Computing and Data requirements from data analysis

Ed Porter
APC/CNRS
ET-PP / EIB Workshop,
Geneva, 26-27 Oct. 2023

ET EINSTEIN TELESCOPE

With thanks to…

Stefano Bagnasco, Tito Dal Canton, Frederique Marion, Benoit Mours, Viola Sordini

Outline

- GW data analysis Primer
- Computing requirements LVK O1 - O3
- Open questions about computing requirements for ET
- conclusion

GW Data Analysis Primer

GW Source Types

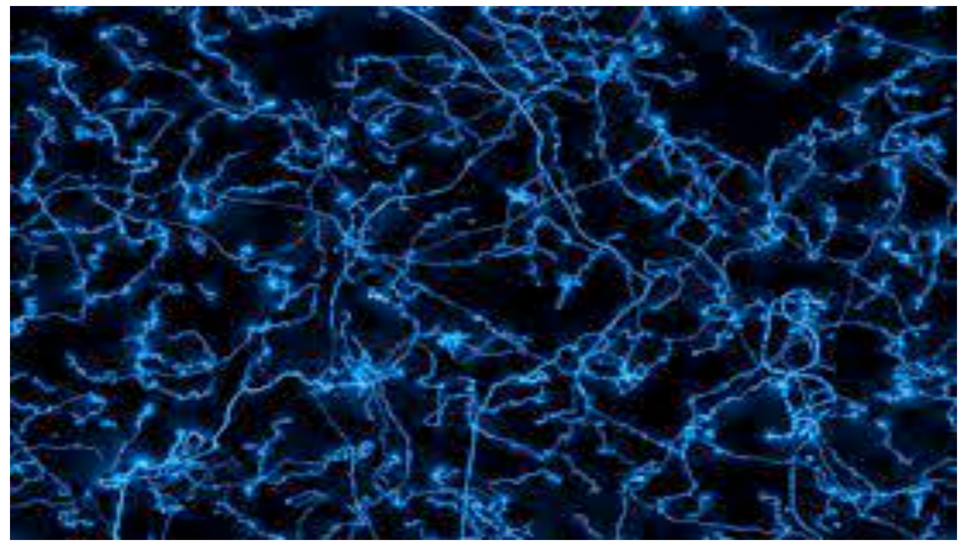
Compact binary coalescence (CBC)

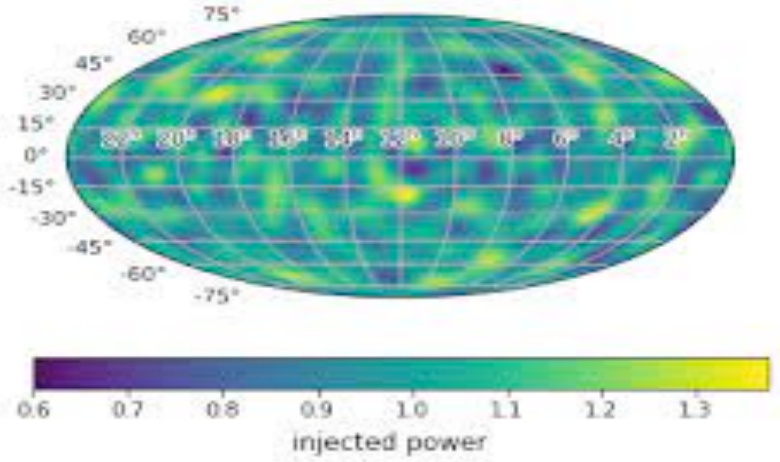
Core collapse SN

Cosmic Strings

Isolated NSs/Pulsars

Stochastic GW Background



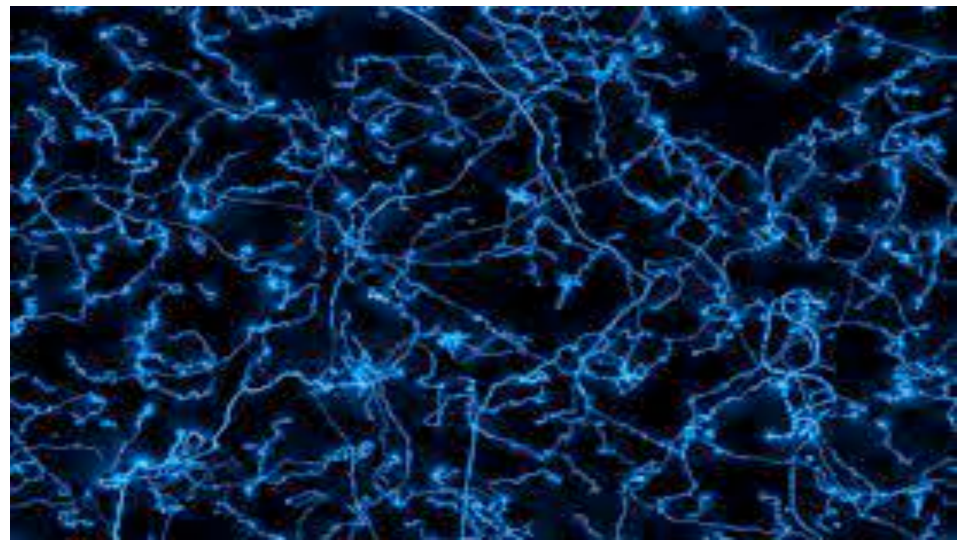


- Primary sources for 2G detectors
- Will also be the primary sources for 3G

GW Source Types

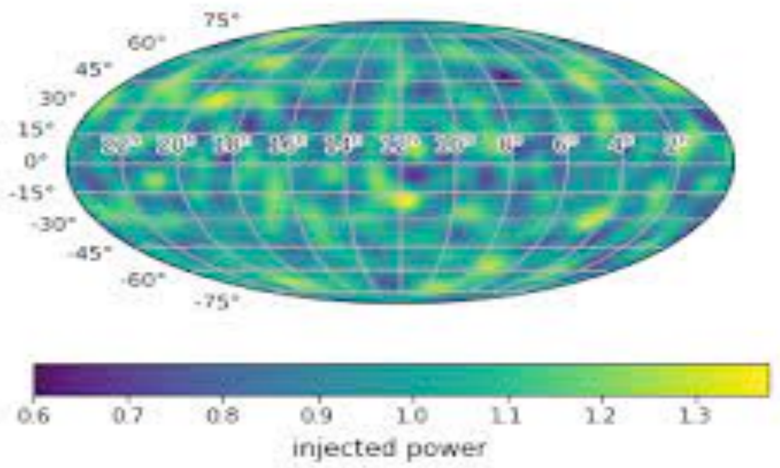
CBC

Burst



CW

Stochastic



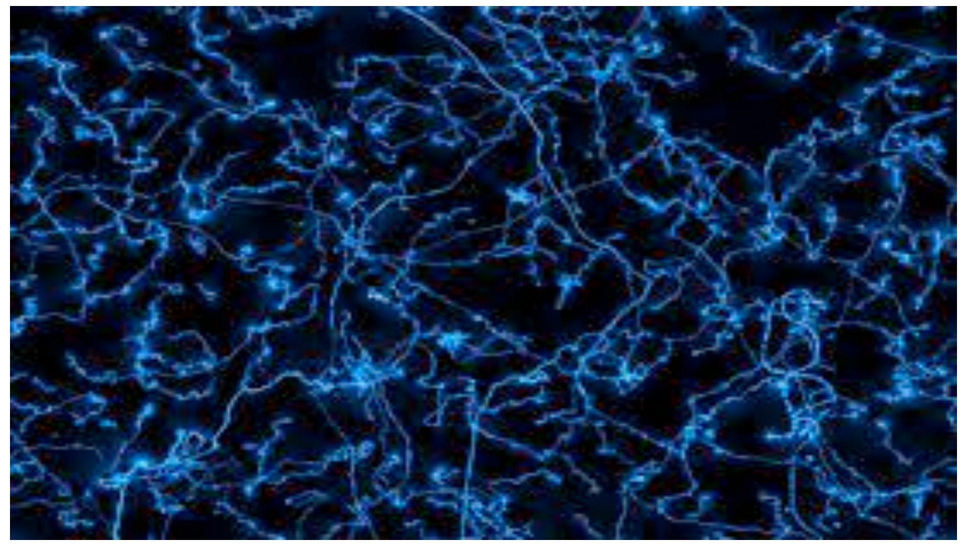
Primary Science Drivers for 2G/3G

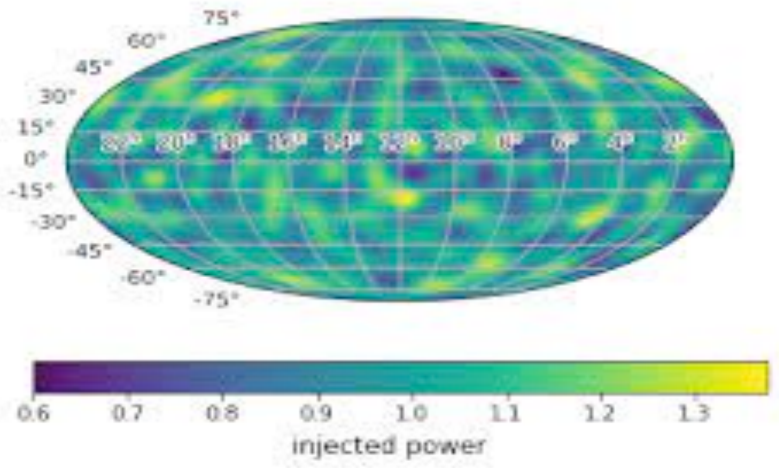
- BBHs/IMBHs
  - populations, tests of GR and fundamental physics, cosmology
- BNS/NSBH
  - populations, tests of GR and fundamental physics, EM/neutrino counterparts, nuclear equation of state, cosmology
- CCSN
  - 1st detection. Coincident EM/neutrino counterpart
- CW
  - 1st detection. Ellipticity of isolated NSs, nuclear equation of state
- Stochastic background
  - proof of an astrophysical/cosmological background of GW, GW energy density
- Cosmological phase transitions / Primordial BHs
  - validation of theoretical predictions of early universe cosmology

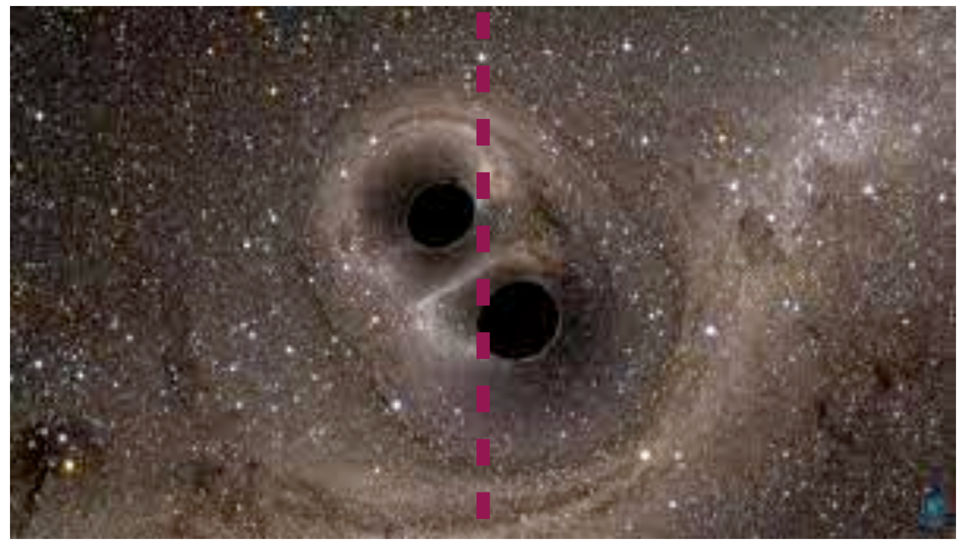
GW Source Types

Modelled

Unmodelled





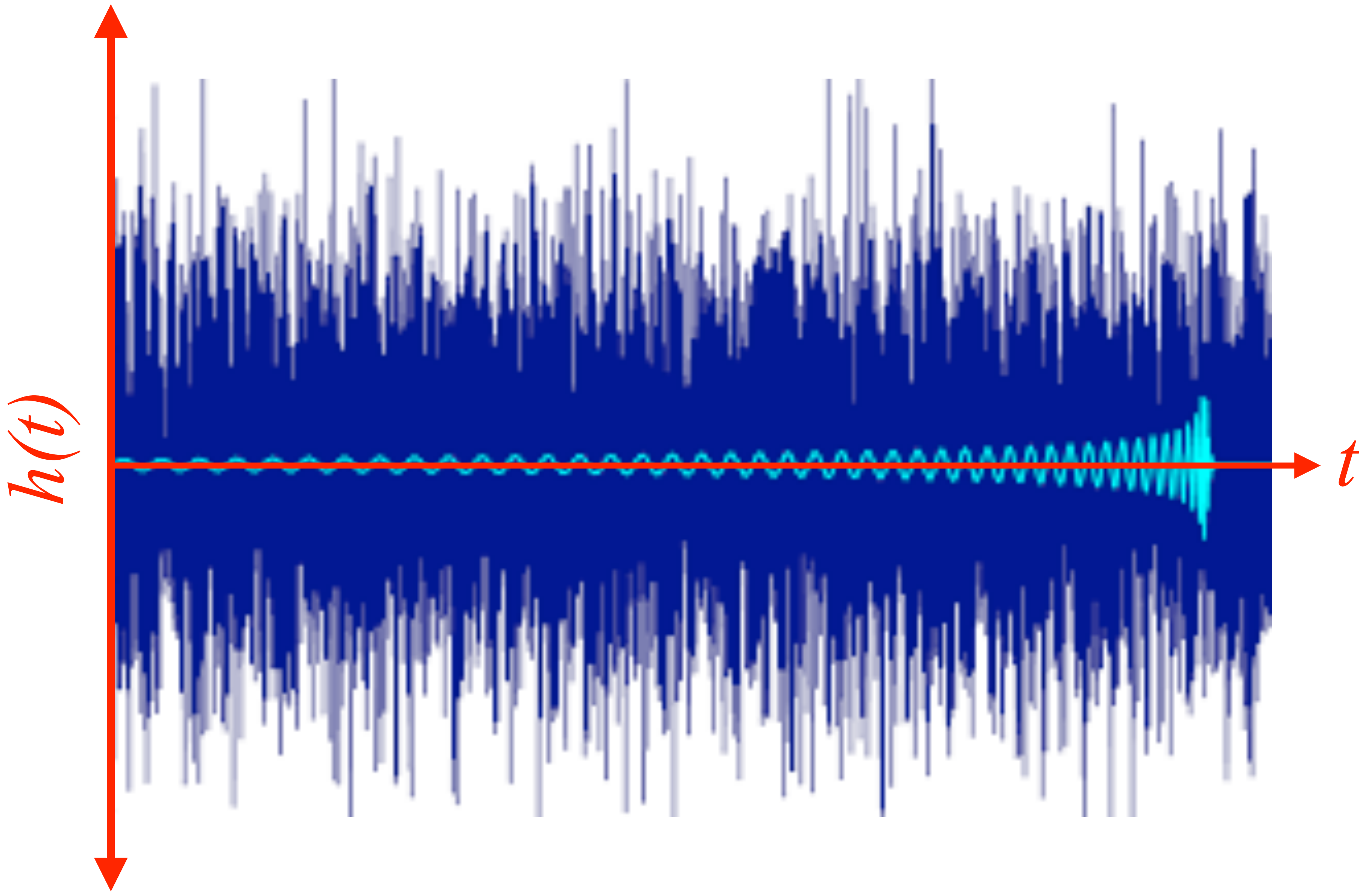


GW Source Types

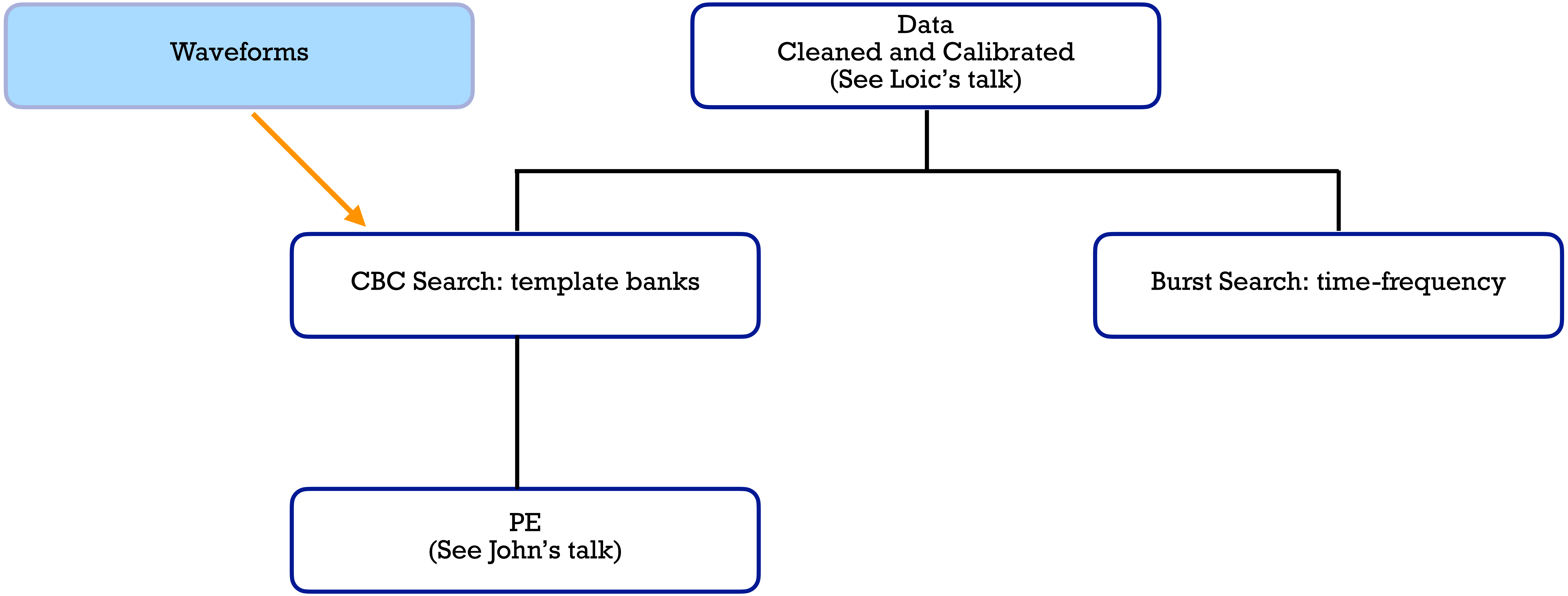
Modelled	Unmodelled
Uses matched filtering	Uses time-frequency maps
Requires a waveform model	Does not require a waveform model
Requires an assumption about GR	No assumption about GR
Requires an assumption about cosmology	No assumption about cosmology
Accurate, but expensive	Less accurate, but cheaper

Modelled Analysis: Matched Filtering

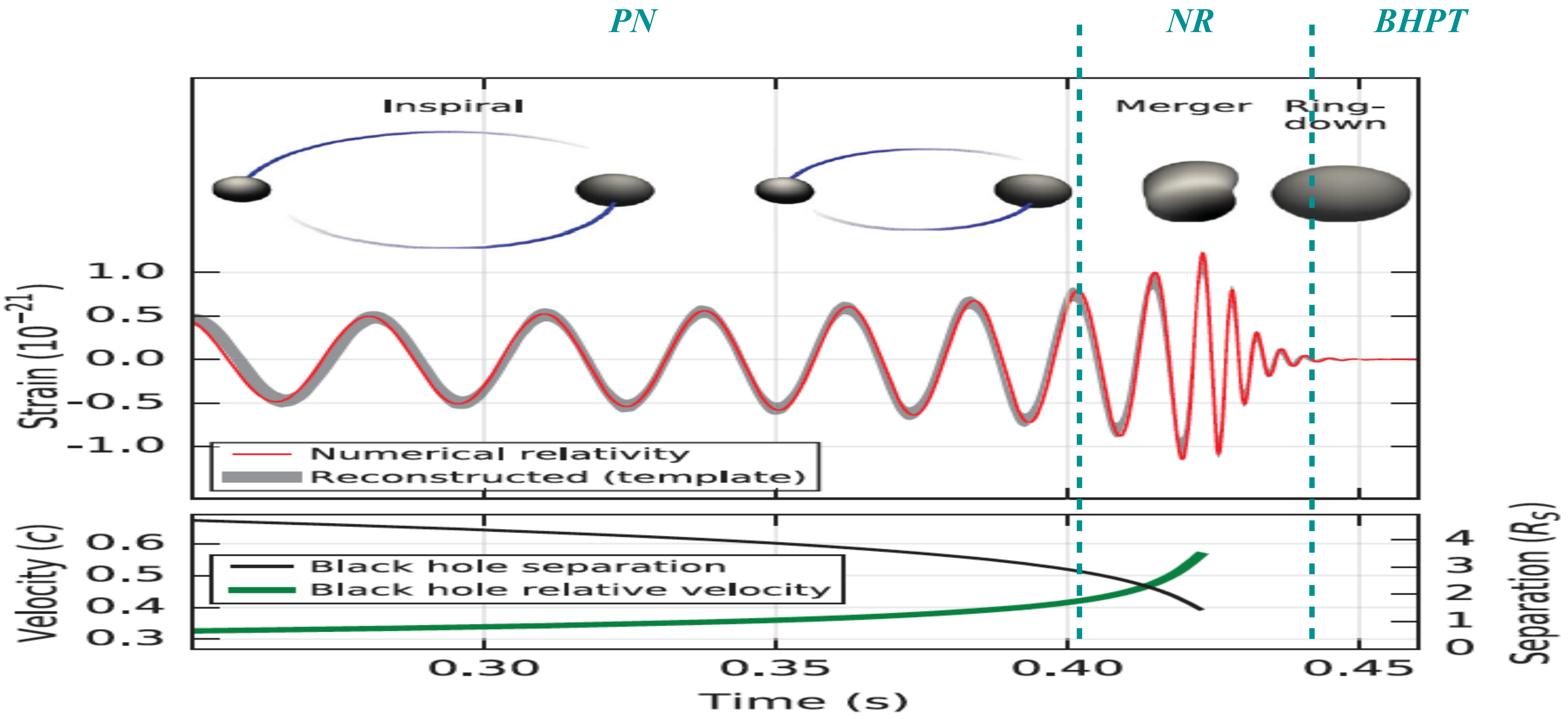
- Matched Filtering:
  - GWs are analogous to 1D sound waves
  - Optimal linear filter for signals buried in noise
  - Also known as optimal or Wiener filtering
  - Correlates a waveform model (template) with the data
  - Used for CBCs, cosmic strings, stochastic background, CWs
  - Phase sensitive



CBC Modelled Analysis: Searches & PE

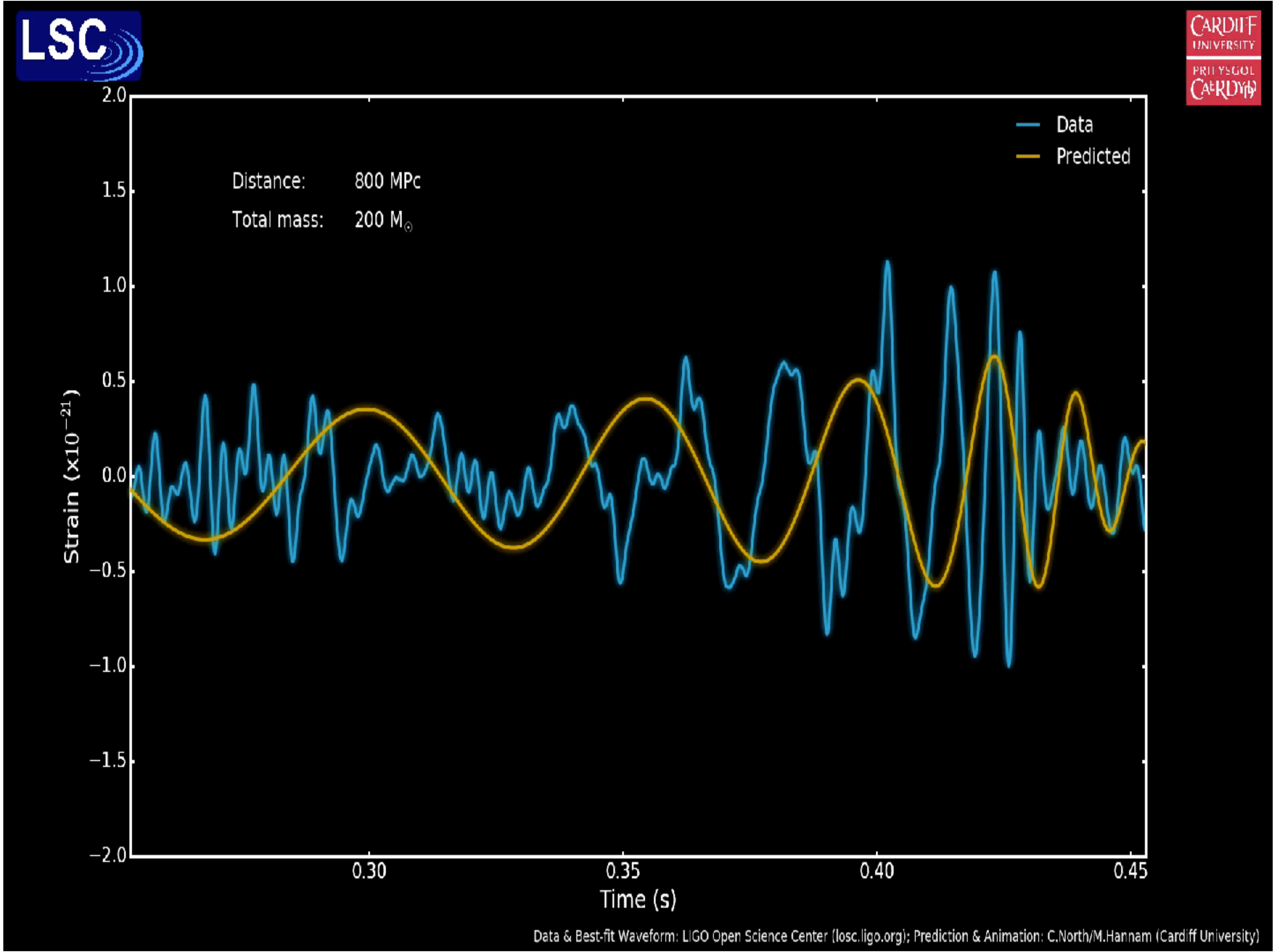


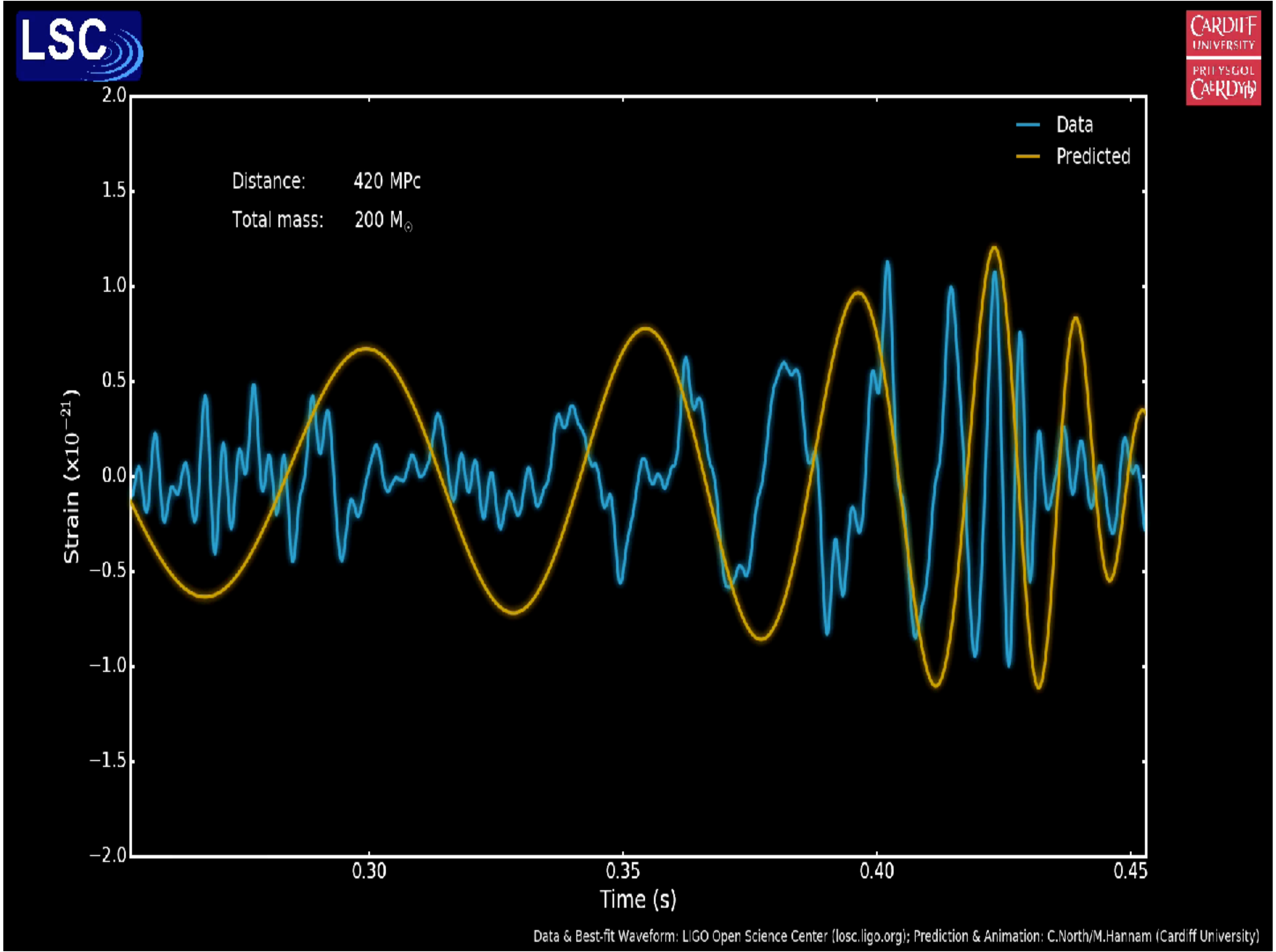
Modelled Analysis: Waveforms

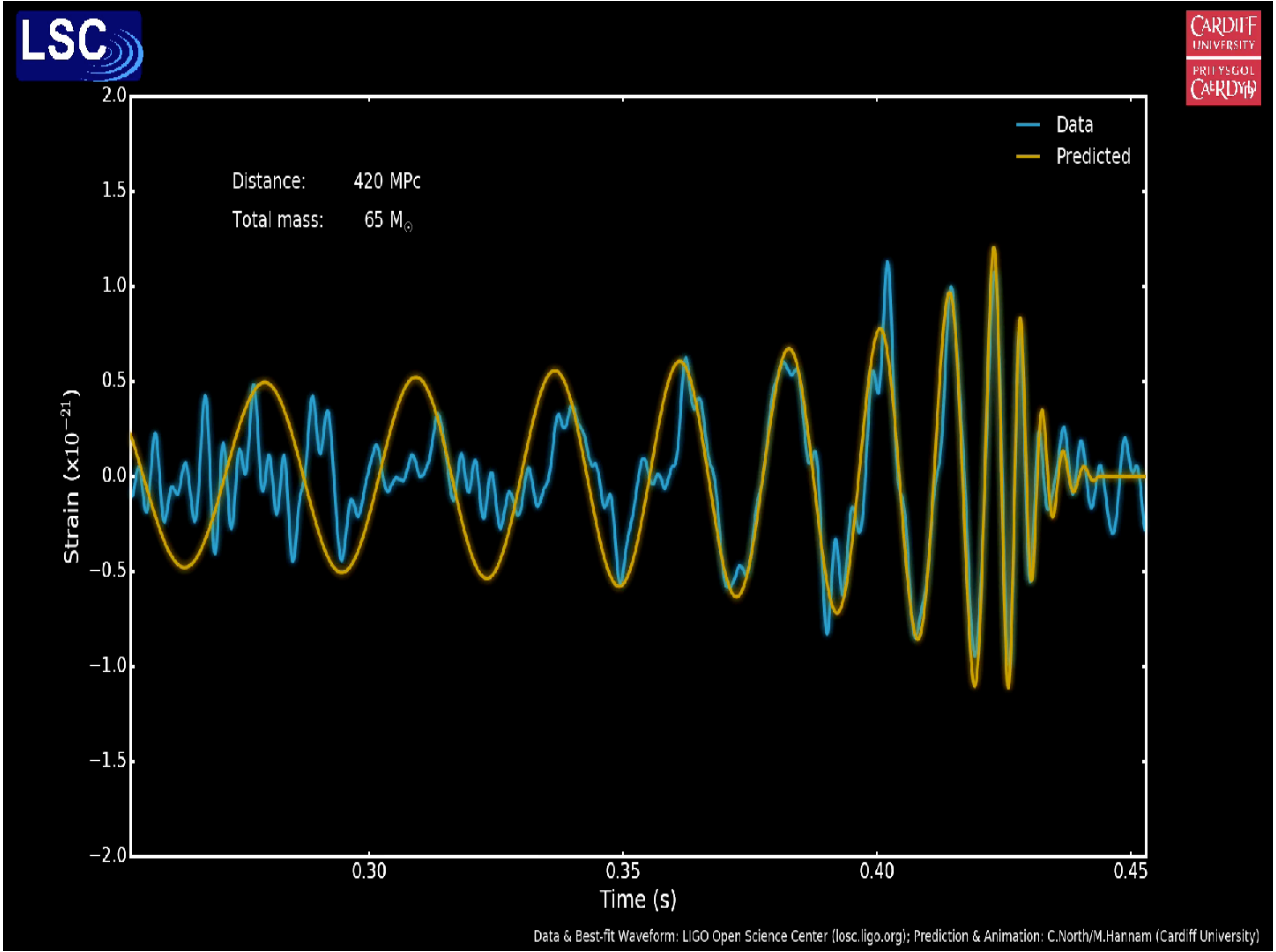


Modelled Analysis: Waveforms

- Inspiral
  - post-Newtonian (analytic methods - Mathematica etc.)
- Merger
  - NR simulations
  - Very expensive
  - At present, completely covered by national agencies/ERCs etc.
  - No known plan to make NR simulations an ET-funded activity at present
- Ringdown
  - BHPT - analytic


LSC
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
Data
Predicted
Distance: 800 MPc
Total mass: 200 $M_\odot$
Strain ($\times 10^{-21}$)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0.30
0.35
0.40
0.45
Time (s)
Data & Best-fit Waveform: LIGO Open Science Center (losc.ligo.org); Prediction & Animation: C.North/M.Hannam (Cardiff University)


LSC
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
Data
Predicted
Distance: 420 MPc
Total mass: 200 $M_\odot$
Strain (x10^{-21})
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0.30
0.35
0.40
0.45
Time (s)
Data & Best-fit Waveform: LIGO Open Science Center (losc.ligo.org); Prediction & Animation: C.North/M.Hannam (Cardiff University)



LSC
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
Data
Predicted
Distance: 420 MPc
Total mass: 65 $M_\odot$
Strain ($\times 10^{-21}$)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0.30
0.35
0.40
0.45
Time (s)
Data & Best-fit Waveform: LIGO Open Science Center (losc.ligo.org); Prediction & Animation: C.North/M.Hannam (Cardiff University)

CBC Modelled Analysis: Searches & PE

Data
Cleaned and Calibrated
(See Loic's talk)

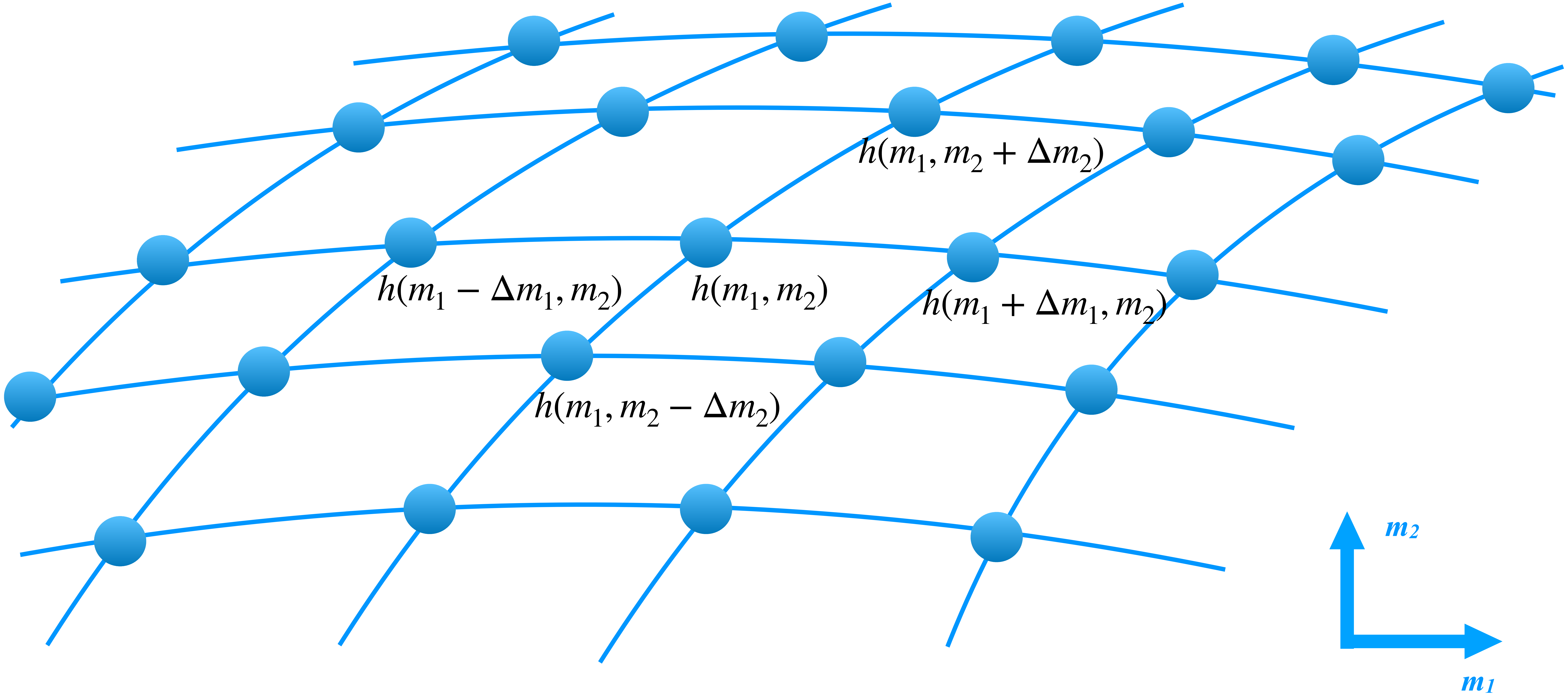
CBC Search: template banks

Burst Search: time-frequency

PE
(See John's talk)

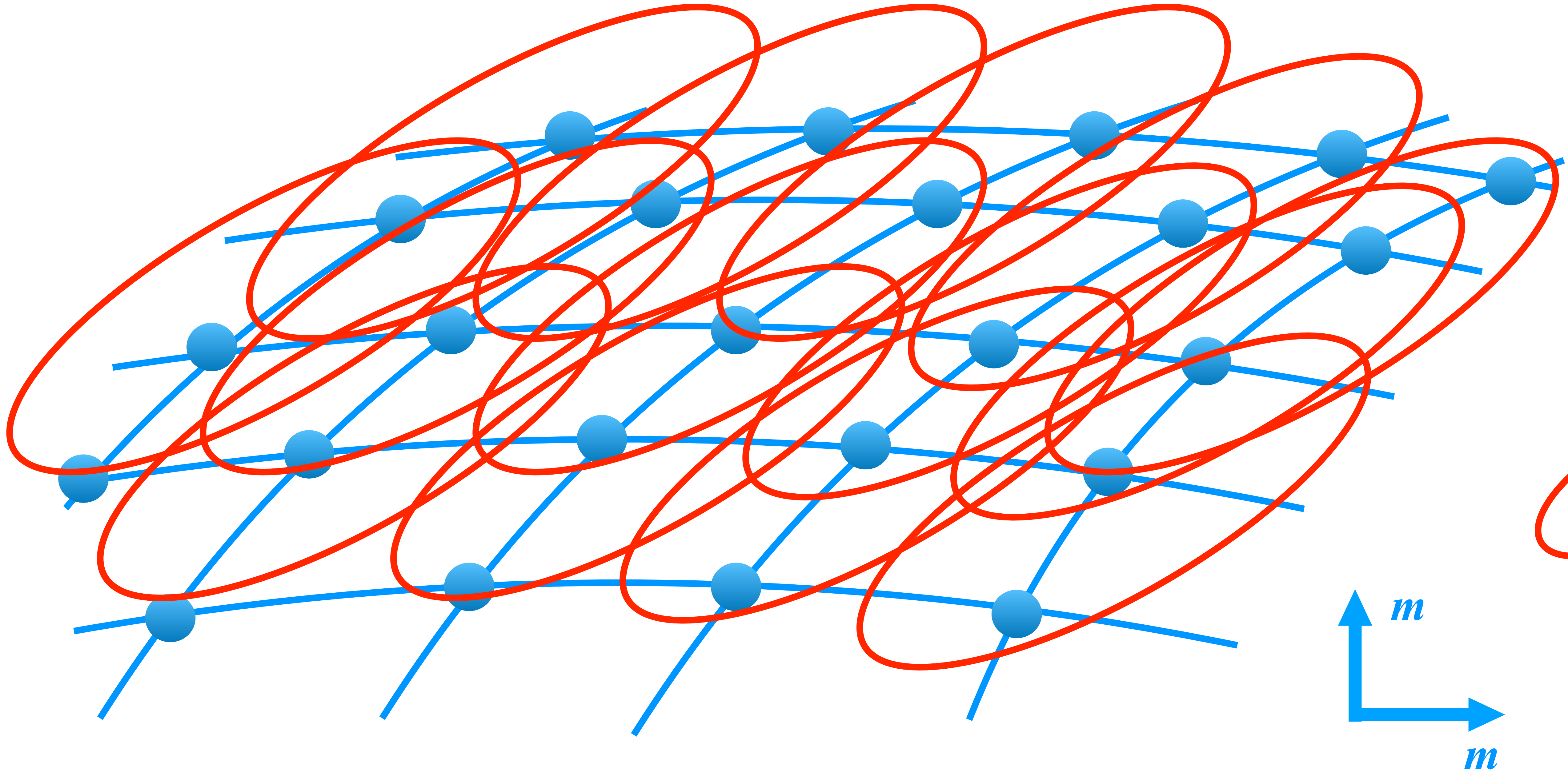
CBC Modelled Analysis: Template Banks

Idea is to cover the parameter space with a regular grid of templates

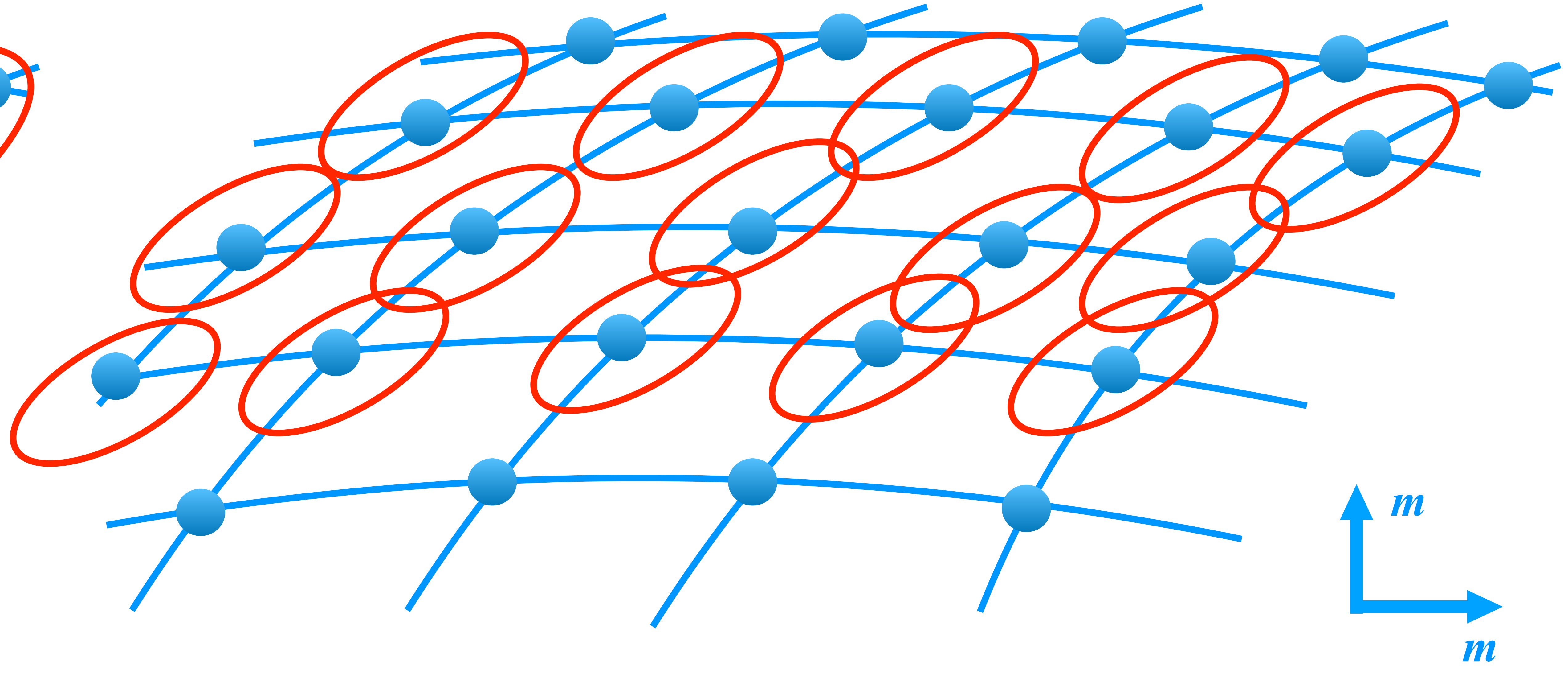


CBC Modelled Analysis: Template Banks

- Too much overlap between templates
- Computationally wasteful

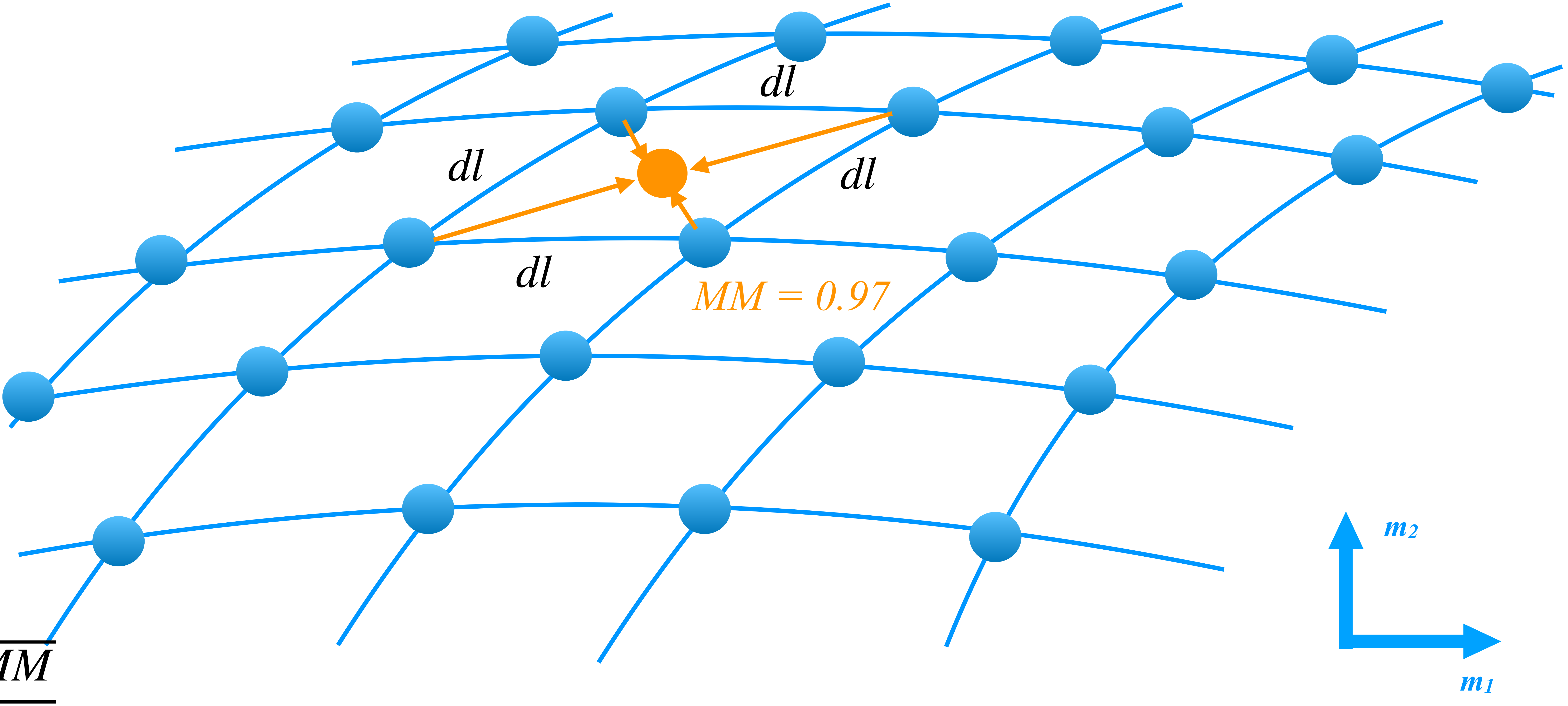


- Too many gaps
- Reduced computational cost
- Risk of missing many sources



CBC Modelled Analysis: Template Banks

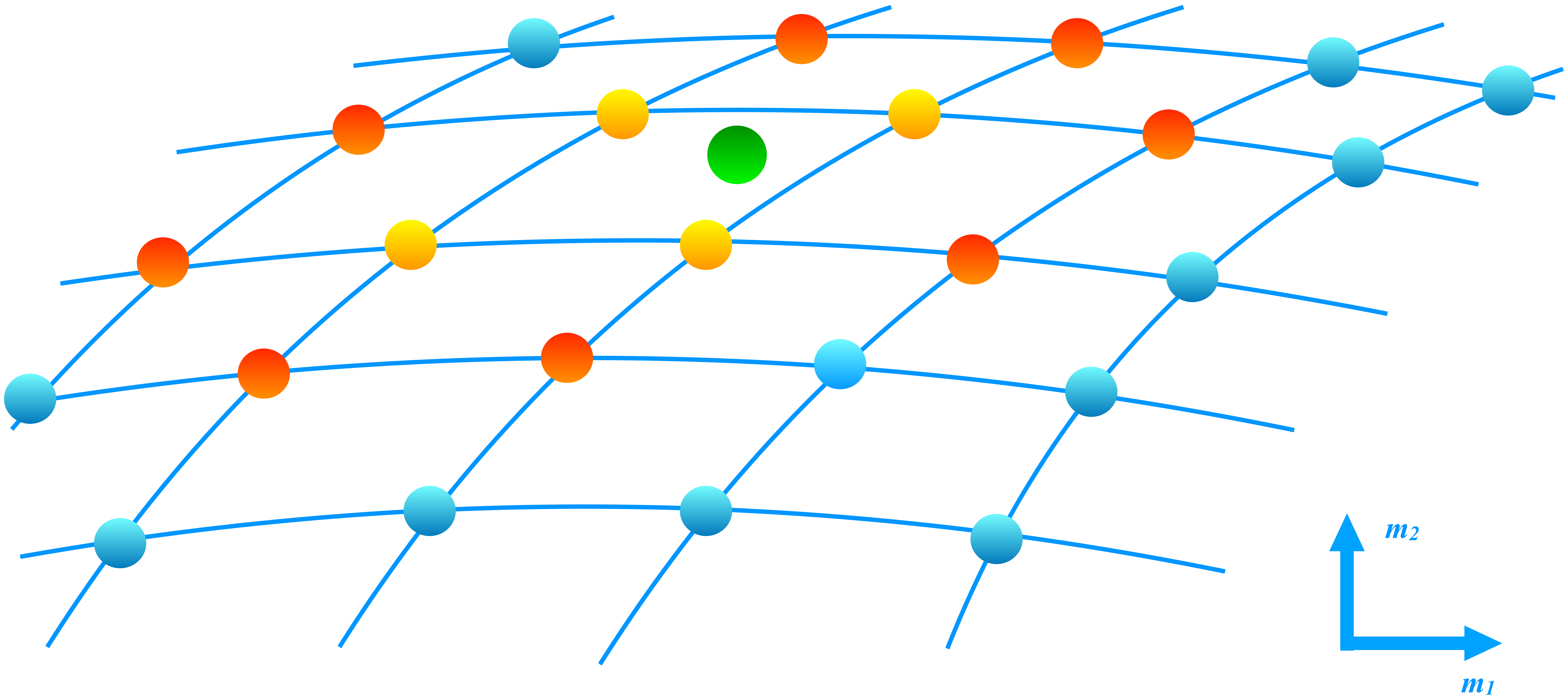
Position the templates such that a signal falling directly in the center of a grid segment has a minimum overlap with the closes templates



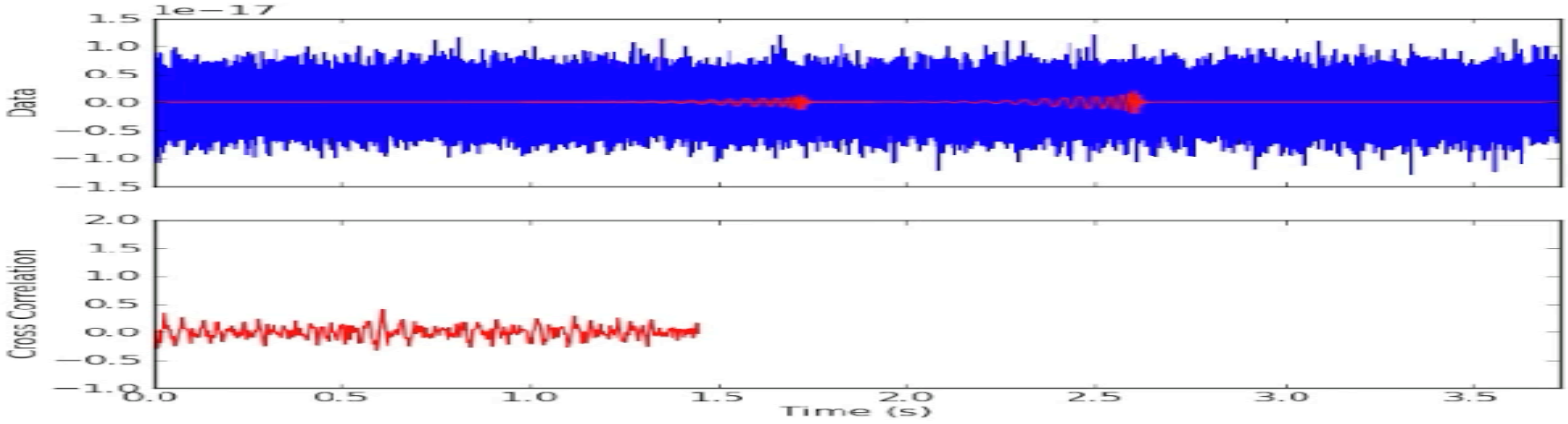
$$dl = 2\sqrt{\frac{1 - MM}{2}}$$

CBC Modelled Analysis: Template Banks

A GW signal will "light-up" the nearest neighbour templates



CBC Modelled Analysis: Template Banks

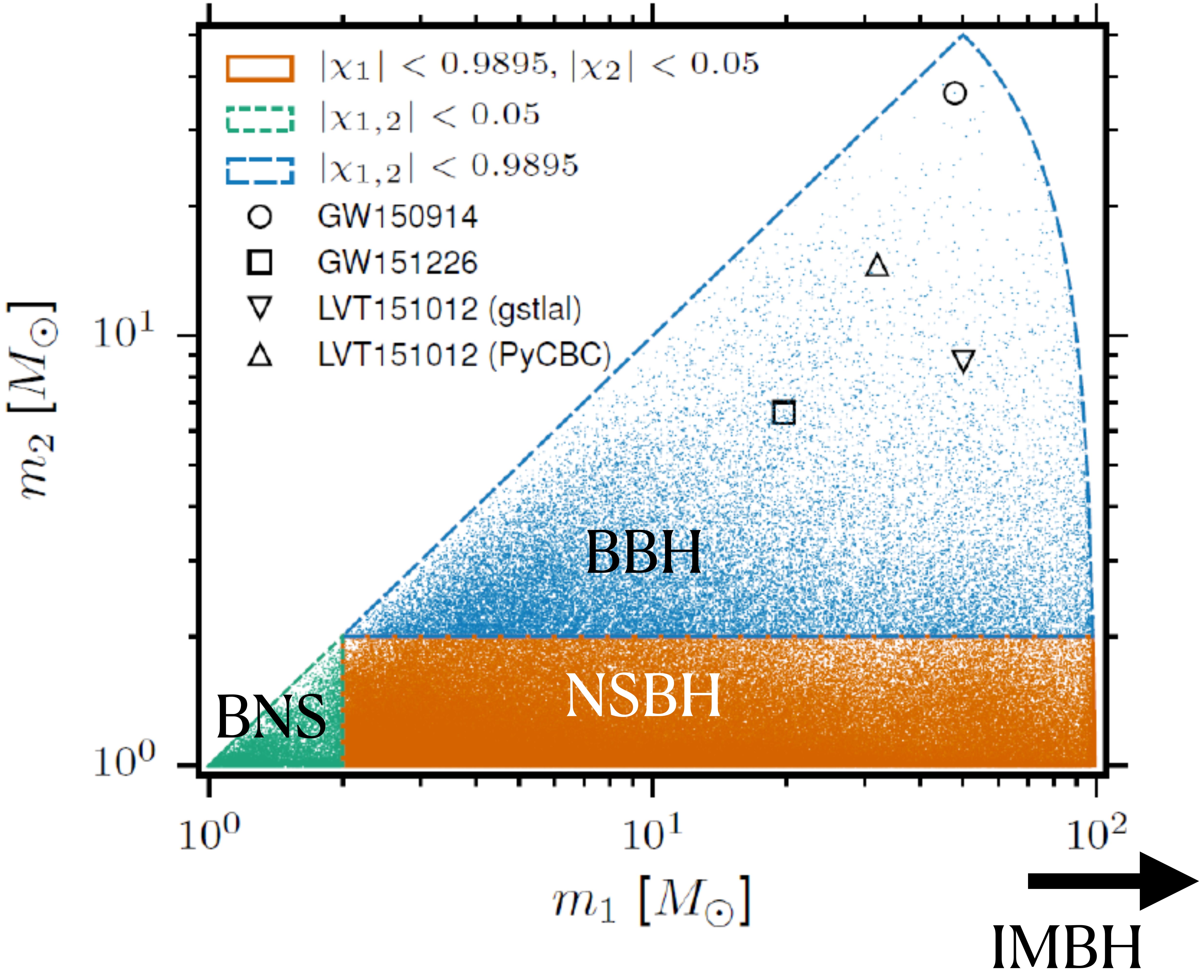


CBC Modelled Analysis: Template Banks

- Today's template banks cover from:
  - ~0.5 solar masses (subsolar mass search)
  - a few hundred solar masses (intemediate mass BHs)

- An ET template bank would increase at the high mass range only

- Result might be x2-x5 times the cost today (more on that later)



CBC Modelled Analysis: Parameter Estimation

ET EINSTEIN TELESCOPE

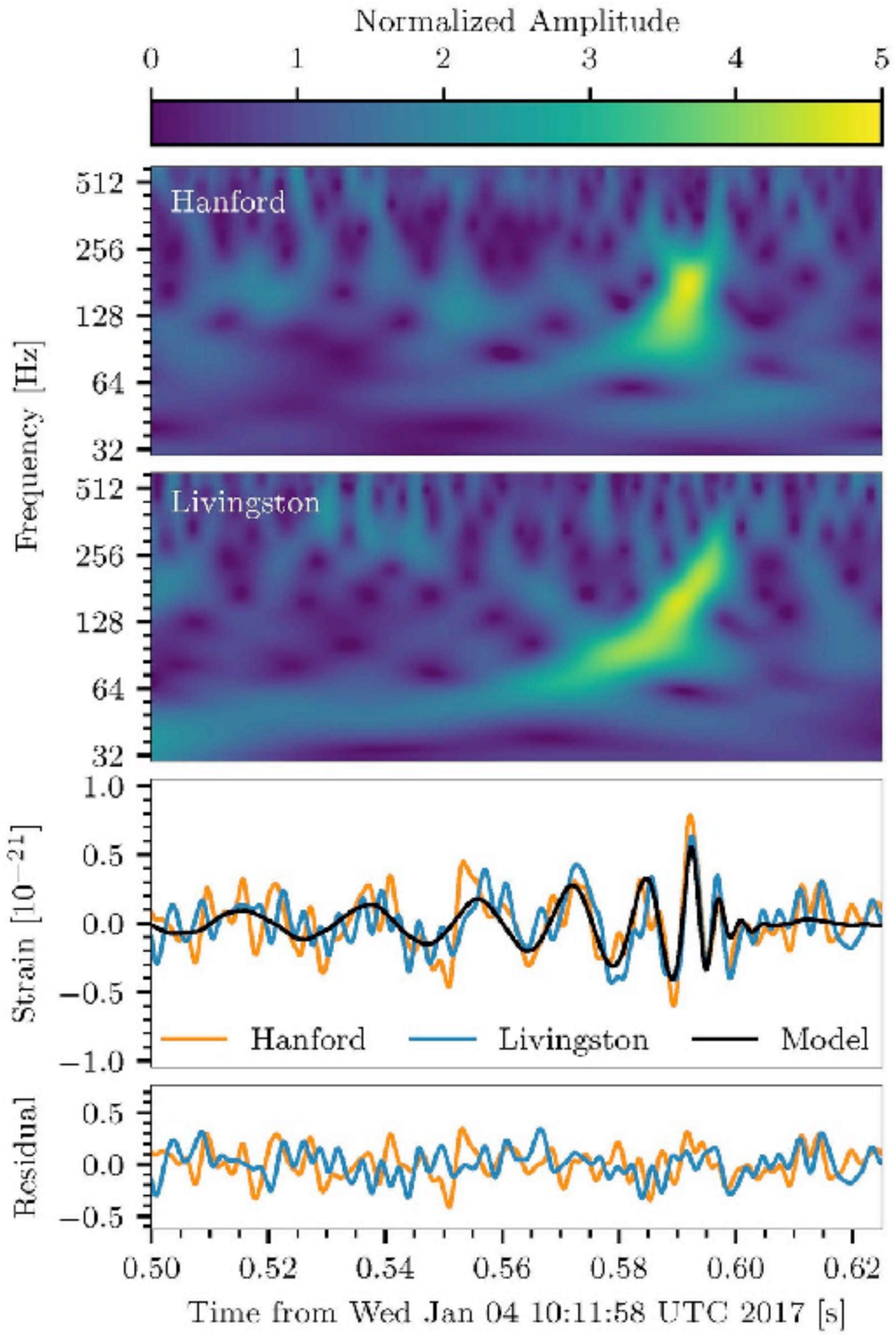
- Uses Bayesian inference
- Two algorithms used (see John's talk for more details):
  - Markov Chain Monte Carlo (Metropolis et al 1953, Hastings 1970)
  - Nested Sampling (Skilling 2006)
- Convergence is slow
- Tens of millions of waveform generations needed
- Runtime takes from days to months to infer parameters, depending on the source type.

CW Modelled Analysis

- Most expensive search
- Monochromatic emission of GWs
- Information scales with time of observation
- Computationally limited as the whole sky has to be covered in different frequency bands
- Data broken up into segments and short FFTs used.

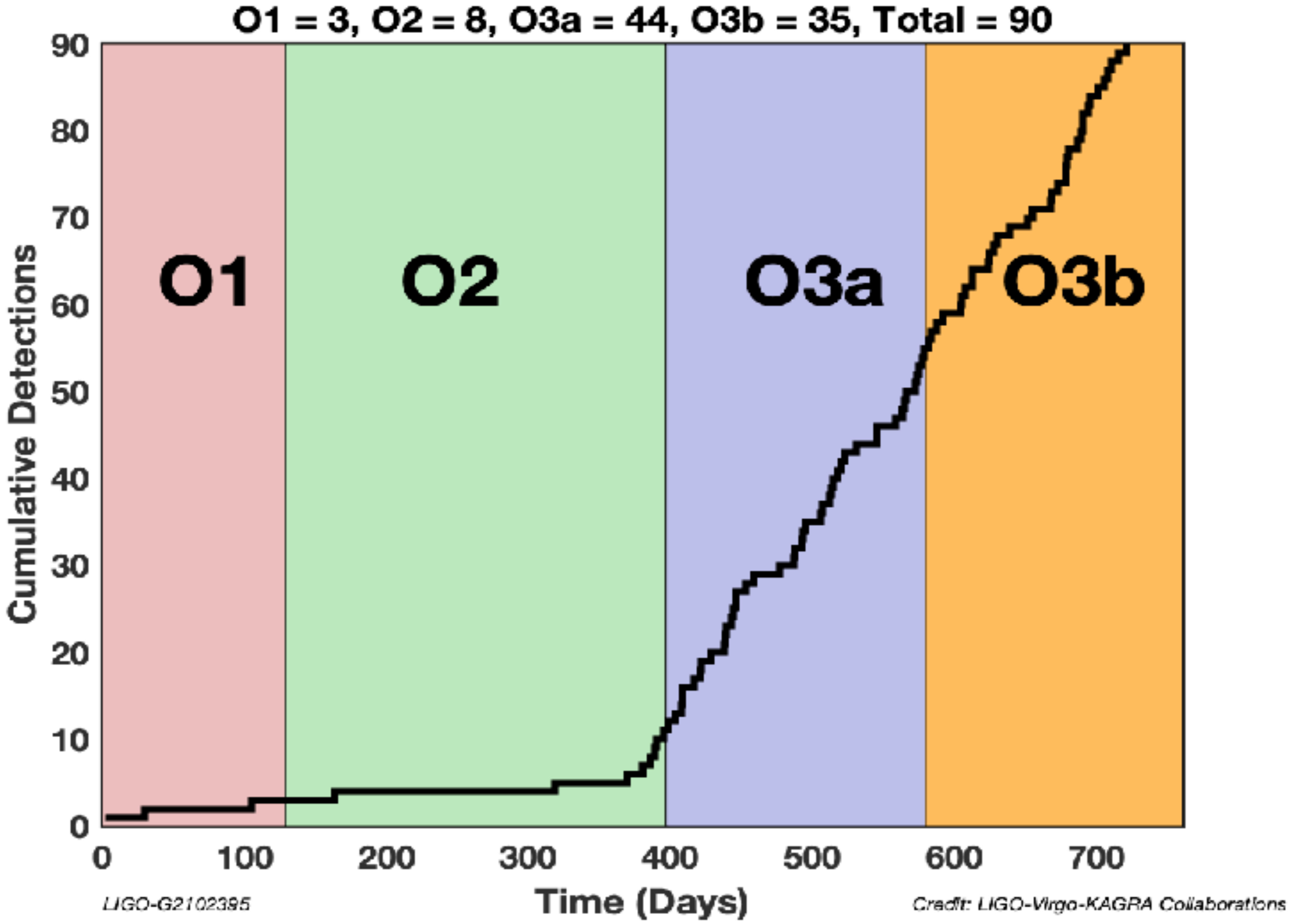
Burst unmodelled Analysis

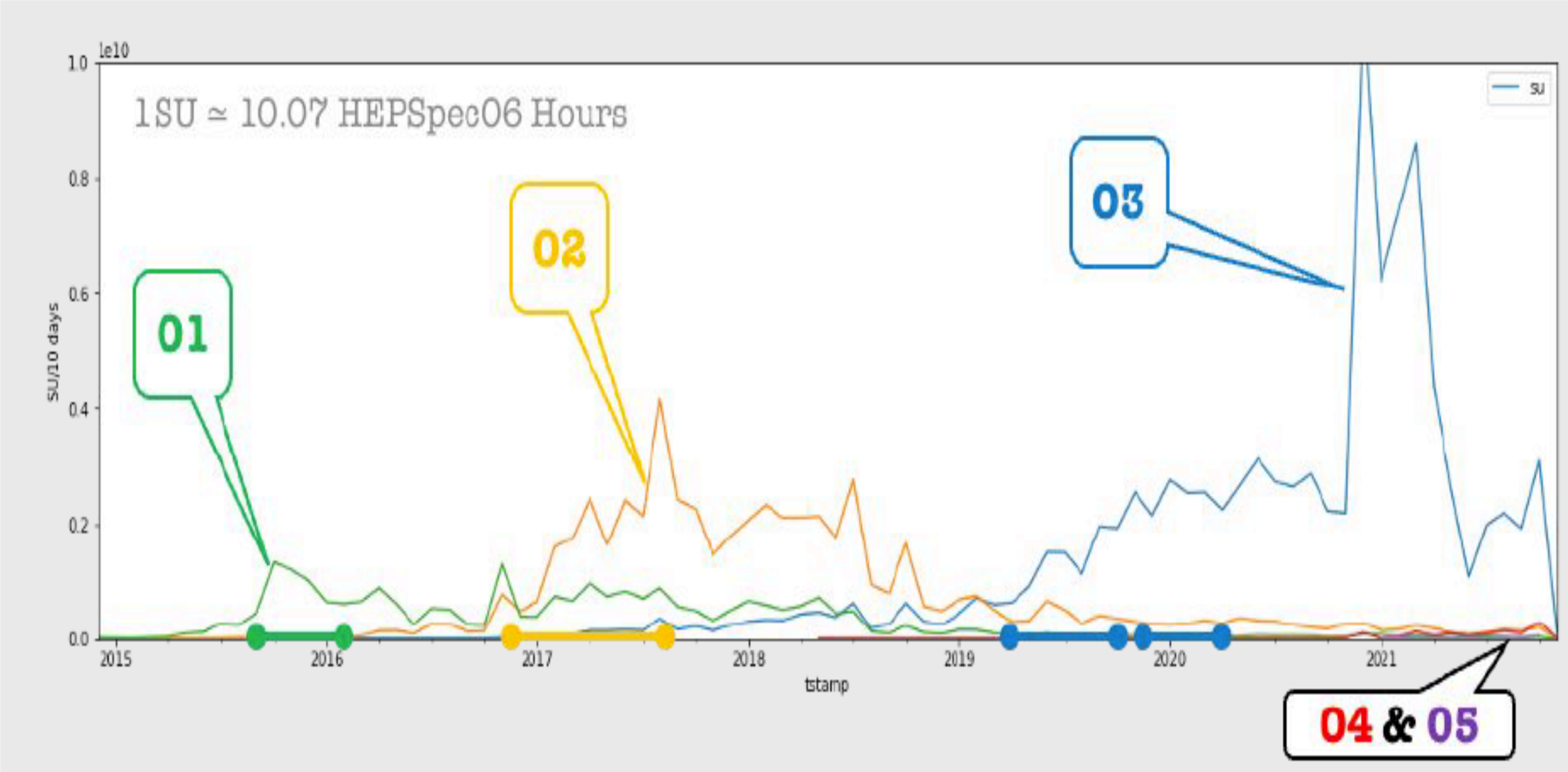
- Creates time-frequency maps with the data
- First BBH originally seen in TF map
- Computationally inexpensive search



LVK Computing Requirements O1 - O3

Computing requirements O1-O3





- Pretty steady evolution from O1 to O3
- No obvious correlation between number of sources and computational cost

Computing requirements O1-O3

ET EINSTEIN TELESCOPE

Total computing MSU per run

“Group”	“O1"	“O2"	Increase
"CBC PE"	24.80	149.58	x6
"CBC non-PE"	106.32	173.56	x1.6
“CW"	175.91	40.31	x0.2
“DetChar"	5.80	5.80	x0
“Burst"	48.14	39.23	x0.8
“SGWB"	1.94	4.77	x2.5
“Other”	0.82	0.00	x0
“Total"	363.72	413.25	x1.14

Courtesy of S. Bagnasco

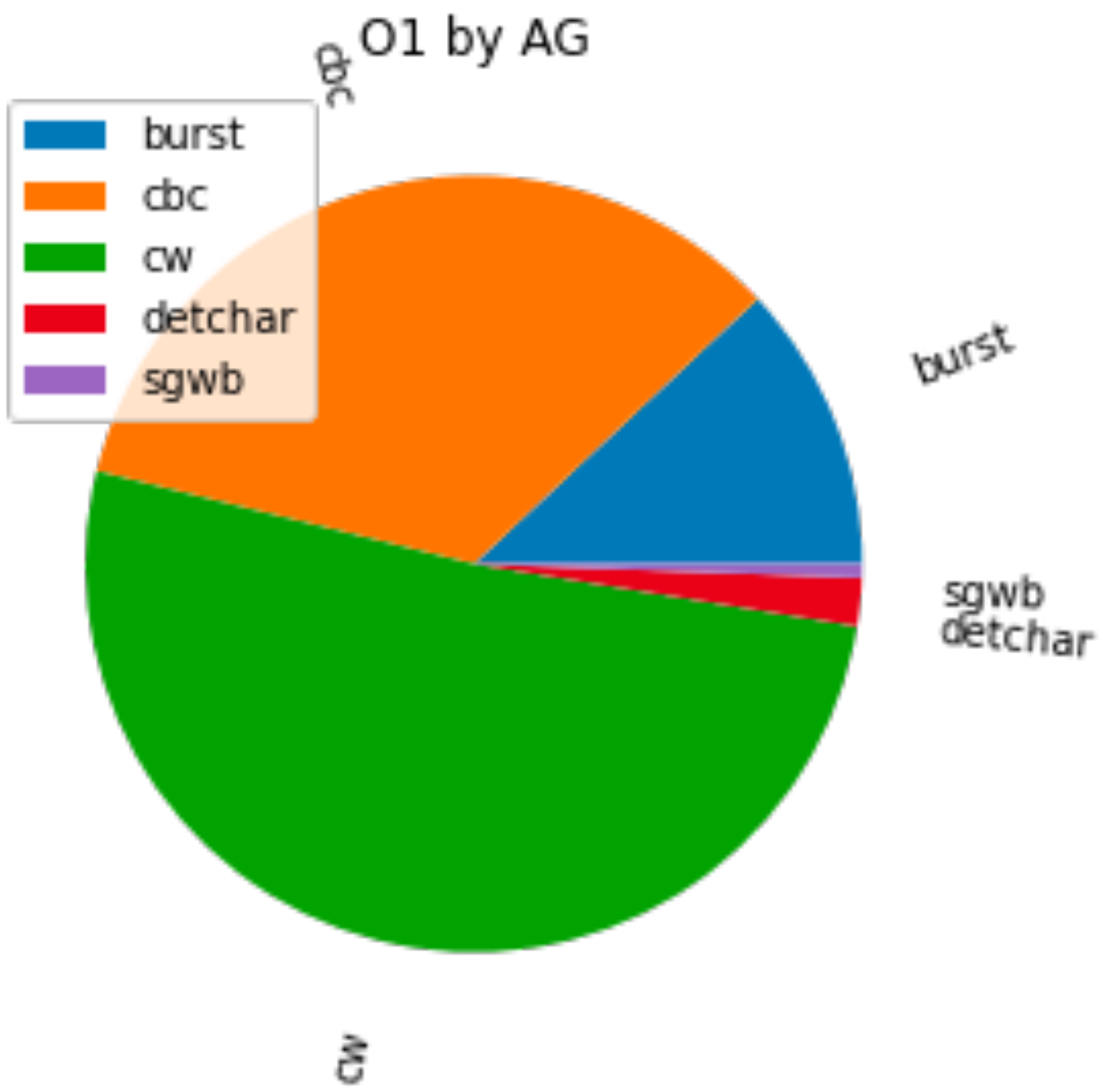
CBC Template Bank Cost

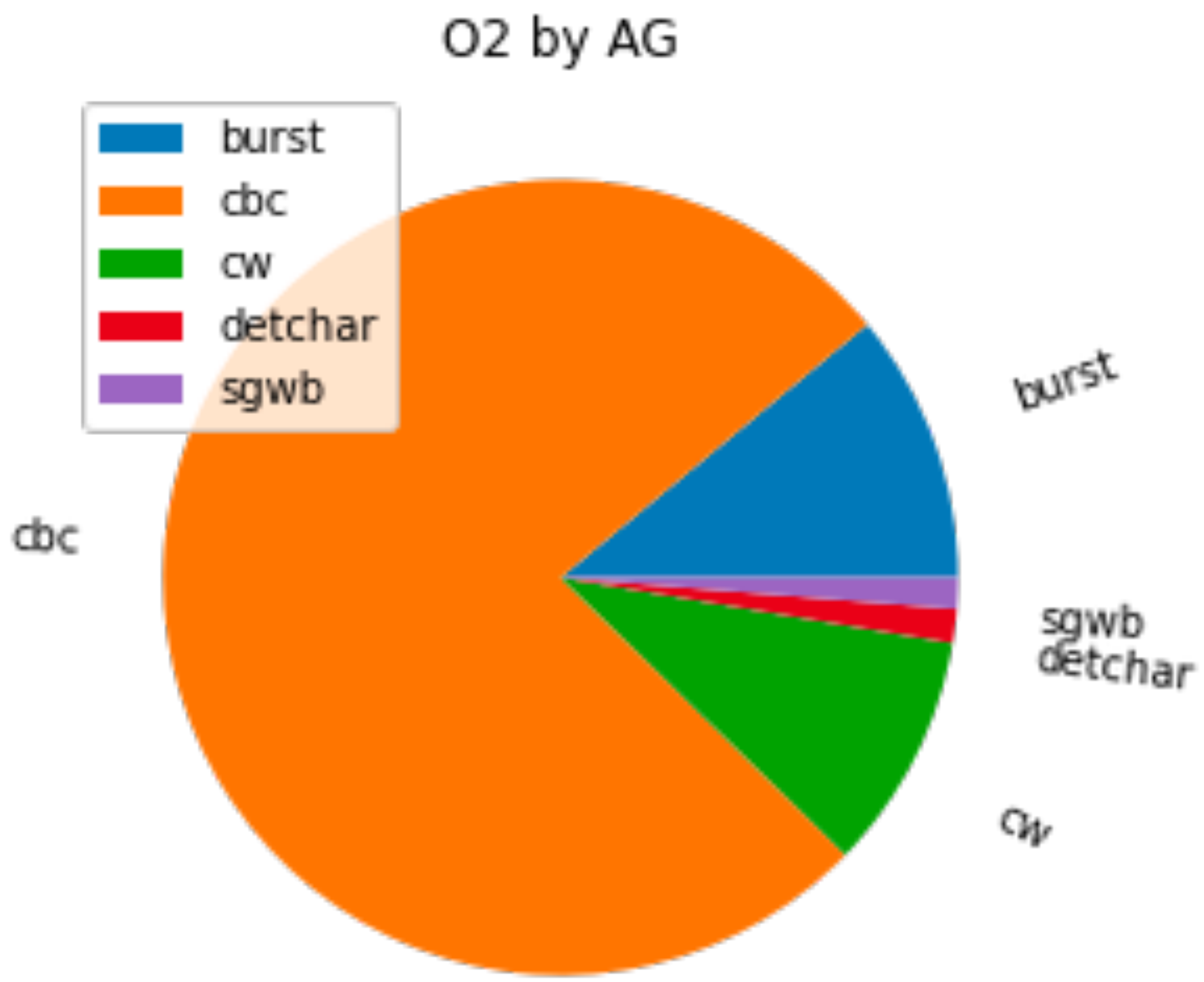
Pipeline	Number of cores (raw)	Number of M640 equivalent cores**	Running searches
gstlal	4404	5866	E.W.; main search; half SSM
PyCBC	1780	1925	E.W.; main search
SPIIR*	3896	1784	E.W.; main search
MBTA	700	645	E.W., main search; full SSM

Courtesy of B. Mours

- Misconception that template banks are a major source of computational cost
- The template banks are generated once at the beginning of each run
- Actual usage is between 10% and 50% of available cores
- Parameter space is pretty well covered

Computing requirements O1-O3



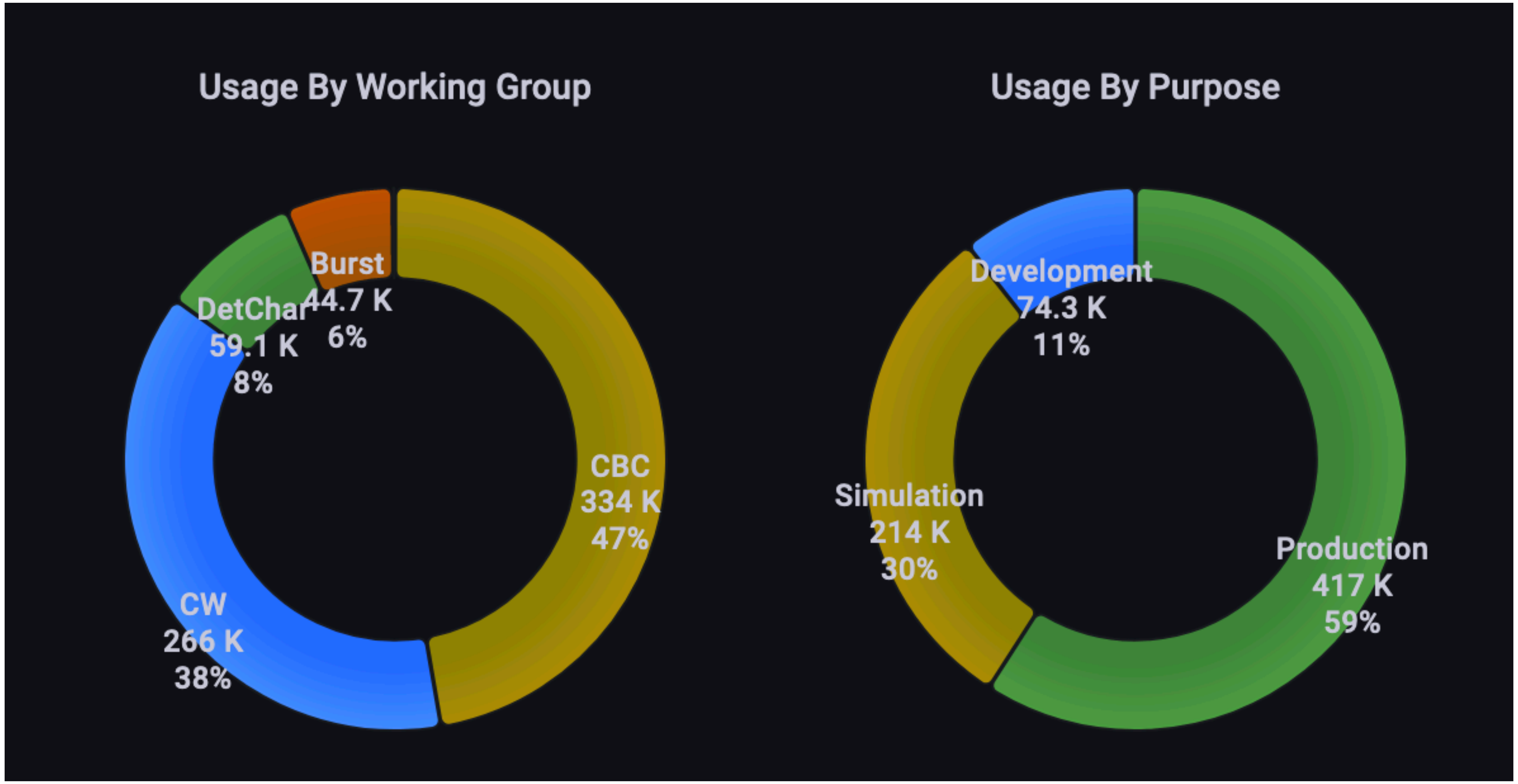


Courtesy of S. Bagnasco

The main changes in CBC and CW usage were due to development more than any other factor

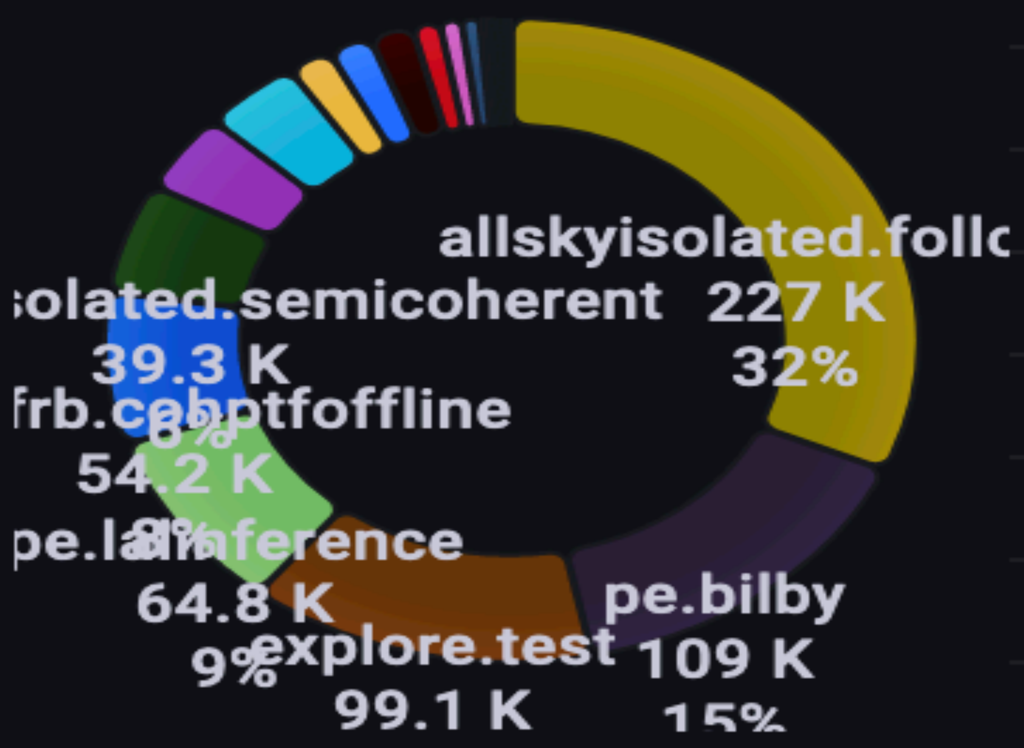
Computing requirements O1-O3

O3



- N.B. Production code only accounts for ~60% of total usage

Computing requirements O1-O3



	Value
allskyisolated.followup	227 K
pe.bilby	109 K
explore.test	99.1 K
pe.lalinference	64.8 K
closefrb.cohptfoffline	54.2 K
directedisolated.semicoherent	39.3 K
closefrb.xoffline	28.1 K
pe.pbilby	27.5 K
pe.lalinferencerapid	12.3 K
bayesianpopulations.parametric	11.9 K
grb.xoffline	11.9 K
chan_mon.ligocam	7.29 K
state.calib	5.65 K
ebbh.cwboffline	4.59 K

bbh.pycbcoffline	1.04 K
directional.stochastic	645
imbh.gstlaloffline	131
offlinedq.idq	130
transient.omicron	75
allsky.cwboffline	53
sgr_qpo.stamp	21
testgr.tiger	7
directedbinary.crosscorr	2
sn.cwboffline	1
uber.gstlaloffline	1

- PE for the 79 O3 detections only 214 K
- Main computational driver is CW search

Computing requirements 03

	03 estimate	03 actual	Ratio	04 estimate
CBC	487	465	95%	680
Burst	141	53	38%	85
DetChar	6	22	367%	9
CW	141	110	78%	349
Stochastic	2	5	250%	25
Total	777	655	84%	**1148**

1SU ≃ 10.07 HEPSpec06 Hours
Updated as of Oct 28

- Already difficult to estimate requirements from one run to another…
- …predicting for ET is doubly difficult

Open questions about computing requirements for ET

Prediction of search costs for ET

- CBC
  - template bank size may change, but it may result in factor of 2-5 difference?
  - conclusion : if we are still using template banks, no-one thinks computing power is an issue
  - Advanced methods may be cheaper again
- Burst
  - Burst search is relatively cheap, and should get cheaper as we move towards 3G
- Stochastic
  - same as burst
- CW
  - Unless there is an algorithmic breakthrough, CW searches will continue to be the main computational cost driver

Open questions on PE/LL for ET

- N.B. This conversation has not yet happened at the OSB level
- PE today is not the main driver of computational cost. Moving towards automation, so O4 numbers may be more of a realistic indicator of cost.
- Main computational bottleneck for PE is waveform generation and/or likelihood calculation runtime
- Faster waveforms or likelihood calculations = faster PE
- For ET PE, a number of outstanding questions exist:
  - only PE for systems with NS, BBHs with large mass ratio, IMBHs?
  - for vanilla BBHs, search parameters enough for populations etc.?
  - what will the EM community require in terms of alert content?
  - any pre-alert PE will be done on partial waveforms, so sampling rate will be low

Open questions for ET

- The future of CE is unclear at present
- If we assume CE is funded, we could probably assume:
  - Joint development of analysis tools
  - Shared computing resources
  - Joint data sharing
  - Joint development of protocols, libraries, data bases etc.
  - Conversation has not yet started in earnest

conclusion

ET EINSTEIN TELESCOPE

- No great explosion in computational needs between LVK O1-O3

- We have no idea of what algorithms will be used for ET…
- …but, no-one foresees any great computational problem for ET data analysis

- There are a lot of things that need to be discussed…
- …but it is still too early to discuss many of the topics

- The first ET-MDC will begin on November 1st 2023…
- …we hope this will provide more concrete initial numbers!