Effects on a Deep-Learning, Seismic Arrival-Time Picker of Domain-Knowledge Based Preprocessing of Input Seismograms

Anthony Lomax[*1], Matteo Bagagli[2,3], Sonja Gaviano[4,5,(now at 2)], Spina Cianetti[4], Dario Jozinović[6], Alberto Michelini[3], Christopher Zerafa[4] and Carlo Giunchi[4]

[1]ALomax Scientific, Mouans-Sartoux, France,

[2]Dipartimento di Scienze della Terra, Università di Pisa, Via Santa Maria, 53, 56126 Pisa, Italy,

[3]Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Nazionale Terremoti, Via di Vigna Murata, 605, Roma, Italy,

[4]Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Pisa, via Cesare Battisti, 53, Pisa, Italy,

[5]Dipartimento di Scienze della Terra, Università degli Studi di Firenze, Via La Pira 4, Florence, Italy,

[6]Swiss Seismological Service (SED), ETH Zurich, Zurich, Switzerland

[*]Corresponding author: Anthony Lomax (anthony@alomax.net)

Supplementary Material

Files S1-3 (File_S1/2/3.zip)

Archives containing figures showing example waveforms and corresponding PhaseNet and DKPN processing and picking for the INSTANCE, ETHZ and PNW datasets. For explanation, see the caption to Figure 4 in the main text.

File S4 (File_S4_TrainDev_losses.pdf)

Figures showing train-validation loss-curves for selected training runs for all dataset sizes. RND: random seed for selected run; B: batch size; LR: learning rate.

Text S1

The individual datasets source-station pairs filtering criteria (applied to training, testing and validation subsets) are listed below:

1. INSTANCE
   a.a. station epicentral distance <= 100.0 km → `path_ep_distance_km`
   a.b. station's channel equal to "HH*" → `station_channels`
   a.c. S-P < 30 sec → fields: `trace_S_arrival_sample`, `trace_P_arrival_sample`
2. ETHZ
   a.a. the whole trace miniseed must be complete >= 0.99 → `trace_completeness`
3. PNW
   a.a. filter for earthquakes → `source_type`
   a.d. trace with no offsets or gaps → `trace_has_offset`
   a.e. the station must have all 3 components recorder → `trace_missing_channel`

Table S1. Parameters for training with INSTANCE dataset

Dataset	Training size	Validation size
NANO 3	*766*	*153*
NANO 2	*1,532*	*306*
NANO 1	*3,064*	*612*
NANO	*6,129*	*1,225*
MICRO	*12,259*	*2,451*
TINY	*24,518*	*4,903*
SMALL	*61,296*	*12,259*
MEDIUM	*153,240*	*30,648*
LARGE	*245,184*	*49,036*

Batch Size = 64
Learning Rate = 0.001
Optimizer = Adam
Max Epochs = 100
Patience = 5 epochs
Loss Improvement delta-stop = 0.0005

Text S2: Precision & Recall

The following figures show the statistics for precision and recall of DKPN and PhaseNet for both in-domain and cross-domain testing.

a)

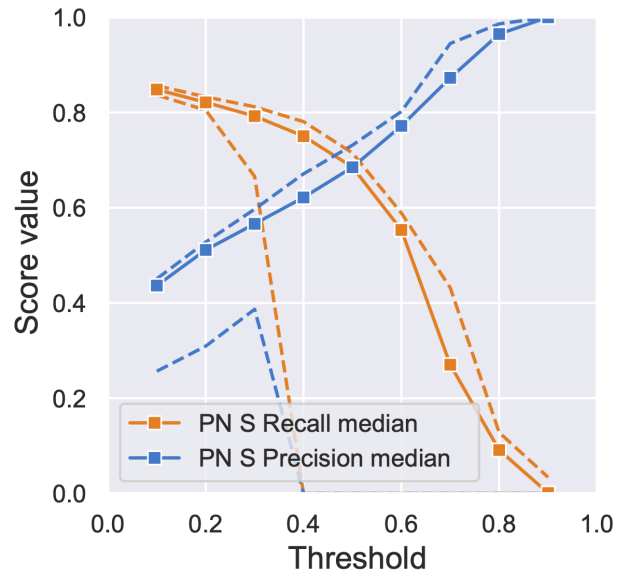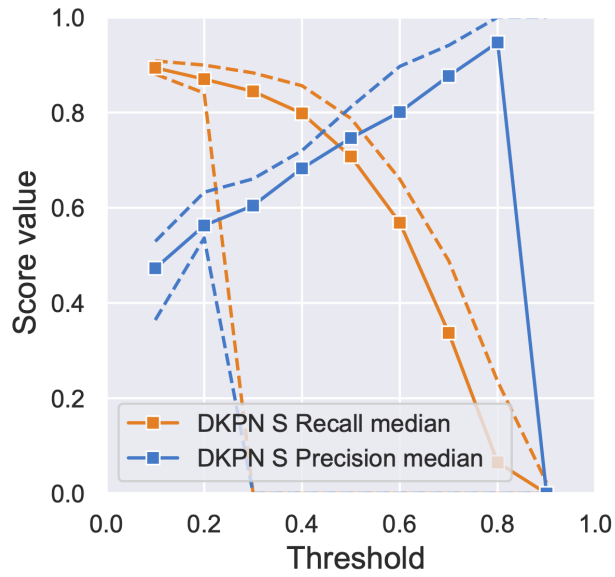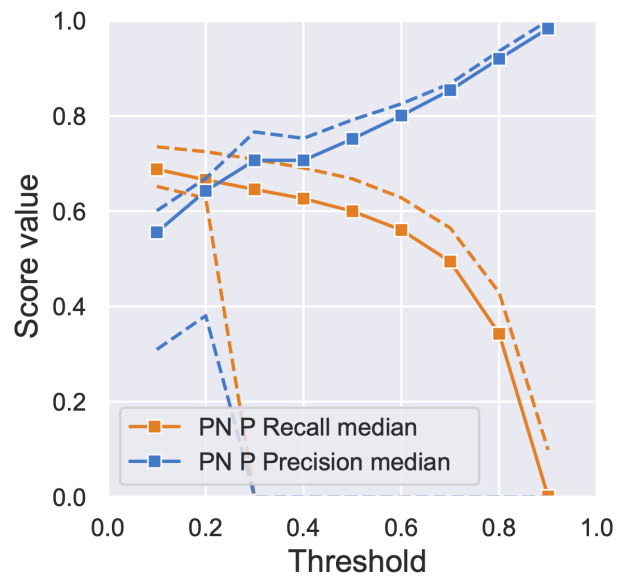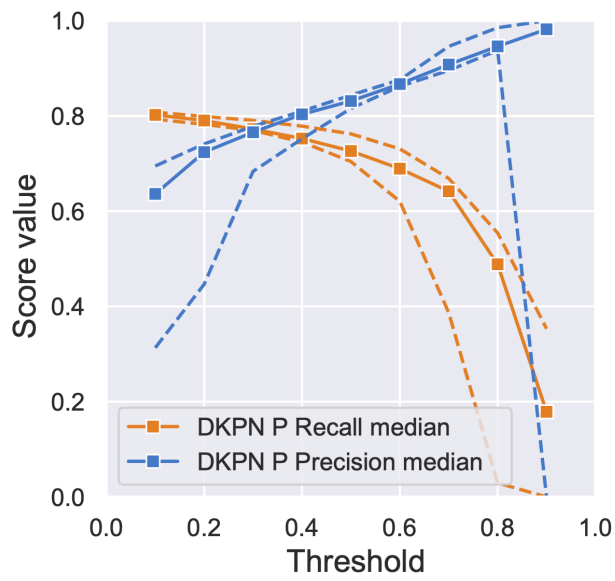INSTANCE - NANO2

b)

INSTANCE - MICRO

c)
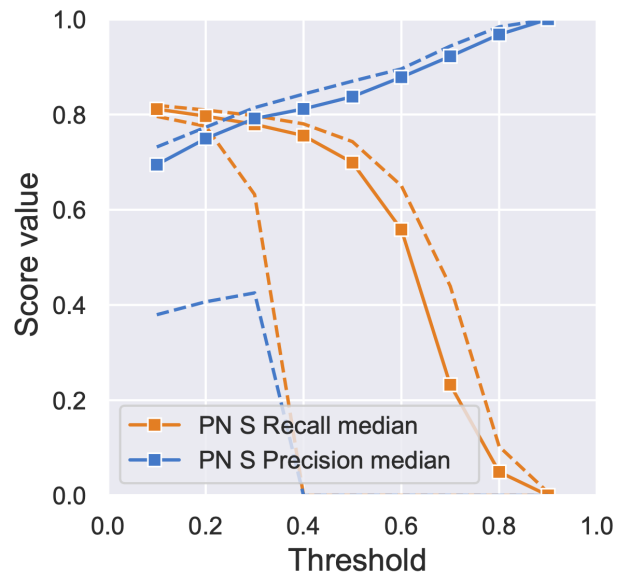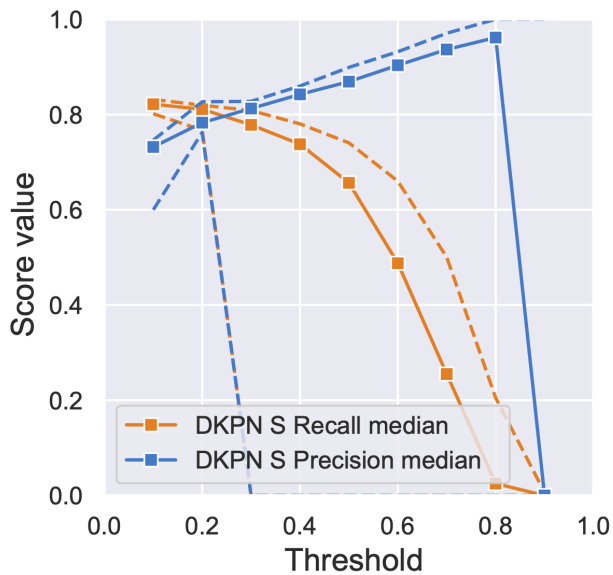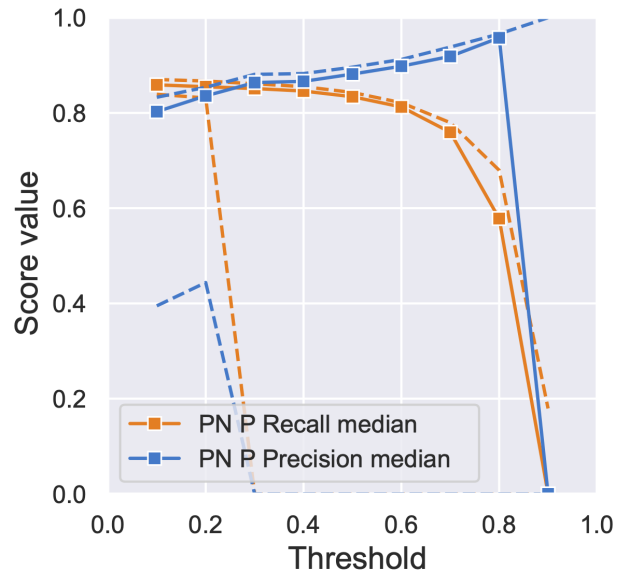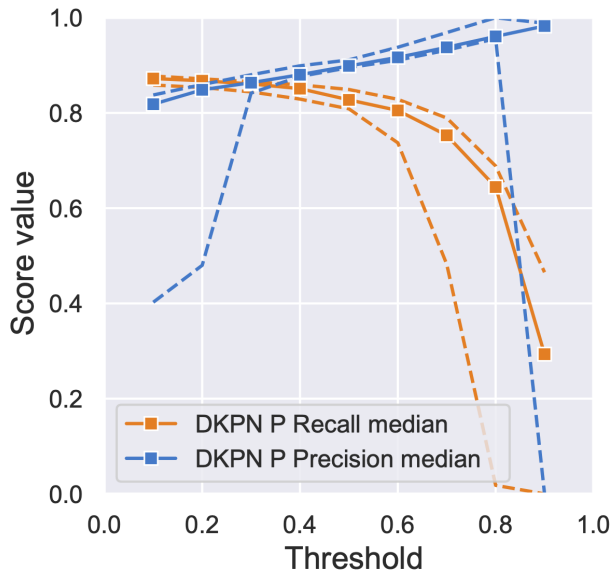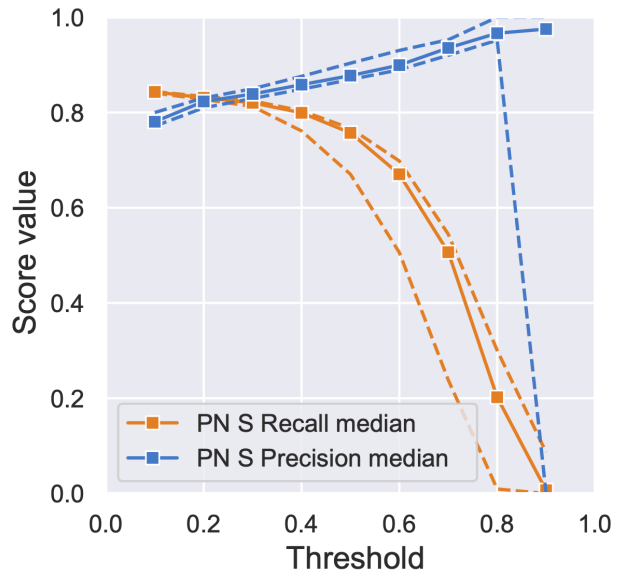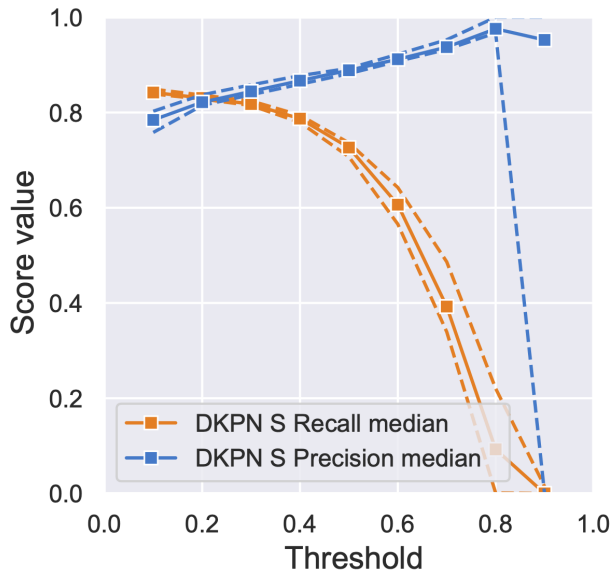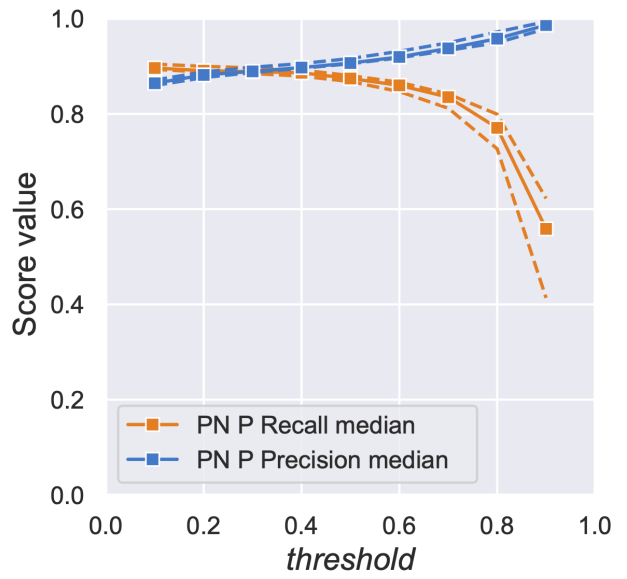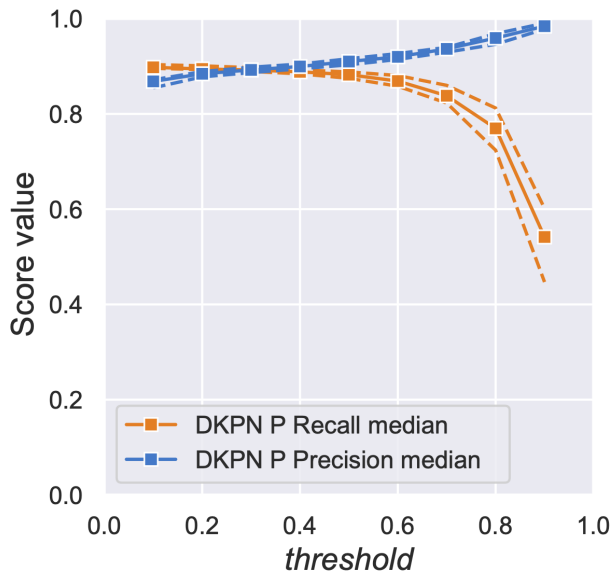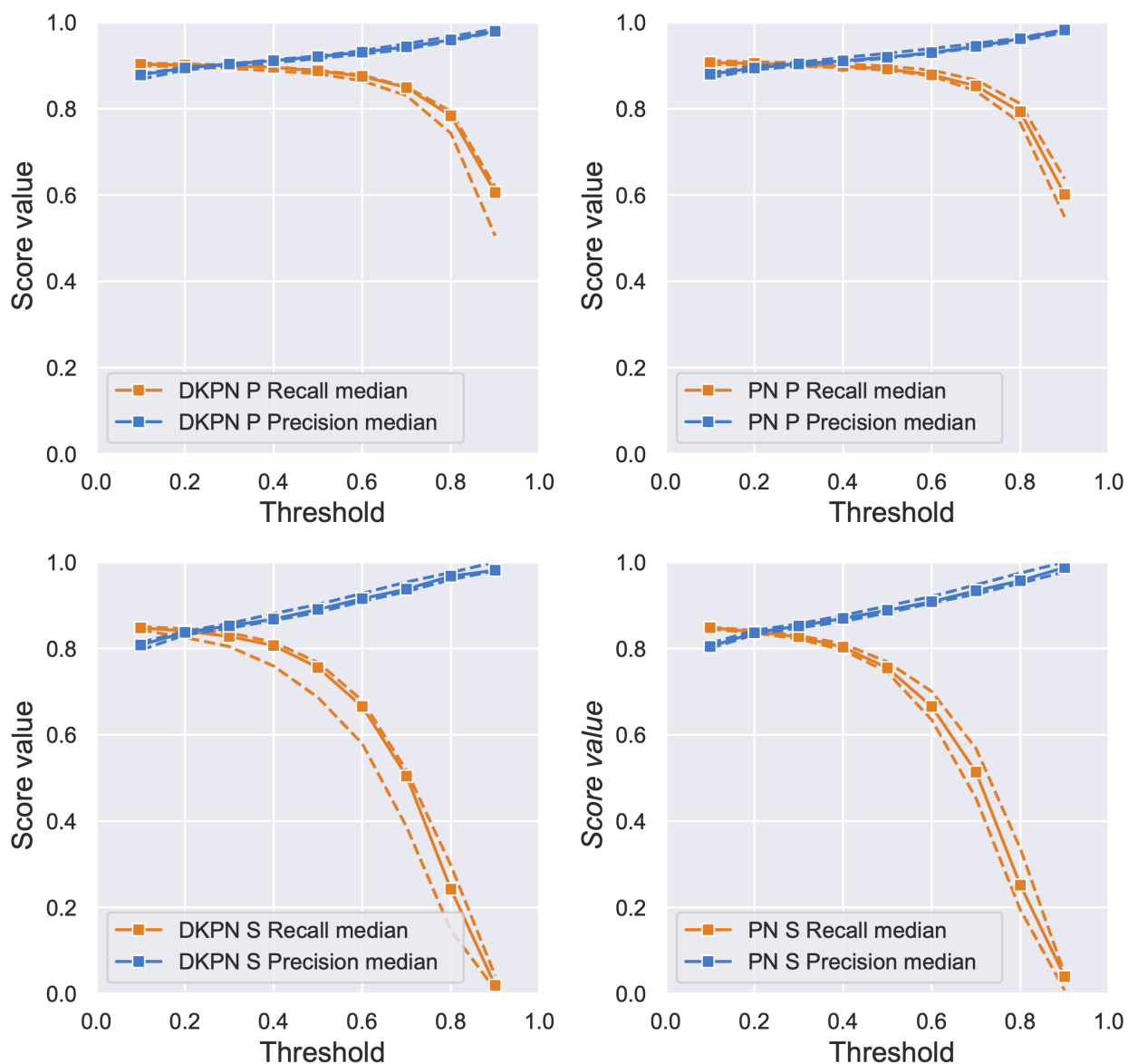
Figure S2.1: Precision and recall scores for DKPN in-domain testing with INSTANCE dataset. We run 7 experiments each with 5,000 evaluation samples drawn from the test dataset. Solid lines show the median score; dashed lines show the upper- and lower- boundary. **a)** the results obtained with the NANO2 model **b)** the results obtained with the MICRO model **c)** the results obtained with the MEDIUM model.
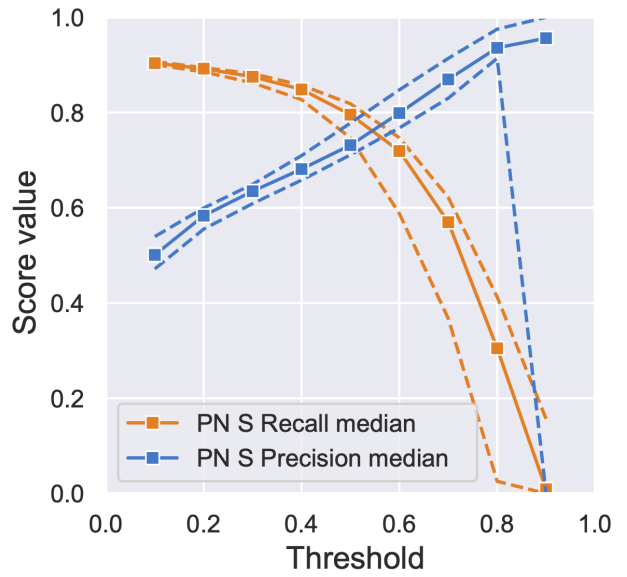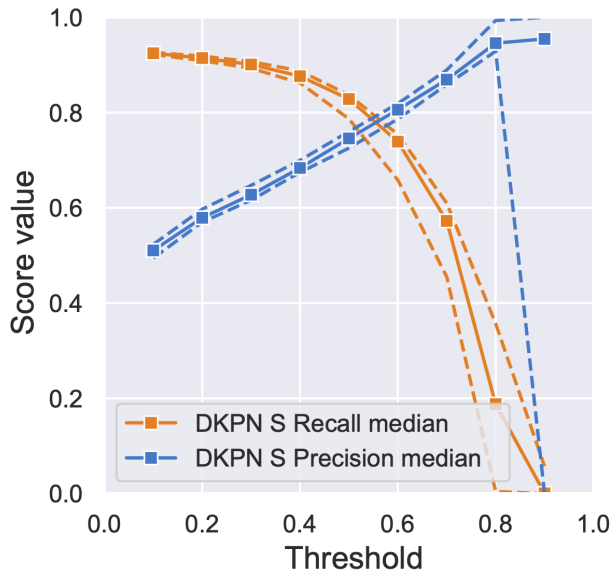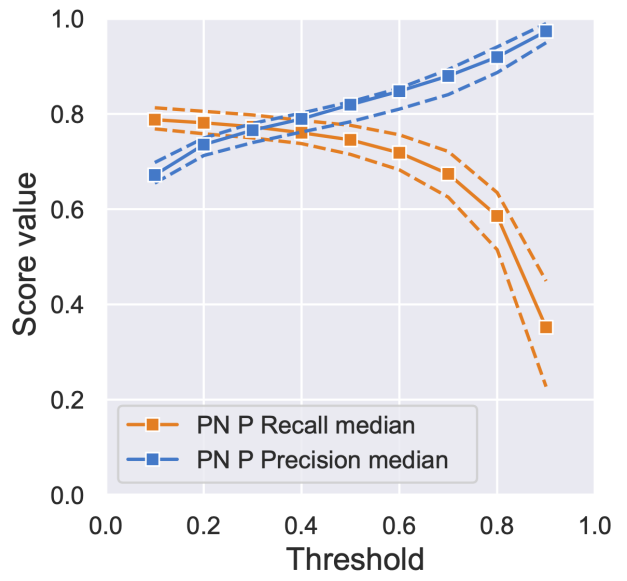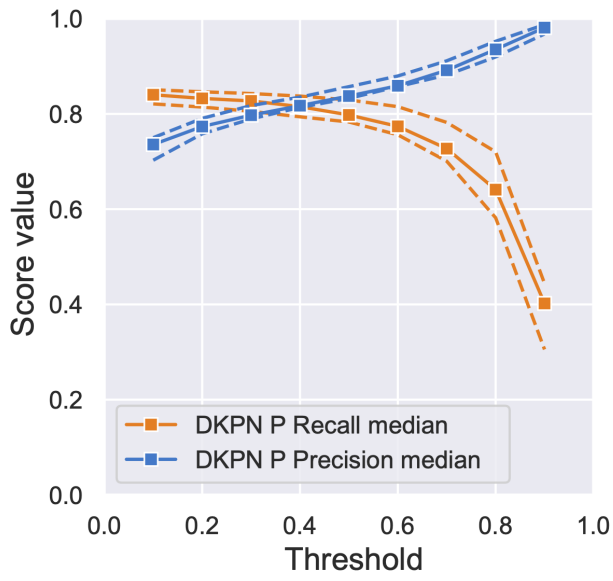
ETHZ - NANO2
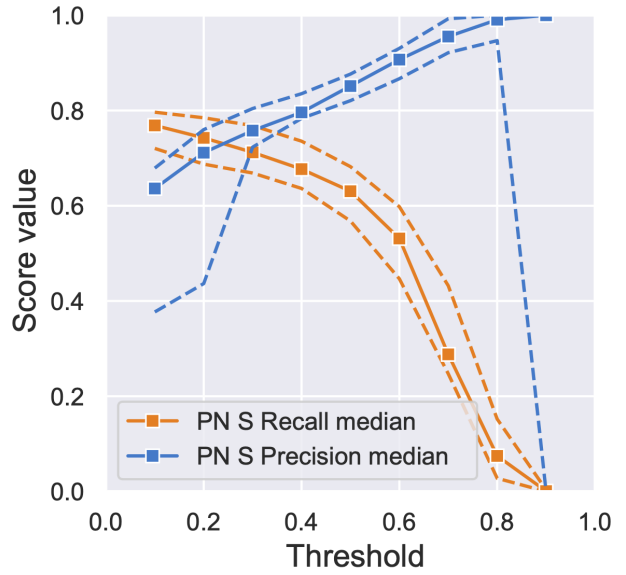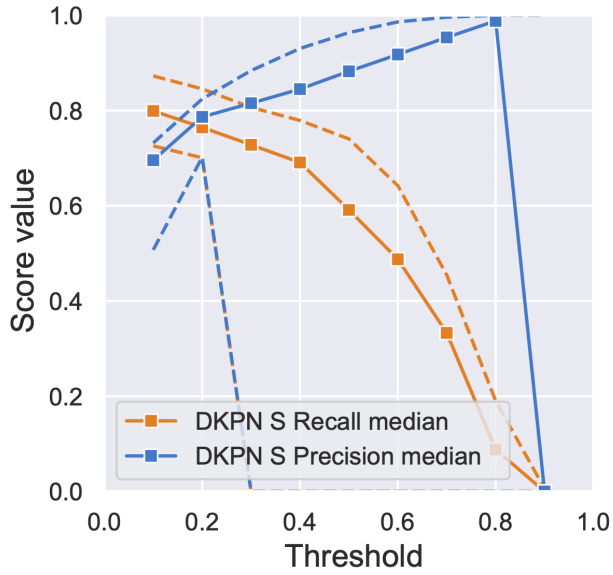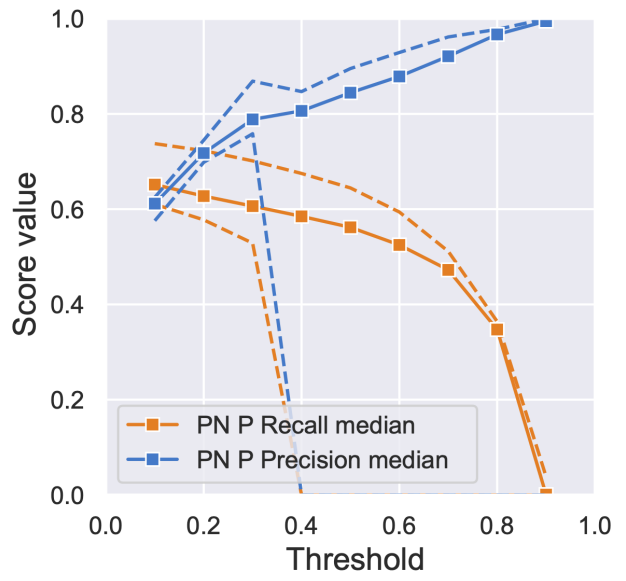
b)

ETHZ - MICRO
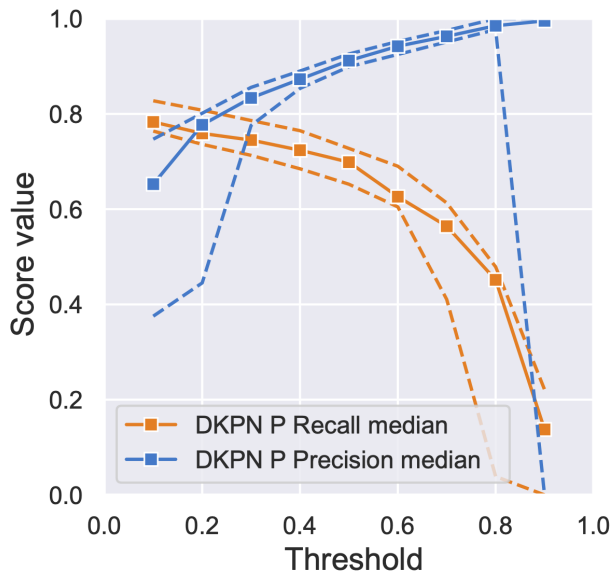
c)

ETHZ - MEDIUM
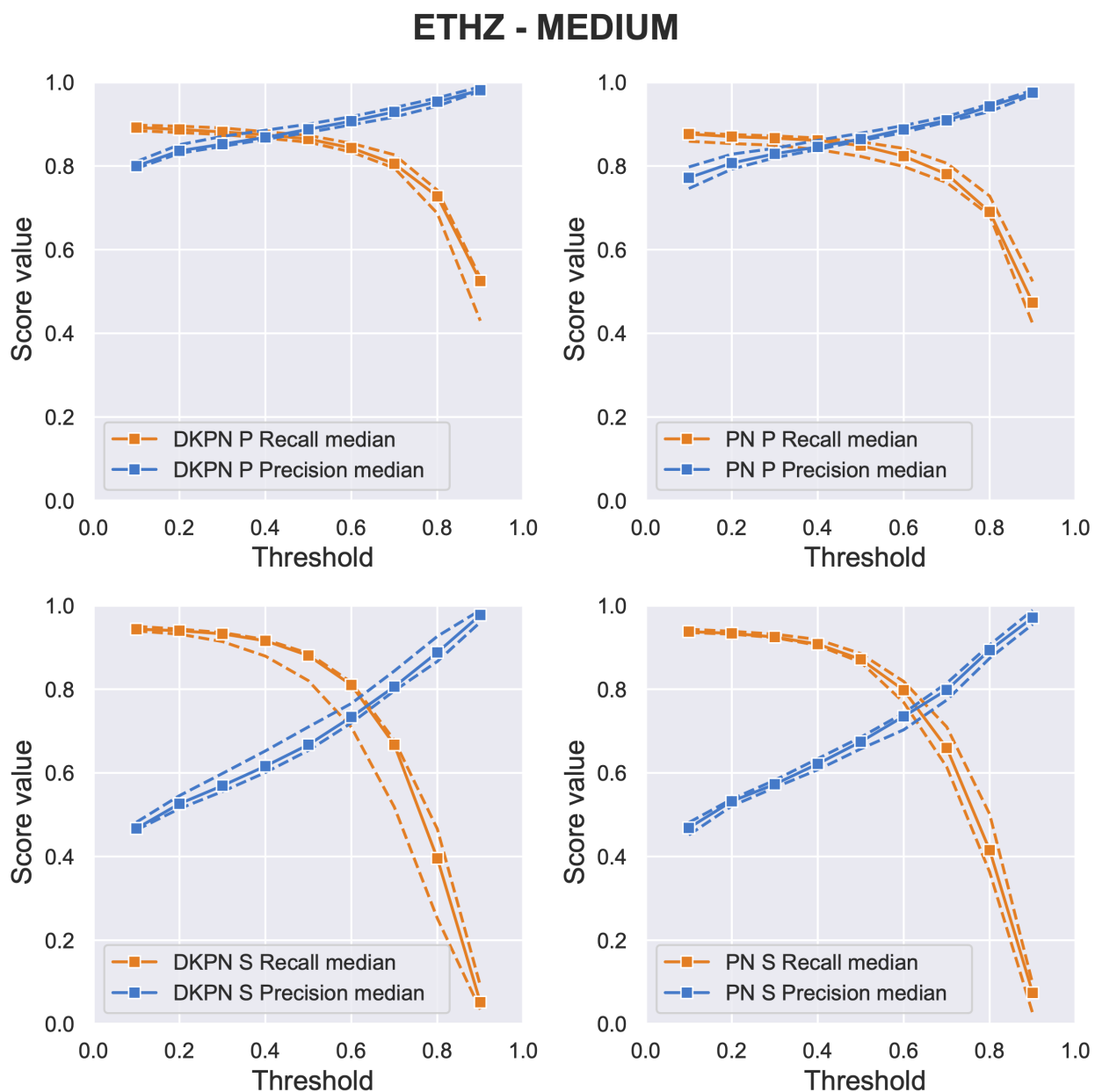
Figure S2.2: Precision and recall scores for DKPN cross-domain testing with ETHZ dataset. We run 7 experiments each with 5,000 evaluation samples drawn from the test dataset. Solid lines show the median score; dashed lines show the upper- and lower- boundary. **a)** the results obtained with the NANO2 model **b)** the results obtained with the MICRO model   **c)** the results obtained with the MEDIUM model.
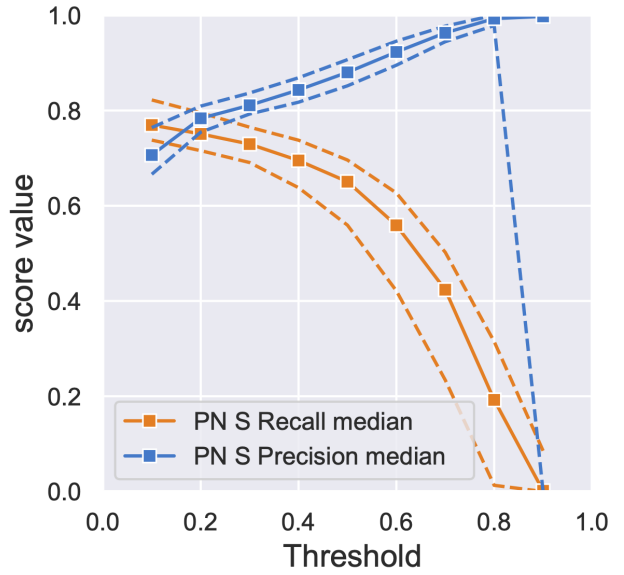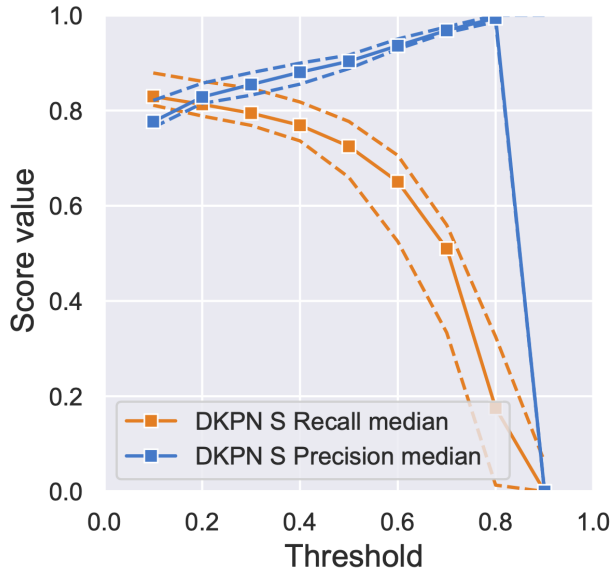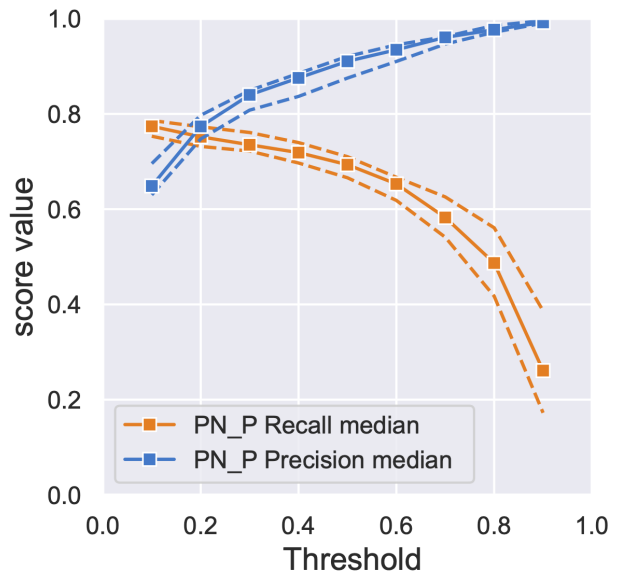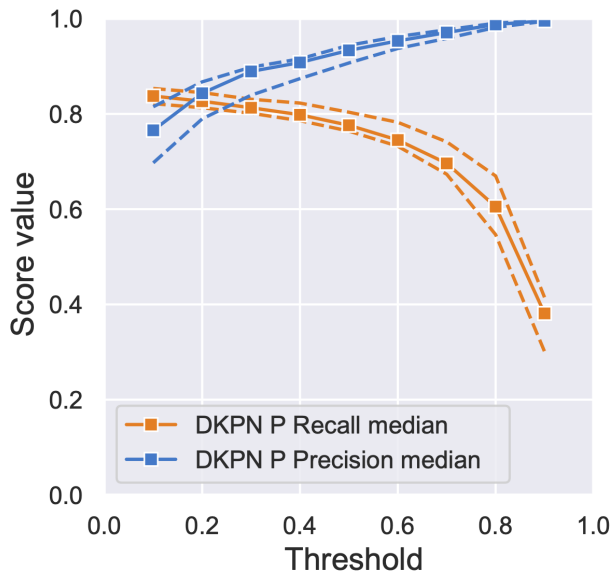
a)

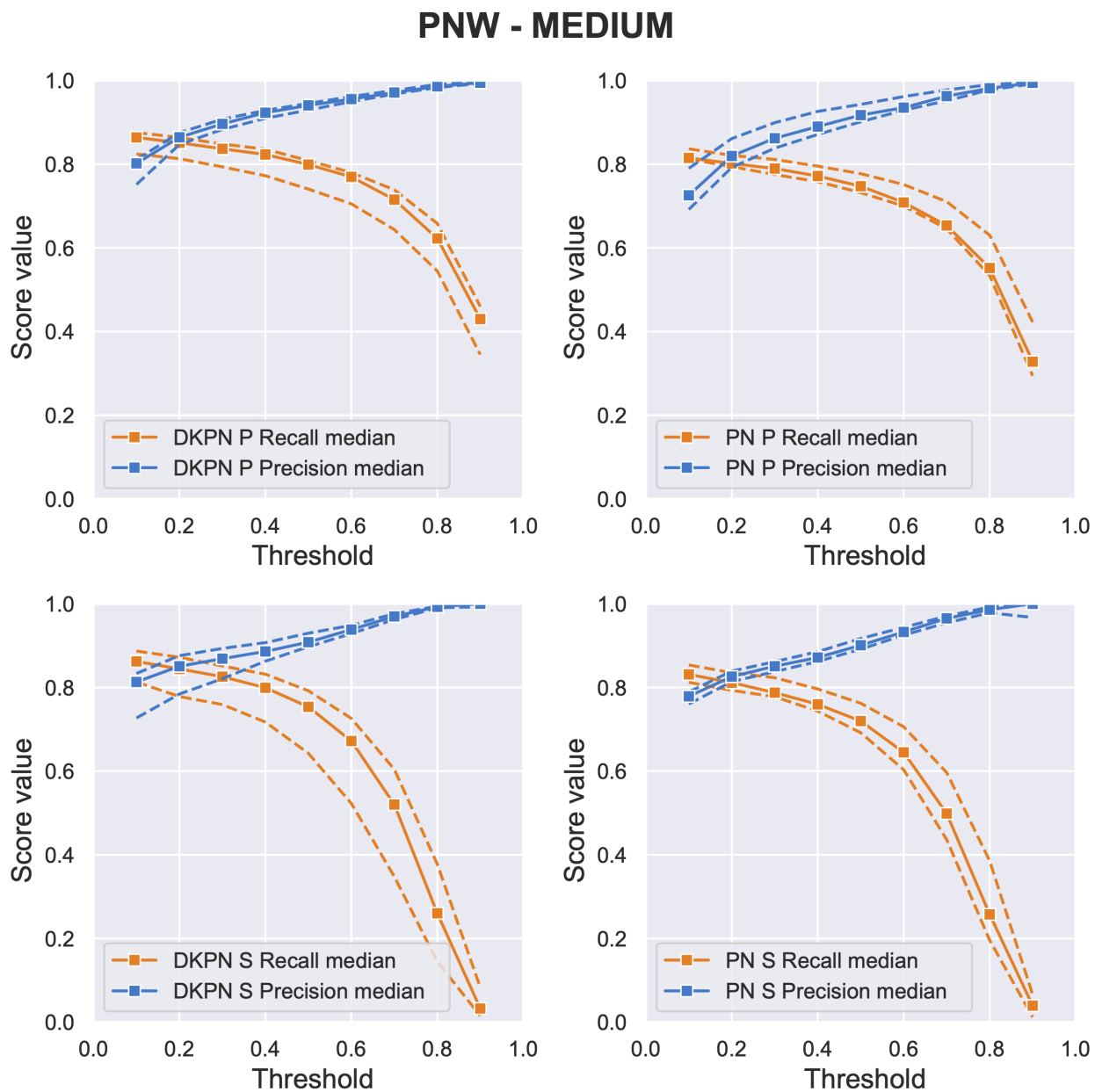
PNW - NANO2

b)

PNW - MICRO

c)

Figure S2.3: Precision and recall scores for DKPN cross-domain testing with PNW dataset. We run 7 experiments each with 5,000 evaluation samples drawn from the test dataset. Solid lines show the median score; dashed lines show the upper- and lower- boundary. **a)** the results obtained with the NANO2 model **b)** the results obtained with the MICRO model **c)** the results obtained with the MEDIUM model.