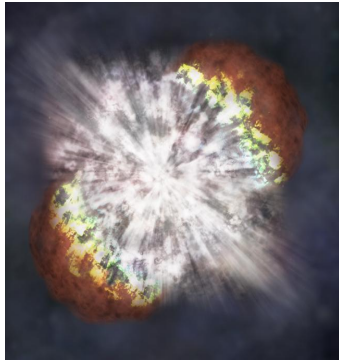
Machine Learning-based Trigger for DUNE

Guanqun Ge, Columbia University
on behalf of DUNE collaboration

CPAD INSTRUMENTATION FRONTIER WORKSHOP
University of Wisconsin-Madison, December 8, 2019



Credits: NASA

DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

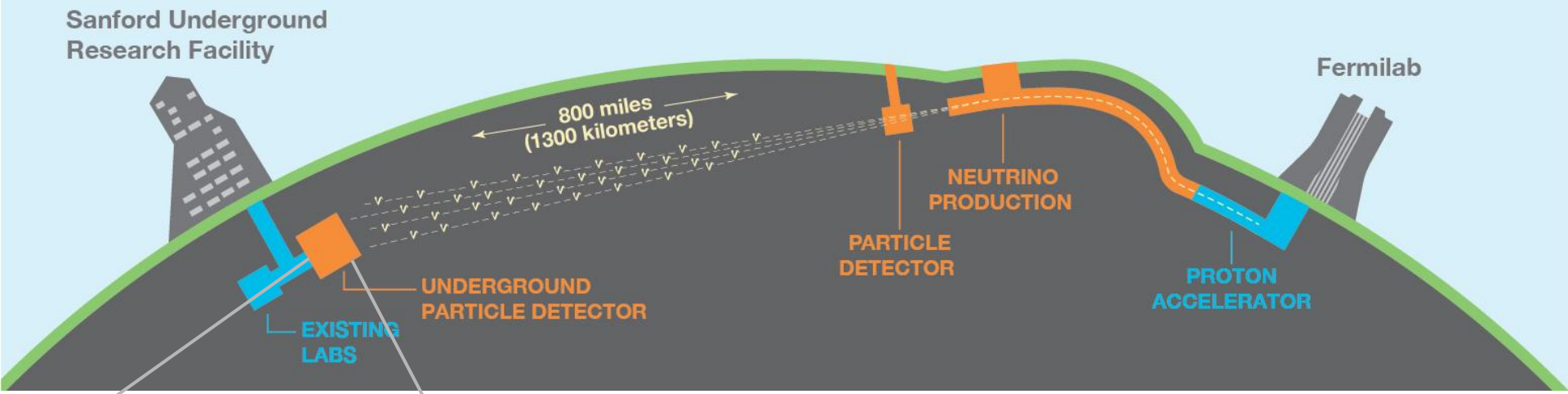
Outline

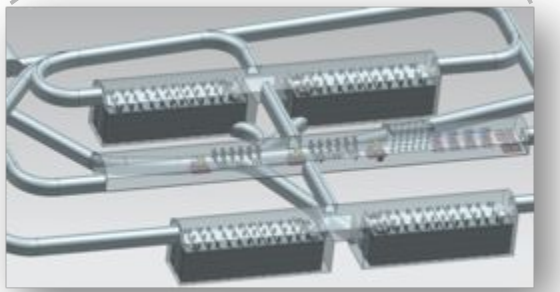
DUNE: How it works, and motivation for ML-based trigger

A two-level, ML-based data selection (trigger) scheme for rare events

Efforts toward a viable, energy-efficient implementation scheme

What is DUNE?





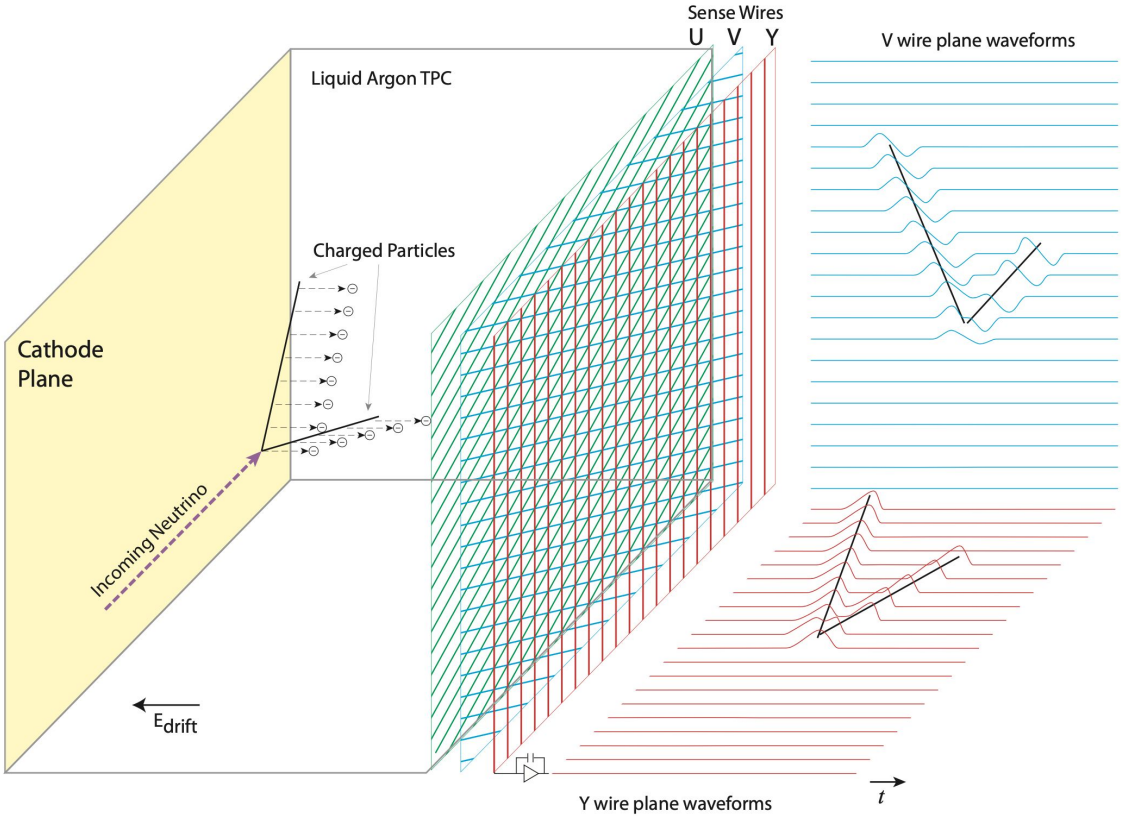
Far detector:

- 4 liquid argon time projection chamber (LArTPC) modules, each with 10kton fiducial mass
- underground (1.5km deep)

Physics goals of DUNE:

- CP violation in the lepton sector
- neutrino mass ordering
- search for rare events, e.g. proton decay, supernova burst neutrinos

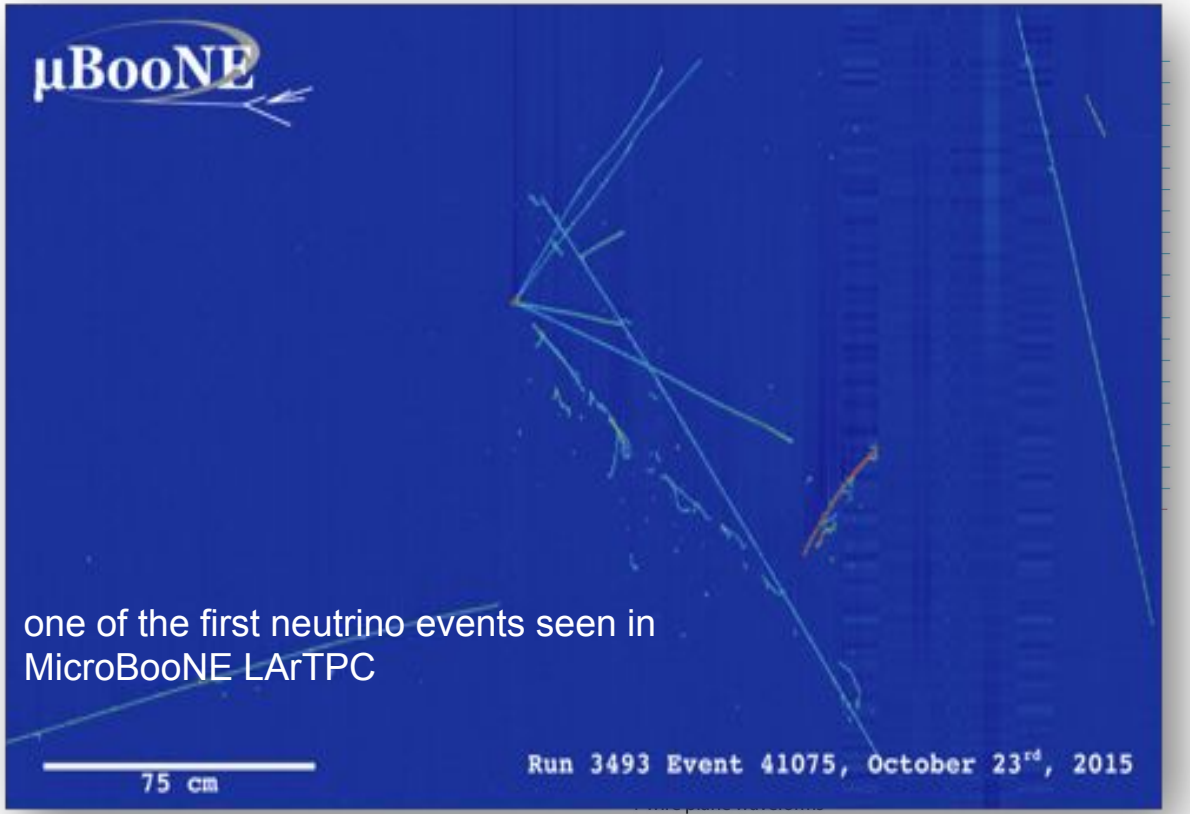
LArTPC detector



*For single phase far detector technology

1. neutrinos interact with argon nuclei, generating charged particles
2. charged particles ionize argon atoms
3. electrons from ionization drift to anode due to the electric field
4. Wire planes record signals from induction or collection. (Wires are reading out 2D projected views of the 3D interaction in the detector.)
5. Also light collection system detects prompt scintillation light, which provides t_0 of interaction

LArTPC detector



one of the first neutrino events seen in MicroBooNE LArTPC

1. neutrinos interact with argon nuclei, generating charged particles
2. charged particles ionize argon atoms
3. electrons from ionization drift to anode due to the electric field
4. Wire planes record signals from induction or collection. (Wires are reading out 2D projected views of the 3D interaction in the detector.)
5. Also light collection system detects prompt scintillation light, which provides t_0 of interaction

*Microboone is an already running LArTPC, and is 500 times smaller than DUNE.

Motivation

- DUNE (or any LArTPC) raw detector data is ideally suited for image analysis for data selection
  - Raw data is streamed out of TPC ‘frame by frame’ in the form of high resolution images
- Recent advances in machine learning allow to extract a lot of information from images
  - e.g. through the use of deep convolutional neural networks for image localization and identification
- Advances in hardware technology and tools enable the acceleration of computationally-intensive algorithms
  - e.g. Fast ML on FPGA

**G. Karagiorgi, Y. Jwa, G. di Guglielmo, L. Carloni;*
DOI: 10.1109/NYSDS.2019.8909784

→ ML-based triggering could be applicable in DUNE, using online (in software) or real-time (in hardware, e.g. FPGA) inference!

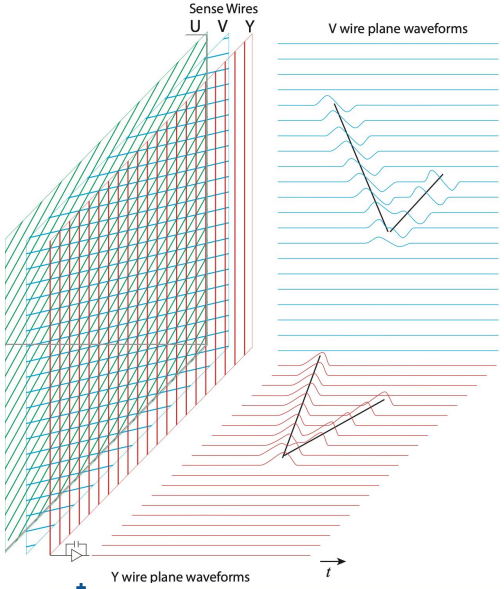
DEEP UNDERGROUND NEUTRINO EXPERIMENT

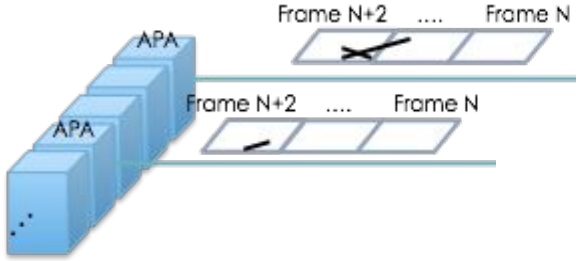
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

CNN-based SN burst trigger

"APA": Anode Plane Array, an array of sensor wires on the anode plane. Each APA is in the middle of a cell of liquid argon volume, and streams out data "frame by frame".
A single DUNE 10kton module has 150 APAs*.





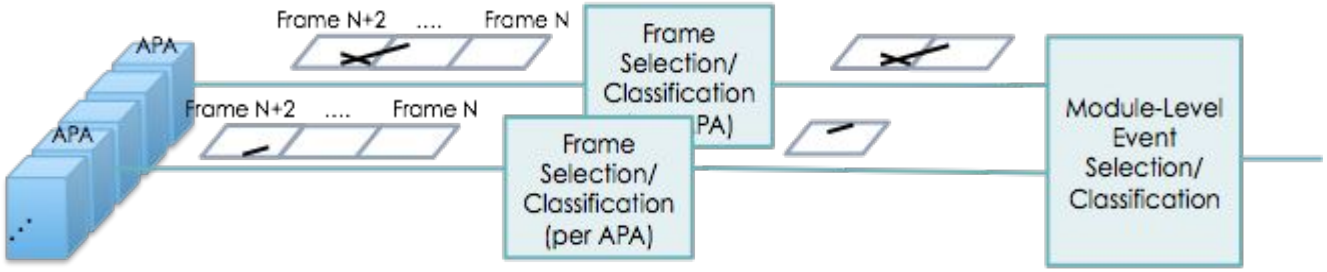
*for single phase LArTPC design.

DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

CNN-based SN burst trigger



1. Low-level:
CNN-based APA-frame selection and reweighting

2. Module-level:
APA-frame coincidence across module and
over 10 seconds

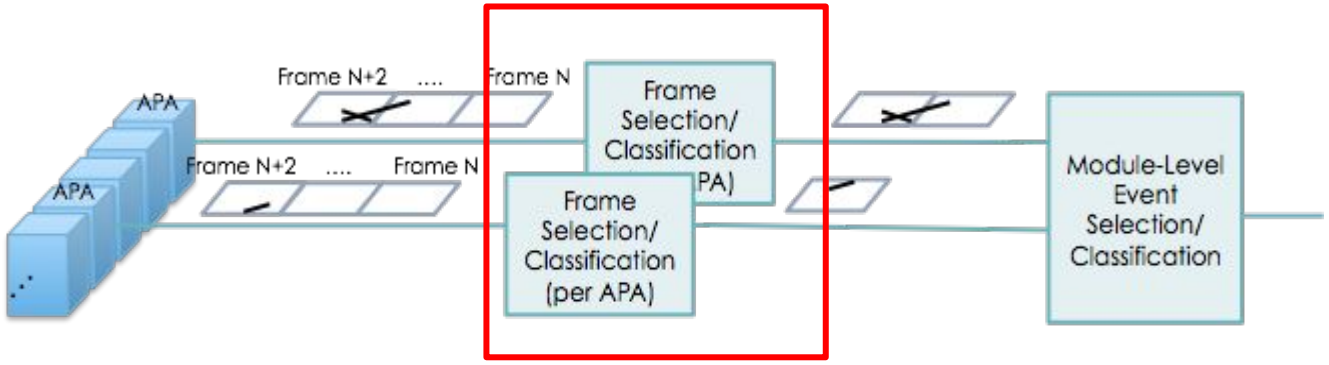
DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Low-level: CNN-based APA-frame selection

1. CNN image classification is used to tag raw TPC data, ‘frame by frame’, as containing three types of activity possible in DUNE:
   a. **SN neutrino interactions (LE)**
   b. **High-energy (HE) interactions**
   c. or just **background (NB)**.
2. Only frames tagged as **SN** and **high-energy** interaction are saved, without lossy compression.

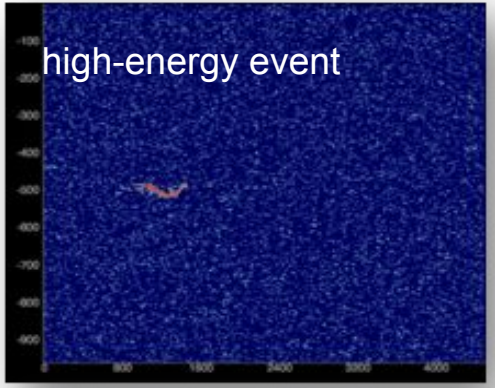


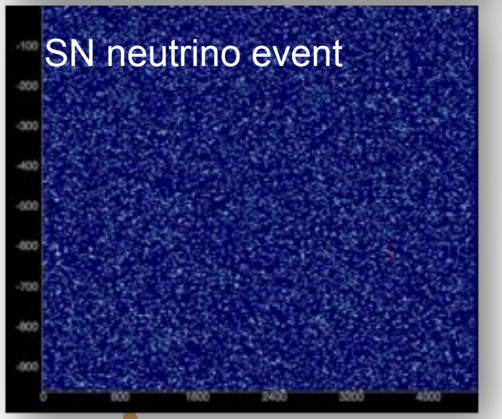
DEEP UNDERGROUND NEUTRINO EXPERIMENT

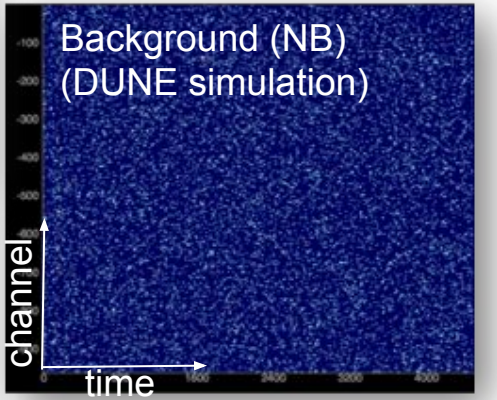
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Low-level: CNN-based APA-frame selection







- The network is trained to give 3 scores (HE, SN, NB) for each frame
- Then frames are kept according to their NB scores (we only keep frames with low NB score)

DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Low-level: CNN-based APA-frame selection

1. Only keep images surviving low NB score cut
2. Efficiencies are shown separately for each exclusive image type (only one interaction per frame assumed)

NB cut	Accuracy (%)						
	ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	$\epsilon_{HE:nnbar}$	$\epsilon_{HE:ndk}$	$\epsilon_{HE:atm}$	$\epsilon_{HE:cosmic}$
0.1	0.73	88.18	96.12	99.98	99.29	92.24	92.57
0.01	0.14	83.27	95.68	99.98	99.18	91.01	92.46
0.001	0.033	77.11	95.21	99.98	99.05	89.76	92.23
0.0001	0.011	69.74	94.61	99.97	98.74	88.39	91.71
0.00001	0.002	60.73	93.79	99.95	98.22	86.61	90.97

Fake rate meets offline data rate requirement of DUNE

*G. Karagiorgi, Y. Jwa, G. di Guglielmo, L. Carloni;
DOI: 10.1109/NYSDS.2019.8909784

DEEP UNDERGROUND NEUTRINO EXPERIMENT

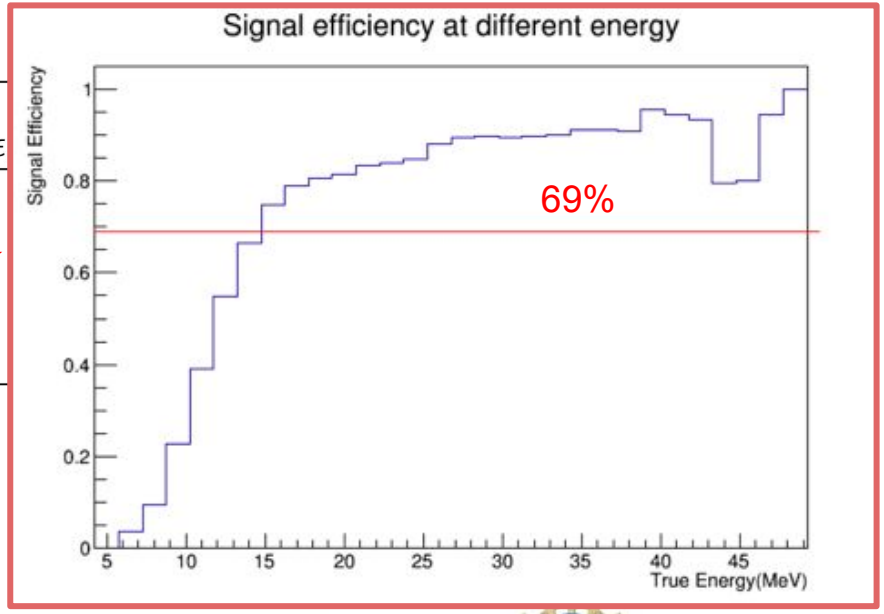
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Low-level: CNN-based APA-frame selection

1. Only keep images surviving low NB score cut
2. Efficiencies are shown separately for each exclusive image type (only one interaction per frame assumed)

NB cut	ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	ϵ
0.1	0.73	88.18	96.12	
0.01	0.14	83.27	95.68	
0.001	0.033	77.11	95.21	
0.0001	0.011	69.74	94.61	
0.00001	0.002	60.73	93.79	



Frame selection and energy-boost

From the single APA-frame selection, we find that **for the needed 1E-4 background reduction rate, an average SN interaction efficiency of 69% can be obtained**.

Are there possible improvements?
If we assume a CNN can also provide an estimate of the energy associated with a SN interaction in a SN tagged frame (R&D in progress), with some given resolution, we could increase SN selection efficiency by employing an “energy-boost” scheme:

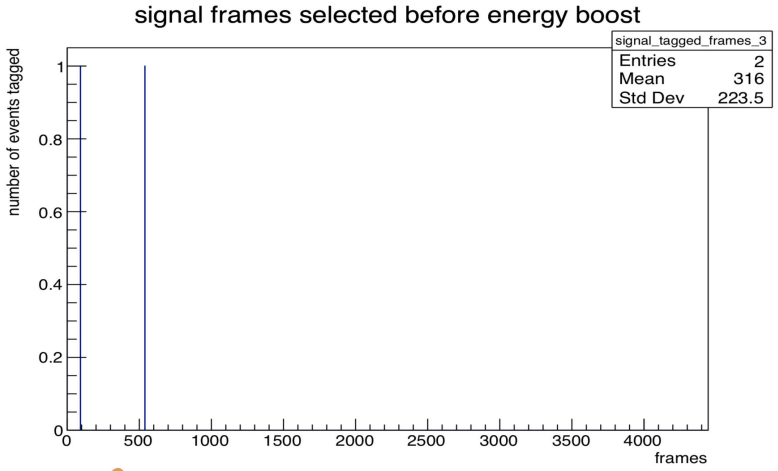
- select an APA-frame as a SN frame (as before)
- **preferentially weigh** the event based on the predicted energy (proportionally to the energy).

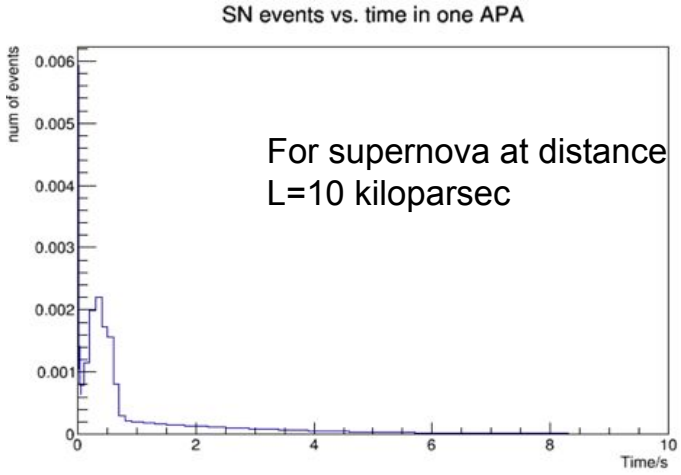
Because most backgrounds are at low energy, this is expected to help signal to background discrimination!

Frame selection and energy-boost: Simulation study

Assuming 10% resolution for CNN energy prediction, in this study, we use:

(1) energy-dependent efficiency for selecting a APA-frame from SN simulation.





DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

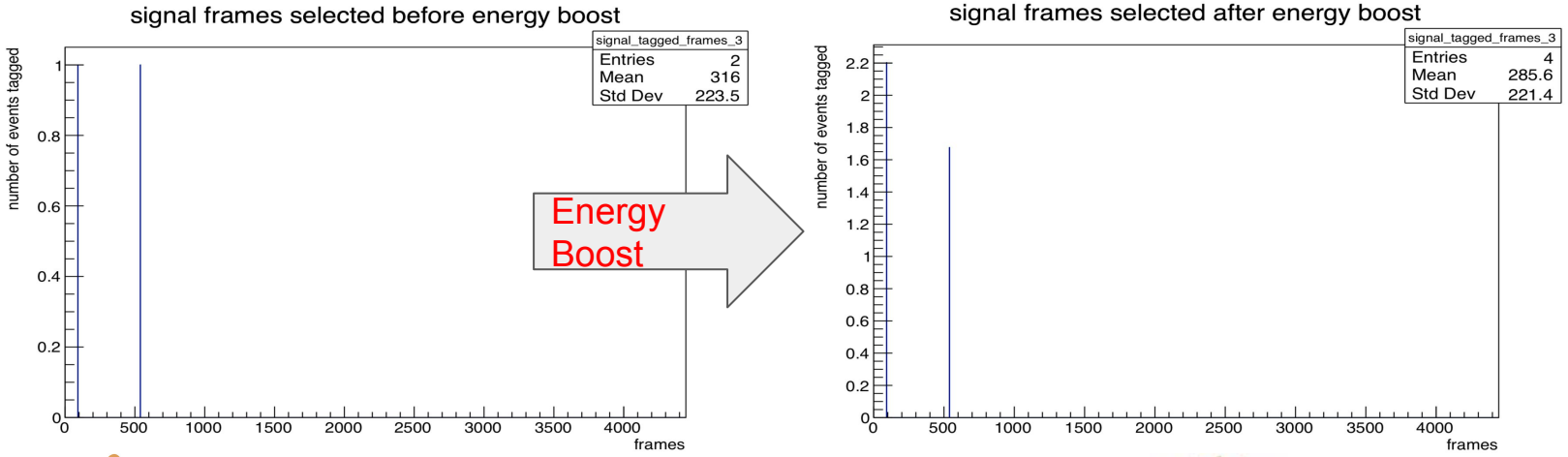
Frame selection and energy-boost: Simulation study

Assuming 10% resolution for CNN energy prediction, in this study, we use:

(1) energy-dependent efficiency for selecting a APA-frame from SN simulation.

(2) for frames selected, apply a 10% energy smearing to mimic CNN energy-prediction resolution.

If its predicted (smeared) energy is >10 MeV, scale the frame by a factor proportional to its energy: smeared energy[MeV]/10.

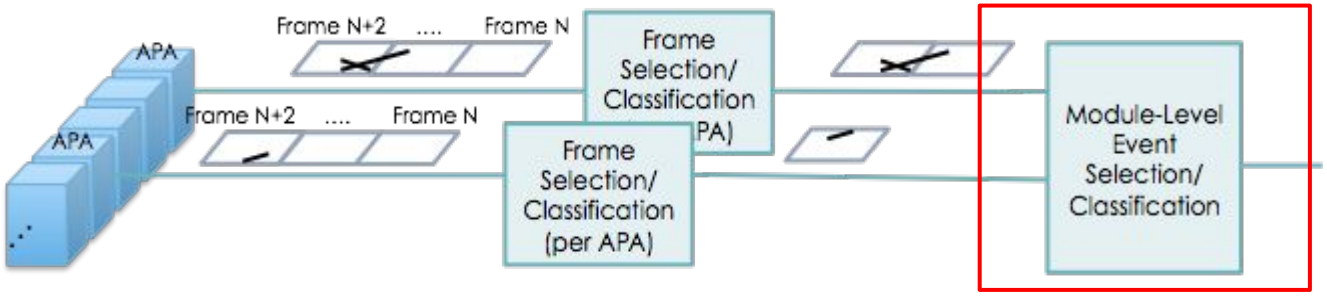


DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Module-level SN burst trigger



Module Level Trigger makes use of the fact that for a galactic supernova we can have up to thousands of neutrino interactions in coincidence over ~10s

2. Module-level:
APA-frame coincidence across module and over 10 seconds

(1) Calculate the APA-frame coincidence (within N-successive-frames window over the 10kton module)
→ defined as “multiplicity”

(2) Signal and background simulation will have different multiplicity distribution
→ place a cut on the multiplicity, to pick out the signal while keep background “fake rate” low

DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Simulation

Network & Training:

- VGG16b* network is used for the training and inference, without initial weights given to the network during the training.
- Images simulated to train/test the network:

Process	SN	NB	HE: nnbar	HE: ndk	HE:atmo	HE:cosmic
Events	74700	150100	75636	76424	74256	60852

*This is a rather big network. Much smaller network (4-layer network with 1 convolutional layer) has been tested as well with similar performance.

DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Simulation

Frame selection & Module-level trigger:

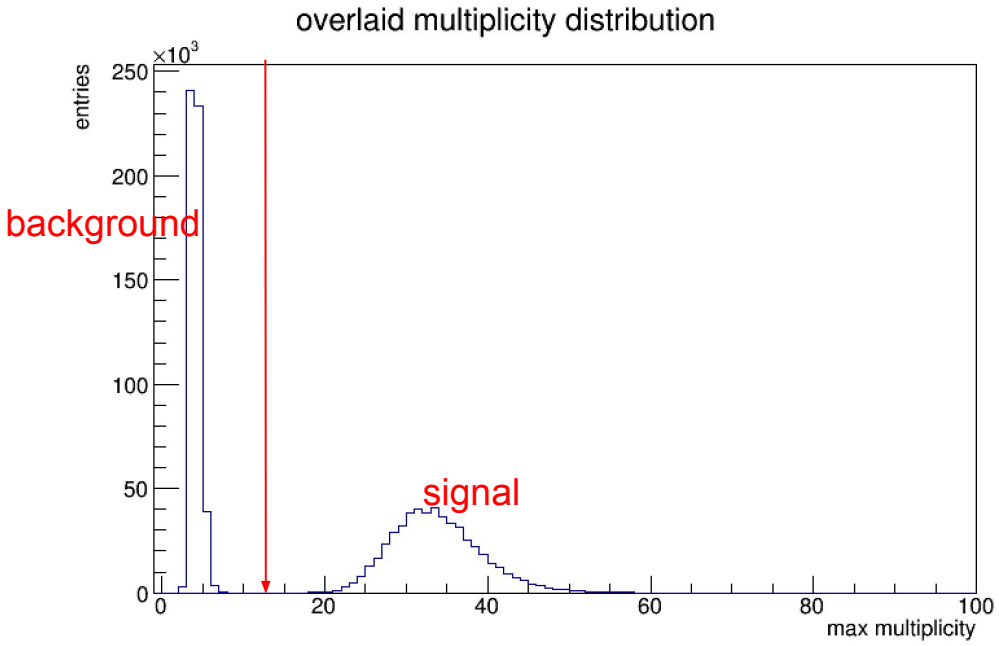
We did 520k simulations for signal (SN burst) and background:

- Each simulation is 10-second (~4445 frames) long, which is the duration of a SN burst, and has the distribution of neutrino events (vs. time) expected in 1 APA plane.

  So that's about: 520k*10 second/(60 s/min * 60 min/hr * 24 hr/d * 30d/month) ~ 2 months worth of background data!

- Signal is simulated based on the neutrino flux distribution (provided by DUNE SNB/LE Physics Working Group), while background is simulated as random distribution based on fake rate.

Example: for SN 15 kiloparsec (kpc) away, with N=20 successive frames



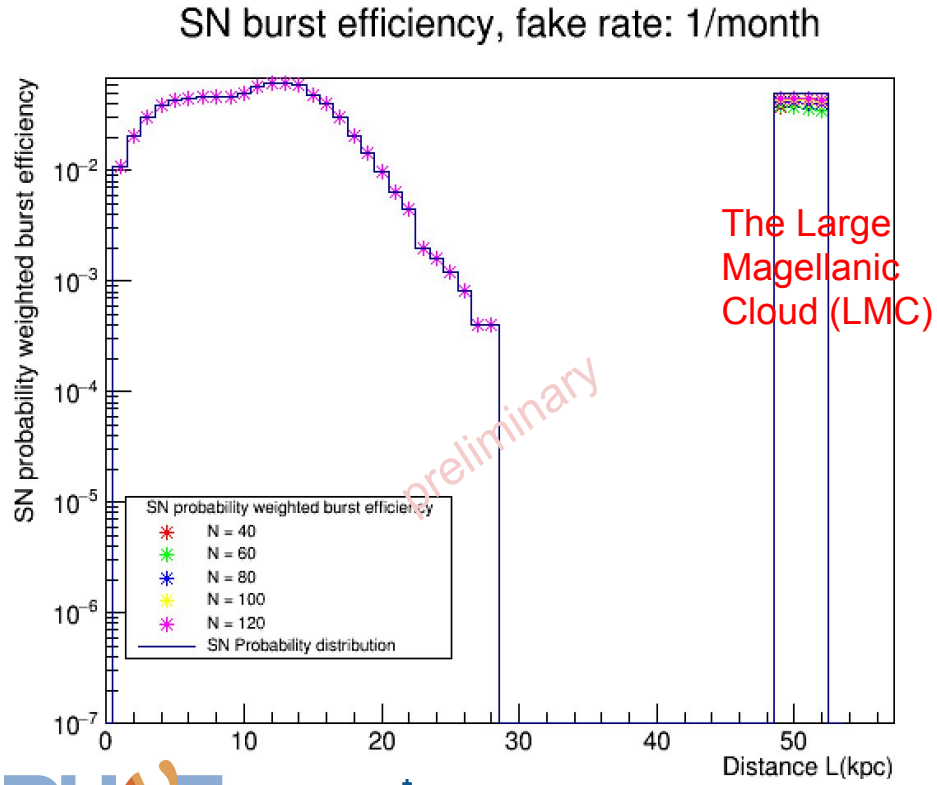
We could place a cut at around 10-15, and achieve 100% SN burst efficiency with fake rate <= 1/month!

DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Performance: Module Level Burst Trigger



Galactic coverage = integral of (burst efficiency x SN probability) graph

N	Galactic Coverage	Uncertainty (10^{-5})
20	0.981	5.96
200	0.997	1.25
400	0.998	0.835
500	0.998	0.895
600	0.998	0.974
4445	0.988	3.44

preliminary

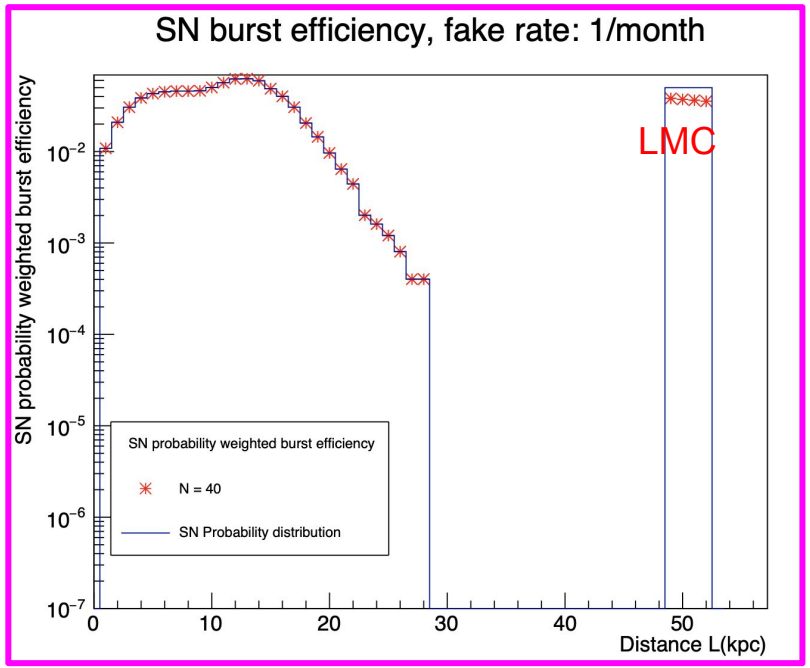
DEEP UNDERGROUND NEUTRINO EXPERIMENT

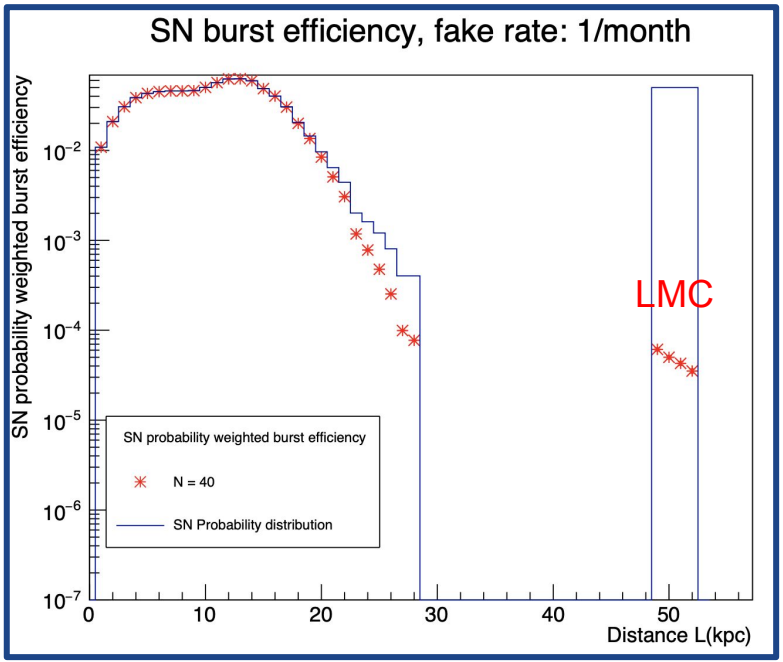
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Performance: Module Level Burst Trigger

Comparison between energy-boosted method and method with no energy boost*.





*Averaged flat selection efficiency (69%) is also used in this case for simulation.

Findings of ML-based trigger simulation study

- A machine learning-based data selection method with a two-level selection scheme for SN burst triggering shows great promise, reaching galactic SN burst coverage of ~100%.
  This assumes the SN neutrino energy associated with SN tagged frames can be determined with 10% resolution.
- We are in the process of training networks to perform energy estimation in order to study performance and refine simulations. In the meantime, trying to understand how quickly the trigger algorithm would work.
- Study of supernova neutrino direction predictor using ML-based regression is also planned.

DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

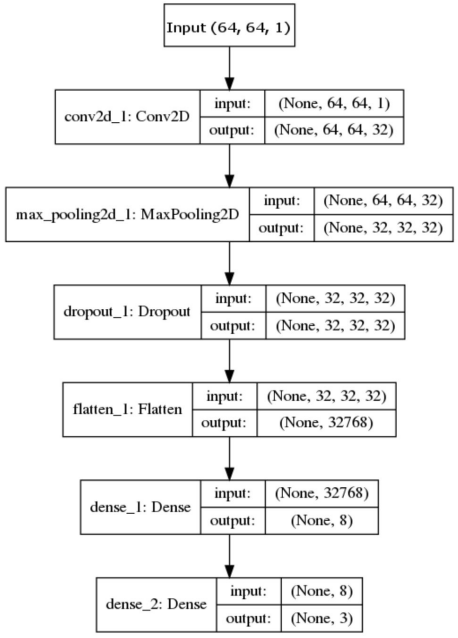
NSF

Implementation options, and hardware acceleration

- The ML-based data selection method could be applied through a number of different hardware implementations: CPU, GPU or FPGA.

- Ongoing efforts at Columbia (Physics+CS collaboration) to demonstrate low-level data selection (image classification) on FPGA:

    - Studies have targeted smaller CNN_s (e.g. 4-layer network).
    - An implementation has been accomplished on FPGA (Xilinx Embedded FPGA that combines both an ARM Cortex-A53 CPU), which can keep up with a reduced frame rate that would be possible from pre-processing (ROI-finding) of APA-frames.



Overview of CNN_s.

Y. Jwa et al, DOI: 10.1109/NYSDS.2019.8909784

DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Implementation options, and hardware acceleration

Performance and power analysis of CNN_s:

Platform	Model	Time (s)	Power (W)	Energy Efficiency (img/s/W)
ARM C-A53	CNN_s	0.0855	2.871	4.074
FPGA	CNN_s	0.0511	1.110	17.630

**G. Karagiorgi, Y. Jwa, G. di Guglielmo, L. Carloni;*
DOI: 10.1109/NYSDS.2019.8909784

- Time in the table is inference time for a single (64x64 ROI) image.
- The energy efficiency of the FPGA implementation is more than 4 times better than the embedded CPU.

DEEP UNDERGROUND NEUTRINO EXPERIMENT

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

NSF

Summary

- DUNE is an ideal application for online or real-time image classification for triggering purposes.
- We have demonstrated that a number of CNNs can be trained (a priori) on simulated events and yield high trigger efficiencies for rare event searches.
- Ongoing efforts are focusing on real-time implementation in FPGA and demonstration.

Thank you!

Back up

CNN-based data selection in more detail

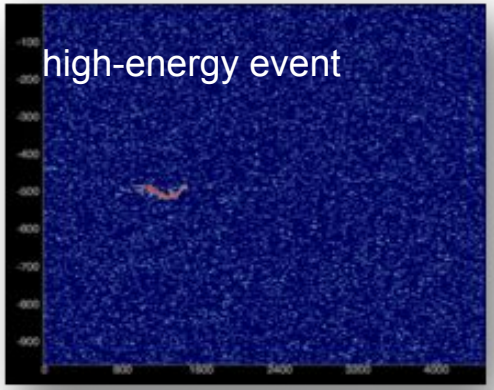
1. CNN (vgg16b) network is trained on frames containing:
   a. (signal frame*) SN neutrino interactions
   b. (signal frame) High-energy off-beam interactions (including proton decay, n-nbar oscillation, cosmic, atmospheric neutrino)
   c. (background frame) Radiologicals and noise only backgrounds
2. Raw data was downsampled to 600X600 pixels to meet CNN input requirements for training and inference.
3. The network will give 3 scores (SN, HE, RAD) for each frame, and then frames are kept according to their RAD scores. (we only keep frames with low RAD score)

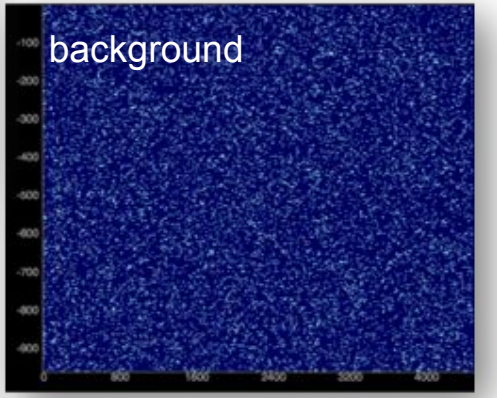
*Signal frame is defined as frames containing the true interaction vertex, not necessarily containing all final states.

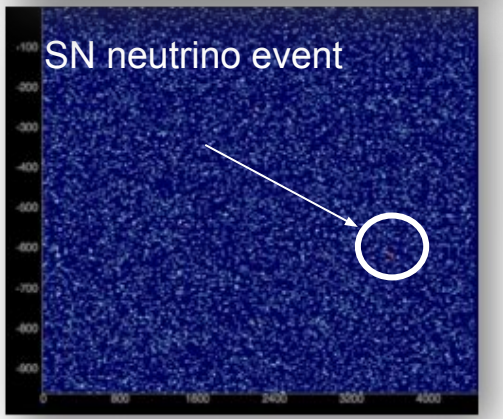
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

DUNE
DEEP UNDERGROUND
NEUTRINO EXPERIMENT

Low-level: CNN-based APA-frame selection



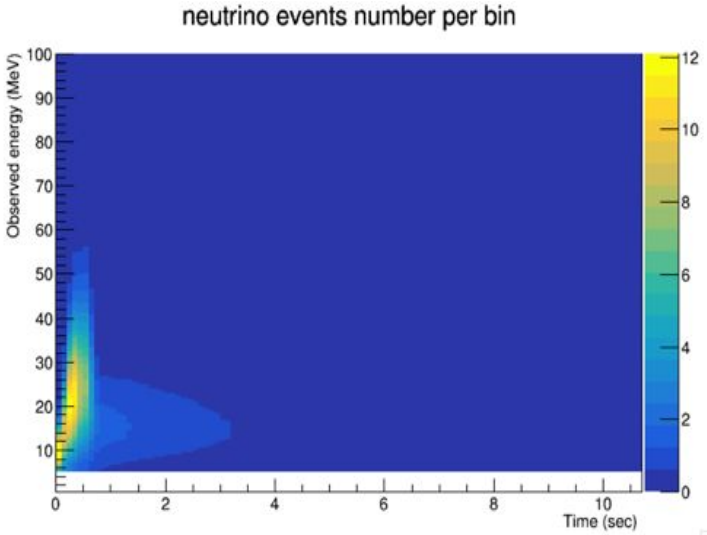


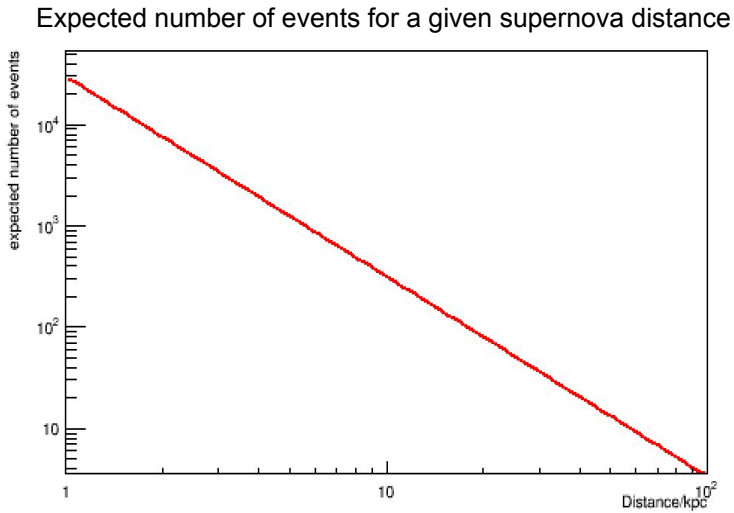


the network is trained to give 3 scores (HE, SN, RAD) for each frame, and then frames are kept according to their RAD scores. (we only keep frames with low RAD score)

Frame selection and energy-boost

- SN neutrino flux vs. energy and time over the burst duration (10s = 4445 frames) are theoretically predicted, provided by Kate Scholberg.
- Given a certain distance, then the number of SN neutrino events vs. energy vs. time for 1 APA-frame could be predicted.



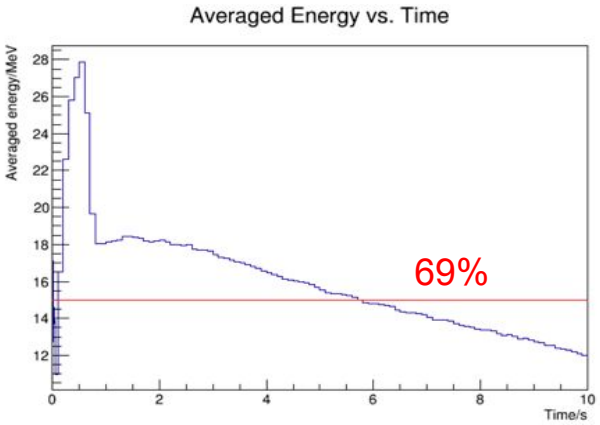


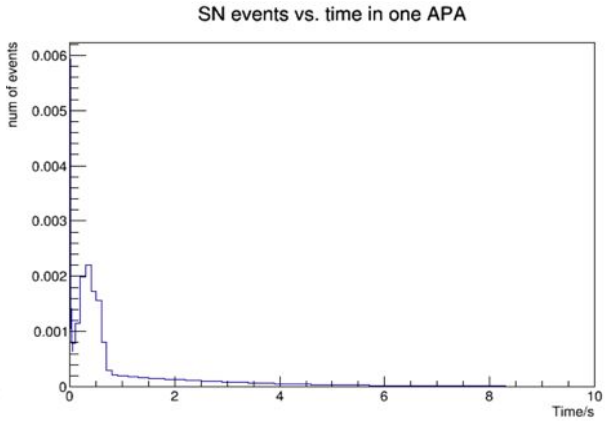
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

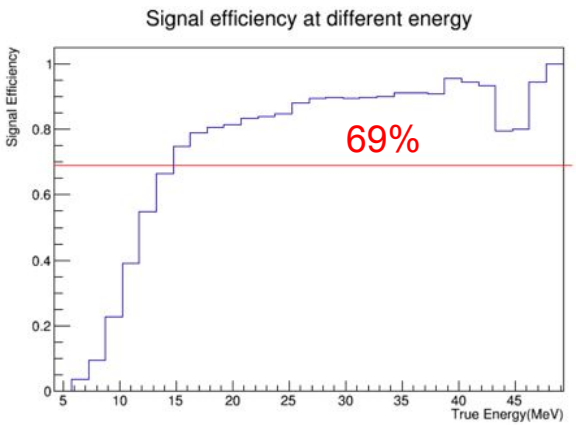
DEEP UNDERGROUND NEUTRINO EXPERIMENT

Frame selection and energy-boost

averaged energy distribution and flux vs. time distribution can be inferred from the 2D flux distribution.







For supernova at distance L=10 kpc

Frame selection and energy-boost

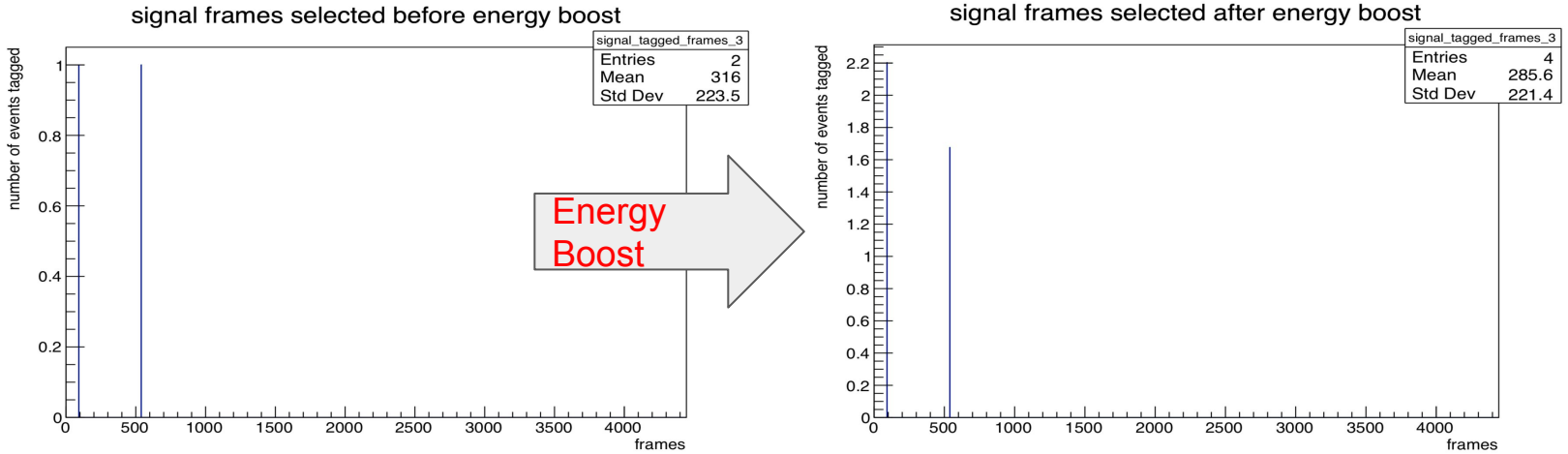
Each frame across 10s is filled with SN+fake event distribution accounting for

(1) expected distribution of neutrinos vs. E vs. time

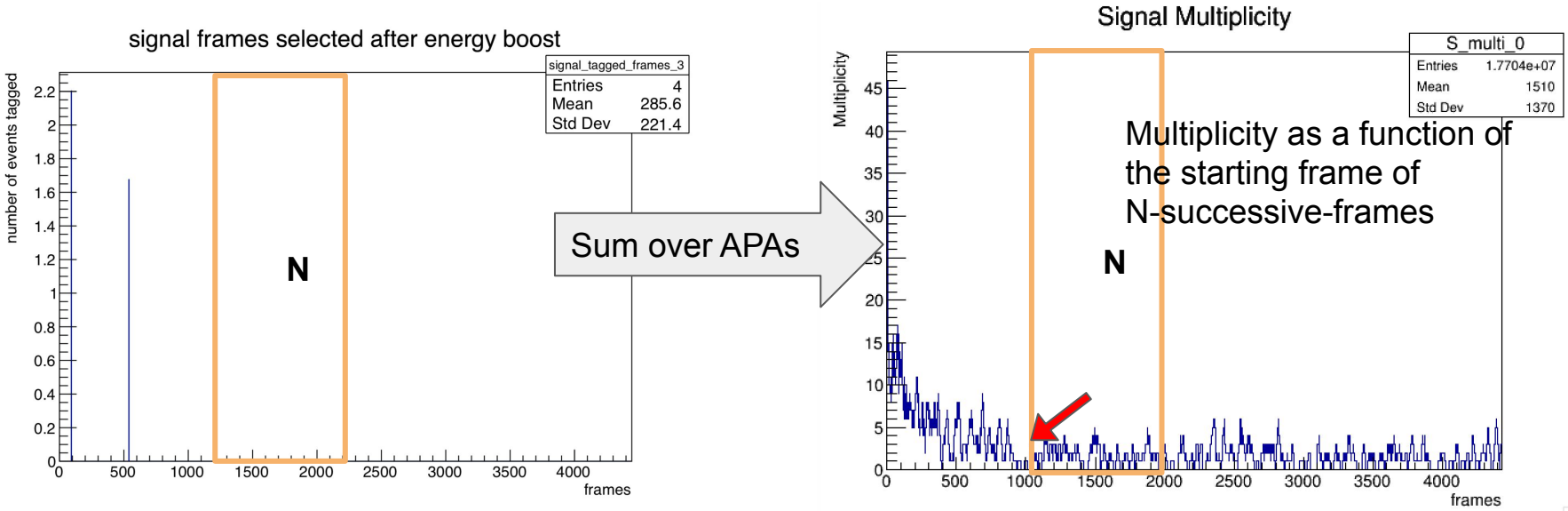
(2) poisson fluctuations

(3) **energy-dependent** efficiency for selecting a SN frame and flat efficiency (0.011%) for selecting a background frame

(4) for frames selected, a 10% energy smearing is applied to approximate low level trigger energy prediction resolution. If its predicted (smeared) energy >10 MeV, the frame is weighted by a factor proportional to its energy (smeared energy[MeV]/10).



Aggregate tagged frame over a single APA (collection planes) and 200 APA (collection planes)



For a SN burst at 10kpc (plus background), a block of N-successive-frames strides from the first frame to the end

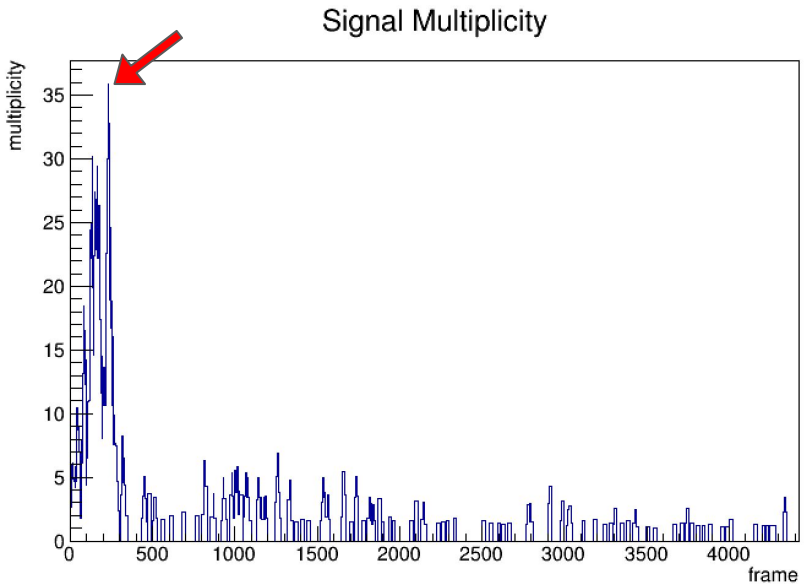
Over 200 APA collection planes, the multiplicity is defined to be:
total number of tagged frames within the window of N-successive-frames over 200 APA collection plane frames.

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

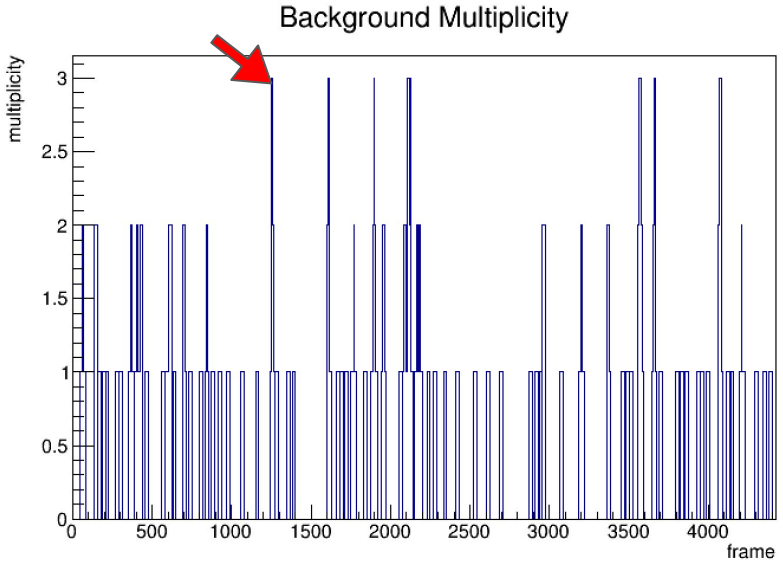
DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

Multiplicity distribution for signal and background

for 1 simulation (10s X 200 APA collection planes)



SN multiplicity over N=20 at 15kpc
Maximum multiplicity of the SN burst : 35.87



background multiplicity over N=20 at 15kpc
Maximum multiplicity of the fake burst : 3

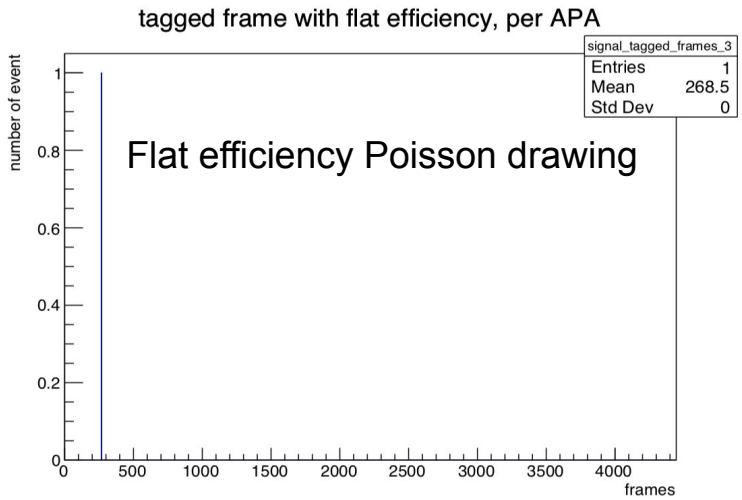
Comparison between flat efficiency & energy dependent efficiency

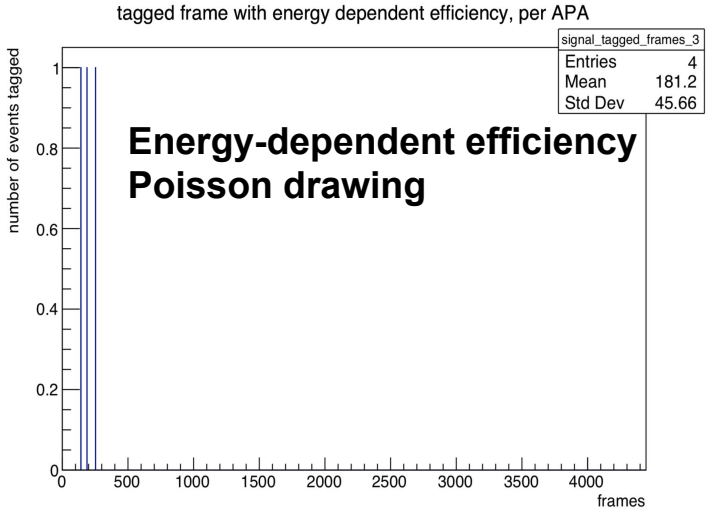
Each frame across 10s is filled with SN+fake event distribution accounting for

(1) expected distribution of neutrinos vs. E vs. time

(2) poisson fluctuations

(3) **energy-dependent** efficiency for selecting a SN frame and flat efficiency (0.011%) for selecting a background frame





Performance

- This result is viable in terms of physics performance. However, the CNN used is a very large network(VGG16b), the inference time on GPU is 27.7 ms.
  - Given the length of frame in DUNE is 2.25ms, this method can't keep up with the frame rate.
  - we tried accelerate the convolution operations on FPGA, but we still don't get the inference times we need to keep up with the DUNE rates for the low-level data selection, but it's viable for high- level filter stage with a lower frame rate.

- We tried pre-processing the APA-frames: noise removal, picking up "region of interest"(ROI) parts of the image; and tried a smaller customized network (CNN_s).
  - It gives similar physics performance results, and we get significant speed up and power improvement!
  - Combined with the fact that ROI finding reduces frame rate by a factor of 50, this is a viable scheme in terms of inference time (1.6ms).