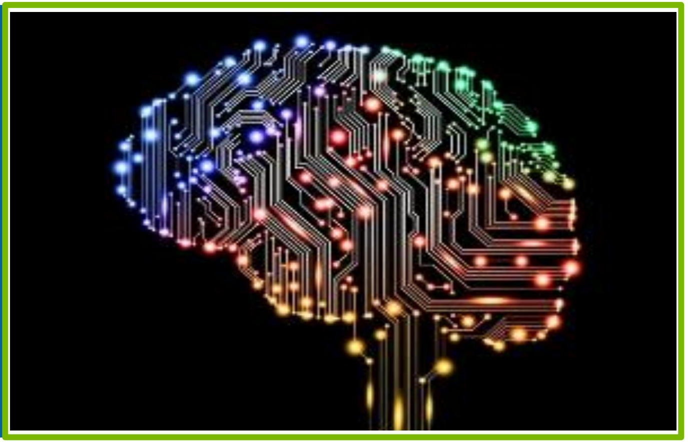
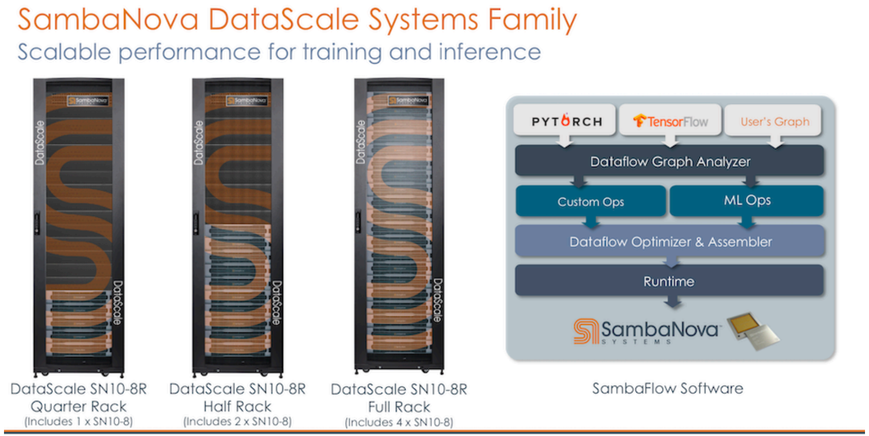
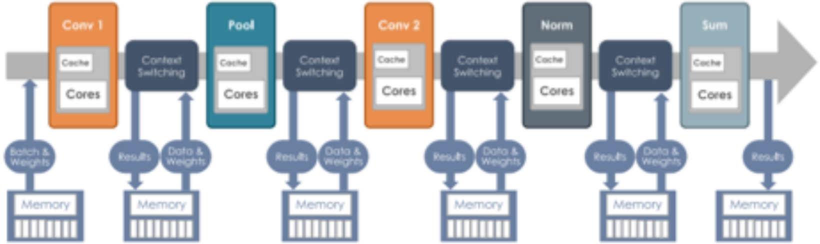
Accelerating Deep Learning for Science with SambaNova



PRASANNA BALAPRAKASH
Computer Scientist
Mathematics and Computer Science Division &
Leadership Computing Facility
Argonne National Laboratory

The 2nd International Workshop on Machine Learning Hardware, ISC 2021





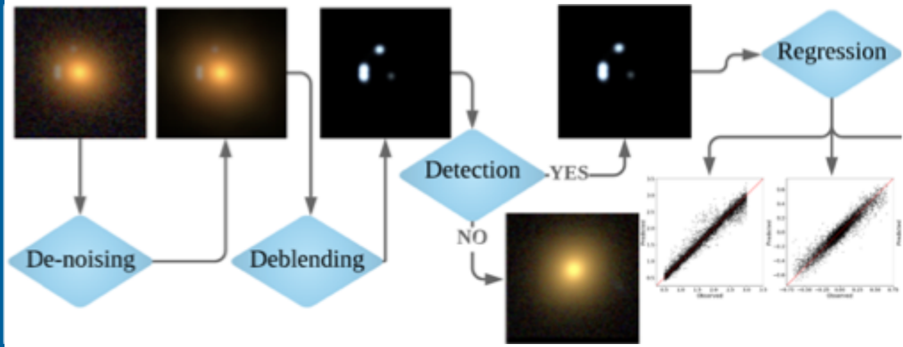
Traditional GPU accelerator

SambaNova dataflow execution

Accelerated Computing with a Reconfigurable Dataflow Architecture, White Paper, 2021
https://sambanova.ai/wp-content/uploads/2021/06/SambaNova_RDA_Whitepaper_English.pdf

A MODULAR DEEP LEARNING PIPELINE FOR GALAXY-SCALE STRONG GRAVITATIONAL LENS DETECTION AND MODELING

https://arxiv.org/abs/1911.03867



SANDEEP MADIREDDY
Assistant Computer Scientist
Mathematics and Computer Science Division

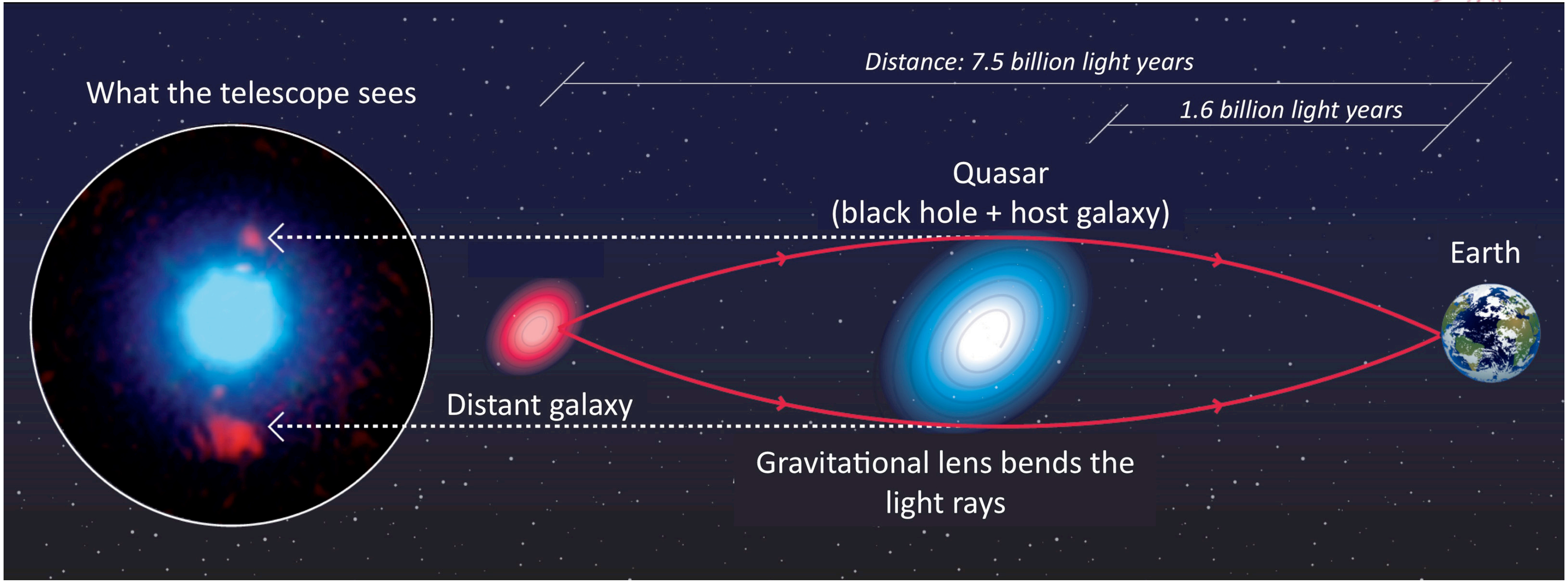
http://www.mcs.anl.gov/~smadireddy/

JOINT WORK WITH

Prasanna Balaprakash[1], Salman Habib[2,3], Katrin Heitmann[3], Nan Li[4], Nesar Ramachandra[3], James Butler[3]

[1]Mathematics and Computer Science (MCS),
[2]Computational Science (CPS),
[3]High Energy Physics Division(HEP),
[4]National Astronomical Observatories of China

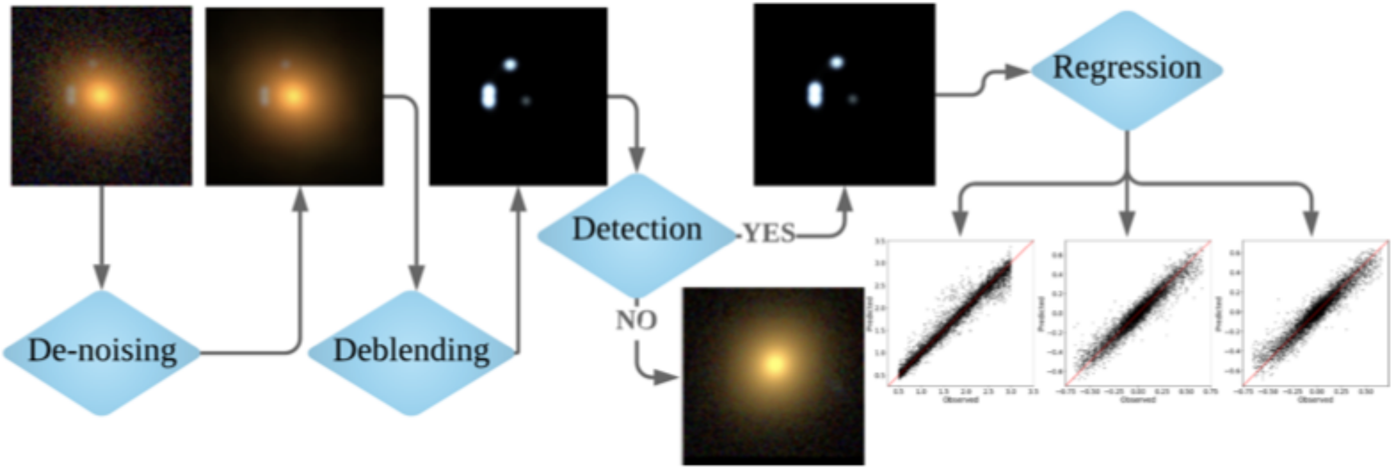
GALAXY-GALAXY STRONG LENSING



- Gravitational lensing: phenomenon by which light rays are deflected as they traverse through curved space caused by the presence of massive astrophysical objects.
- Galaxy-galaxy strong lensing (GGSL) - background source and foreground lens are both galaxies

DEEP LEARNING PIPELINE

Training

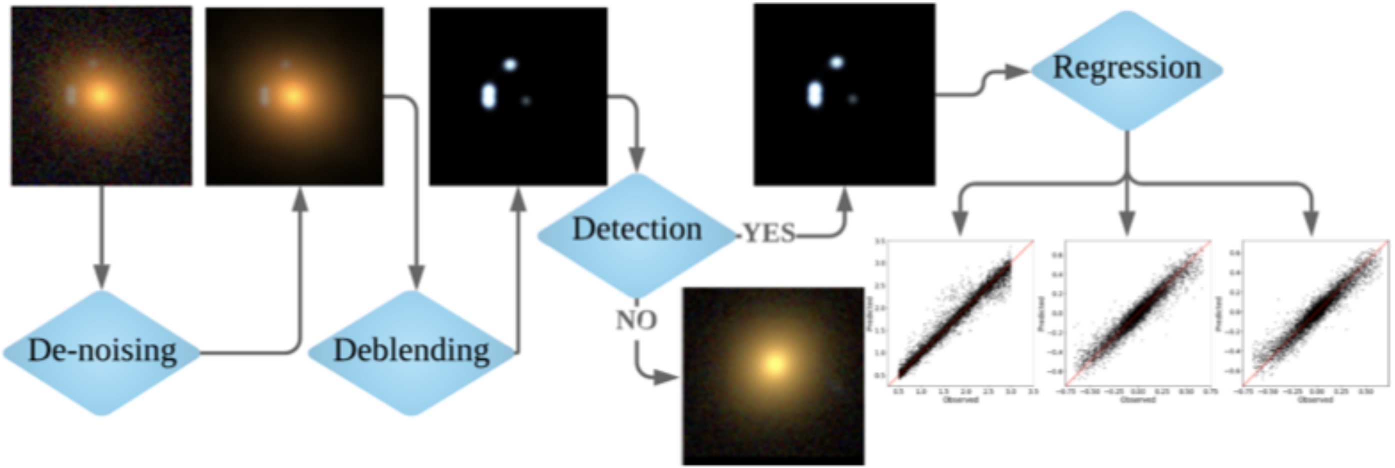


Denoising is an image restoration approach used to recover a clean image from a noisy observation

- No prior assumption on the noise form is required

DEEP LEARNING PIPELINE

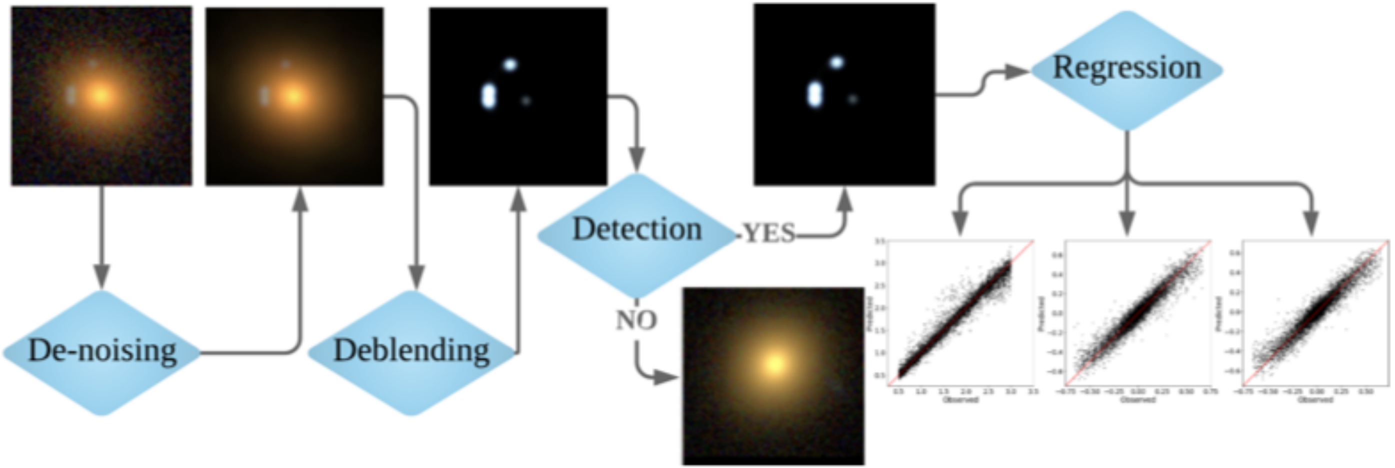
Training



Deblending refers to decoupling the lensed light and the source galaxy from the observations.

DEEP LEARNING PIPELINE

Training

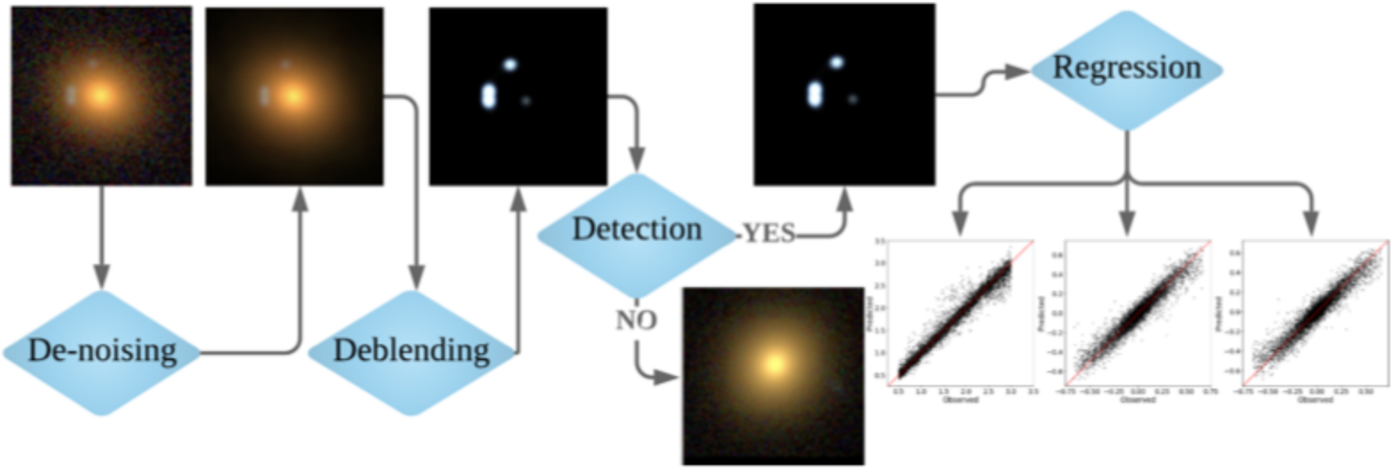


Lens Detection/Finding refers to classifying the lensed and non-lensed systems from the source separated images

DEEP LEARNING PIPELINE

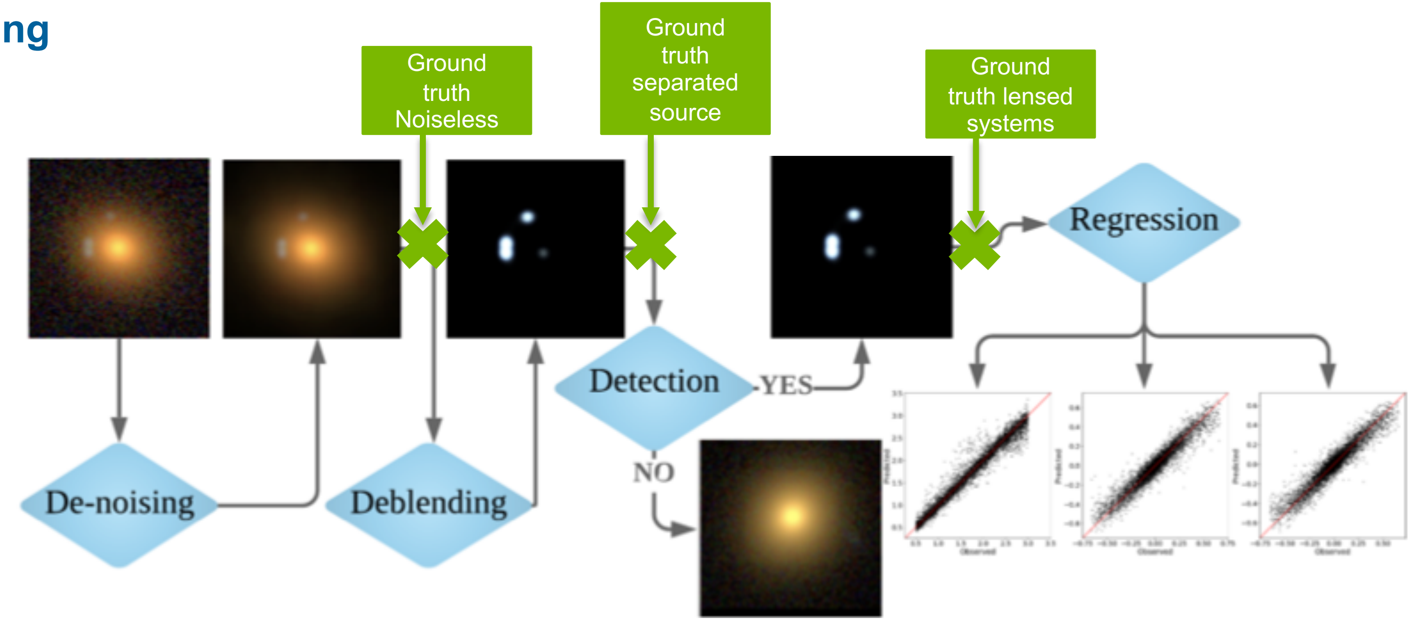
Training

Lens modeling is a regression module that takes the source separated lensed galaxies and predicts its characteristics: [Einstein Radius, Axis Ratio and Position Angle].



DEEP LEARNING PIPELINE

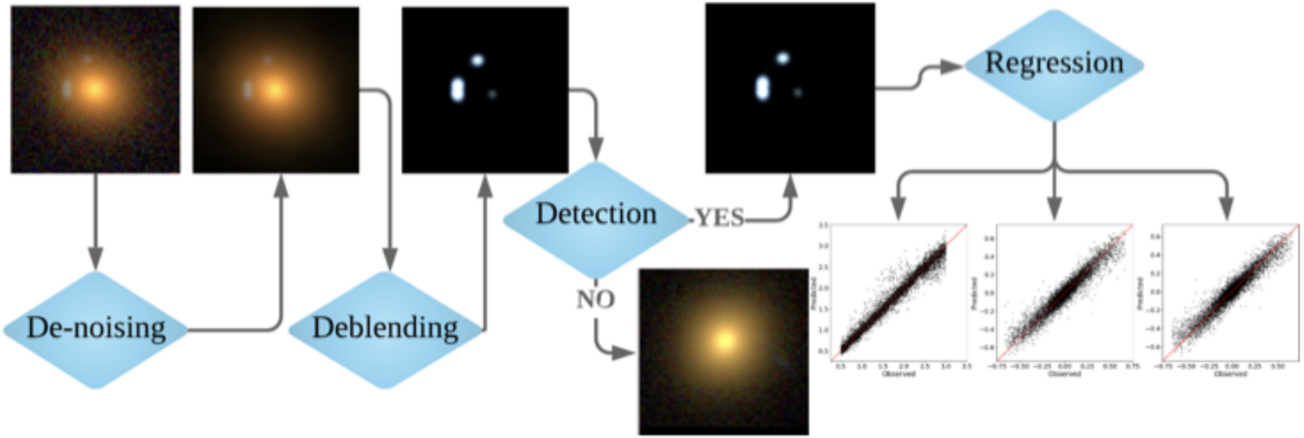
Training



- The denoising and the deblending modules essentially preprocess the images in the pipeline to enhance the lens searching and modeling tasks further down the pipeline.
- Each of the four modules are trained with corresponding simulated data as the training set
    - Ex: Denoising model is trained to output noiseless images generated from simulation with the corresponding noisy counterparts as input

DEEP LEARNING PIPELINE

Inference



- All the trained models weights are frozen
- Only the noisy blended image is fed to the pipeline
  - All the subsequent steps – Denoising, Deblending, Lens Detection and Finding are done sequentially

DENOISING AND DEBLENDING PIPELINE ON SAMBANOVA

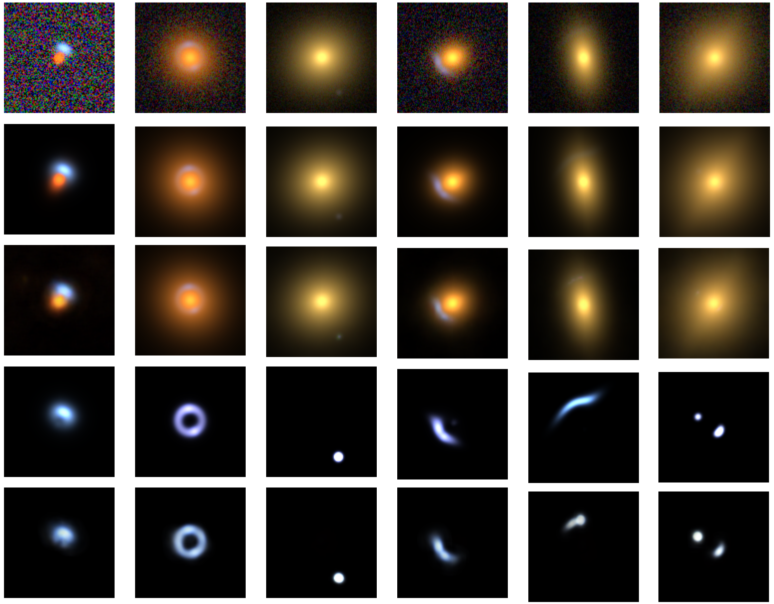
Noisy blended (S_1)

Noiseless blended (S_2)

Output from denoising module (T_2)

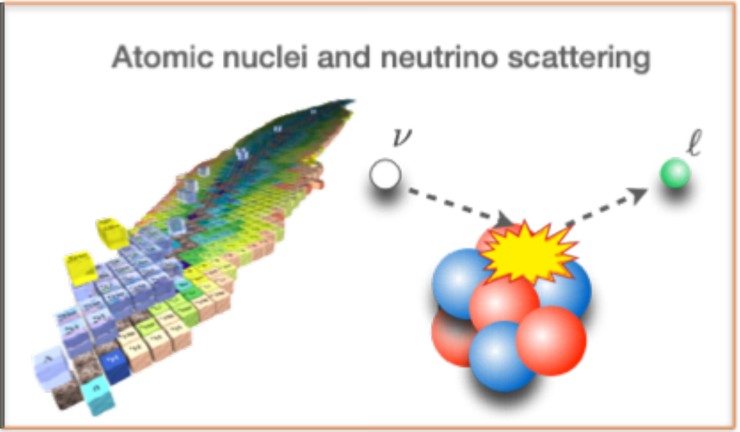
Noiseless deblended (S_3)

Output from deblending module (T_3)



MACHINE LEARNING-BASED INVERSION OF NUCLEAR RESPONSES

https://journals.aps.org/prc/abstract/10.1103/PhysRevC.103.035502



KRISHNAN RAGHAVAN
Mathematics and Computer Science Division

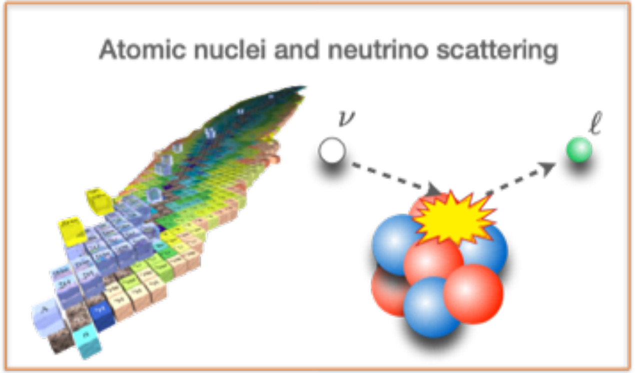
https://www.anl.gov/profile/krishnan-raghavan

JOINT WORK WITH

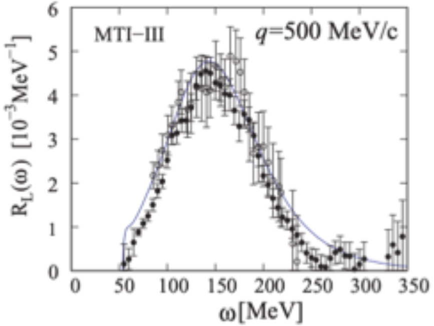
Prasanna Balaprakash[1], Alessandro Lovato[2], Noemi Rocco[2] and Stefan Wild[1]

[1]Mathematics and Computer Science (MCS),
[2]Physics Division (PHY)

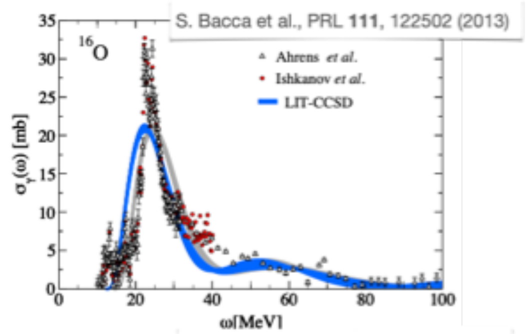
SciDAC Scientific Discovery through Advanced Computing

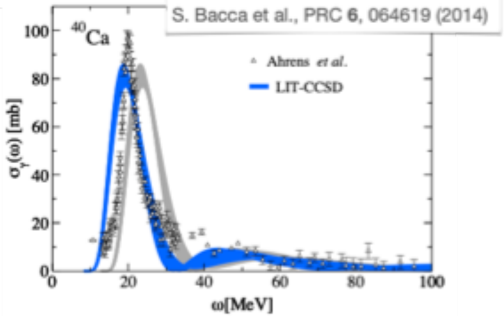


The response functions contain all information on the structure and dynamics of the target



Measured by electron-scattering experiments



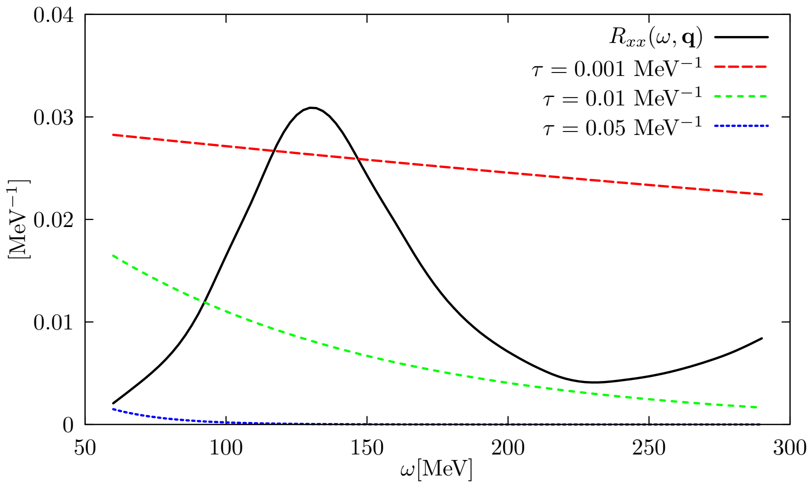


$$R_{\alpha\beta}(\omega,\mathbf{q}) = \sum_f \langle\Psi_0|J_\alpha^\dagger(\mathbf{q})|\Psi_f\rangle\langle\Psi_f|J_\beta(\mathbf{q})|\Psi_0\rangle\delta(\omega - E_f + E_0)$$

LAPLACE TRANSFORM

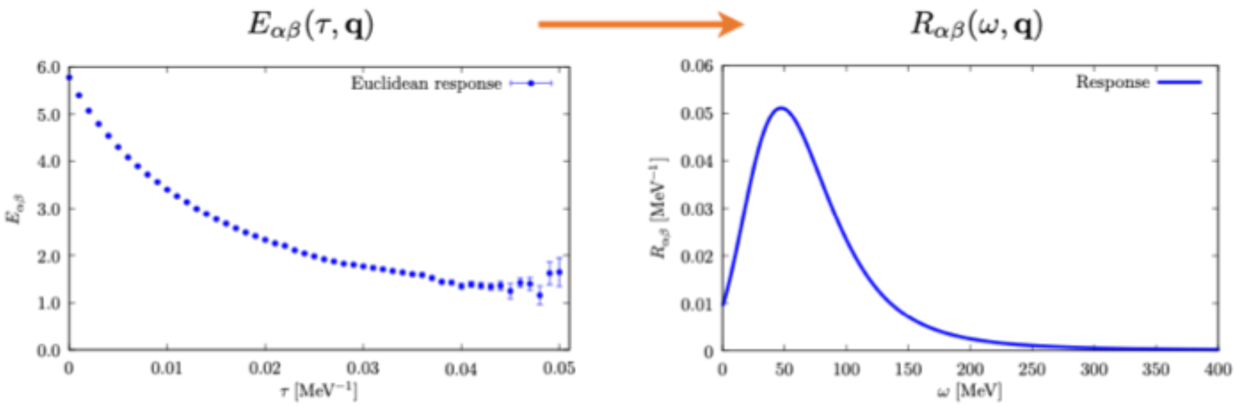
Valuable information on the energy dependence of the response functions can be inferred from their **Laplace transforms**

The system is first heated up by the transition operator.
Its cooling determines the **Euclidean response** of the system



$$\underbrace{E_{\alpha\beta}(\tau,\mathbf{q})}_{\text{Euclidean Response}} \equiv \int d\omega e^{-\omega\tau} \underbrace{R_{\alpha\beta}(\omega,\mathbf{q})}_{\text{Response Function}}$$

LAPLACE TRANSFORM

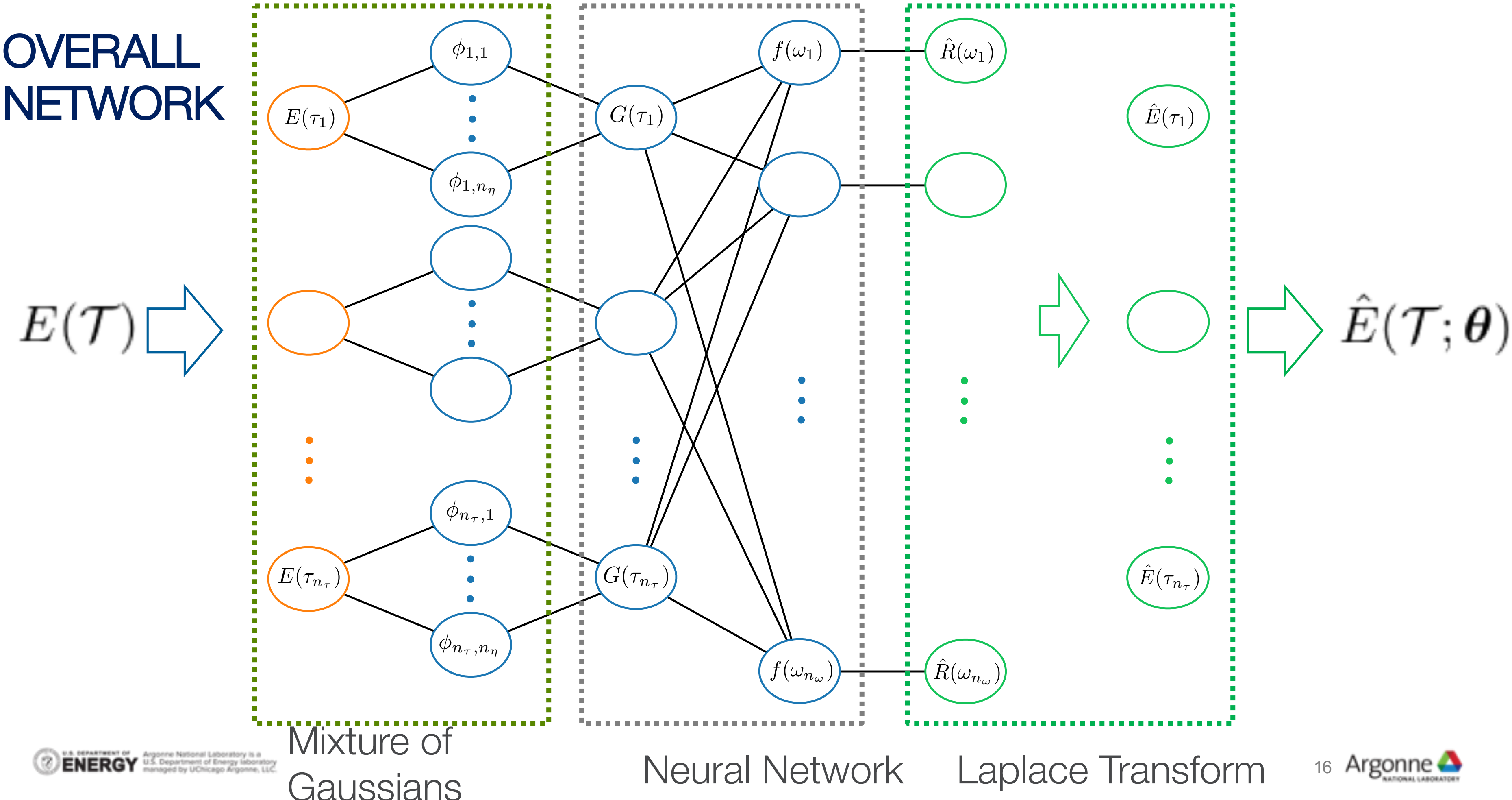


$$R(\Omega) = K(\Omega, \mathcal{T})^{-1} E(\mathcal{T}).$$

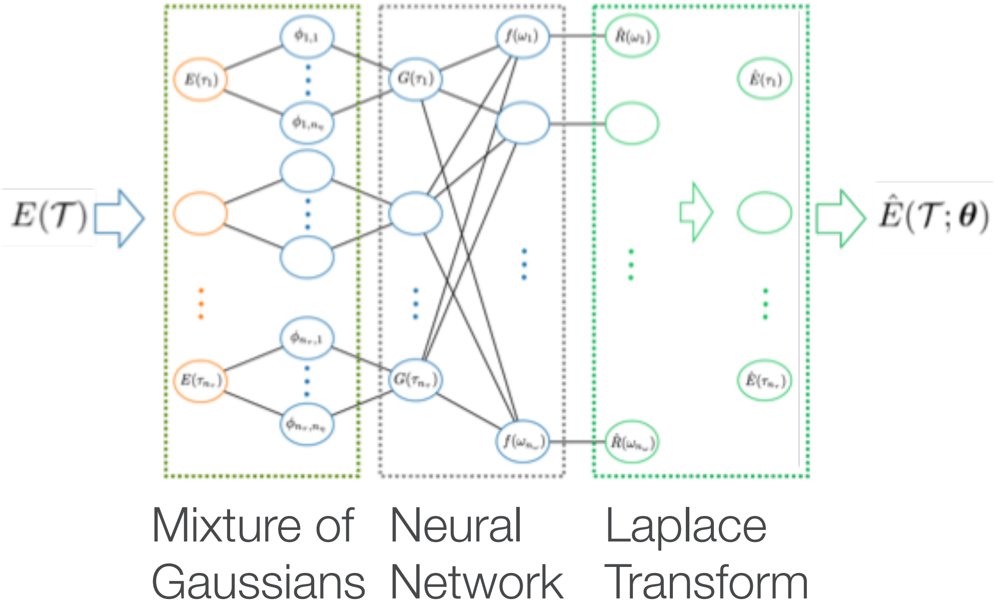
- The inverse is ill-posed
  - Multiple response functions can have the same Euclidean response (within errors)
- High noise in the Euclidean response results in unstable inversions
- Response: smooth, positive, and Laplace integration

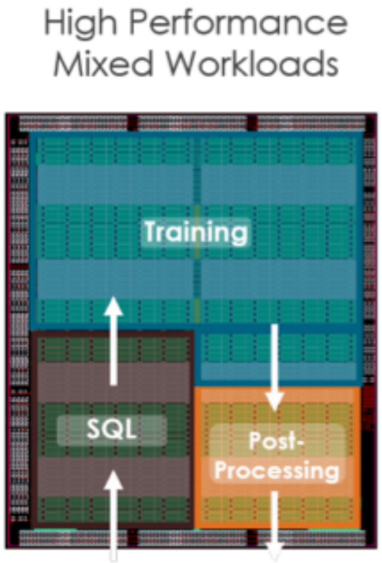
$$\hat{R}(\Omega; \boldsymbol{\theta}) = \frac{1}{\mathcal{N}_0} e^{f(E(\mathcal{T}); \boldsymbol{\theta})}$$

OVERALL NETWORK



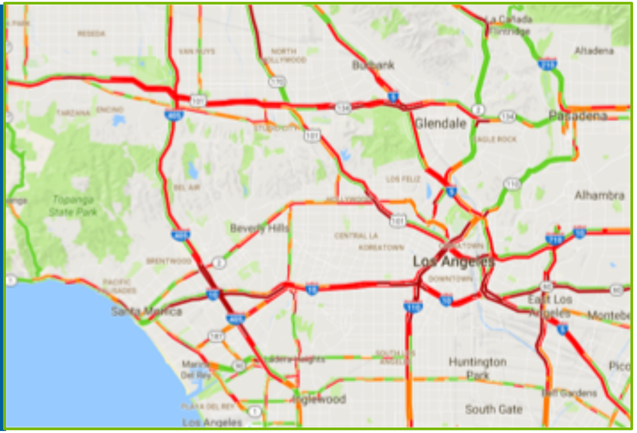
MAPPING INVERSION PIPELINE ON SAMBANOVA





ACCELERATING GRAPH CONVOLUTION BASED DEEP LEARNING FOR LARGE SCALE HIGHWAY TRAFFIC FORECASTING

https://journals.sagepub.com/doi/abs/10.1177/0361198120930010



TANWI MALLICK
Mathematics and Computer Science Division

https://www.anl.gov/profile/tanwi-mallick

JOINT WORK WITH
Prasanna Balaprakash[1] and Jane Mcfarlane[2]

[1]Mathematics and Computer Science (MCS)

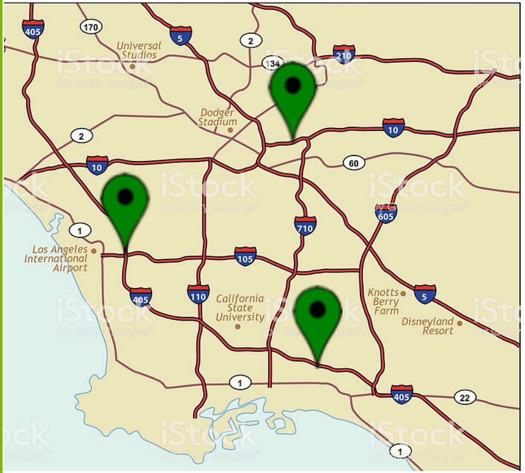
[2]Lawrence Berkeley Lab

U.S. DEPARTMENT OF
ENERGY

Energy Efficiency &
Renewable Energy

TRAFFIC FORECASTING

Road network



Loop detectors

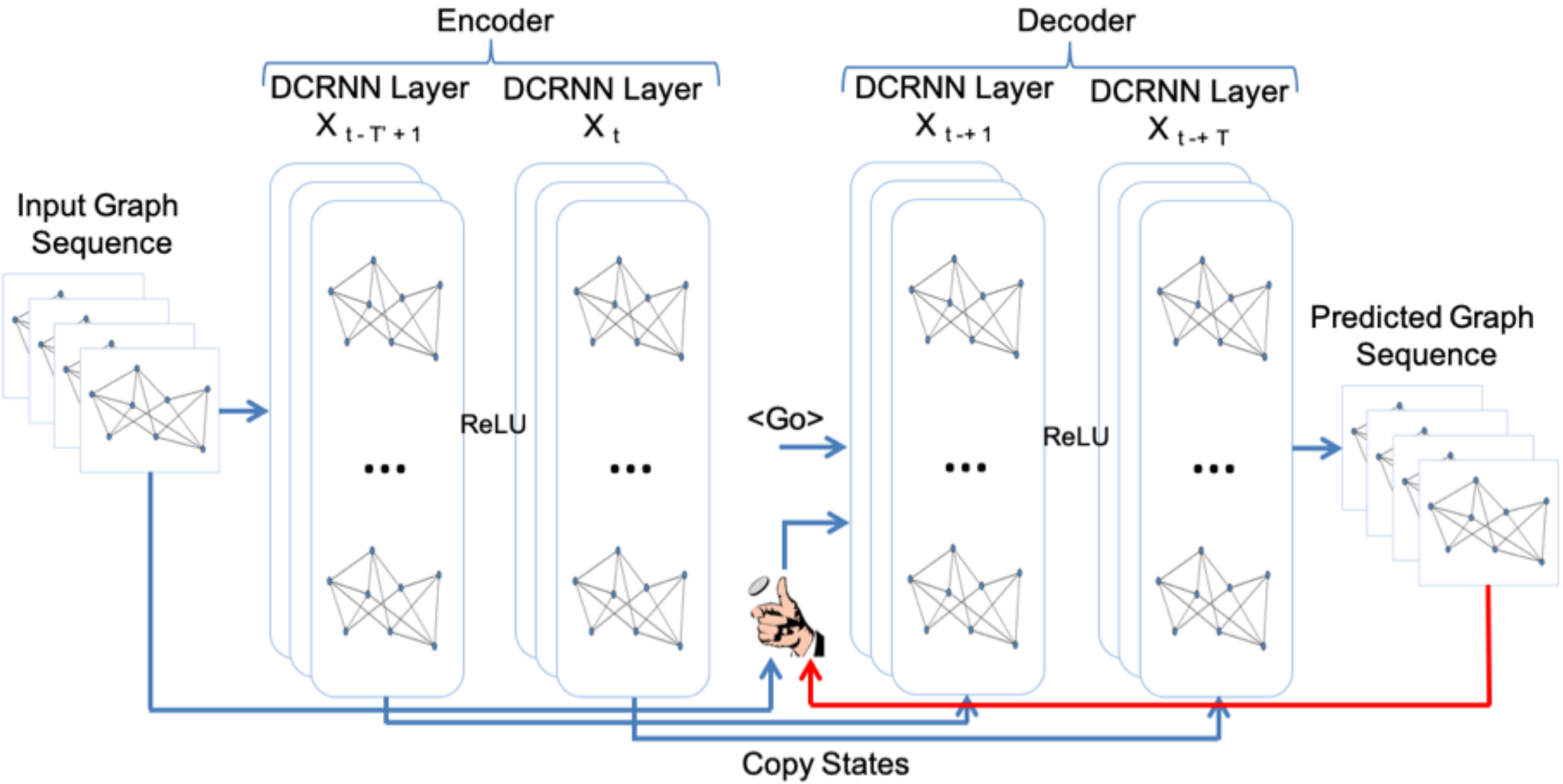
Historic traffic metrics

8.00 AM … 9.00AM

Future traffic metrics

9.00 AM … 10.00AM

TRAFFIC FORECASTING USING DCRNN



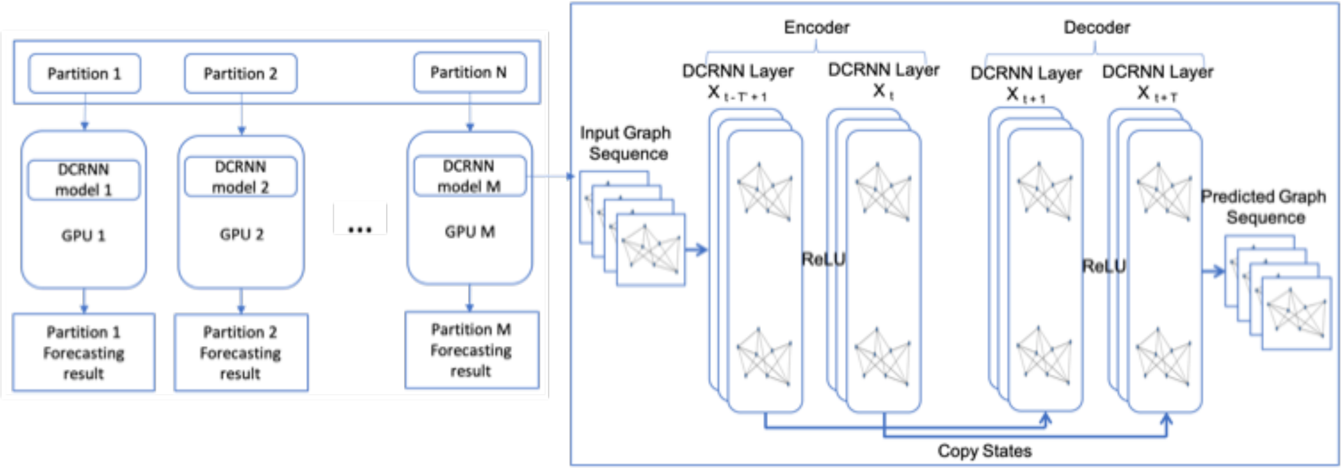
HANDING LARGE GRAPHS USING DCRNN

- Partition large graph into number of sub-graphs
- Run DCRNN for each sub-graph
- Combine the results and forecast traffic

Traffic of the north of the state is different from south

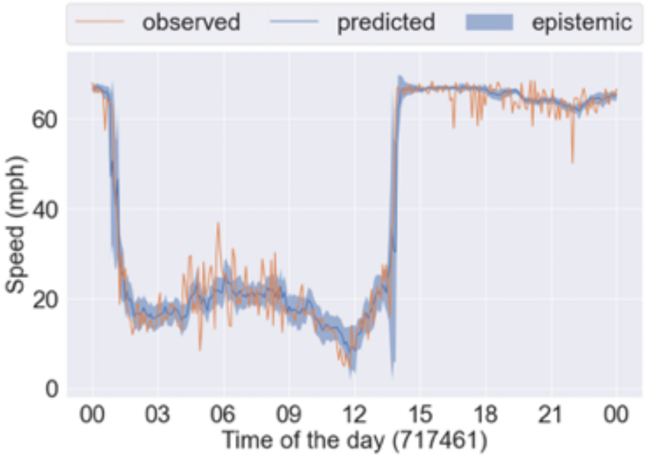
https://geology.com/state-map/california.shtml

GRAPH-PARTITIONING-BASED DCRNN

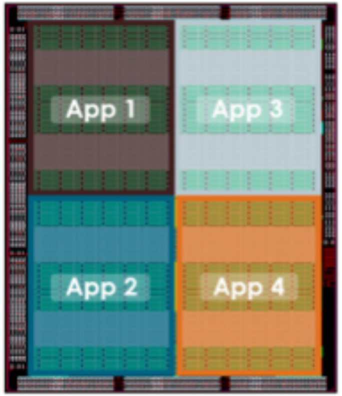


UNCERTAINTY ESTIMATION USING DEEP ENSEMBLE LEARNING ON SAMBANOVA

- Estimate the model uncertainty from M models running using different random seed



Concurrent Application Isolation



SUMMARY

- AI for science applications are complex
- Novel AI architectures can significantly accelerate AI for science
- Many different ways to configure and use
- Novel mapping strategies

- Not all AI for science applications require vision and language
- Support for custom models and experimentation
- Data movement will become a bottleneck
- Performance and power/energy tradeoffs
- Need for co-design and adaptation