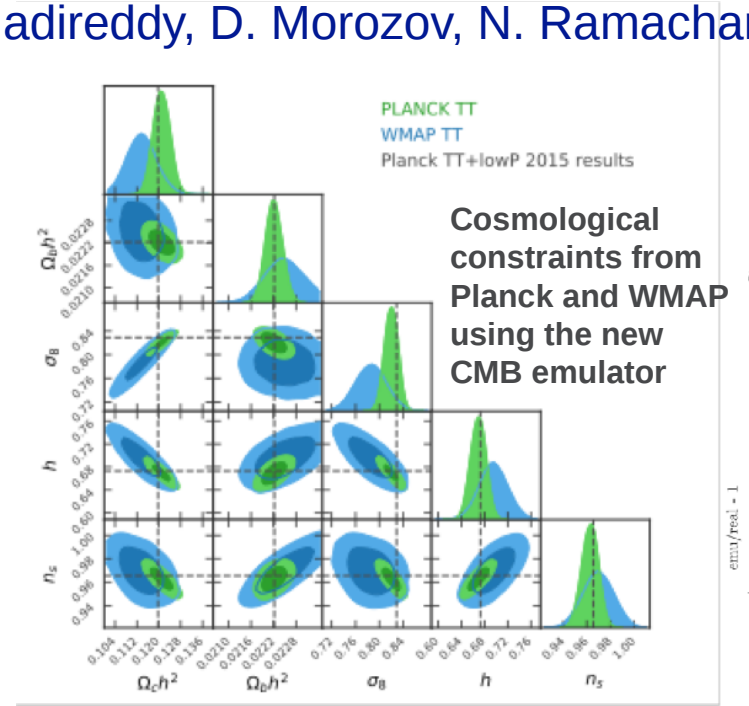
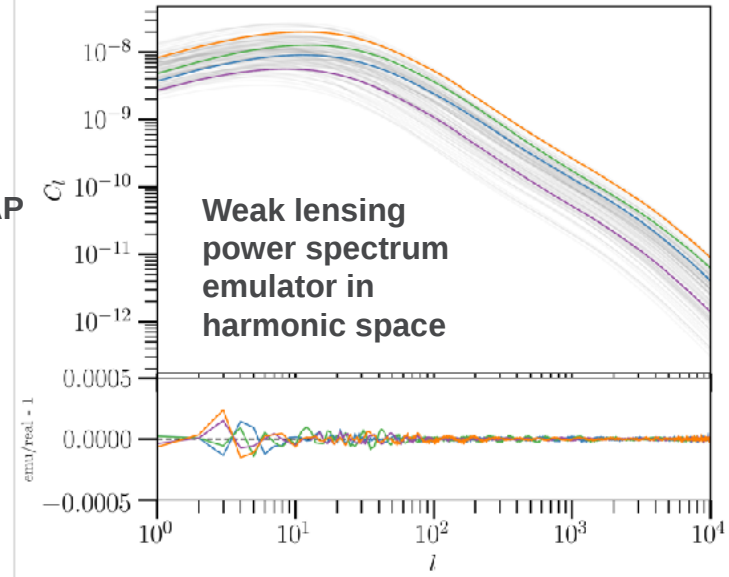
SciDAC-4: Inference and Machine Learning at Extreme Scales

P. Balaprakash, M. Binois, J. Chaves-Montero, A. Fadikar, R. Gramacy, S. Habib (PI), K. Heitmann, D. Higdon, P. Larsen, E. Lawrence, Y. Lin, Z. Lukic, S. Madireddy, D. Morozov, N. Ramachandra, A. Slosar, S. Wild, S. Yoo

- **Opportunity:** Use of HPC resources for data-intensive statistical and machine learning (ML) methods for discovery science
- **Project Thrusts:** Melding HPC and observational datasets with ML/stats to solve scientific inference problems in Cosmic Frontier experiments — CMB-S4, DESI, LSST, SPT
- **Applications/Impact:** Cosmological parameter estimation, strong lens image classification/characterization, synthetic sky catalog generation, fast prediction of summary statistics, fast likelihood estimation, improved photo-z estimation





- **Examples — Cosmic Microwave Background and Weak Lensing**
  - Large training data sets using direct computation (CMB) or first-level emulation (lensing)
  - Dimensional reduction via unsupervised learning
  - High-dimensional non-parametric regression
  - Emulators yield **~1000** speed-up with 0.5% errors over the desired dynamic range, technique used by all future surveys

http://press3.mcs.anl.gov/cpac/projects/scidac/

Argonne NATIONAL LABORATORY · BROOKHAVEN NATIONAL LABORATORY · BERKELEY LAB · Los Alamos NATIONAL LABORATORY EST. 1943 · VIRGINIA TECH

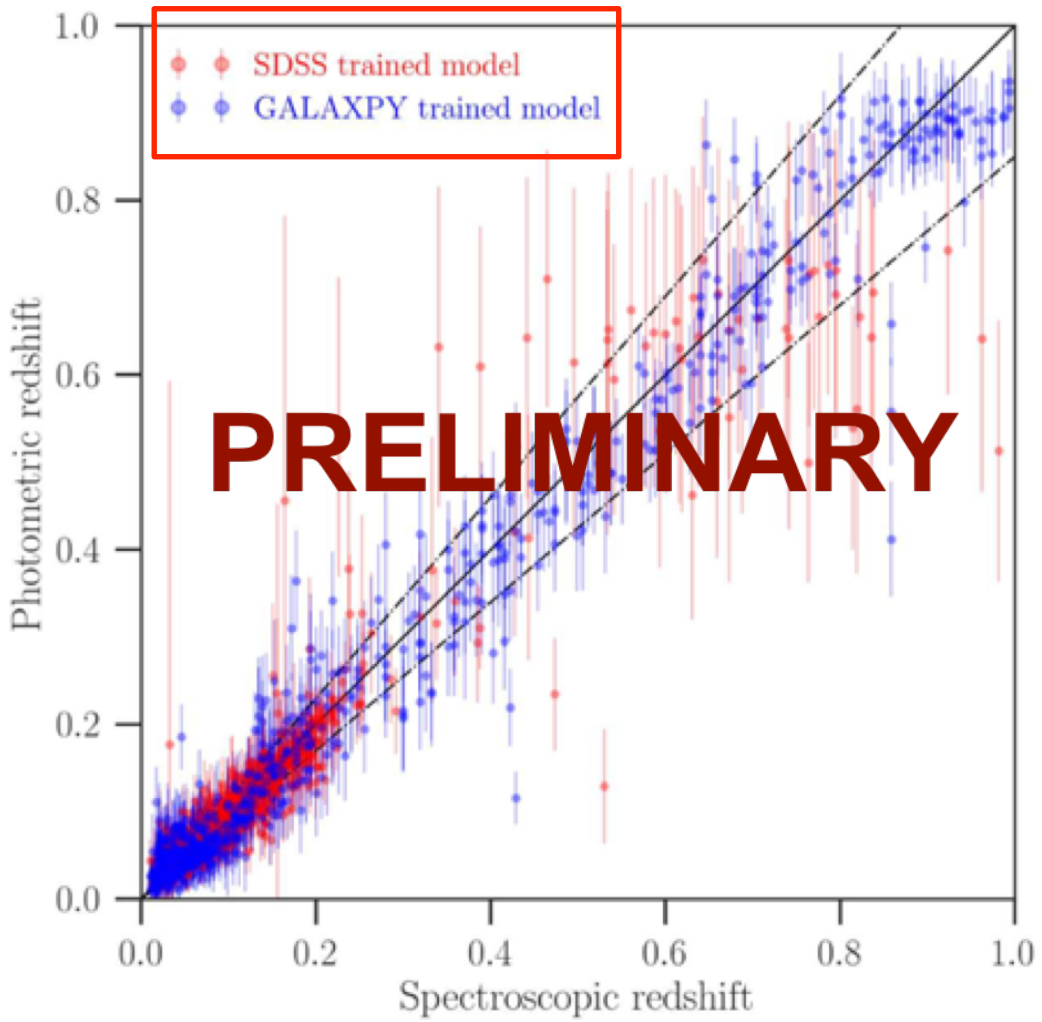
Unraveling the Cosmic Puzzle: HPC and ML Opportunities

- **Idea:** Use of HPC to generate high-fidelity, large data-volume, ***training sets*** for statistical and ML methods
- **Forward Modeling:** Use fast, ***high-accuracy forward models*** to attack a host of inverse problems with Bayesian methods
- **Uncertainty Quantification:** Error analysis and systematics modeling to develop a new set of ***robust analysis techniques***
- **Impact:** Methods developed by this team, especially the ***'emulation'*** concept are becoming the preferred approach for next-generation surveys

Cosmological scientific inference process showing forward modeling and systematic error exploration/control loop

Argonne NATIONAL LABORATORY · BROOKHAVEN NATIONAL LABORATORY · BERKELEY LAB · Los Alamos NATIONAL LABORATORY EST. 1943 · VIRGINIA TECH

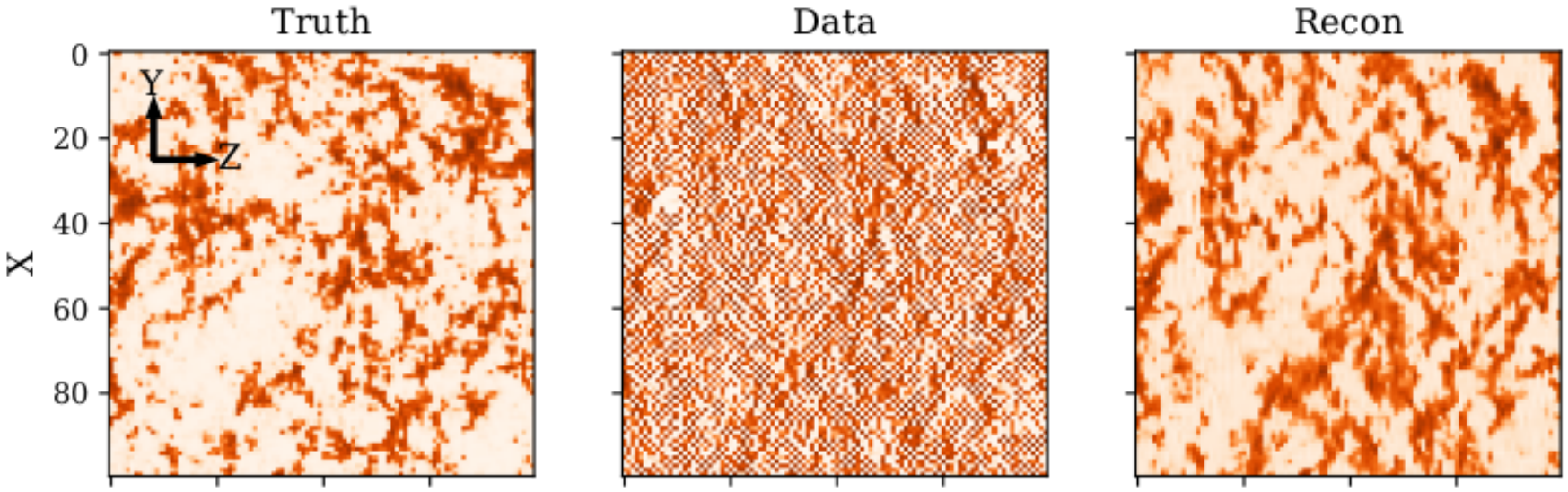
New Technique for Photometric Redshift Estimation



- **Redshift Estimation:** Next generation surveys will have billions of galaxies with image and “color” information (2-d); need to go from color information to the galaxy redshift (3-d) necessary for cosmology
- **Mixed Density Networks**: Method for mapping colors to redshifts, yields a p(z) for each object; degeneracies handled using Gaussian Mixtures
- **Improved Training Set:** Traditional use of ML has suffered from incomplete training sets for faint sources; we introduce GALAXPY, a detailed and robust generative model for emulating galaxy colors using Gaussian processes, fills data space not covered by observations
- **Tests on Real Data:** Clear reduction of outliers when using the GALAXPY training set versus using actual data

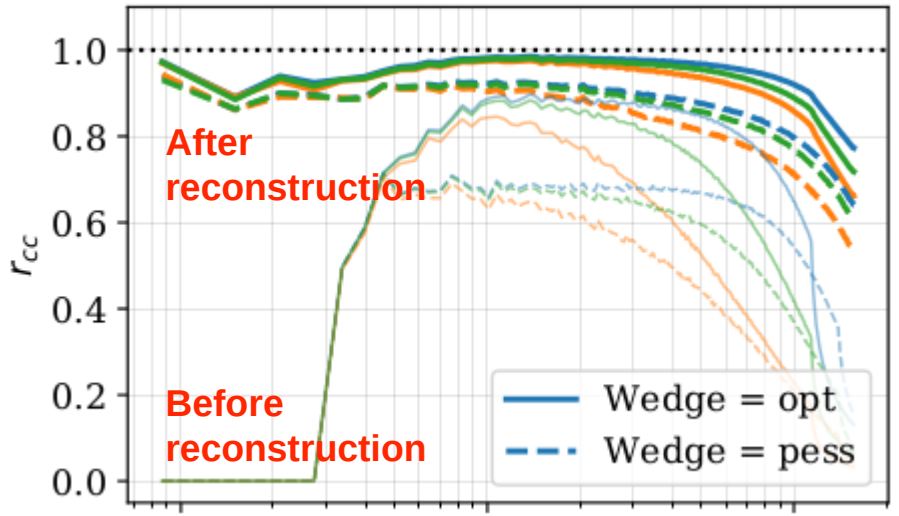
Argonne NATIONAL LABORATORY
BROOKHAVEN NATIONAL LABORATORY
BERKELEY LAB
Los Alamos NATIONAL LABORATORY EST. 1943
VIRGINIA TECH

Information Recovery: The Case of 21cm Intensity Mapping



- **Problem:** Lose cosmologically important low-k modes in 21cm intensity mapping due to foreground filtering; can the information be recovered?

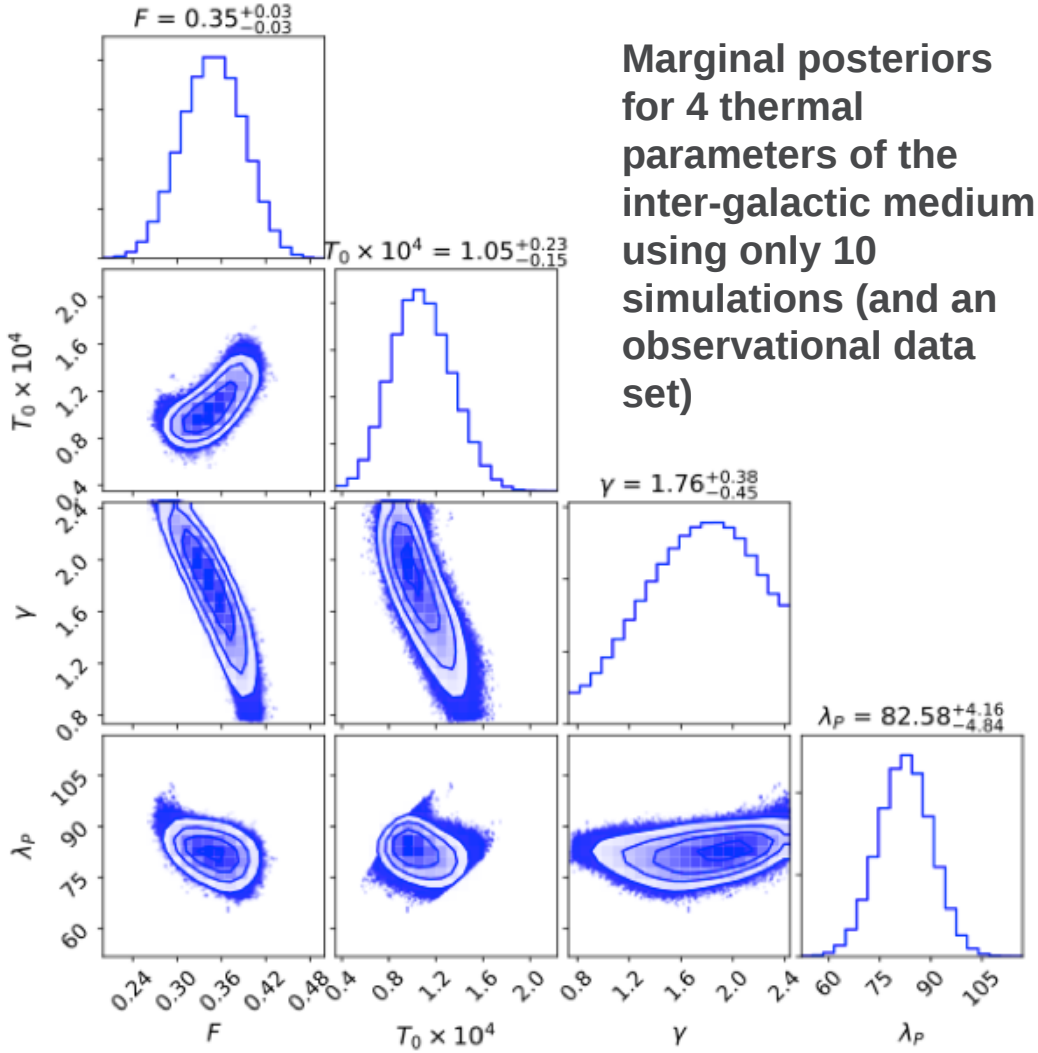
- **Nonlinear Flow:** Non-linear cosmological evolution moves low-k power to higher-k modes, less susceptible to filtering; suggests use of a ML-based reconstruction method
- **Results:** Use of ML/back-propagation methods enables solution for millions of parameters (works due to simple nature of bias and high redshift)
- **Impact:** The success of this technique will enable future 21cm experiments



Argonne NATIONAL LABORATORY
BROOKHAVEN NATIONAL LABORATORY
BERKELEY LAB
Los Alamos NATIONAL LABORATORY EST. 1943
VIRGINIA TECH

Inverse Problems with Expensive Simulations: Adaptive Exploration

- **Adaptive Exploration of Parameter Space:** HPC simulations can be very expensive; in some cases, they can be run beforehand to generate emulators, but not always — if observational data is available, efficient adaptive methods can be constructed
- **Methodology:** Used multi-output Gaussian Process emulators, adaptively constructed using Bayesian optimization
- **Science Case:** Used the Ly-alpha flux power spectrum observations to demonstrate power of new technique, demonstrating a significantly reduced number of simulations compared to non-adaptive methods



Marginal posteriors for 4 thermal parameters of the inter-galactic medium using only 10 simulations (and an observational data set)

Argonne NATIONAL LABORATORY · BROOKHAVEN NATIONAL LABORATORY · BERKELEY LAB · Los Alamos NATIONAL LABORATORY EST. 1943 · VIRGINIA TECH

Solving Many Other Inference and ML Problems

http://press3.mcs.anl.gov/cpac/projects/scidac/

- **Deep Learning:** Lensing image classification and characterization; generative models for constructing sky maps at multiple wavebands; use of AI methods for domain translation (e.g., mapping hydrodynamic properties on gravity-only simulations)
- **Likelihood-Free Methods and Likelihood Emulation:** Next steps for full forward modeling based approaches
- **Cross-Connection with HEP-CCE (HEP Center for Computational Excellence):** Summer student program for work on HPC and data-intensive computing problems
- **Collaborations with Experiments:** Working with DESI, LSST, and SPT science working groups
- **SciDAC Institutes:** Strong connection with RAPIDS

Synthetic sky catalog for LSST — source for deep learning training data sets

Argonne NATIONAL LABORATORY
BROOKHAVEN NATIONAL LABORATORY
BERKELEY LAB
Los Alamos NATIONAL LABORATORY EST. 1943
VIRGINIA TECH