Self-Supervised Learning for Multi-Channel Optical Waveforms in LArTPCs

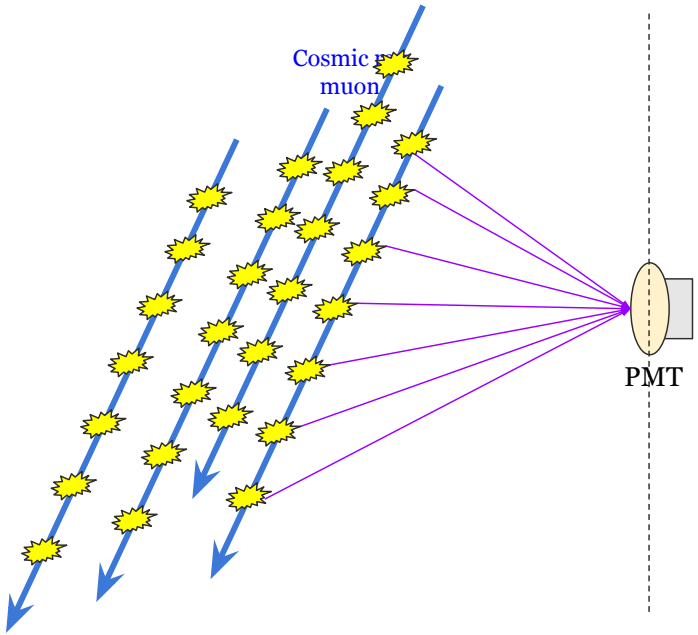
Self-Supervised Learning for **Multi-Channel Optical Waveforms** in LArTPCs

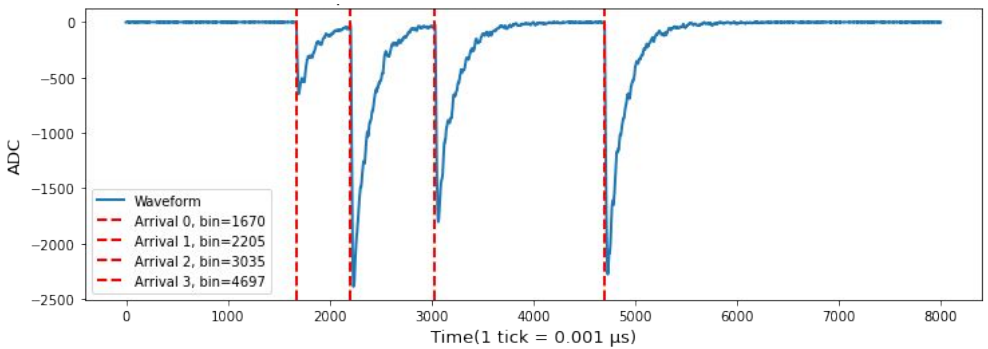
Stanford University

Stanford University

Multi-Channel Optical Waveforms

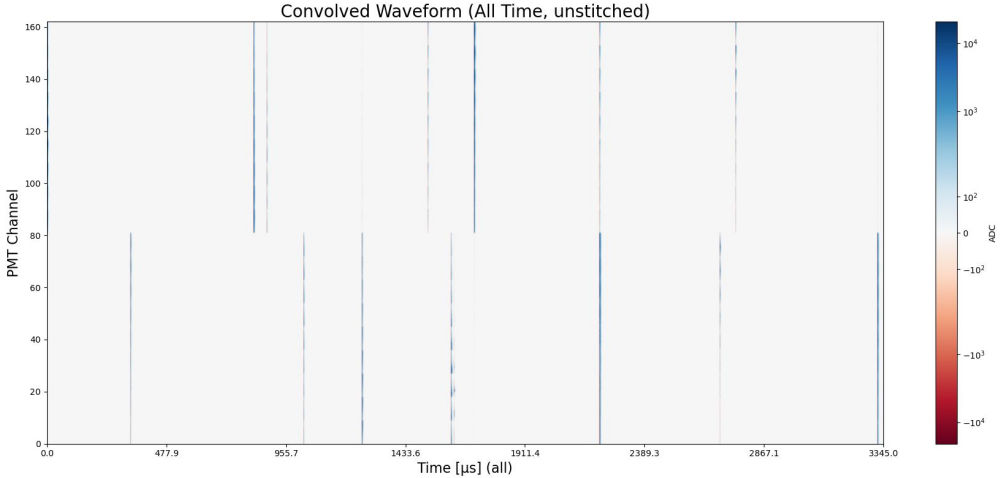
- O(ns) resolution on interaction timing





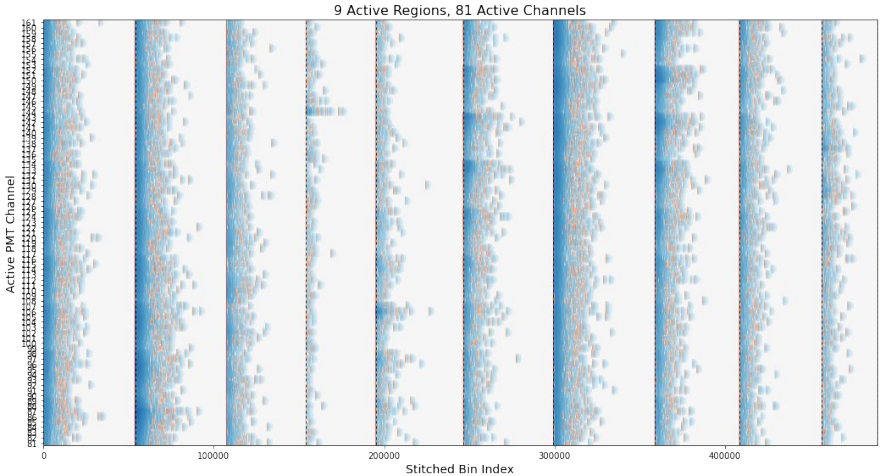
Challenge: Fully-Dense Data Format

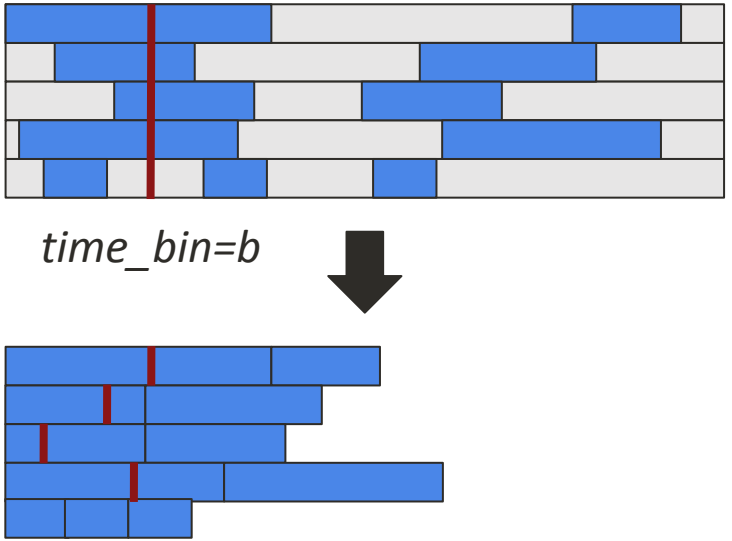
no



- A calm and normal image size is ~1080x1080 ⇒ 1.1M pixels^2
  - ~4 MB
- A raw optical waveform from GOOP is ~ 162x1000000 ⇒ 162M pixels^2
  - ~650 MB
- This is a bottleneck for any computation you want to do on these

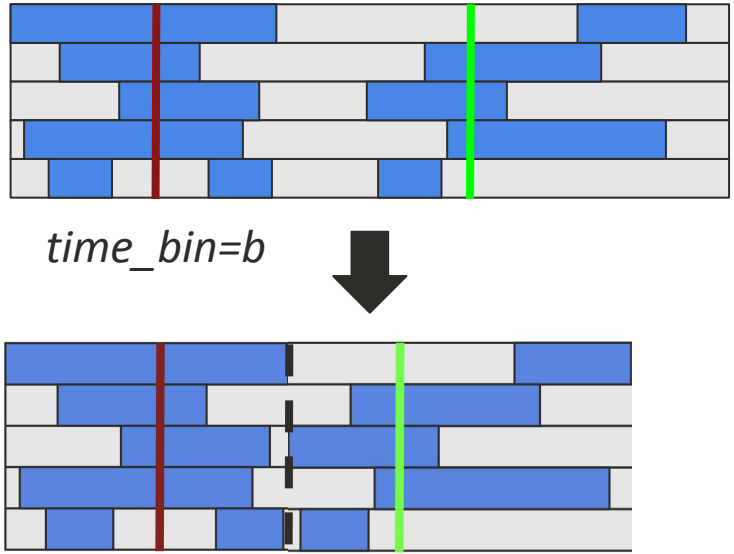
Fully Compressed Data Format





- Compressing along every PMT channel independently takes you from ~1M time bins ⟶ 40K time bins
- Local correlations in the data (if you look at all PMTs at the same nominal time bin) are not preserved

Semi-Compressed Data Format



- Cut out empty space while saving data in all channels for active regions
- Gets you from ~1000000 time bins ⇒ 60K time bins
- Can save each active region separately (chunks of ~10K time bins)

Self-Supervised Learning for Multi-Channel Optical Waveforms in LArTPCs

Stanford University

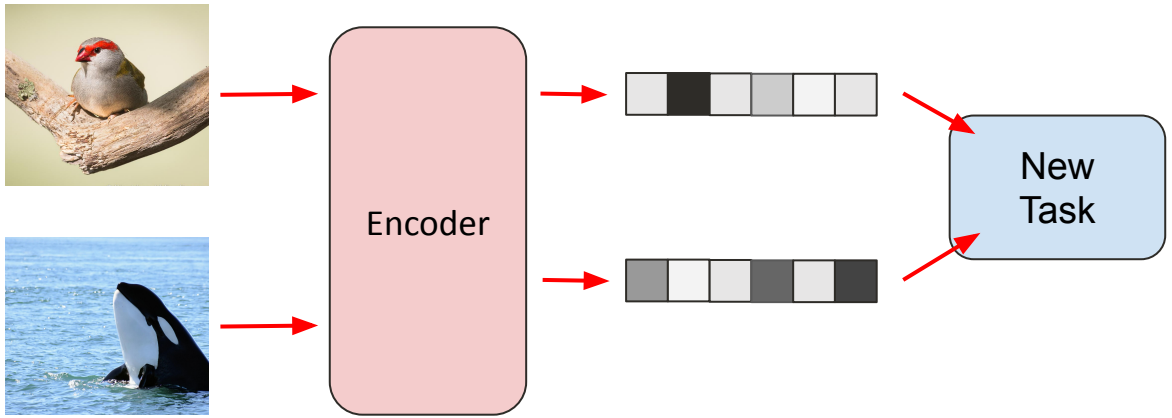
Stanford University

Problem: Domain Adaptation

- Some HEP models rely on simulation data for training, which do not include all unmodeled noise present in real data. They come from **different underlying distributions**.
- Could improve generalisation of models by
  - Directly training on real data w/out labels to learn high-level representations
  - Fine-tune on labelled dataset (simulation (or hand-labelled real) data with rich metadata)
- Goal: Create data representations mostly invariant to this domain shift

What is Self-Supervised Learning?

- ML technique where a model learns useful feature representations from raw data rather than relying on manually designed features.
- Goal: transform **raw data** into a meaningful and compact representation that makes downstream tasks (e.g., classification, clustering, or regression) easier
- Different methods:
  - Autoencoders
  - PCA
  - Masked-Based Reconstruction
  - Contrastive Learning



Self-Supervised Learning for Multi-Channel Optical Waveforms in LArTPCs

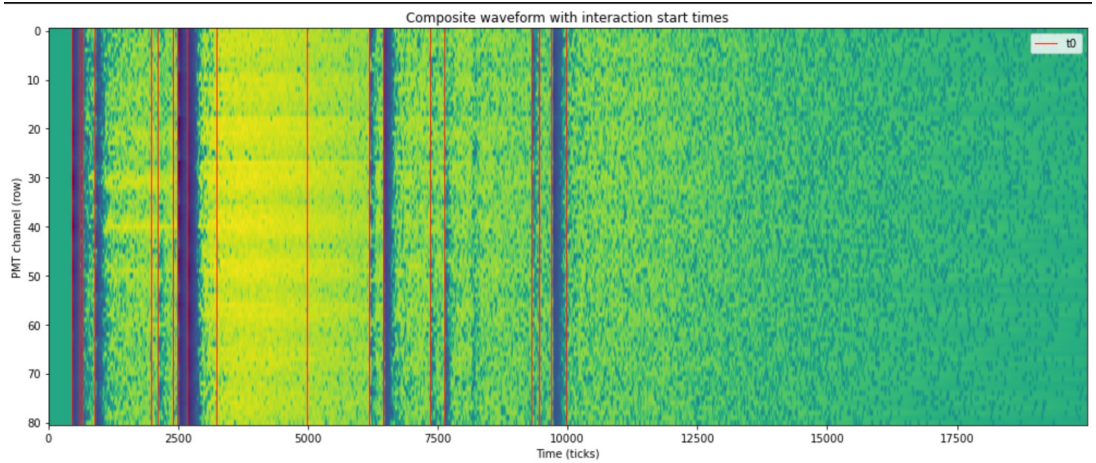
Stanford University

Stanford University

What distribution of optical data are we interested in?

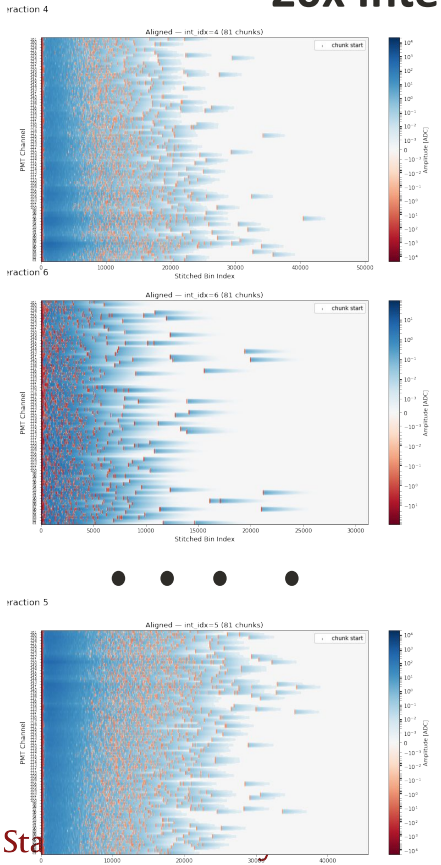
- See Sam & Junjie's project for more physics motivation
- But, generally motivated by ND pileup environment
- We roughly expect to see ~20 interactions in 10 microseconds

Goal: I want to do self-supervised learning on data with this rate of pileup



Generating a Data Sample

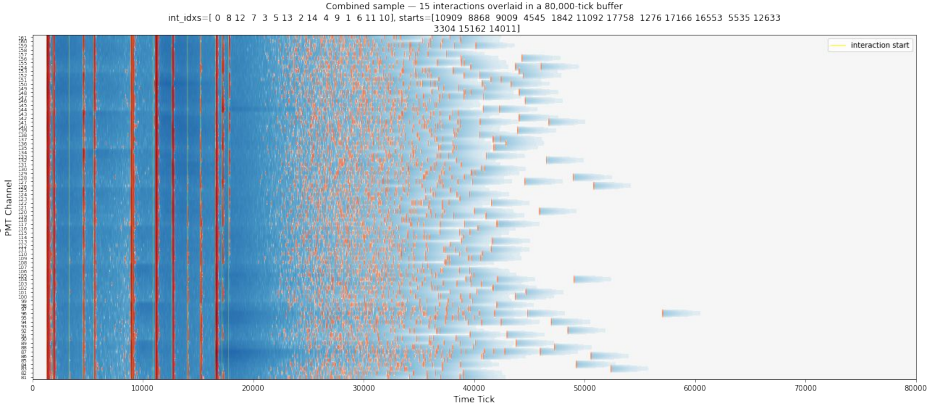
20x Interactions



Sample a *t0* in [0, 10 µs]

Sample a *t0*

Sample a *t0*

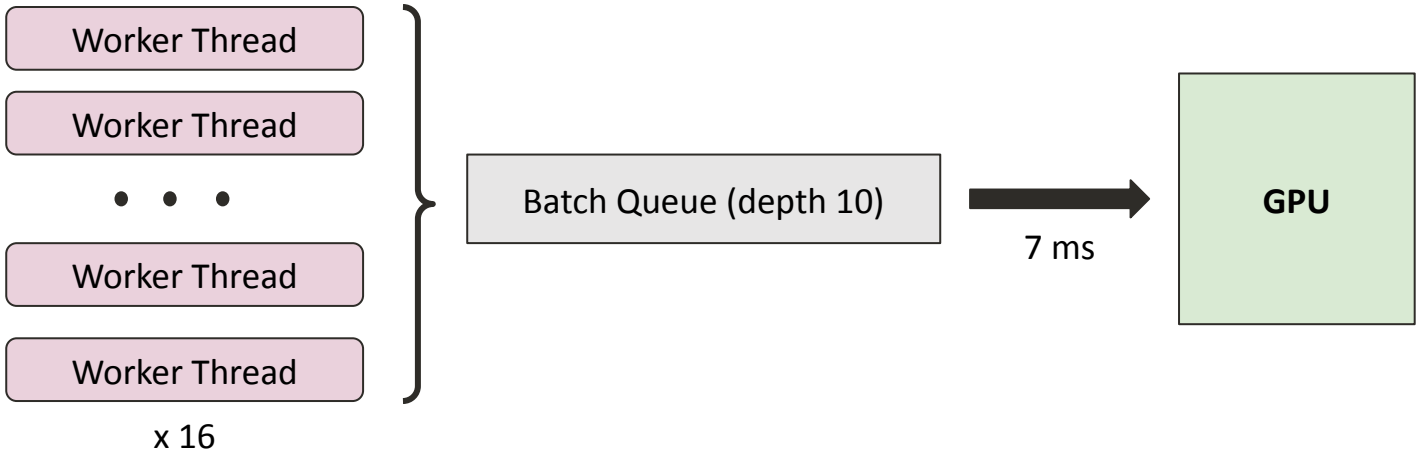


Accumulate ADC into 20 microsecond buffer

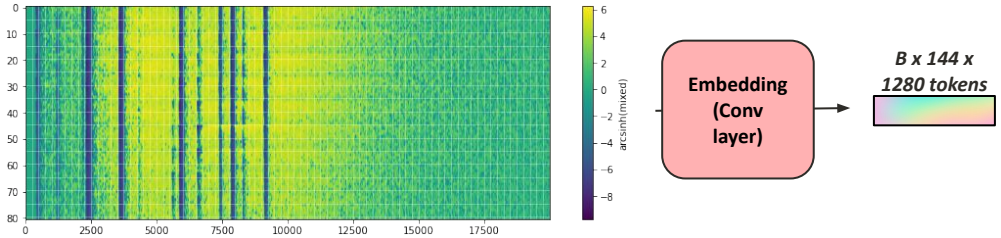
Sta

On-The-Fly Batch Generation

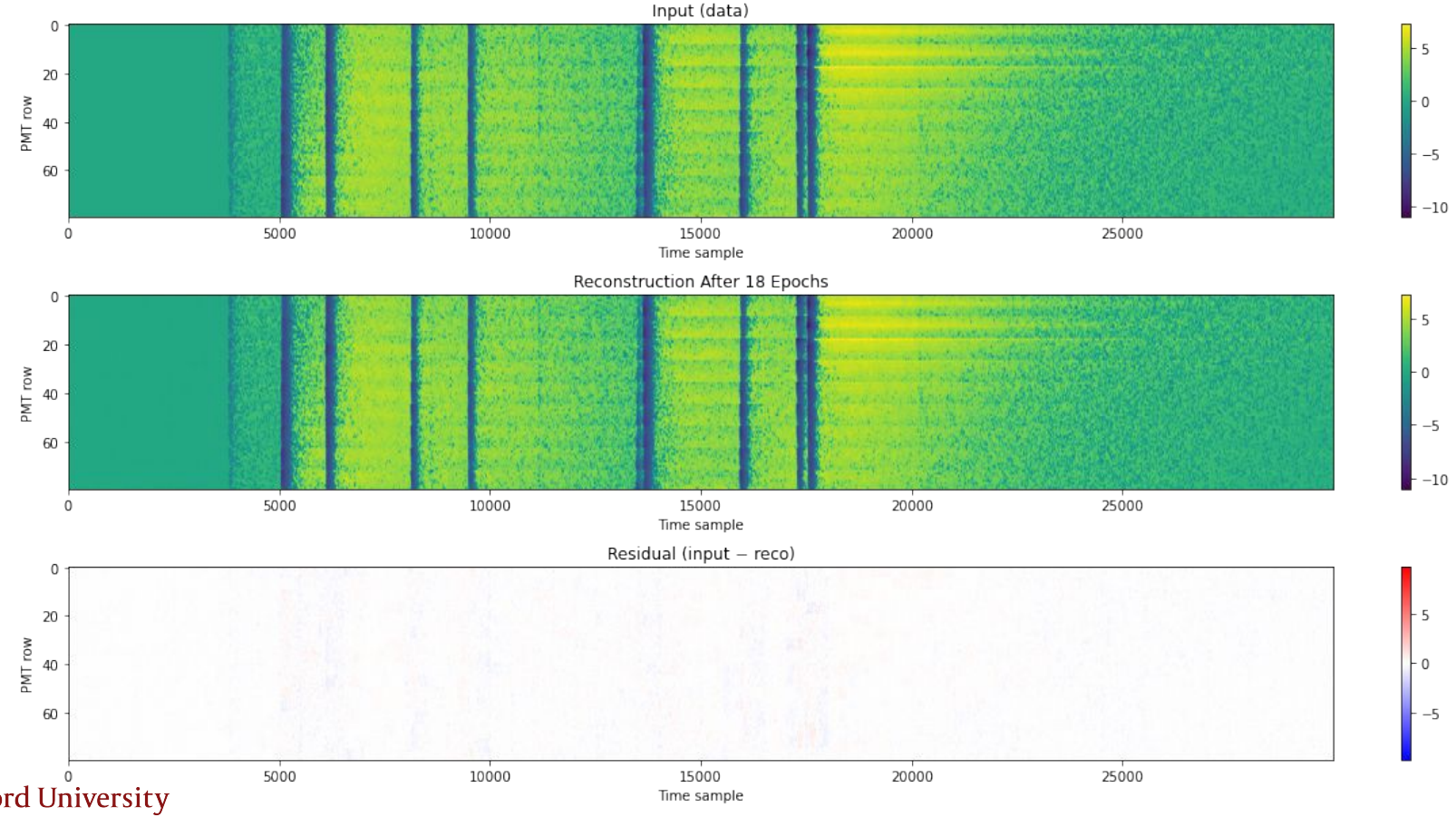
- Because I am making synthetic data in this way, I am not constrained by a dataset size.
- Can generate training data dynamically during model training
- **GPU downtime between batches is ~7 ms (transfer time)**
- **Forward + Backward is ~300 ms**



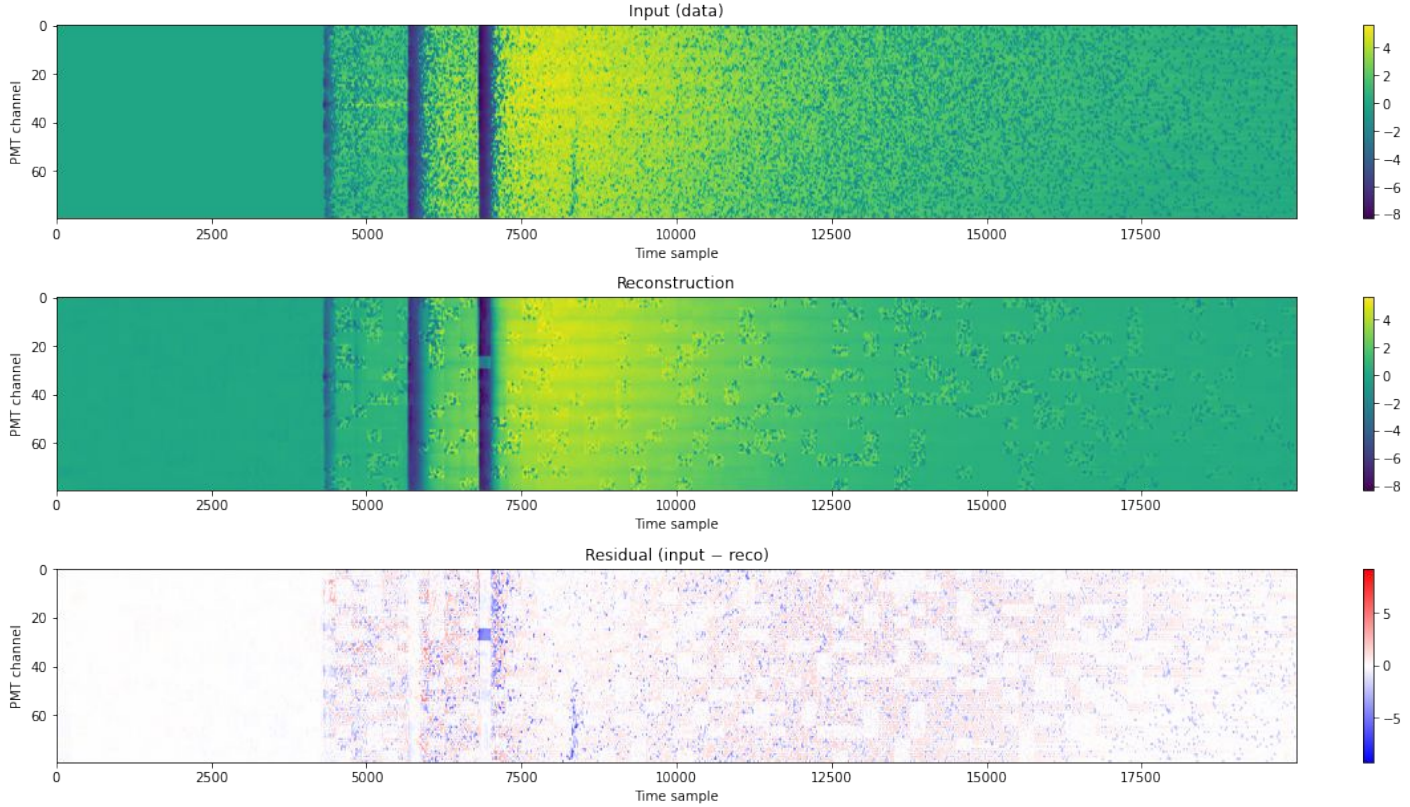
Masked Autoencoding



Working Autoencoder (Model Capacity Proof of Concept)

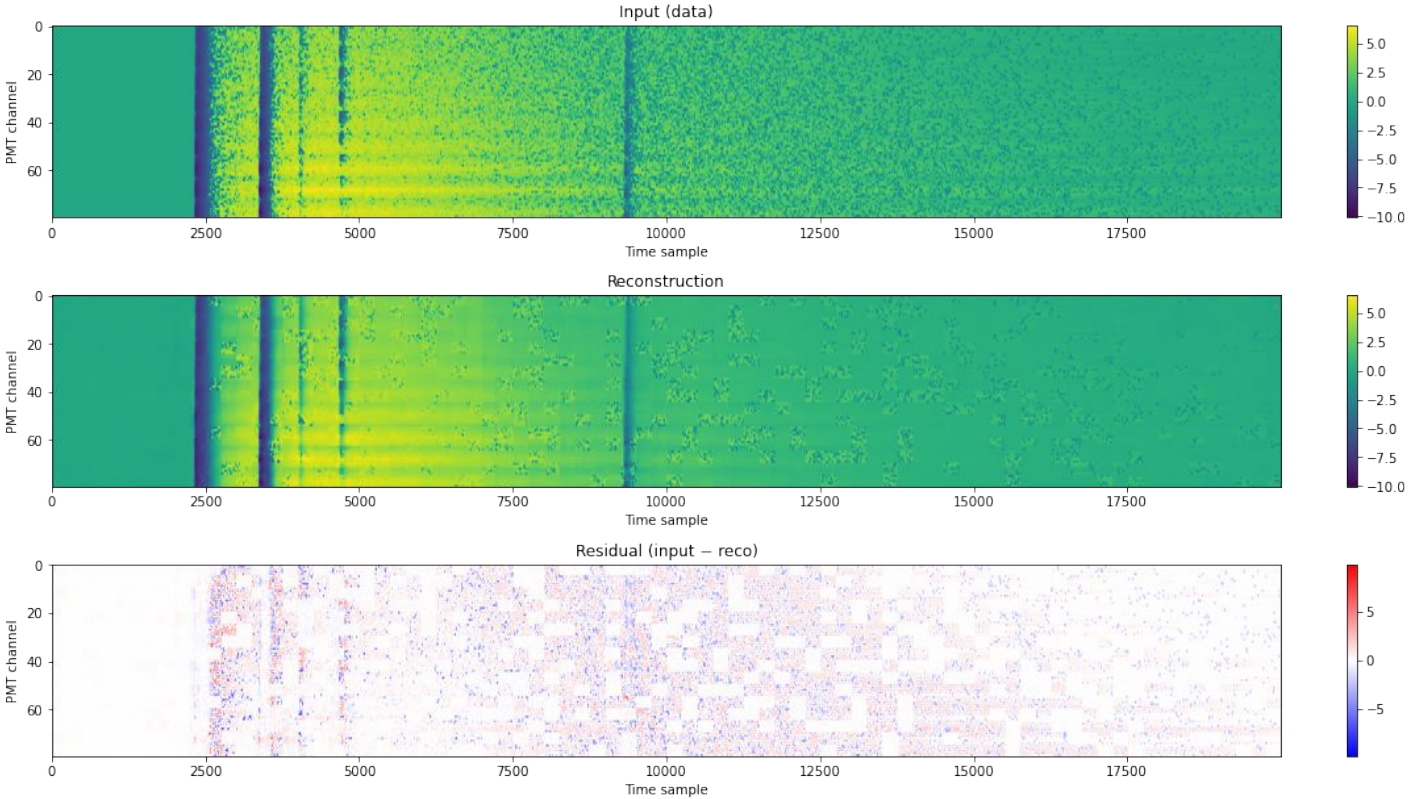


Masked Autoencoding



- Masking ratio 0.8

Masked Autoencoding



- Masking ratio 0.7