Autoencoding Good Bandpass Solutions: A Scalable, Self-Supervised Framework for Anomaly Detection in Bandpass Solutions

Omkar Bait (NRAO, Charlottesville)
CosmicAI Fellow

Brian Mason, Ryan Loomis, Ci (Ceci) Xue, Tristan Ashton, Ignacio Toledo, Eric Murphy, Jeff Philips and CosmicAI Observable Universe WG

2nd Annual Cosmic Horizons Conference, 15th July, 2026

ALMA Interferometer

- A **66-antenna** mm/sub-mm interferometer; every antenna pair (a baseline) is correlated to form the image.
- Calibration produces a **bandpass solution**: **amplitude & phase response vs. frequency**.
- Scale: up to **~128 - 2048 channels** per solution, for each antenna, polarization and spectral window.
- Bandpass shape reflects **complex receiver / electronics behaviour** — very hard to model from first principles.
- The **Wideband Sensitivity Upgrade (WSU)**:

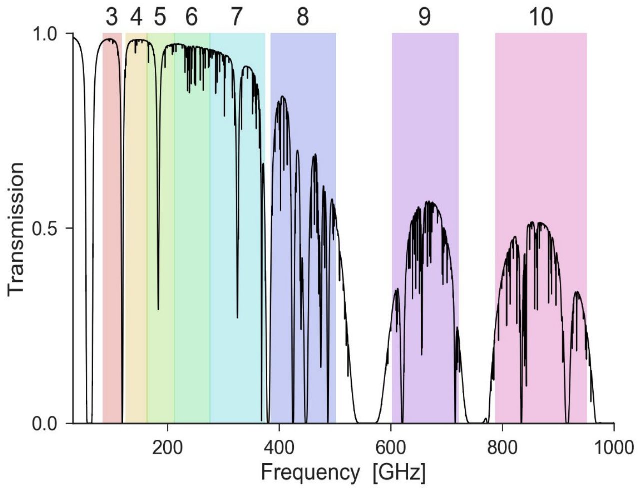
  ×2-4 bandwidth.

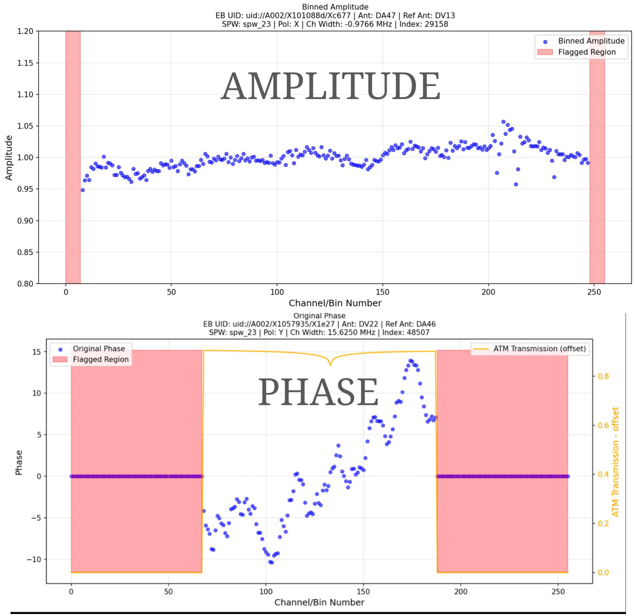
Carpenter+22

ALMA array

Credit: Juan Carlos Rojas - ALMA (ESO/NAOJ/NRAO)

ALMA Data is High-Dimensional

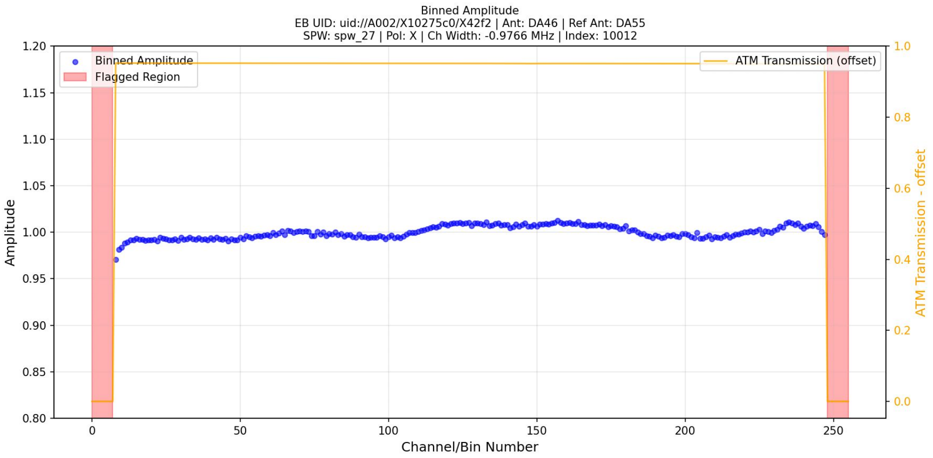


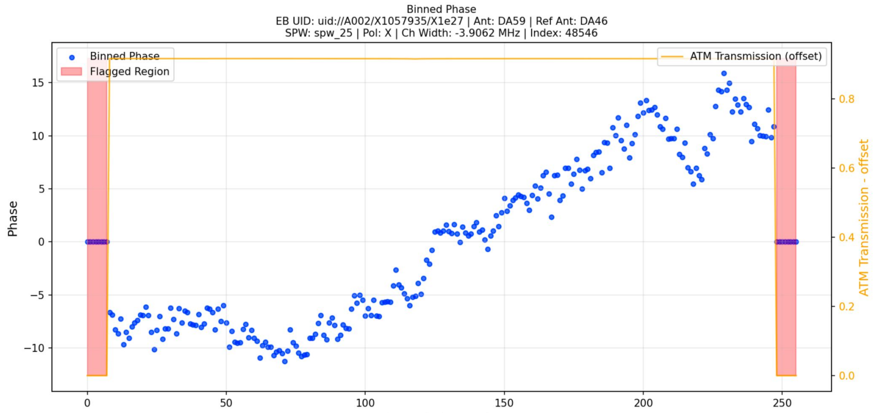


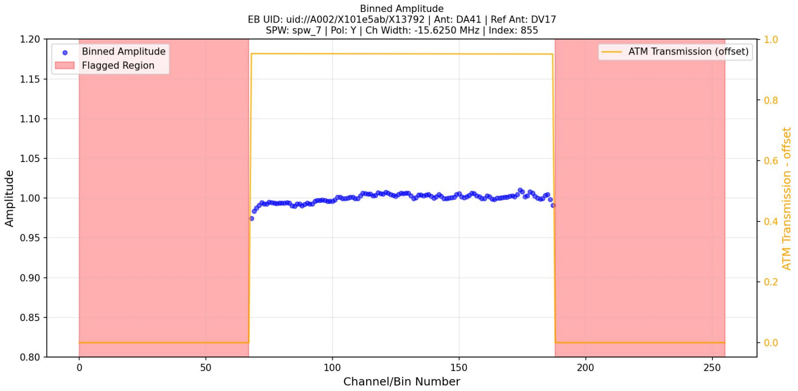
- Different observing modes:
  - Continuum mode **(TDM; 128 channels)** , Spectral mode **(FDM; 512, 1024, 1920, 2048 channels)** with different channel resolutions and polarizations.
- Observations cover broad science topics: exoplanets, star-forming regions, black holes, galaxies, cosmology etc.

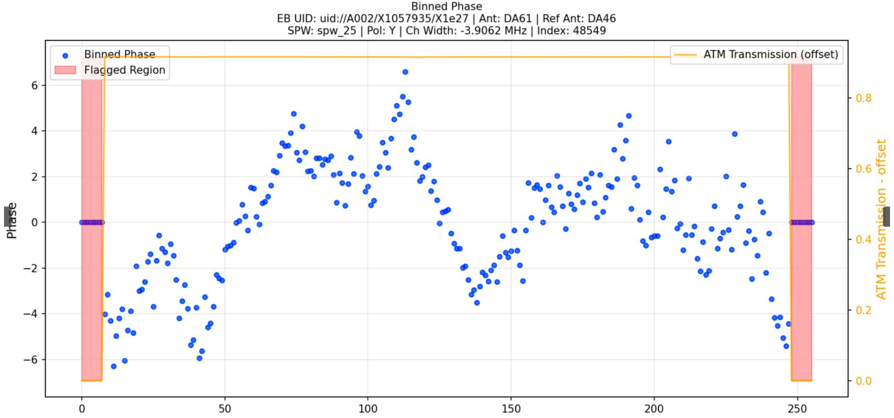
AMPLITUDE

PHASE









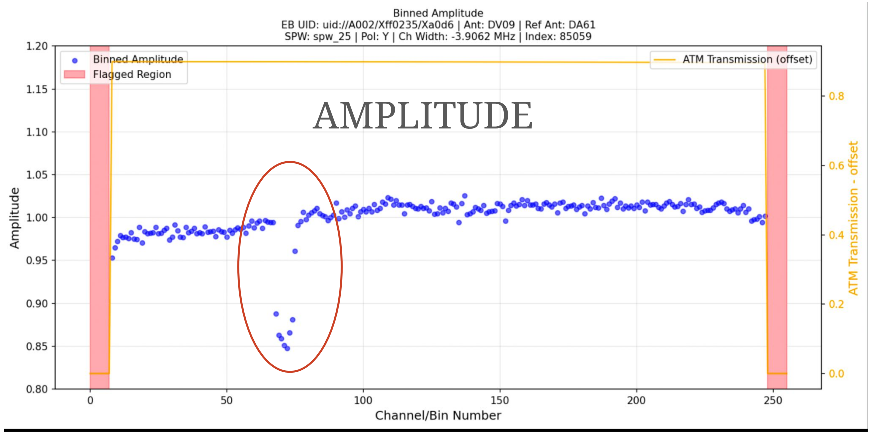
Anomalies in Bandpass Solutions

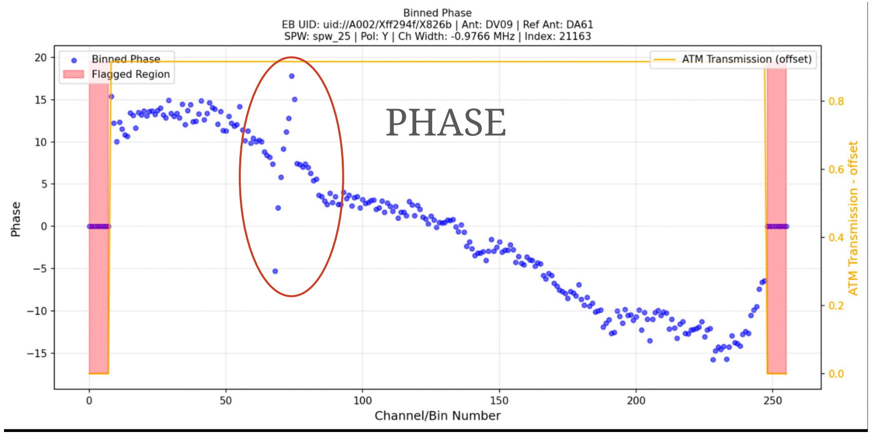
Goal: **automatically flag potentially problematic** calibration solutions with minimal human intervention.

The ALMA archive has **large sample of good, pipeline-processed, expert-vetted** calibrations — high-confidence “normal” data.

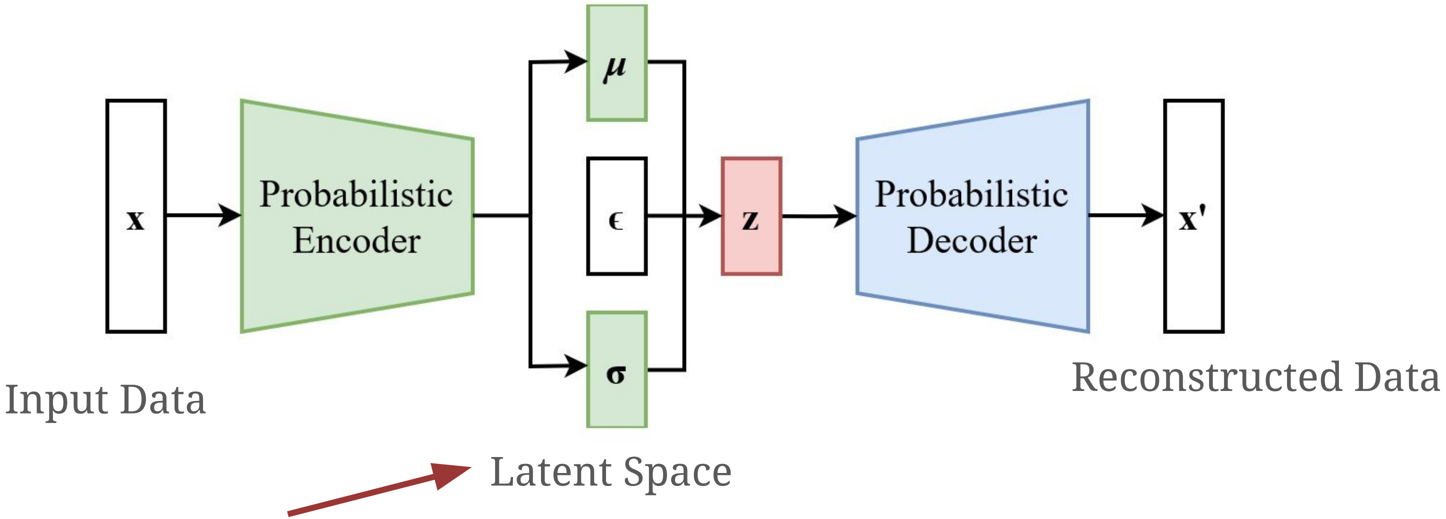
This good data can be used to train an AI model using self-supervised techniques.

Detect anomalies as **departures from learned “normal”**: high reconstruction error, outliers in the latent space.





VAE for Anomaly Detection

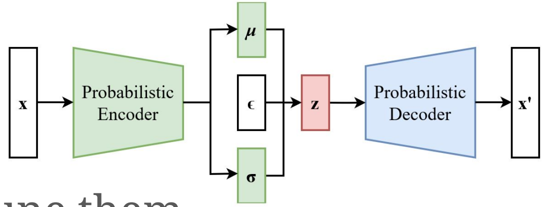


Train the autoencoder on **good** data → **self-supervised learning.**

Latent space → low-dimensional representation of data (good solutions).

Anomalous data → bad reconstructions, outliers in the latent space.

Why Autoencoder?



Self-supervised learning is generalizable:

- Single model across observing bands, modes etc.
- Do not need to define **manual heuristics** or fine-tune them.
- Completely data-driven **representation** of good bandpass solutions.
- **Good data** is easy to produce → Better/general representation of the data.

Multi-modal:

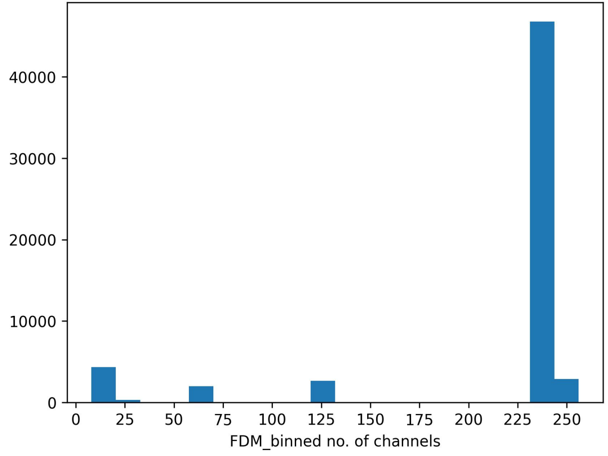
- Conditioning the latent space using **atmospheric lines, phases**.

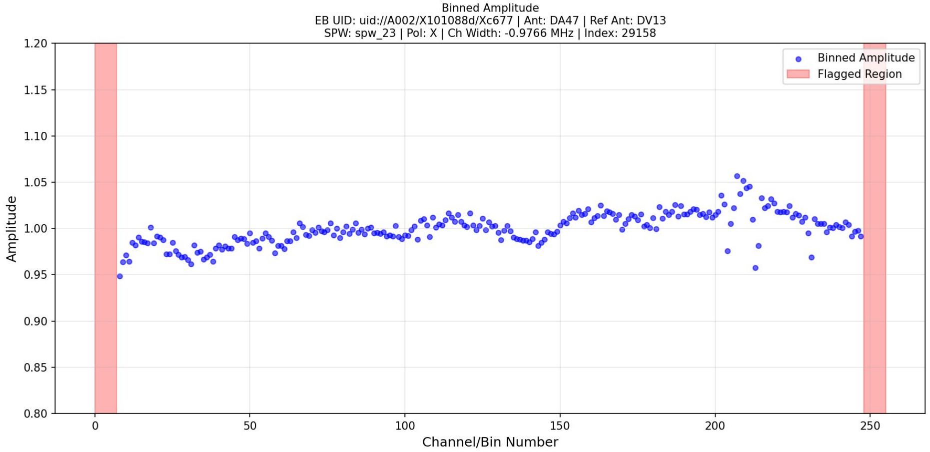
Scalable

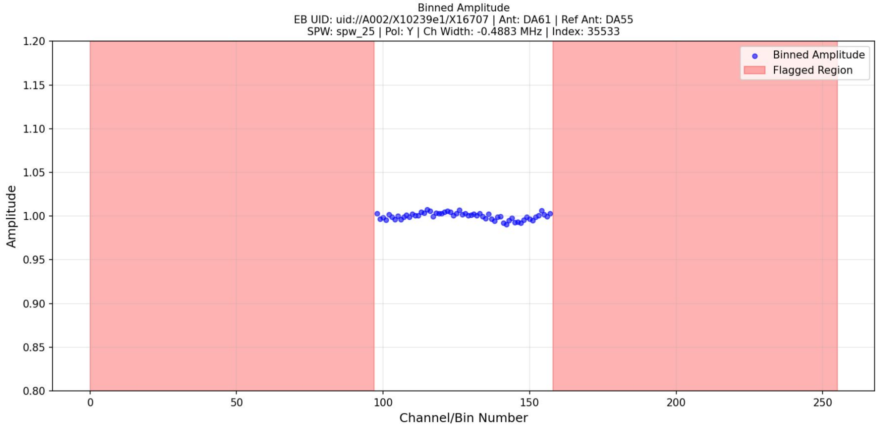
- Once trained inference is extremely efficient.

Pre-processing the Bandpass Solutions

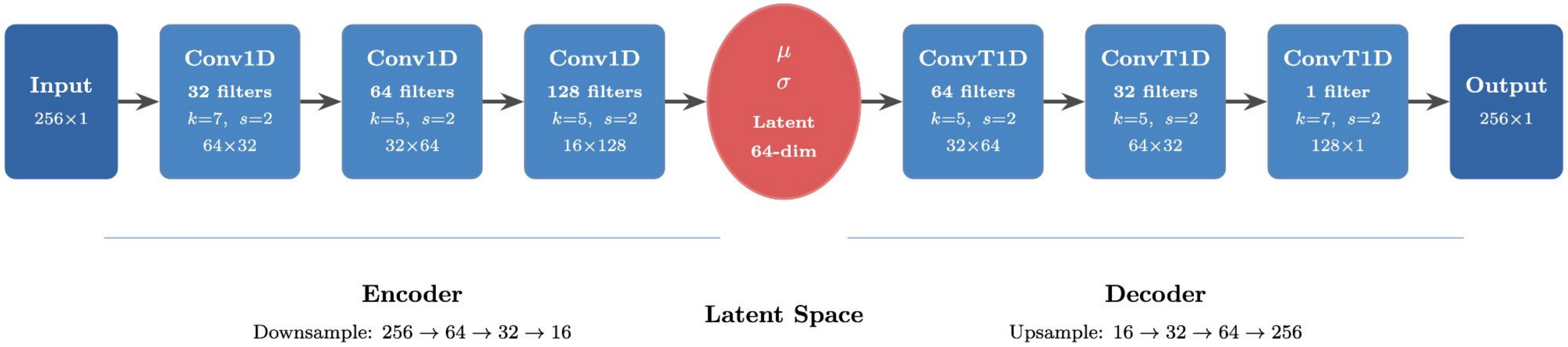
1. Cycle 9 Data with anomaly labels.
2. Bin native channels to fixed 7.8125 MHz
3. Mask bins with flagged channels
4. If no. of bins < 256, add masked channels upto 256.
5. **83093** independent solutions.







VAE: Architecture & Training



$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \beta \cdot \mathcal{L}_{KL}$$

$$\mathcal{L}_{recon} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_i - \hat{x}_i)^2}{\sigma^2}$$

Training done on Vista TACC

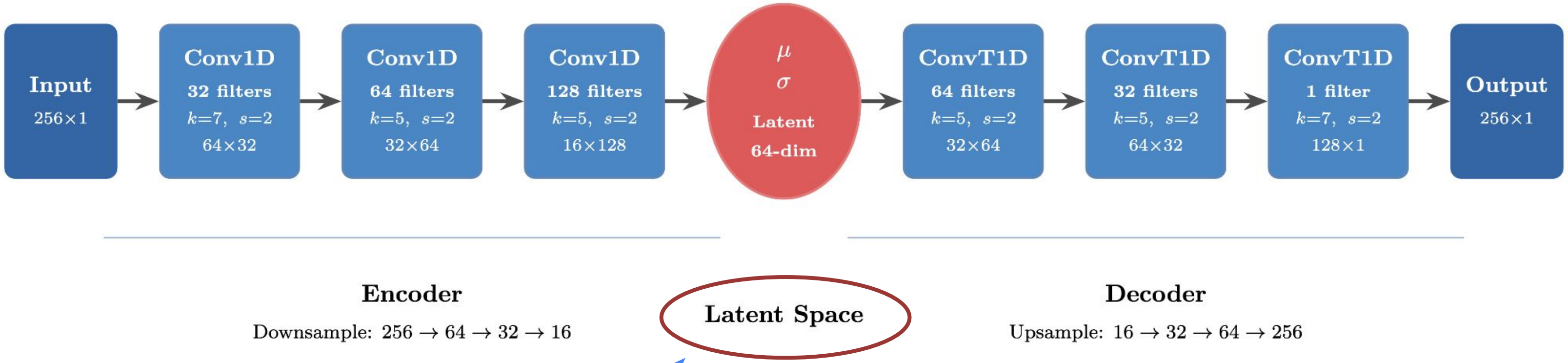
Single NVIDIA H200 GPU node

Takes ~**20 mins to train**

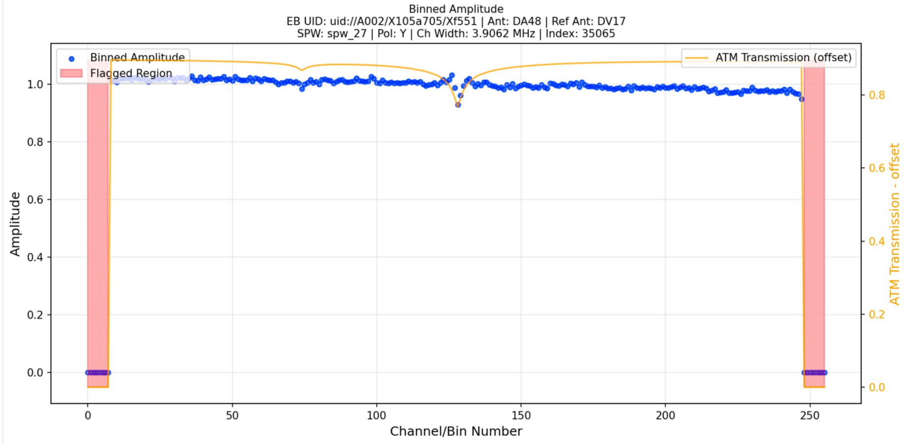
Masked channels are ignored in the loss function

Bait+26, in prep

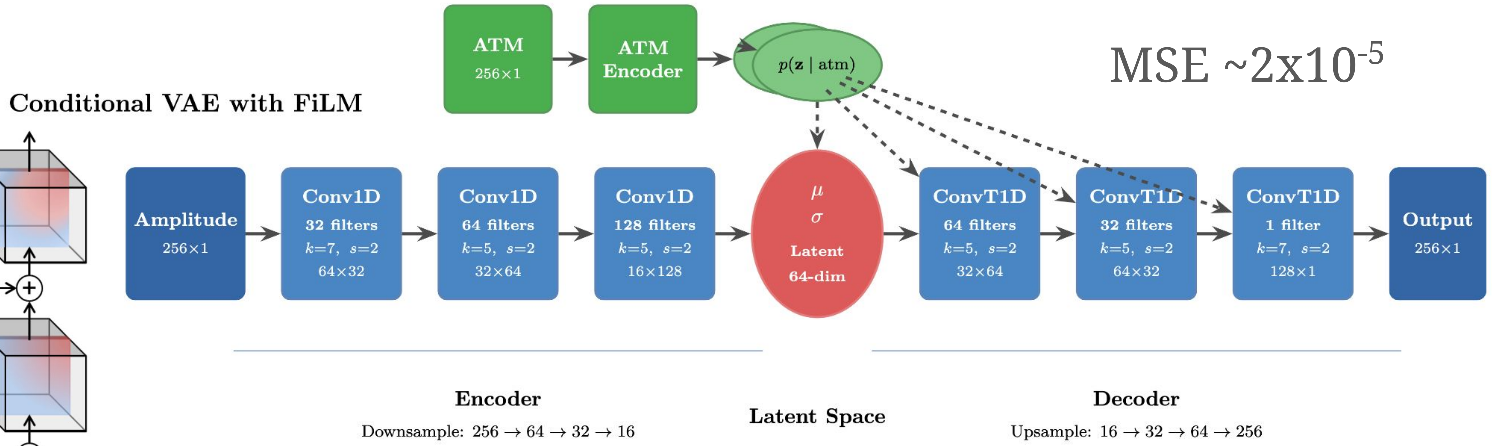
How to incorporate the atmospheric information?

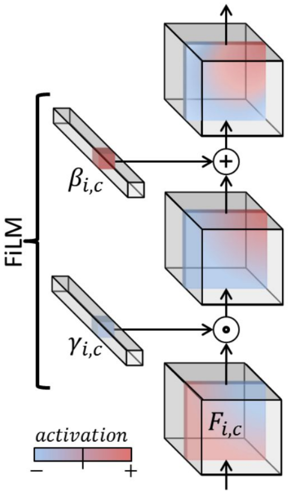


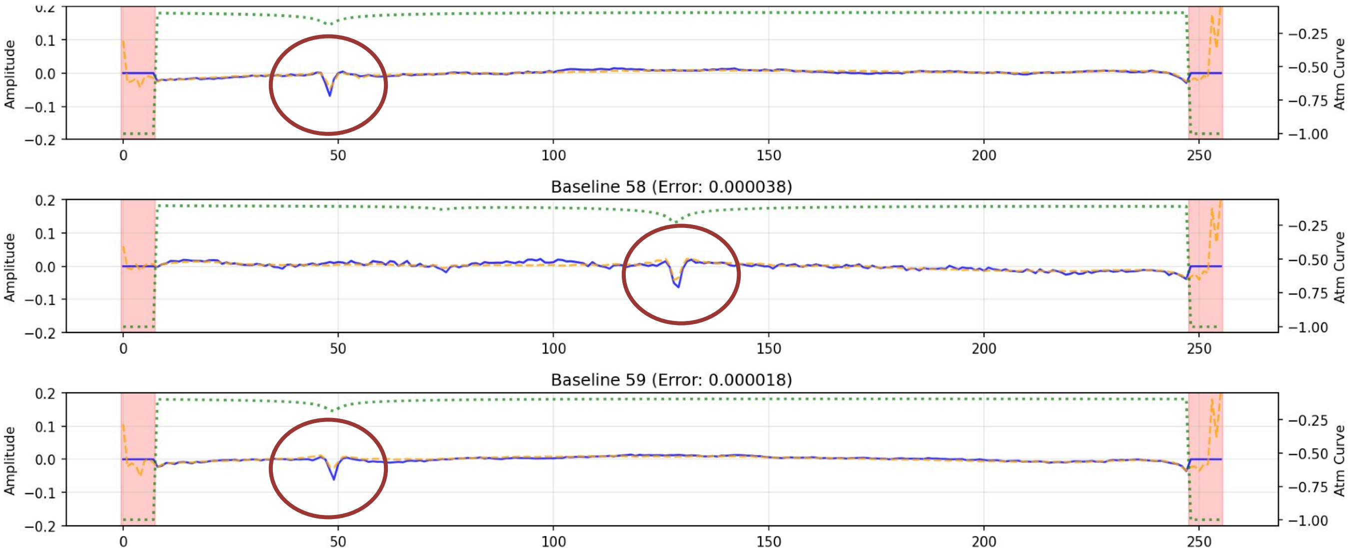
Condition the Latent space using the ATM curves



MSE ~2x10^{-5}

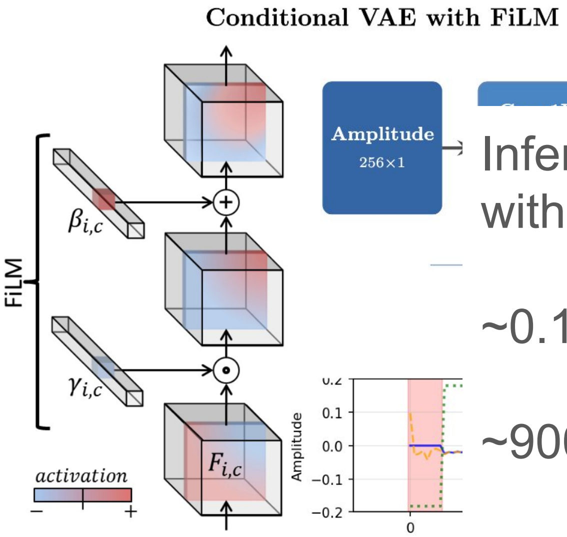


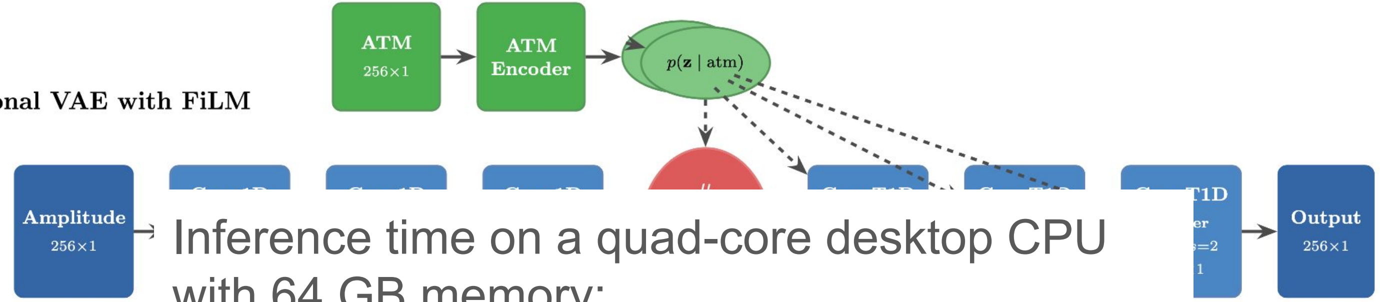




Perez+17,
arxiv:1709.07871

Bait+26, in prep

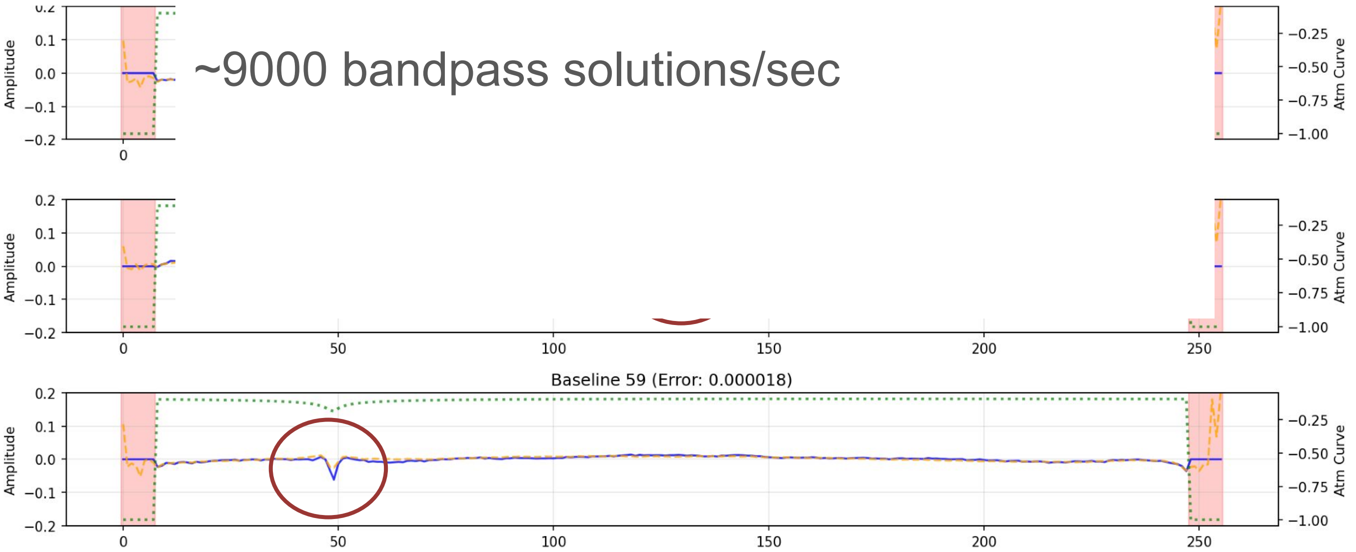




Inference time on a quad-core desktop CPU with 64 GB memory:

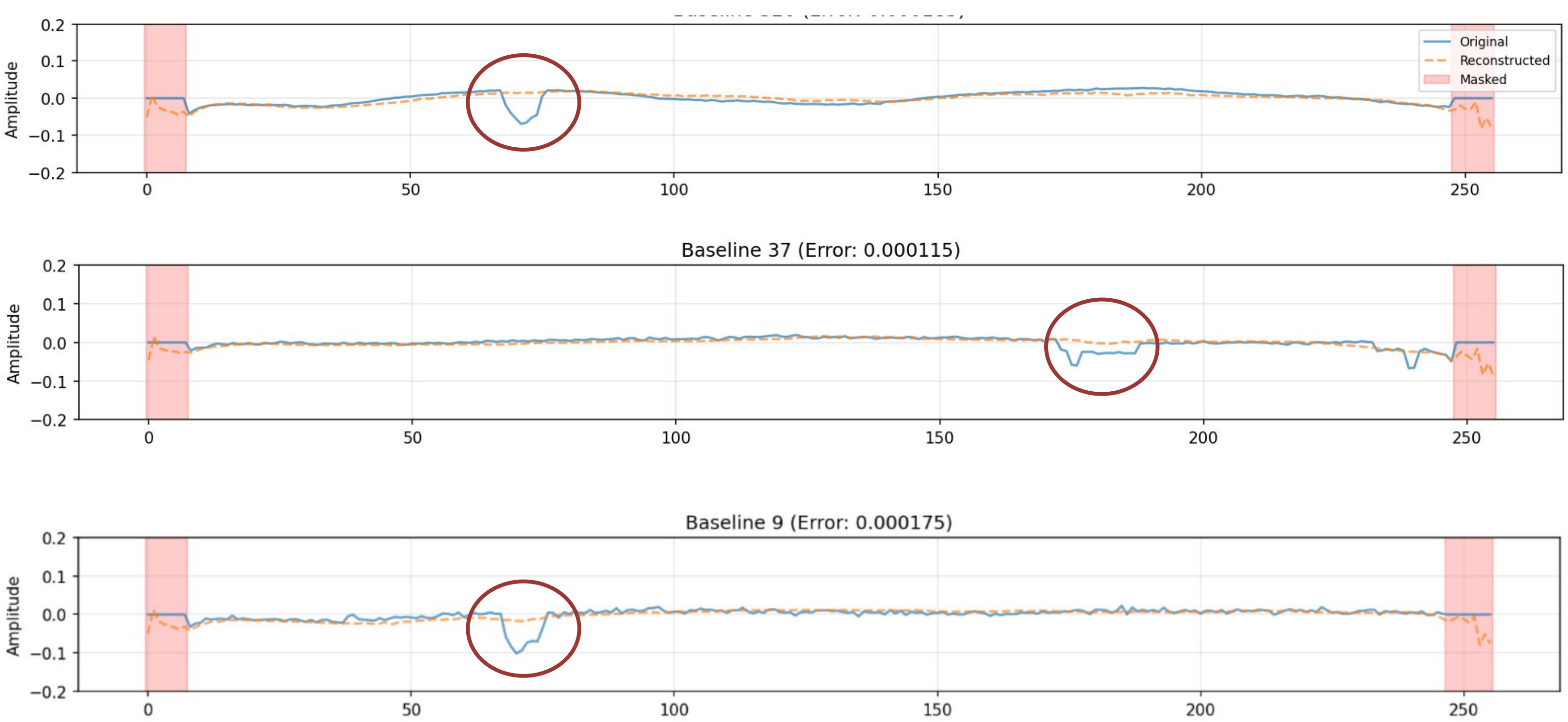
~0.11 ms per baseline

~9000 bandpass solutions/sec



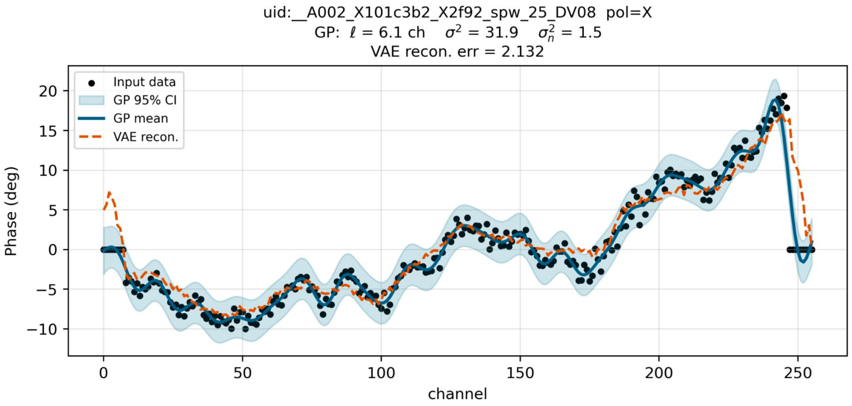
Bait+26, in prep

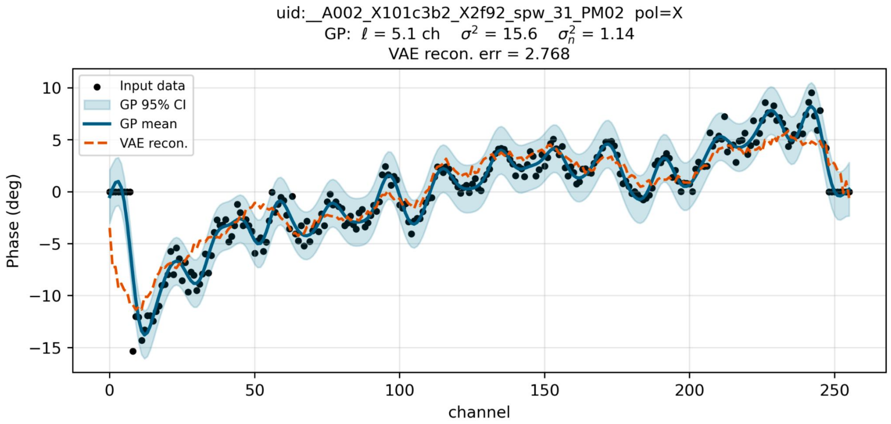
VAE on Anomalous Bandpass Solutions

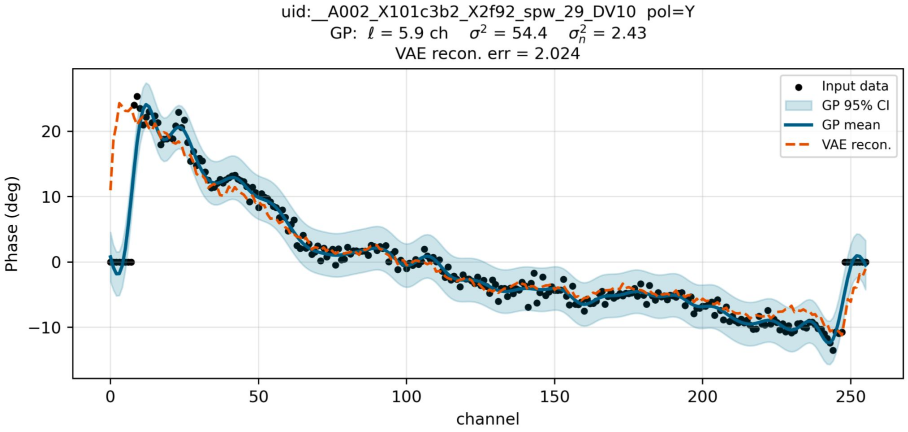


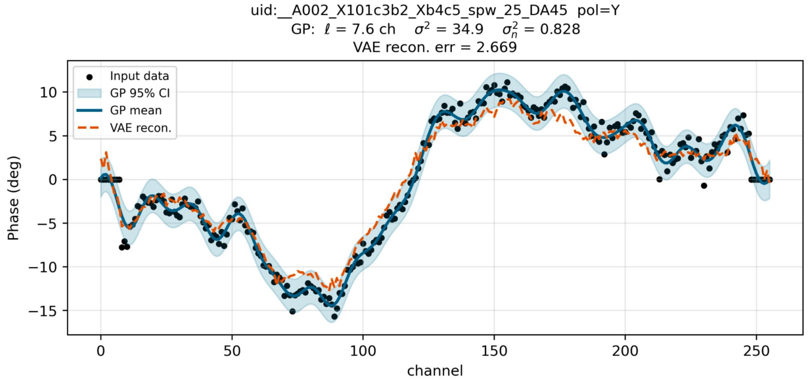
Bait+26, in prep

VAE on Good Phase solutions



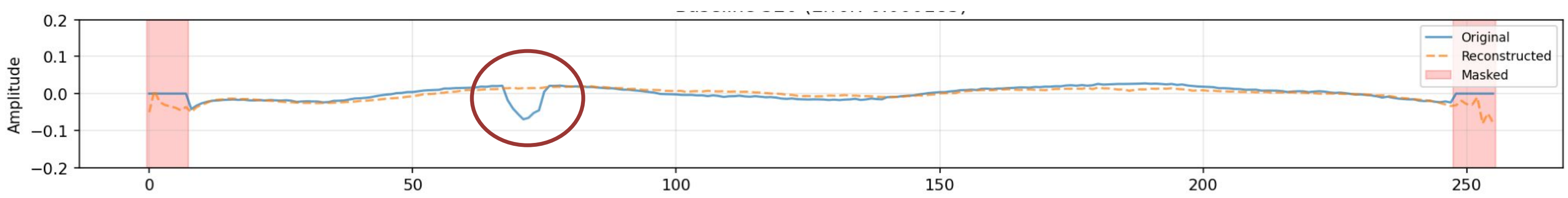






Bait+26, in prep

A classifier to identify anomalies



Input features

Max. Residual

Mean Residual

Residual Noise etc.

Residual scan stats

ML Classifier: Random Forest, XGBoost

Anomaly detection

True

False

See Ci Xue's talk tomorrow

Summary and Current Status

- Built a self-supervised, **ATM-conditioned (FiLM) VAE** that learns good ALMA bandpass solutions trained on 83,093 Cycle-9 solutions.
- Extends to phase: VAE reconstructions on good phase solutions.
- Fast: ~9K solutions/sec on a desktop CPU scalable to WSU data volumes.
- Shows promising results in identifying anomalies.

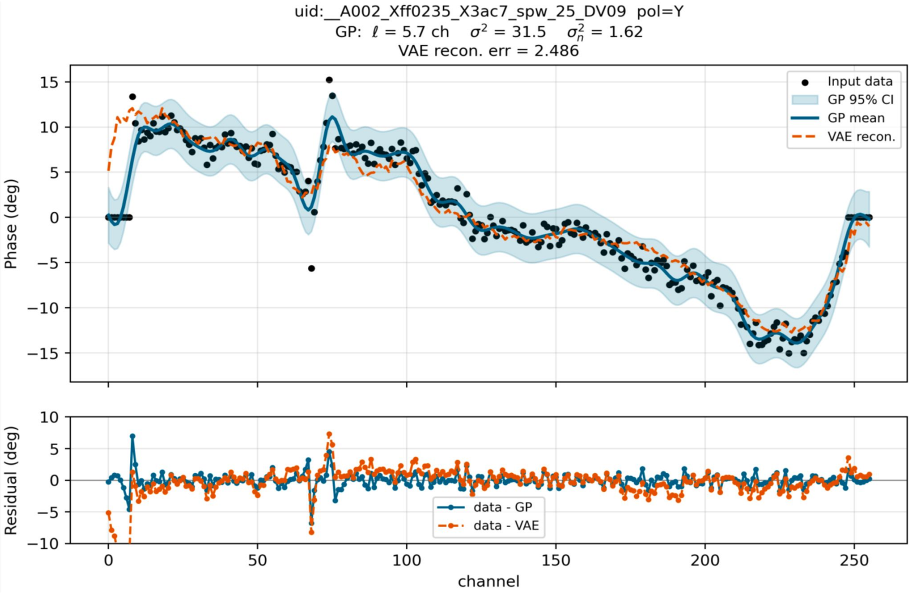
Future Work:

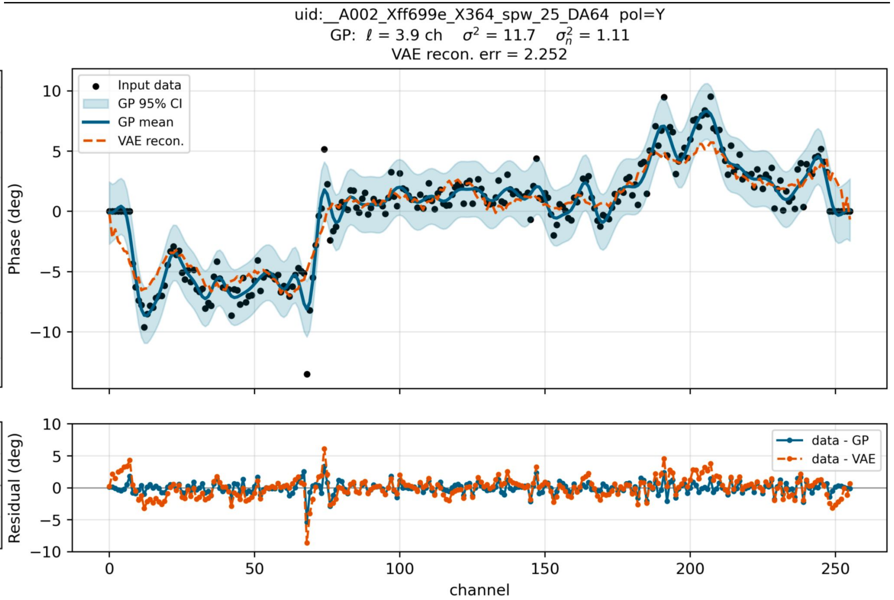
- Detect anomalies using a residual-feature classifier (in progress).
- Use anomaly scores from the latent space.
- Extend to calibrated visibilities.

Thank you

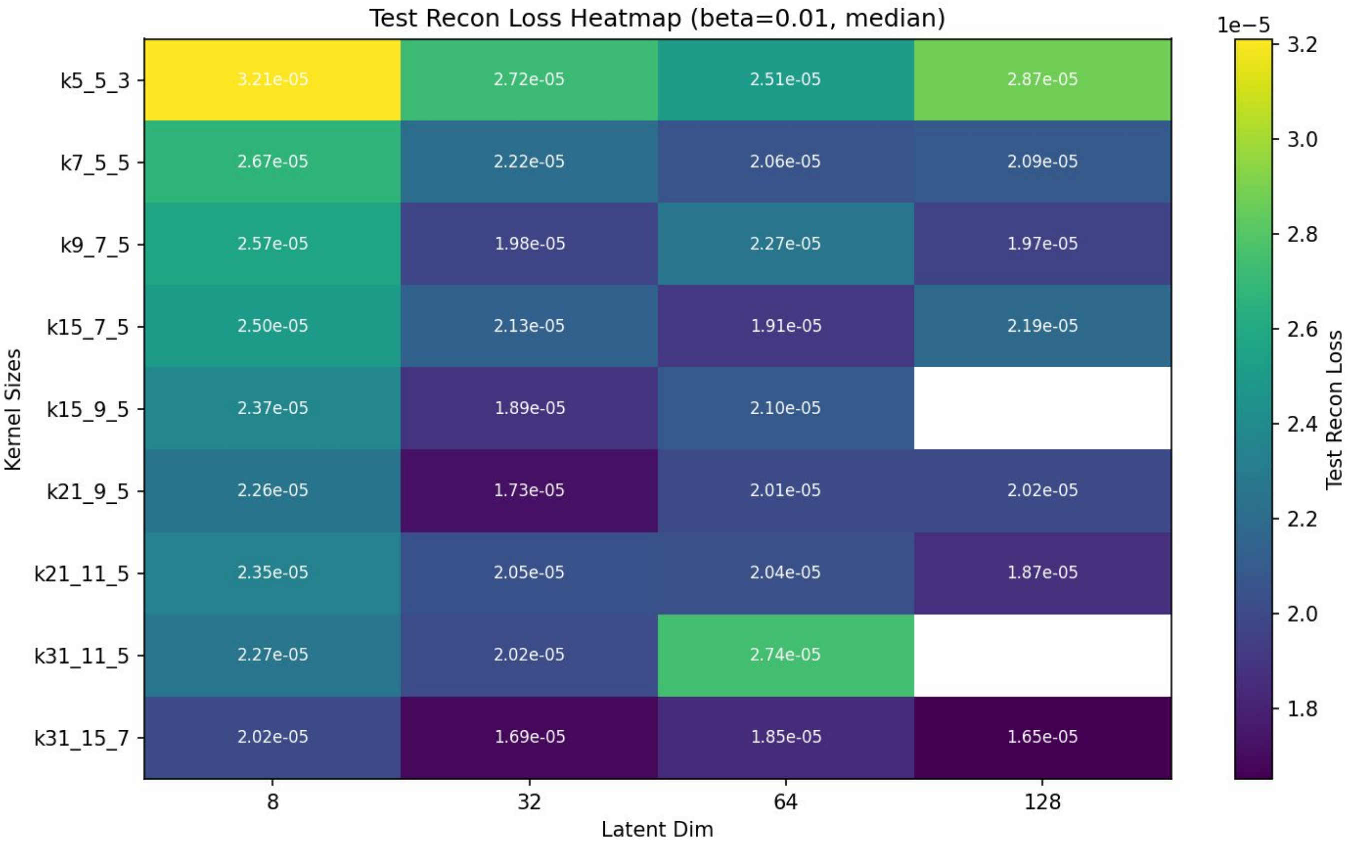
Extra Slides

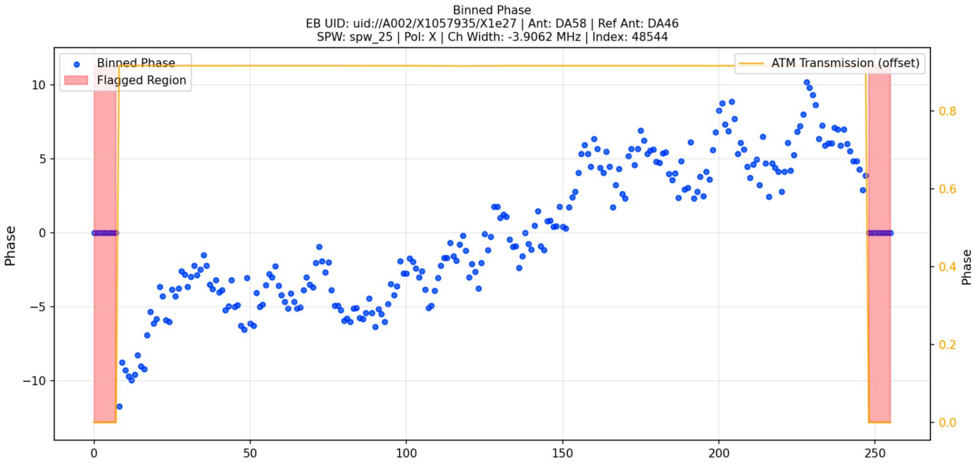
GPR and VAE overplot on Bad Phase solutions

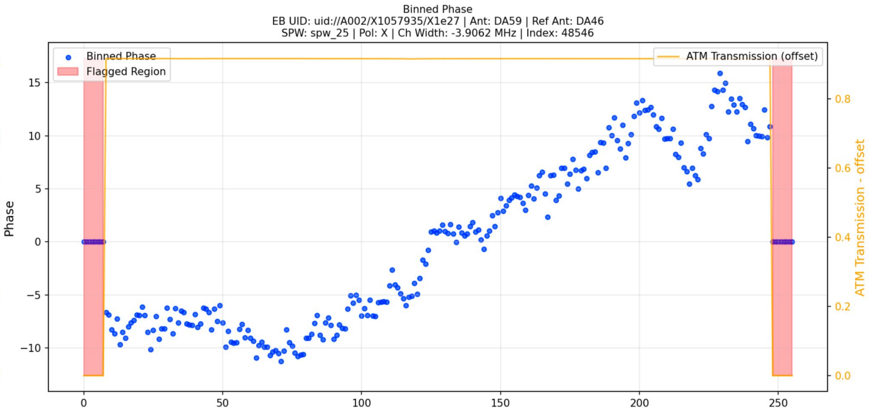


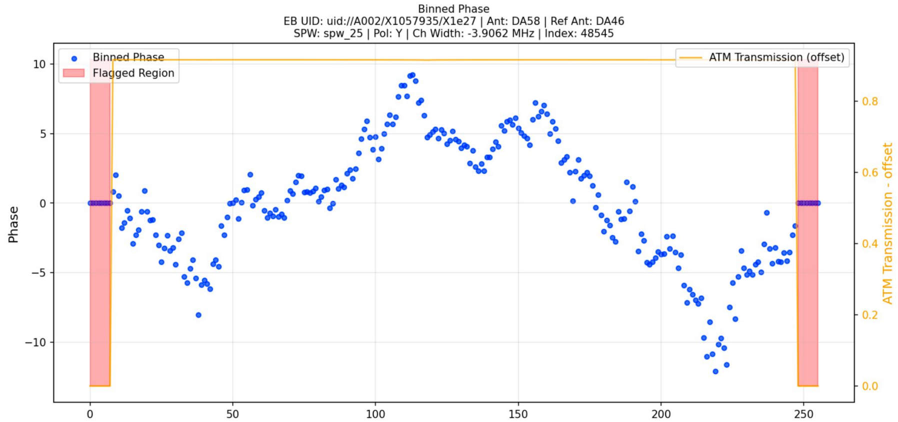


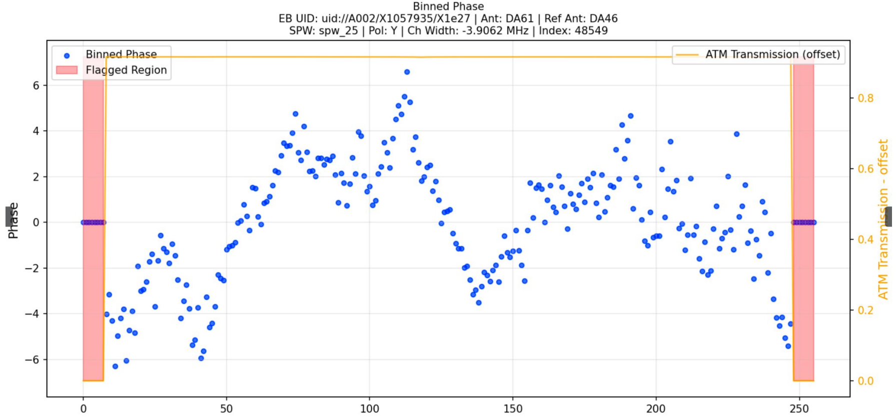
β-VAE: Ablation Study – Median Recon_error



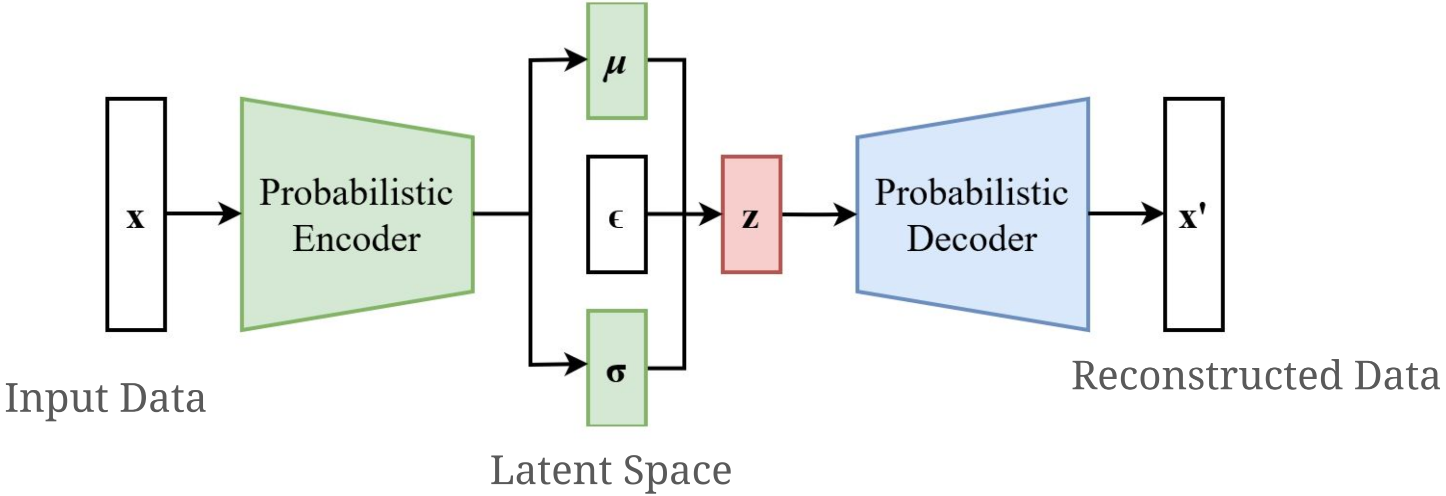

Binned Phase
EB UID: uid://A002/X1057935/X1e27 | Ant: DA58 | Ref Ant: DA46
SPW: spw_25 | Pol: X | Ch Width: -3.9062 MHz | Index: 48544
Binned Phase
Flagged Region
ATM Transmission (offset)
Phase


Binned Phase
EB UID: uid://A002/X1057935/X1e27 | Ant: DA59 | Ref Ant: DA46
SPW: spw_25 | Pol: X | Ch Width: -3.9062 MHz | Index: 48546
Binned Phase
Flagged Region
ATM Transmission (offset)
Phase
ATM Transmission - offset


Binned Phase
EB UID: uid://A002/X1057935/X1e27 | Ant: DA58 | Ref Ant: DA46
SPW: spw_25 | Pol: Y | Ch Width: -3.9062 MHz | Index: 48545
Binned Phase
Flagged Region
ATM Transmission (offset)
Phase
ATM Transmission - offset


Binned Phase
EB UID: uid://A002/X1057935/X1e27 | Ant: DA61 | Ref Ant: DA46
SPW: spw_25 | Pol: Y | Ch Width: -3.9062 MHz | Index: 48549
Binned Phase
Flagged Region
ATM Transmission (offset)
Phase
ATM Transmission - offset

Roadmap: Towards Visibility data



Use VAE → low-dimensional representation of well calibrated visibility data

Identify anomalies in the visibilities.

Useful for consequentiality studies.