COSMOLOGY USING SCIENTIFIC MACHINE LEARNING

Nesar Ramachandra
nramachandra@anl.gov
HEP Division, Argonne National Laboratory

INTRODUCTION

This talk highlights a subset of the AI/ML efforts within the Cosmology group at Argonne.

Common themes here are:

- Synthetic/simulation data to enhance/replace real astronomical observations.
- Bayesian/probabilistic schemes rather than point-predictions.
- Explainability of the ML algorithms.

Different case studies:

- Generative models using Gaussian Processes, PCA, Auto-encoders
- Probabilistic classification and regression
- Image processing pipelines for de-noising, de-blending etc.

WHY USE SYNTHETIC DATA?

Argonne NATIONAL LABORATORY

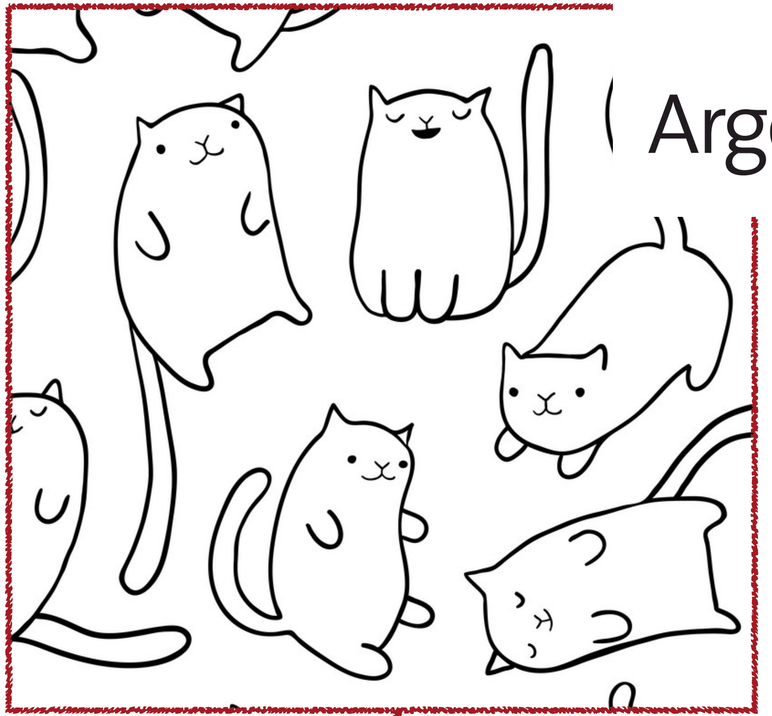
- Typically the training and testing data come from the same ‘set’
  - Assumes completeness, representation
  - Captures data-prior, biases in the training set
- In addition, data dealt in industry tends to be ‘low-cost’, and available in large volumes (required to train highly parametrized models like Deep NNs)
- Scientific Data on the other hand typically are high cost, have to be carefully sampled and may be incomplete or not-representative.

GARBAGE IN, GARBAGE OUT

Data size and complexity: Usually requires a large amount of high fidelity representative data - particularly in methods that are feature agnostic before training (like deep CNNs)

- Tricks: Transfer learning, Data augmentation, space filling/active sampling, realistic synthetic data.

Data quality: Observed data also tends to be incomplete/biased (in the parameter space of interest), noisy, and systematics may not be obvious.



TRAIN

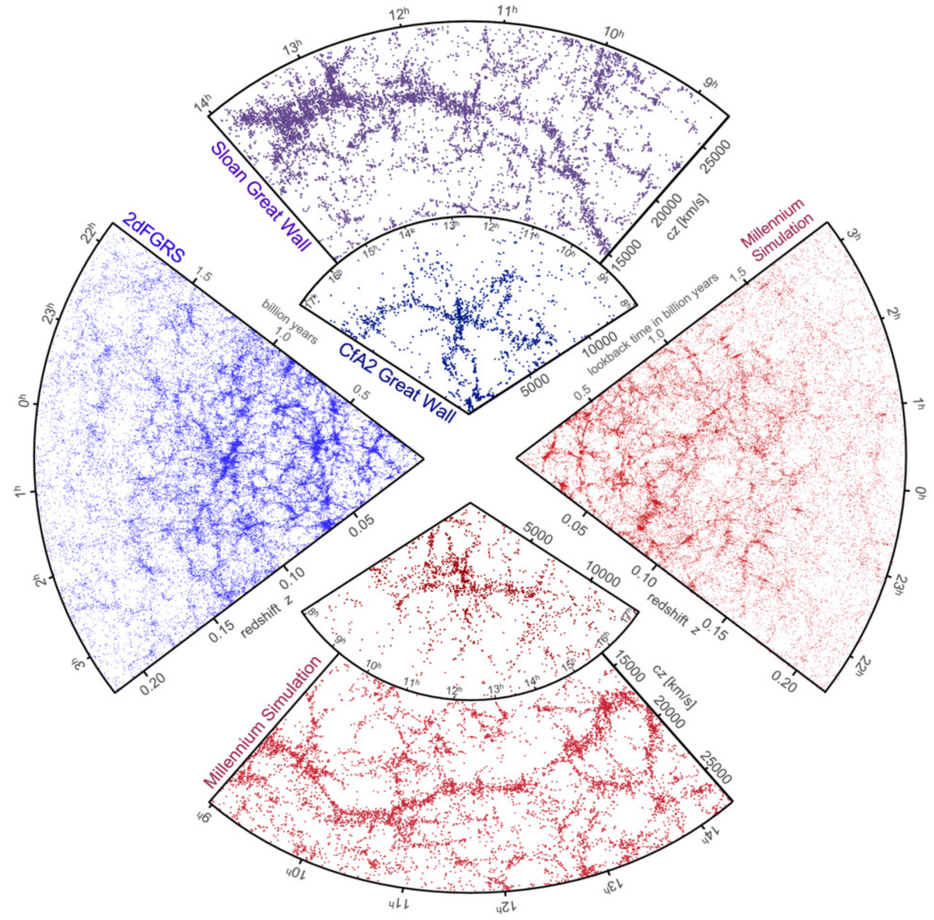
TEST

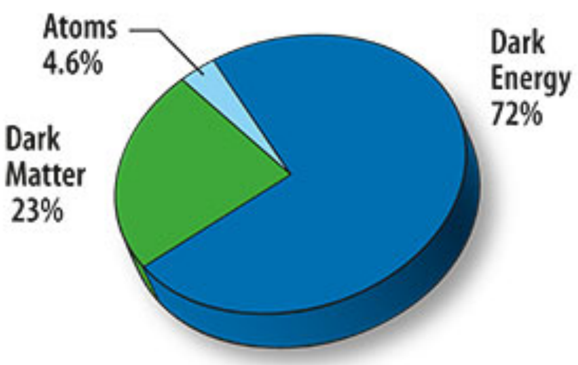
Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY | Office of Science

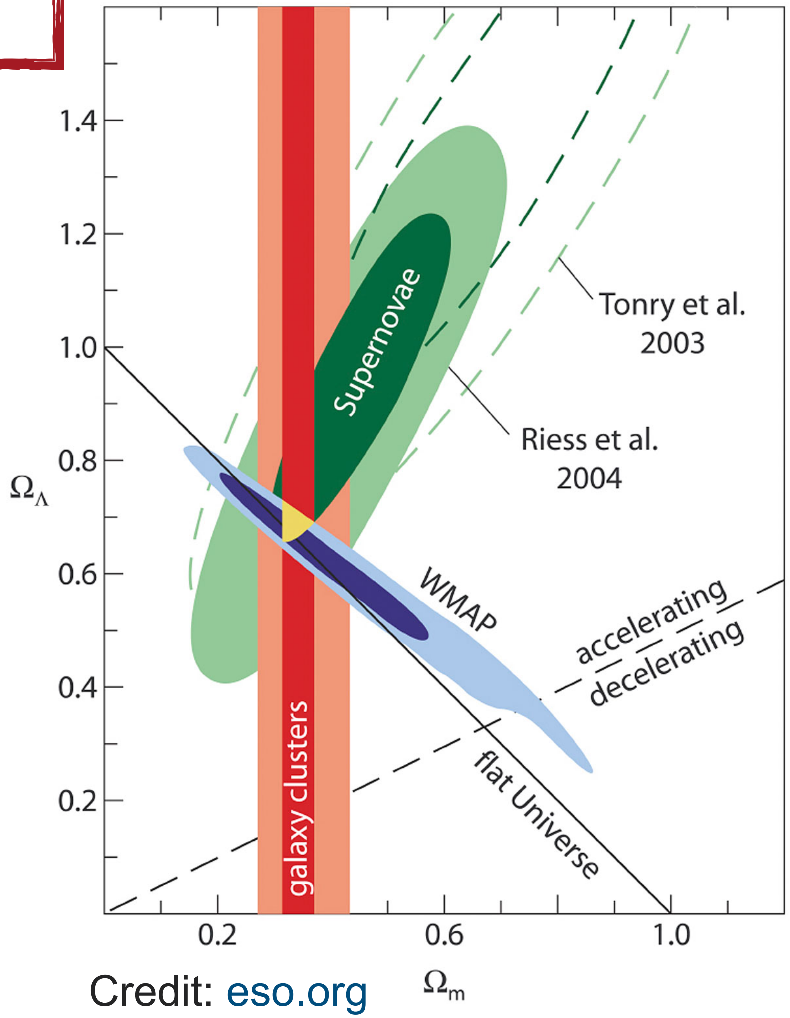
STUDYING THE COSMOS

Recent progress in Cosmology a largely data-driven
- due to numerical and observational data





Credit: nasa.gov



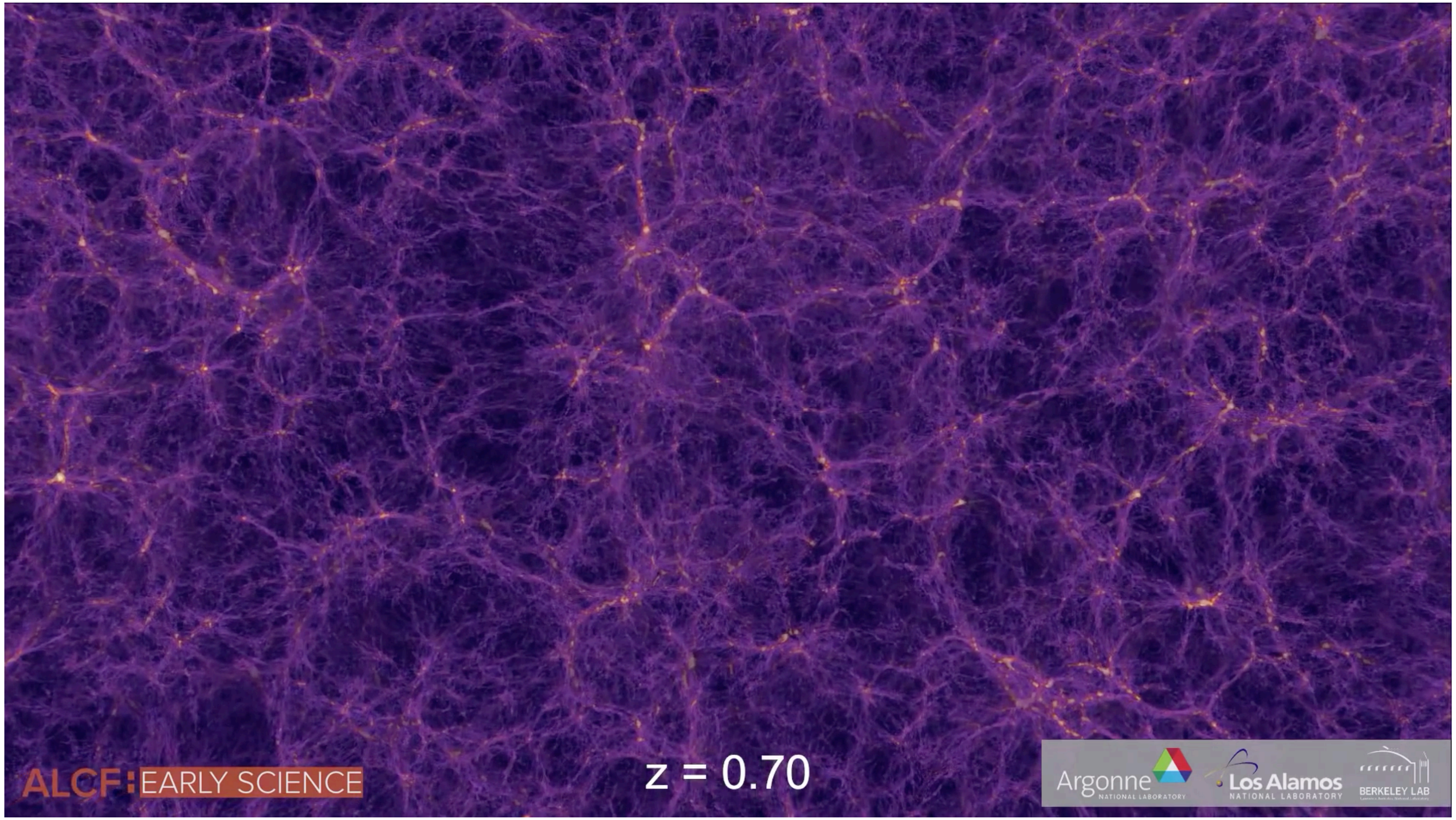
Credit: eso.org

Zavala, J.; Frenk, C.S. Dark Matter Haloes and Subhaloes. *Galaxies* **2019**, *7*, 81.

SYNTHETIC DATA FOR COSMOLOGICAL PARAMETER CALIBRATION

Motivation:

- Unfortunately we only have one observable Universe
- Expensive Cosmological simulations or summary statistics are essential



Outer Rim simulation: youtu.be/rtBIZJ6gNiI

FAST GAUSSIAN PROCESS EMULATORS

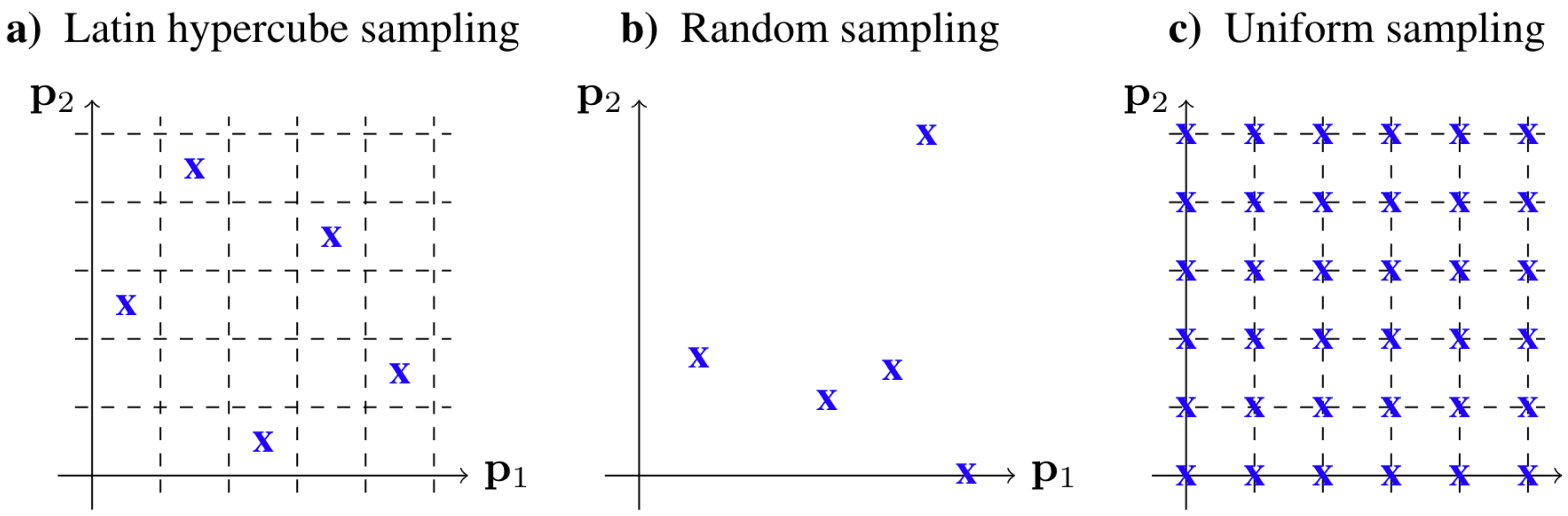
Argonne NATIONAL LABORATORY

Motivation:

Simulations themselves can be very expensive, one may replace their summary statistics with cheap emulators

Cosmic Emu - Heitmann et al 2006 and others: hep.anl.gov/cosmology/CosmicEmu)

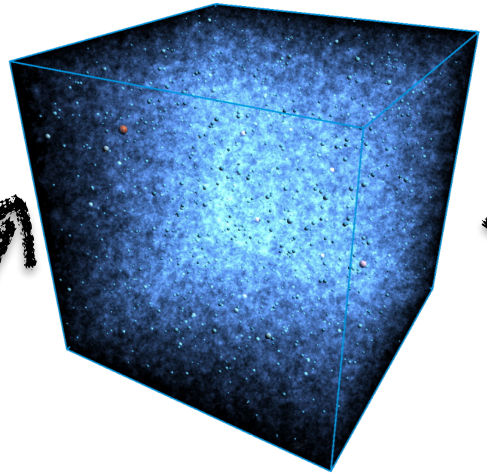
Sampling schemes for synthetic data is very important while dealing with expensive simulations

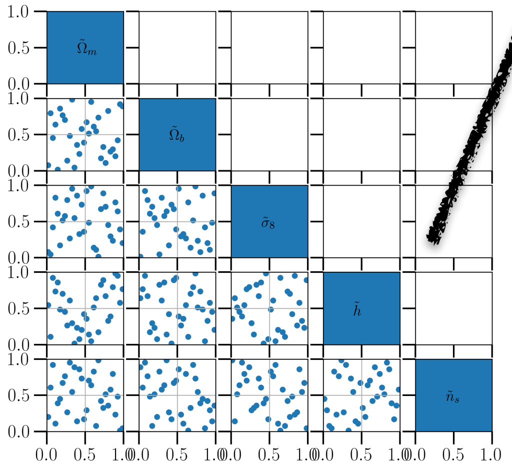


	a)	b)	c)
Sampling guaranties	**Representative**	**No guaranty**	**Representative**
Number of samples	M	M	N^M

GP-PCA EMULATION PIPELINE

Run training simulations, generate summary statistics



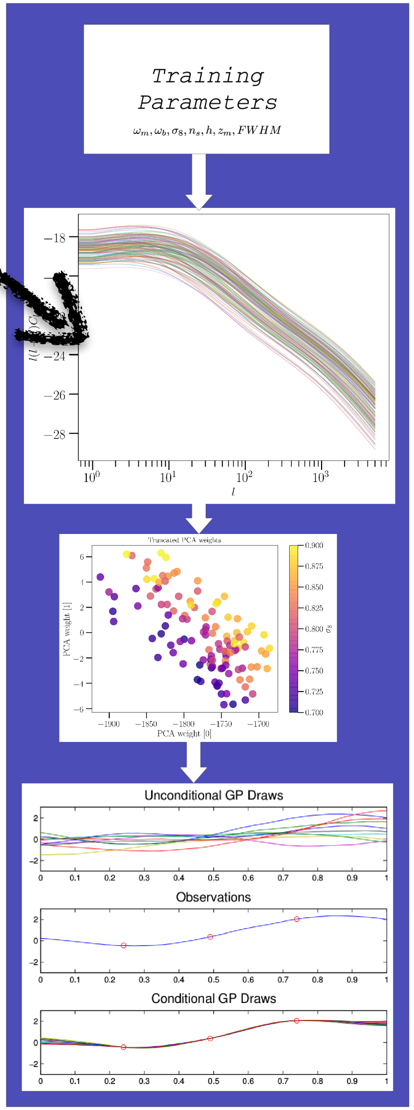


Experimental design: space filling latin hypercube

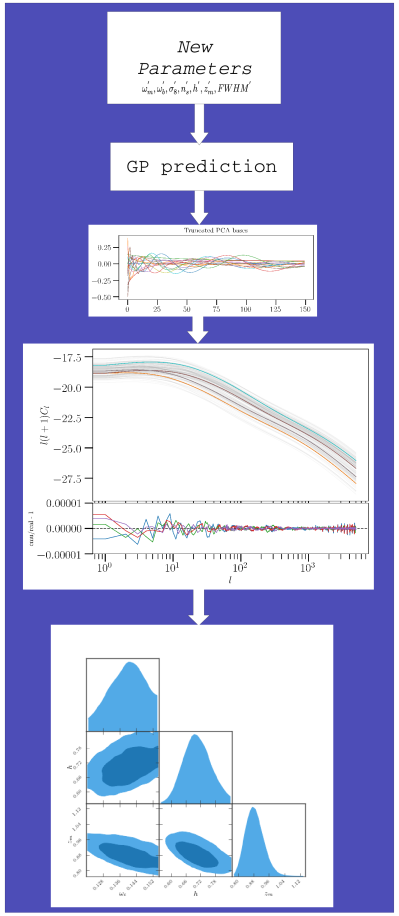
Emulation:

$$\chi(k;\theta) = \sum_{i=1}^{p_n} \phi_i(k) w_i(\theta) + \epsilon$$

PCA bases GP weights Error



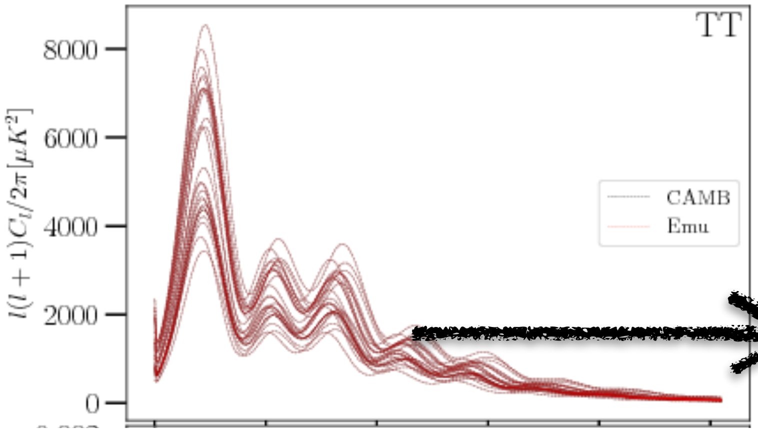
PCA reduction, GP training



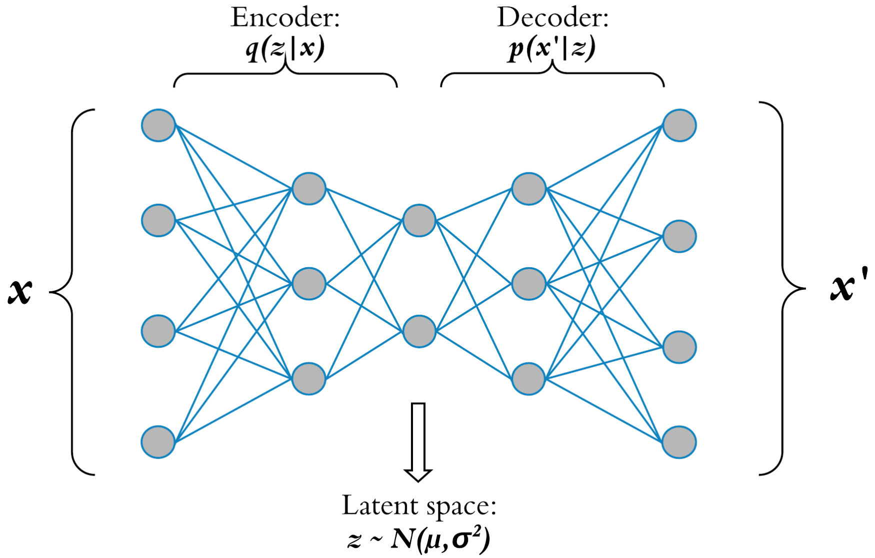
Emulation at new parameters, used in an inference pipeline

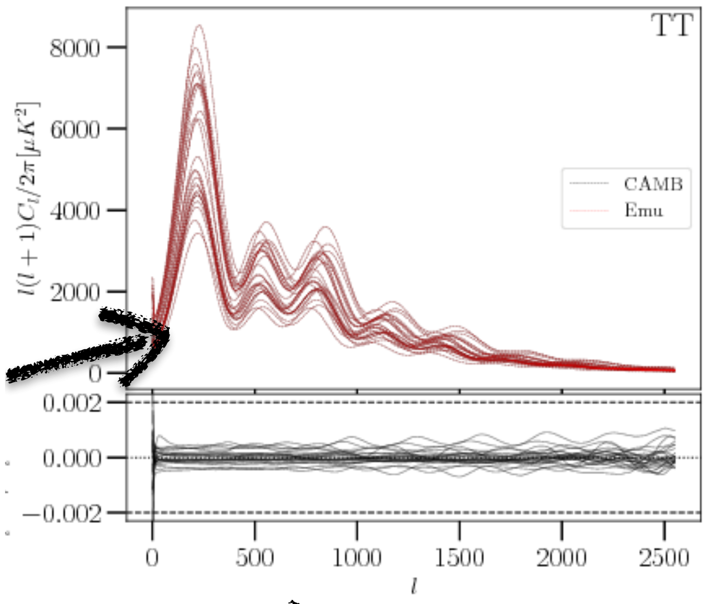
GP EMULATION WITH VARIATIONAL AUTO-ENCODERS

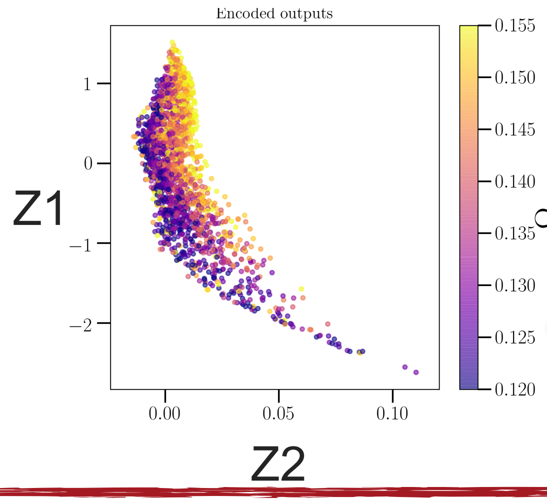
Argonne



Unsupervised VAE learning





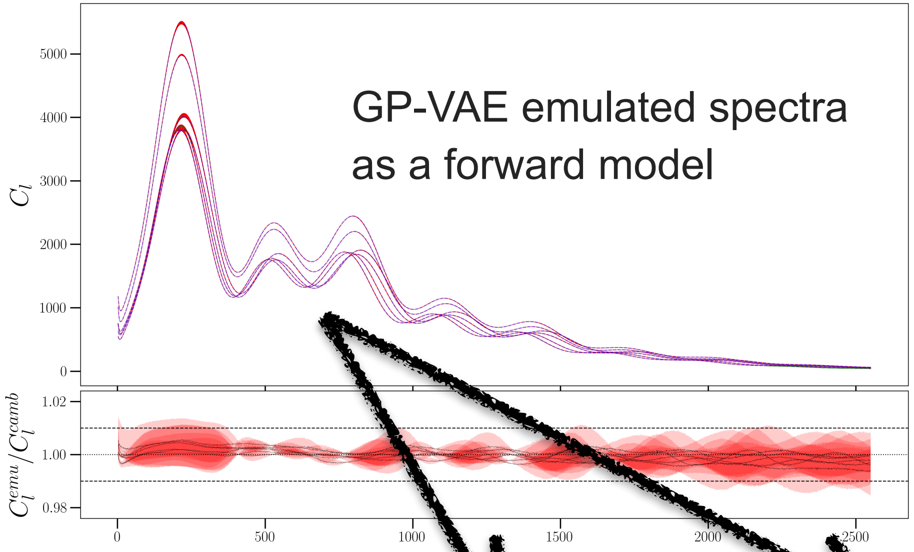


Emulation = Profit !!

$$w(\theta) = GP(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^{'}); \theta)$$

Supervised GP/
representation learning

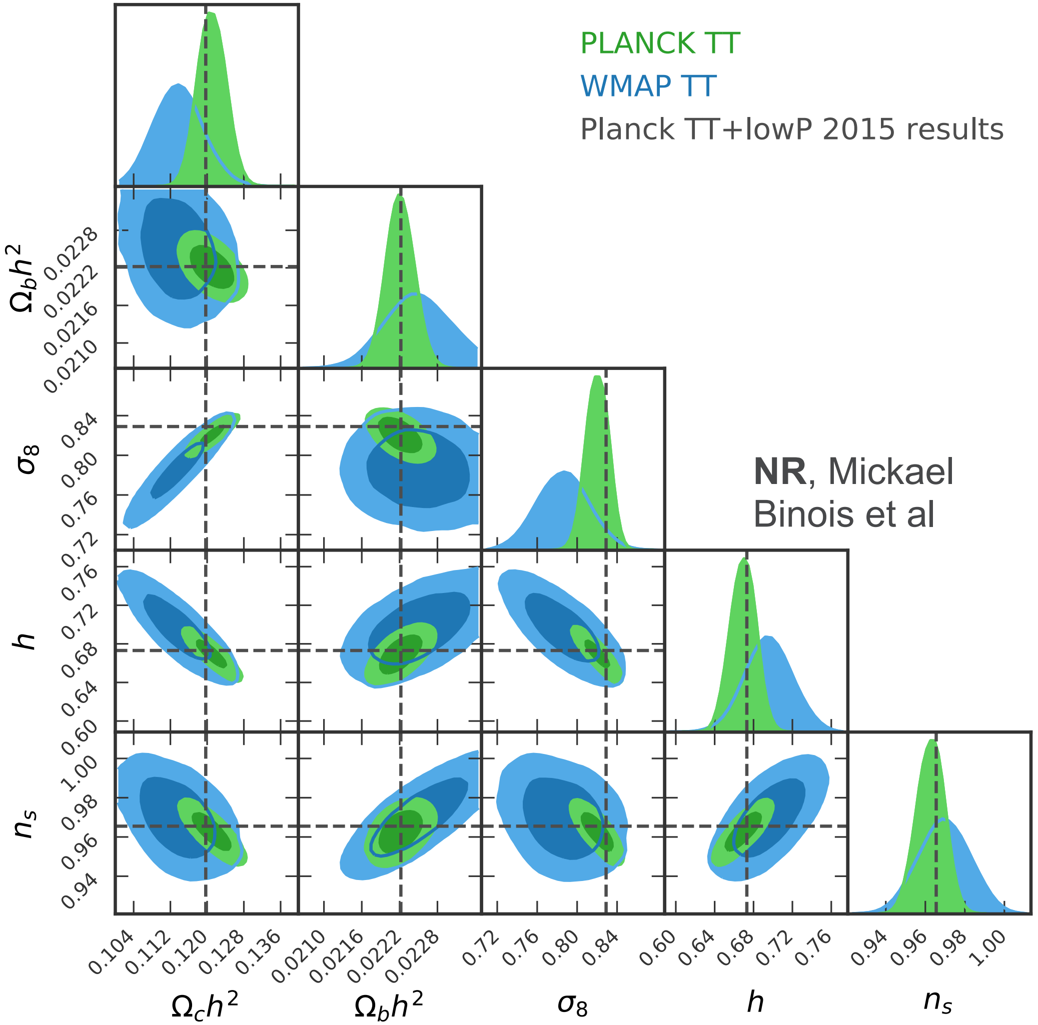
BAYESIAN INFERENCE WITH EMULATORS



$$\mathscr{L}(D|\theta) \propto \exp\left[-\frac{1}{2}\sum_{i,j}\left(D-f(\theta)\right)_i C_{ij}^{-1}\left(D-f(\theta)\right)_j\right]$$

$$P(\theta|D) \propto \mathscr{L}(D|\theta)P(\theta)$$

MCMC sampling
for PLACK/WMAP
data

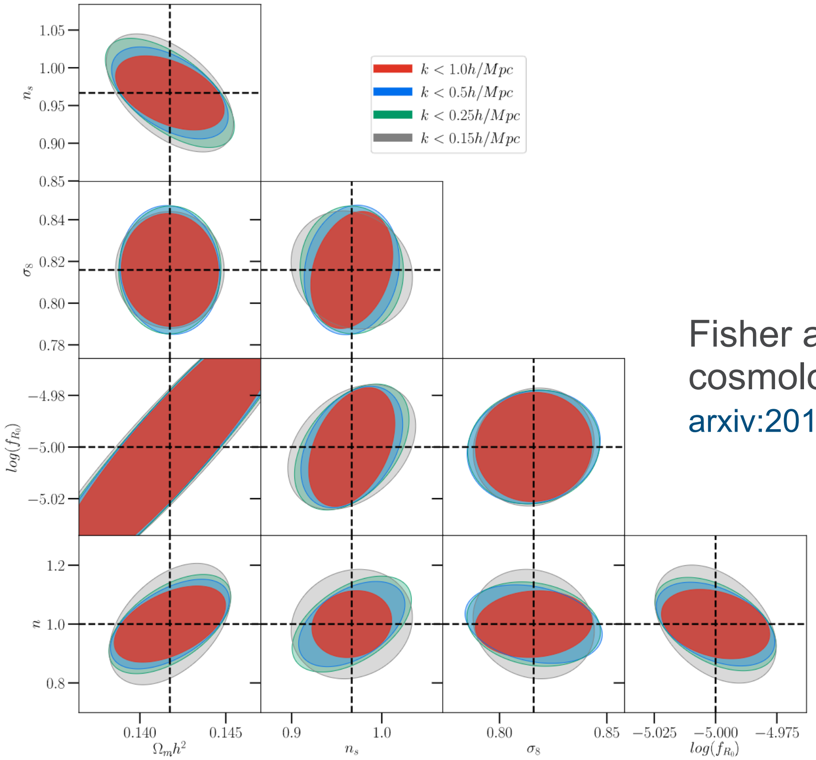


NR, Mickael
Binois et al

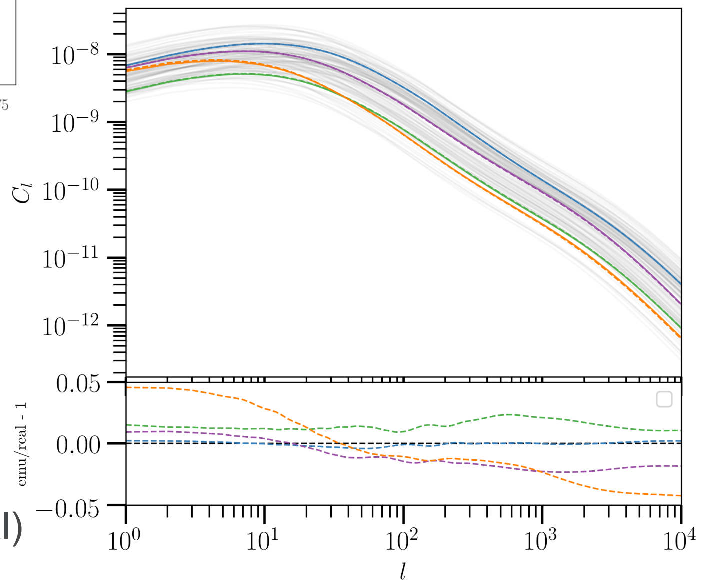
SUITE OF EMULATORS!

Emulators created for

- Dark matter power spectrum
- Dark energy evolution reconstruction from supernovae data,
- Halo mass function,
- Modified gravity observables,
- Weak lensing observables,
- CMB power spectra etc.



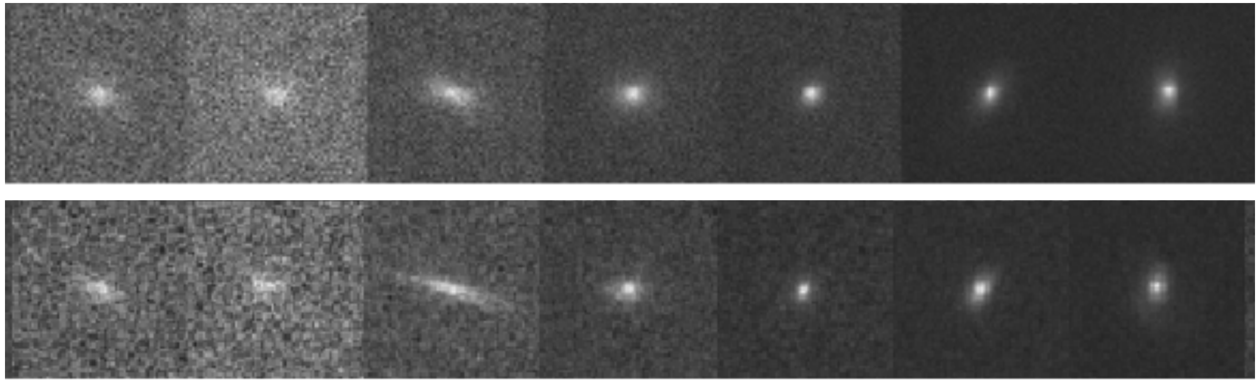
Fisher analysis for beyond General Relativity cosmologies (NR, Georgios Valogiannis et al: arxiv:2010.00596)



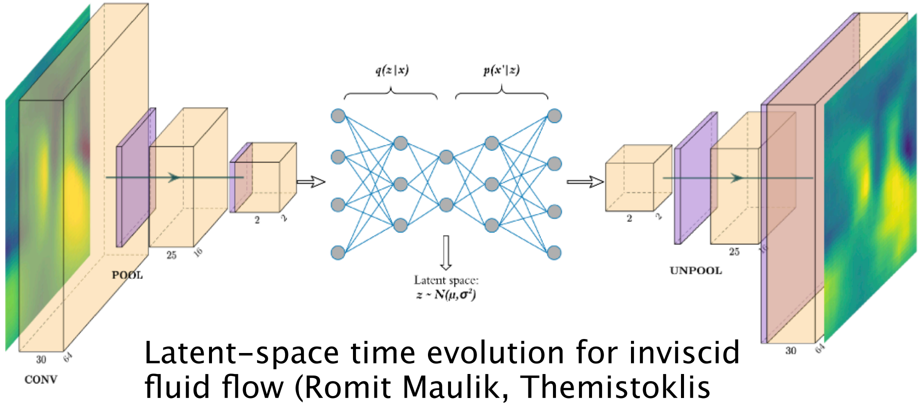
Weak lensing shear power spectra emulation (**NR**, Patricia Larsen et al)

SUITE OF EMULATORS!

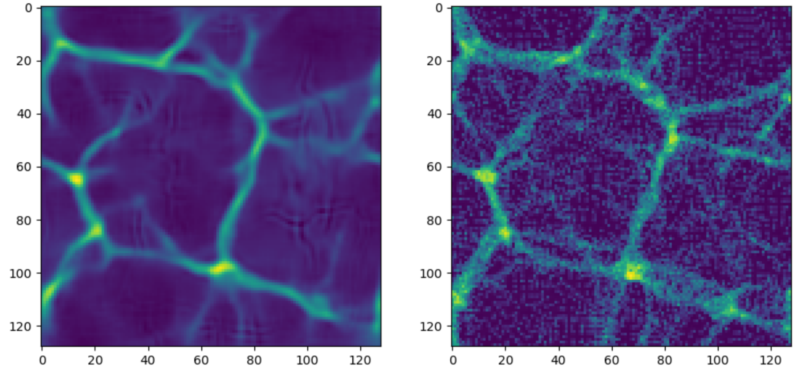
- Deep learning, especially convolution operation enables feature-important extraction and non-linear compressions



Galaxy image emulation (Claire Guilloteau, **NR** et al)



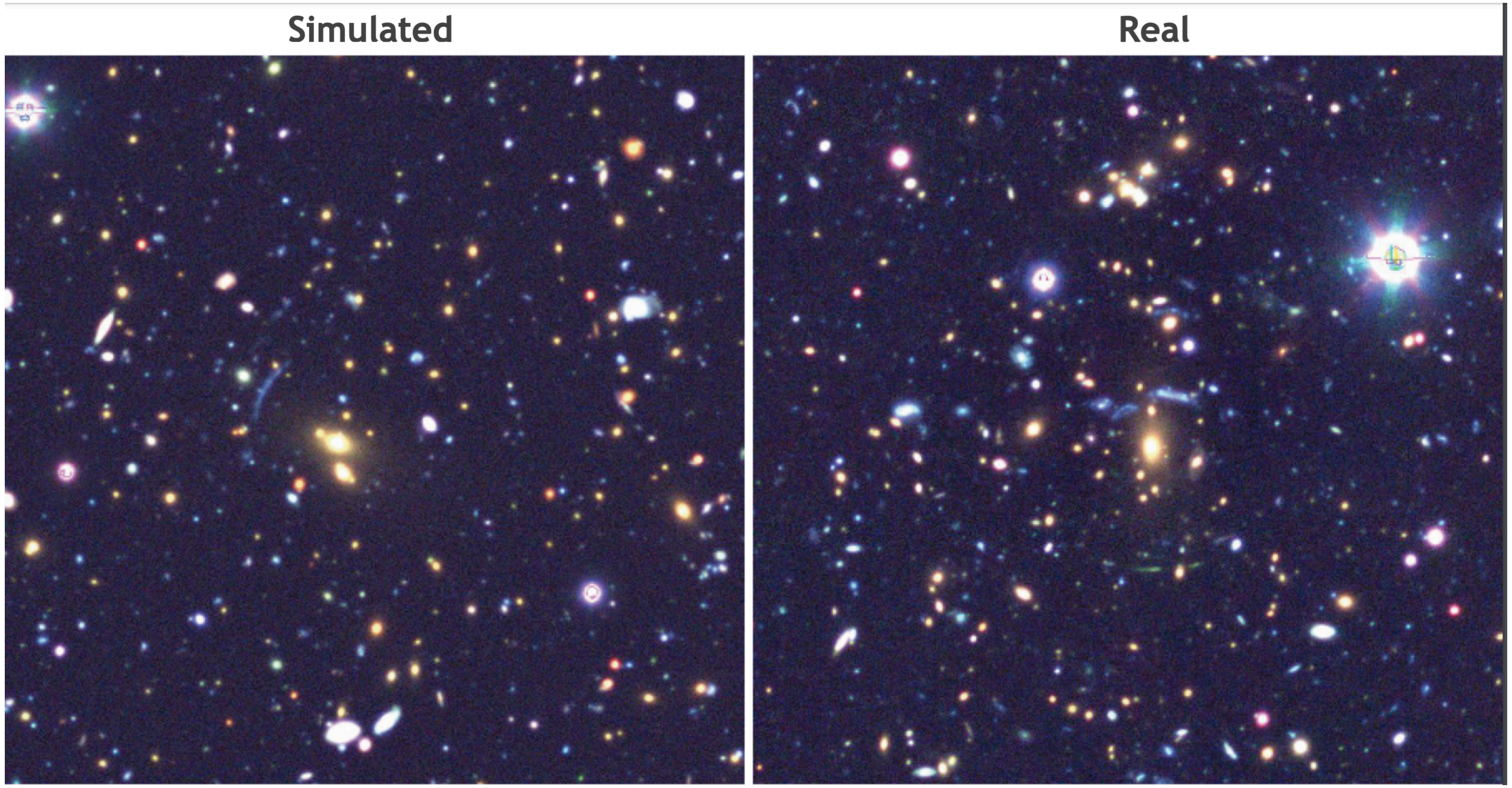
Latent-space time evolution for inviscid fluid flow (Romit Maulik, Themistoklis Botsas, **NR** et al: arxiv:2007.12167)



3D cosmic density field reconstruction (Xiaofeng Dong, **NR** et al)

WHY USE SYNTHETIC DATA?

- Tractable fundamental physics principles may help in synthetic data generation.

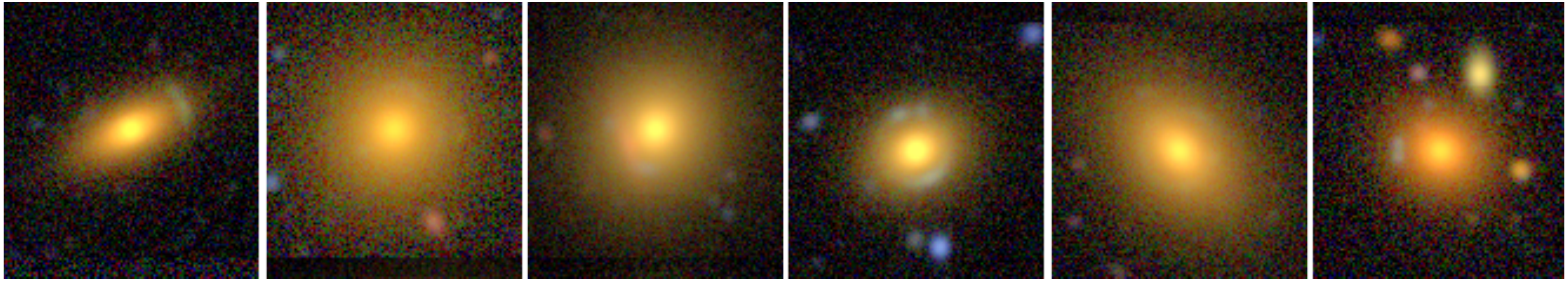


Simulated strong lens image to match SPT cluster observations taken with the MegaCAM camera on Magellan, in collaboration L. Bleem, M. Florian, S. Habib, M. Gladders, N. Li, S. Rangel N.
Li et al., arxiv:1511.03673

SYNTHETIC DATA FOR STRONG LENSING ANALYSIS

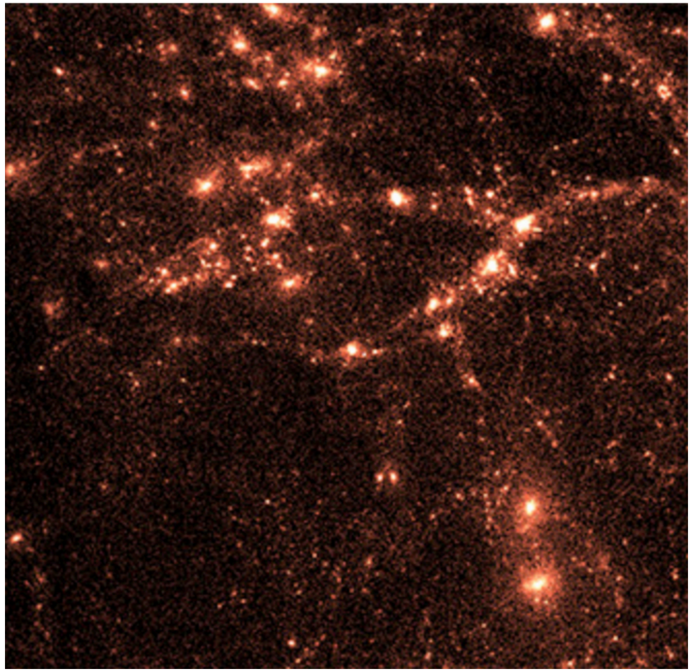
Motivation:

- Discrepancy with current amount of observed data vs future data
- Observed data is/will be a highly imbalanced dataset
- Relative ease of modeling with physical toy models



Credit: Nan Li. Strong Lenses created with the line of sight galaxies

GALAXY-SCALE STRONG LENSING CATALOG



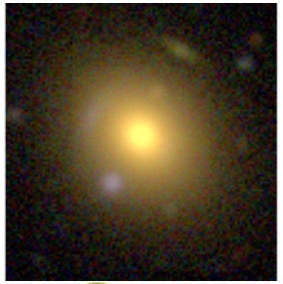
Outer-rim simulation

Galaxy modeling,
Ray tracing,
lensing pipeline

CosmoDC2
synthetic sky catalog
(arXiv:1907.06530)

Central galaxies and
high redshift sources
from cosmoDC2

Mass model for lens galaxy
(Singular Isothermal Ellipsoid,
Collett 2015), PSF and noise

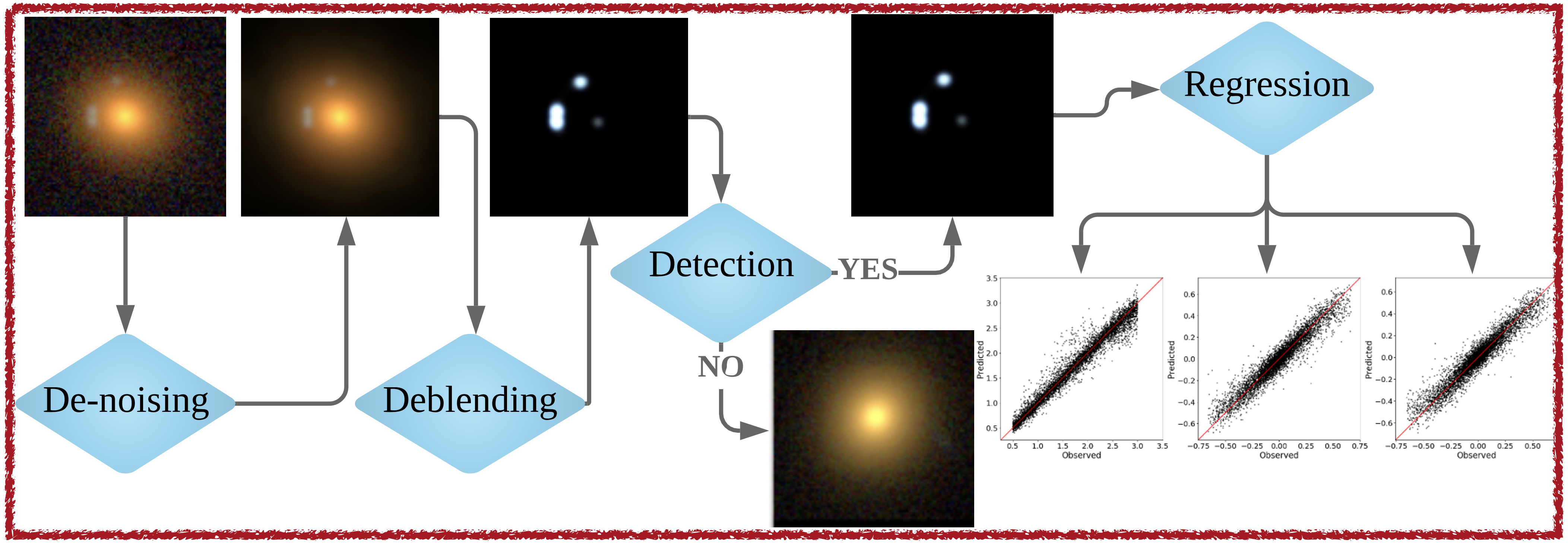


LSST DESC Dark Energy Science Collaboration

Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY Office of Science

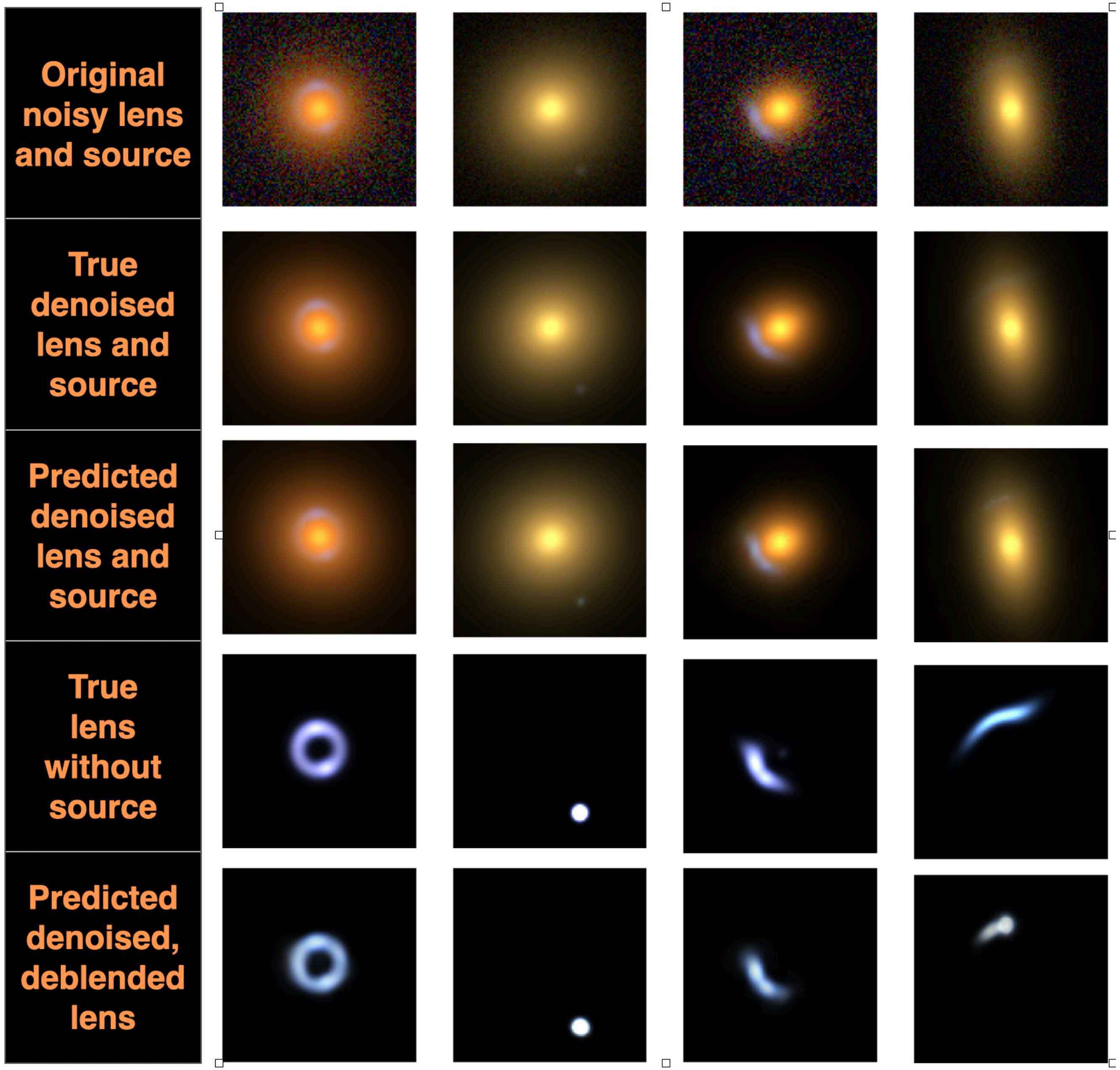
INTERPRETABLE LEARNING PIPELINES



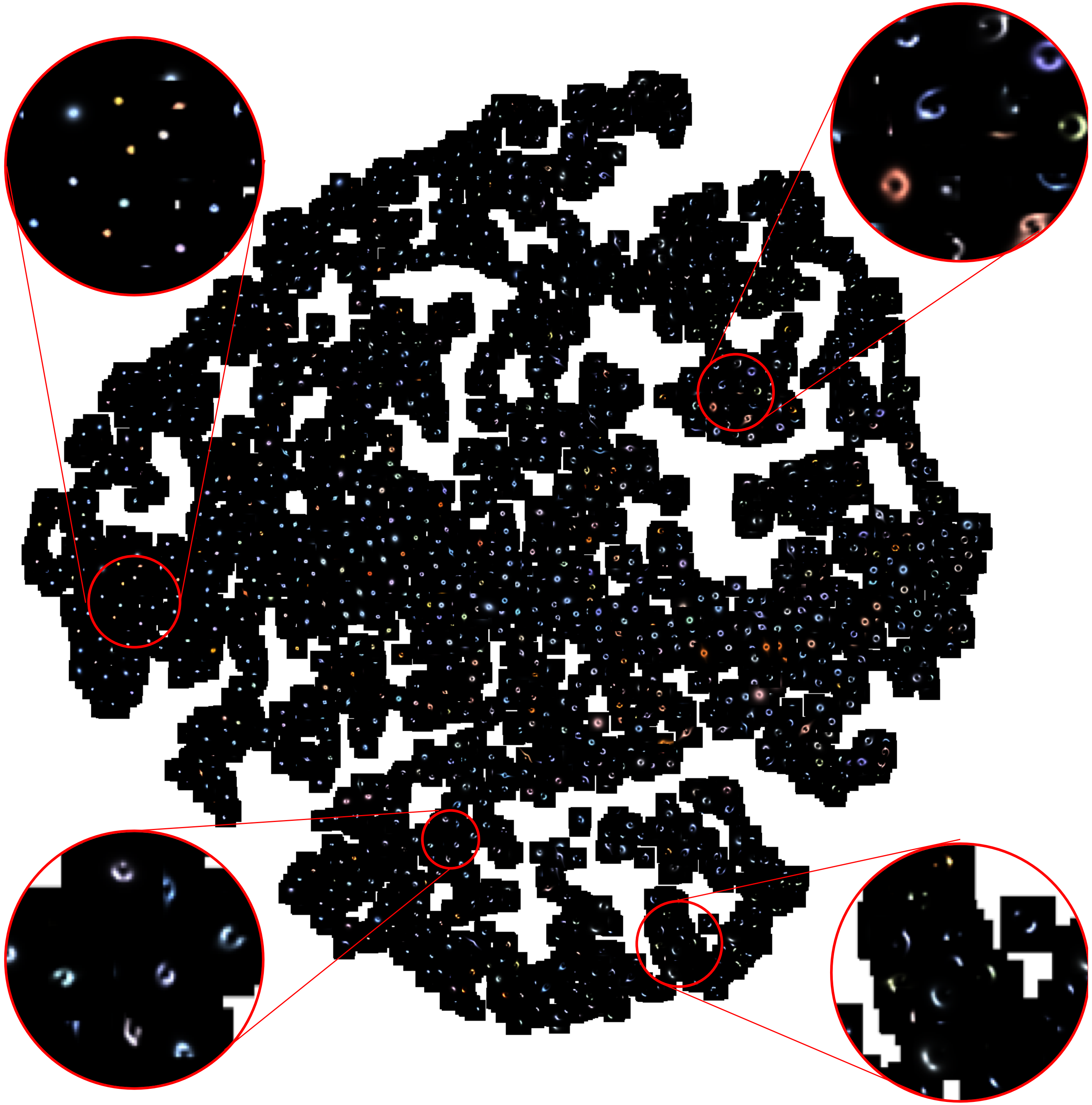
Added bonus:

- Synthetic data allows one to train modular pipelines that enable better control over systematics than end-to-end training methods
- Increase in classification and regression accuracy

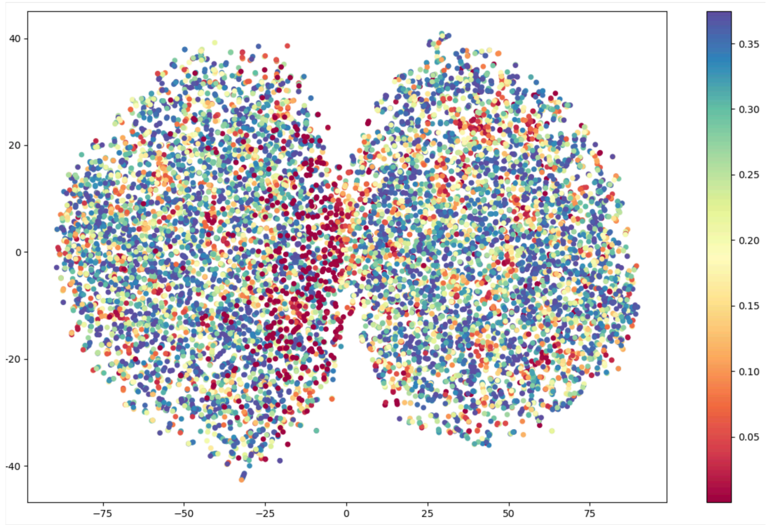
INTERPRETABLE STRONG LENS END-TO-END ANALYSIS PIPELINE



Sandeep Madireddy, Nan Li, **NR** et al: arxiv.org:1911.03867

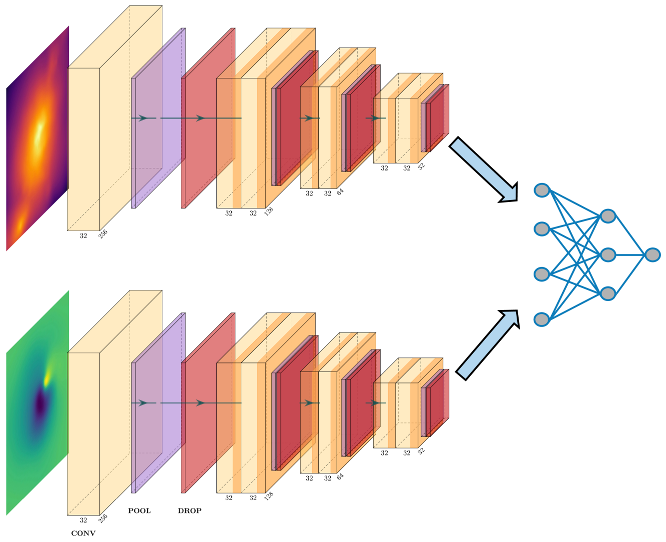


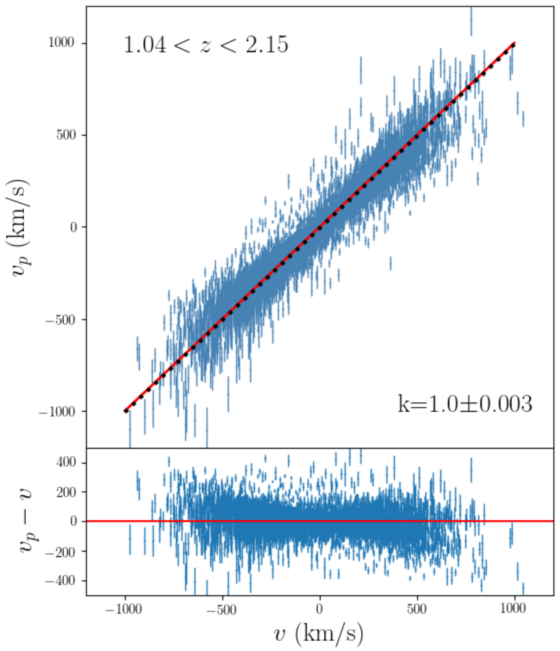
Variational Information Bottleneck and representation learning



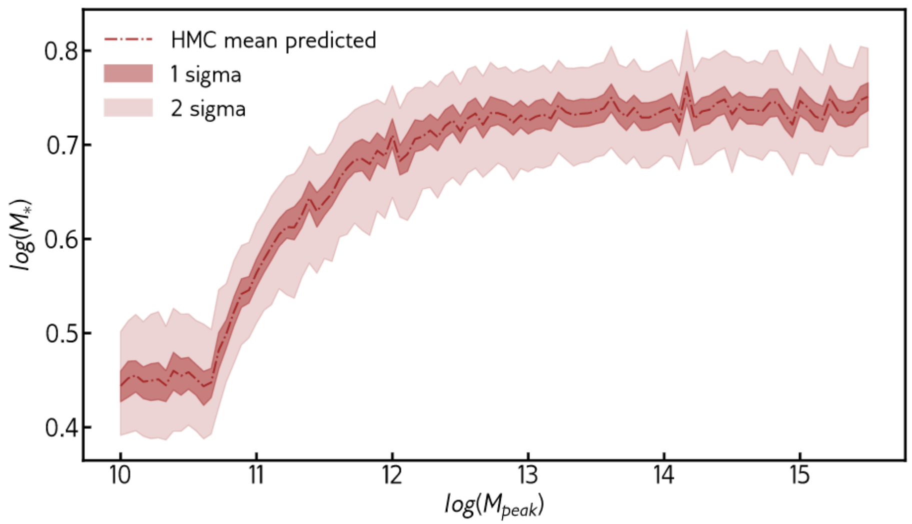
Uncertainty quantification for classification

SEVERAL APPROACHES TO UQ IN ML

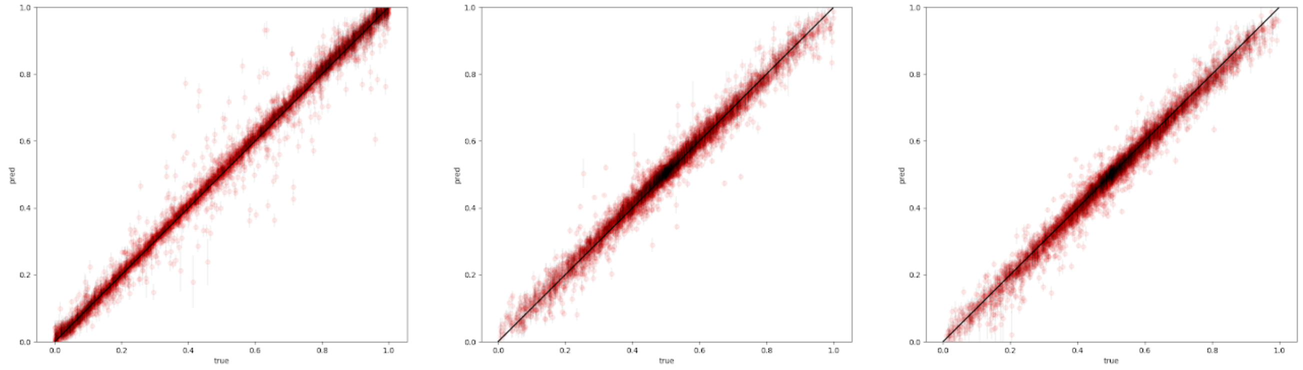




Monte-Carlo Dropout uncertainty quantification for galaxy peculiar velocity estimation (Yuyu Wang, **NR** et al
arxiv.org:2010.03762



Hamiltonian Monte Carlo sampling for weights of Neural Networks (Andrew Hearin, **NR** et al)



Variational Inference for Einstein radius, axis ratio, position angle for Strong Lensing problem (Sandeep Madireddy, Nan Li, **NR,** James Butter et al)

SYNTHETIC TRAINING IN PHOTOMETRIC REDSHIFT ESTIMATION

Motivation:

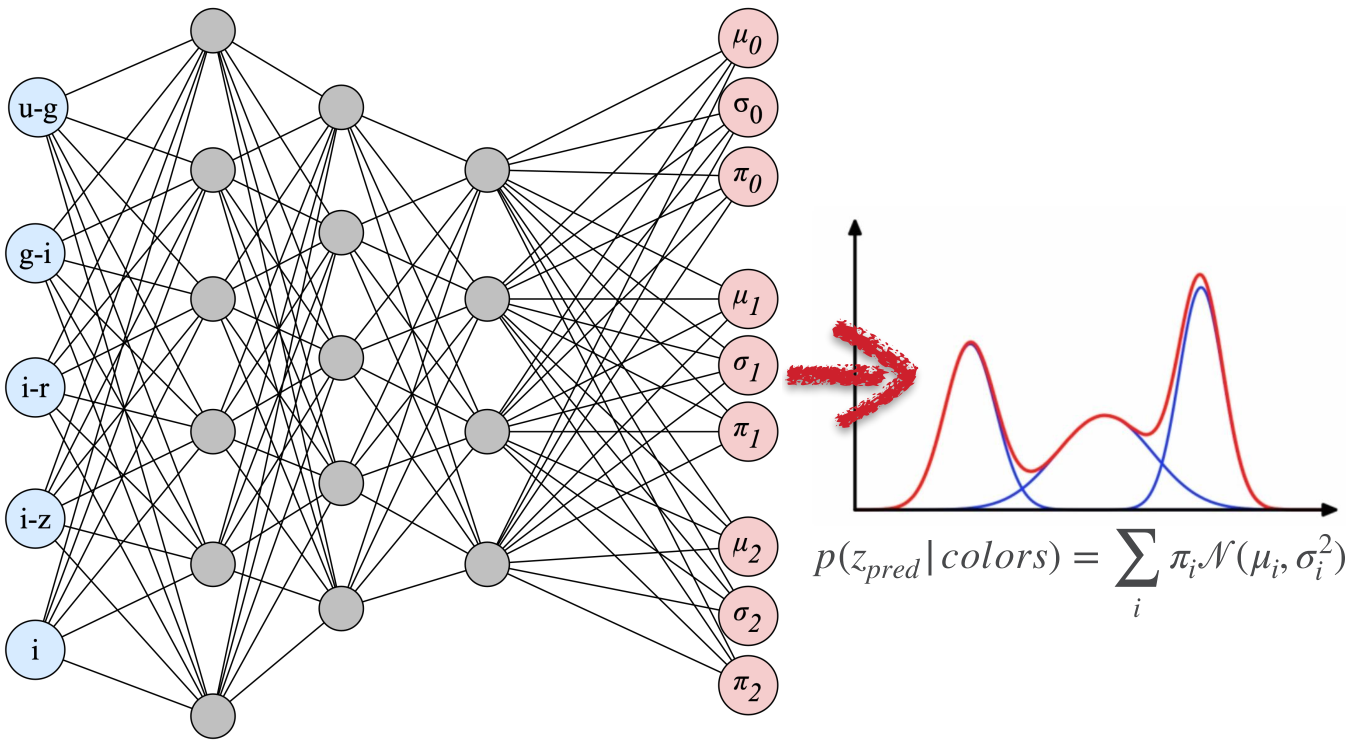
Real data maybe biased, gaps in color space, and fewer high-z galaxies.

- Development of robust generative modeling tool GALAXPY for emulating SEDs using a Gaussian Processes
  - Capture effects of star formation histories, metallicities, initial mass functions, dust attenuations and emission line ratios.

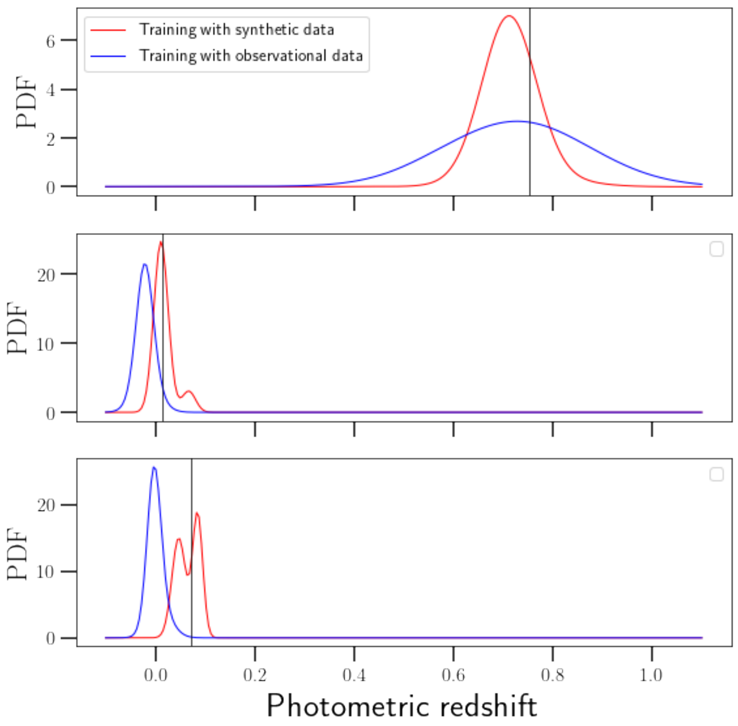
Added bonus:
Ability to create large amount of training samples, with uncertainties in the sample.

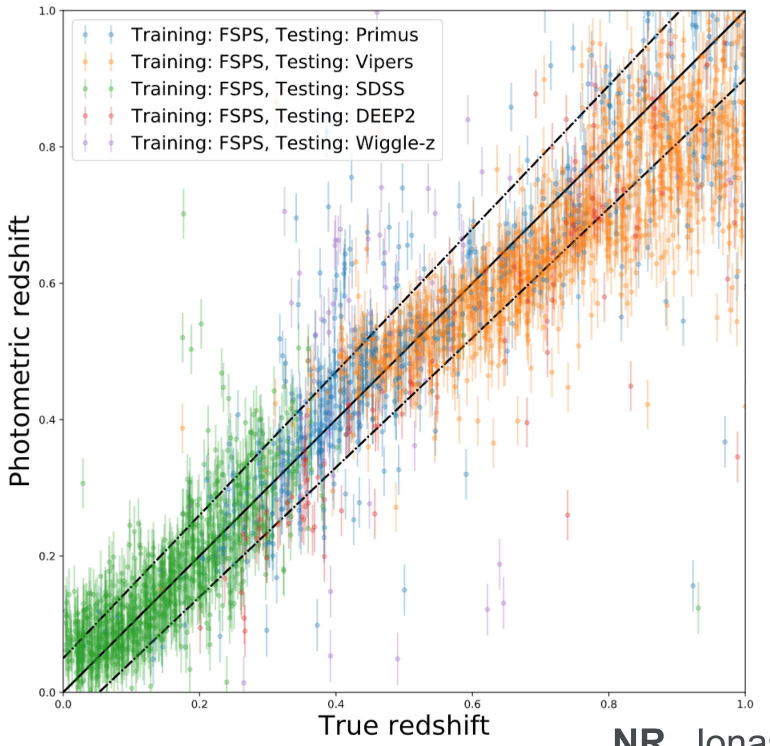
BAYESIAN NEURAL NETWORKS: APPLICATION IN PHOTOMETRIC REDSHIFT ESTIMATION

- Mixed Density Network for mapping LSST-like color magnitudes to redshifts
  - Allows for Uncertainty quantification in photo-z estimates
  - Allows for degeneracy in the data using Gaussian Mixture models
  - For comparison, training done with observed data and synthetic data (large number of training samples)



PHOTOMETRIC REDSHIFT ESTIMATION: OBSERVED AND SYNTHETIC TRAINING



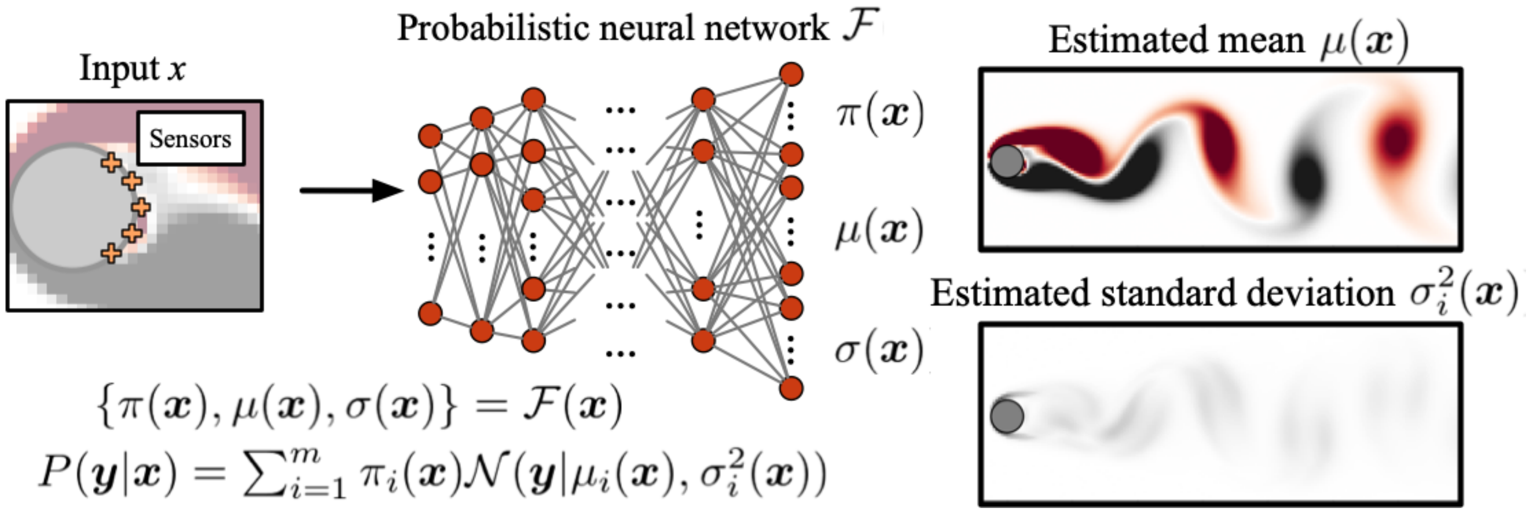


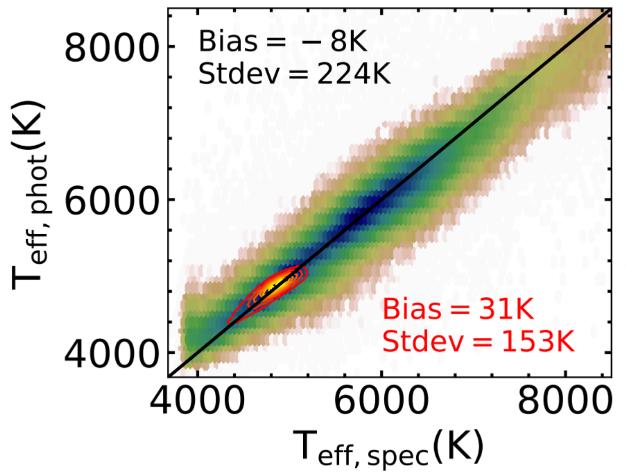
NR, Jonas Chaves-Montero, Arindam Fadikar et al

Photo-z estimates for SDSS galaxies. The synthetic training results in fewer prediction outliers compared to the SDSS-trained model. Fewer data in larger z: error bars are larger, predictions are worse.

DATA RECOVERY USING PROBABILISTIC NEURAL NETWORKS

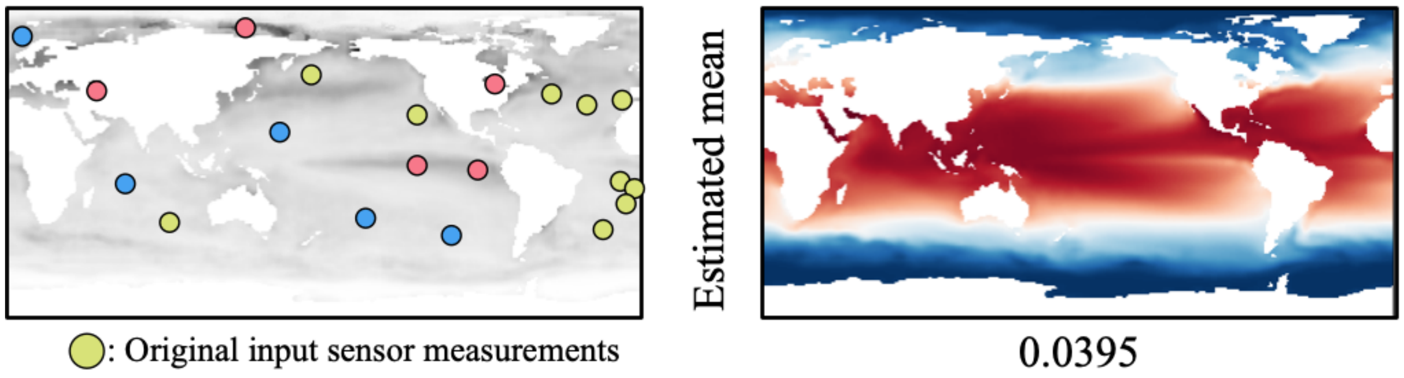
Fluid flow models





Madeline Lucey, Yuan-Sen Ting, **NR,** Keith Hawkins, arxiv:2002.02961

Extracting a pristine sample of red clump stars in the Milky Way



Sea surface temperature

Romit Maulik, Kai Fukami, **NR** et al arxiv:2005.04271

CONCLUSIONS

Argonne
NATIONAL LABORATORY

- Synthetic datasets are sometimes a necessity (cosmological simulations), sometimes a convenience (photometric data analysis)

- Careful experimental design, robust data creation, extensive validations are all required while dealing with synthetic data.

- Interpretable, Uncertainty quantified models are still very important, probably even more so while using synthetic data in training.