21cm foreground removal with machine learning

Tianyue Chen

Postdoctoral Researcher

Ecole Polytechnique Fédérale de Lausanne (EPFL)

SKA SWG meeting 2023

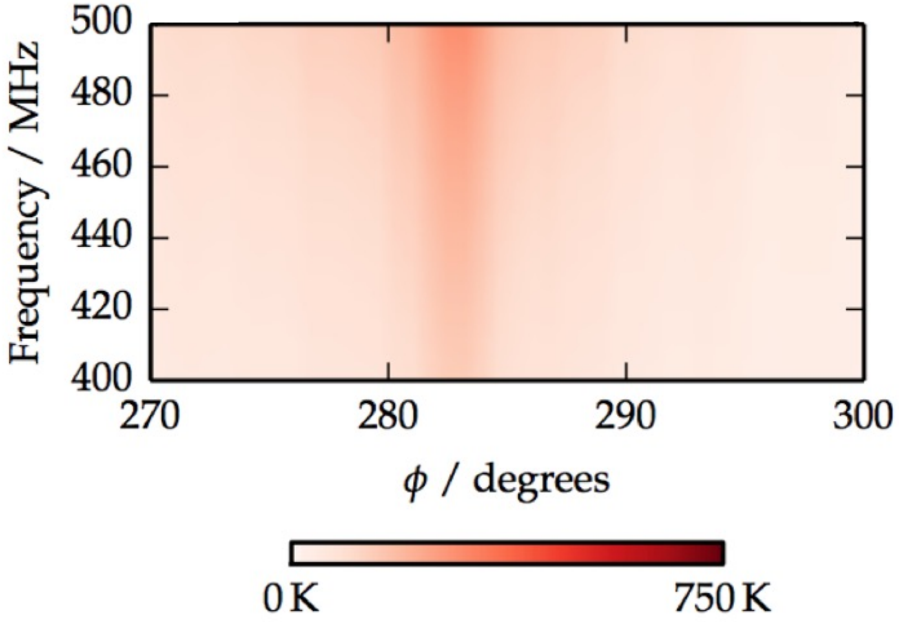
Introduction

- IM as one main goal of SKA
- Foreground critical for 21cm detection
- Large SKA dataset incoming

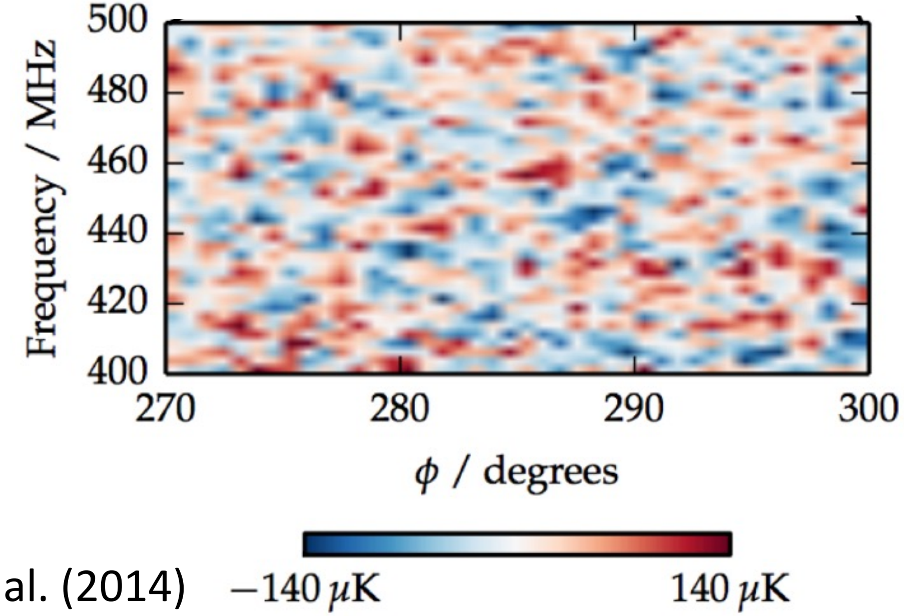
- Traditional approach:
  - Sensitive to systematics (e.g., KL filter)
  - Signal loss (e.g., PCA)

- Can we design a machine learning algorithm?
  - Effectively remove FG
  - Robust againt systematics
  - Handle large dataset

Synchrotron



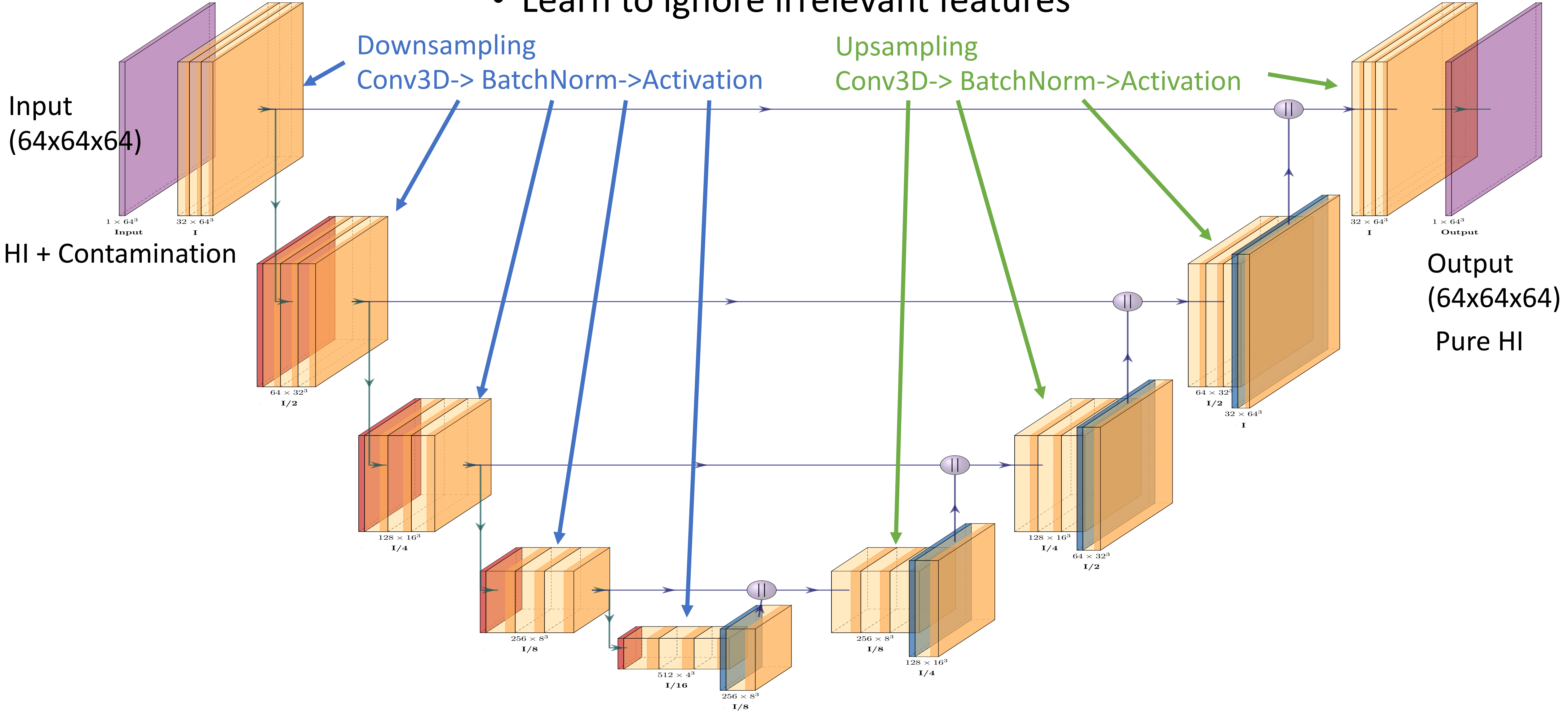
HI



Shaw et al. (2014)

U-net for IM

- One type of artificial neutral network
- Learn to ignore irrelevant features



Sky models

- Foreground models:
  - Santos et al. (2005) $$C_\ell(\nu_i, \nu_j) = A\left(\frac{1000}{\ell}\right)^\beta \left(\frac{\nu_{\rm ref}^2}{\nu_i \nu_j}\right)^\alpha I_\ell^{ij}$$
  - Planck Sky Model: Synchrotron → Haslam 408 map; Free-free → Hα template
- HI model:
  - Battye et al. 2013 $$\bar{T}_{\rm obs}(z) = 44\mu{\rm K}\left(\frac{\Omega_{\rm HI} h}{2.45^{-4}}\right)\frac{(1+z)^2}{E(z)}$$
  - PINOCCHIO light cone (Spinelli et al. 2022)
  - CoLoRe (lognormal fields)

- Beam: SKA-mid single dish beam
- Frequency range: 700-1020 MHz, 64 channels

- Format: Healpix full sky maps → 192 equal-size patches (64x64x64)

Instrumental systematics

- Bandpass error G_v = 1+ΔG_v
  - ΔG_v- Gaussian random of Nfreq

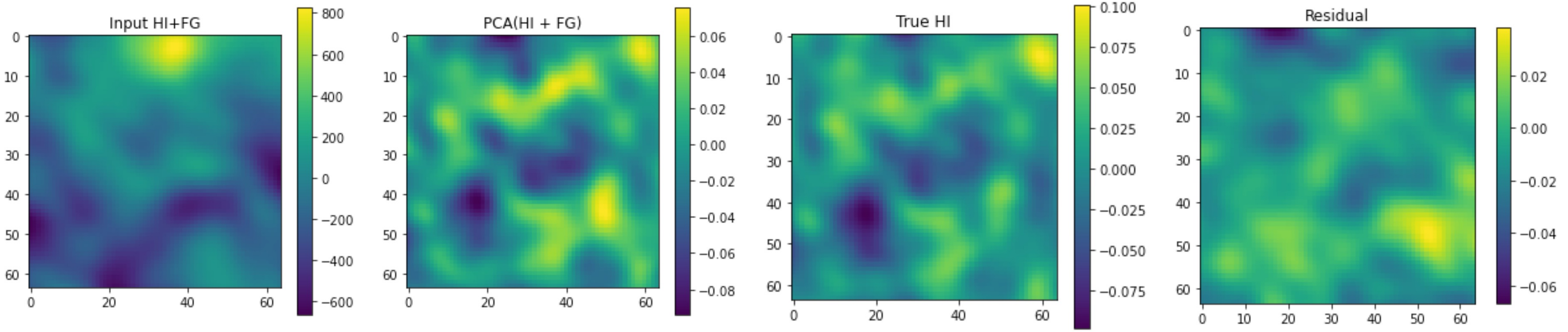
- Frequency-dependent beam

$$\theta_{\mathrm{B}}(z_i) = \theta_{\mathrm{FWHM}}(\nu_{\mathrm{mid}})\frac{\nu_{\mathrm{mid}}}{\nu_i}$$

- Radio Frequency Interference
  - Flaging random channels

PCA Pre-processing

- Network can't handle large dynamic range
- Apply PCA to pre-process the data (mode = 2)
- Use ML for fine tuning



Based on Santos (FG) + Battye (HI) model
No systematics

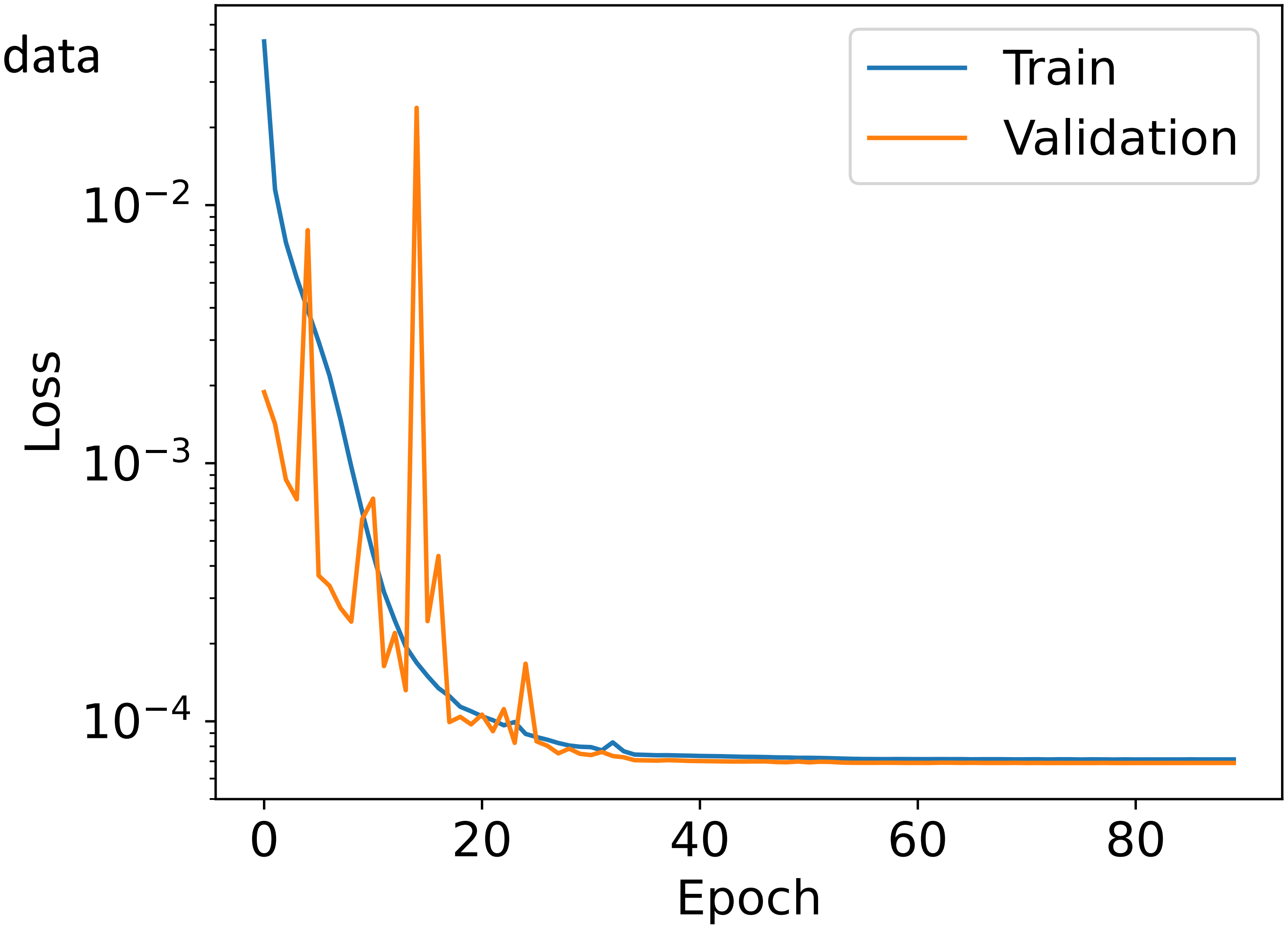
Loss function

Minimise reconstruction errors (loss)

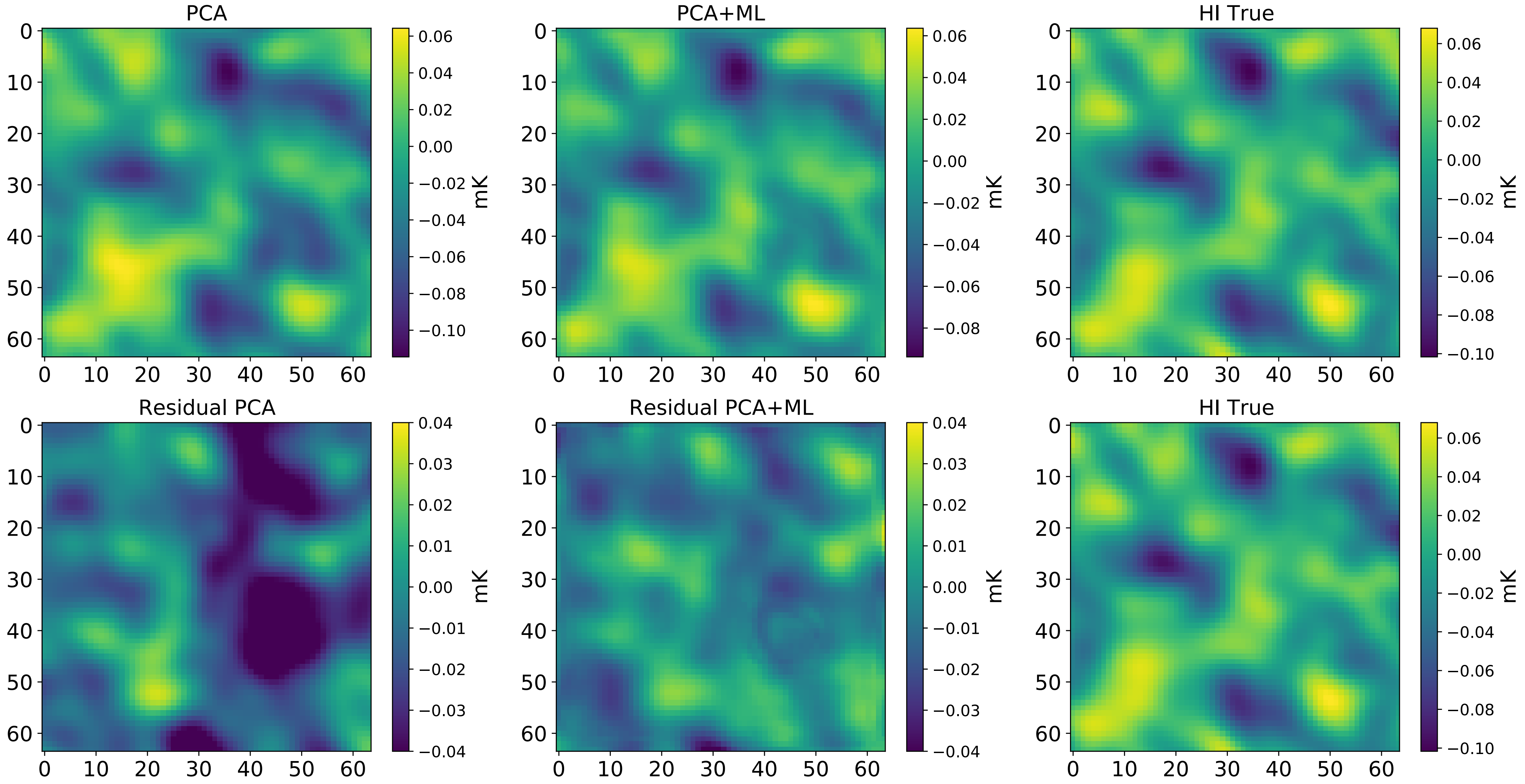
$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2$$

Train loss: how model fits the training data
Validation loss: how model performs

- Training: 50 healpix maps
- Validation: 10 healpix maps
- Test: 10 healpix maps



Map results



PCA
PCA+ML
HI True
Residual PCA
Residual PCA+ML
HI True
mK
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
0.03
0.01
−0.01
−0.03
0
10
20
30
40
50
60

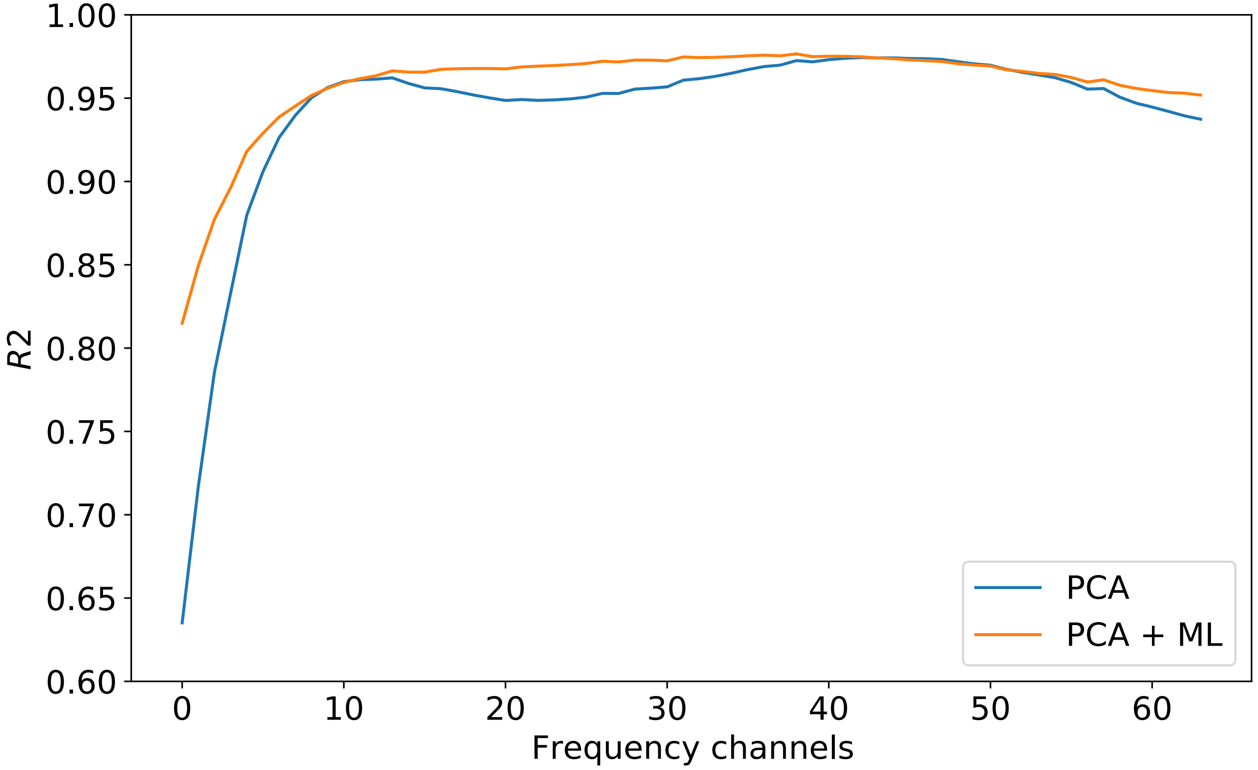
R^2 Score

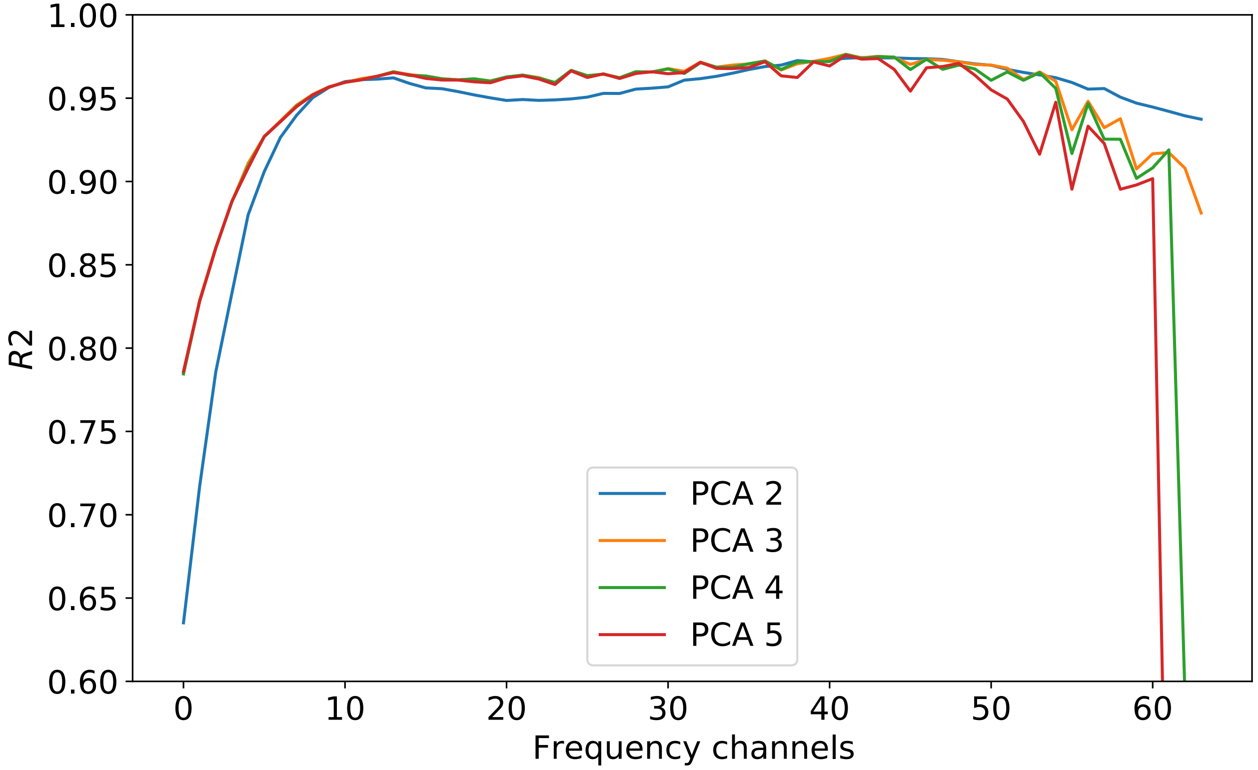
Coefficient of determination

$$R^2 = 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}$$

Evaluate the performance of the ML model

Accuracy measurement of predictions v.s. target





Conclusions and ongoing work

- Consistent results with traditional methods

- Test the network on different simulation models
  - Reliable performance independent of sky models

- Introduce systematics in the data
  - Robust against systematics