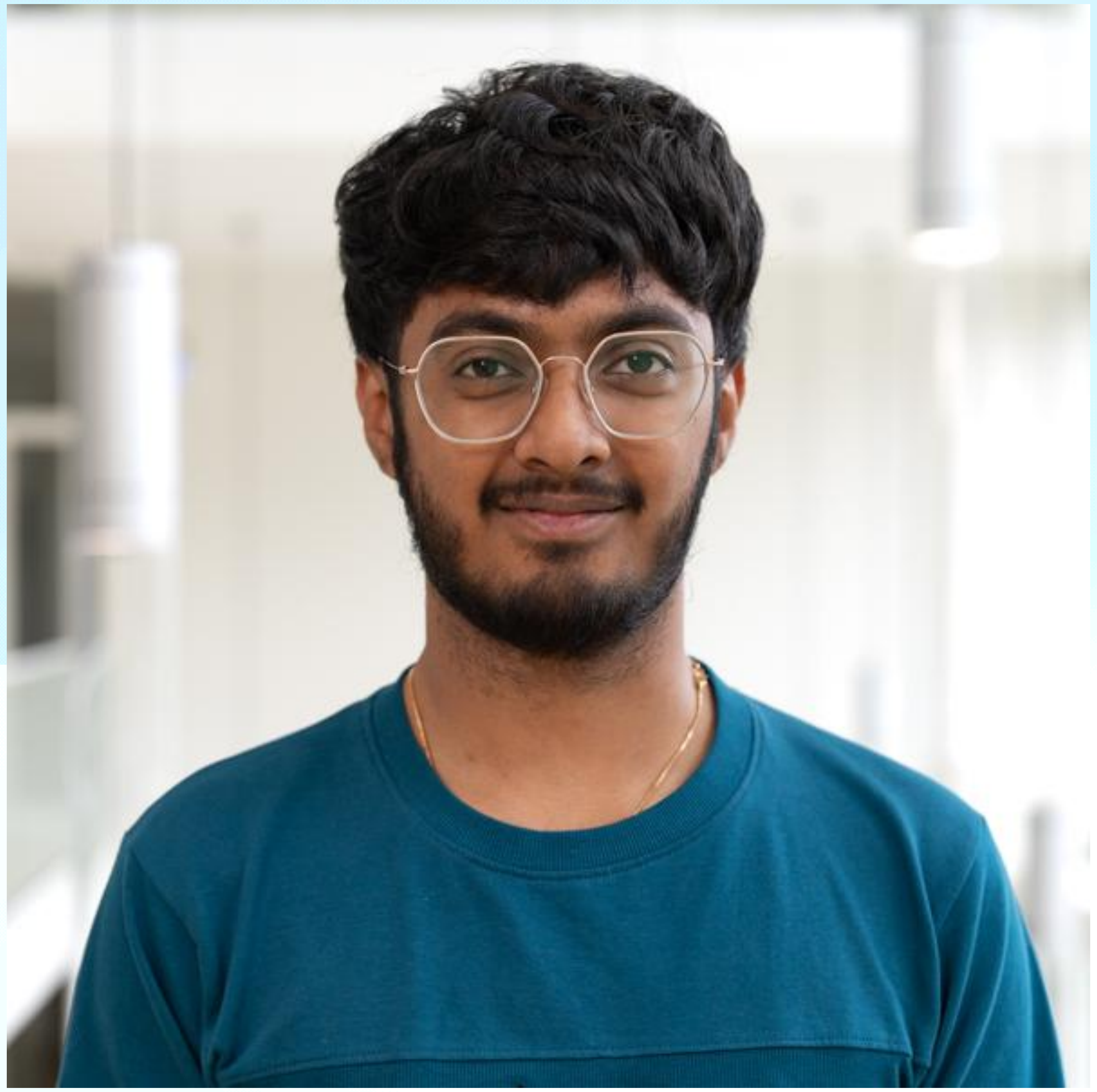
Optimizing Foundation models for anomaly detection

Humberto Reyes González

With Soumya Shaw, Salome Fresenbet, Anna Hallin, Gregor Kasieczka, Michael Krämer, Louis Moreaux, Alexander Mück.



Soumya Shaw

Saarland Univerity

Goal of the Study

- **CAN WE OPTMIZE A BIT THE FOUNDATION MODEL FRAMEWORK FOR FULLY AGNOSTIC ANOMALY DETECTION PURPOSES?**

- Which pretraining strategies work better?
- Which model selection criteria works better?

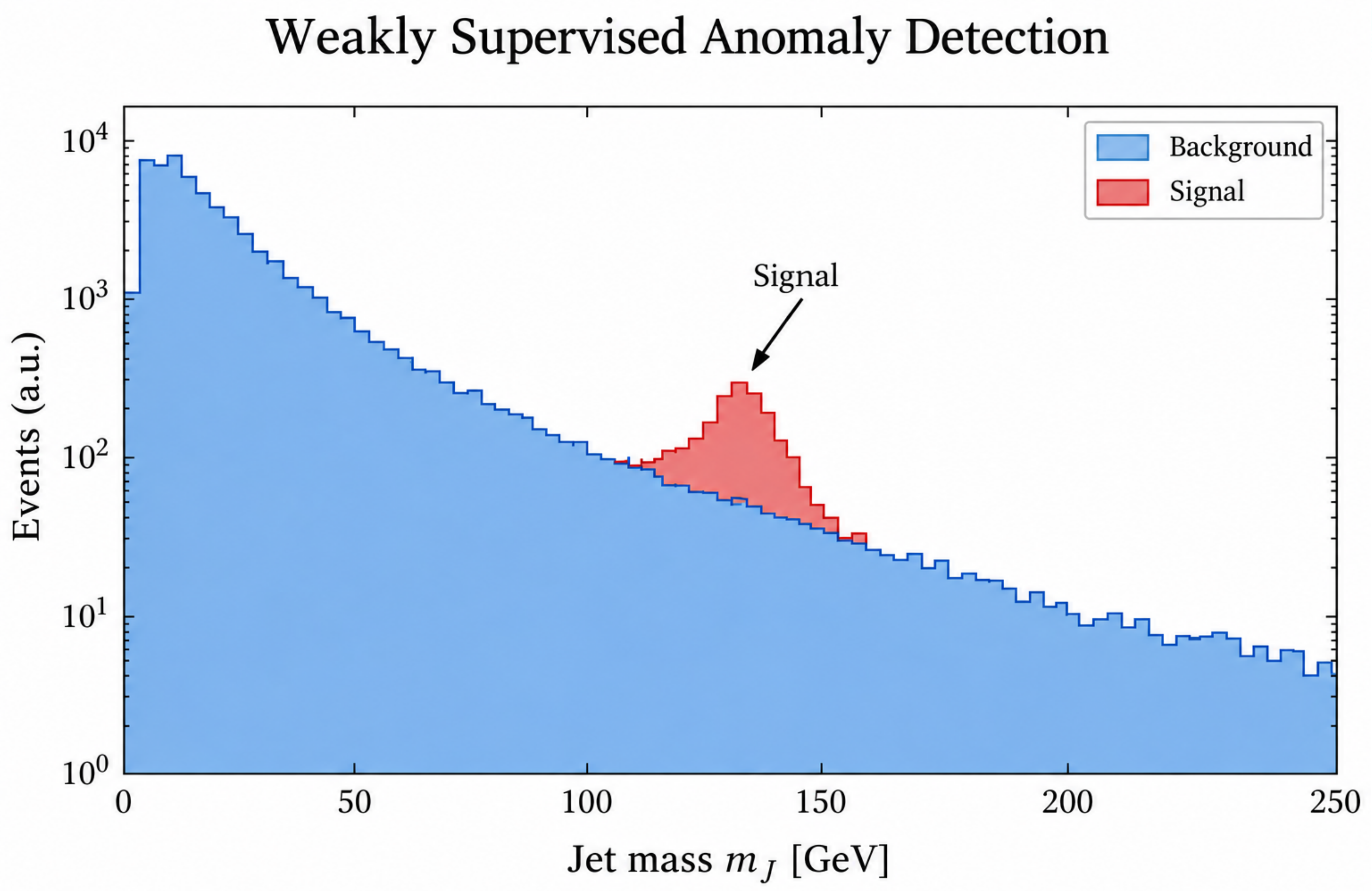
Outline

- Idealised Anomaly Detection
- The Datasets
- The architecture of OmniJet-α
- Experimental Setup
- Pre-training Strategies
- Results
- Conclusions

Idealized Anomaly Detection

The Setup:

- **Signal Region (SR)**: contains mostly background + a tiny fraction of signal
- **Background template (BT)**: contains pure Background (simulating a perfect density estimator)



Low Level Features Used

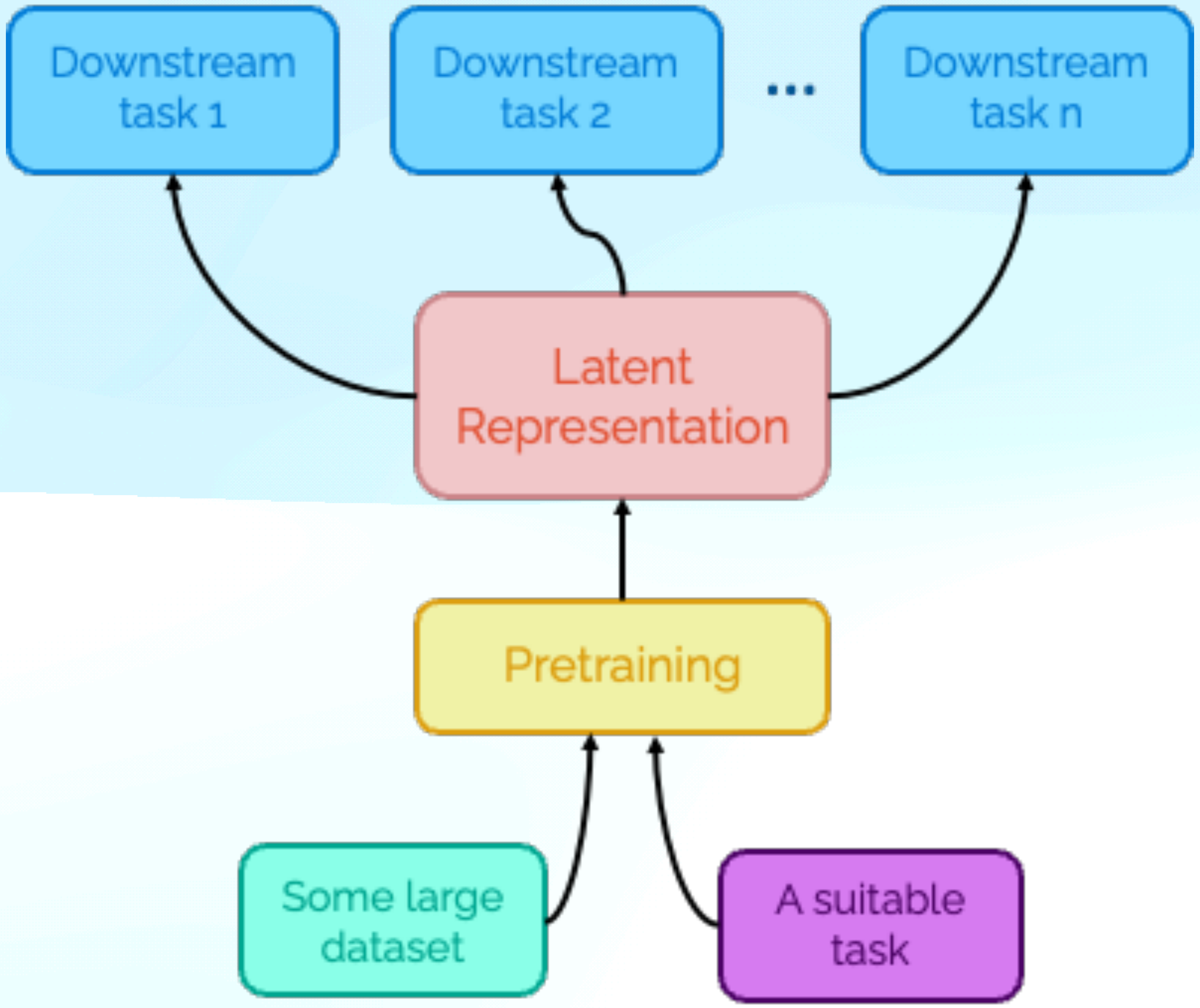
Previous approaches

- High-level features
- Engineered jet-level observables
- Pt jet, subjettiness, EFP

Our approach

- Low-level particle features
- P_T, η_{rel} and ϕ_{rel} for each constituent
- OmniJet-α learns representations directly from the constituents

What are Foundation Models?



[2509.21434]

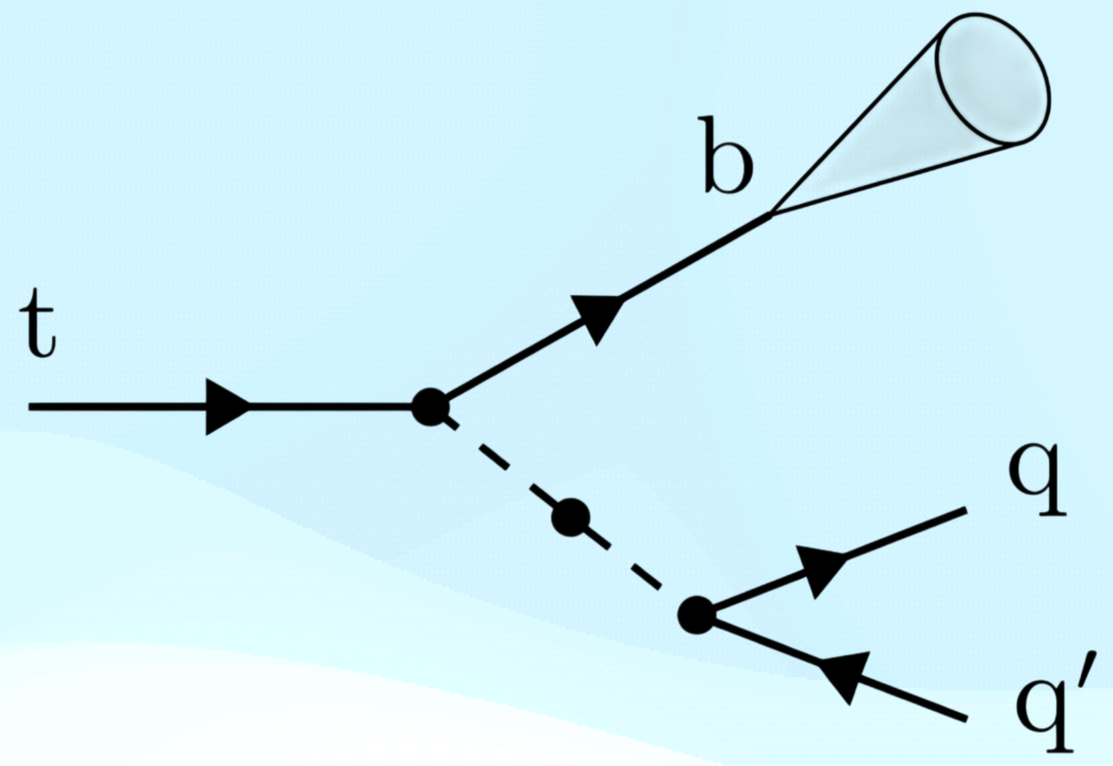
- Large neural networks **pre-trained on broad data**.
- Transfer to many downstream tasks with **fine-tuning**.
- Enables good performance even on smaller datasets.
- Captures general patterns in high-dimensional data.
- Let's try **Anomaly Detection**.

The Datasets

JetClass dataset:

- simulated **single** jet events
- 100 million jets in **10 classes**

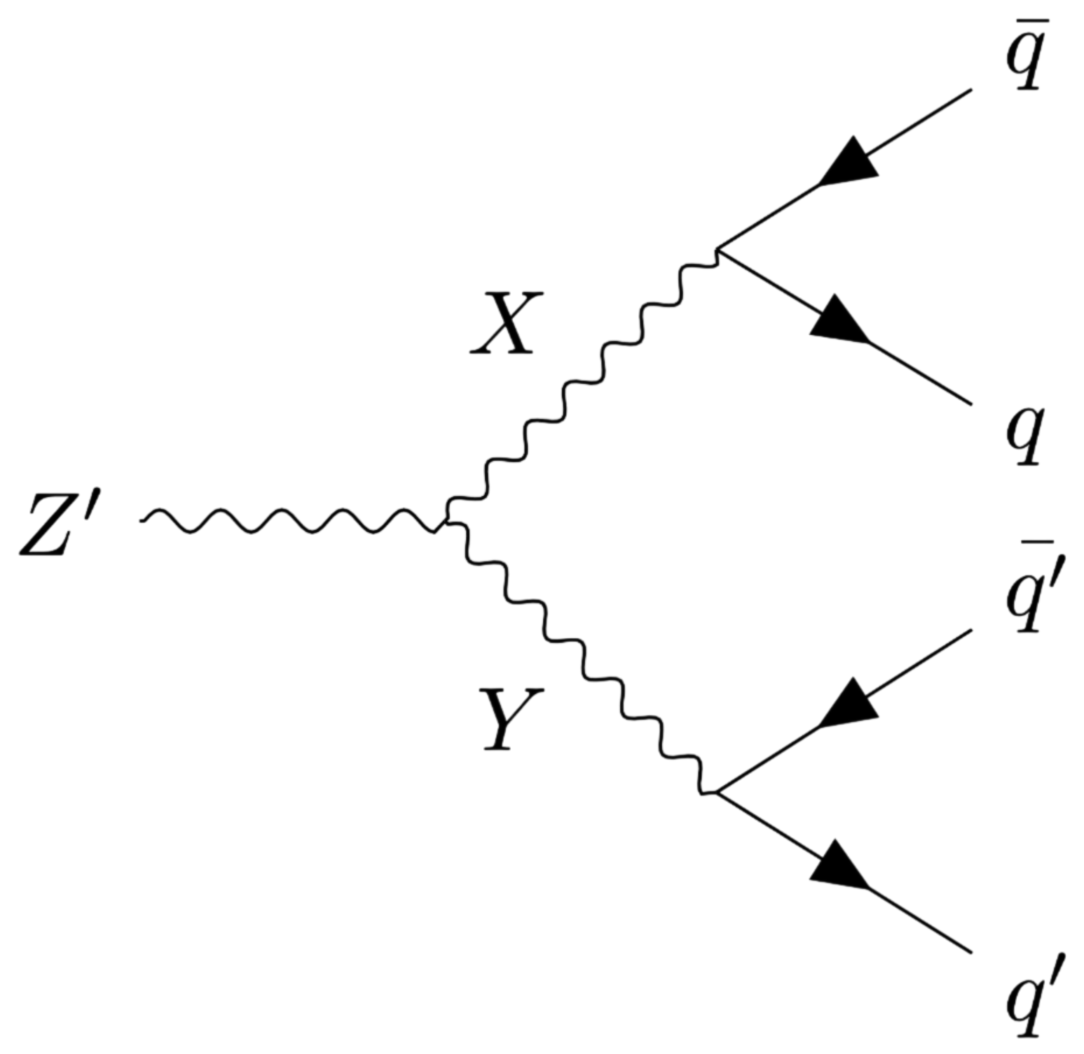
Used for **Pre-training**



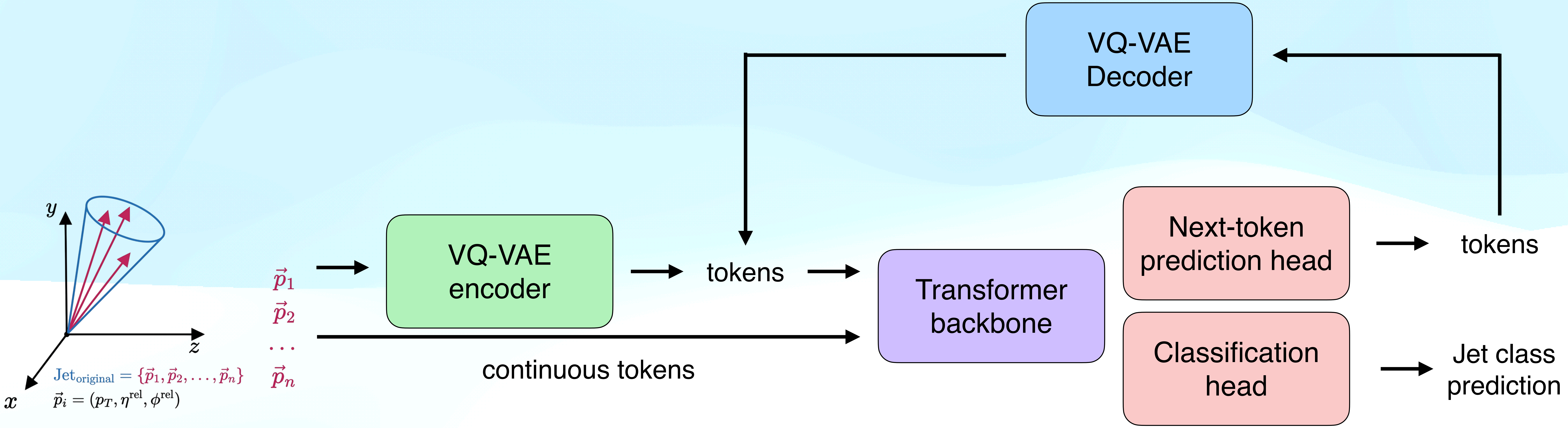
LHCO R&D dataset:

- simulated **dijet** events
- 1 million SM **QCD** jets as background
- 100k BSM jets **Z’ → XY → 4q**

Used for **fine-tuning**



Introducing OmniJet-α



The architecture of OmniJet-α
[2403.05618]

Experimental Setup

Idealised Anomaly Detection (IAD)

Background
QCD-only

350k jets

Signal
QCD + x% Z'

$0.1\,\% \leq x \leq 10\,\%$

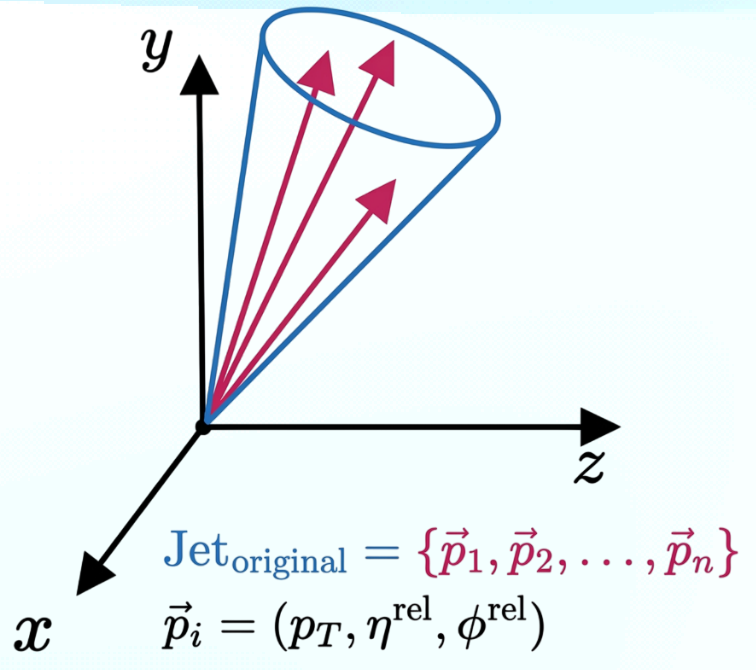
100k jets

OmniJet-α: Pre-training architectures

Generative Pre-training

start token → Transformer backbone → Next-token prediction head → tokens

Supervised Pre-training

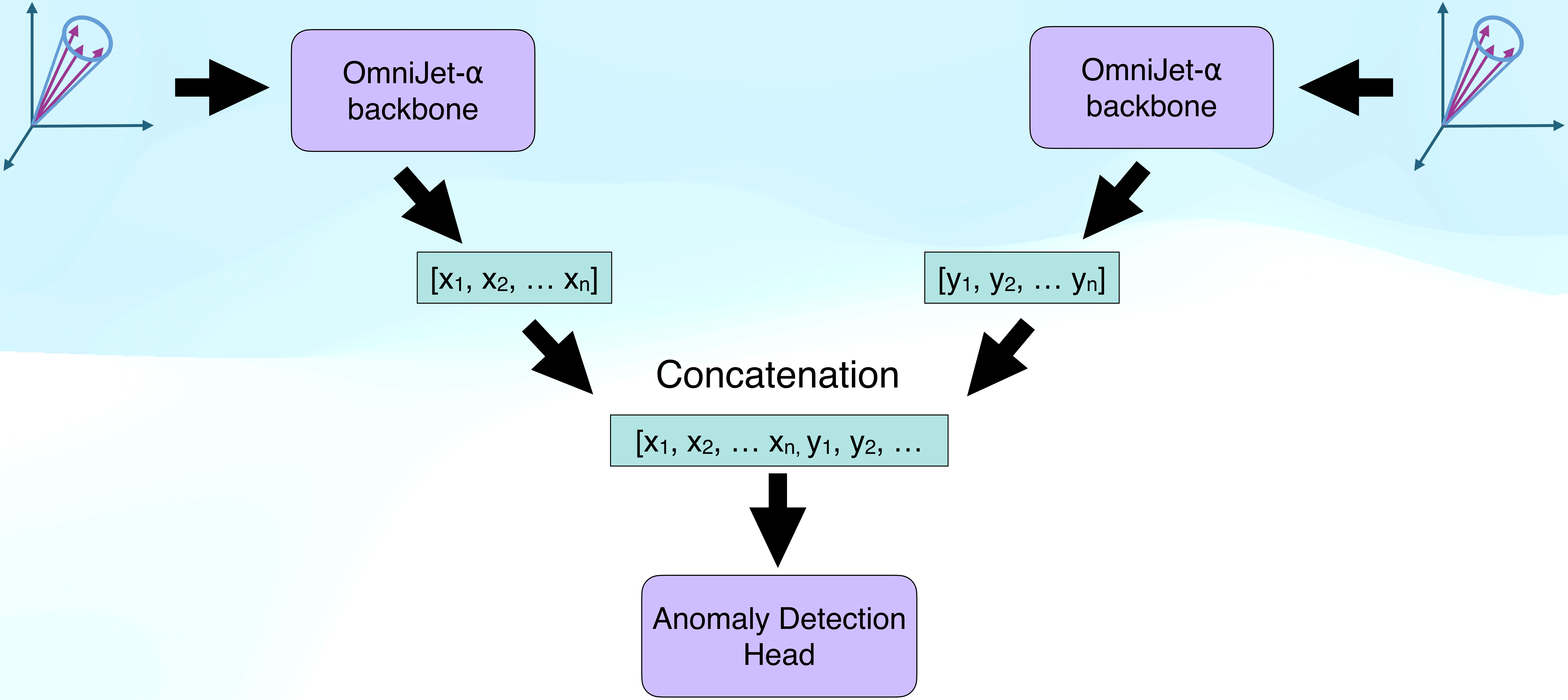


$\vec{p}_1$
$\vec{p}_2$
$\cdots$
$\vec{p}_n$

→ continuous tokens → Transformer backbone → Classification head → Jet class prediction

[2403.05618]

Incorporating DiJet Context



ARGOS metric

arXiv:2511.14832

Above Random Gain of SIC

$$\mathrm{ARGOS}(t) = \frac{\epsilon_{\mathrm{SR}}(t)}{\sqrt{\epsilon_{\mathrm{BT}}(t)}} - \sqrt{\epsilon_{\mathrm{BT}}(t)}$$

- $\epsilon_{\mathrm{SR}}(t)$: fraction of signal-region events passing threshold (t)
- $\epsilon_{\mathrm{BT}}(t)$: fraction of background-template events passing the same threshold

Interpretation

- $\mathrm{ARGOS} \approx 0$: no excess relative to the background template
- $\mathrm{ARGOS} > 0$: anomalous enrichment in the signal region
- $\mathrm{ARGOS} < 0$: preference for background-template events

Epochs with the largest maximum ARGOS are selected

$$\max_t \mathrm{ARGOS}(t)$$

Key advantage: does not require signal truth labels, but does require a representative background template.

Pre-training Setups

Generative Pre-training	Generative Pre-training	Supervised Pre-training
1 class - QCD only	10 classes	10 classes
10 million jets	1 million jets each	1 million jets each

Pre-training Setups

Generative Pre-training

1 class - QCD only

10 million jets

Generative Pre-training

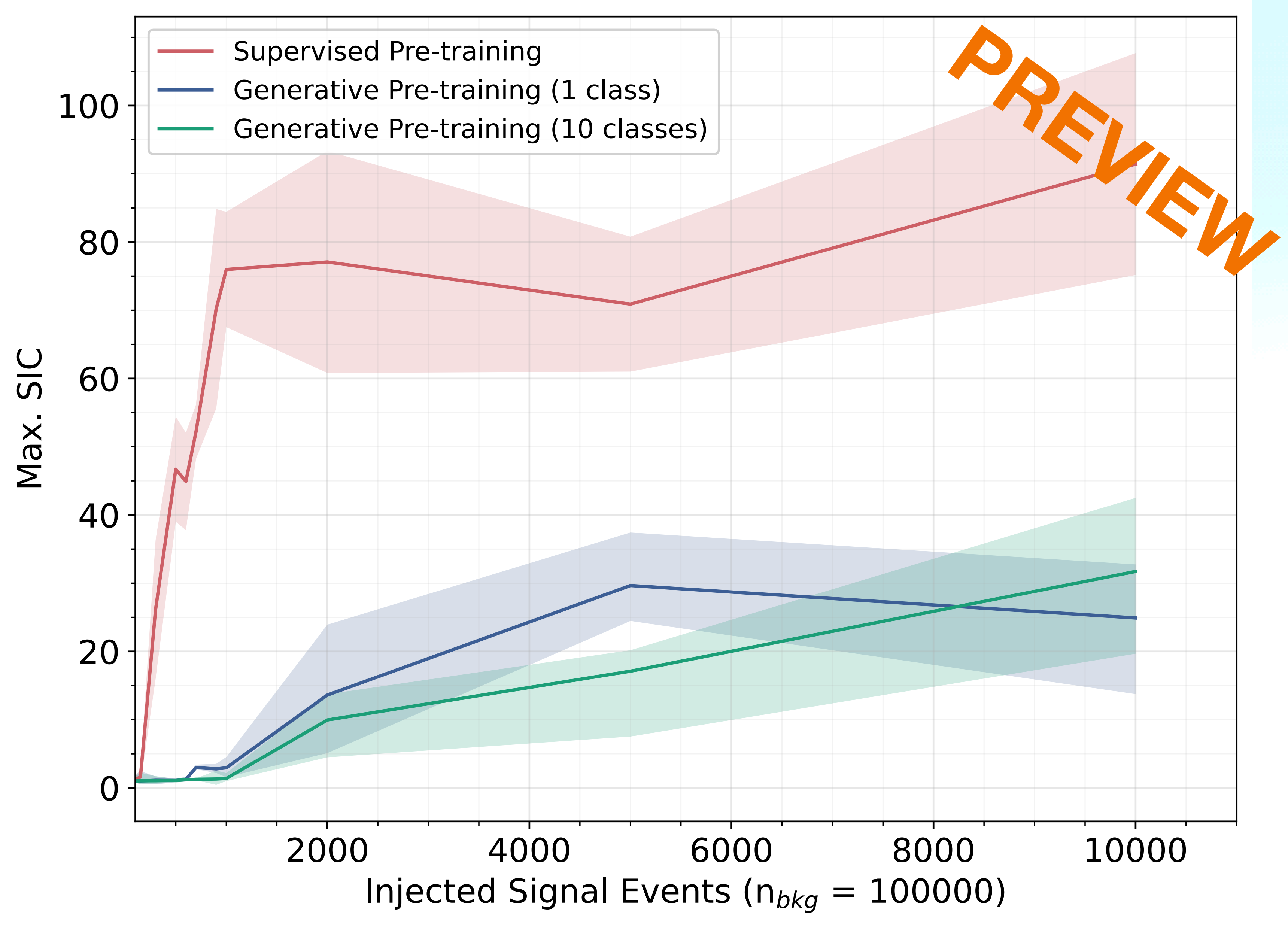
10 classes

1 million jets each

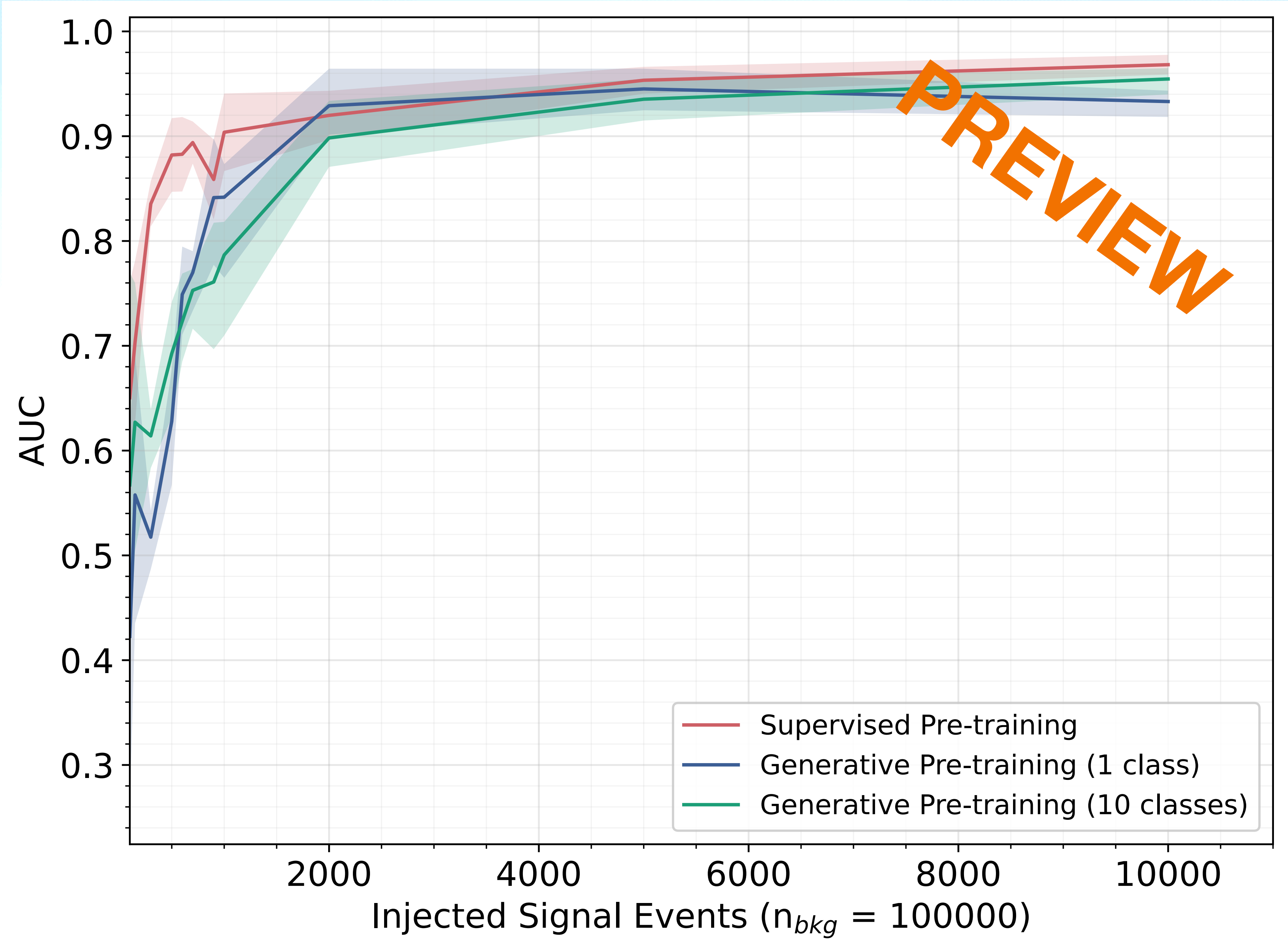
Supervised Pre-training

10 classes

1 million jets each

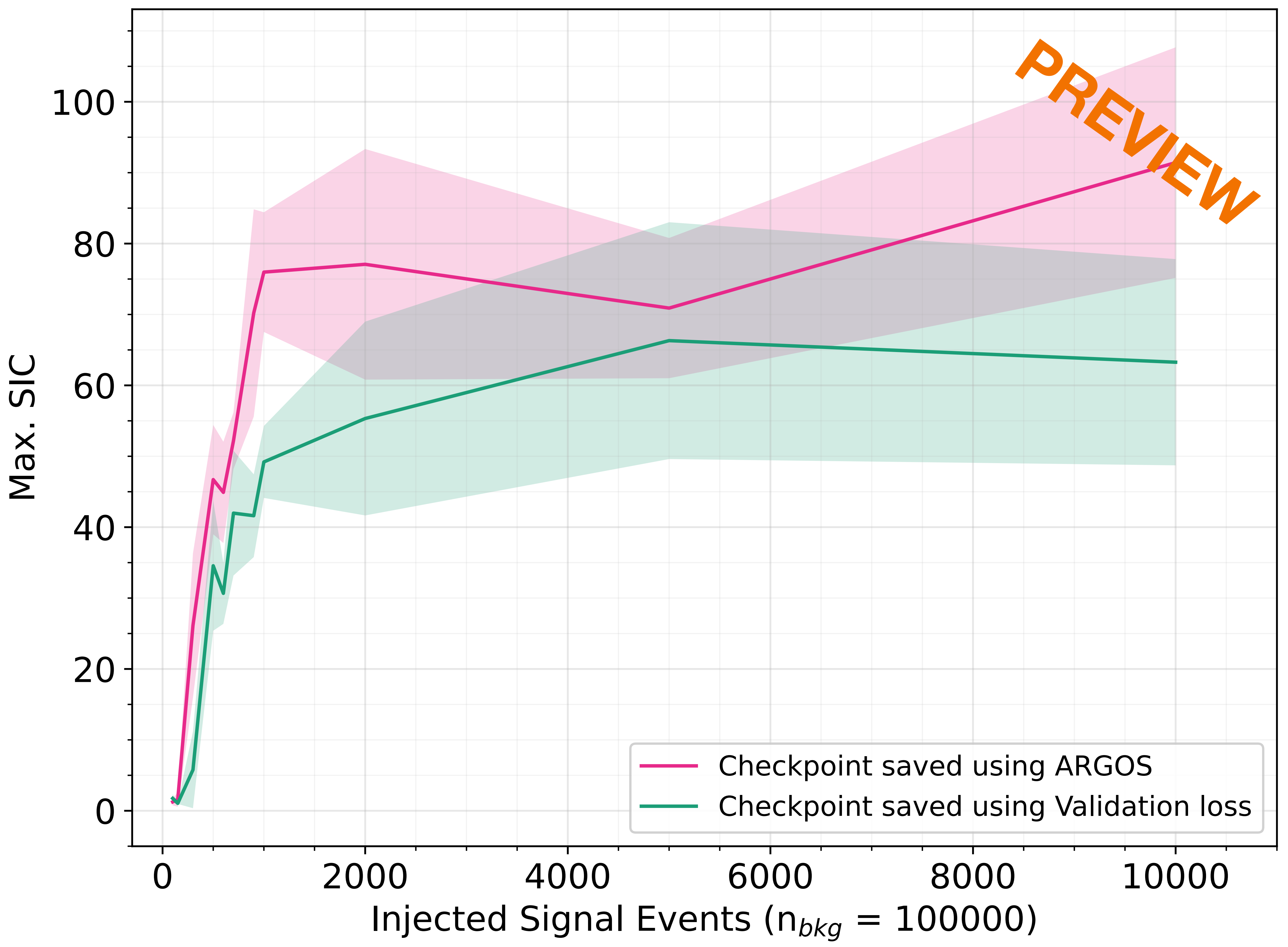


Max SIC Curve

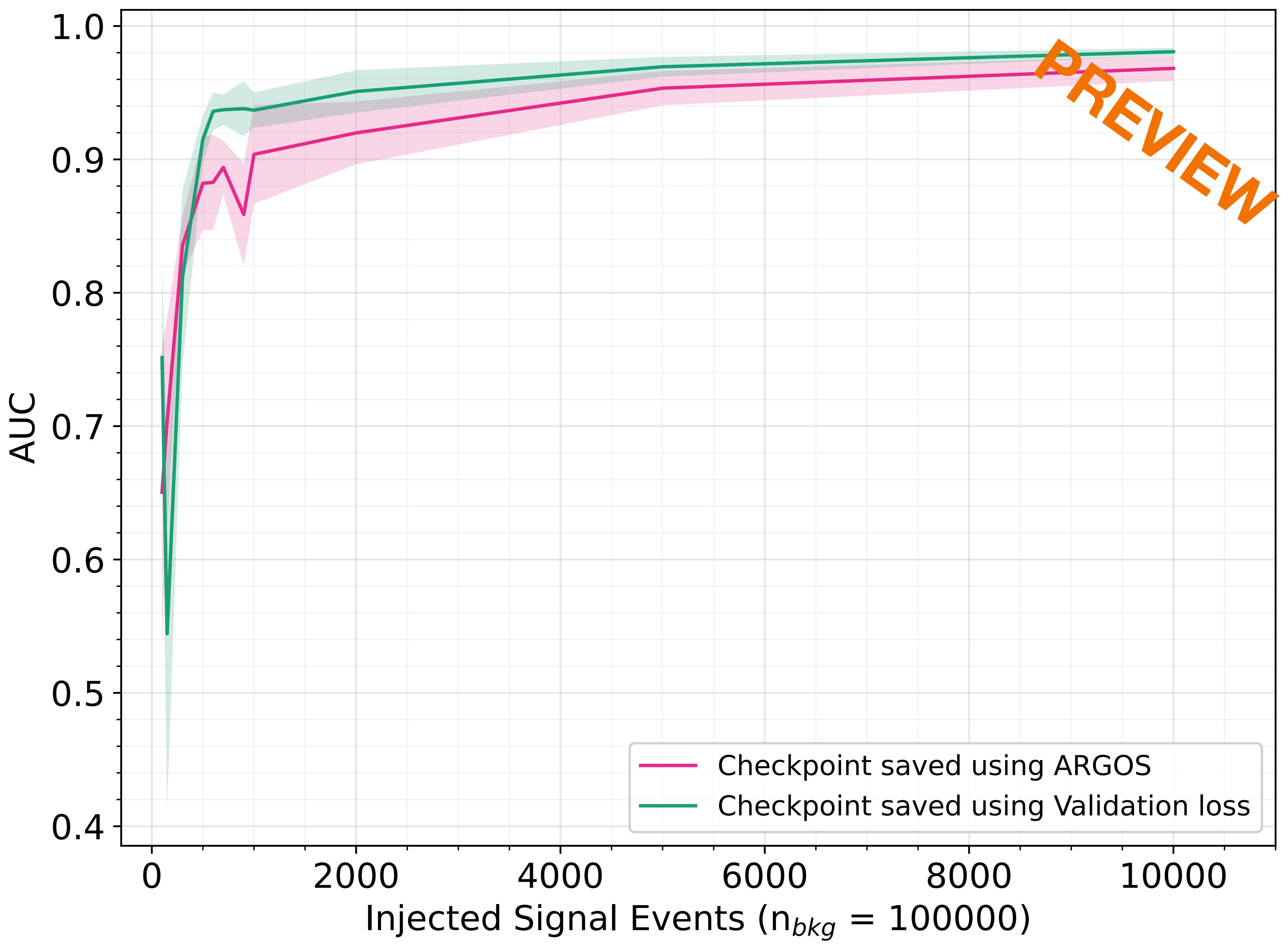


AUC Curve

ARGOS vs Validation loss

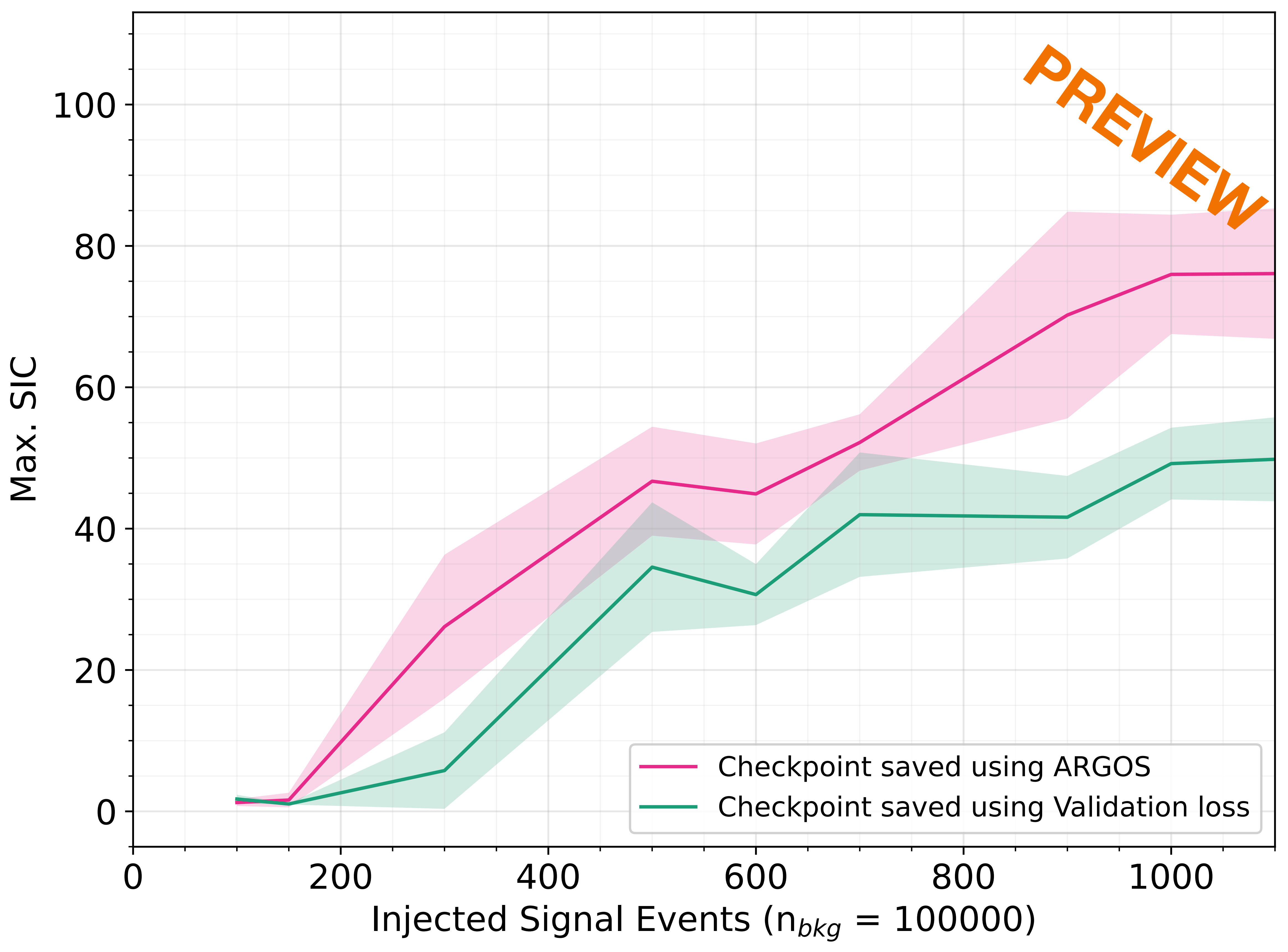


Max SIC Curve

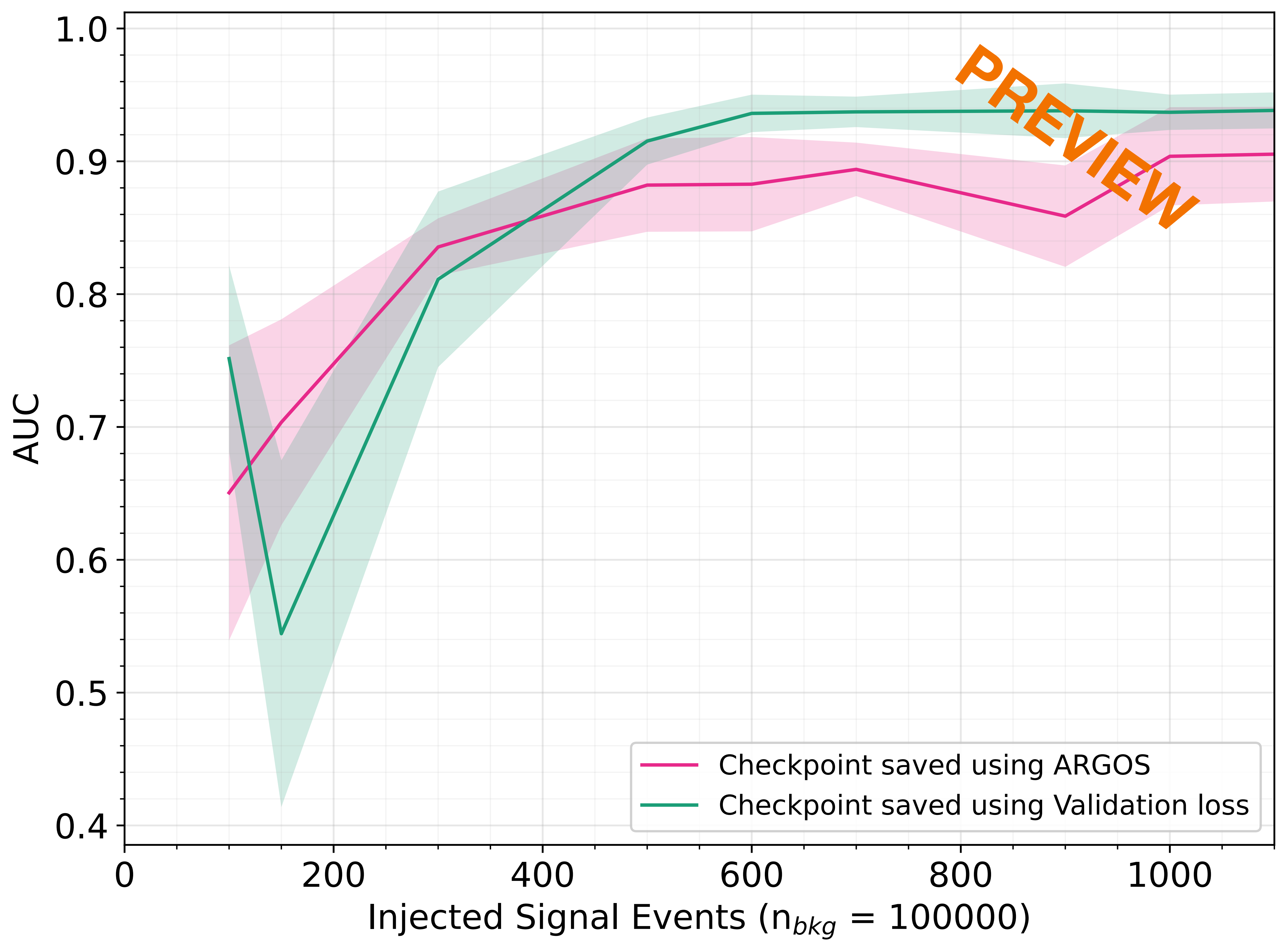


AUC Curve

ARGOS vs Validation loss: Zoomed in



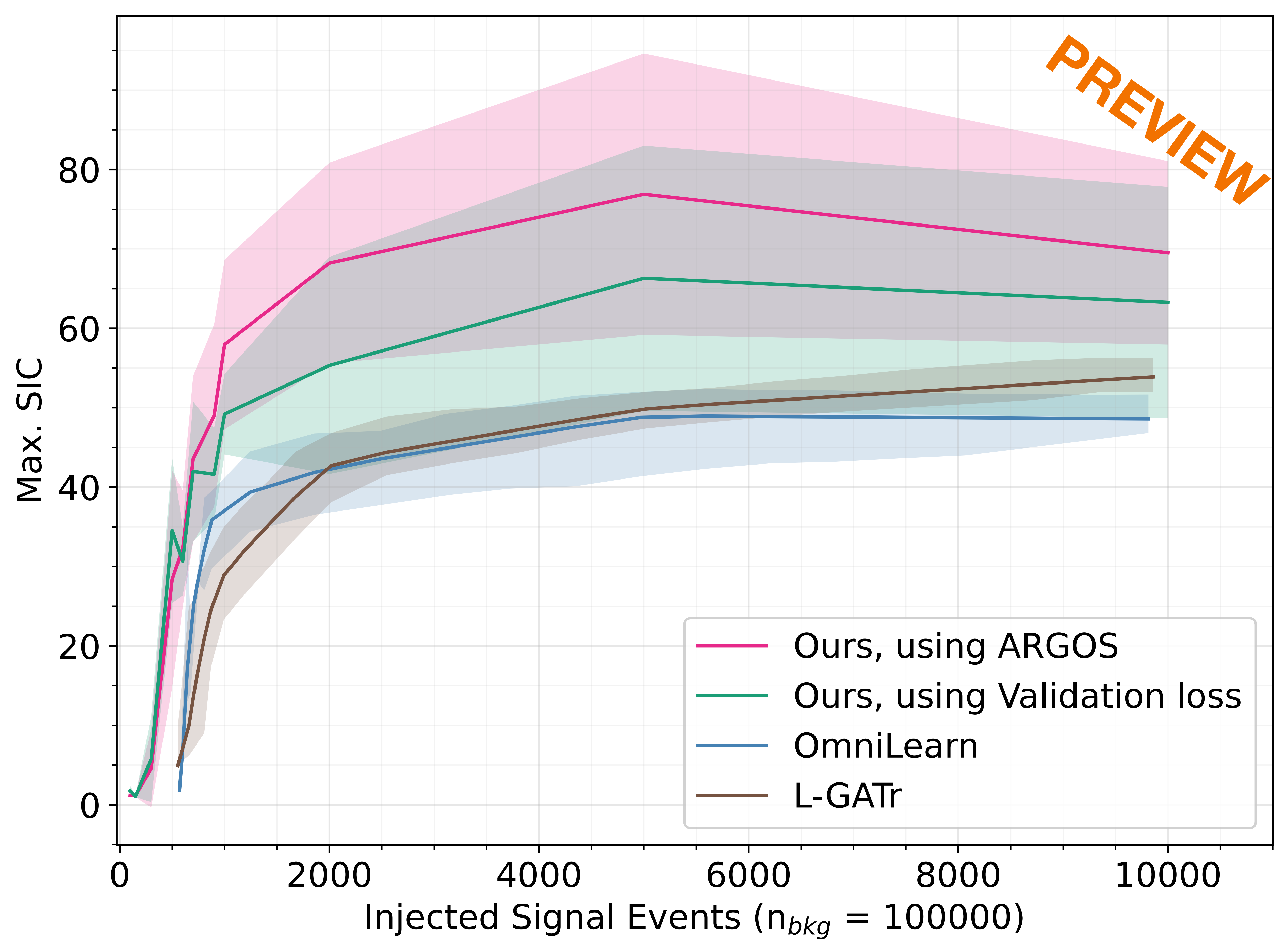
Max SIC Curve



AUC Curve

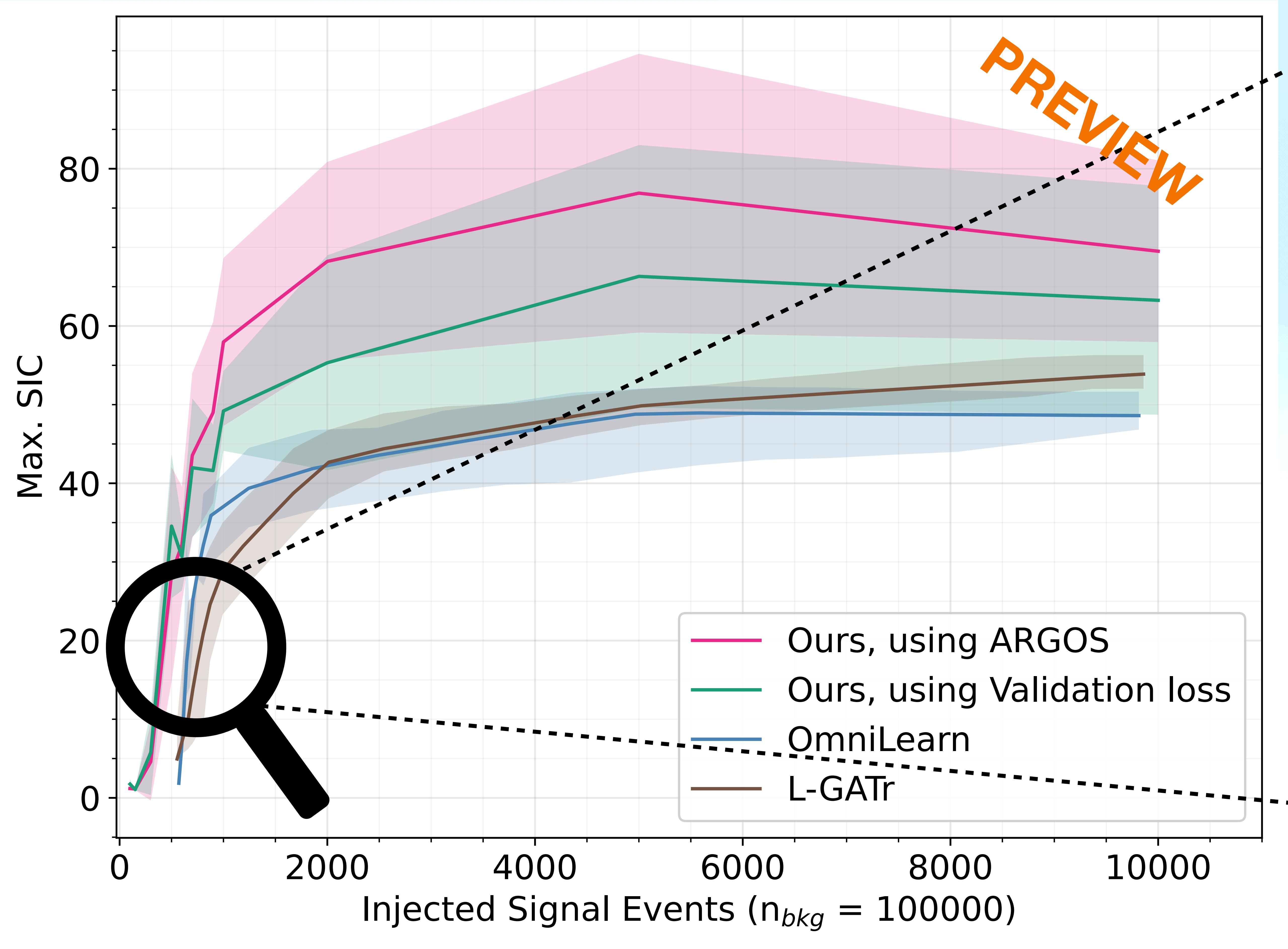
UNIVERSITÄT DES SAARLANDES RWTH AACHEN UNIVERSITY Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

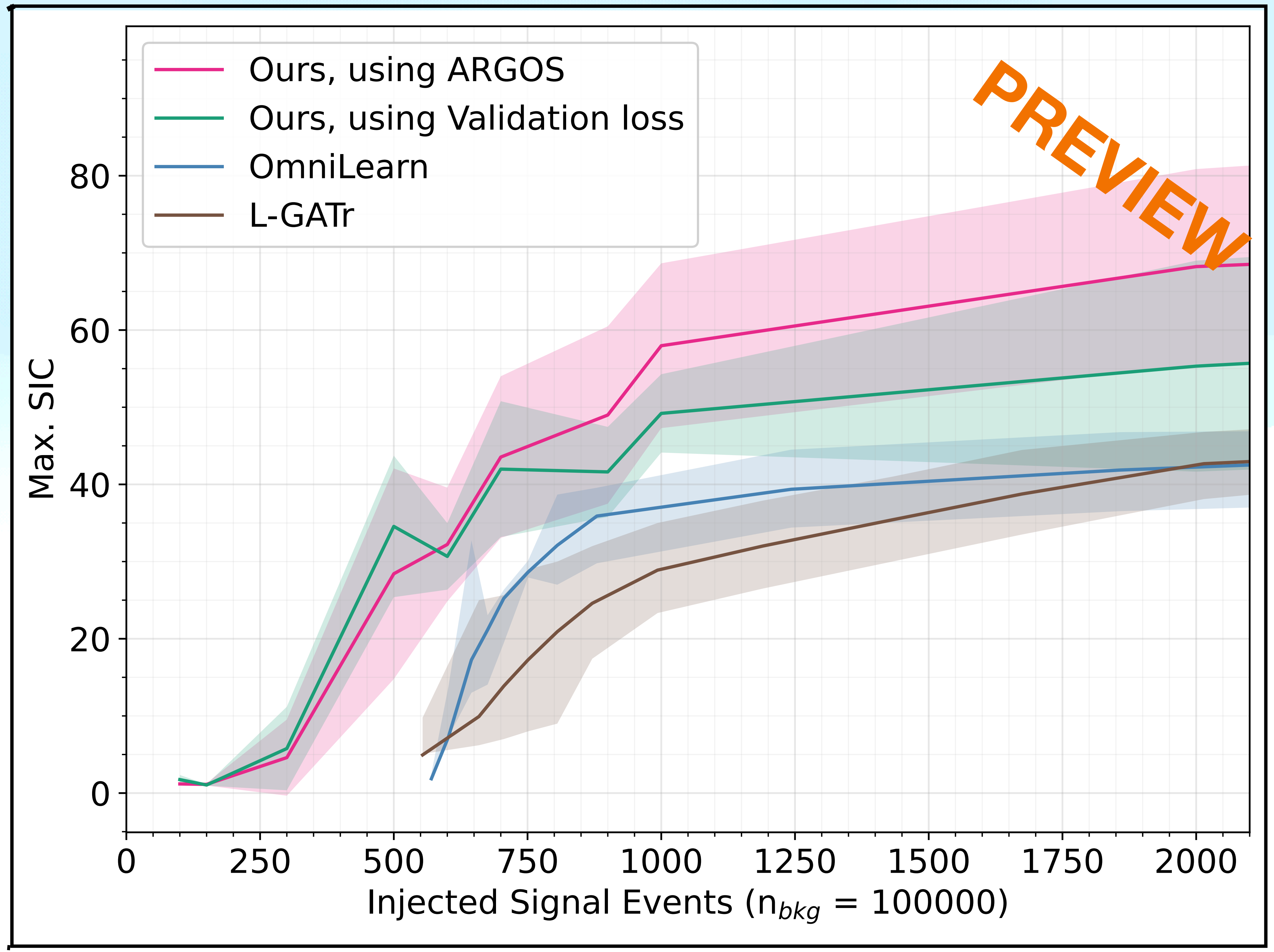
Comparing with State-of-the-art



Max SIC Curve

Comparing with State-of-the-art





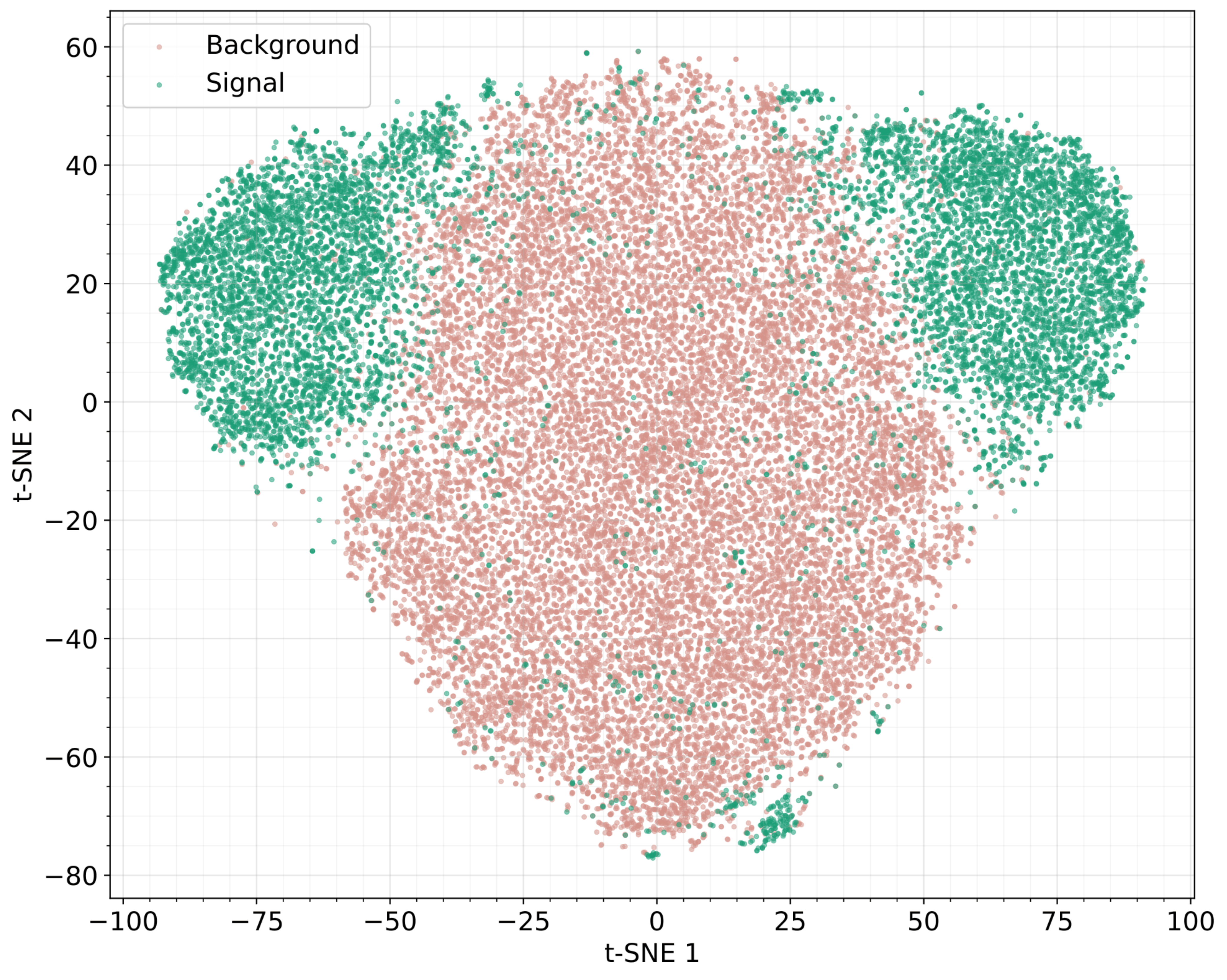
Max SIC Curve

Conclusions

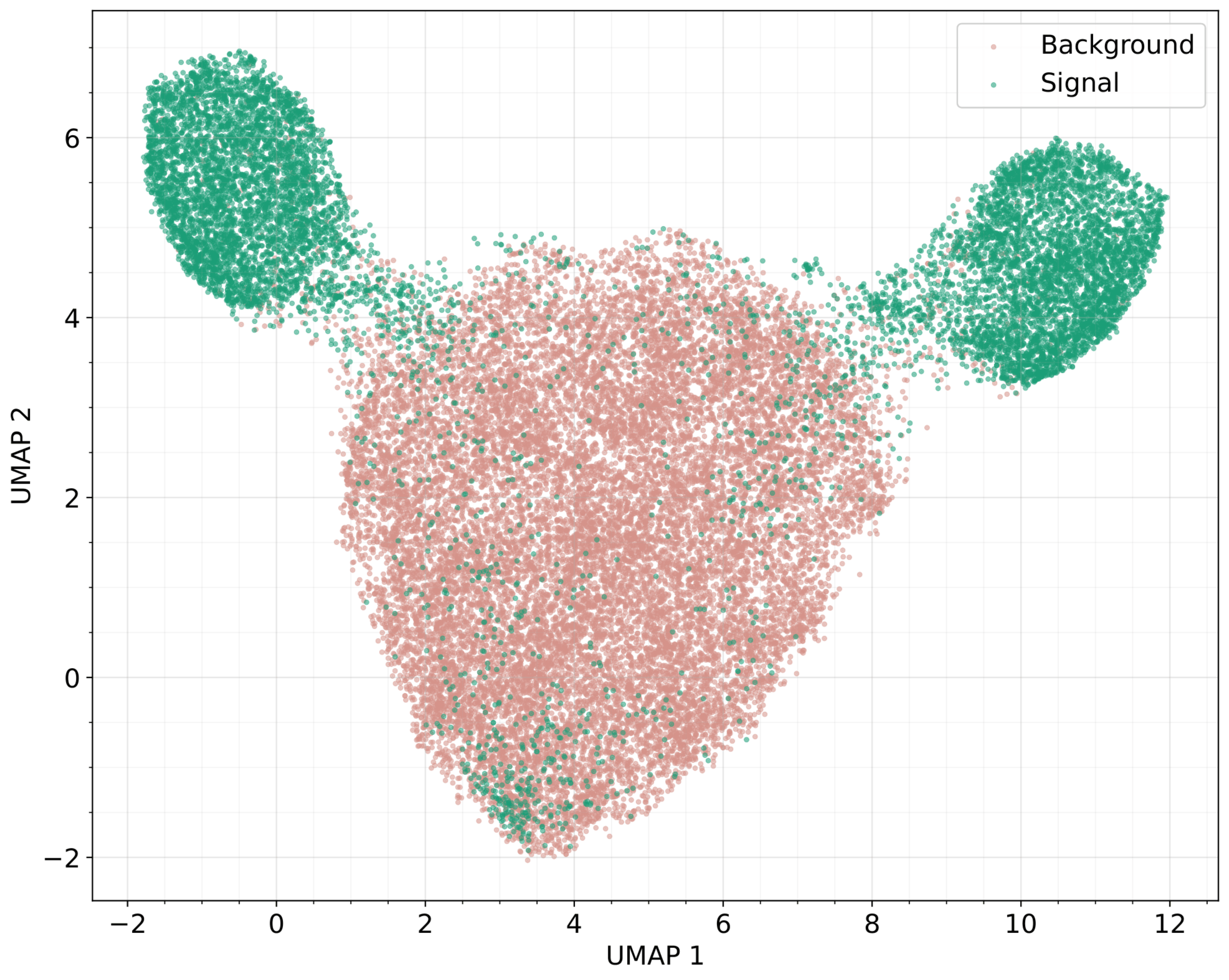
- **OmniJet-α** can support **anomaly detection** task
- **Pre-training improves** the **performance** on the anomaly detection task at hand
- **Supervised classification** pre-training remains the **most-effective for the weak supervised task.**
- **ARGOS** is preferable for saving the **better checkpoint**
- Generalisation to other datasets and to CWola ongoing.

Thank You

Supervised Pre-training: Embeddings



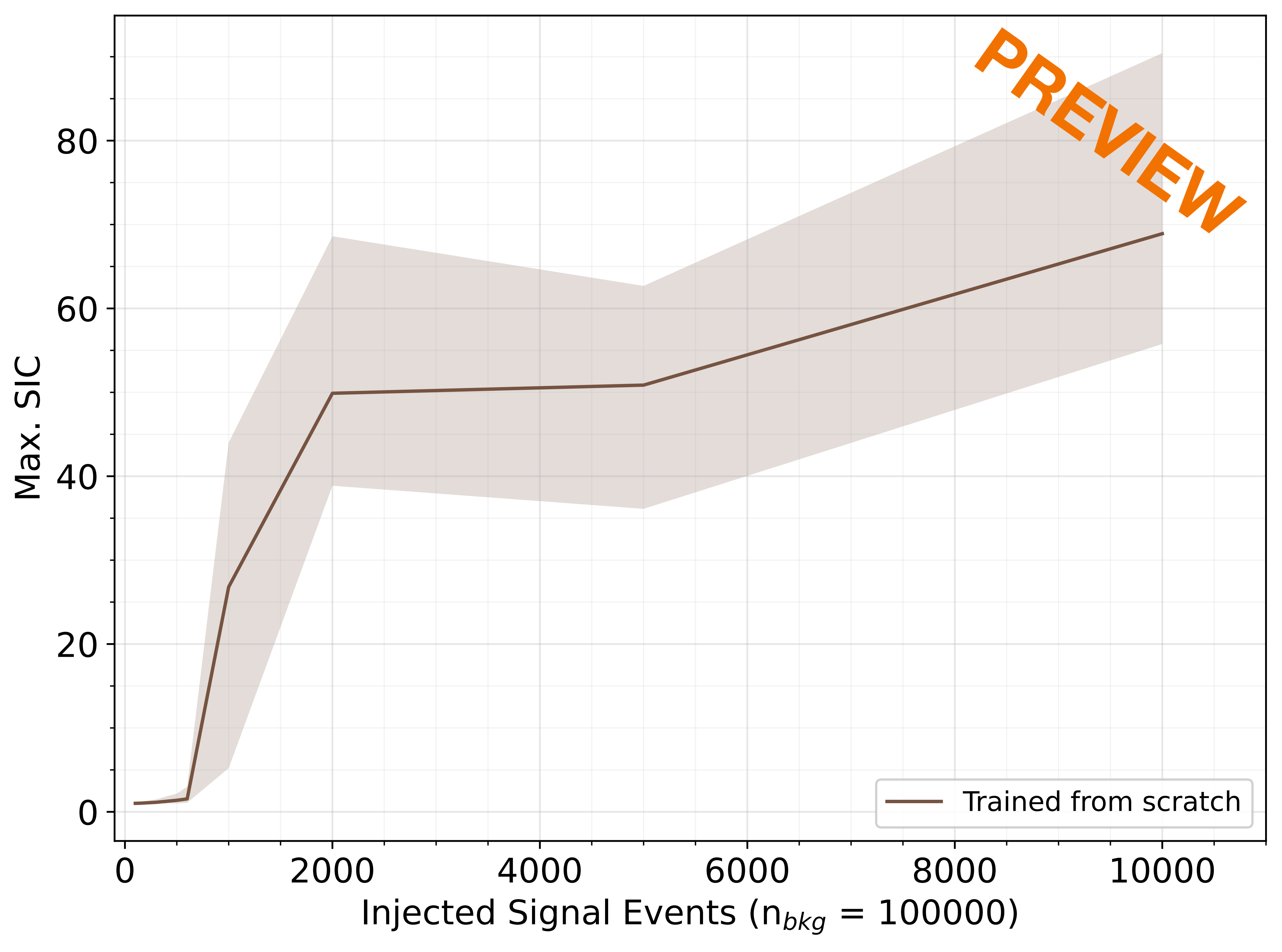
t-SNE visualisation



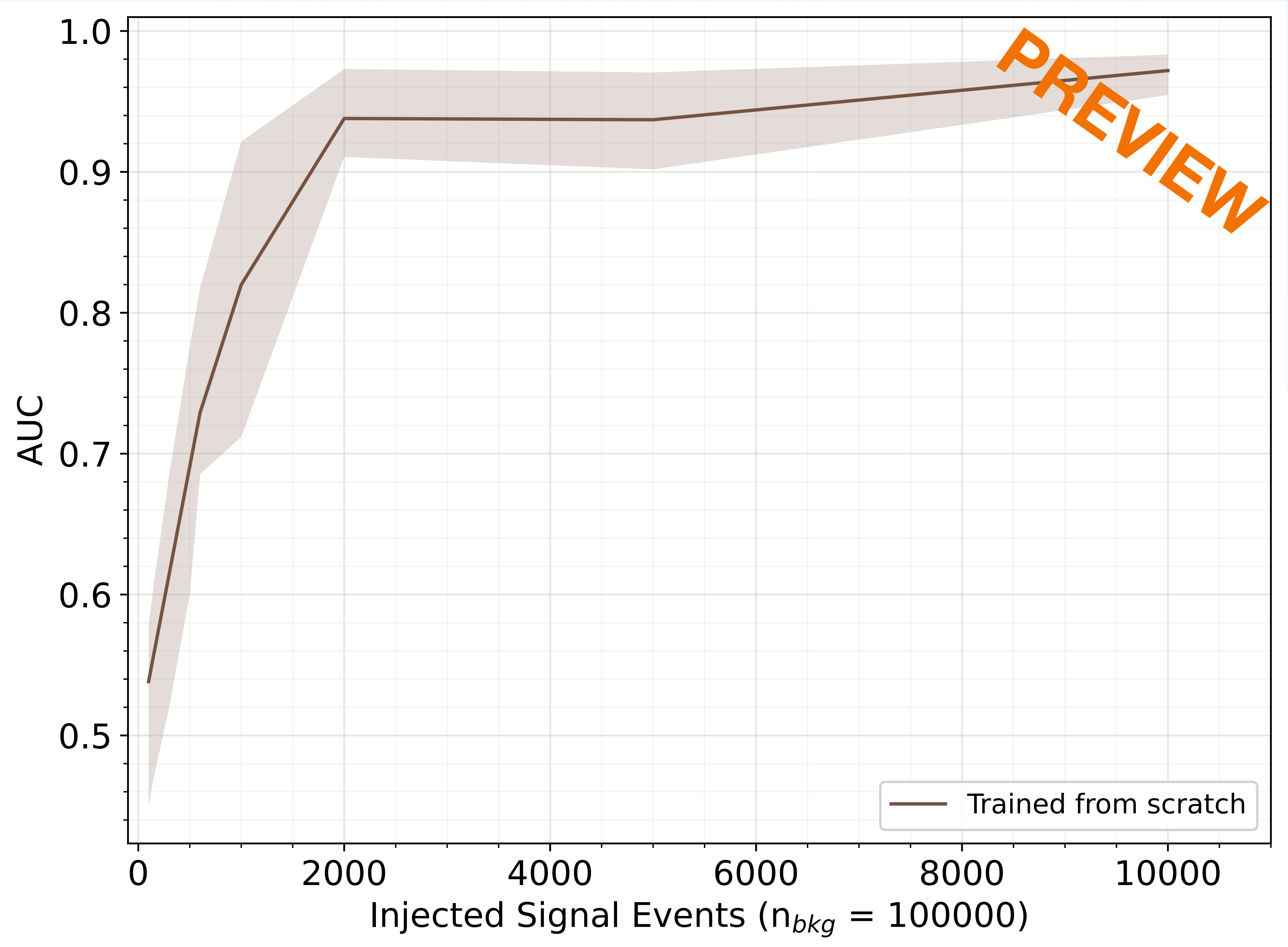
UMAP visualisation

Training from Scratch

Idealised Anomaly Detection (IAD)



Max SIC Curve



AUC Curve