2021		01-01-2021		23-06-2021	
	End Date	31-12-2020		31-12-2023		31-12-2023		18-01-2024	
	Duration (Days)	1094		1094		1094		939	
	Size Zipped (GB)	0.26		8.48		1.31		8.98	

Official repo: https://github.com/cruiseresearchgroup/DIEF_BTS/

This is a part of NSW's DIEF project https://research.csiro.au/dch/projects/nsw-dief/

Motivation:
Importance of building analytics

Buildings are responsible for a **third** of global **energy** consumption and a **quarter** of CO_2 emission.



(Gtoe: Giga tonnes of oil equivalent)

Table 1. Share of direct and indirect CO_2 emissions by sector in 2019.

Sector	*Direct*	*Indirect*	*Total*
Industry	19%	19%	38%
Buildings	9%	19%	28%
Transport	25%	3%	28%
Other	2%	4%	6%

M. Gonz.lez-Torres, L. P.rez-Lombard, J. F. Coronel, I. R. Maestre, and D. Yan. A review on buildings energy information: Trends, end-uses, fuels and drivers. Energy Reports, 8:626–637, 2022.

This research is supported by the NSW Government through the CSIRO's NSW Digital Infrastructure Energy Flexibility (DIEF) project, funded under the Net Zero Plan Stage 1: 2020-2030.

This project is also funded by the Reliable Affordable Clean Energy for 2030 (RACE for 2030) Cooperative Research Centre.

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

Literature Gap

Despite the importance and urgency in advancing building analytics, there is a lack of dataset on buildings with properties that are required. These properties include:

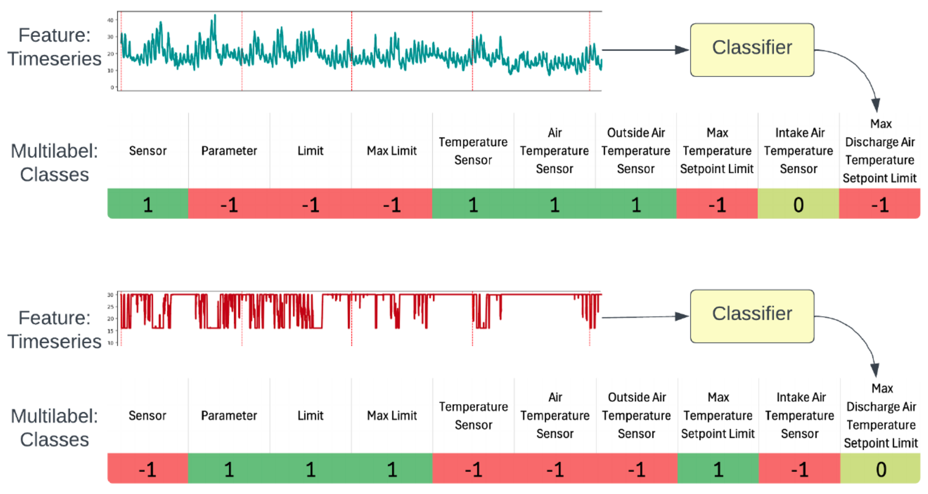
- **publicly available**, does not require permission from the data provider
- **freely accessible**, not paid
- on building **operations** (e.g. not blueprints)
- **real-world**, not simulation
- and **comprehensive.**

This is shown by a table of representative samples on the right.

	Datasets
Private	HVAC [70, 35, 79, 72, 32, 31, 21], energy use [62, 63], timeseries ontology classification [36, 37, 25, 44, 45, 68], and simulation [78].
Paid	Pecan Street [15].
Upon discretion of the data provider	ecobee [22]. Mortar [24] (Not freely available from the website (`https://mortardata.org/`) as per 13 August 2024, awaiting imrpovement in infrastructure).
Static	EUBUCCO [55], PLUTO [18], GBMI [10], Roofpedia [90], HBD3D [9], and Google Research's Open Buildings [76].
Corase temporal granularity (more than daily)	CBECS [17], BERTOOL [77], CENED+2 [69].
Simulation-based	BEM4CBECS [2, 94, 95, 93], ResStock [87], ComStock [60], CityLearn Challenge Series [84, 56, 59, 58], BuildingBench [23], and hardware-in-the-loop laboratory [67, 66].
Limited scope	SLRHOME [5], LCLD [81], and UCI [80]
NILM	Non-intrusive load monitoring (NILM) is task and many dataset have been made for this task check this recent survey [61] that list publicly available dataset. However, since the datasets are only made for this specific task in mind, the scope is limited to only electricity submetering. Other datasets with focus on submetering: BDG [54] and BDG2 [53].
Occupant behaviour	From AshraeOB [19, 49] website: "The ASHRAE Global Occupant Behavior Database aims to advance the knowledge and understanding of realistic occupancy patterns and human-building interactions with building systems. This database includes 34 field-measured occupant behavior datasets for both commercial and residential buildings, contributed by researchers from 15 countries and 39 institutions covering 10 different climate zones. It includes occupancy patterns, occupant behaviors, indoor and outdoor environment measurements."
Comprehensive	Lawrence Berkeley National Laboratory building 59 (LBNL59) [38, 51] and **BTS (ours)** `https://github.com/cruiseresearchgroup/DIEF_BTS`.

Benchmark 1:
Multi-label Classification

Brick schema was developed to aids in data interoperability across buildings, to ensure the scalability of smart building solutions. However, constructing the Brick schema for each building requires expensive and error prone manual expert labour to classifying timeseries data into the correct Brick classes. To automate this process, we formulate it as a multi-label classification task visualised below:



Method	Accuracy		F1		mAP	
Zero	0.8484	±N/A	0.0000	±N/A	0.0000	±N/A
Mode	0.8592	±N/A	0.1296	±N/A	0.0990	±N/A
Random Proportional	0.8147	±0.0001	0.1487	±0.0002	0.1520	±0.0001
Random Uniform	0.4999	±0.0002	0.1813	±0.0002	0.1520	±0.0001
One	0.1516	±N/A	0.2234	±N/A	0.1516	±N/A
LR	0.2366	±N/A	0.0882	±N/A	0.0497	±N/A
XGBoost	0.8593	±N/A	0.2697	±N/A	0.2627	±N/A
Transformer (default)	0.7807	±0.0139	0.3360	±0.0116	0.3171	±0.0078
Transformer (HP tuned)	0.8052	±0.0074	0.3615	±0.0079	0.3489	±0.0057
Informer	0.7627	±0.0010	0.3162	±0.0019	0.2849	±0.0030
DLinear	0.7030	±0.0042	0.2499	±0.0020	0.2494	±0.0010
PatchTST	0.7534	±0.0017	0.2981	±0.0014	0.2721	±0.0013

Benchmark 2:
Zero-shot forecasting

The advent of building digitalization presents significant opportunities for leveraging deep learning methods in building management systems for accurate forecasting. In practical applications, it is crucial for well-trained models to be applicable across diverse building scenarios without retraining costs. However, specific building constraints, operational variances, functionality differences, and data heterogeneity pose significant challenges in real-world settings. Models must adapt to dynamic ontology changes when applied to different buildings. Previous studies often rely on identical features and well-processed data, not reflecting the complexity of real-world scenarios. LBNL59, involving only one building, is insufficient for transfer learning studies. This study establishes a baseline for zero-shot forecasting using the BTS multivariate time series.

		BTS-A		BTS-B		BTS-C	
		MAE	SMAPE	MAE	SMAPE	MAE	SMAPE
Previous Day Persistence		0.5377	48.1539	0.4976	43.2985	0.5458	45.7014
Previous Week Persistence		0.6190	57.2713	0.5918	51.3867	0.6499	58.1922
BTS-A	DLinear	N/A		0.4324	35.9846	0.4262	36.2734
	PatchTST	N/A		0.3748	29.2570	0.3712	29.5552
	Informer	N/A		0.5968	49.2217	0.5920	51.9745
	iTransformer	N/A		0.4026	31.1924	0.3842	30.1102
BTS-B	DLinear	0.4940	41.2264	N/A		0.4206	35.3121
	PatchTST	0.4575	36.7689	N/A		0.3711	29.2135
	Informer	0.5233	45.9279	N/A		0.4592	39.7068
	iTransformer	0.4783	37.5907	N/A		0.3901	29.9940
BTS-C	DLinear	0.4858	40.7421	0.4158	34.1473	N/A	
	PatchTST	0.4542	36.9451	0.3723	28.9325	N/A	
	Informer	0.5213	46.6112	0.4602	39.7162	N/A	
	iTransformer	0.4859	39.5158	0.4262	32.6550	N/A	