Iterative adaptive learning for parameter and population inference

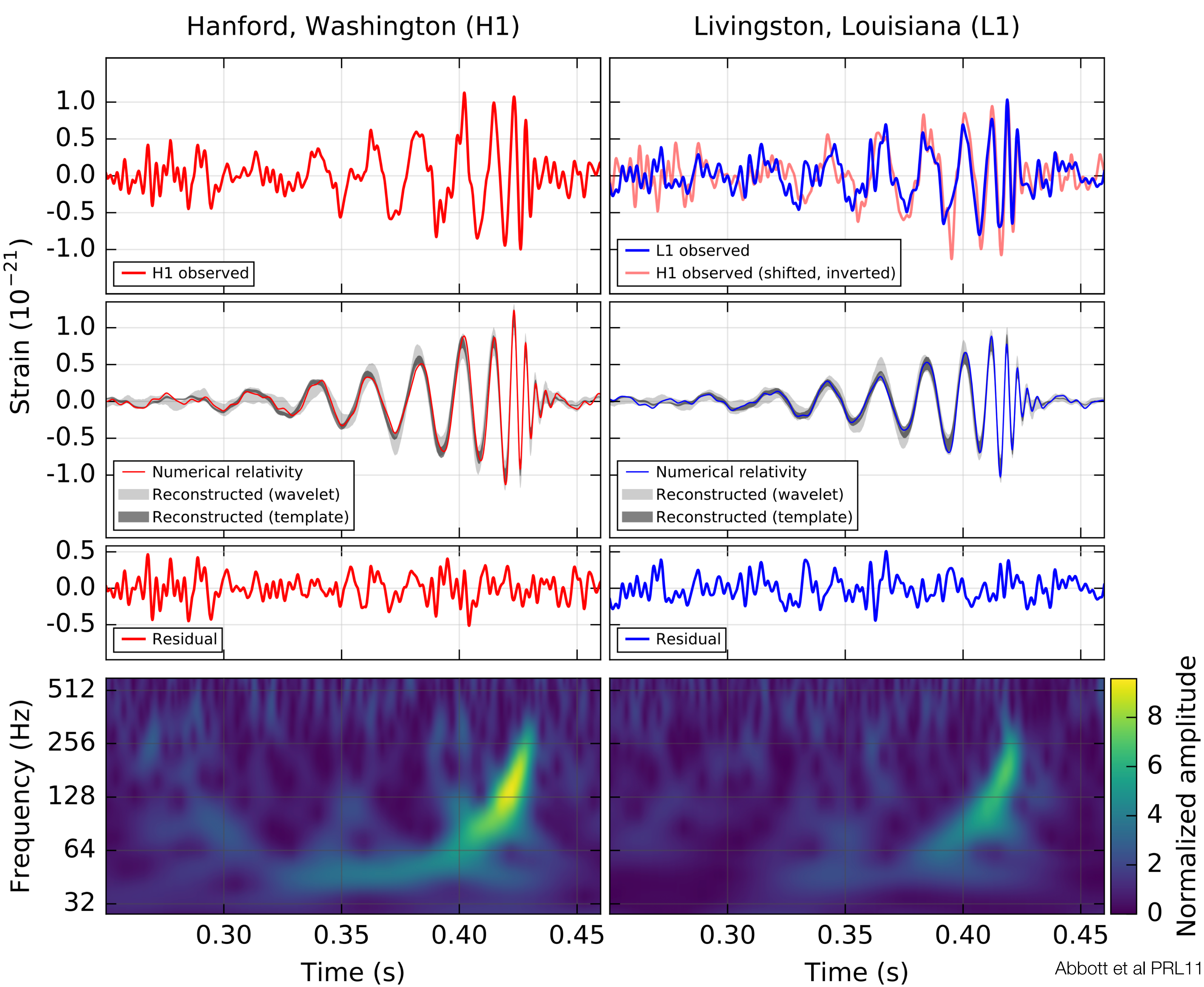
Richard O'Shaughnessy

2020 11 20 ICERM meeting

Outline

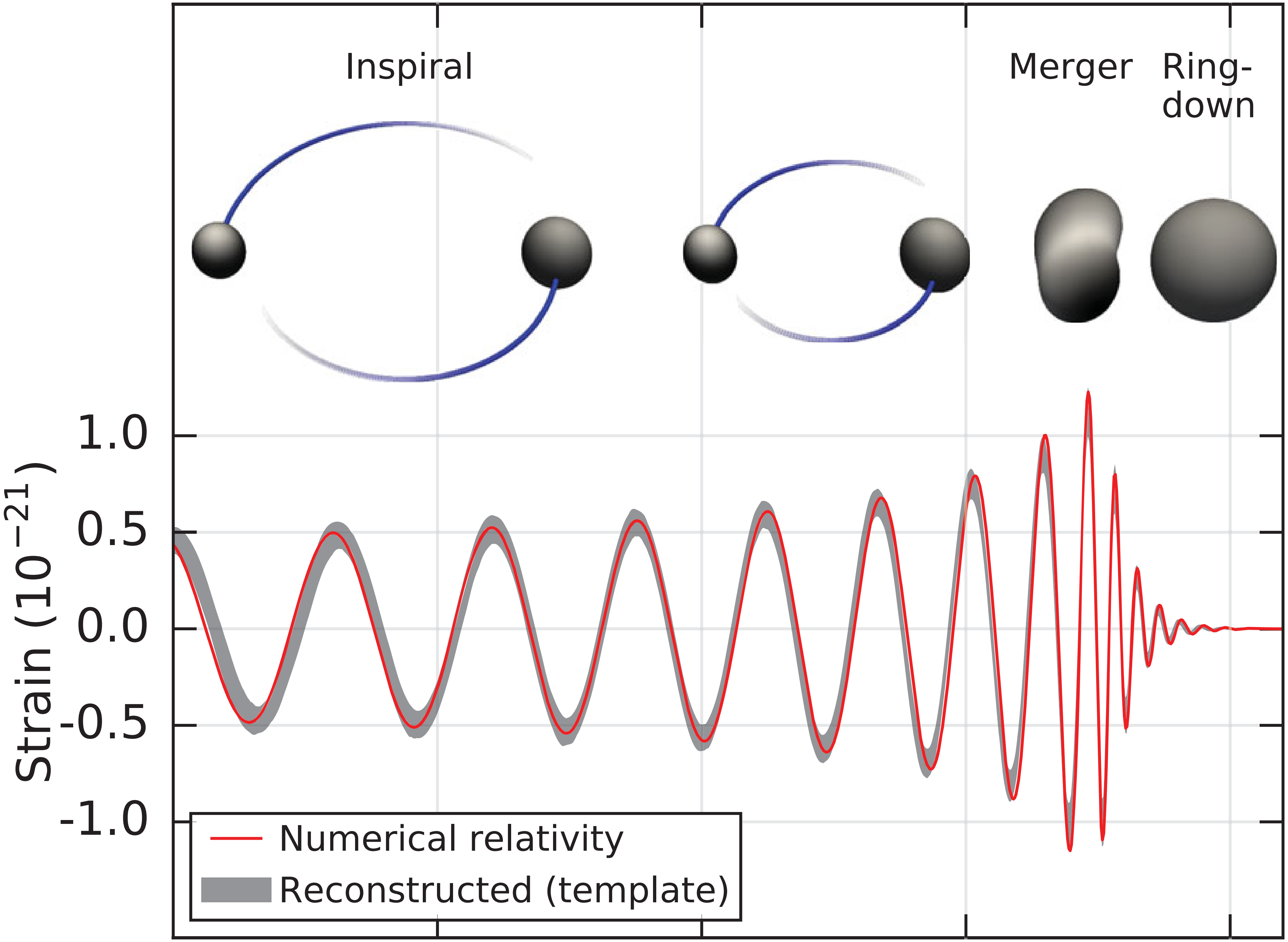
- Motivation: Expensive forward modeling
- Surrogate likelihoods for GW parameter inference: RIFT [J. Lange]
- Surrogate likelihoods for population inference [D. Wysocki, V. Delfavero]
- Surrogate kilonova light curves [M. Ristic]
- Lessons learned

Inference challenge 1: GW parameters



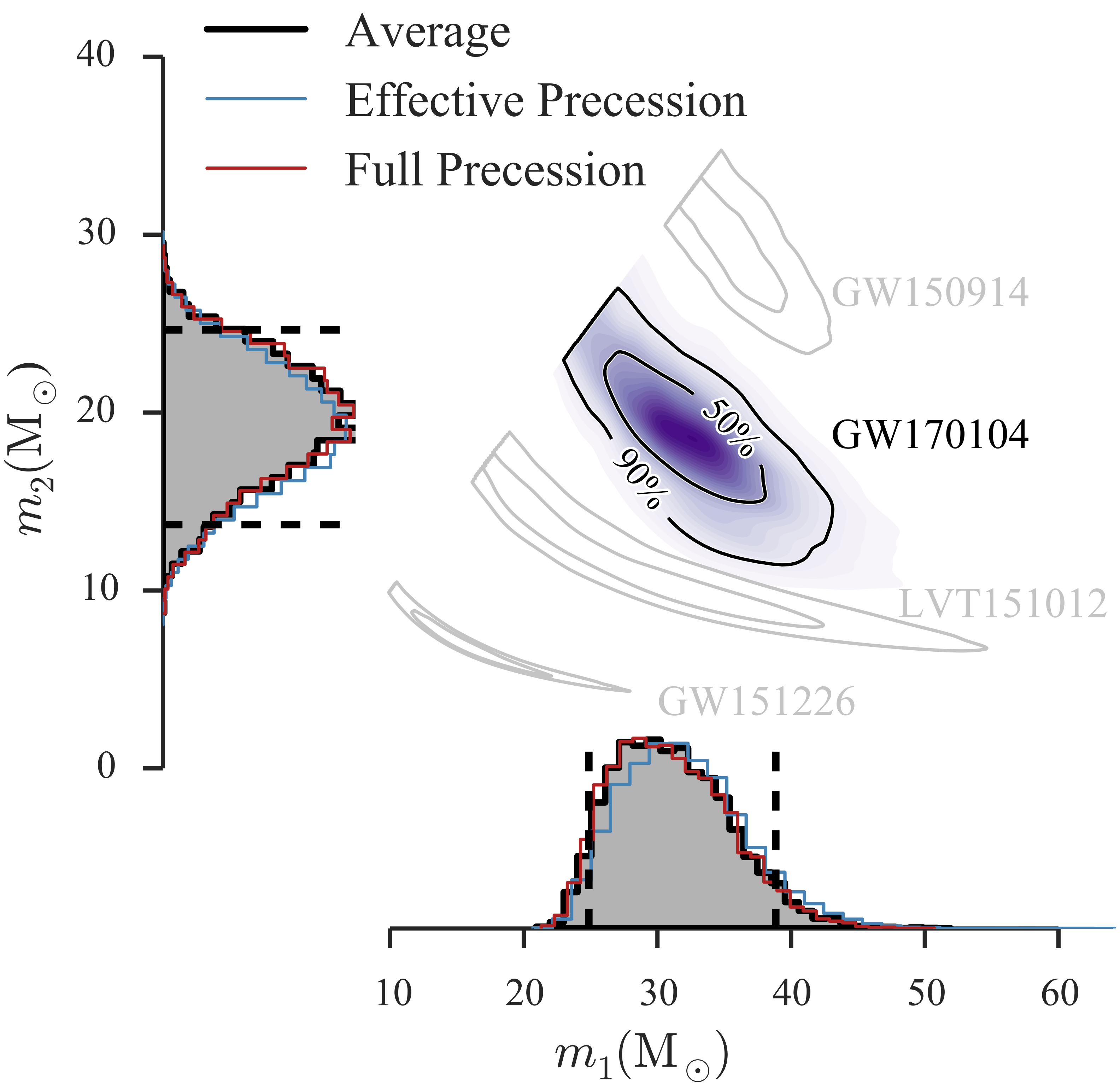
Abbott et al PRL116, 061102 (2016)

Comparison to source models



Abbott et al PRL116, 061102 (2016)

… tells us about each source



Abbott et al PRL 118, 221101 (2017)

RIFT: Rapid Iterative Fitting (for inference)

- Problem to solve:

$$p(x) = \frac{\mathcal{L}(x) p_{prior}(x)}{\int \mathcal{L}(x) p_{prior}(x) dx}$$

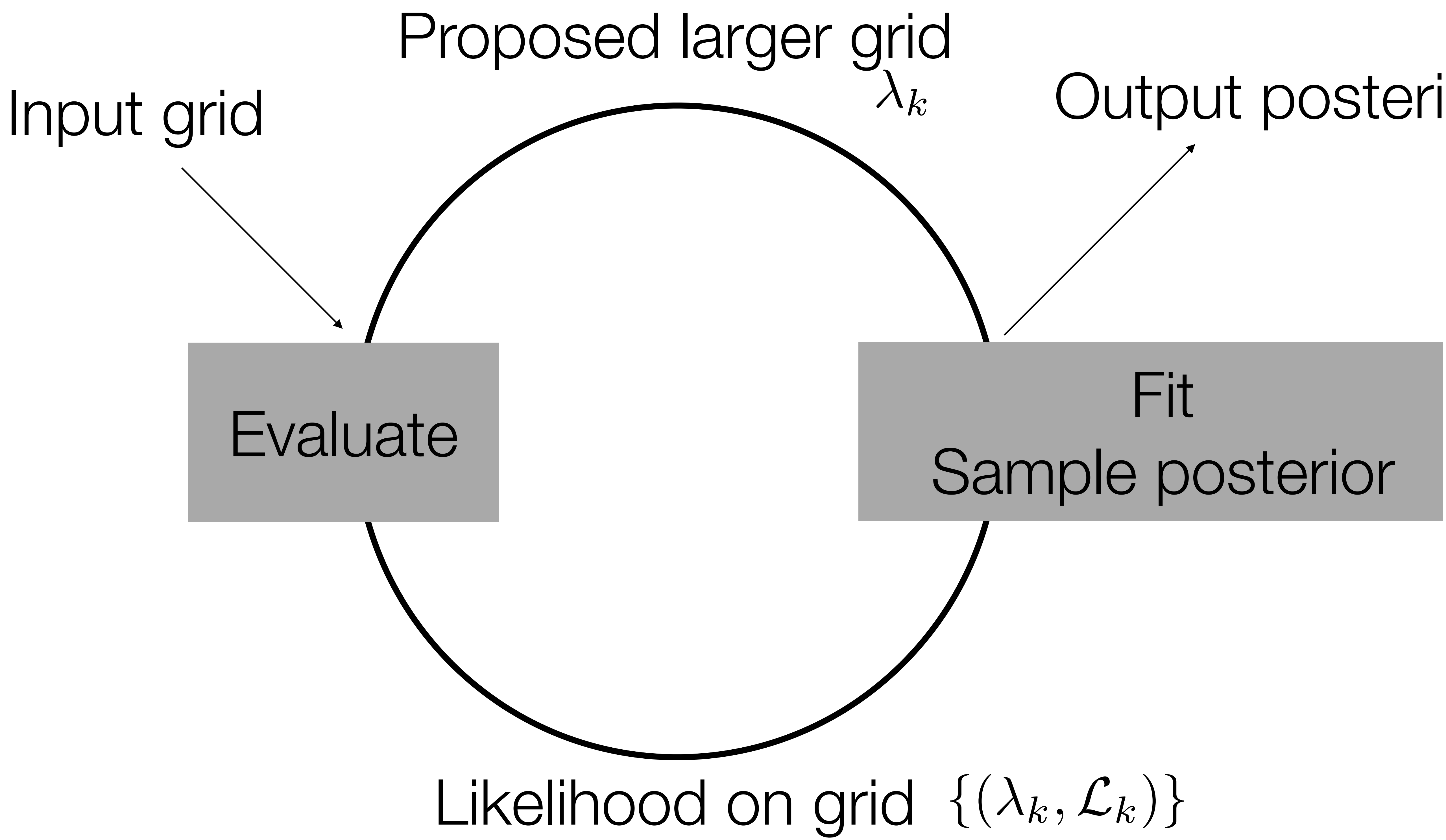
  - but $\mathcal{L}(x)$ is very expensive

- Method: Evaluate $\mathcal{L}(x)$ and train an approximation (“surrogate”) $\hat{\mathcal{L}}(x)$
  - Iteratively refine
  - Target reliable posterior in ‘bulk’ of support (e.g.,, 99.9…% probability)

- Fit method: Anything flexible (e.g. random forests, …). We usually use gaussian processes in modest-sized problems
- Posterior generation: Monte Carlo integration. (Preferably adaptive)

RIFT: Rapid Iterative Fitting (for inference)

- **Repeat until stable**



Abbott et al 2016, PRD 94 4035

Under the hood: Intrinsic/extrinsic split

- RIFT works with GW data
  - Assumes likelihood is gaussian noise, with signal h, in detectors k

$$\ln \mathcal{L}(\lambda;\theta) = -\frac{1}{2}\sum_k \langle h_k(\lambda,\theta) - d_k | h_k(\lambda,\theta) - d_k\rangle_k - \langle d_k|d_k\rangle_k$$

  - **Very** efficient likelihood calculation (re-uses many things) **when "intrinsic" parameters (λ) fixed** (i.e., same binary seen different ways)

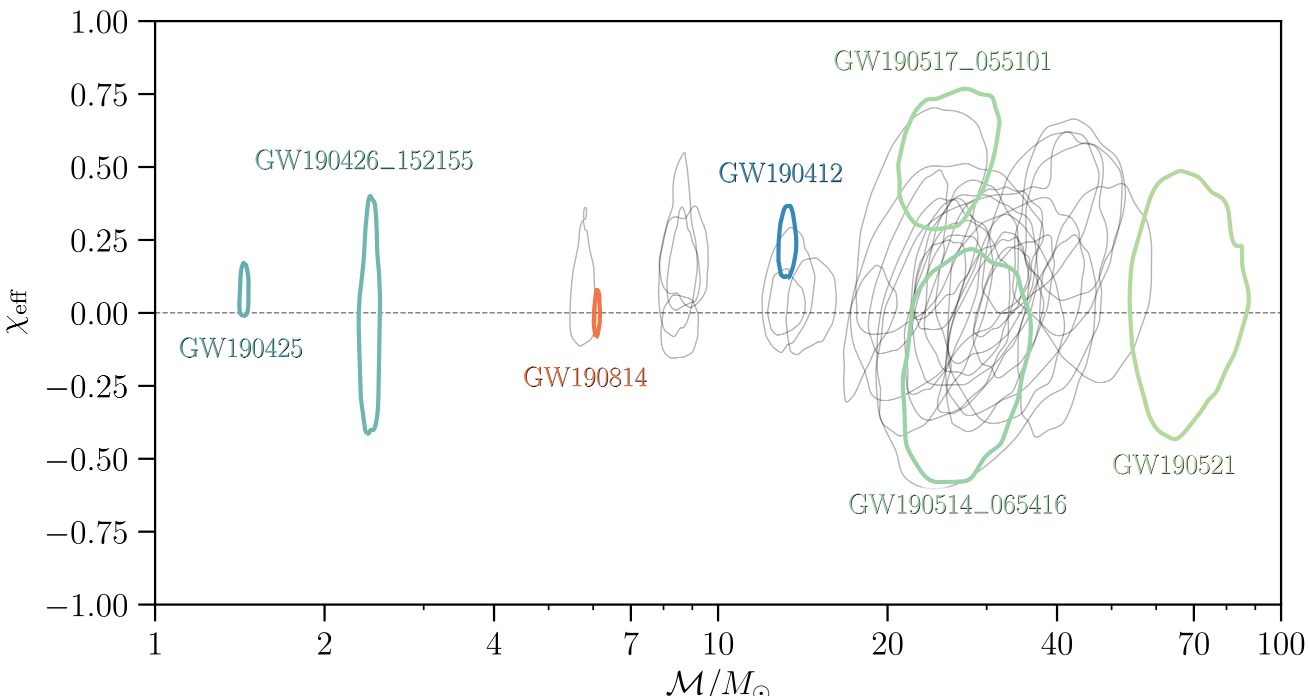
- RIFT splits the calculation into two parts: intrinsic and extrinsic

$$\mathcal{L}_{\rm marg}(\lambda) \equiv \int \mathcal{L}(\lambda,\theta)p(\theta)d\theta$$

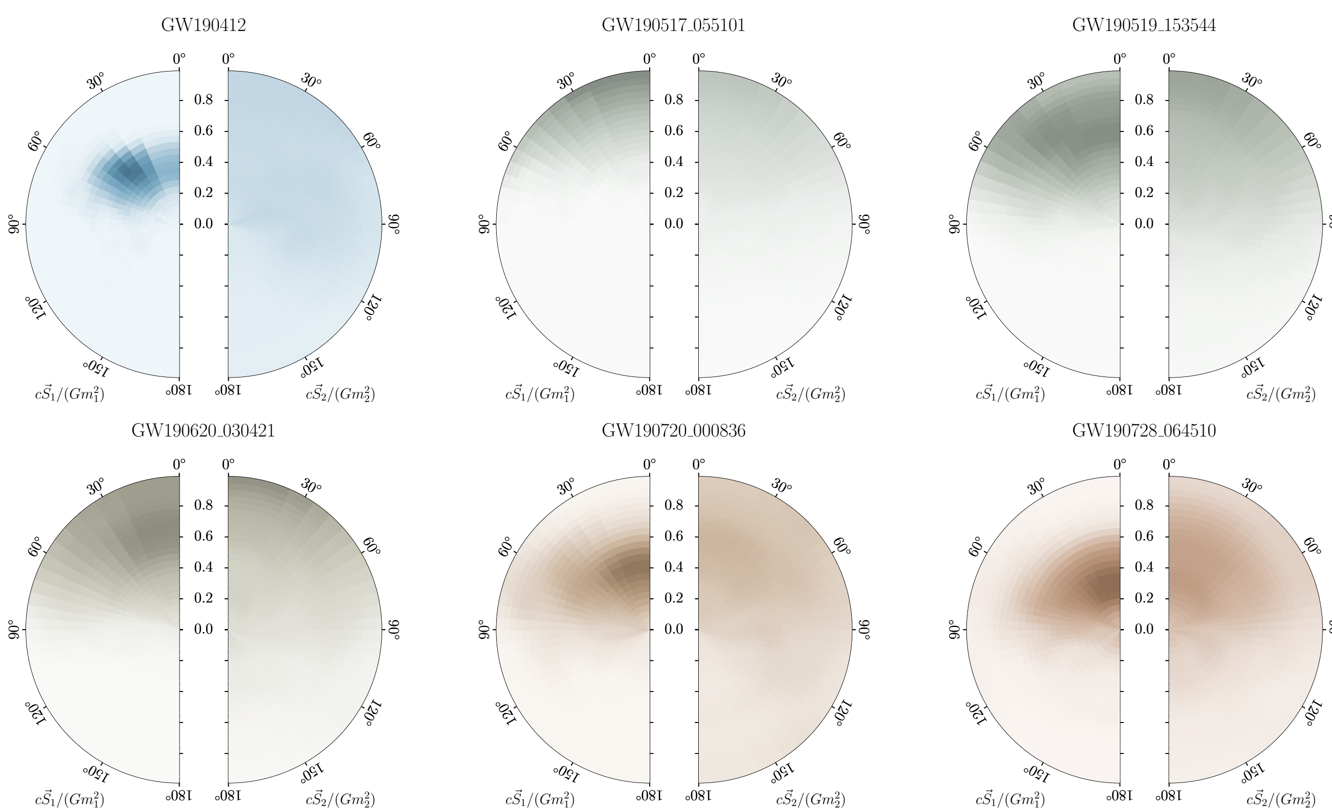
  - Iterative fitting applied to **marginal** likelihood

Why use RIFT? Very fast, cheap

- Optimized and low cost — good for expensive models
  - Example: LIGO GWTC-2 : Most HM calculations with RIFT

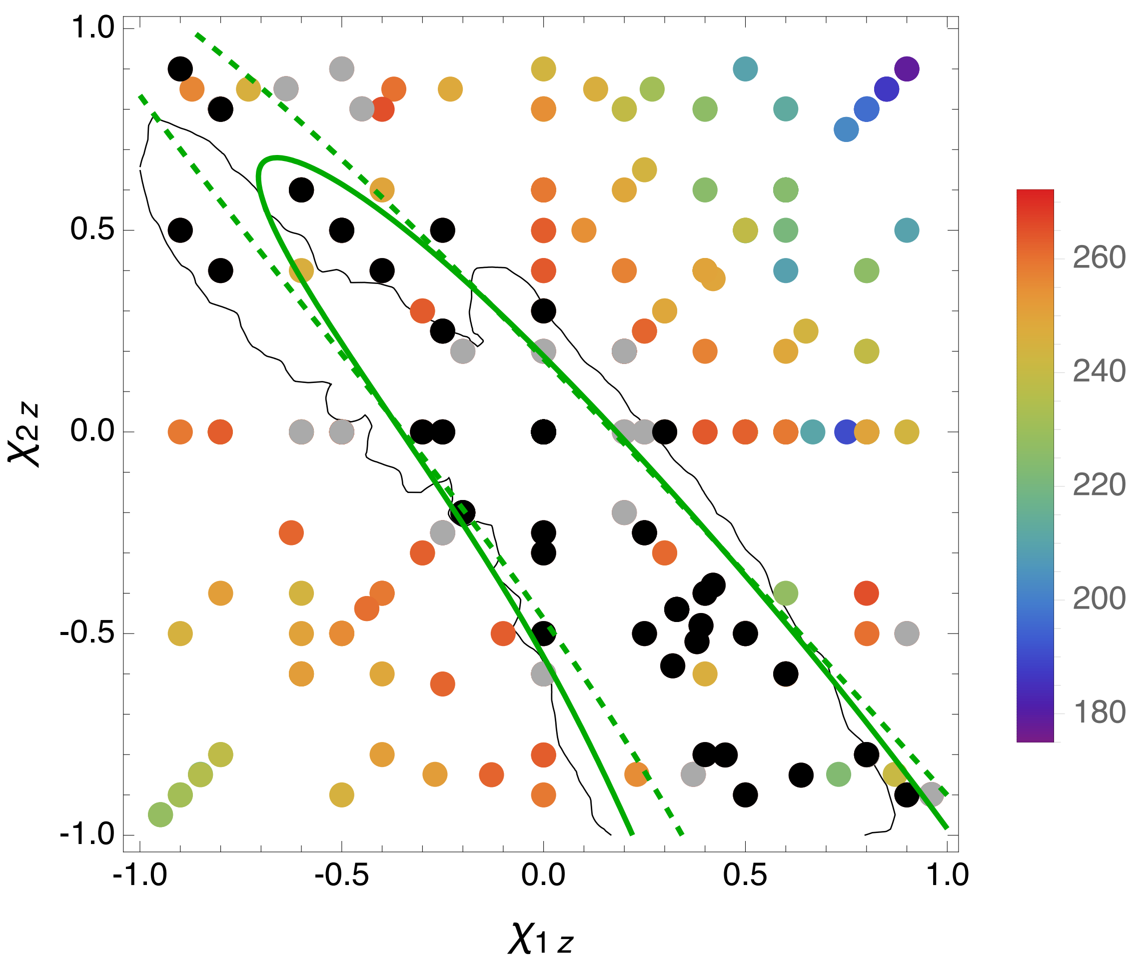


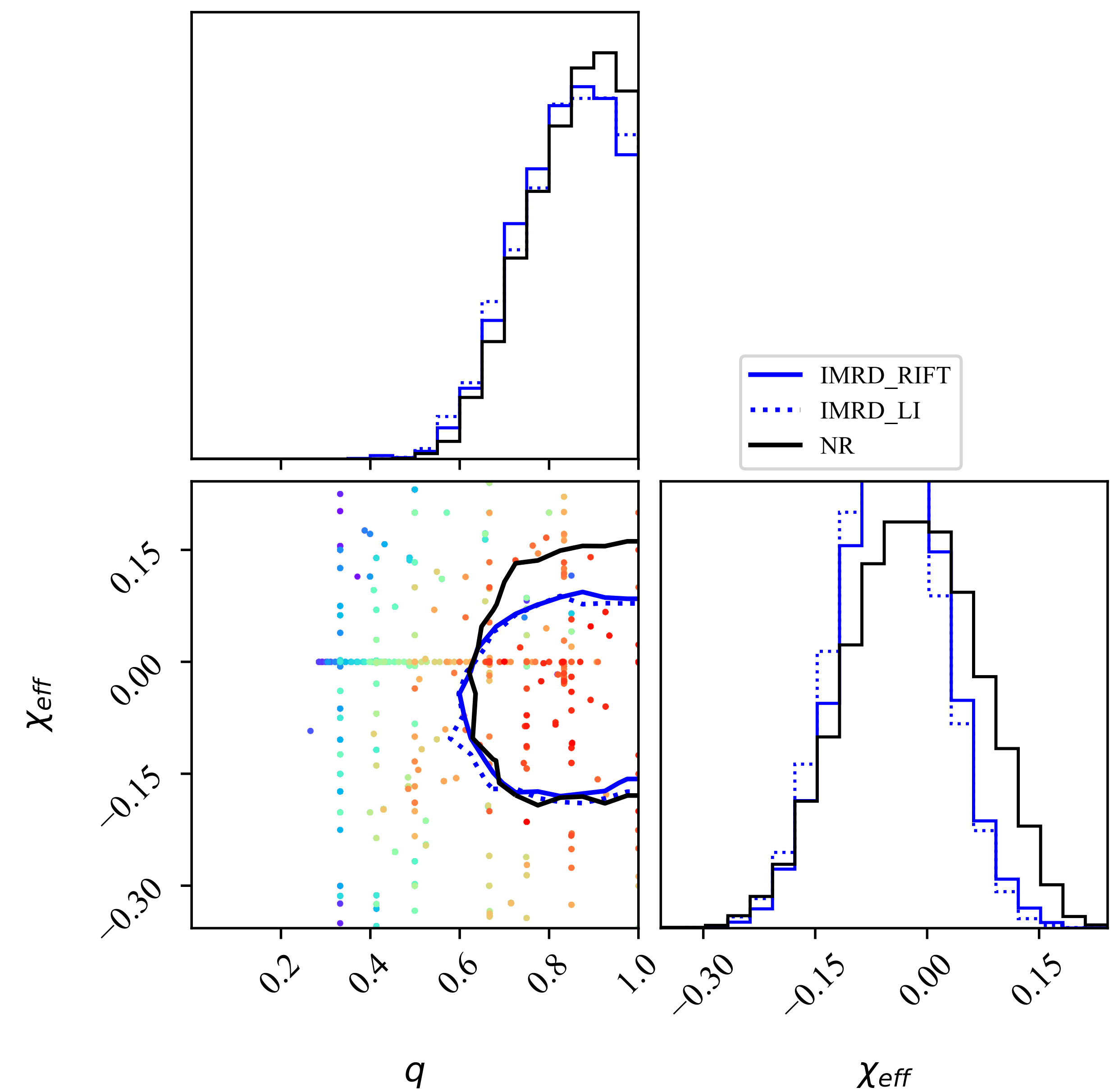
<u>GWTC-2</u>



Why use RIFT? Models not continuously available

- Some good models (NR) only available on discrete sets





Abbott et al 2016, PRD 94 4035

see also Healy et al arxiv:2010.00108

Why use RIFT? Models not continuously available

- NR: Same answer if using same priors
  - Unless low-mass (long) - sim duration
  - Even just using one group's NR sims!
- Doing NR followup studies again now

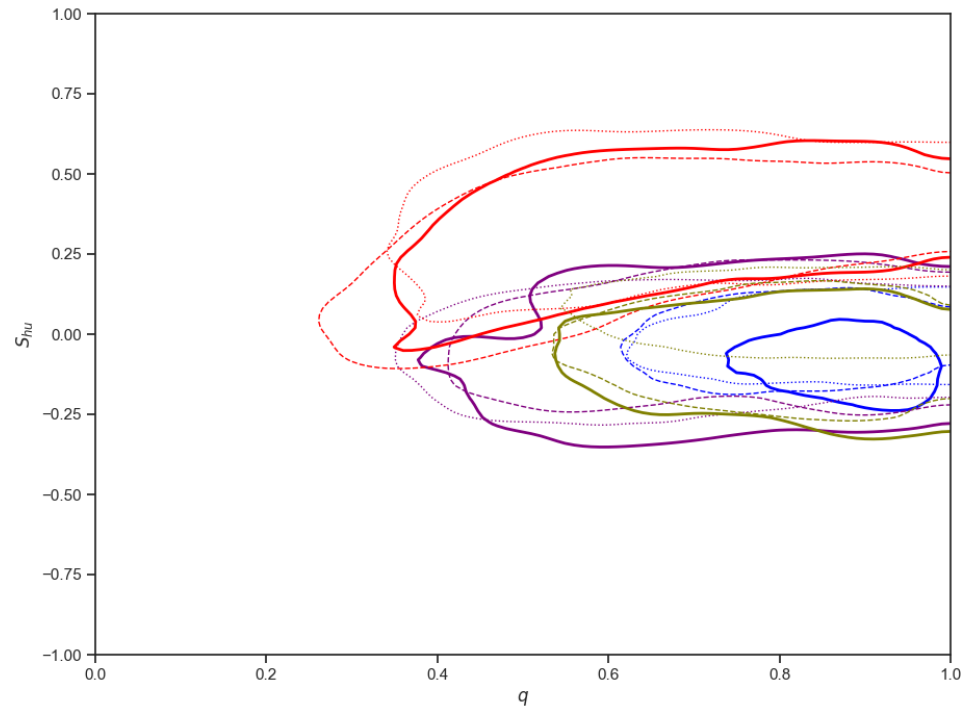


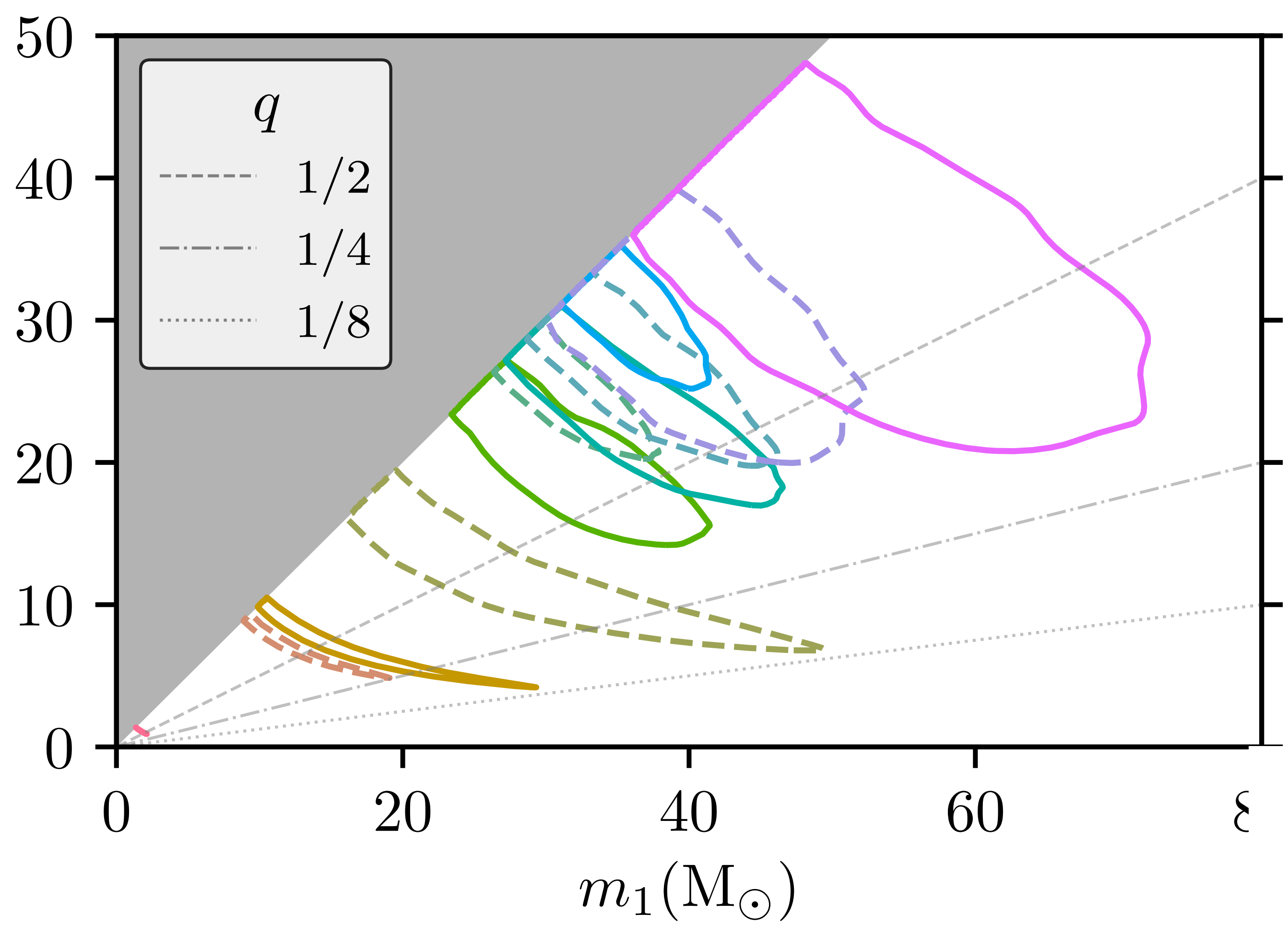
TABLE IV. Jensen-Shannon divergence for the mass-ratio, spin magnitudes, and total mass in the detector frame between our preferred analysis with uniform priors and GWTC-1 LIGO posteriors (first number in each column) and between the CIP analysis and GWTC-1 using the same priors (second number in each column).

Event	q	a_1	a_2	$M_{total}/M_\odot$
GW150914	0.1501, 0.0201	0.4061, 0.0203	0.4453, 0.0113	0.0135, 0.0147
GW151012	0.0263, 0.0167	0.1616, 0.0029	0.1999, 0.0271	0.0840, 0.0461
GW151226	0.0551, 0.0287	0.1478, 0.0038	0.1279, 0.1193	0.2004, 0.3512
GW170104	0.2758, 0.0186	0.2954, 0.0033	0.2615, 0.0012	0.0280, 0.0188
GW170608	0.0261, 0.0106	0.2249, 0.0387	0.2465, 0.0804	0.3281, 0.1553
GW170729	0.0118, 0.0478	0.0562, 0.0070	0.0545, 0.0224	0.0041, 0.0699
GW170809	0.1145, 0.0056	0.2846, 0.0060	0.2221, 0.0004	0.0352, 0.0234
GW170814	0.2945, 0.0508	0.2228, 0.0042	0.5244, 0.0023	0.6033, 0.3592
GW170818	0.0704, 0.0104	0.1671, 0.0003	0.1278, 0.0054	0.0439, 0.0229
GW170823	0.0506, 0.0137	0.1644, 0.0006	0.1450, 0.0038	0.0291, 0.0366

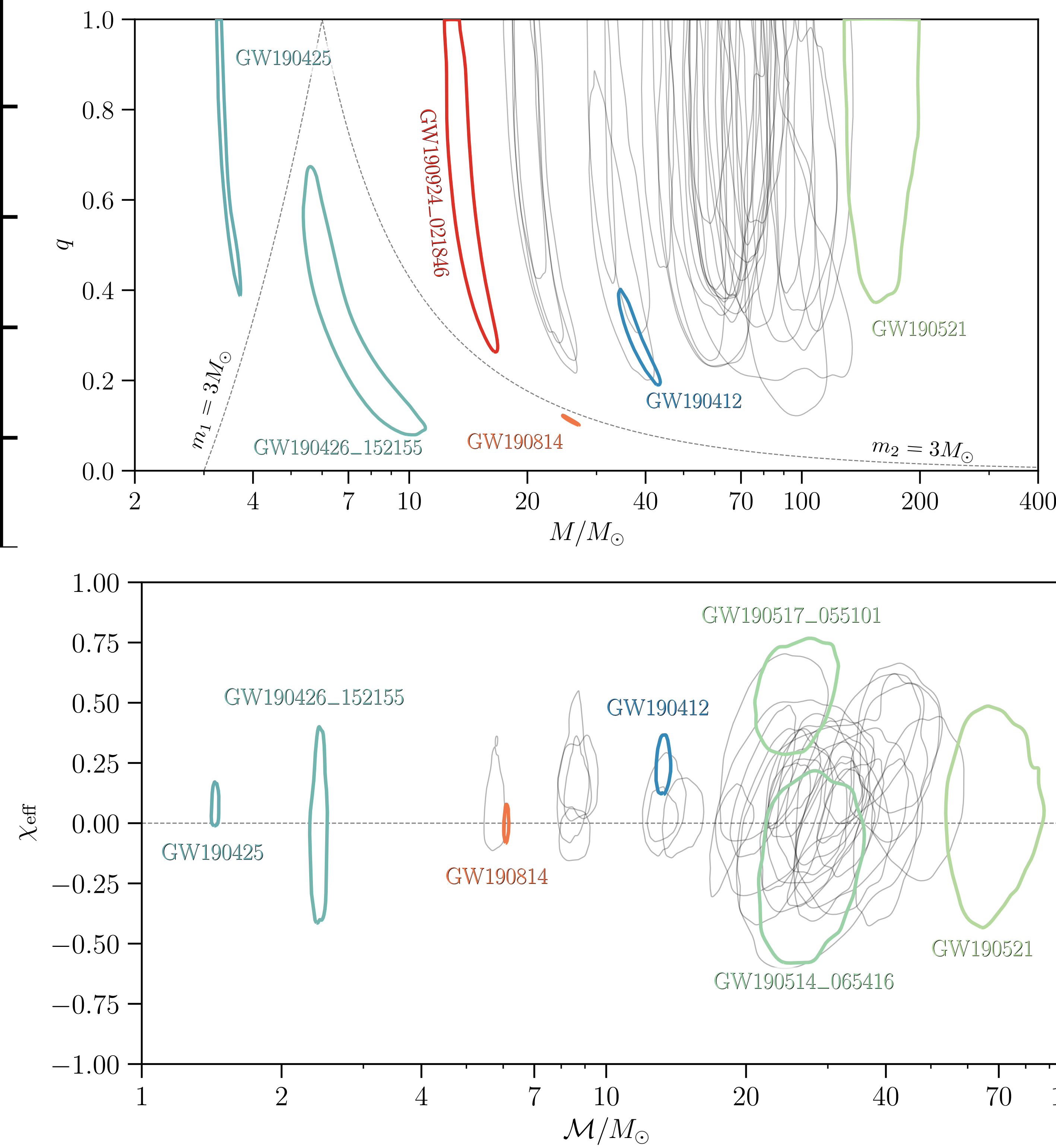
Healy et al arxiv:2010.00108

Inference challenge 2: GW source populations

Abbott et al 1811.12907

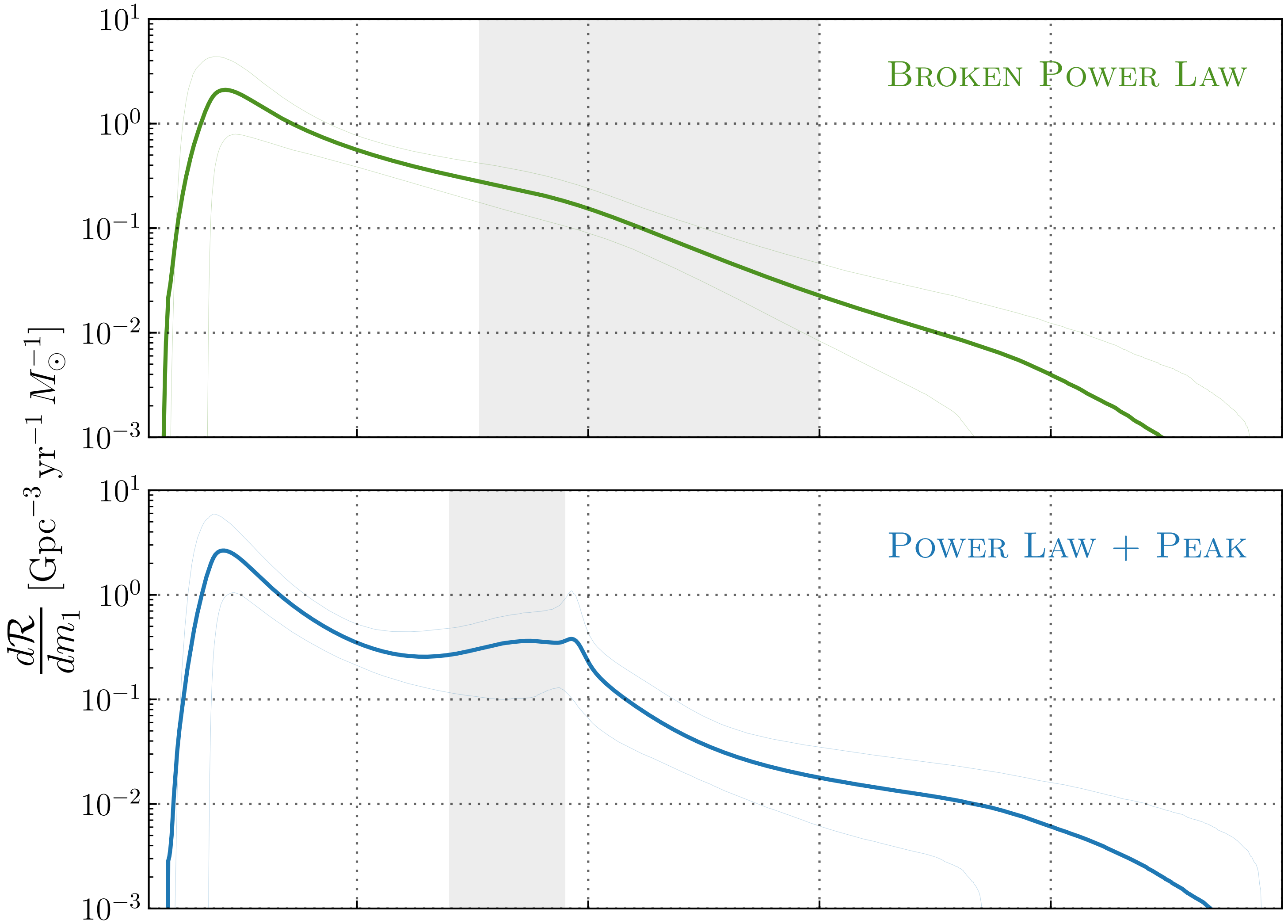


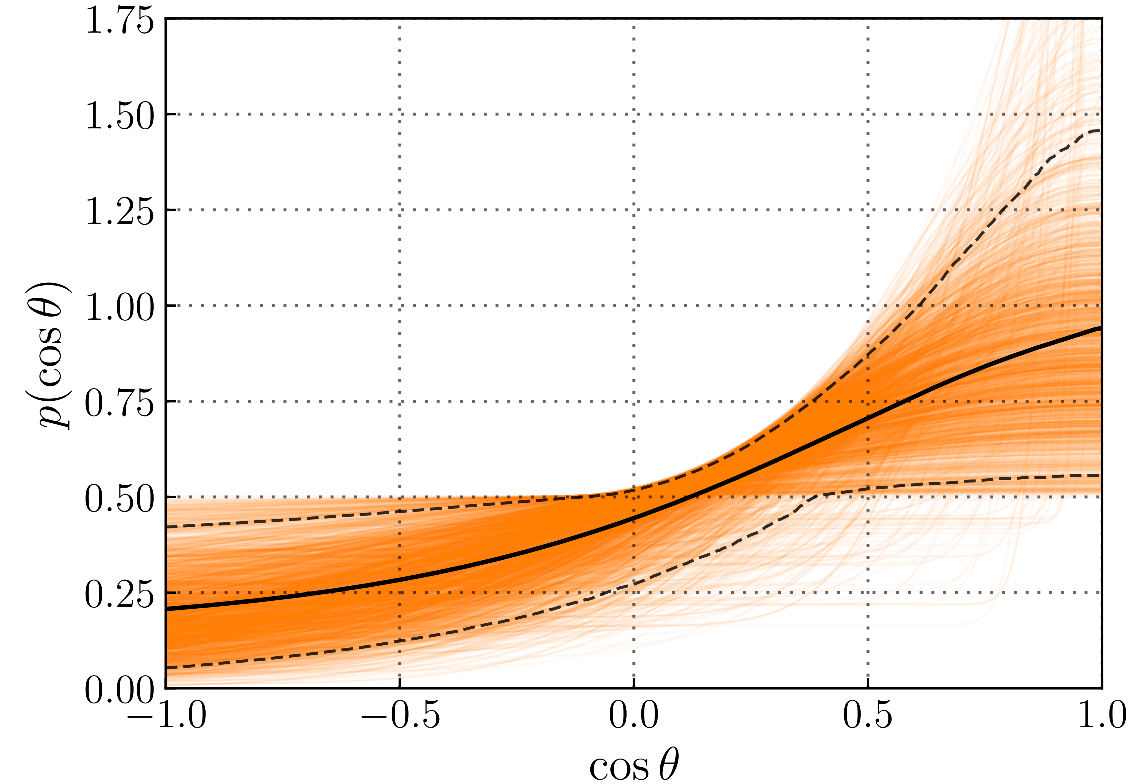
GWTC-2

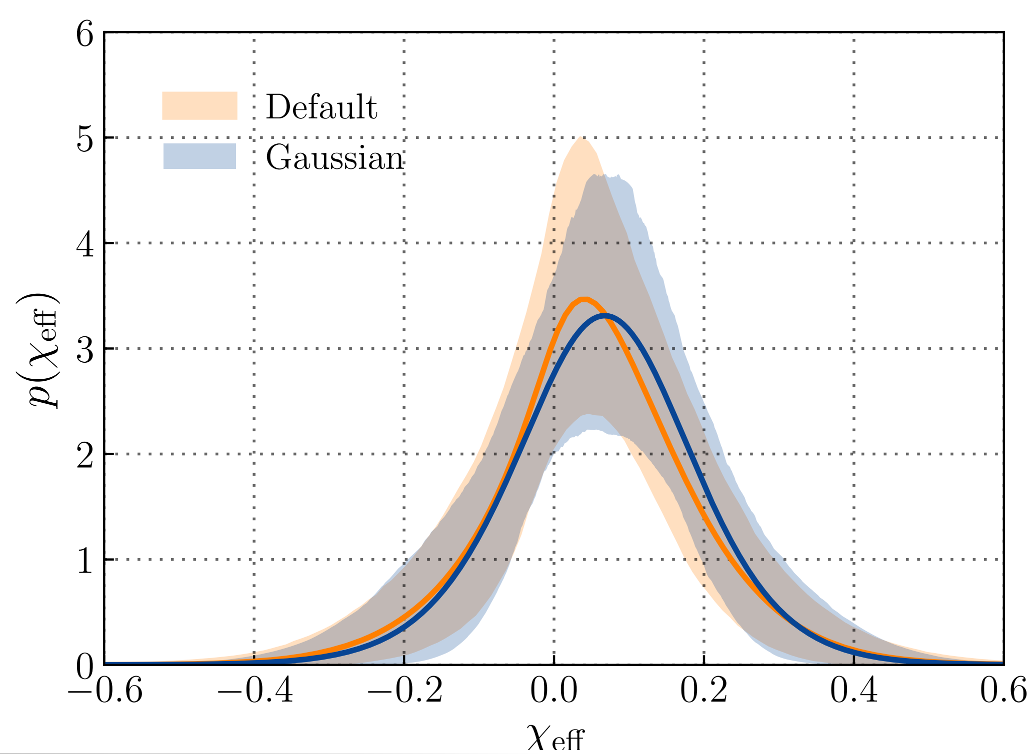


…how to infer population responsible

GWTC-2







Joint likelihood straightforward

- Joint likelihood for all selected events [inhomogeneous poisson]

Observation Model prediction

$$\mathcal{L} = e^{-\mu(\Lambda)} \prod_k \mathcal{R} \int d\lambda_k p(d_k|\lambda_k) p(\lambda_k|\Lambda)$$

$$p(\Lambda, \mathcal{R}) \propto \mathcal{L}(\Lambda, R) p(\Lambda, \mathcal{R})$$

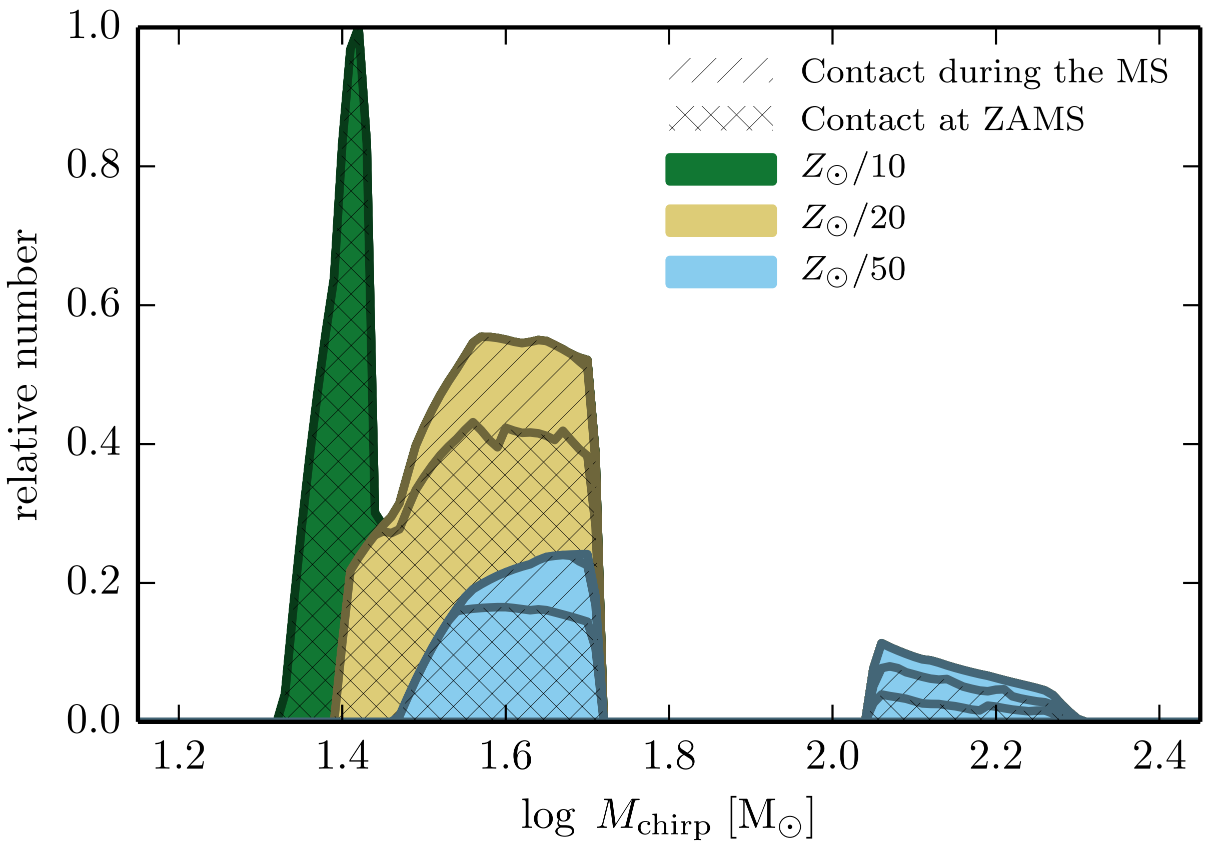
- Problem is **model availability**:
  - Simple analytic models easy [eg, Wysocki et al 2018 [PopModels] arXiv:1805.06442]
  - Real simulations (e.g., binary evolution) hard

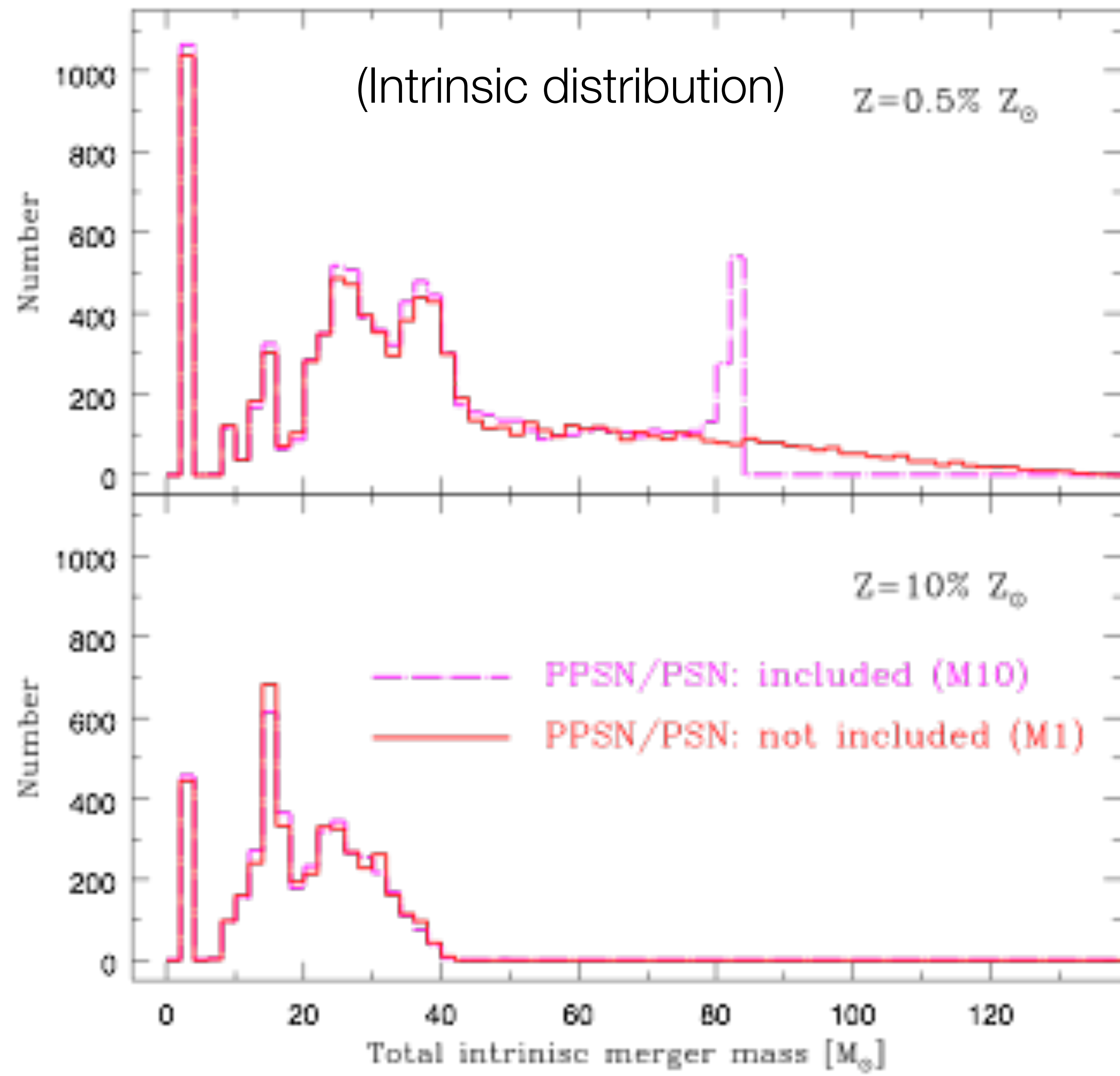
Wysocki et al 2018 PRD 97 3014

Observed distribution reflects formation physics

- Mass distribution example: sharp features from SN

Marchant 2017



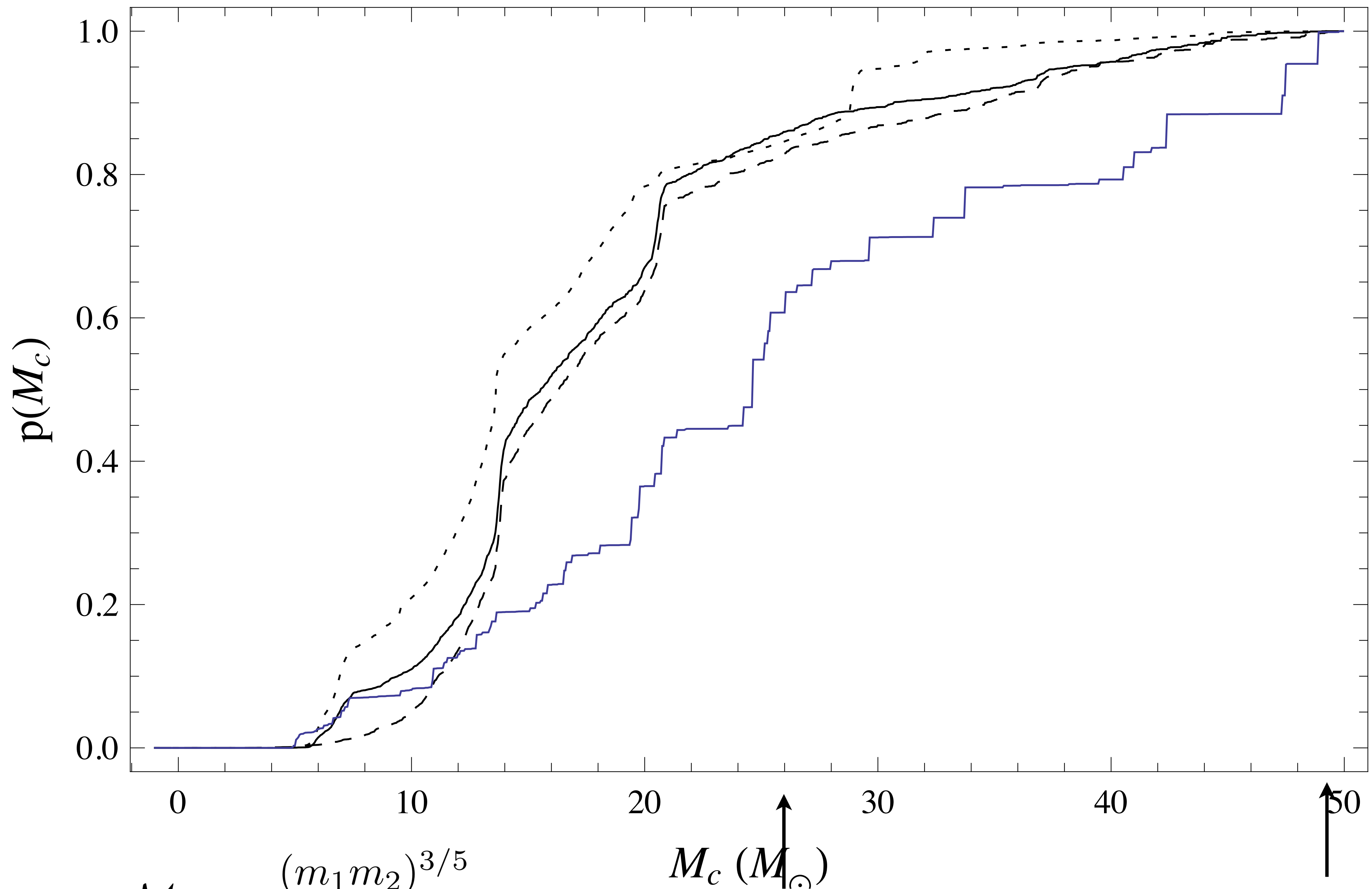


Belczynski et al 1607.03116

Distributions vary significantly…

(**Detected** distribution)

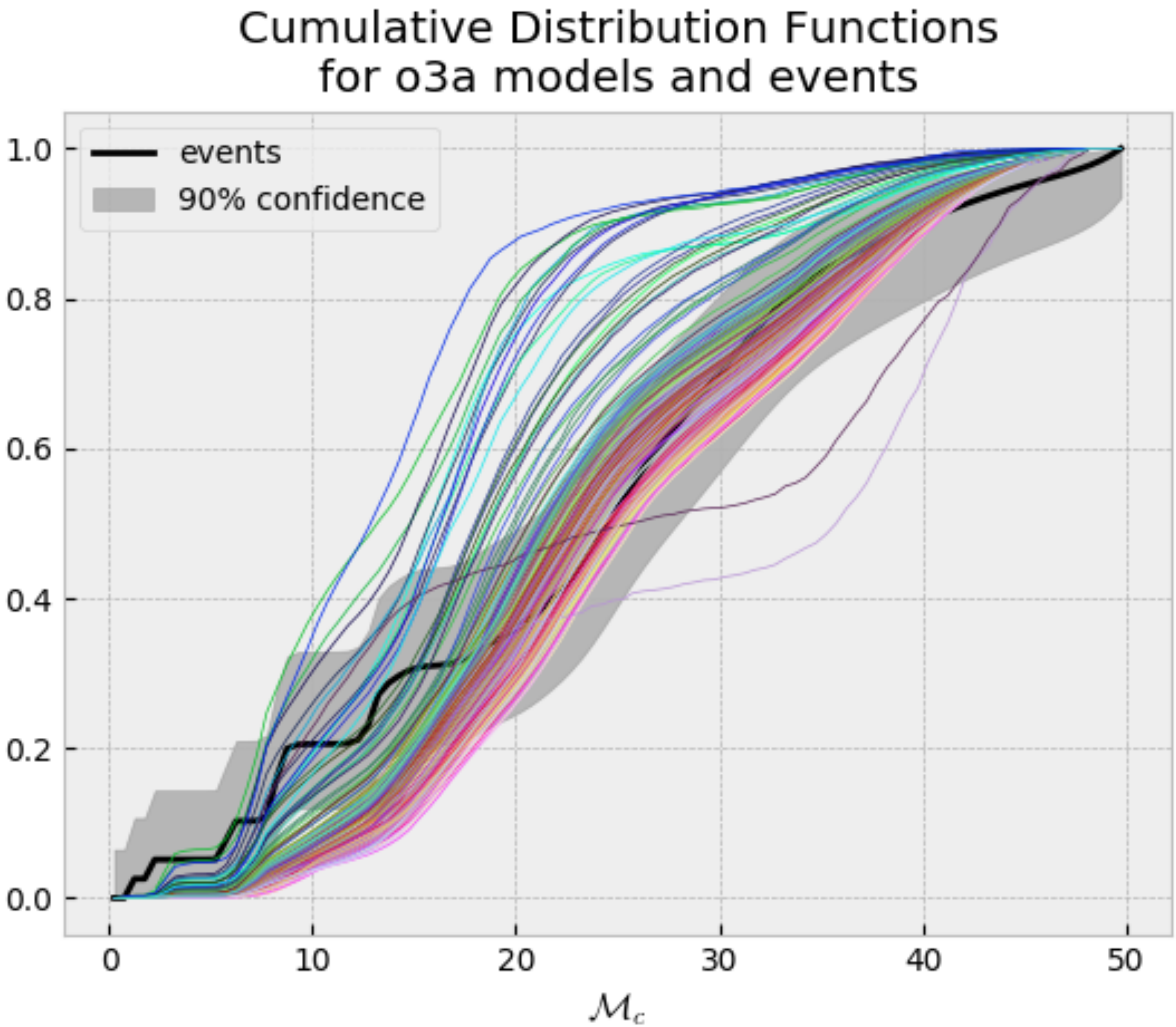
Dominik et al (2015: 1405.7016)



$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}$$

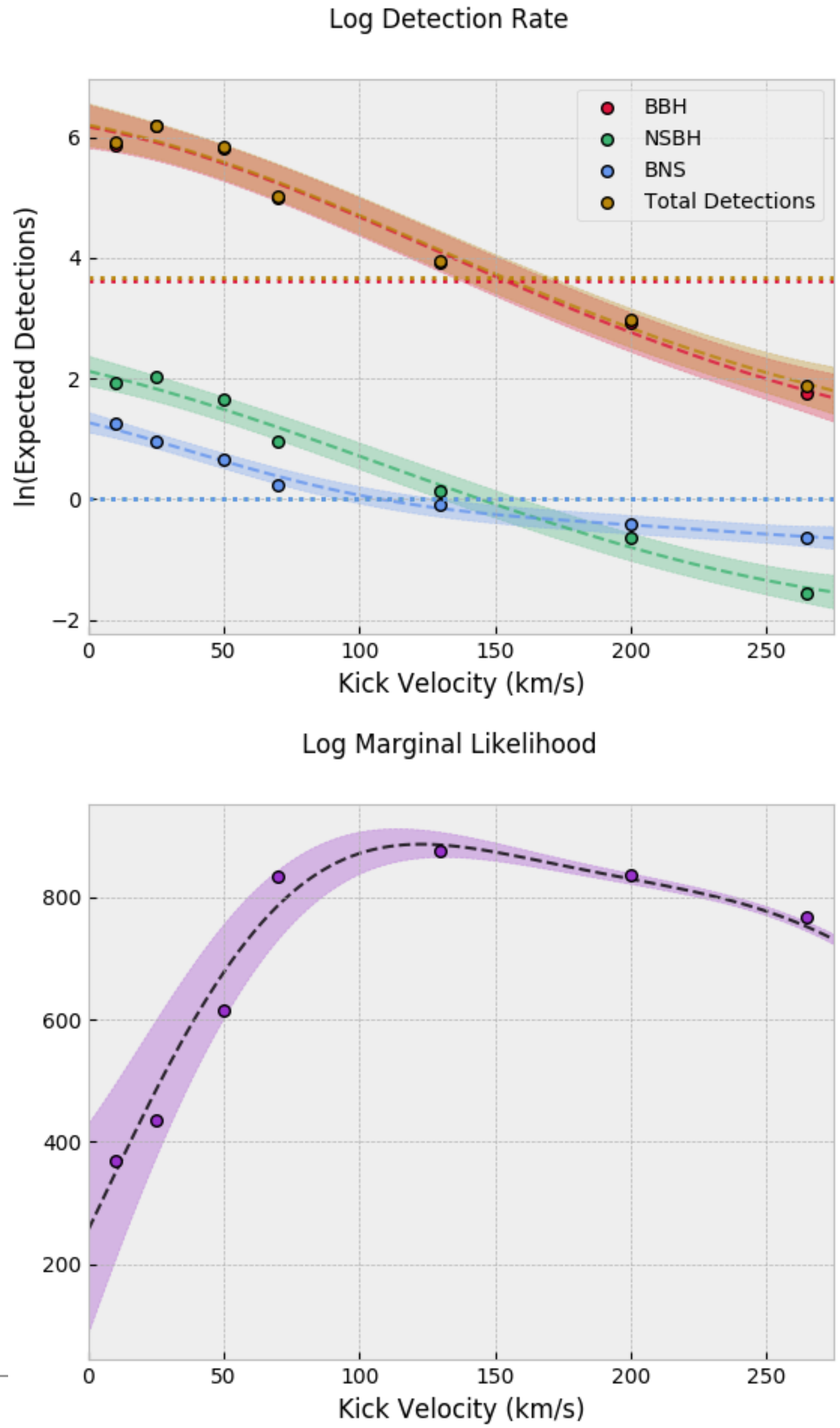
Distributions vary significantly…

Delfavero et al (in prep)

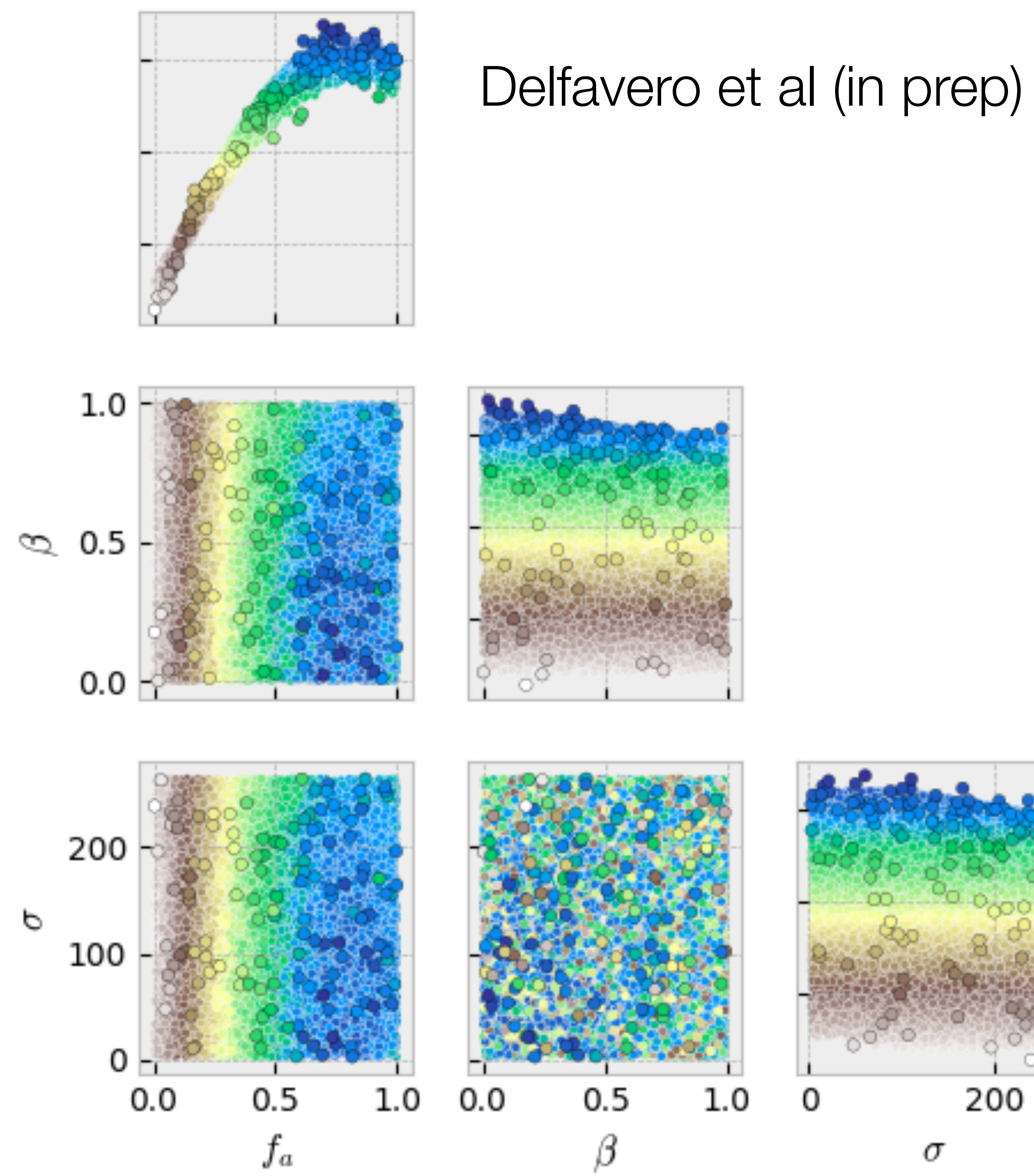


Likelihood fitting for populations

Versus kick alone



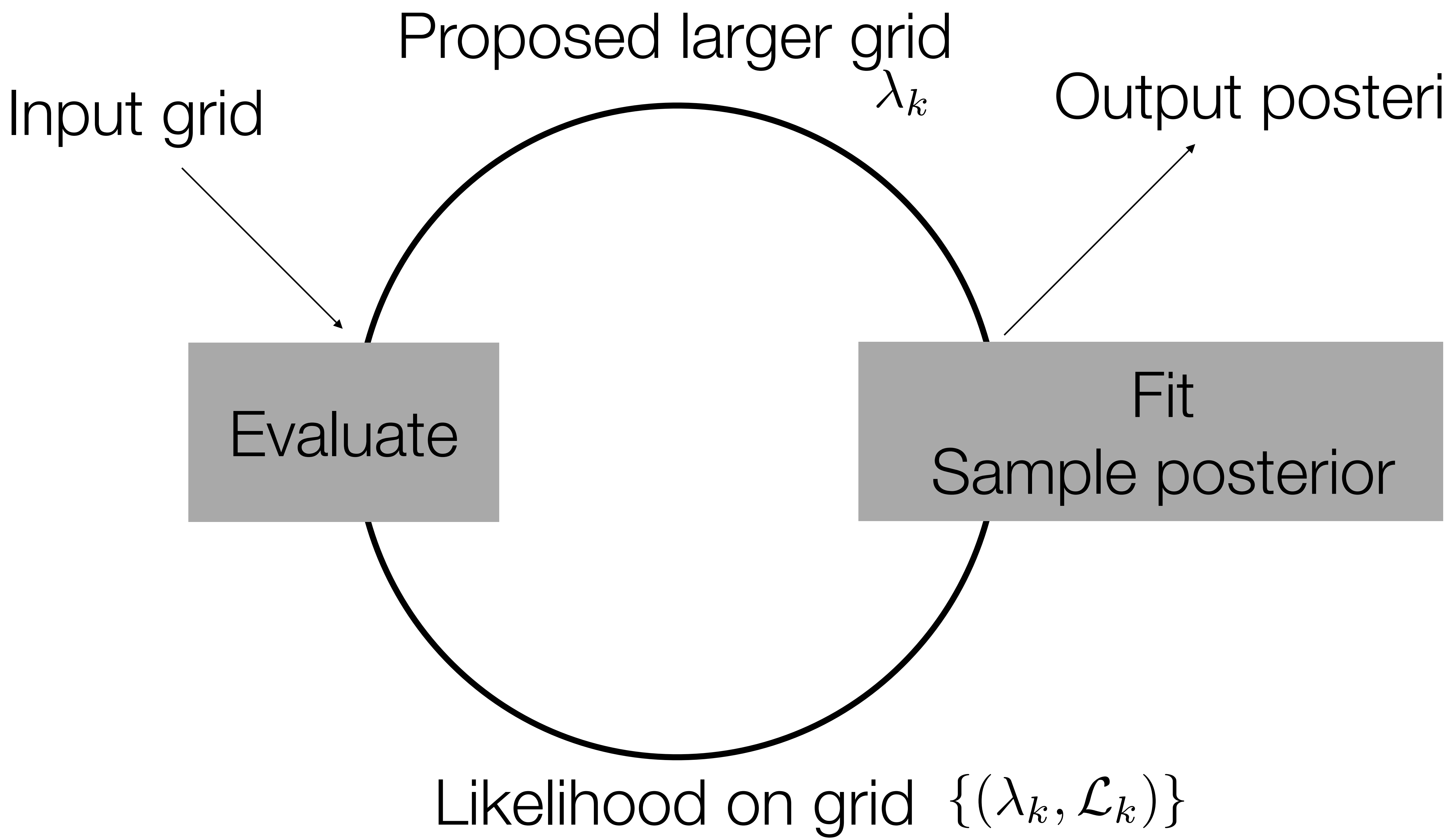
BBH detection rate (3d)



Delfavero et al (in prep)

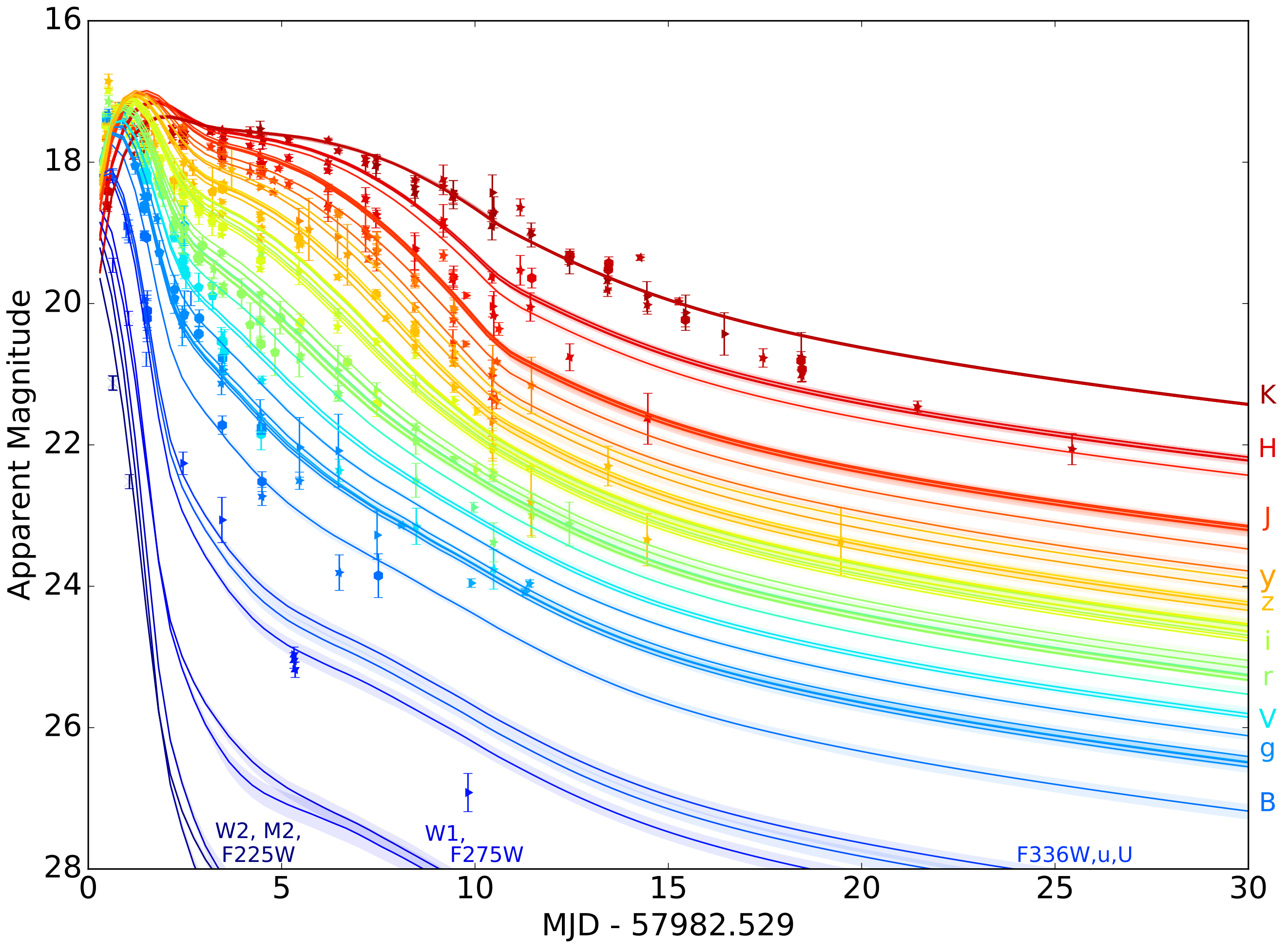
Iterative Fitting (for inference)

- **Repeat until stable**



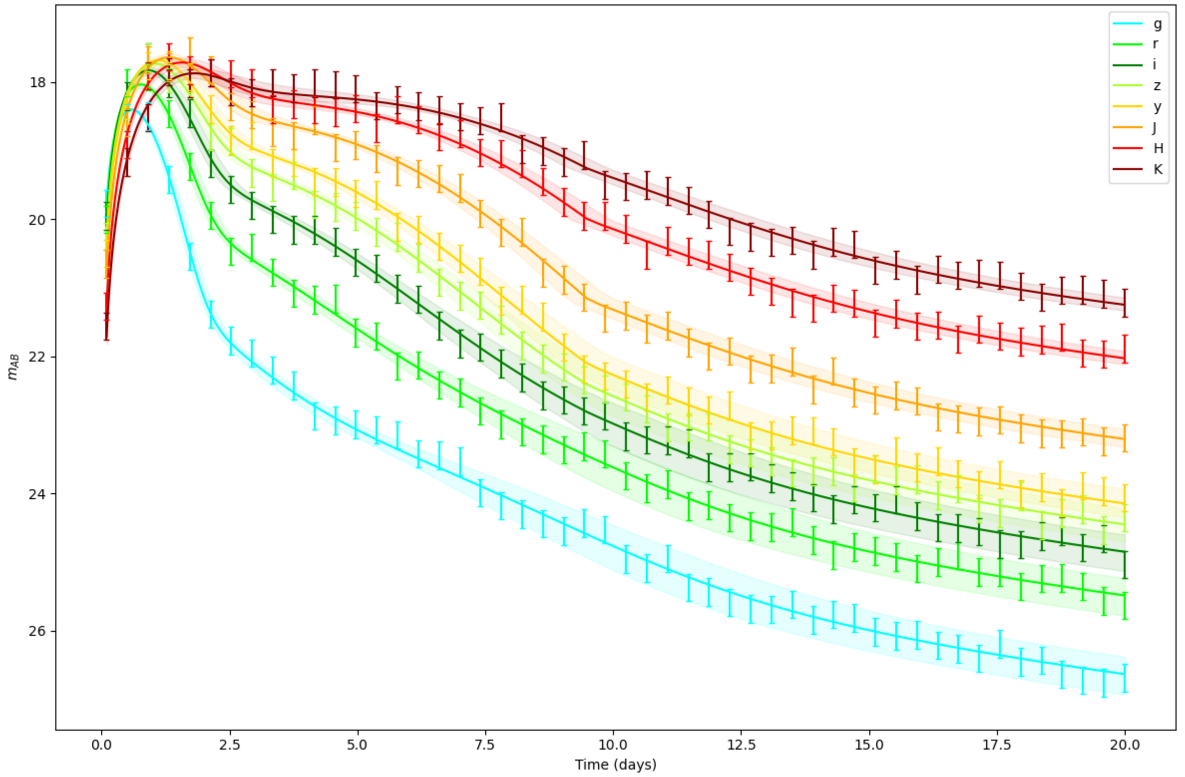
Inference challenge 3: Kilonova inference

GW170817



Villar et al 1710.11576

Synthetic kilonova

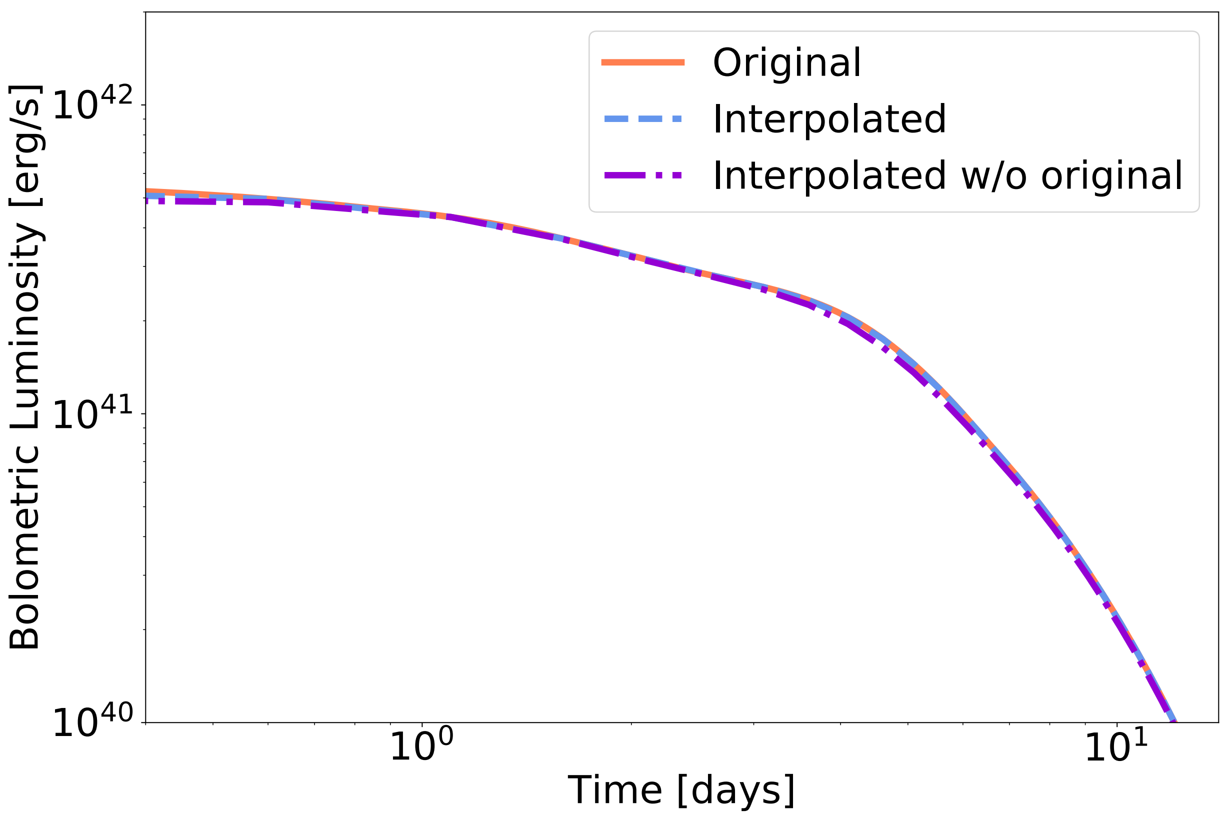


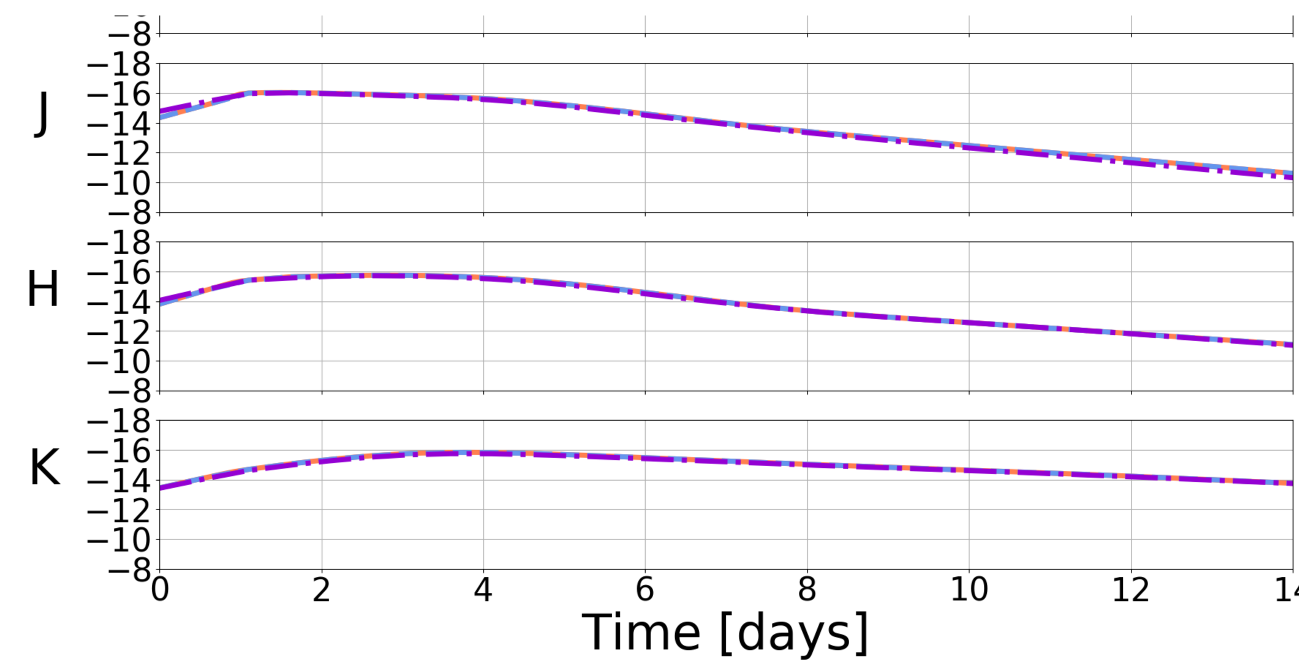
B. Champion

Interpolating simulations

- Spherically symmetric models historically used (analytic and not)
- First simulation surrogate: GP for single-component ejecta (m,v,X)
    - Grid 7x5x6

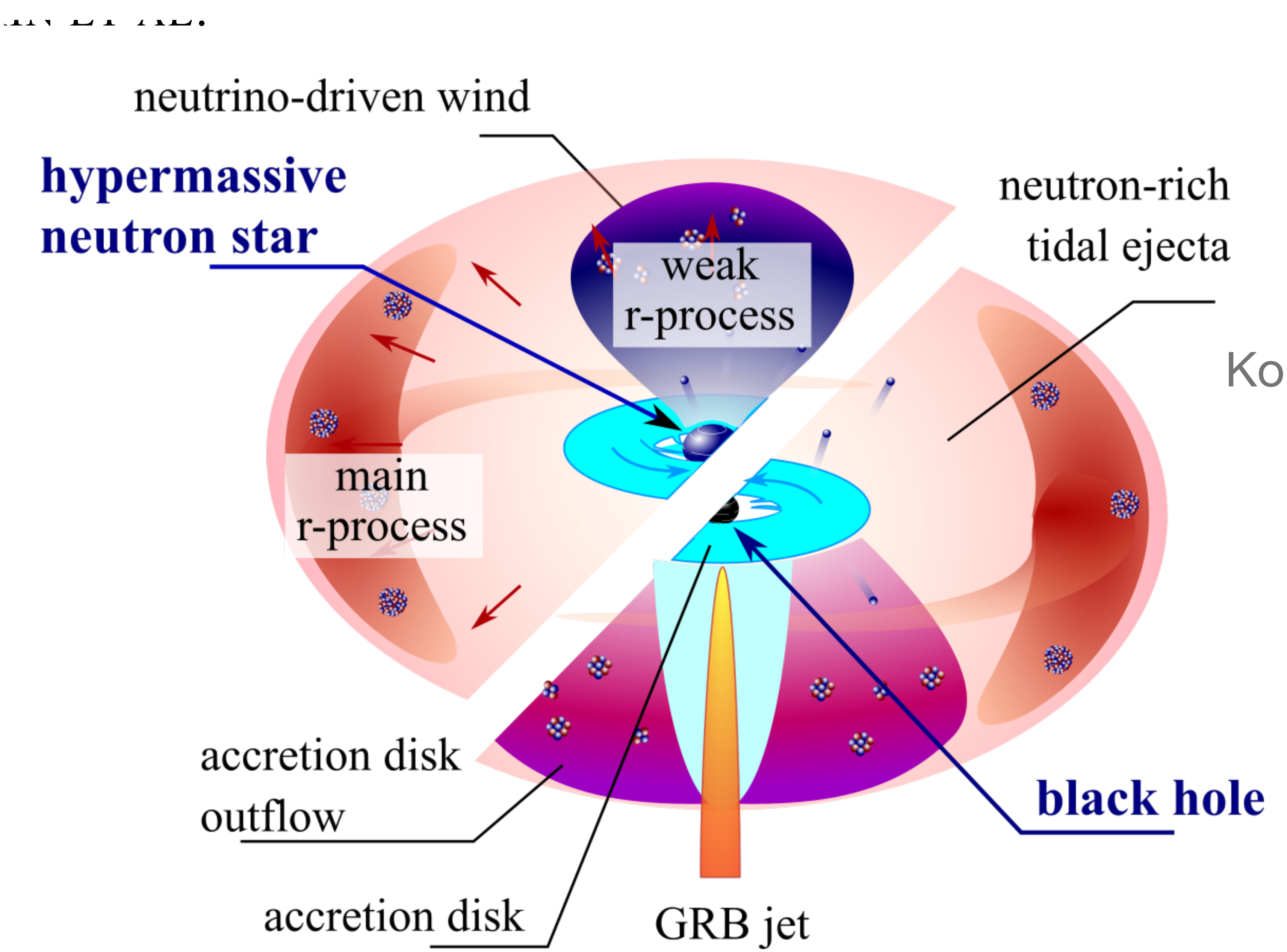
M. Coughlin et al 2018 arxiv:1805.09371





- Two-component by **flux addition (!!)**

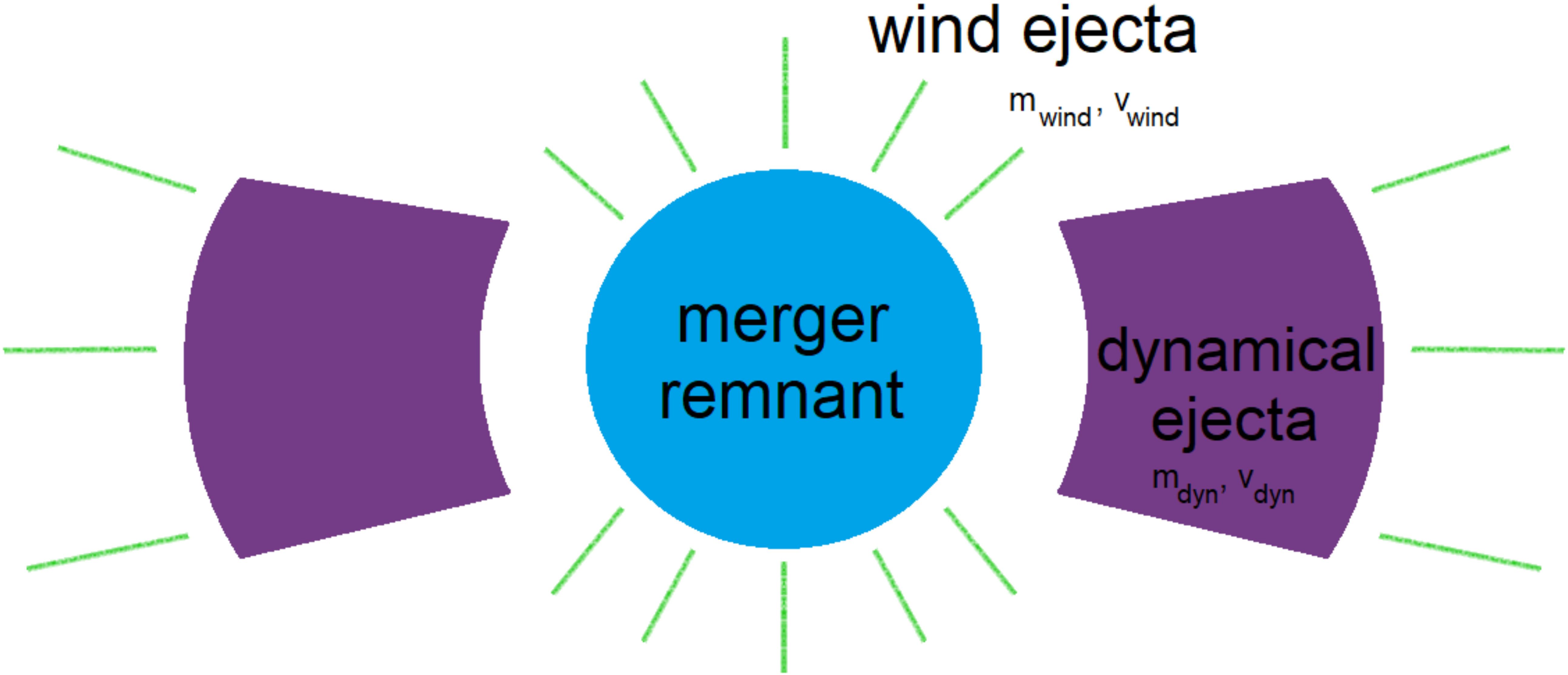
Real kilonovae are not spherically symmetric



Korobkin et al 2004.00102

- **Orientation dependence** and **reprocessing of multiple components** matters [eg Korobkin et al 2004.00102; J. Miller et al 1905.07477]

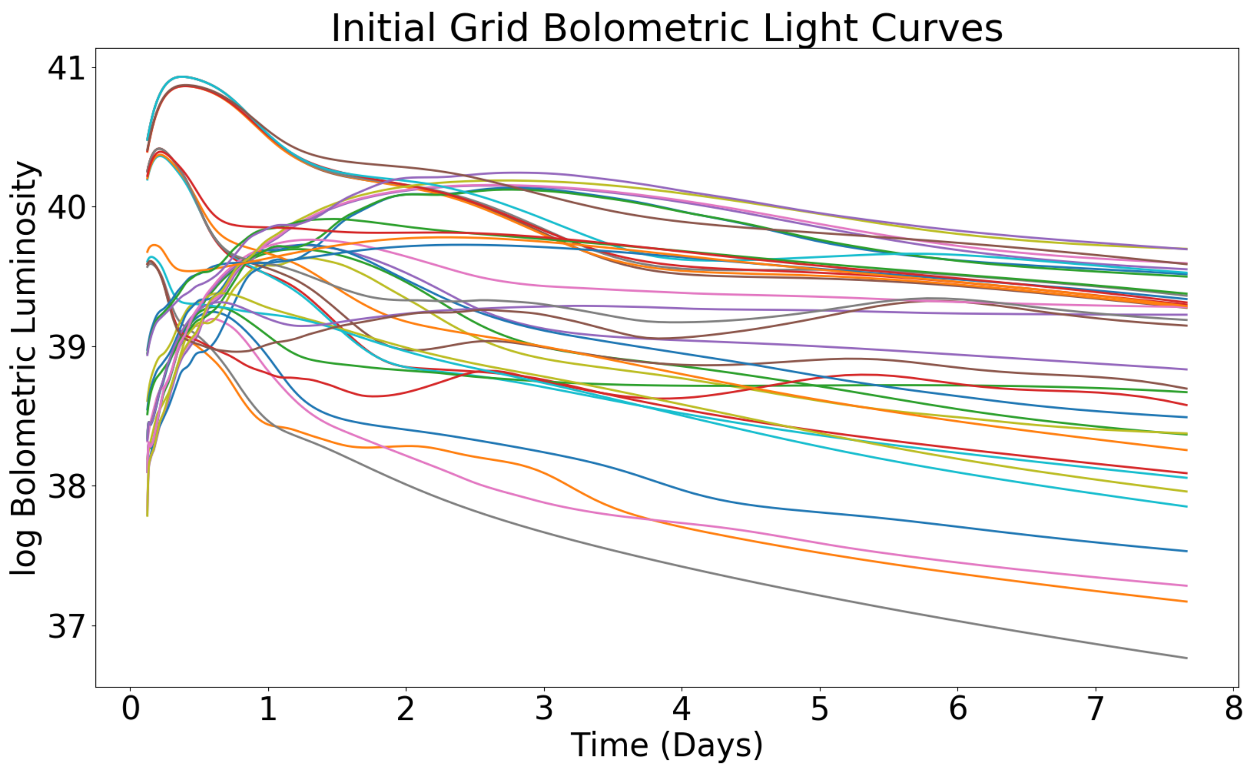
Real kilonovae are not spherically symmetric

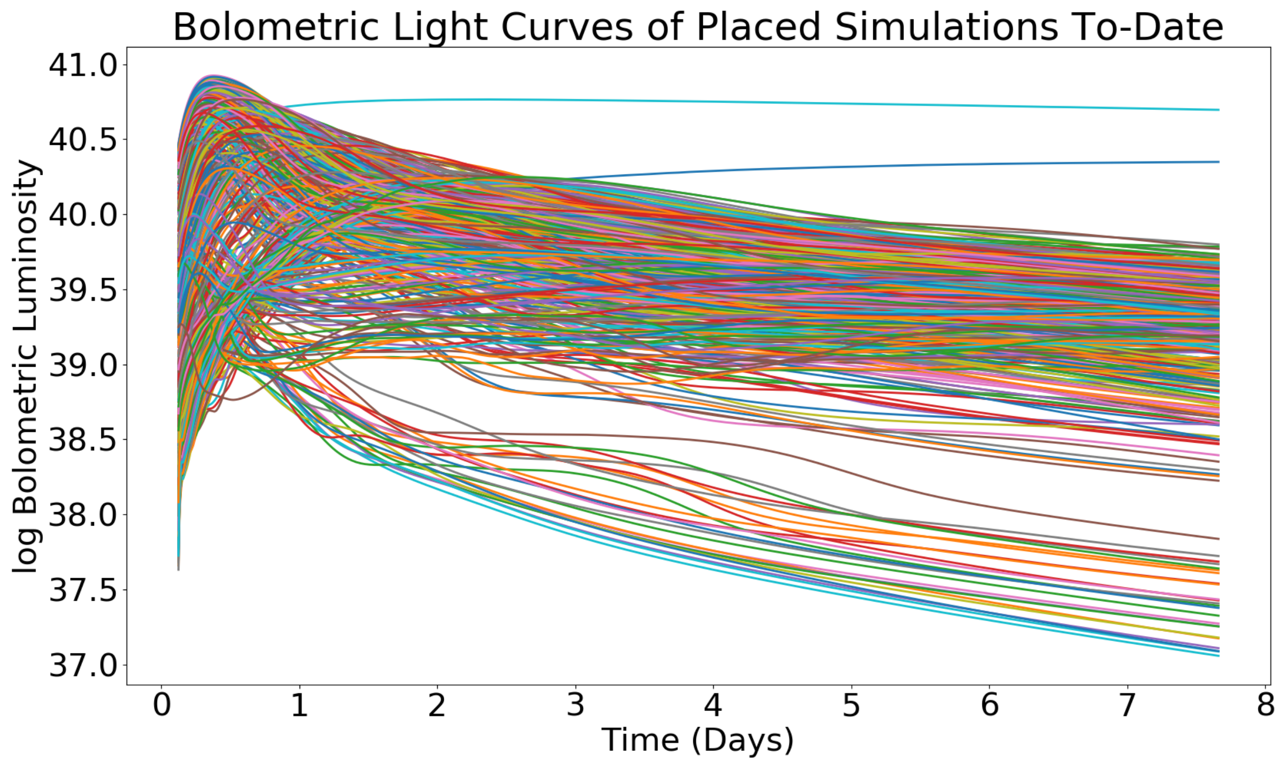


- **Orientation dependence** and **reprocessing of multiple components** matters [eg Korobkin et al 2004.00102; J. Miller et al 1905.07477]

Adaptive simulation placement

- Two-component anisotropic model (m1,m2,v1,v2)
- GP and GP error to target new sets of two-component sims
  - Uses **limited** data (orientation-averaged, etc) so **GP can run**
- Running continuously, automatically

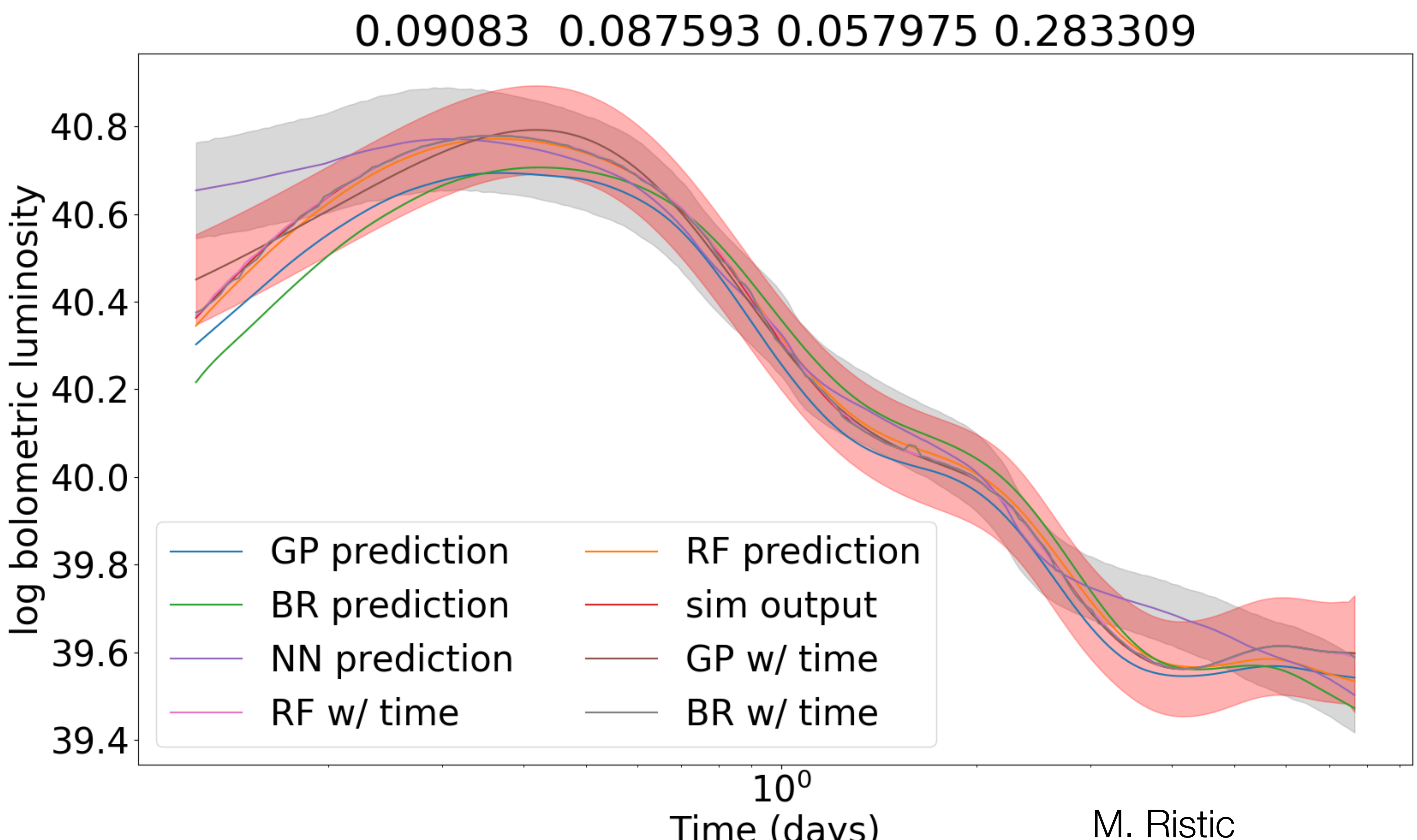




M. Ristic

Surrogate light curves

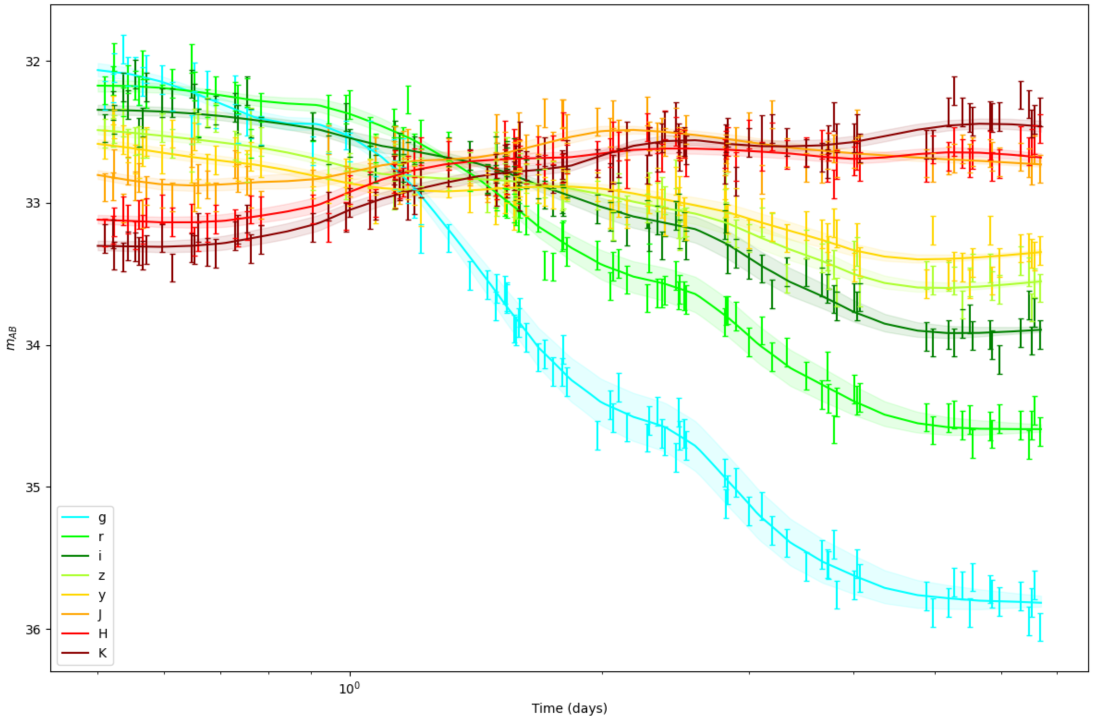
- Several surrogate methods, some fast and scaling well to large training sets



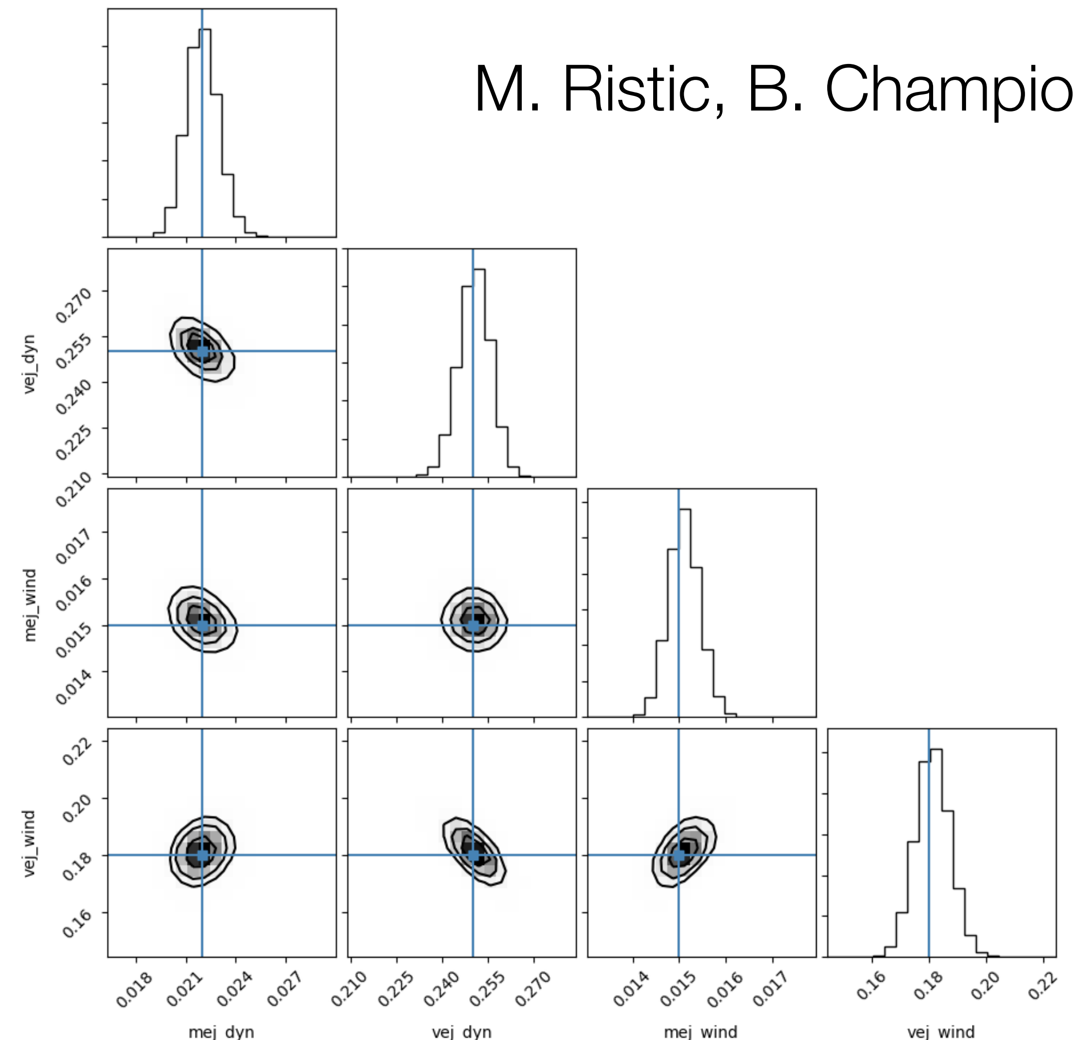
M. Ristic

Surrogate light curves can be used for inference

- Plug into inference code



M. Ristic, B. Champion



Simulations needed, very sensitive to many inputs

- Composition (Even et al 1904.13298)

- Nuclear physics (Zhu et al 2010.03668)

- Shape of outflow (Wollaeger et al; Korobkin et al 2004.00102; …)

Lessons learned

- Gaussian processes **do not scale well**
  - Great for small problems, very hard to apply to large training sets (n^3 scaling)
  - Moving towards other methods
- Surrogates [likelihood or model]
  - Domain expertise hugely important to pick “slow” degrees of freedom/ parameterization. Fit more easily to slowly-varying quantities.
- Likelihood interpolation
  - Easy, robust … but data-specific, no surrogate **model** at end
- Simulation placement
  - Hard simulations need **optimal** placement…

References

- Surrogate likelihoods for GW parameter inference: RIFT [J. Lange]
  - RIFT
    - arxiv:1805.10457
    - arxiv:1902.04934
  - NR applications
    - https://arxiv.org/abs/1606.01262
    - https://arxiv.org/abs/1705.09833
    - https://arxiv.org/abs/2010.00108

- Surrogate likelihoods for population inference [D. Wysocki, V. Delfavero]

  Wysocki et al 2018 PRD 97 3014
  Delfavero et al in prep

- Surrogate kilonova light curves [M. Ristic]
  - Ristic: RIT thesis
  - M. Coughlin et al 2018 arxiv:1805.09371

Inference notation

- Notation

$$Z(d|H_1) \equiv \frac{p(\{d\}|H_1)}{p(\{d\}|H_0)} = \int d\lambda p(\vec{\lambda}|H_1)\frac{p(\{d\}|\vec{\lambda},H_1)}{p(\{d\}|H_0)}$$

posterior distribution

H_1 : with signal
H_0 : no signal

- Inputs:
  - Prior knowledge $p(\lambda|H_1)$ about distribution of λ
  - Signal model $h(\lambda)$
  - Noise model $p(\{d\}|H_0)$ $p(\{d\}|\vec{\lambda},H_1) = p(\{d-h(\vec{\lambda})\}|H_0)$
  - Data
  - Algorithm for integral/exploration in many dimensions

- Noise model: Gaussian

$$\langle n^*(f)n(f)\rangle = \frac{1}{2}S_h(|f|)\delta(f-f')$$

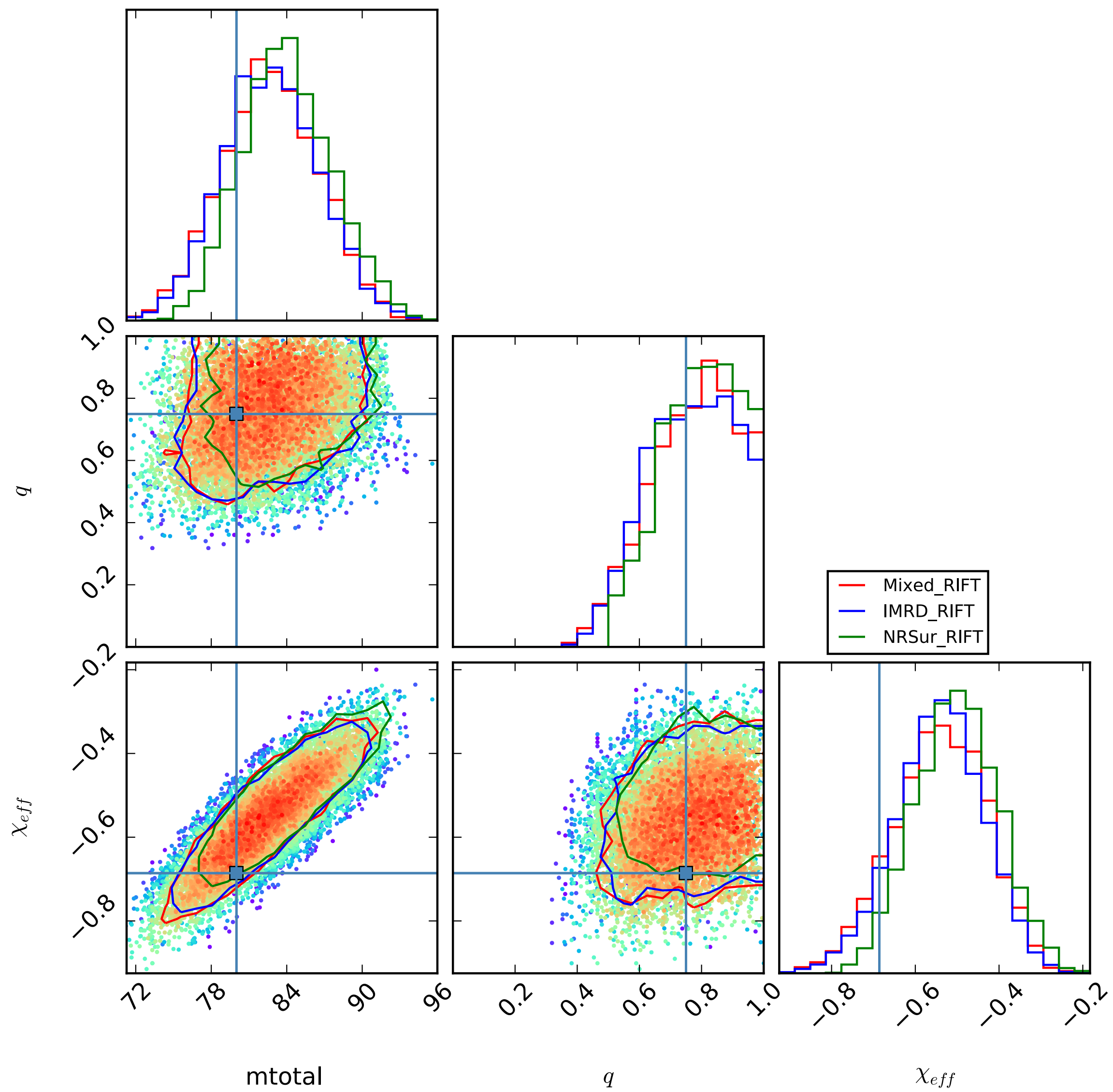
$$p(\{d\}|H_0) \propto \exp -\frac{\langle d|d\rangle}{2} dd_1 dd_2 \ldots dd_N$$

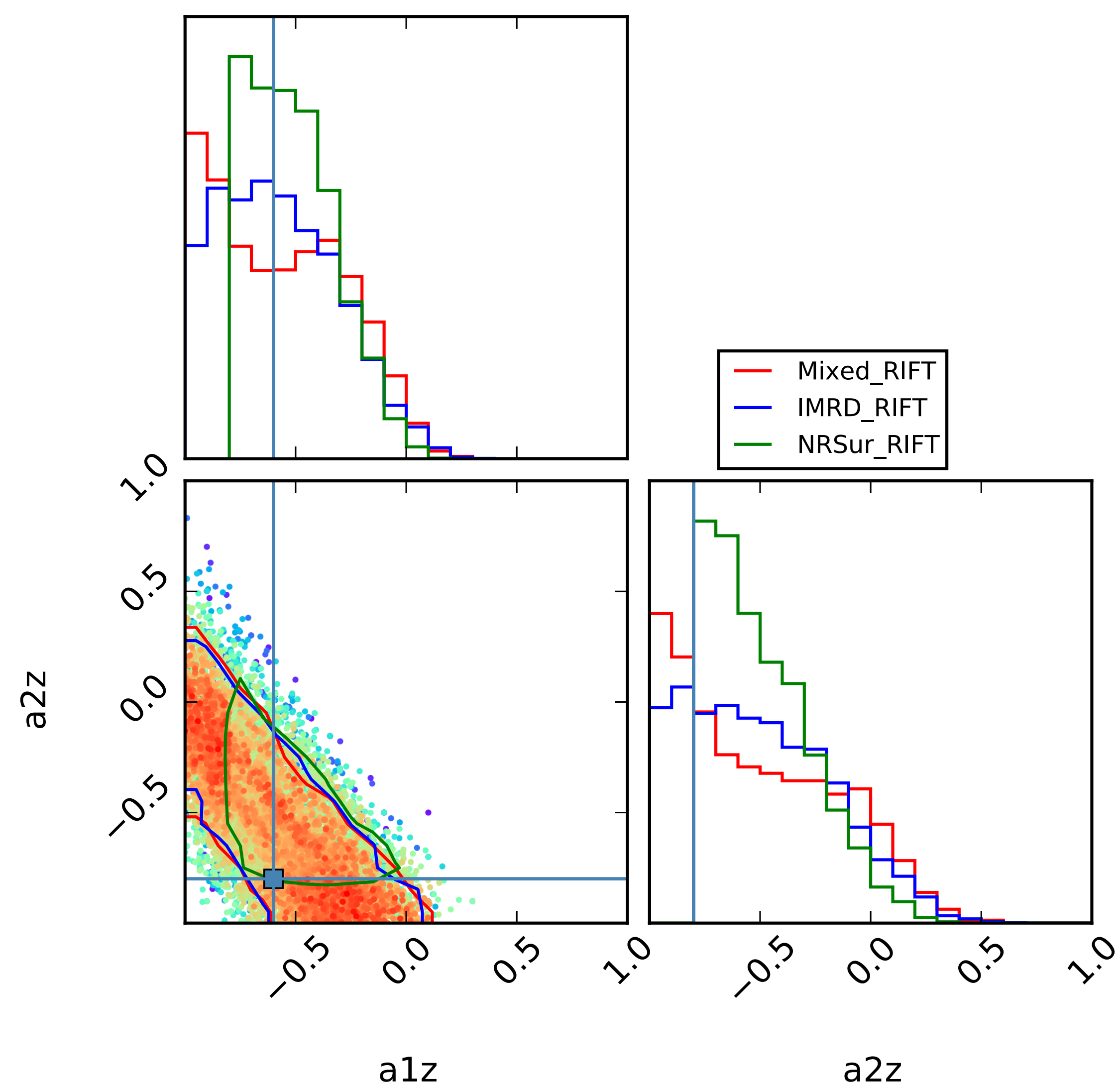
$$\langle a|b\rangle \equiv 2\int_{-\infty}^{\infty} df \frac{a^*(f)b(f)}{S_h(|f|)}$$

$$\mathcal{L} \equiv p(\{d\}|\vec{\lambda},H_1)/p(\{d\}|H_0) = \frac{e^{-\langle h(\lambda)-d|h(\lambda)-d\rangle/2}}{e^{-\langle d|d\rangle/2}}$$

Why use RIFT? Mixing models

- Some good models (NR-calibrated surrogates) have limited validity: mass ratio
  - Fill in the gaps with other models (PN; NR; …)



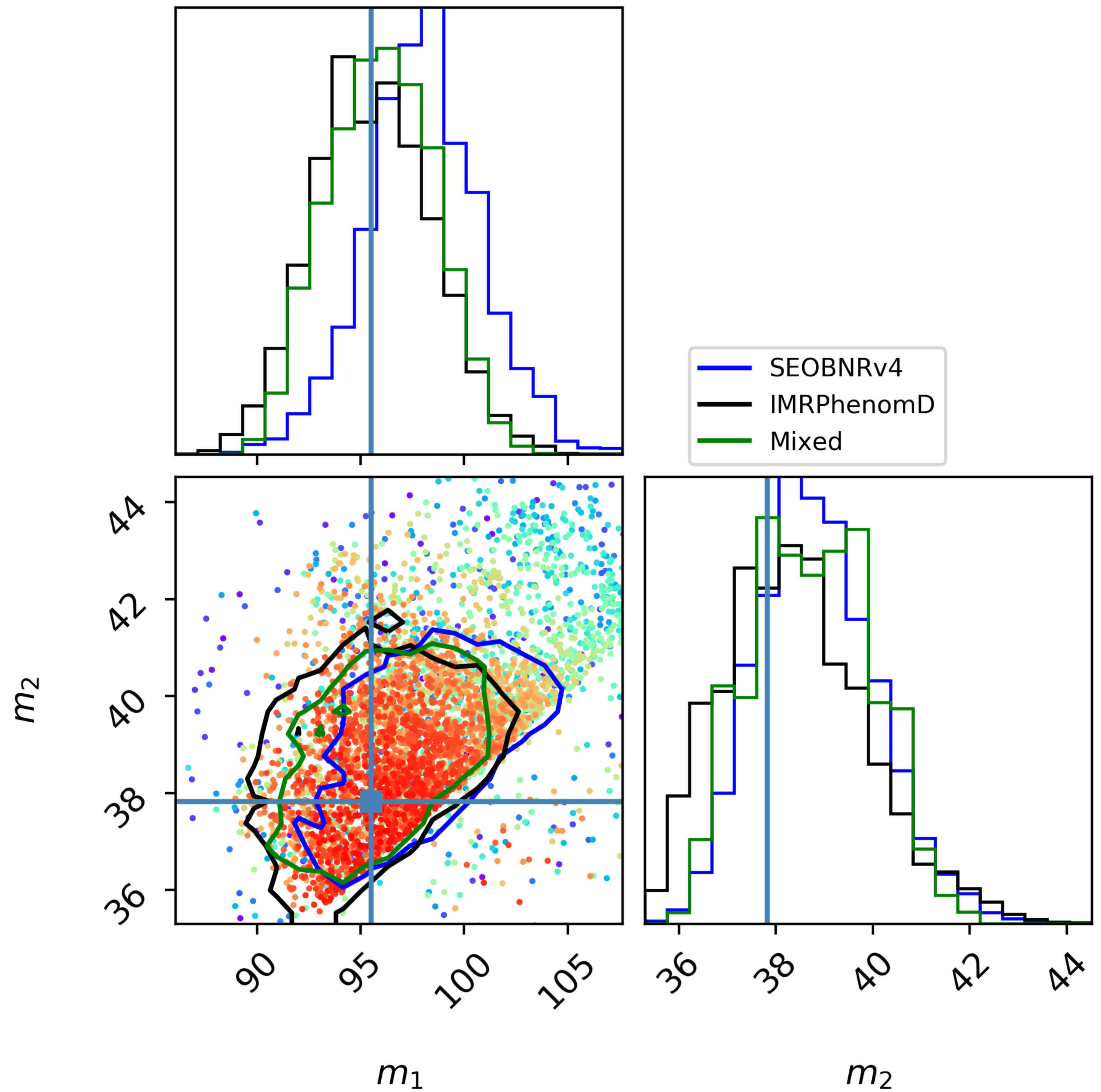


Lange et al arxiv:1805.06442

Why use RIFT? Model marginalization

- Natural technique to average over model differences

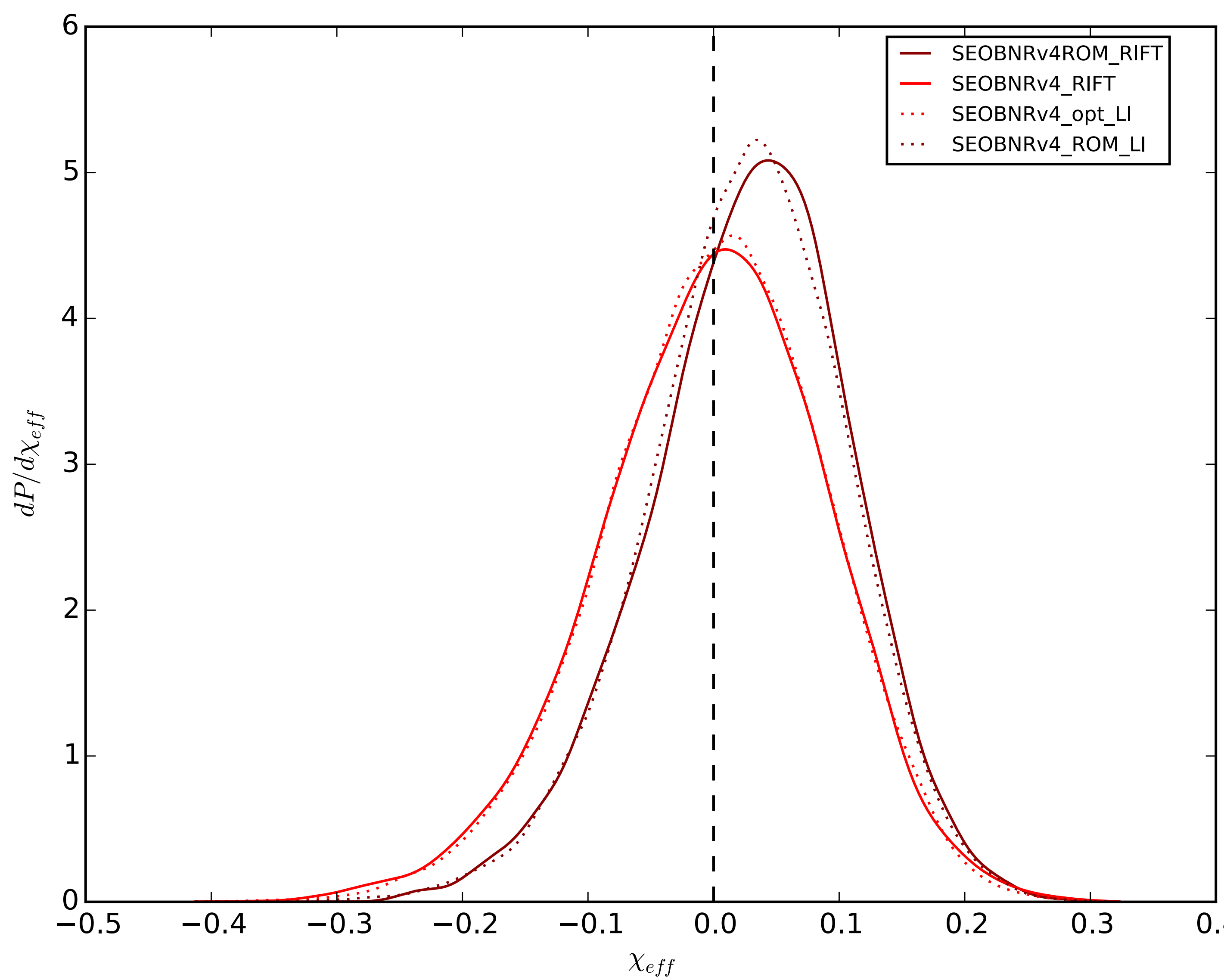
$$\mathcal{L}_1 p_1 + \mathcal{L}_2 p_2 + \ldots$$



Jan et al arXiv:2011.03571

Why use RIFT? Useful investigations easier

- Example of waveform systematics: GW150914-like synthetic source
  - One model designed as a fast approximation to another (match >0.997)



q=1,SNR~20

$$\chi_{1,z} = -\chi_{2,z} = 0.25$$
$$M_z = 80 M_\odot$$

Lange, LIGO-G1801107

- **Other examples:** 170729 paper; ROQ bias (mchirp for 170729);…