ML4GW: An AI-based pipeline for Real-time Gravitational Wave Analysis

Erik Katsavounidis (MIT, LIGO Laboratory,A3D3 Institute)

on behalf of the ML4GW team: Will Benoit, Deep Chatterjee, Michael Coughlin, Malina Desai, Katya Govorkova, Alec Gunny, Phil Harris, Ethan Marx, Eric Moreno, Saleem Muhamed, Rafia Omer, Ryan Raikman

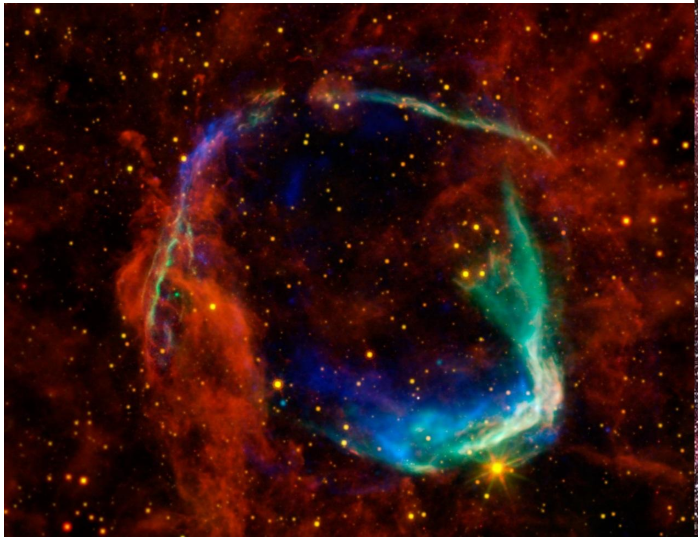
at the University of Minnesota, MIT and the A3D3 Institute

17th Marcel Grossmann Meeting

July 7-12, 2024

Award PHY-2117997

Observing our Universe



SN185

Image Credit: X-ray: NASA/CXC/SAO & ESA; Infared: NASA/JPL-Caltech/B. Williams (NCSU)

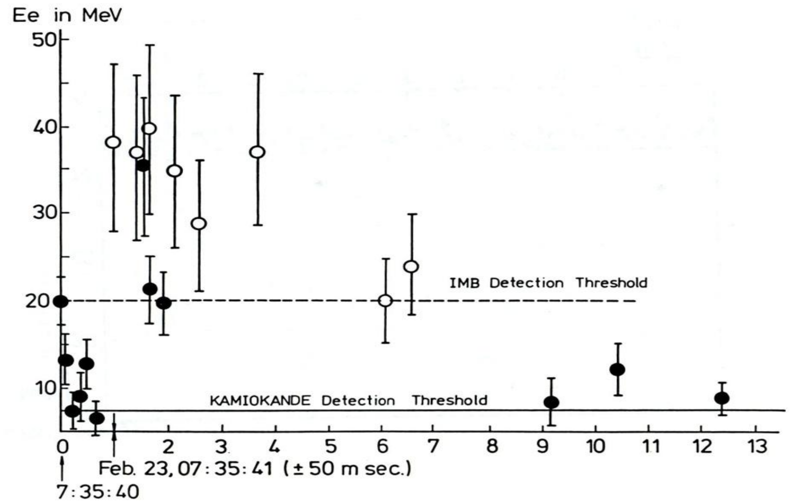
SN1987A

Electromagnetic waves

Observing our Universe

SN185

Image Credit: X-ray: NASA/CXC/SAO & ESA; Infared: NASA/JPL-Caltech/B. Williams (NCSU)



SN1987A

Koshiba, M. et al. 1988

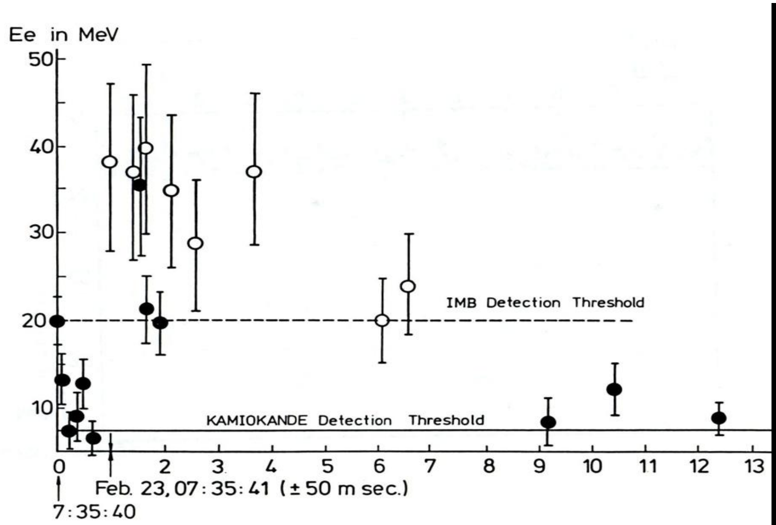
Electromagnetic waves

Particles: neutrinos, cosmic rays

Observing our Universe

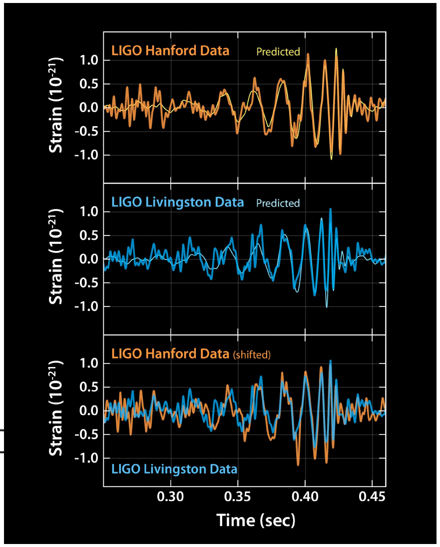
SN185

Image Credit: X-ray: NASA/CXC/SAO & ESA; Infared: NASA/JPL-Caltech/B. Williams (NCSU)



SN1987A

Koshiba, M. et al. 1988



GW150914

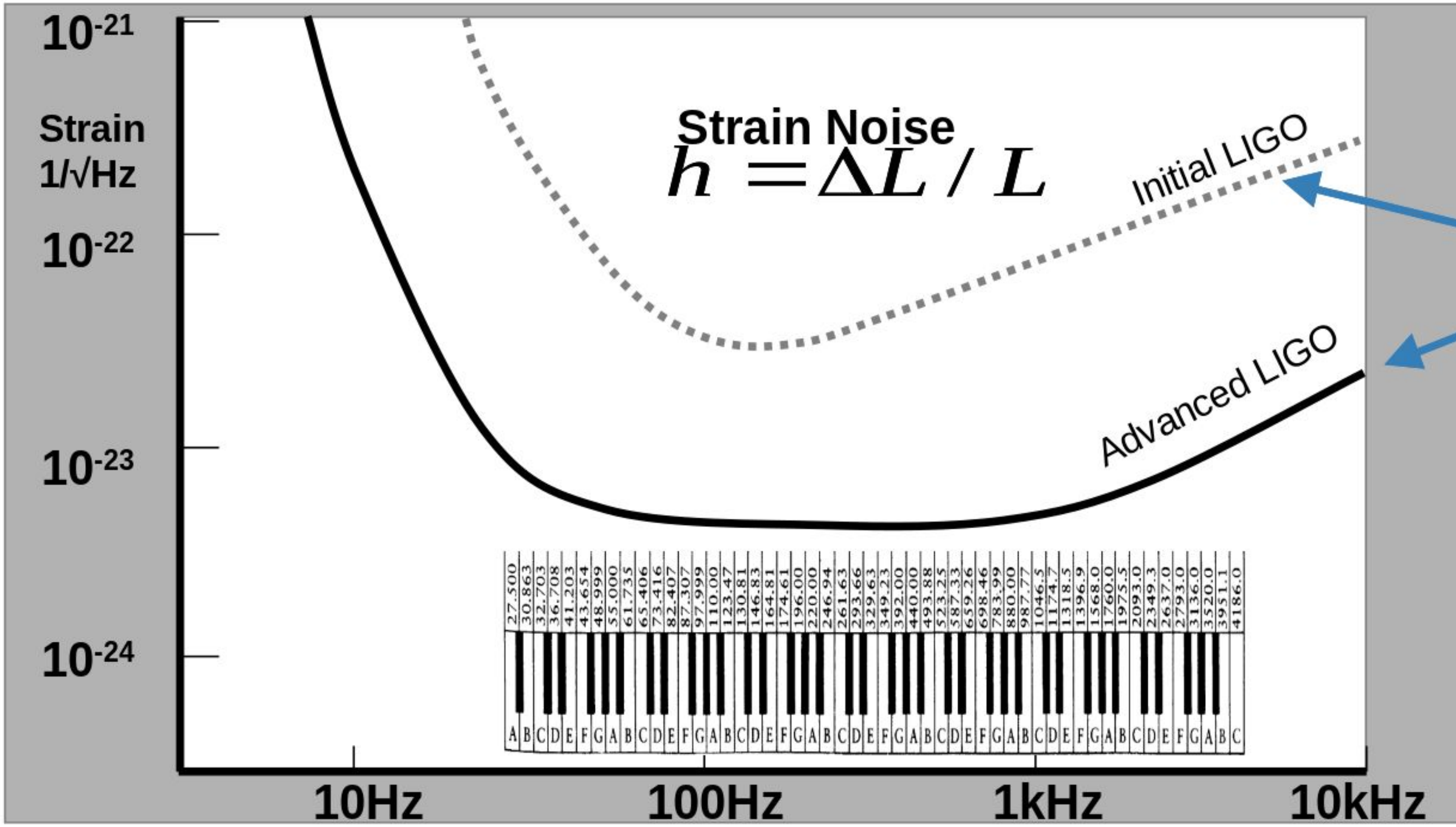
PRL 116, 061102, 2016

Electromagnetic waves

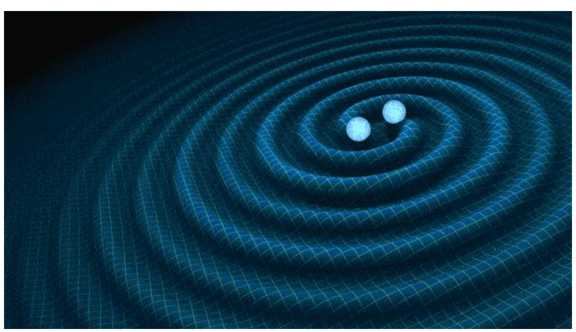
Particles: neutrinos, cosmic rays

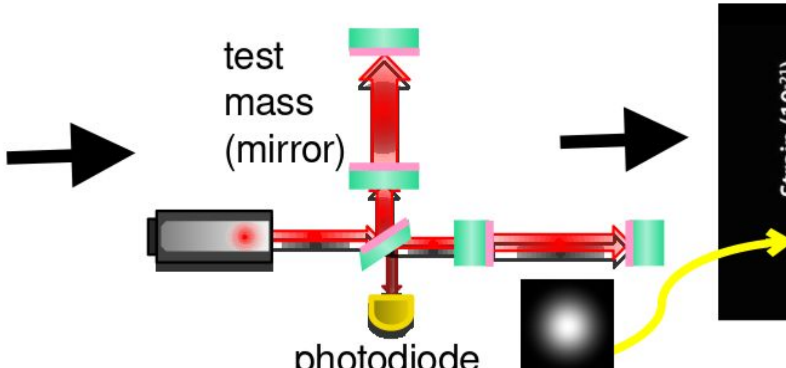
Gravitational waves

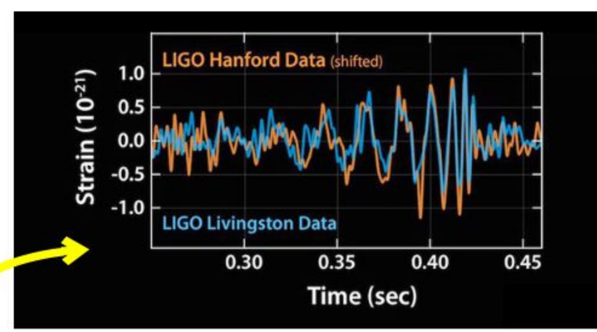
Gravitational wave detectors



The lower noise gets, the better the ability to detect GWs







(after calibration)

GW170817: The first Binary Neutron Star Merger

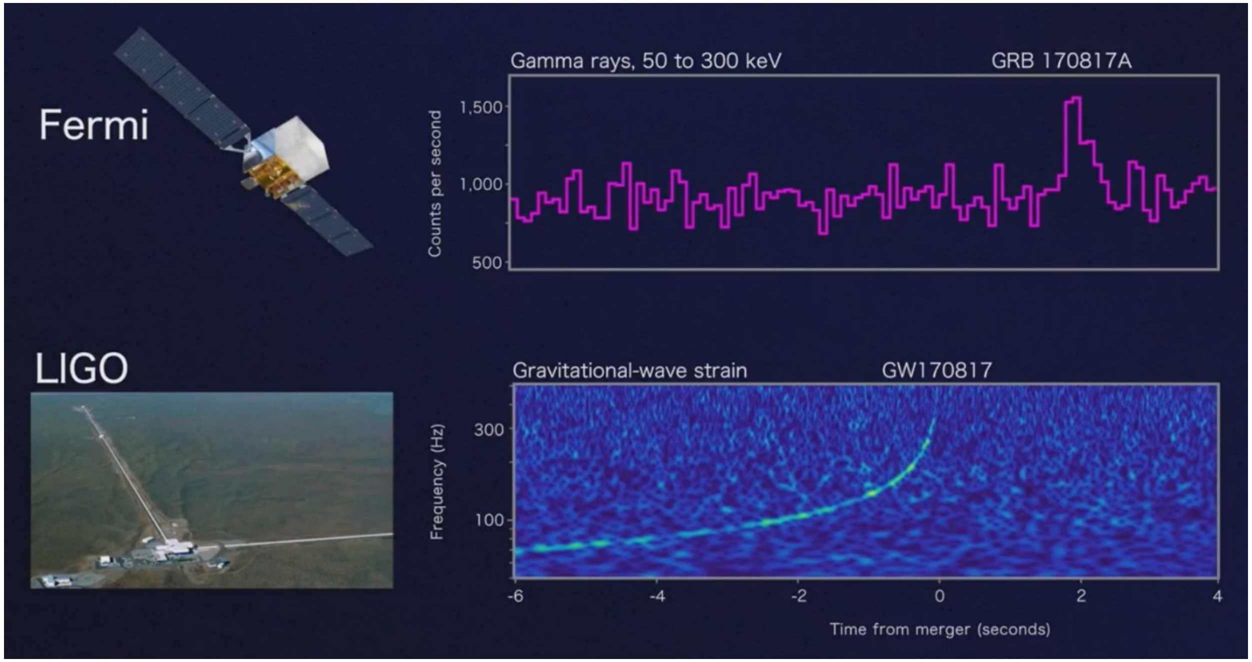
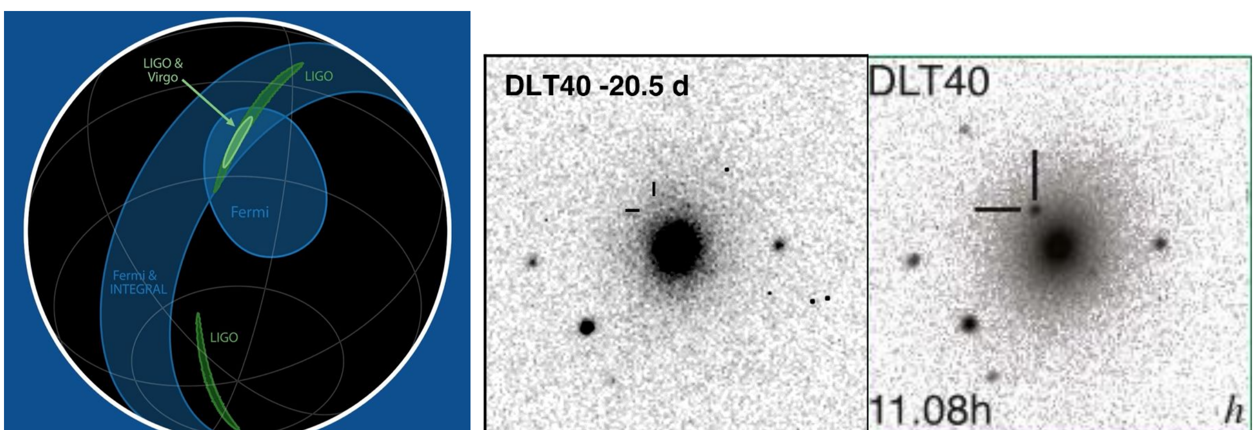


Image credit: NASA GSFC & Caltech/MIT/LIGO Lab



EM Partners with LIGO-Virgo, *Astrophys. J. Lett.* 848, L12 (2017)

sGRB progenitors

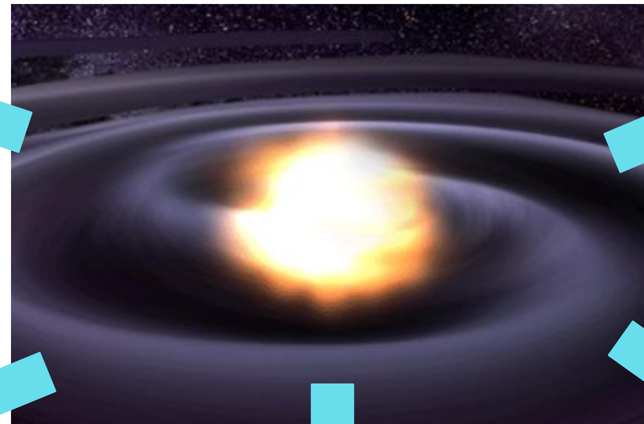
Kilonova and the origins of heavy elements

'Standard siren' measurement of the Hubble constant

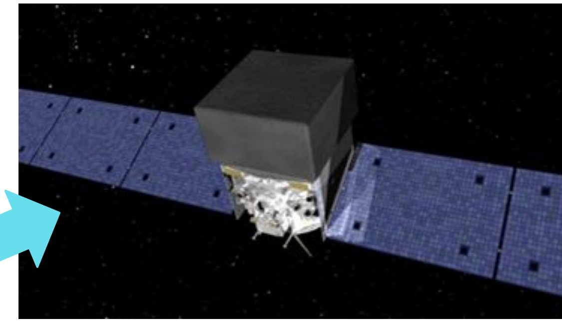
Speed of gravity

Multi-Messenger Astrophysics

Image credit: NASA Goddard Space Flight Center/ Dana Berry



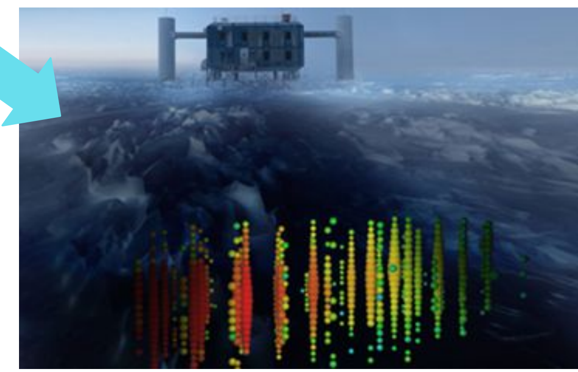
Gravitational waves



X-rays/Gamma-rays

Visible/infrared light

Radio waves



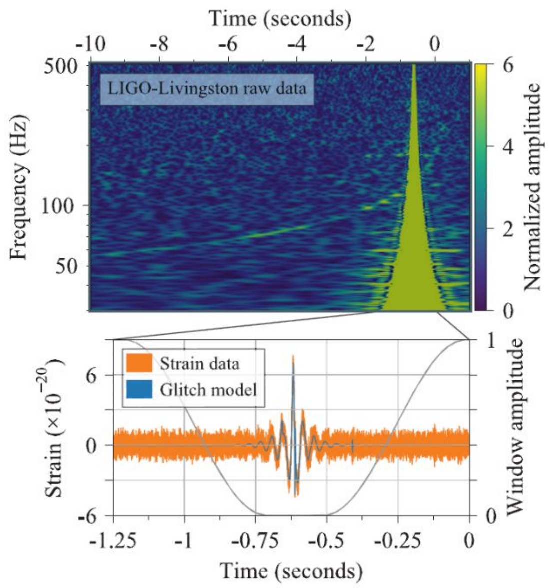
Neutrinos

The Challenge: the 3 deadly F's

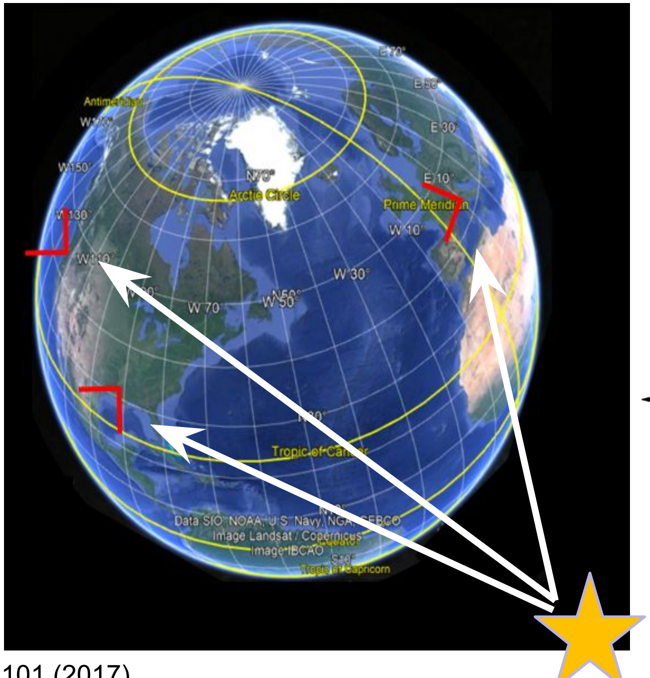
Fast:
need to identify GW transients as quickly as possible in order to have a chance to catch the earliest light

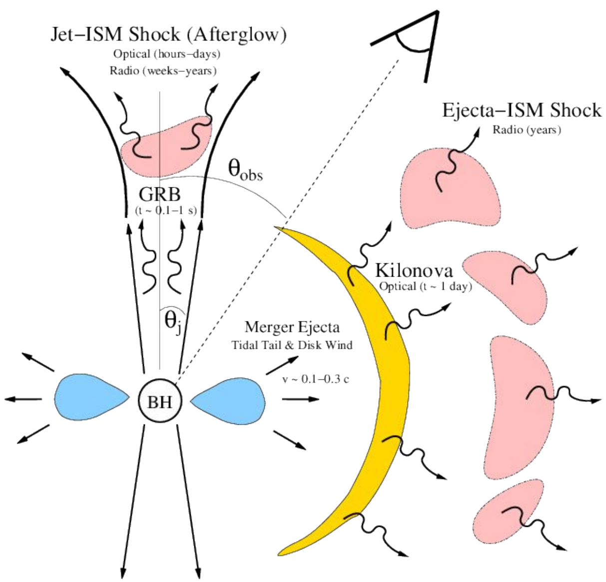
Fuzzy:
gravitational-wave detectors are more like radio receivers than telescopes

Faint:
for EM counterparts at the nominal BNS merger range of 200Mpc and BBH ranges out to Gpc



LIGO-Virgo Collaborations, *Phys. Rev. Lett.* 119, 161101 (2017)

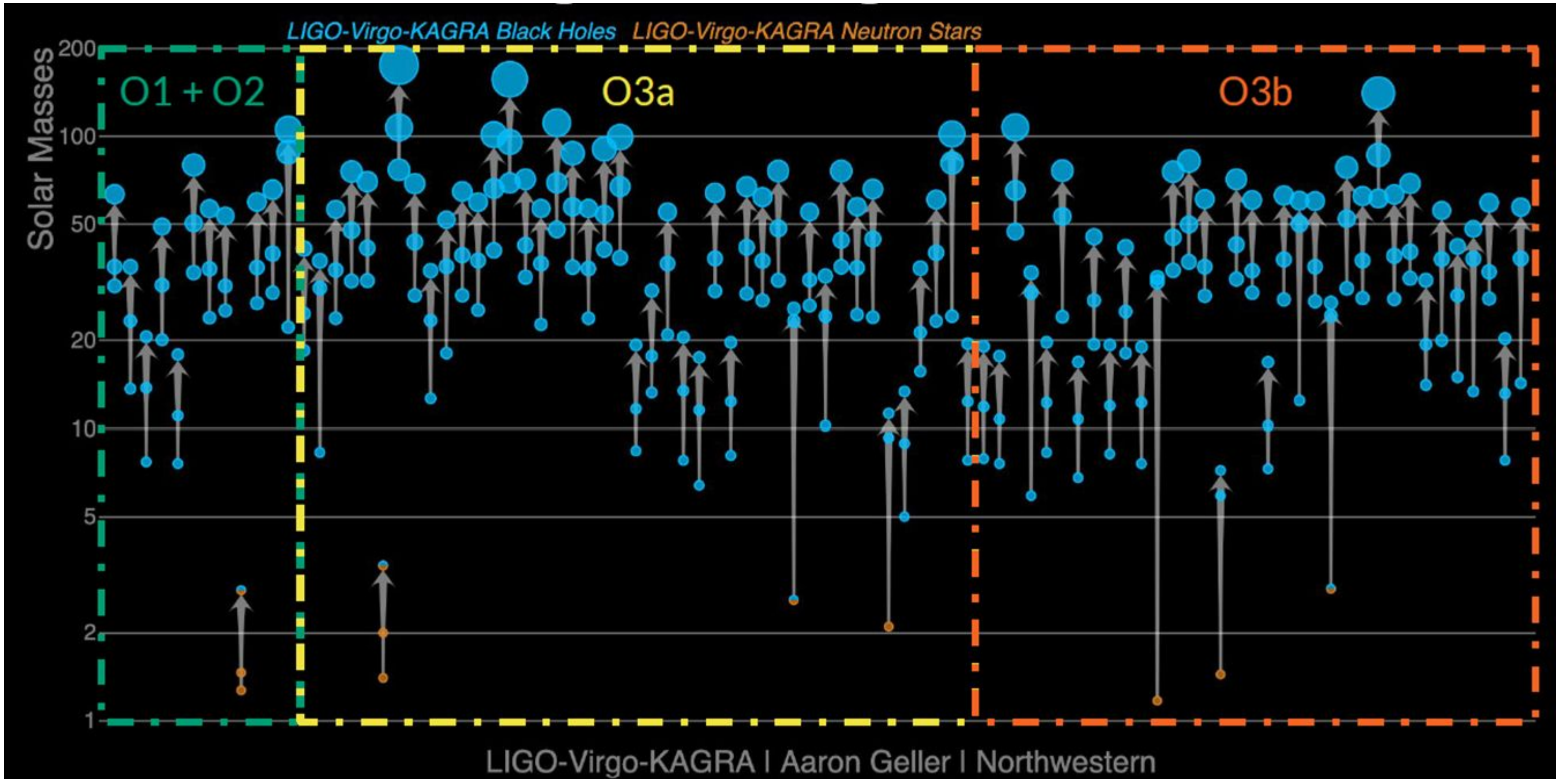




Metzger and Berger, *Astrophys. J.* 746, 1 (2012)

GW transient events

Image: https://www.zdnet.com/article/kafka-channels-the-big-data-firehose/



11 events from O1+O2

44 events in O3a, 55 total
1041 “subthreshold” events in O1,O2,O3a

35 events in O3b, 90 total (catalogs are cumulative)

Gravitational-wave detector data

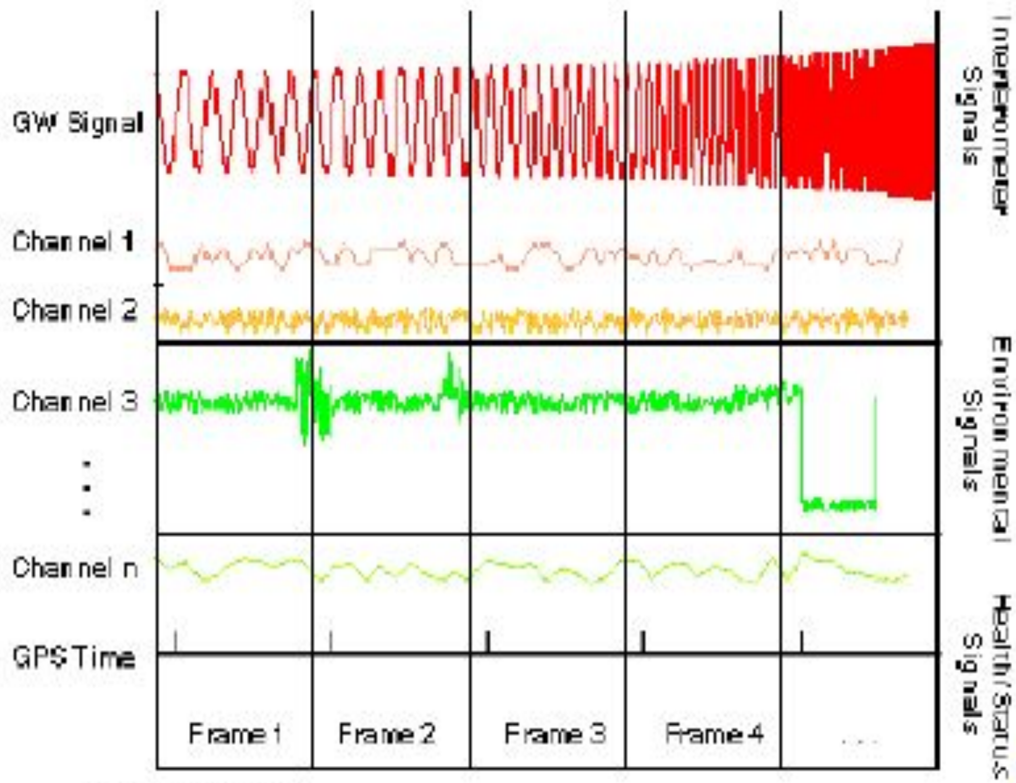
Continuous **time series** (1Hz, 128Hz … 16kHz)

Gravitational Wave channel: ~20GB/day (per instrument)

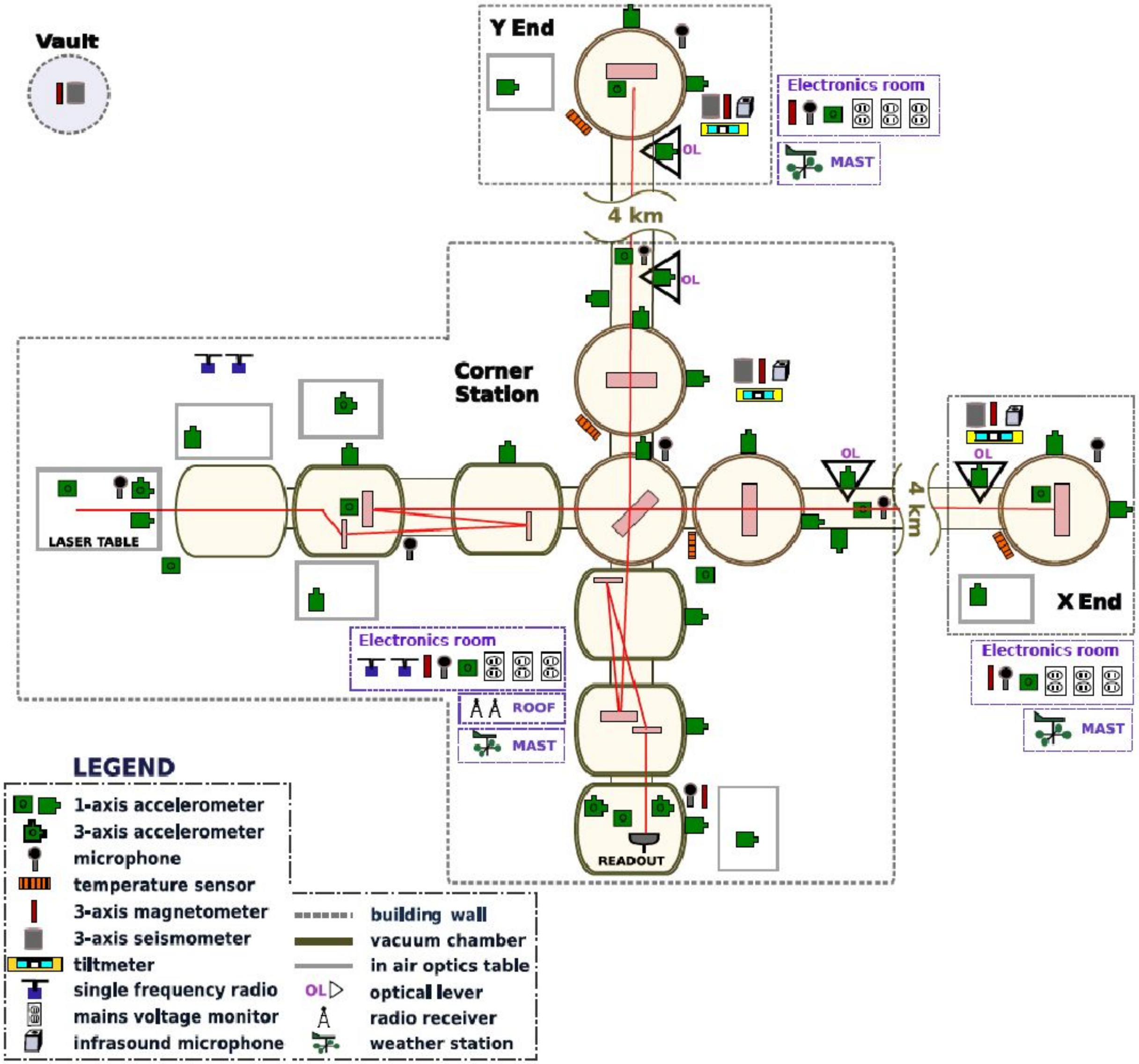
Physical Environment Monitors (seismometers, accelerometers, magnetometers, microphones etc)

Internal Engineering Monitors (sensing, housekeeping, status etc)

Together with various intermediate data products >2TB/day (per instrument)



Interefometric and environmental sensors



LIGO and Virgo Collaborations, *CQG* **33**, 134001 (2016)

Each LIGO detector records over 200,000 auxiliary channels that monitor instrument (interferometric) behavior and environmental conditions.

Enables the study of correlations (couplings) of the gravitational wave channel with the environment (including global events, e.g. lightnings).

Machine Learning for Gravitational-wave data

Lots of data

Rich, complex signal space

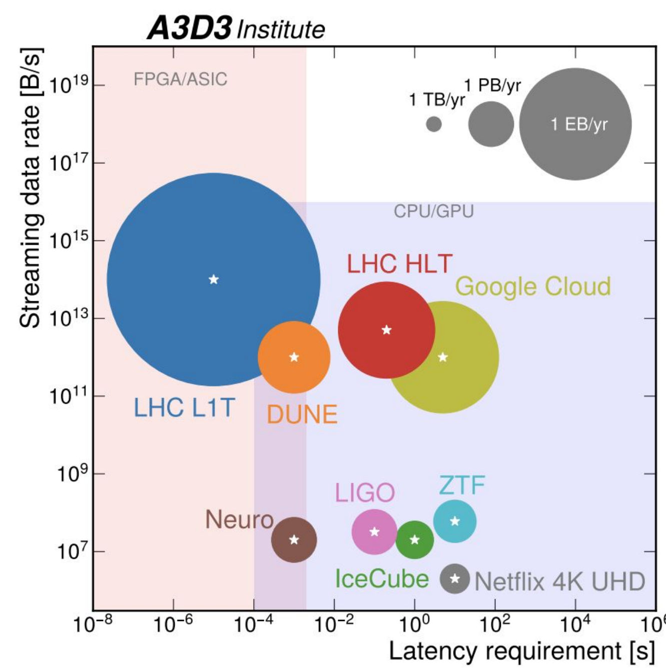
Rich, complex noise space

Low-latency/real-time requirements

→ Neural Networks!

Computing revolution:

- Success of deep learning has led to sophisticated algorithms
- Rise of heterogeneous computing has enabled deep learning
- Developing ML+GPU integration has enabled large throughput computing
- Developing ML+FPGA/ASIC for low latency computing



Accelerated AI Algorithms for Data-Driven Discovery

[FastML for Science]   [A3D3 Institute]

Requirements for ML deployment in GW searches

Training

Load time-series data from disk and efficiently move to GPU

Leverage simulations to create robust datasets

Implement signal processing operations on GPU

Inference

Offline - produce predictions on O(100+years) of background data

Online - produce transient detections on real-time data in O(1s) and estimate parameters in O(1s)

Stream time-series into NN

Heterogeneous computing backends/data-types

Infrastructure design goals

Intuitive - maps on to familiar, physically meaningful concepts

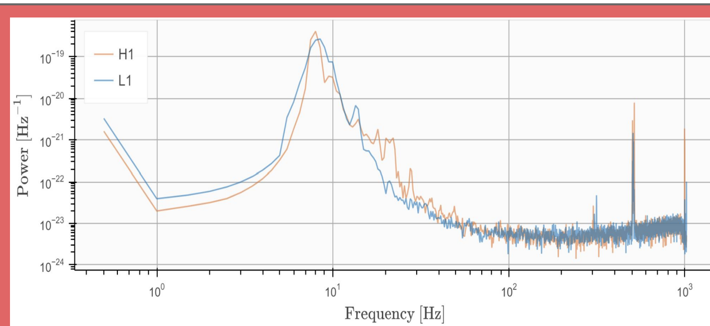
Composable - hierarchical layers of abstraction support new use cases seamlessly

Integrated - ecosystem of tools following same standards and nomenclature

Efficient - make the most out of parallel computing resources

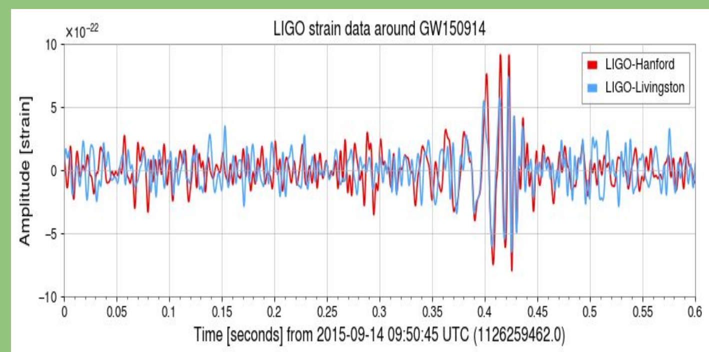
Gravitational-wave data analysis workflows

Data quality: identify and mitigate noise sources (“detector characterization”)



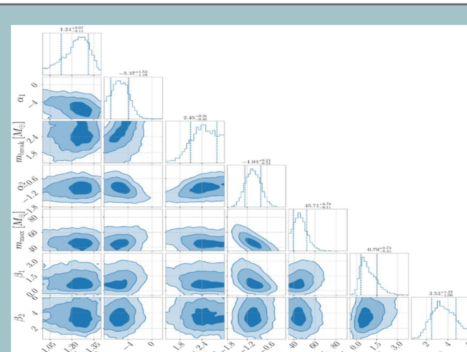
Noise subtraction
nonlinear regression

Detection: identity data instances that stand out statistically as deviating from noise



Modelled transients
supervised
Unmodelled transient
semi-supervised
Vetoes Glitch identification

Inverse problem: extract intrinsic and extrinsic signal/source parameters

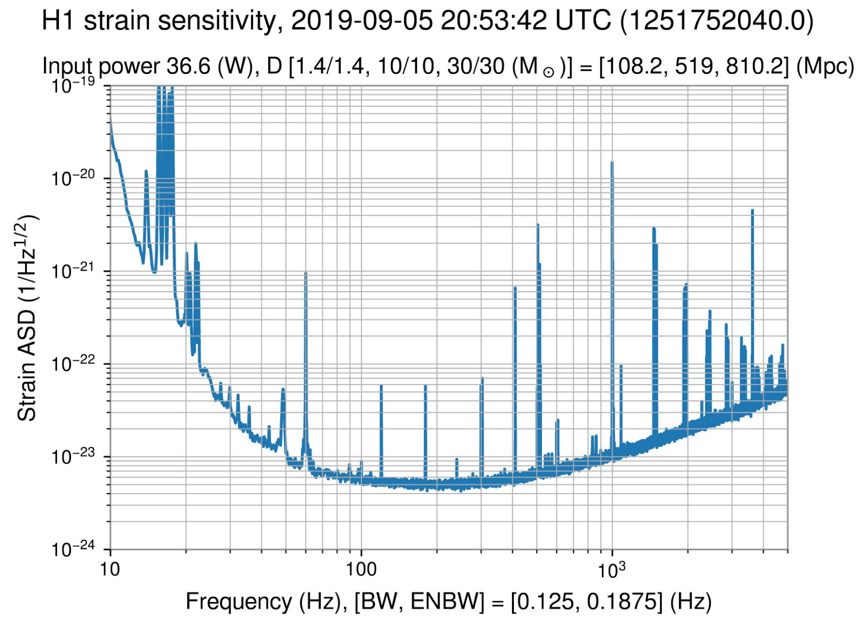


Parameter estimation
Normalizing flows

[ML4GW/HERMES](): a new ecosystem for end-to-end ML-based GW searches

DeepClean: a noise subtraction platform

Non-fundamental noise in interferometers can be subtracted, when such noise is “witnessed” by auxiliary channels:



NON-REMOVABLE (FUNDAMENTAL NOISE)

$$h(t) = s(t) + n(t)$$

$$n(t) = n_{nw}(t) + n_w(t)$$

POSSIBLE GW SIGNAL

DETECTOR NOISE

REMOVABLE WITNESS NOISE

$h(t)$

$\{w_i(t)\}$

$$n_w(t) = \mathscr{F}(\{w_i\}; \vec{\theta})$$

$$r(t) = h(t) - n_w(t)$$

$r(t)$

Residual strain
after noise subtraction

[Ormiston et al. “Noise Reduction in Gravitational-Wave Data via Deep Learning.”]

DeepClean: a noise subtraction platform

Non-fundamental noise in interferometers can be subtracted, when such noise is “witnessed” by auxiliary channels:

$$h(t) = s(t) + n(t)$$

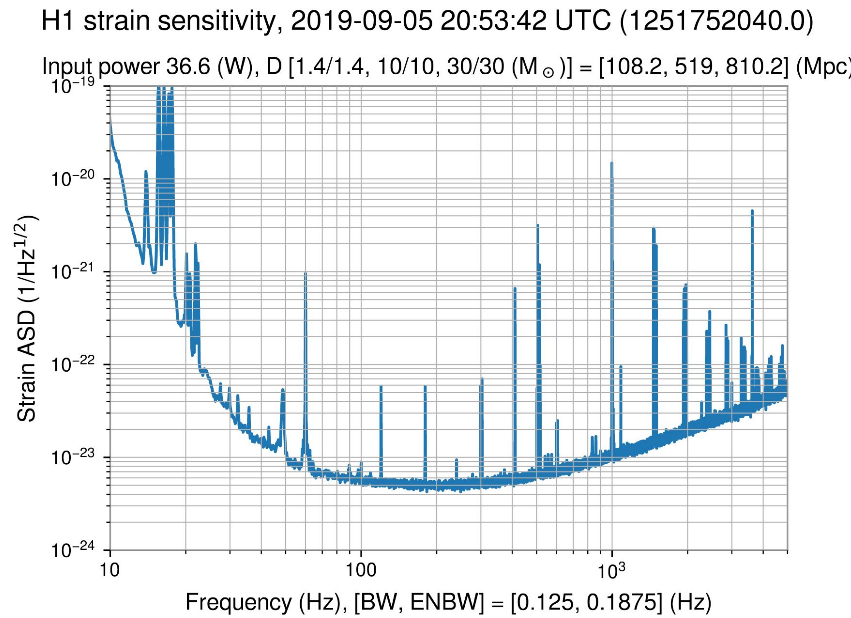
$$n(t) = n_{nw}(t) + n_w(t)$$

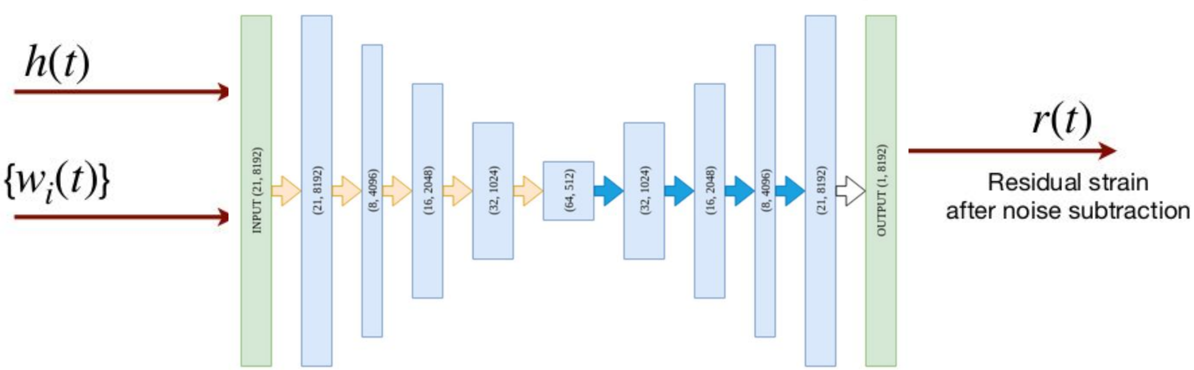
NON-REMOVABLE (FUNDAMENTAL NOISE)

POSSIBLE GW SIGNAL

DETECTOR NOISE

REMOVABLE WITNESS NOISE





Convolutional auto-encoder

Real-time implementation with ~1s latency

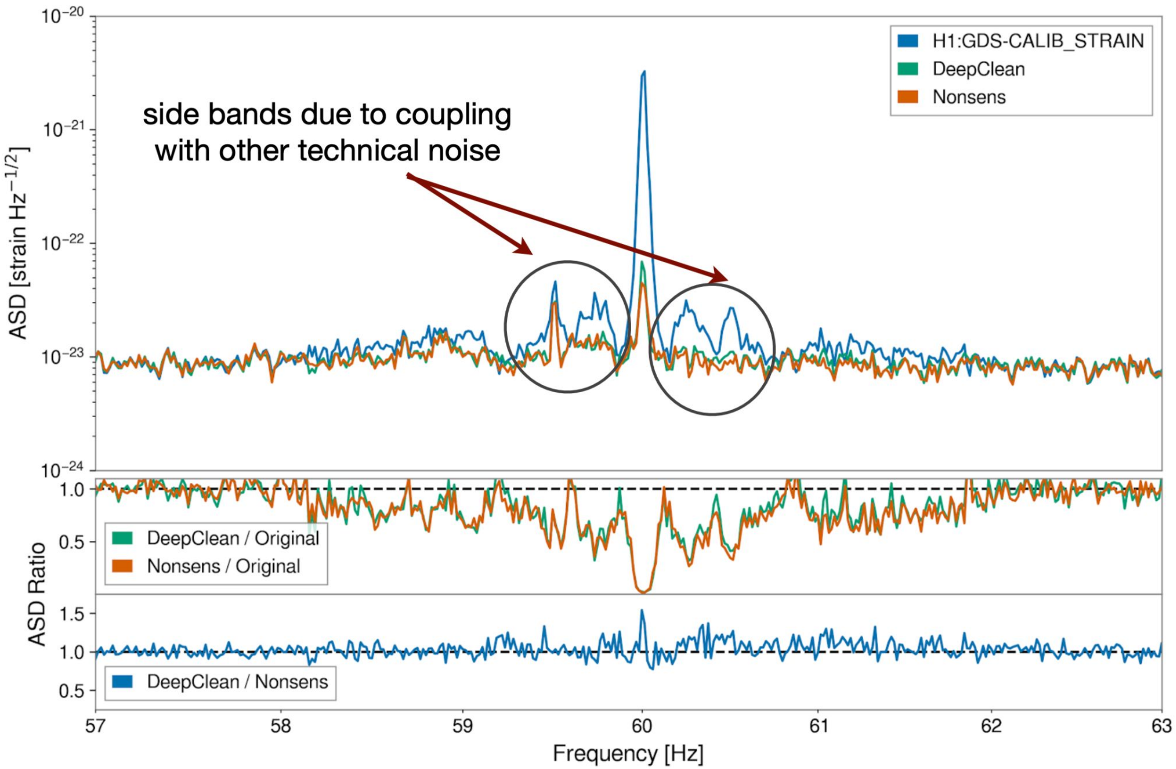
Provided to analyses downstream

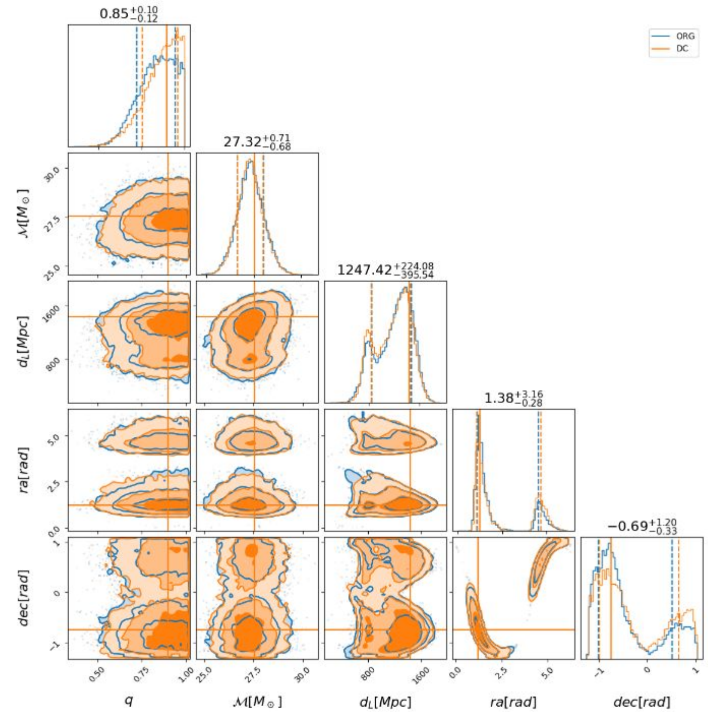
Able to go beyond linear couplings/algorithms

Also implemented on an FPGA

[Ormiston et al. “Noise Reduction in Gravitational-Wave Data via Deep Learning.”]

DeepClean performance in O3: Amplitude Spectral Densities and Parameter Estimation





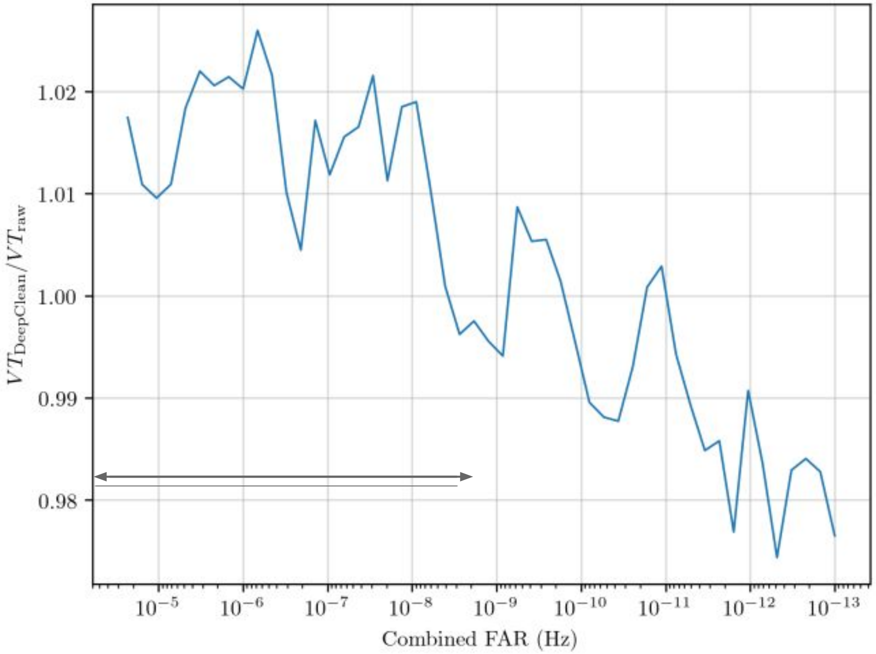
Demonstrated non-linear subtraction on 60 Hz power lines and sidebands!

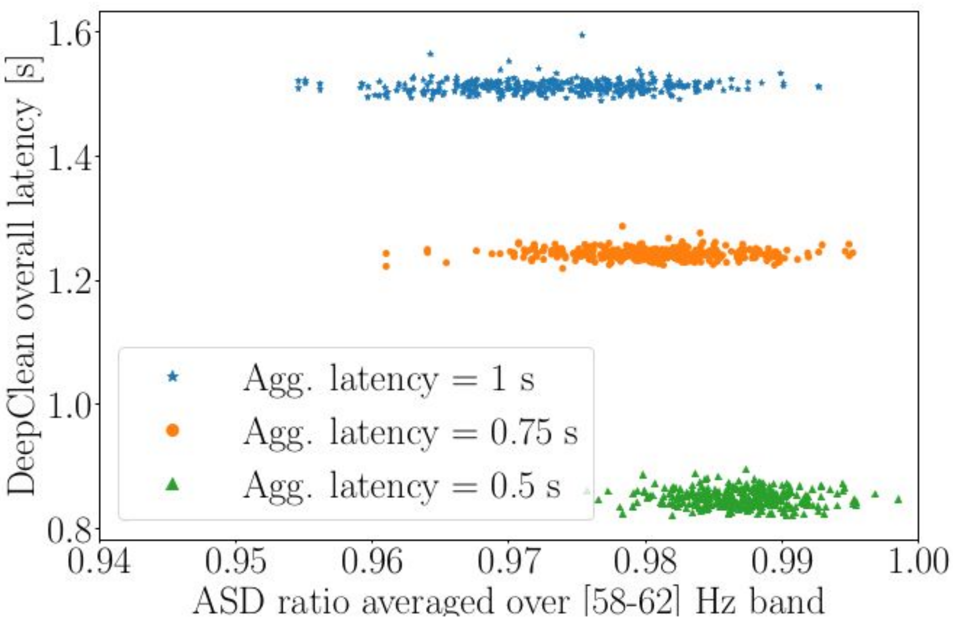
Multiple tests on BBH injections to demonstrate unbiased recovery of astrophysical signals

Ormiston et al. "Noise Reduction in Gravitational-Wave Data via Deep Learning."

Saleem et al (2023). **Demonstration of Machine Learning-assisted real-time noise regression in gravitational wave detectors.**

DeepClean performance in O3: time-volume reach and latencies





Sensitive volume (V*T) fractional gain or loss (with/without DeepClean) as a function of the false alarm rate in a GstLAL search.

Trade-off between latency and quality: the ASD ratio improves with higher aggregation latency, at the cost of increased overall latency.

Saleem et al (2023), **Demonstration of Machine Learning-assisted real-time noise regression in gravitational wave detectors**,

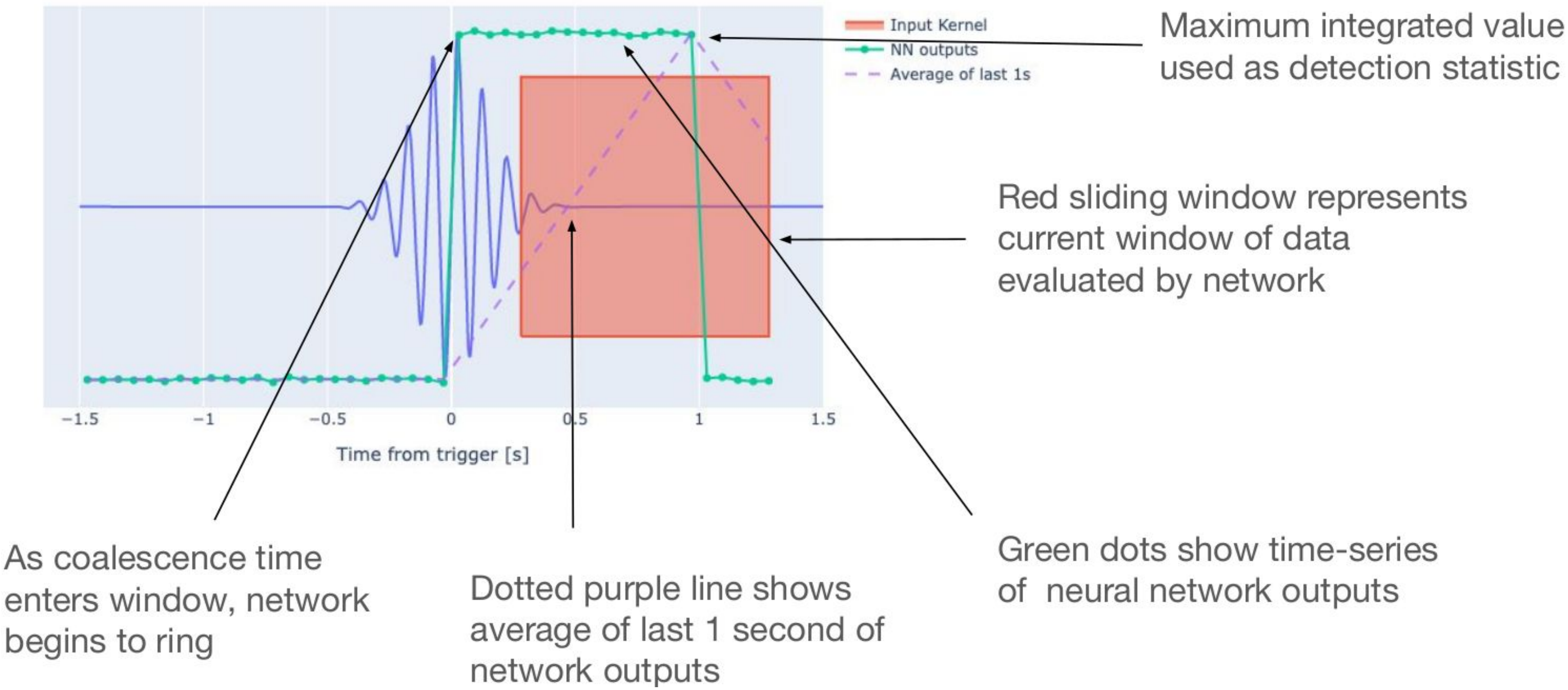
Aframe: detecting compact binary coalescences

Residual Network architecture trained to map 1.5 second windows of h(t) time-series data to scalar value that indicates likelihood of signal being present in the window.

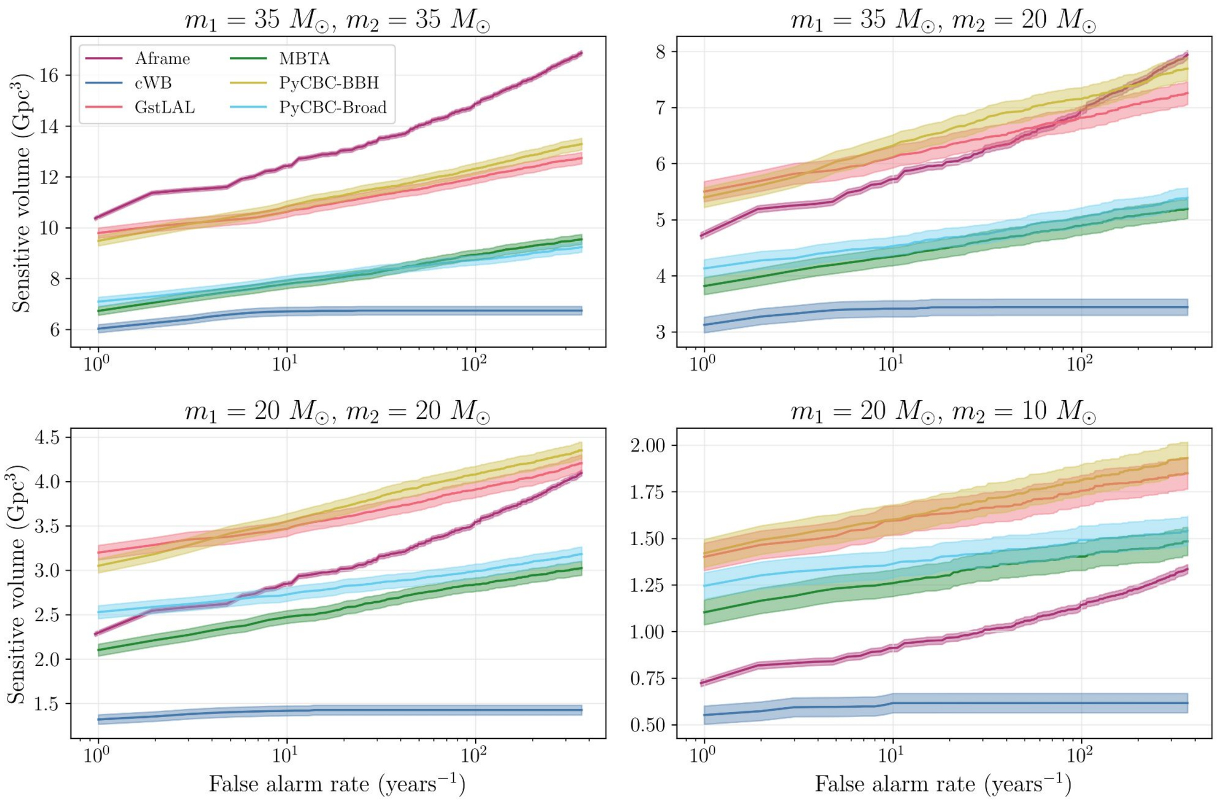
Training on 10 days of coincident H1 and L1 strain from beginning of LIGO-Virgo-KAGRA's O3a run; 100,000 IMRPhenomPv2 waveforms from astrophysical prior used for training search space: 5 - 100 M_solar.

Extensive data augmentation to show the model as diverse a training set as possible.

Event identification:



Aframe: pipeline sensitivity



Sensitive volume calculation is the same as the one used to measure performance of LVK pipelines in GWTC-3 catalog

Competitive performance on higher-mass catalog distributions

Work remains to be done for lower masses – alternative architectures or smarter training techniques

https://zenodo.org/records/7890437

E. Marx, W. Benoit et al (2024), **A machine-learning pipeline for real-time detection of gravitational waves from compact binary coalescences**

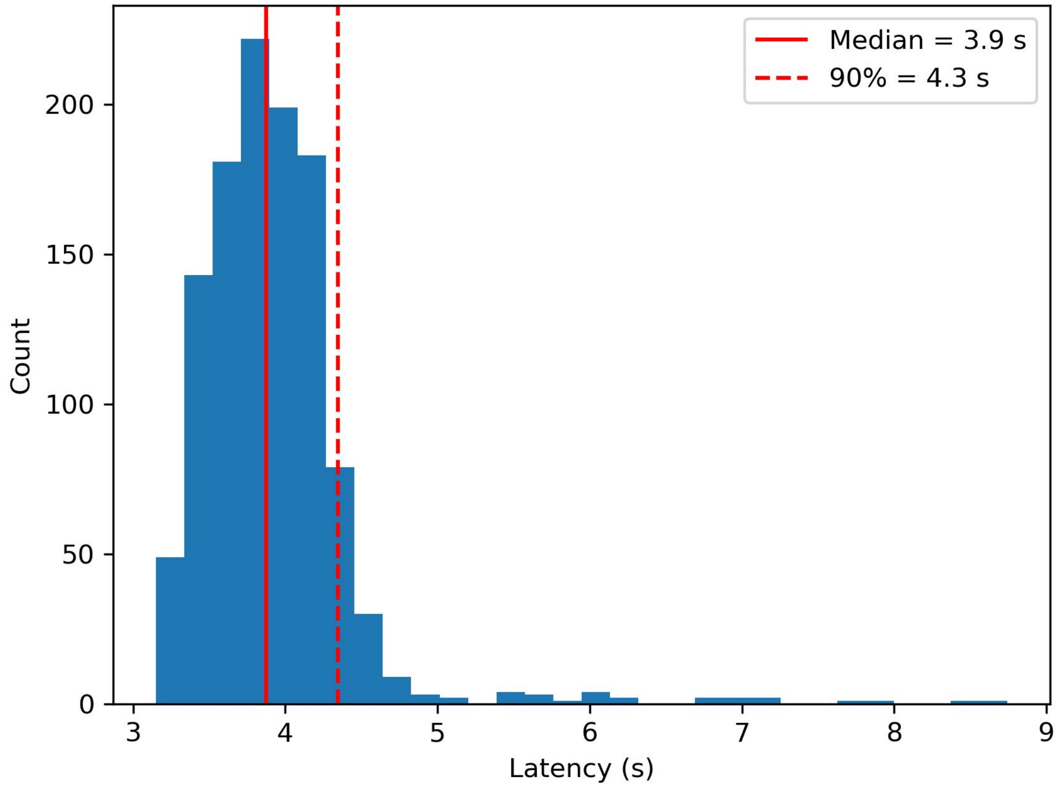
Computational cost and throughput

Training: **~44 hours** on 1 GPU on a 16GB V100 GPU

Evaluation: Analysis of O3 took ~30 hours for 21 years of livetime on 8 GPUs, roughly **750** seconds of data processed per second per GPU

Total time scales roughly linearly with number of GPUs

Ran pipeline online over ~1 month of “replayed” O3 data emulating real-time environment: median (90%) latency is 8.4 (37.1) seconds faster than rest of CBC pipelines (as reported here: https://arxiv.org/abs/2308.04545)



AMPLFI: from detections to astrophysics

Source parameter estimation:

$$\underbrace{d_i(t)}_{\text{Detector output}} = \underbrace{n_i(t)}_{\text{noise}} + \underbrace{F_{+,i}}_{\text{Antenna factor}} \underbrace{h_+(t;\theta)}_{\text{Our goal (its parameters)}} + \underbrace{F_{\times,i}}_{\text{Antenna factor}} \underbrace{h_\times(t;\theta)}_{\text{Our goal (its parameters)}}$$

Calculate posteriors via likelihood estimation:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{L(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z} = \frac{L(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Likelihood-Free (simulation-based) Inference: train neural network to estimate posteriors $p(\boldsymbol{\theta}|\mathbf{d})$
Model true posterior distribution with a normalizing flow:

$$p(\theta|d) \sim q_\phi(\theta|d) = p_u(T_d^{-1}(\theta))|det J_{T_d^{-1}}(\theta)|$$

G Papamakarios et al, JMLR Vol. 22, Art. 57 2617-26 (2021)

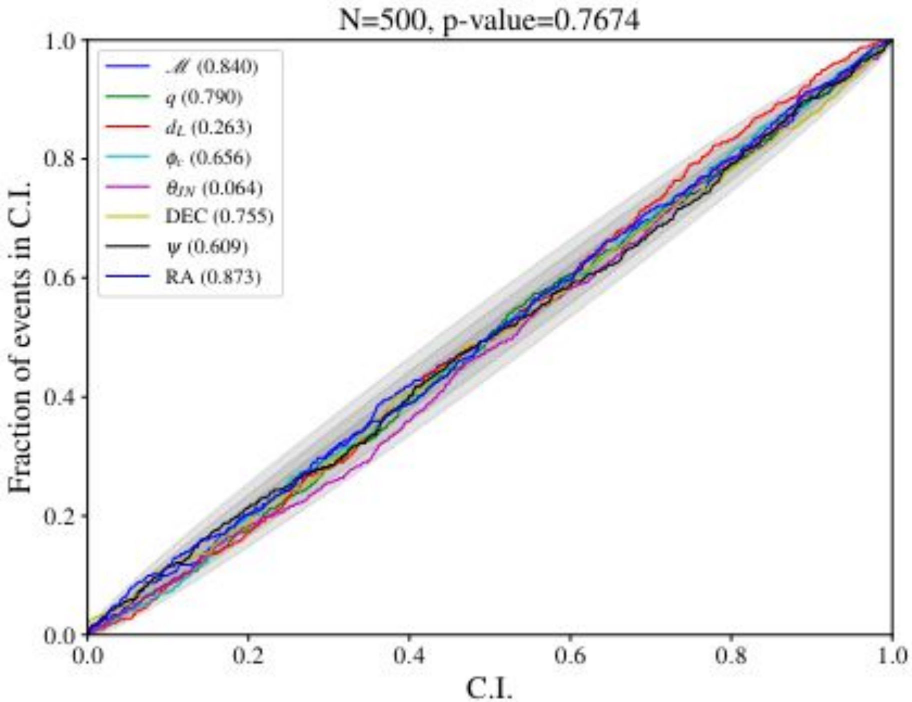
The flow is parameterized via a neural network, and trained by minimizing the Kullback-Leibler divergence

$$L \approx -\frac{1}{N}\sum_{i=1}^{N} \log q_\phi(\theta^{(i)}|d)$$

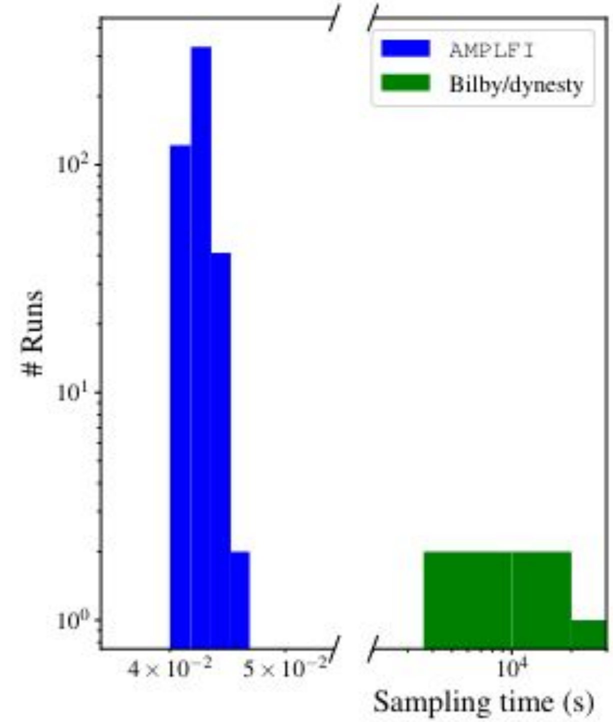
Train neural network to estimate posteriors for any assumed signal morphology (e.g. sine-Gaussian, binary coalescences etc) embedded in <u>real instrument noise</u> and that's how you arrive at **<u>AMPLFI: Accelerated Multi-messenger Parameter-estimation using LFI !</u>**

Deep Chatterjee, Ethan Marx et al, submitted for publication (2024)

AMPLFI performance



P-P plot from inferences with AMPLFI over 500 Binary Black Hole systems



Sampling times for AMPLFI (1 GPU) vs. nested sampling runs (24 CPUs)

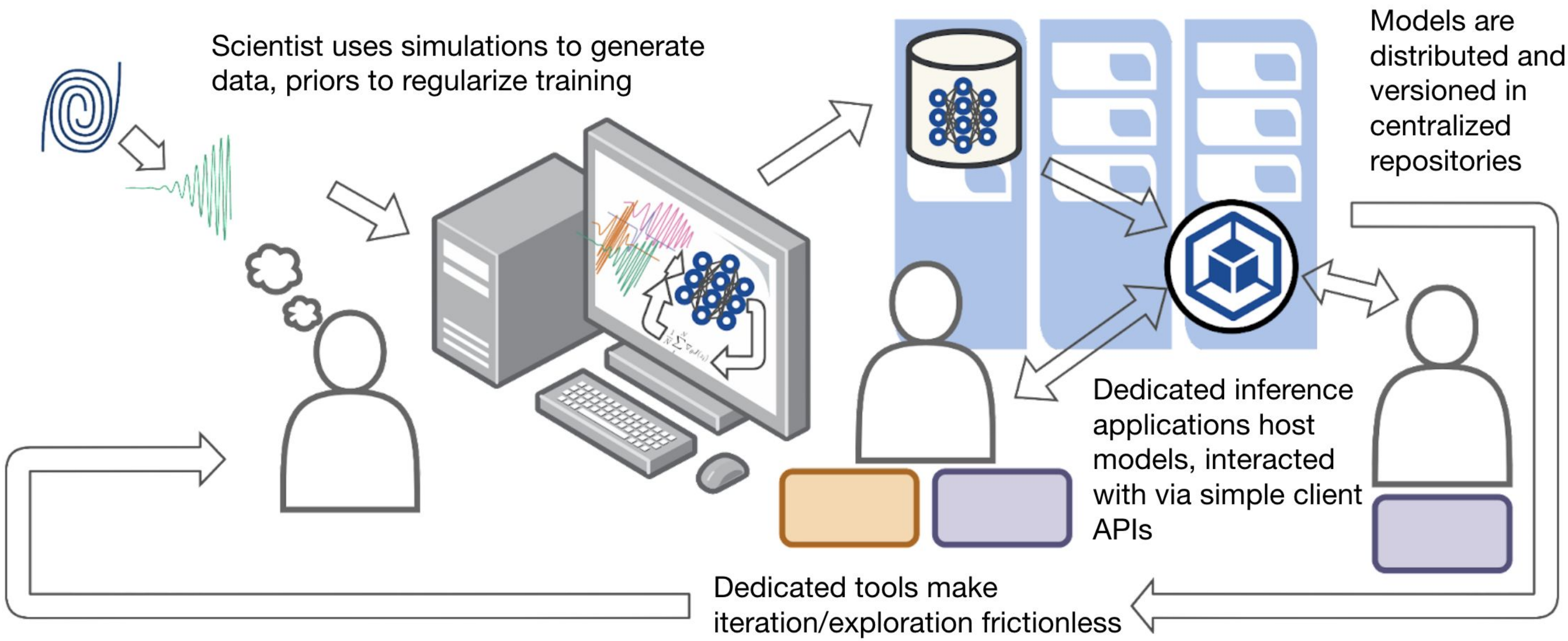
Deep Chatterjee, Ethan Marx et al, submitted for publication (2024)

Putting all these together: “Who will bell the cat?”

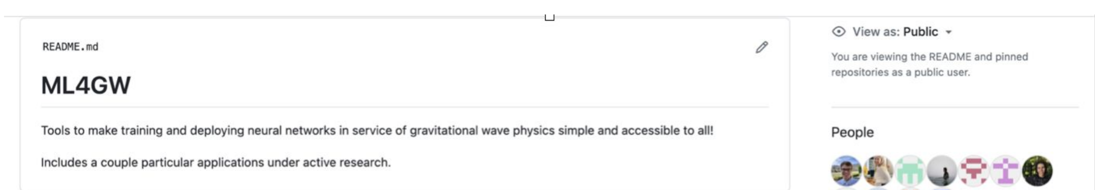
Cartoon adopted from Alec Gunny

`ml4gw/HERMES`: https://github.com/ML4GW

Bringing AI into GW and MMA data analyses



[ML4GW/HERMES](): an ecosystem for ML applications in GW enabling fast deployment, fast inference, small computation footprint and optimized for computing heterogeneity

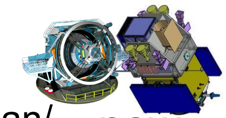


Hardware-accelerated Inference for Real-Time Gravitational-Wave Astronomy
Alec Gunny et al Nature Astronomy (2022)

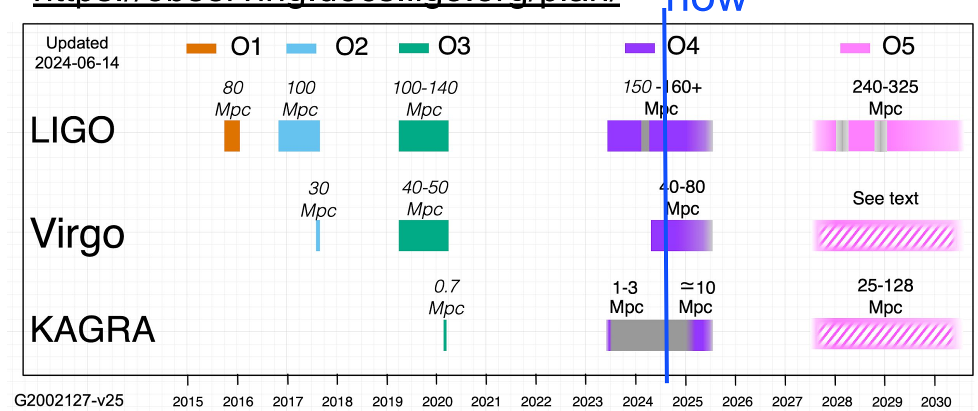
Summary and outlook

- `ml4gw`: A new computing ecosystem for ML applications in GW data analyses
- Emphasis on real-time processing for improving multi-messenger prospects of the GW observatories
  - focus on latency
  - minimal computational footprint (a
  - couple of GPUs to keep up with
real-time)
- Offline implementation
  - portable, robust pipelines
  - emphasis on throughput
  - extensible
- End-to-end ML-based workflows have been implemented addressing:
  - data cleaning
  - transient event detections
  - parameter estimation
- Expecting to deploy in production for real-time use during LIGO-Virgo-KAGRA's current O4 observing run

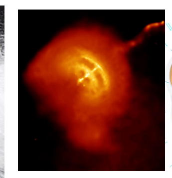
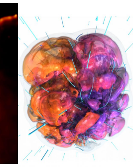
https://www.svom.eu/
https://rubinobservatory.org/



https://observing.docs.ligo.org/plan/



BBH populations **Known Unknowns** **Unknown Unknowns**

???

Summary and outlook

`ml4gw`: A new computing ecosystem for ML applications in GW data analyses

Emphasis on real-time processing for improving multi-messenger prospects of the GW observatories

- focus on latency
- minimal computational footprint (a couple of GPUs to keep up with real-time)

Offline implementation

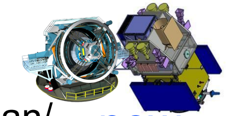
- portable, robust pipelines
- emphasis on throughput
- extensible

End-to-end ML-based workflows have been implemented addressing:

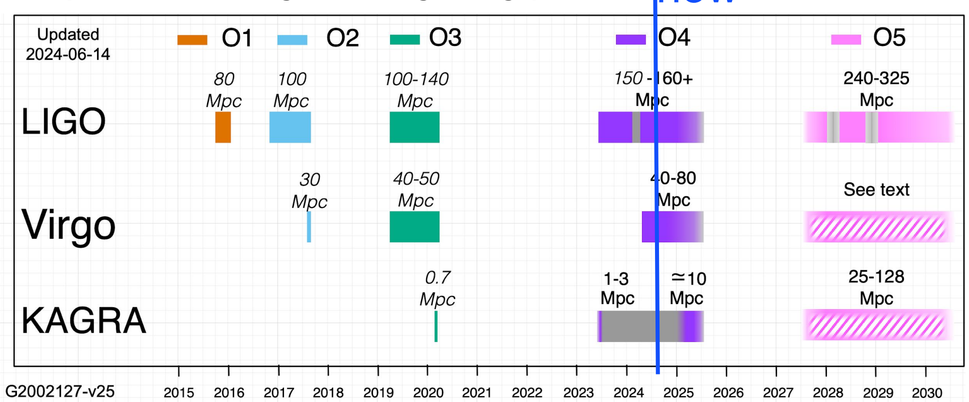
- data cleaning
- transient event detections
- parameter estimation

Expecting to deploy in production for real-time use during LIGO-Virgo-KAGRA's current O4 observing run

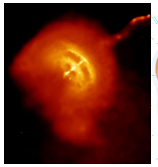
https://www.svom.eu/
https://rubinobservatory.org/

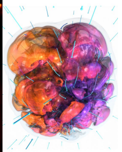


https://observing.docs.ligo.org/plan/



BBH populations **Known Unknowns** **Unknown Unknowns**





???

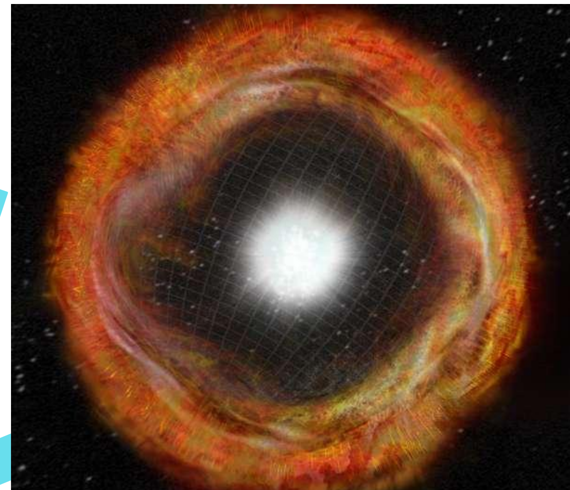
Come join!

https://github.com/ML4GW

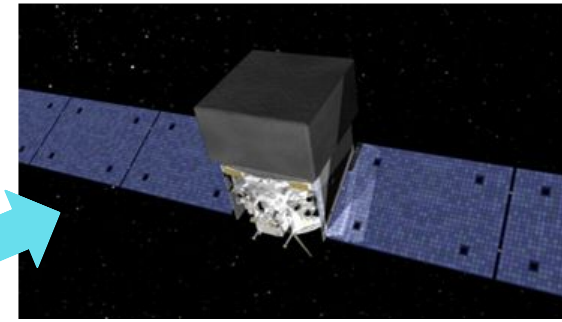
EXTRA SLIDES

Multi-Messenger Astrophysics

Image credit: Bill Saxton, NRAO



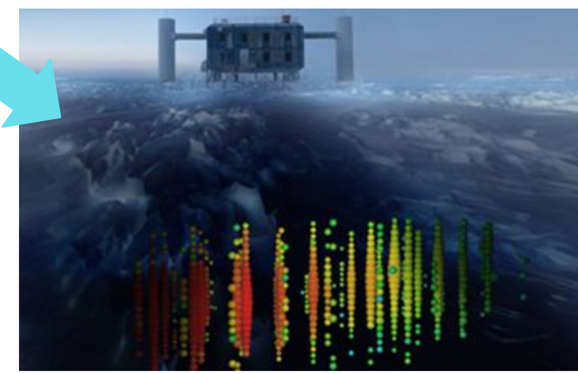
Gravitational waves



X-rays/Gamma-rays

Visible/infrared light

Radio waves

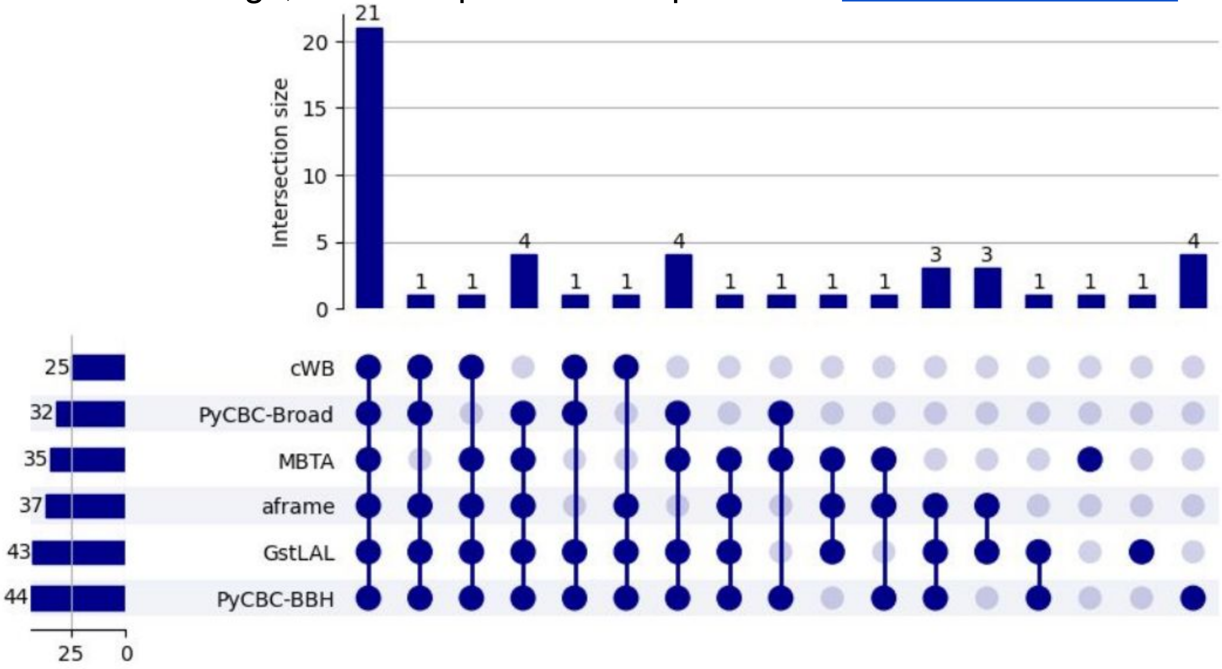


Neutrinos

Comparison to LVK's O3 detections

Reanalyzing LVK's O3 observing run, 37/50 candidates detected by Aframe at false alarm threshold of 1 per 5 months; missed events have network matched filter SNR<13.1 or chirp mass < 10 M_solar. The latter is consistent with Aframe's sensitive volume measurements.

Also the next 10 most significant Aframe events have no overlaps with the LVK catalogs, but have partial overlap with the [Olsen et al event list]



gpstime	FAR (1 / yr)
1262635012.75	3.6
1246523564.75	4.0
1264333383.00	4.2
1238351045.00	4.4
1251010355.50	4.7
1264246793.25	5.9
1262163593.25	7.8
1249032684.75	11.0
1253452013.50	11.7
1259411705.25	12.0

[E. Marx, W. Benoit et al, in preparation]

GWAK: an anomaly detection framework

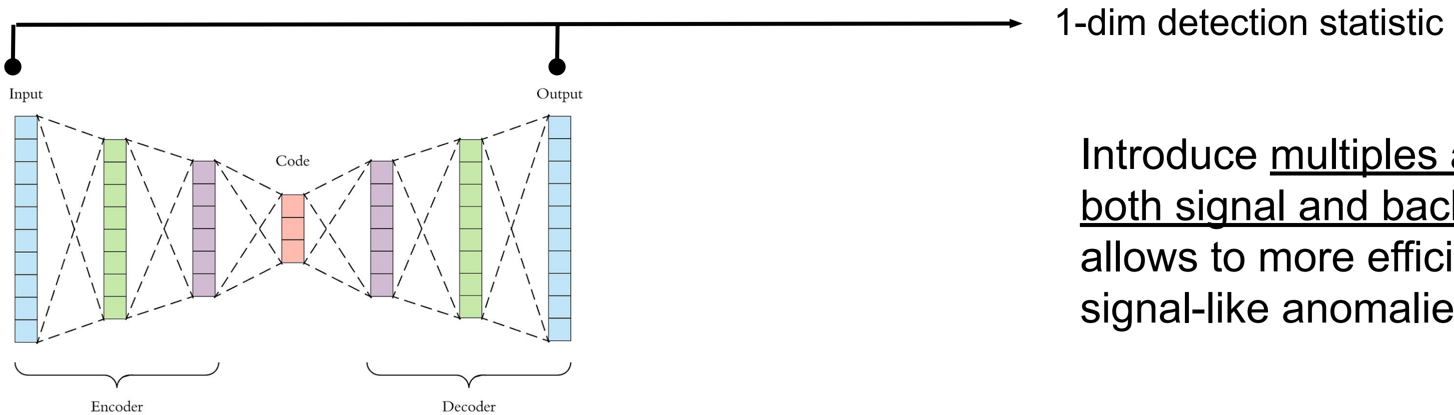
Strong astrophysical motivation to look beyond modeled binary coalescences: supernova, neutron star glitches, magnetars, GRBs, FRBs, cosmic strings and cusps, unknown unknowns may emit GWs that we can not fully modeled currently and thus can not be searched with a matched-filter approach

We refer to them as anomalous and aim to develop a semi-supervised approach which would let us to discover such anomalous signals without explicit modeling

- use multiple autoencoders to create embedded space
- use real background and inject signals
- verify on anomalous signals that aren't included in training

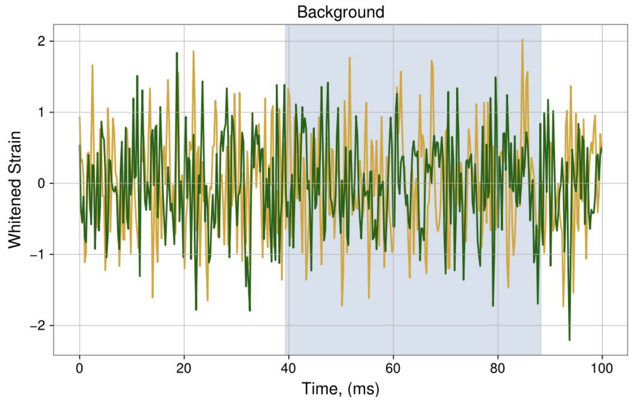
GWAK is the Gravitational Wave Anomalous Knowledge, an algorithm using recurrent autoencoders inspired by similar approaches ([QUasi Anomalous Knowledge, by Sang Eon Park et al. https://arxiv.org/abs/2011.03550)](https://arxiv.org/abs/2011.03550) taken in performing anomaly detection in LHC data

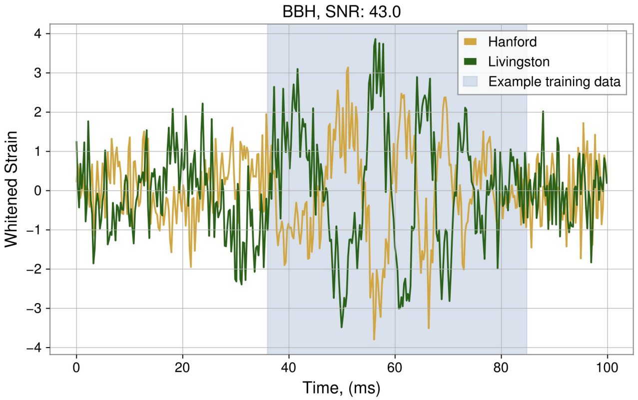
Core idea: go beyond vanilla anomaly detection in 1-dimensional approach where the distance between the input and output is used as a metric for anomaly detection:

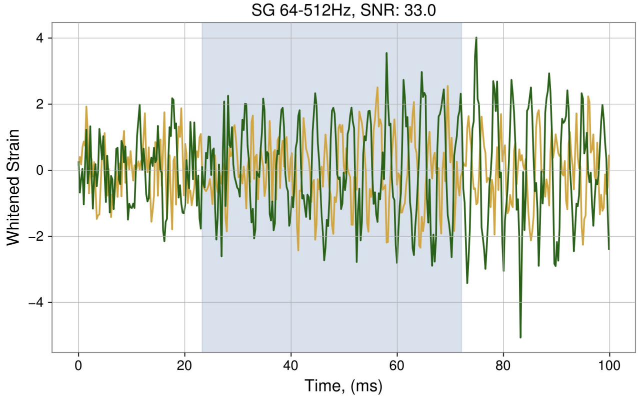


Introduce <u>multiples axes, for both signal and background</u> ⇒ allows to more efficiently select signal-like anomalies

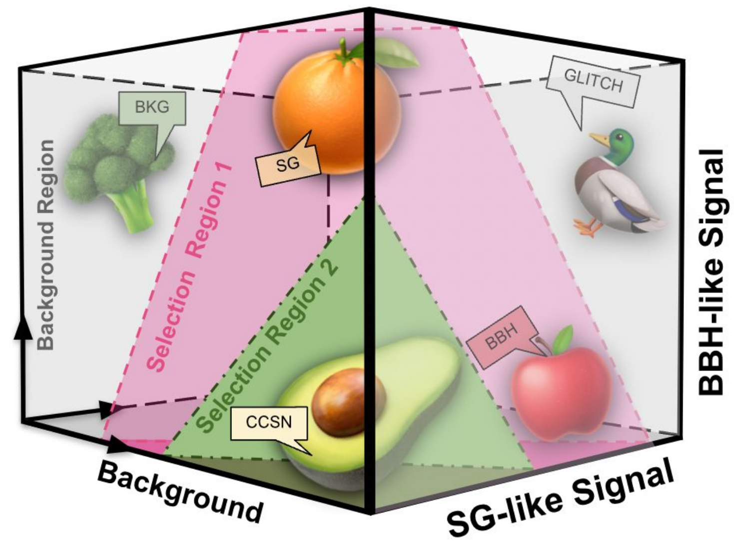
GWAK: multi-dimensional approach





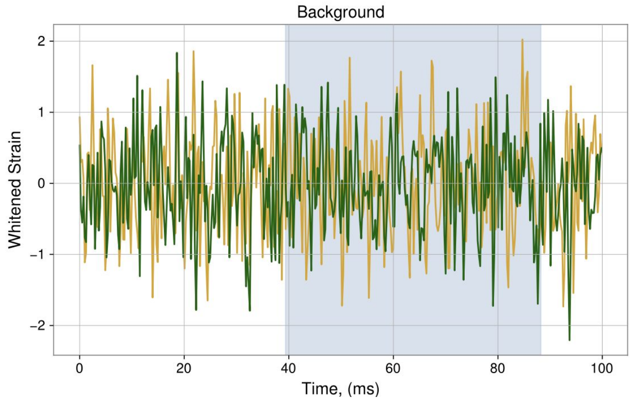


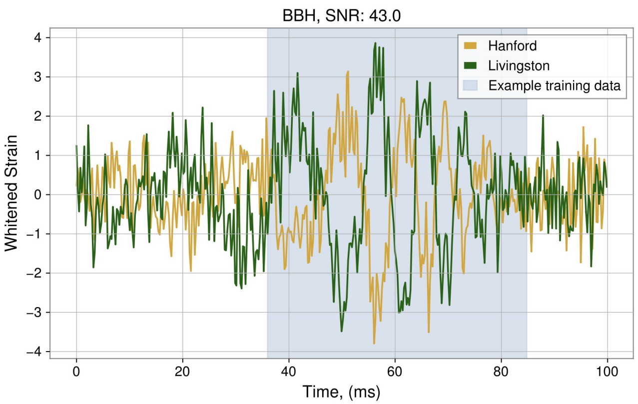
A 3-dim GWAK space example



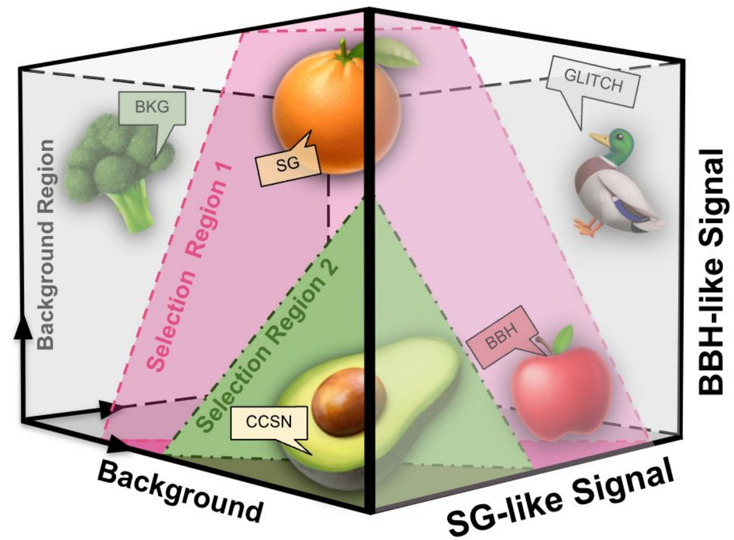
Raikman et al (2023) , **GWAK: Gravitational-Wave Anomalous Knowledge with Recurrent Autoencoders**, https://arxiv.org/abs/2309.11537

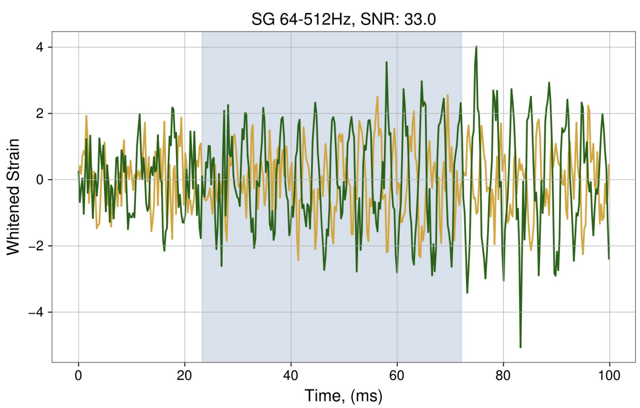
GWAK: multi-dimensional approach

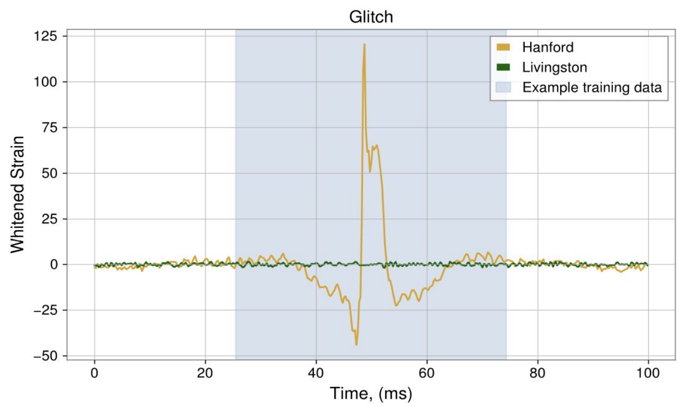


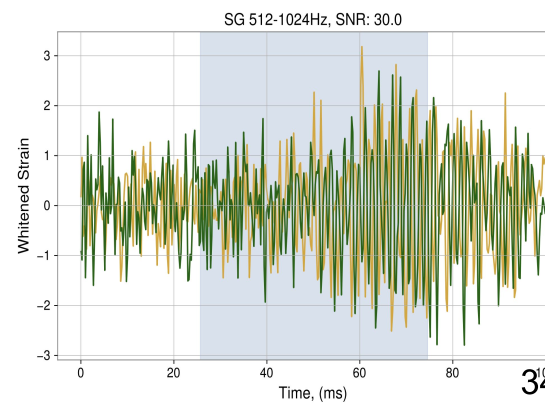


A 3-dim GWAK space example









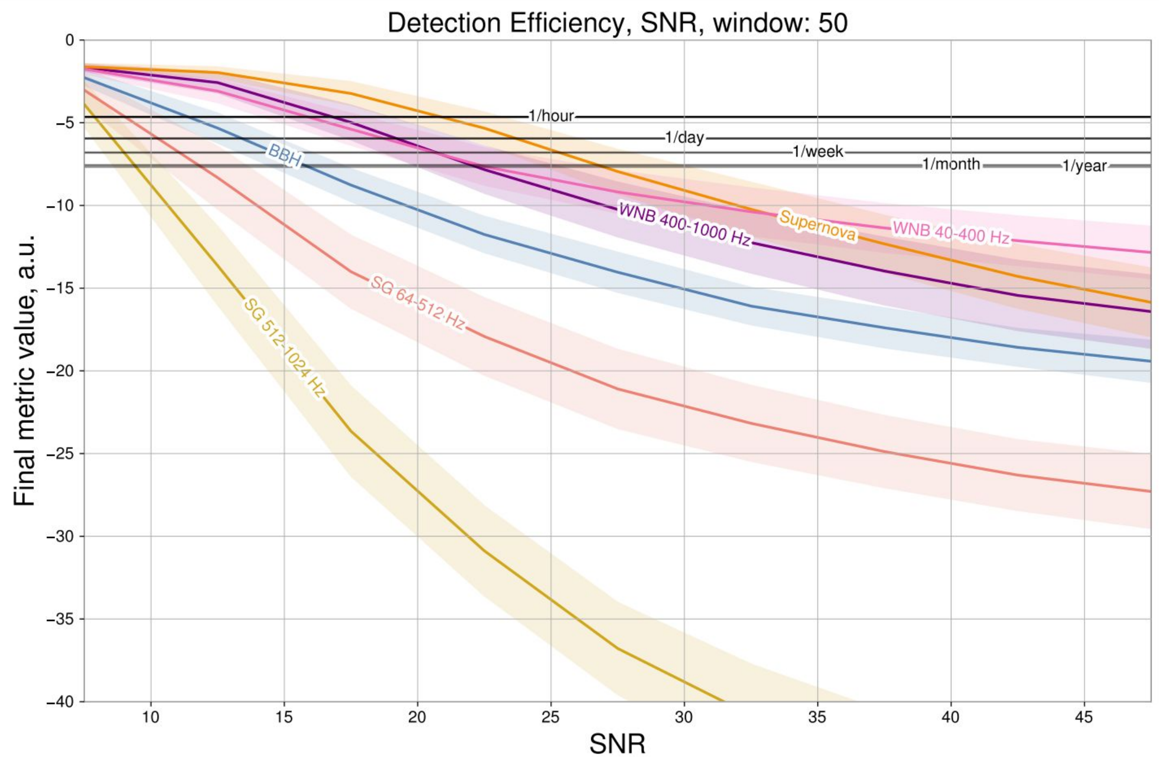
GWAK efficiency during O3

The final metric as a function of SNR for GWAK axes training signals, BBH (blue), SG 64-512 Hz (yellow), SG 512-1024 Hz (salmon)

and for potential (unseen) anomalies, WNB 40-400 Hz (pink), WNB 400-1000 Hz (purple), and Supernova (orange)

The black lines of varied width correspond to different FARs, from the FAR of 1 per hour to 1 per year

For each of the lines, the events that are below that line would be detected.



Raikman et al (2023) , **GWAK: Gravitational-Wave Anomalous Knowledge with Recurrent Autoencoders**, https://arxiv.org/abs/2309.11537

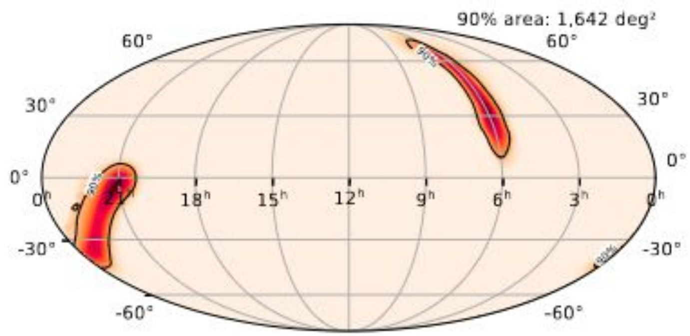
Low latency sky localization comparisons on O3 alerts

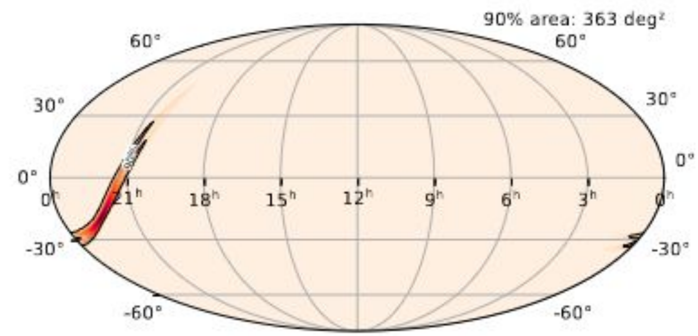
AMPLFI (2-LIGO detector network)

Bayestar (2-LIGO detector network)

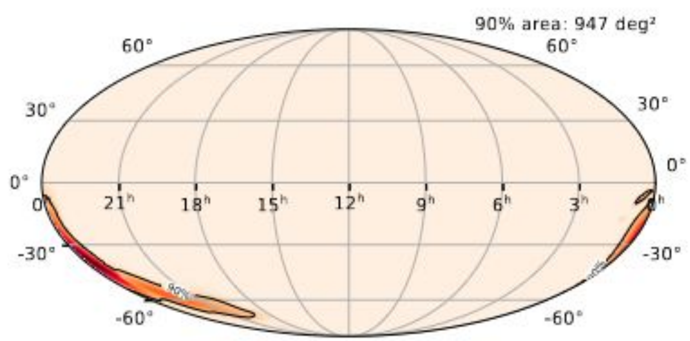
Singer & Price PRD 93, 024013 (2016)

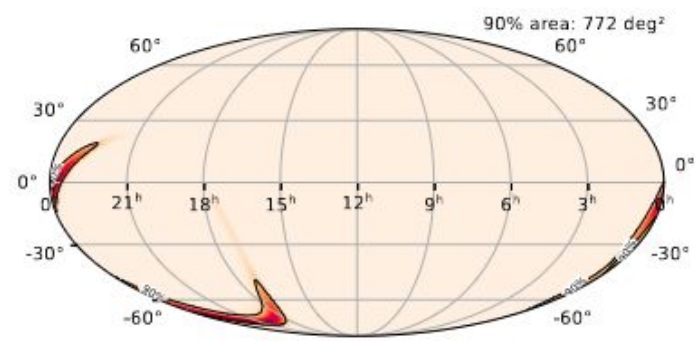
S200129m





S200311bg





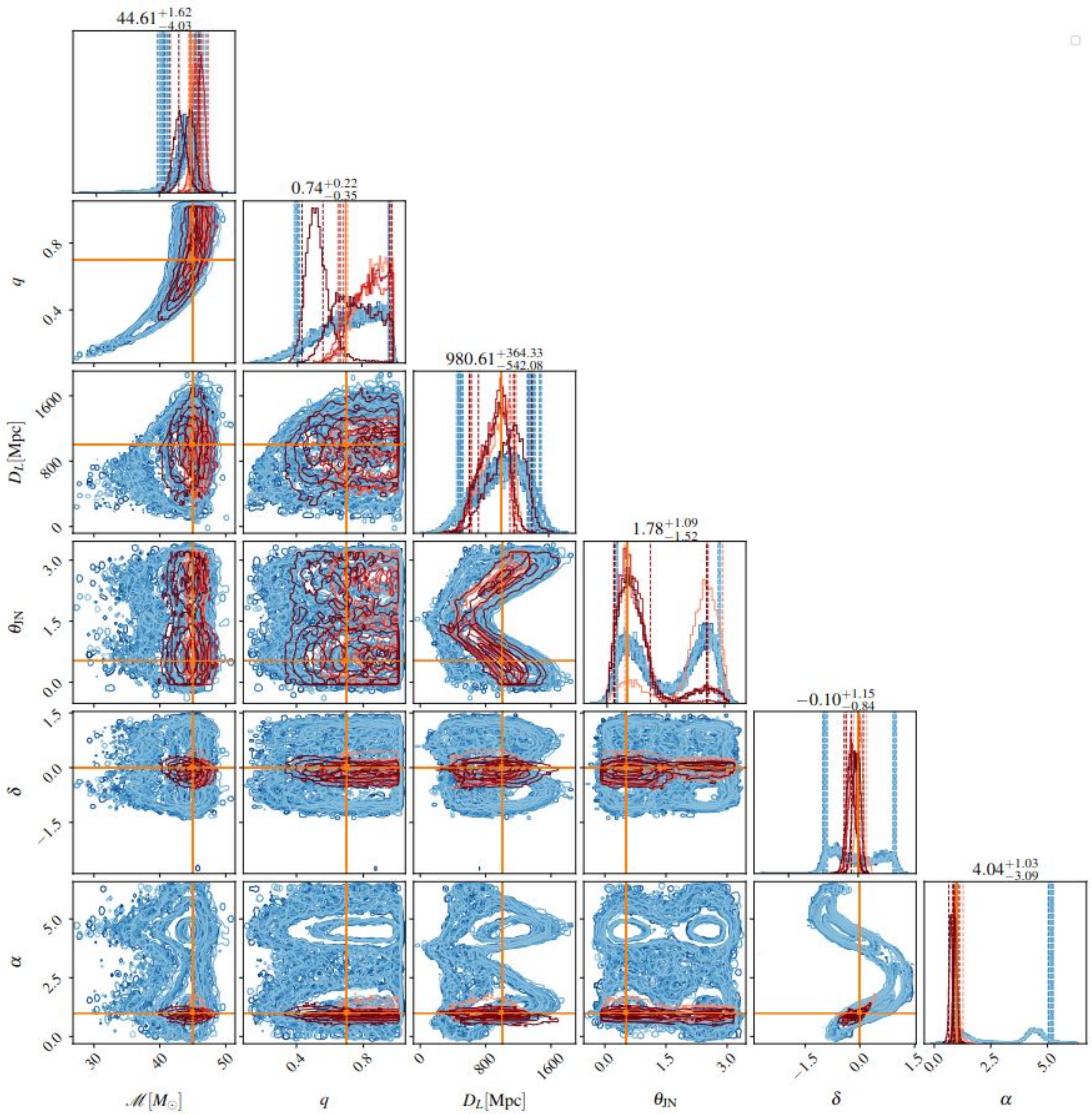
Deep Chatterjee, Ethan Marx et al, submitted for publication (2024)

Sample parameter estimation result

Injected signal: Mchirp = 45 solar masses, q=0.7, D_L = 1 Gpc and optimal SNR~20 added on 20 different background segments

Posterior samples: AMPLFI in blue, Bilby/Dynesty [Ashton et al. ApJS 241, 27 (2019)] in red

Parameter recovery is consistent with injections and stochastic samplers, although posterior widths are tighter with the later



Deep Chatterjee, Ethan Marx et al, submitted for publication (2024)

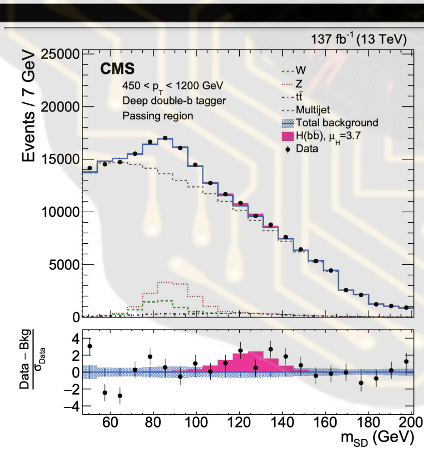
The A3D3 Institute (www.a3d3.ai)

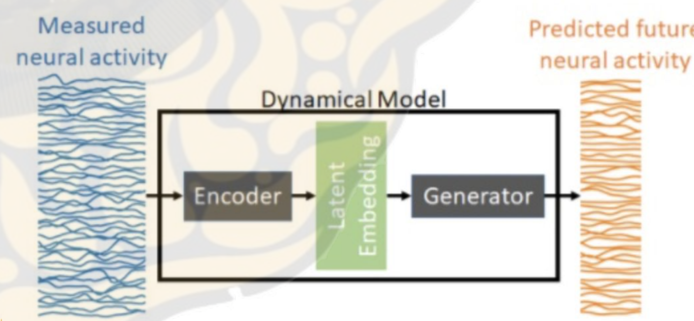
Accelerated AI Algorithms for Data Driven Discovery

Explore real-time AI in MMA, HEP and NeuroScience

New Types of Computing

LHC Physics





NeuroScience