Transformers + Normalizing Flows for parameter estimation of overlapping gravitational waves in next generation detectors

EuCAIFCon, Cagliari – 18 June 2025

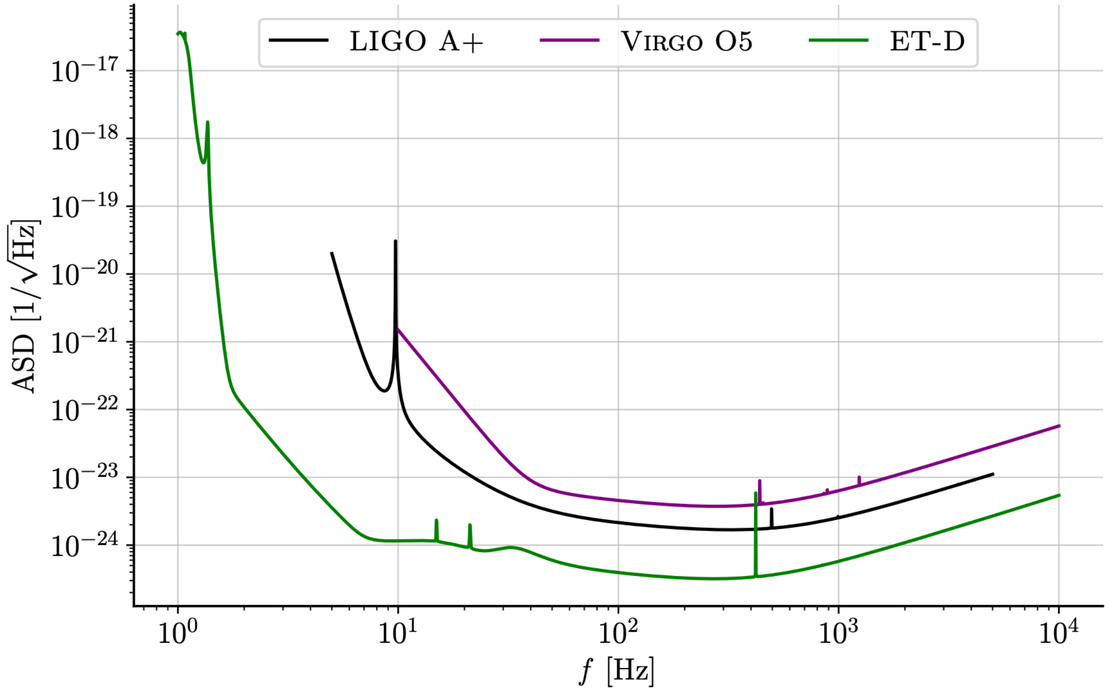
Papalini et al., arXiv:2505.02773, *submitted to CQG*

Lucia Papalini[1,2], Federico De Santi[3,4], Massimiliano Razzano[1,2], Ik Siong Heng[5], Elena Cuoco[6]

[1]University of Pisa, [2] INFN Pisa, [3] University of Milano Bicocca, [4] INFN Milano, [5] University of Glasgow, [6] University of Bologna

The challenge of 3rd Generation Detectors

How do we expect to observe compact binary coalescences (CBC) in the Einstein Telescope?



Papalini et al., arXiv:2505.02773, *submitted to CQG*

Curves publicly available at
https://dcc.ligo.org/LIGO-T2000012/public https://apps.et-gw.eu/tds/?r=14065

- Longer signals
- Higher detection rate

$R_{\text{BBH}} \sim 10^5 - 10^6/\text{yr}$

arXiv:1912.02622

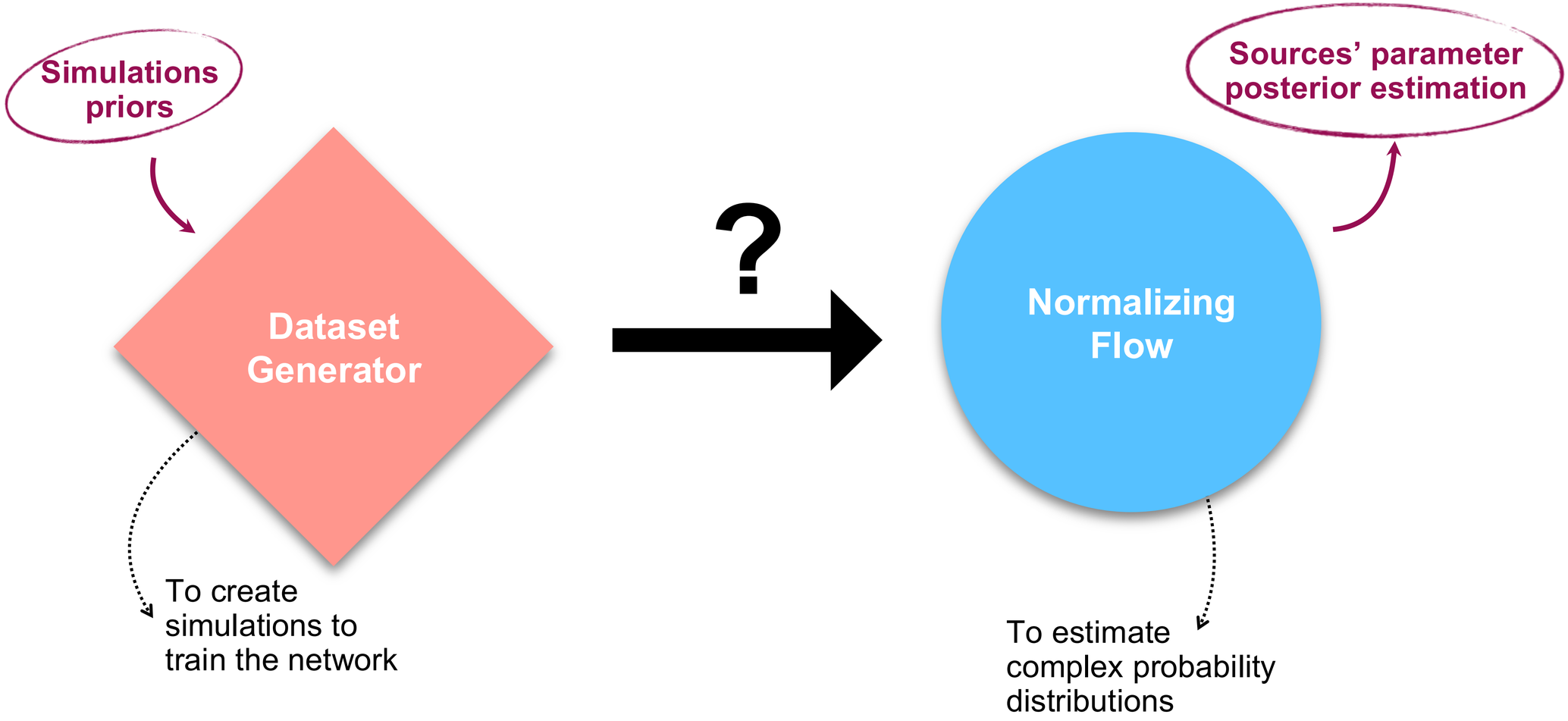
Overlapping signals

- Computationally challenging
- Biases in parameter estimation (PE)

arXiv:2102.07692

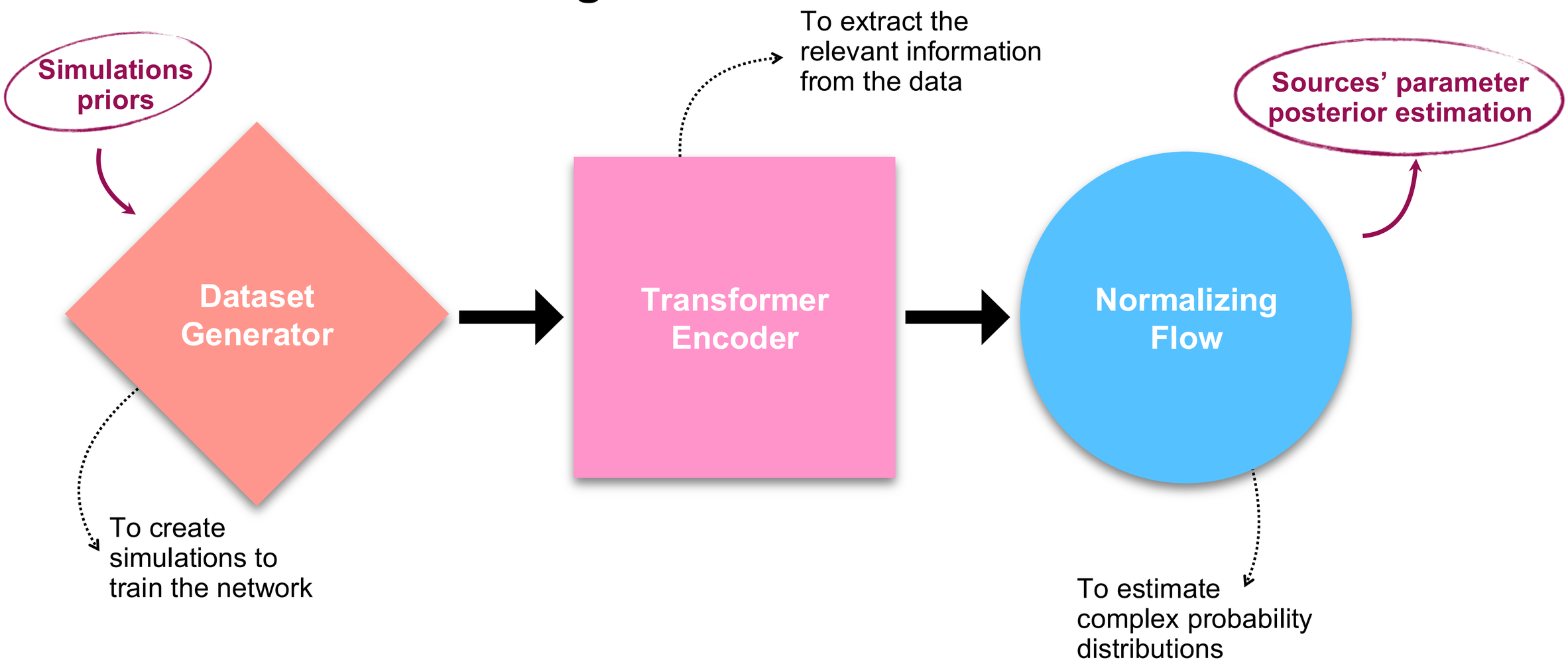
Ingredients for a PE pipeline

Based on machine learning



Ingredients for a PE pipeline

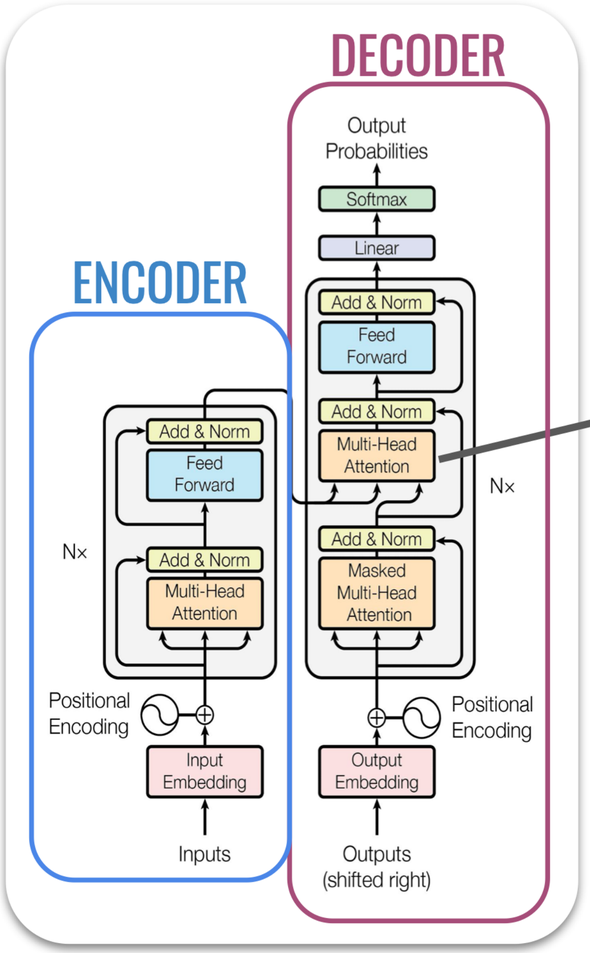
Based on machine learning

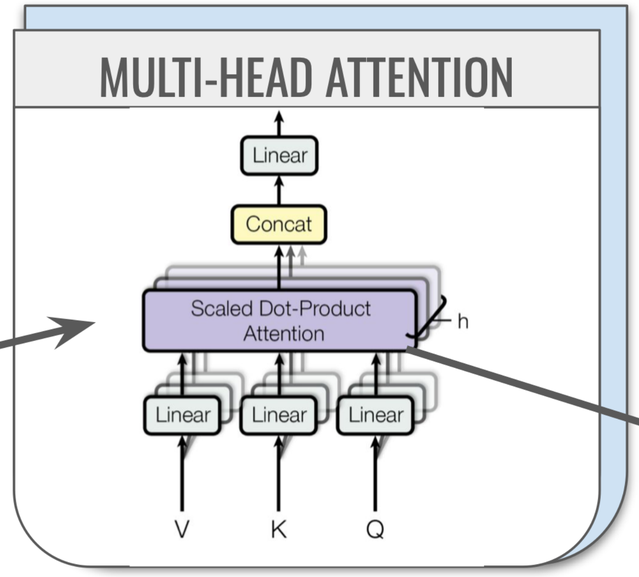


Transformers

Very briefly

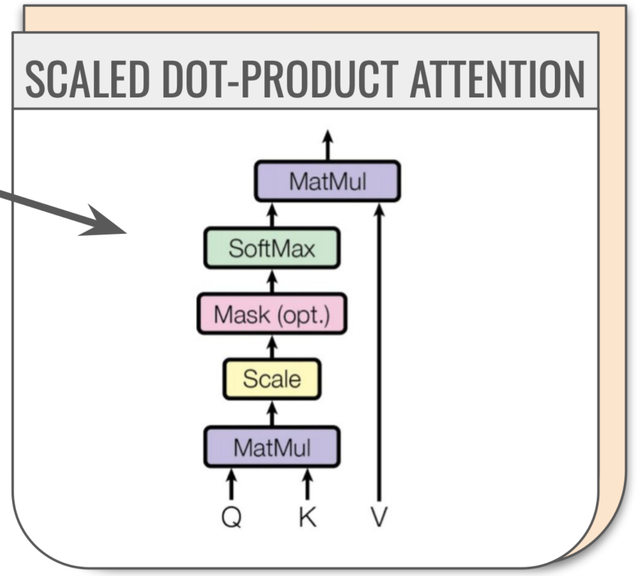
Images taken from: Attention is all you need (2017)





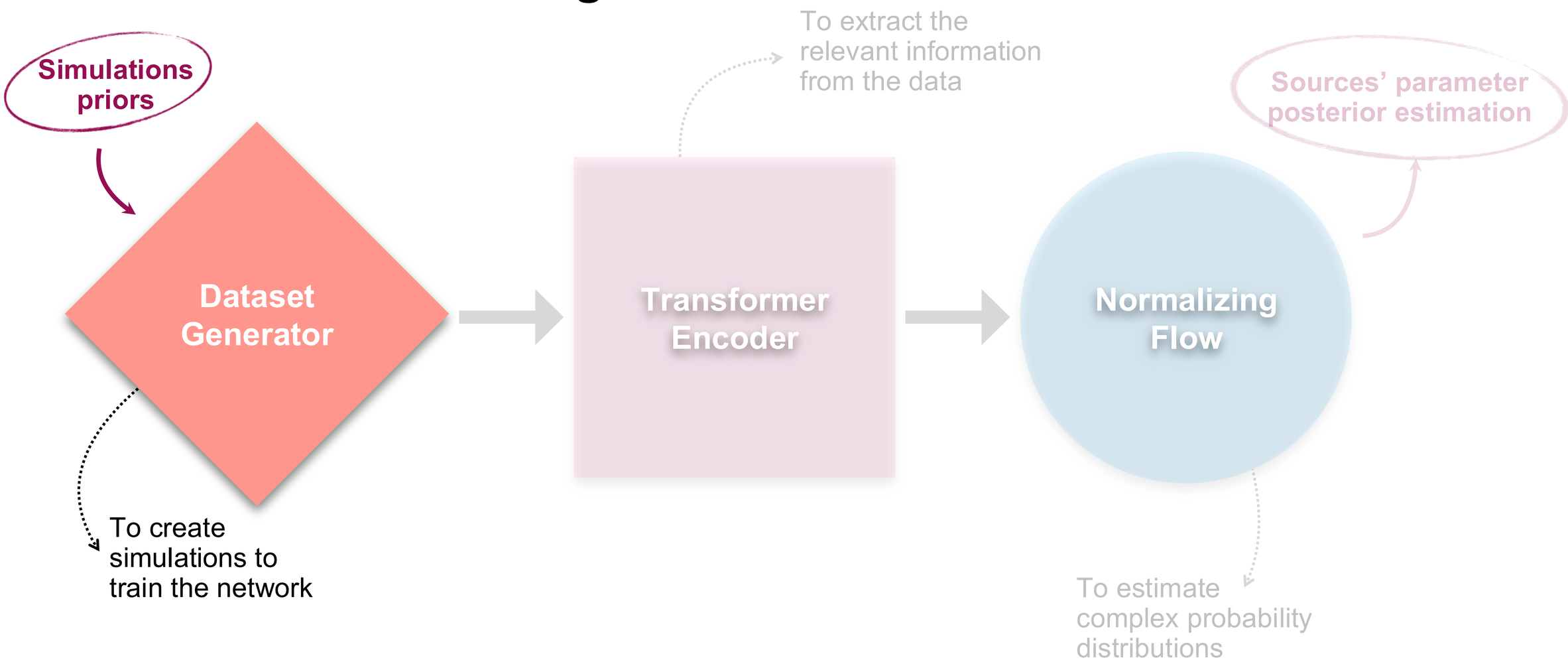
$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Ingredients for a PE pipeline

Based on machine learning



Dataset

Simulations priors

Mainly motiveted by hardware reasons

Our samples:
- ✦ Duration *16 s*
- ✦ Sampling rate *1024 Hz*
- ✦ *IMRPhenomXPHM*
- ✦ *ET triangular shape in Sardinia site*

- High-mass range (to assure shorter durations)
- Masses are sampled in the detector frame
- We implicitly assume z_max=10

- We associate each signal with a *random time shift* in the signal window

parameter	prior
Intrinsic	
$m_1\ [M_\odot]$	$\mathcal{U}(100, 800)$
$m_2 \leq m_1\ [M_\odot]$	$\mathcal{U}(100, 650)$
inclination ι	$\sin(0, \pi)$
coalescence phase ϕ	$\mathcal{U}\,(0, 2\pi)$
Extrinsic	
SNR $\rho_{\rm opt}$	$\mathcal{U}(10, 150)$
polarization ψ	$\mathcal{U}\,(0, 2\pi)$
right ascension α	$\mathcal{U}\,(0, 2\pi)$
declination δ	$\cos\,(-\pi/2, \pi/2)$
$t_{\rm merger}$	$\mathcal{U}(-\frac{2}{3}T_{\rm obs}, 0)$
$t_{\rm GPS}$	fixed 1399075278

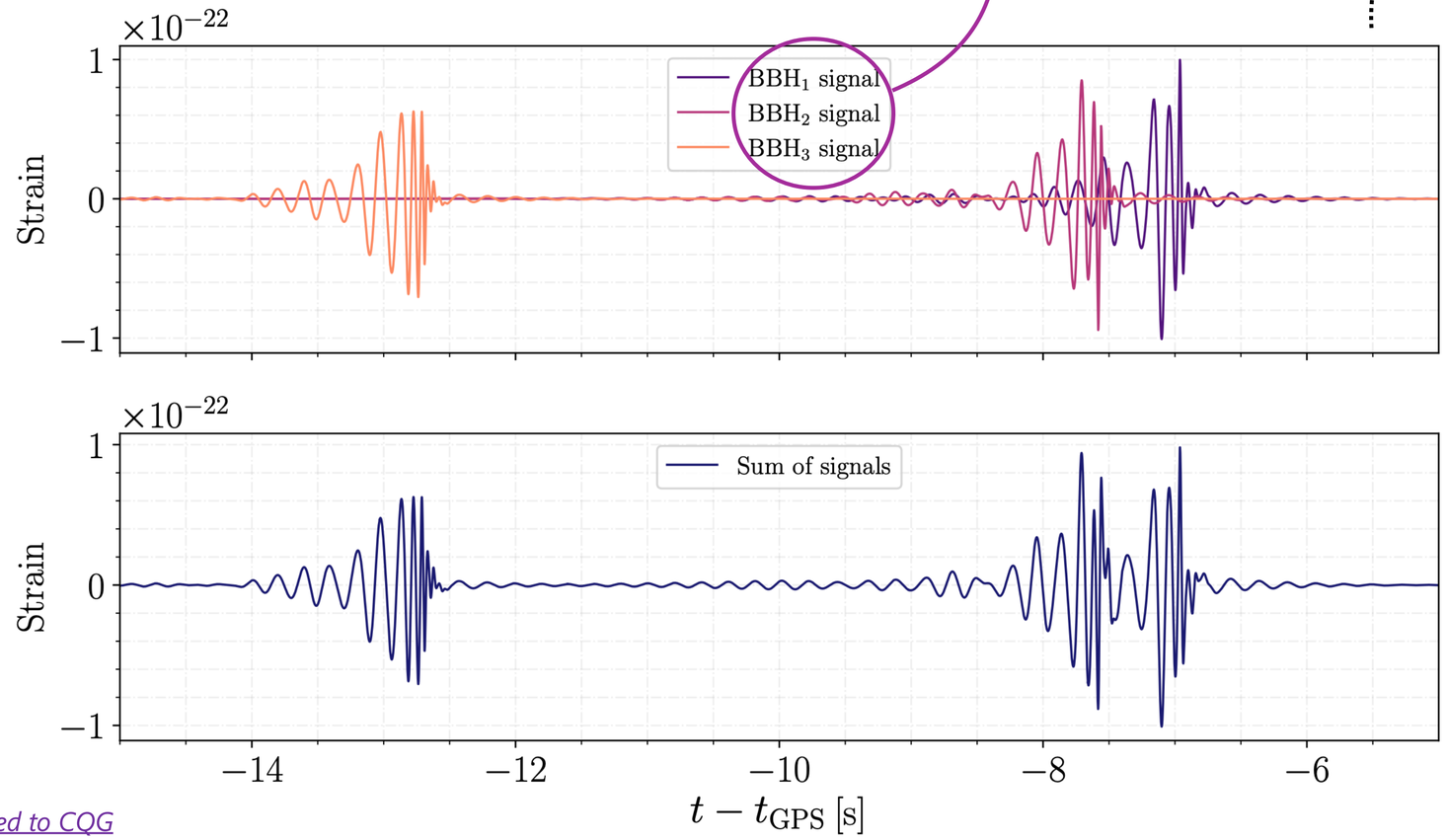
Dataset

3 overlapping BBHs

The sorting is not random!
(Remind this for later)

1 out of 3
ET channels

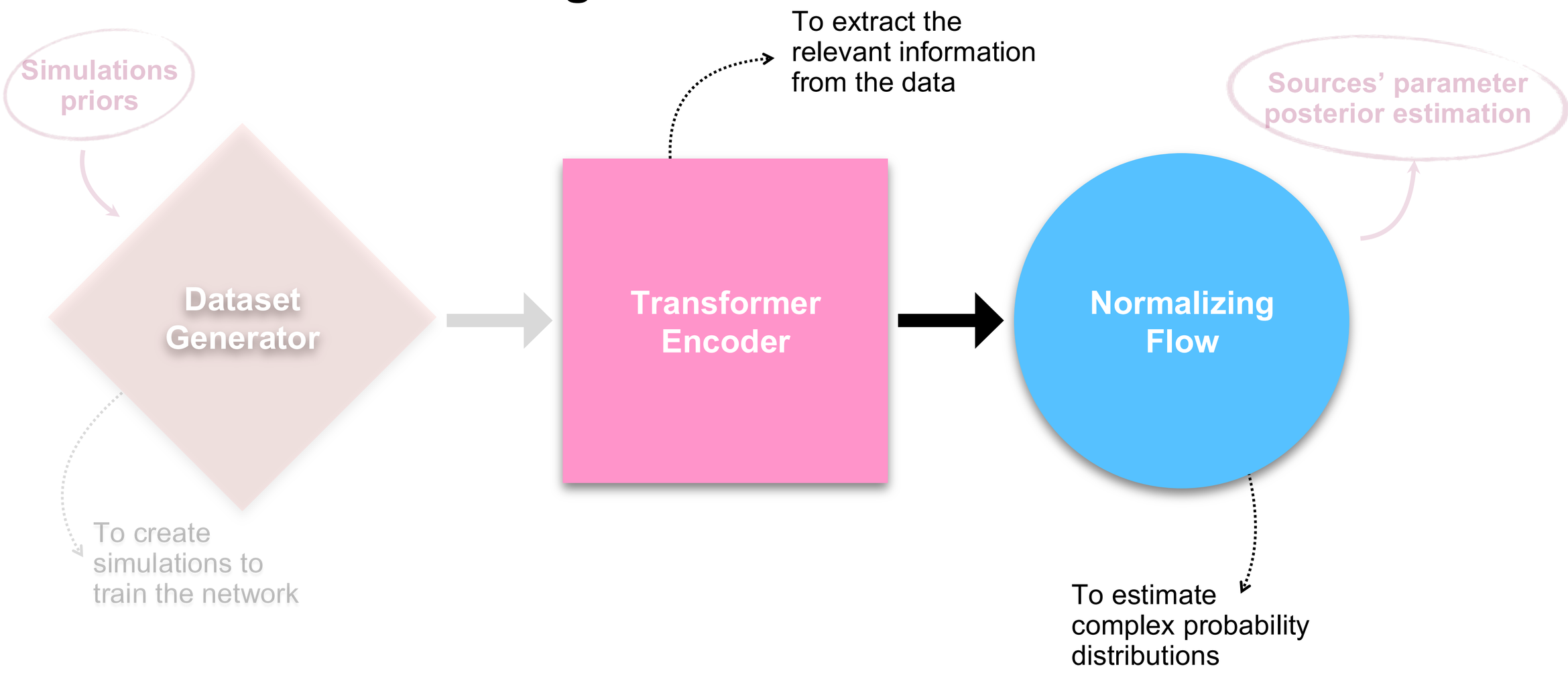
SNR = 108, 87, 70



Papalini et al., arXiv:2505.02773, *submitted to CQG*

Ingredients for a PE pipeline

Based on machine learning

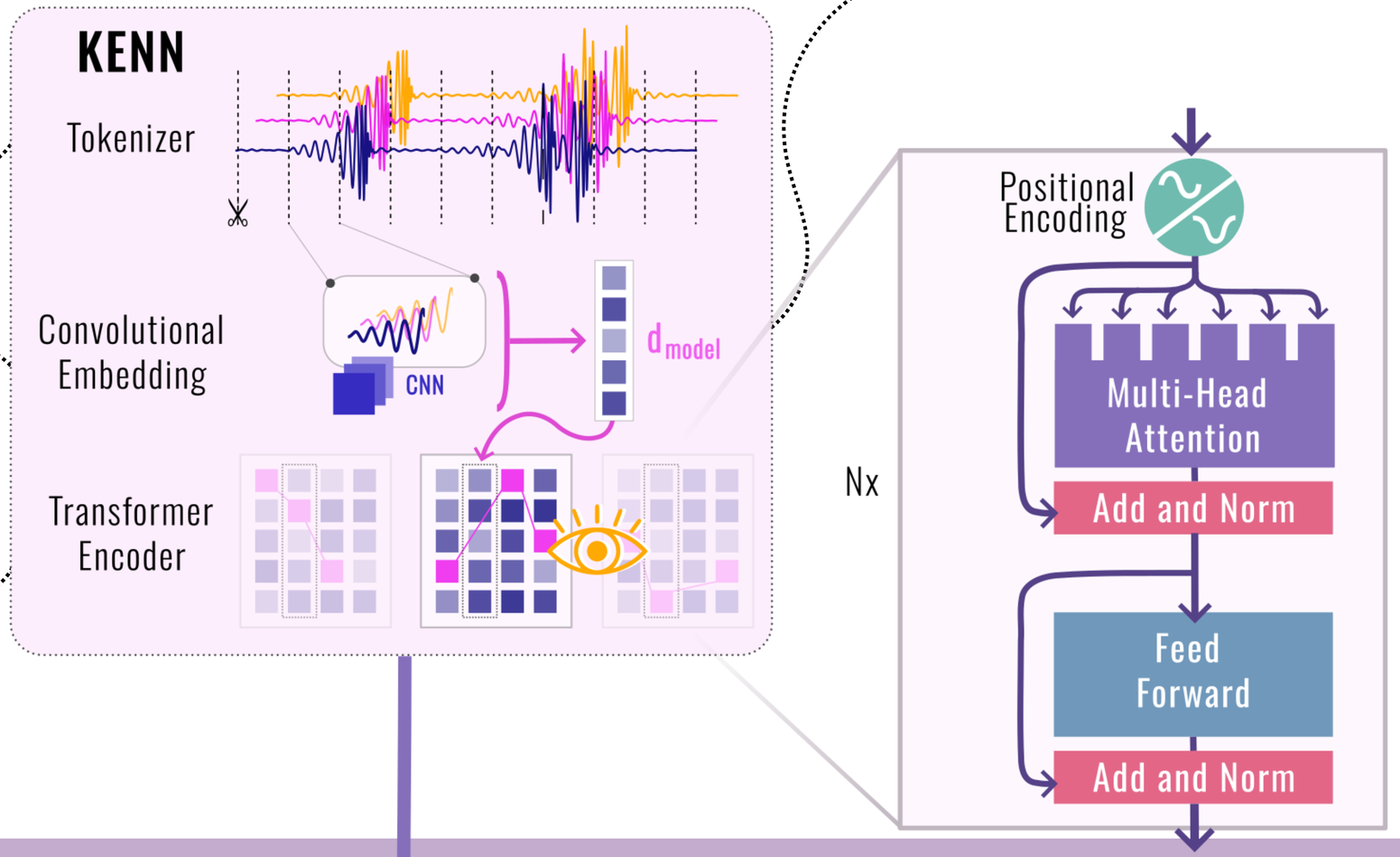


KENN
Knowledge Extractor Neural Network

We split the time series in chunks

We map every chunk in the embedding space, in which the Transformer works

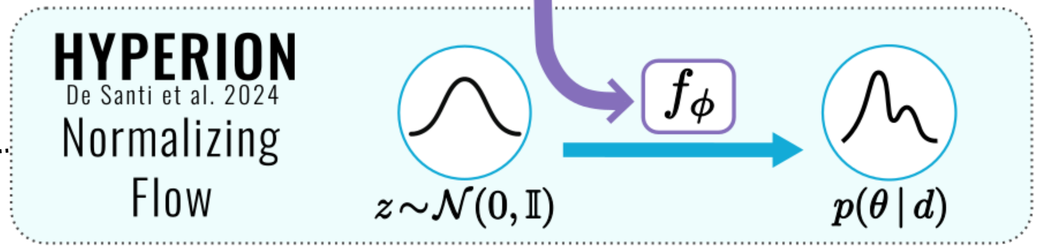
The Transformer Encoder uses the Multi-Head Attention mechanism to create a more informative representation of the signal



Papalini et al., arXiv:2505.02773, *submitted to CQG*

HYPERION

Our normalizing flow architecture



First developed for Close Encounters

(**De Santi et. al 2024 Phys. Rev. D 109**)

Flow Architecture:

- Coupling Layers
- Affine Transformations

The training proceeds with the minimisation of the Kullback-Leibller cost function:

$$\mathrm{KL}[p||q_\phi] \simeq -\frac{1}{N}\sum_{i=1}^{N} \log q_\phi(\theta_{\mathrm{gw}}|d)$$

Training of the network

Parameters estimated:

BBH 1 : $[M_1, \mathcal{M}_1, q_1, t_{\mathrm{merger},\,1}]$

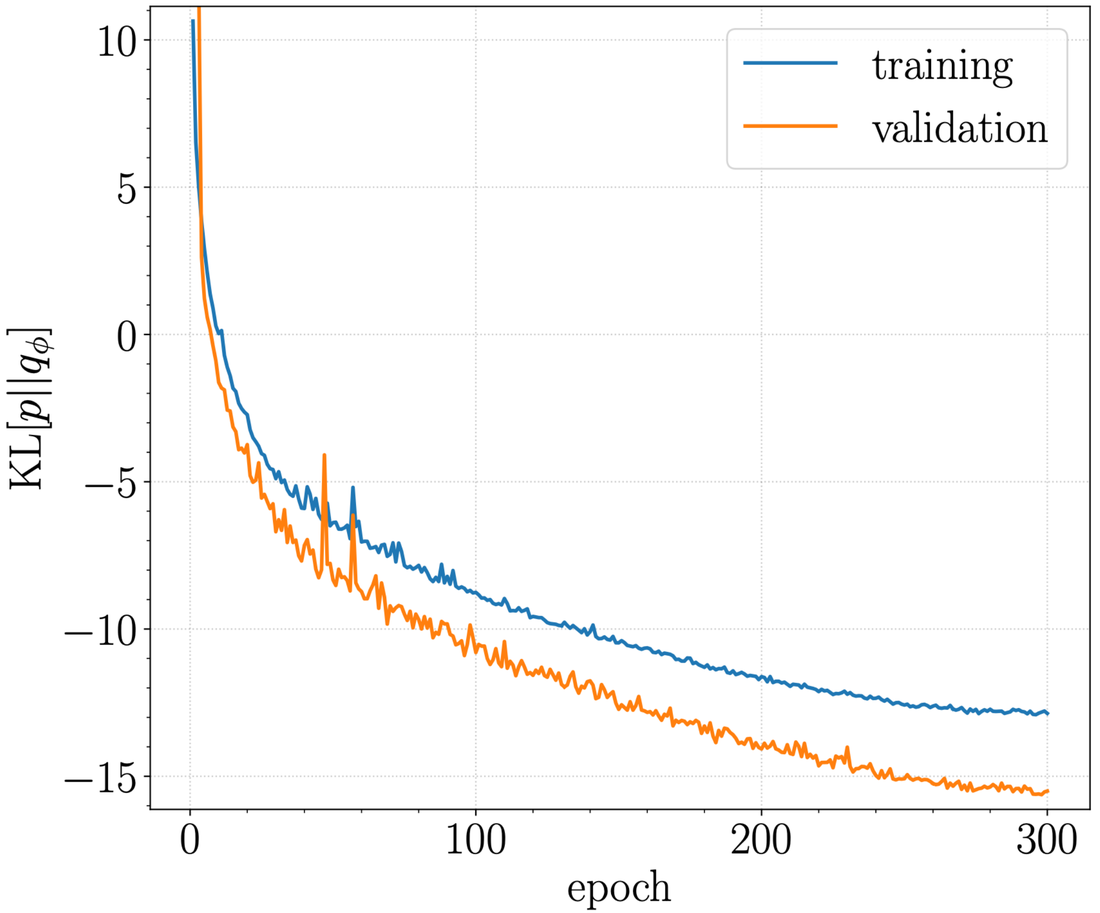
BBH 2 : $[M_2, \mathcal{M}_2, q_2, t_{\mathrm{merger},\,2}]$

BBH 3 : $[M_3, \mathcal{M}_3, q_3, t_{\mathrm{merger},\,3}]$

How does the network know who is 1, 2, 3?

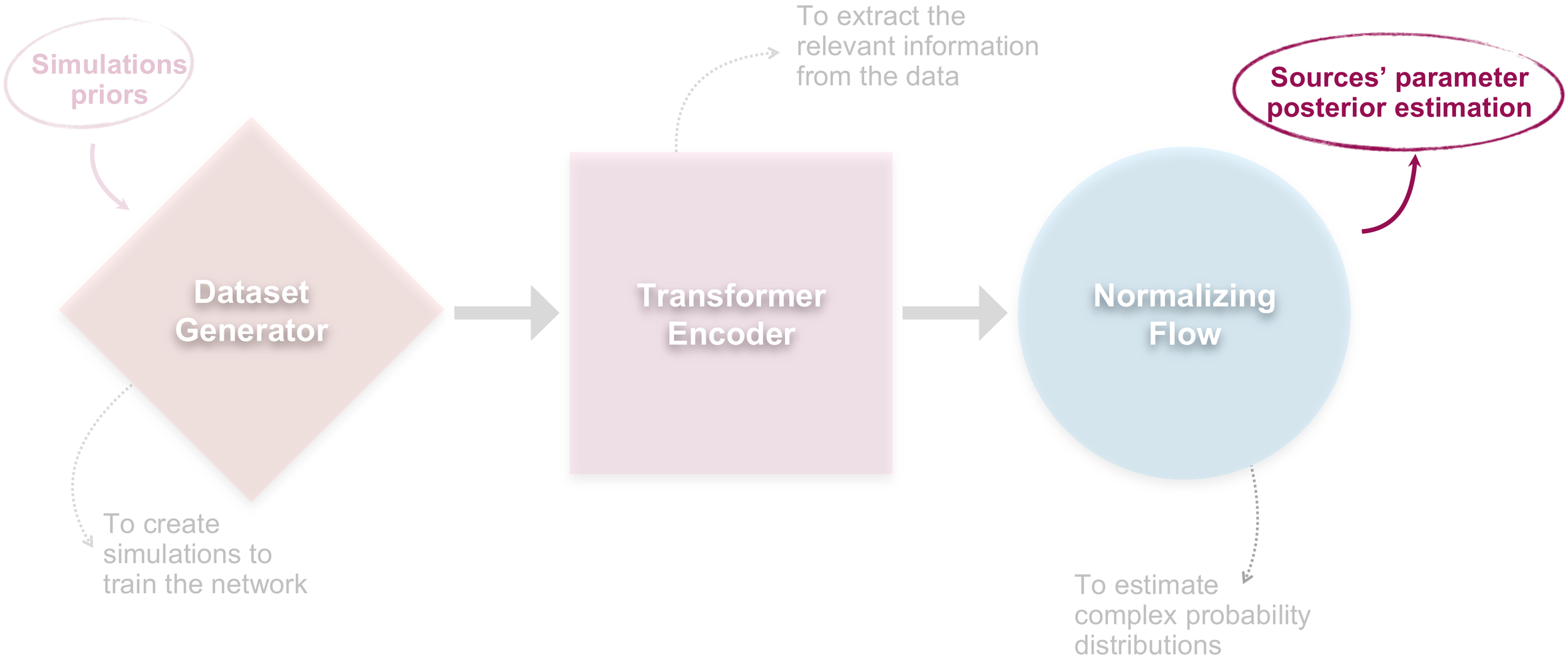
We sort the signals by SNR: the network needs a univoque way of sorting the signals

- Training dataset “size”: *3.87e7*
- Trainable parameters: *18e6*
- Training duration: *3 days on NVIDIA A30 GPU*



Ingredients for a PE pipeline

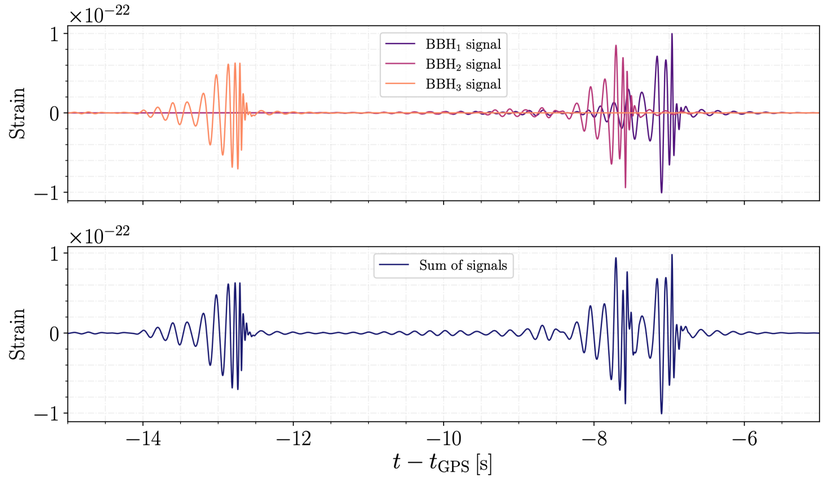
Based on machine learning

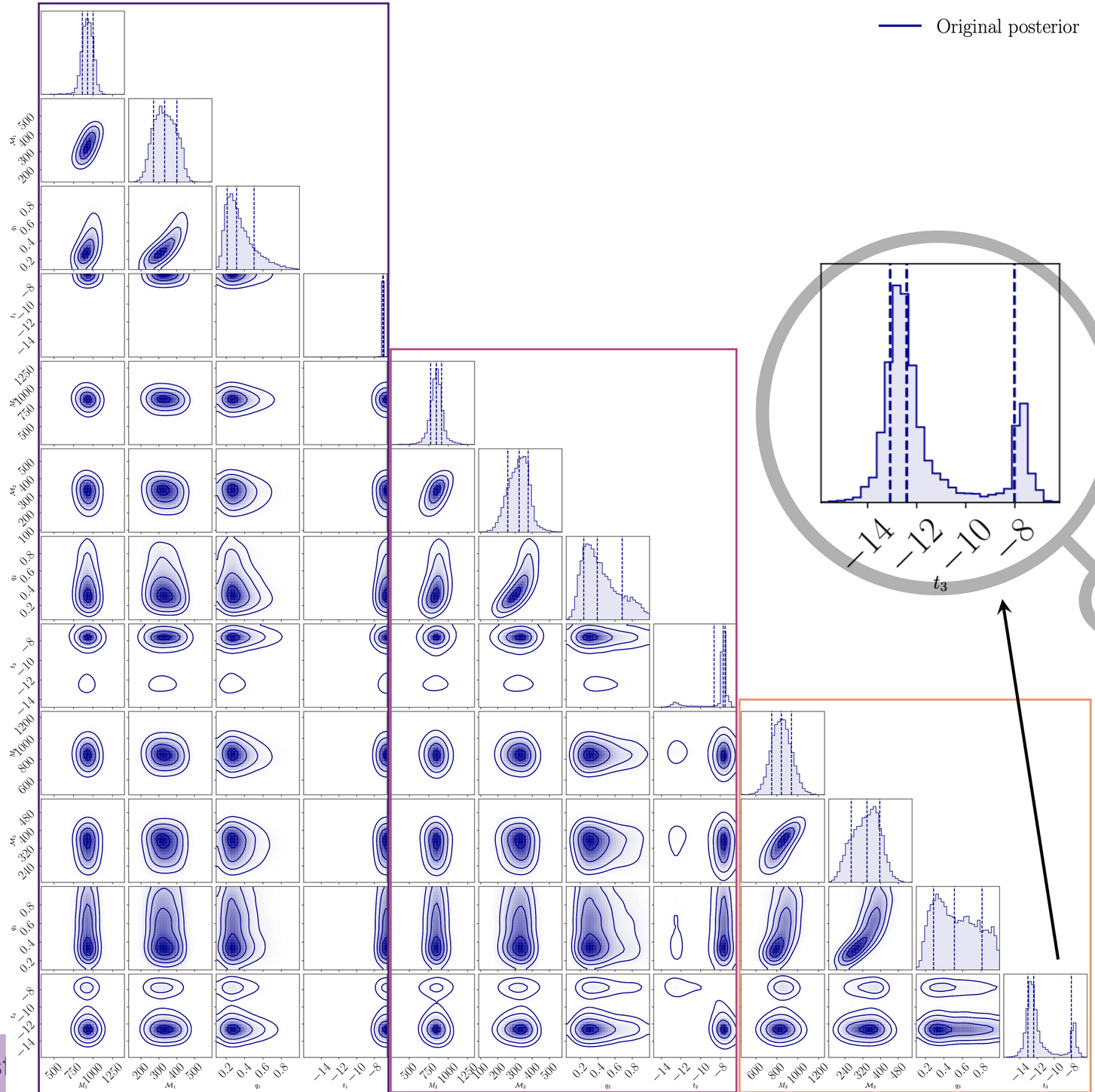


Results

A journey in degeneracies

SNR = 108, 87, 70

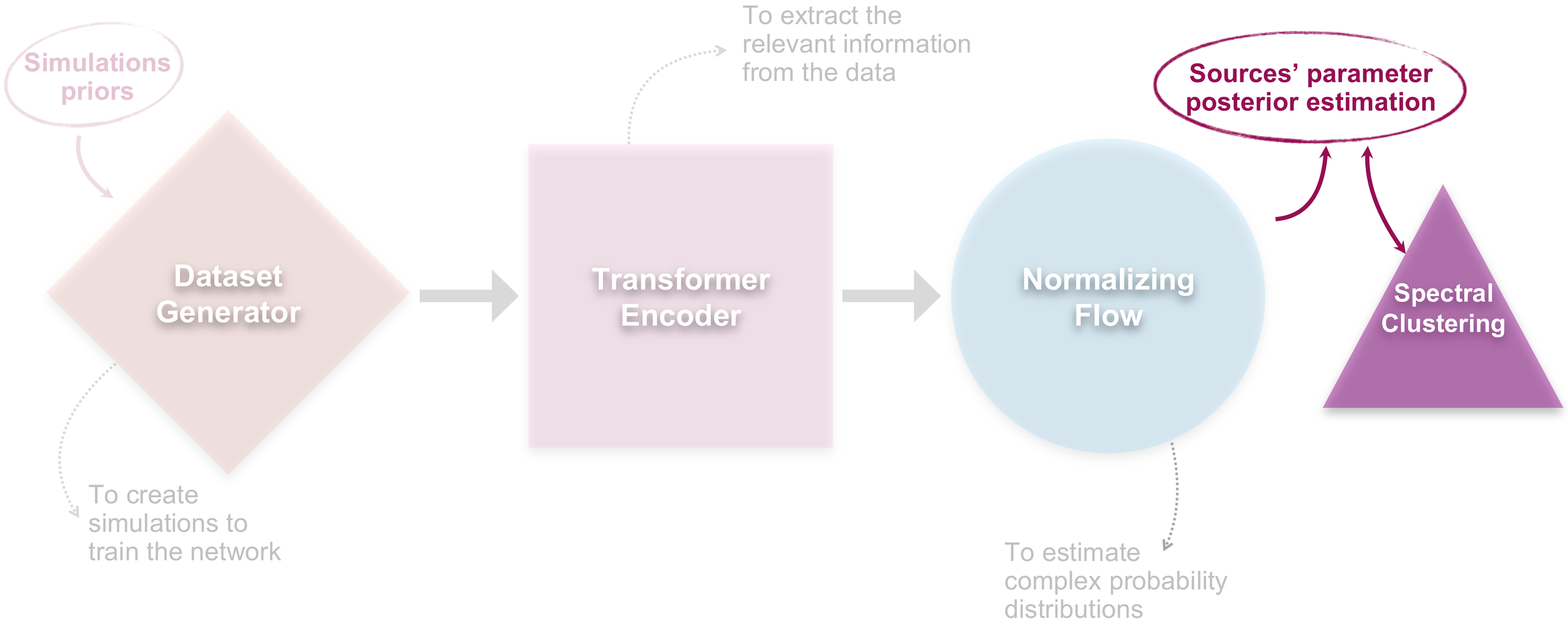




[Papalini et al., arXiv:2505.02773, *submitted to CQG*]

Ingredients for a PE pipeline

Based on machine learning

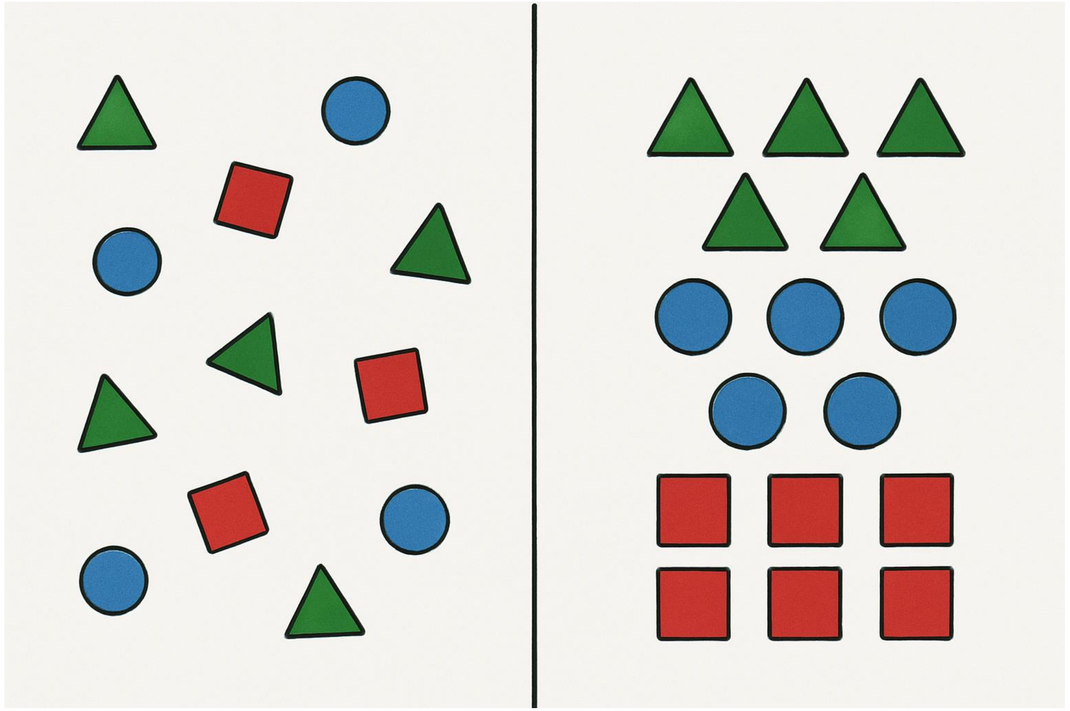


Spectral Clustering

Gerosa D et al. 2025 “Which Is Which?” Phys. Rev. Lett. 134 121402

What’s spectral clustering?

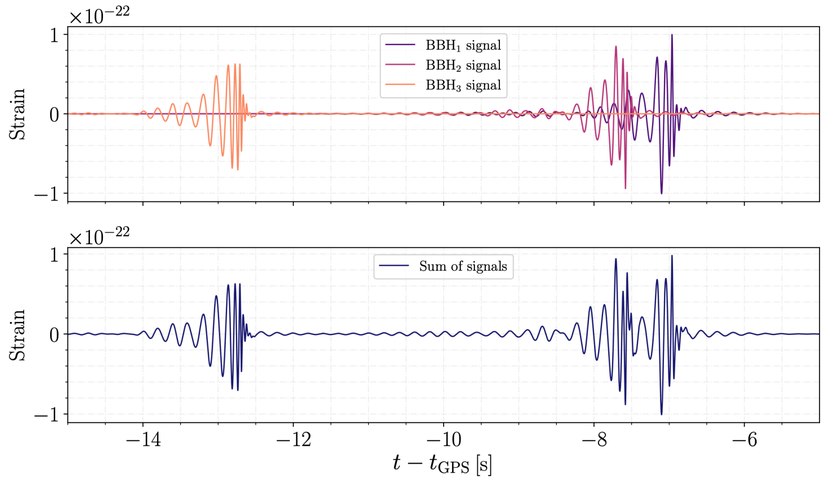
- **Purpose:** Refine posterior samples using Spectral Clustering
- **Method:** Cluster based on eigenvectors of an affinity matrix built via k-NN
- **Adjustment:** Map new labels back to the original using the Hungarian Algorithm

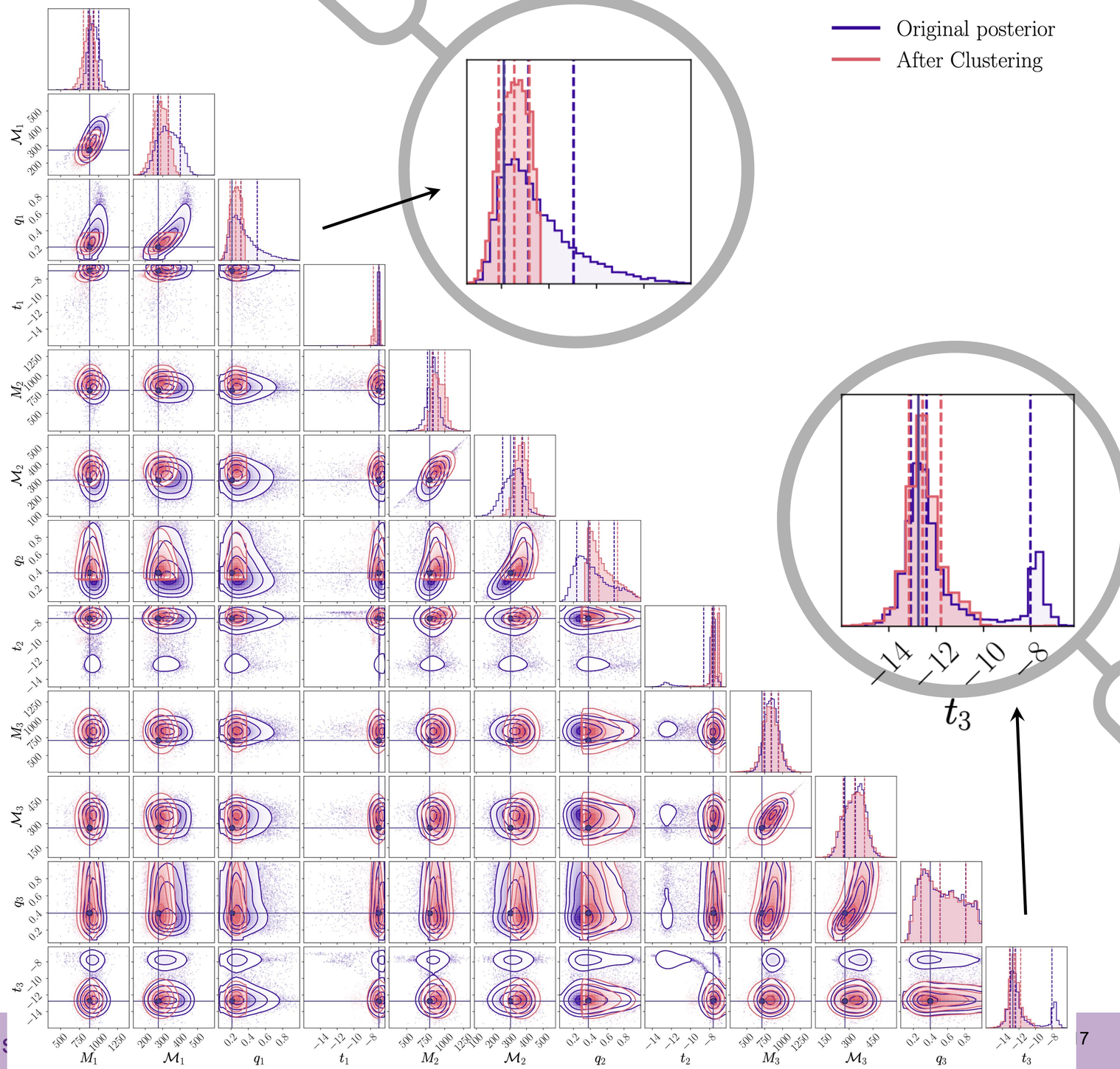


Postprocessed results

Degeneracies are successfully removed and distributions are narrower

SNR = 108, 87, 70



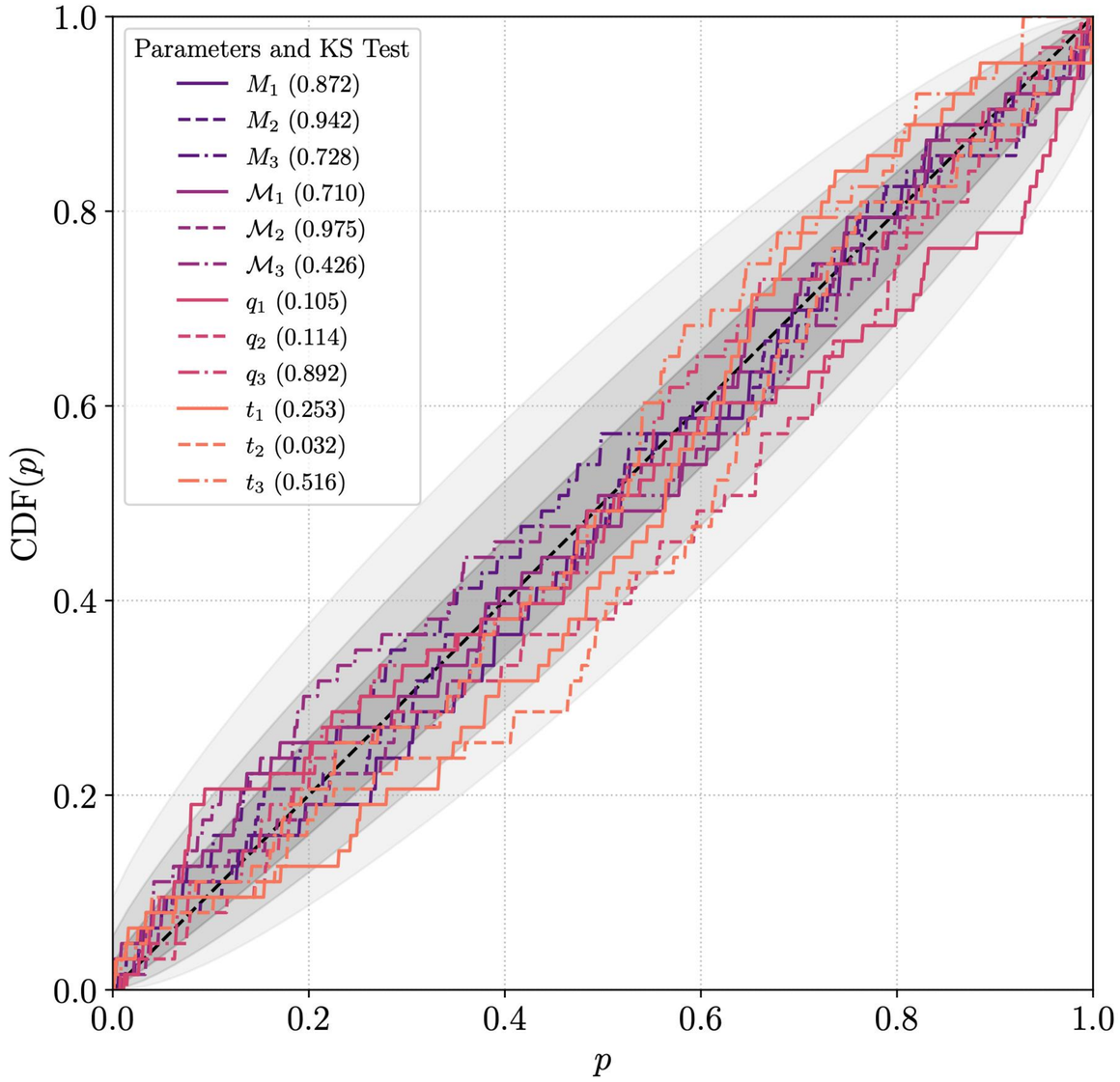


Statistical analysis

Of network performances

Probability-probability plot

- N samples = 64
- Good diagonal behaviour



Papalini et al., arXiv:2505.02773, *submitted to CQG*

Statistical analysis

Of network performances

Relative error vs correlation

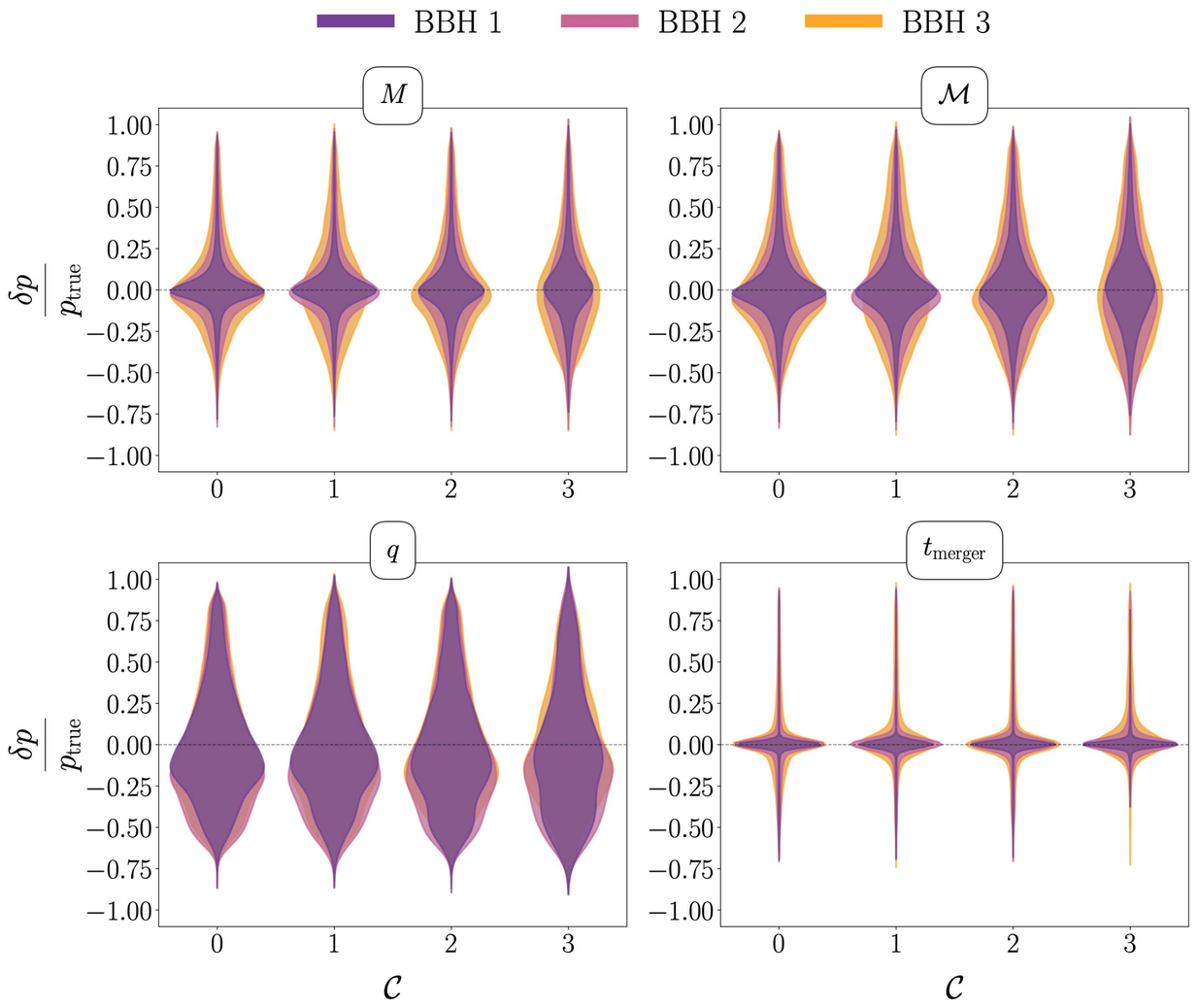
Correlation is calculated with this discrete statistics

$$\mathcal{C}_{k,\,\mathrm{det}} = \mathcal{P}\left[h_k^{\mathrm{det}} \,\middle|\, \sum_{j\neq k} h_j^{\mathrm{det}}\right]$$

$$\mathcal{C}_k = \frac{1}{\sqrt{N_{\mathrm{det}}}}\left(\sum_{\mathrm{detectors}} \mathcal{C}_{k,\,\mathrm{det}}^2\right)^{1/2}$$

$$\mathcal{C} = \sum_k \Theta(C_k - 0.05)$$

Number of BBH signals whose correlation with at least one other is > 5%



Papalini et al., arXiv:2505.02773, *submitted to CQG*

Thank you for your

$$\text{Attention}(Q, K, V) = Softmax\left(\frac{Q \cdot K^\top}{\sqrt{d_k}}\right)V$$

Backup slides

Data

A custom generator immune to overfitting

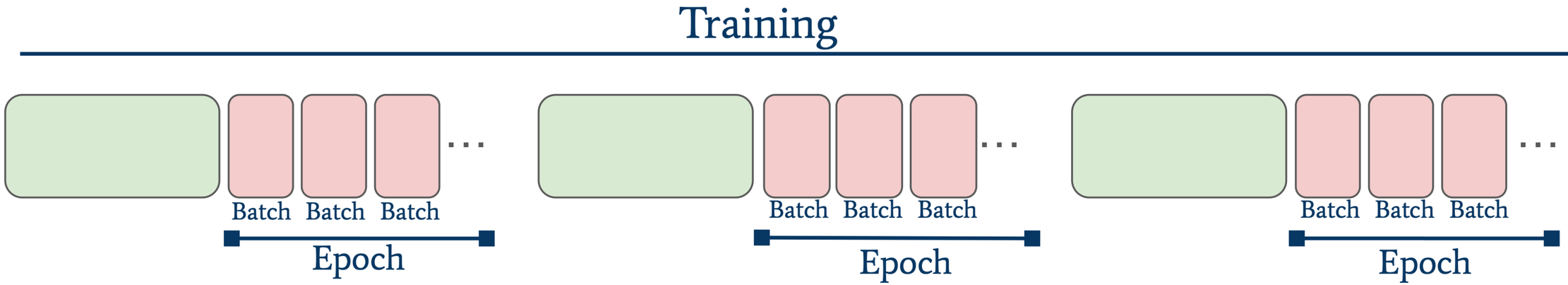
We generate new signals samples every epoch while the training is ongoing, so the network never sees the same sample twice.

Simulation of source-frame waveforms

- Sampling of *intrinsic* parameters

Extraction of random simulations + projection

- Sampling of *extrinsic* parameters
- Noise addition + whitening
- Add signals, sort them by SNR (labels)



Transformers

Very briefly

Attention is all you need (**2017**, A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaisr, I. Polosukhin)

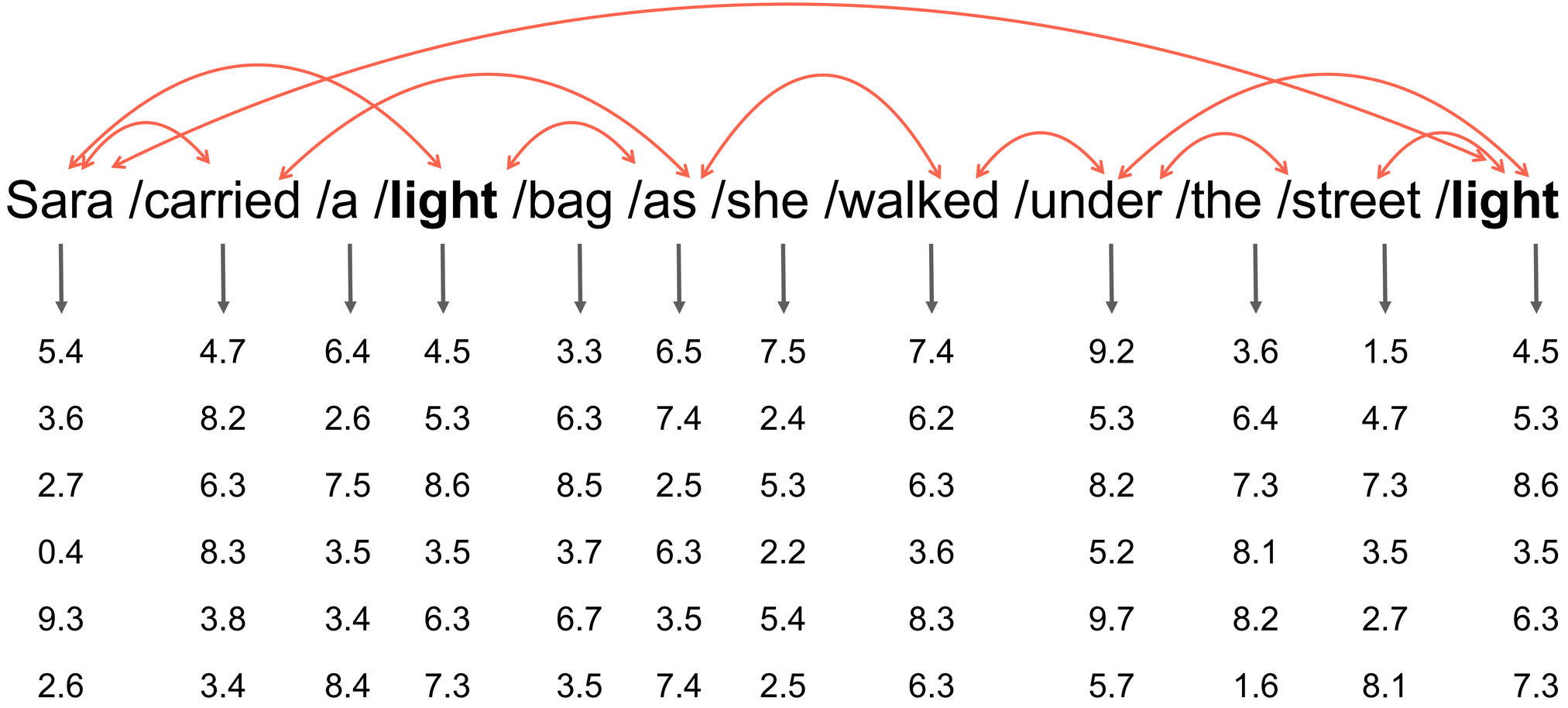
Sara /carried /a /**light** /bag /as /she /walked /under /the /street /**light**

5.4	4.7	6.4	4.5	3.3	6.5	7.5	7.4	9.2	3.6	1.5	4.5
3.6	8.2	2.6	5.3	6.3	7.4	2.4	6.2	5.3	6.4	4.7	5.3
2.7	6.3	7.5	8.6	8.5	2.5	5.3	6.3	8.2	7.3	7.3	8.6
0.4	8.3	3.5	3.5	3.7	6.3	2.2	3.6	5.2	8.1	3.5	3.5
9.3	3.8	3.4	6.3	6.7	3.5	5.4	8.3	9.7	8.2	2.7	6.3
2.6	3.4	8.4	7.3	3.5	7.4	2.5	6.3	5.7	1.6	8.1	7.3

Transformers

Very briefly

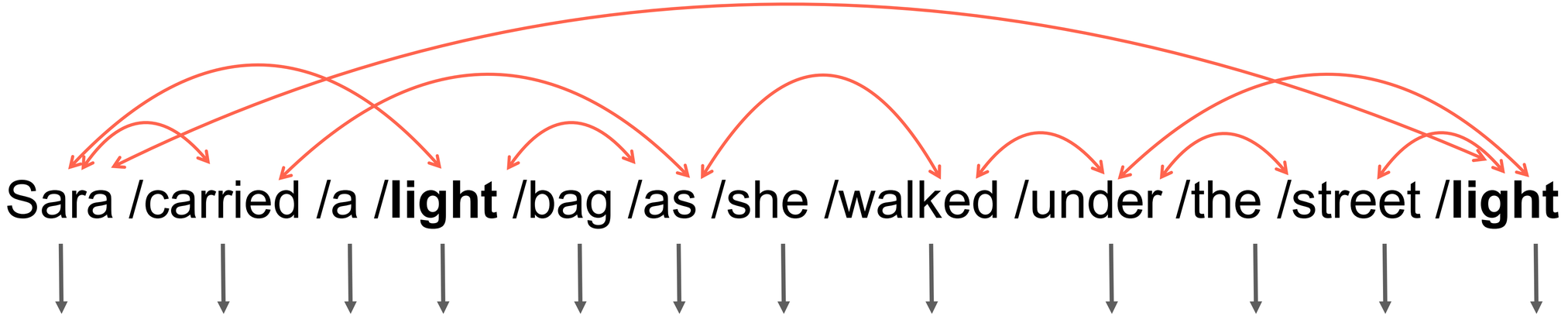
Attention is all you need (**2017**, A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaisr, I. Polosukhin)



Transformers

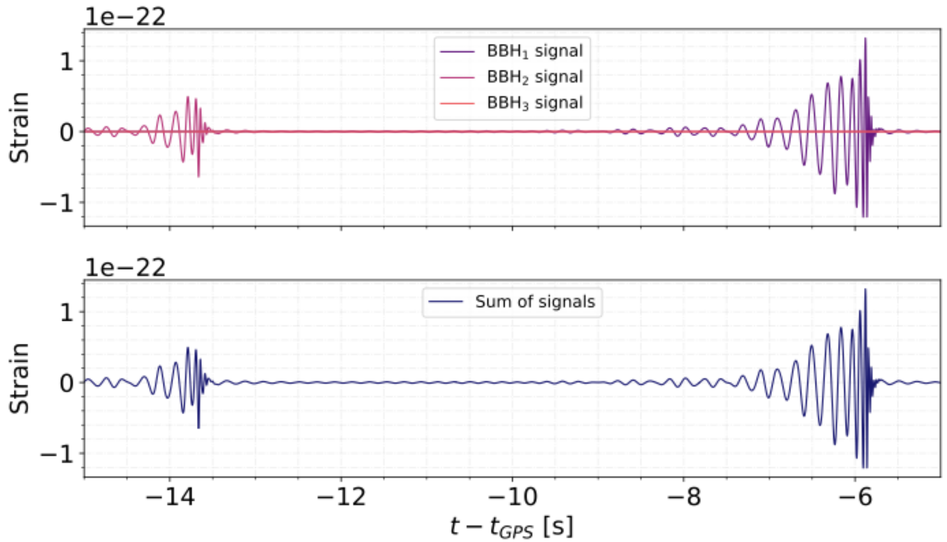
Very briefly

Attention is all you need (**2017**, A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaisr, I. Polosukhin)

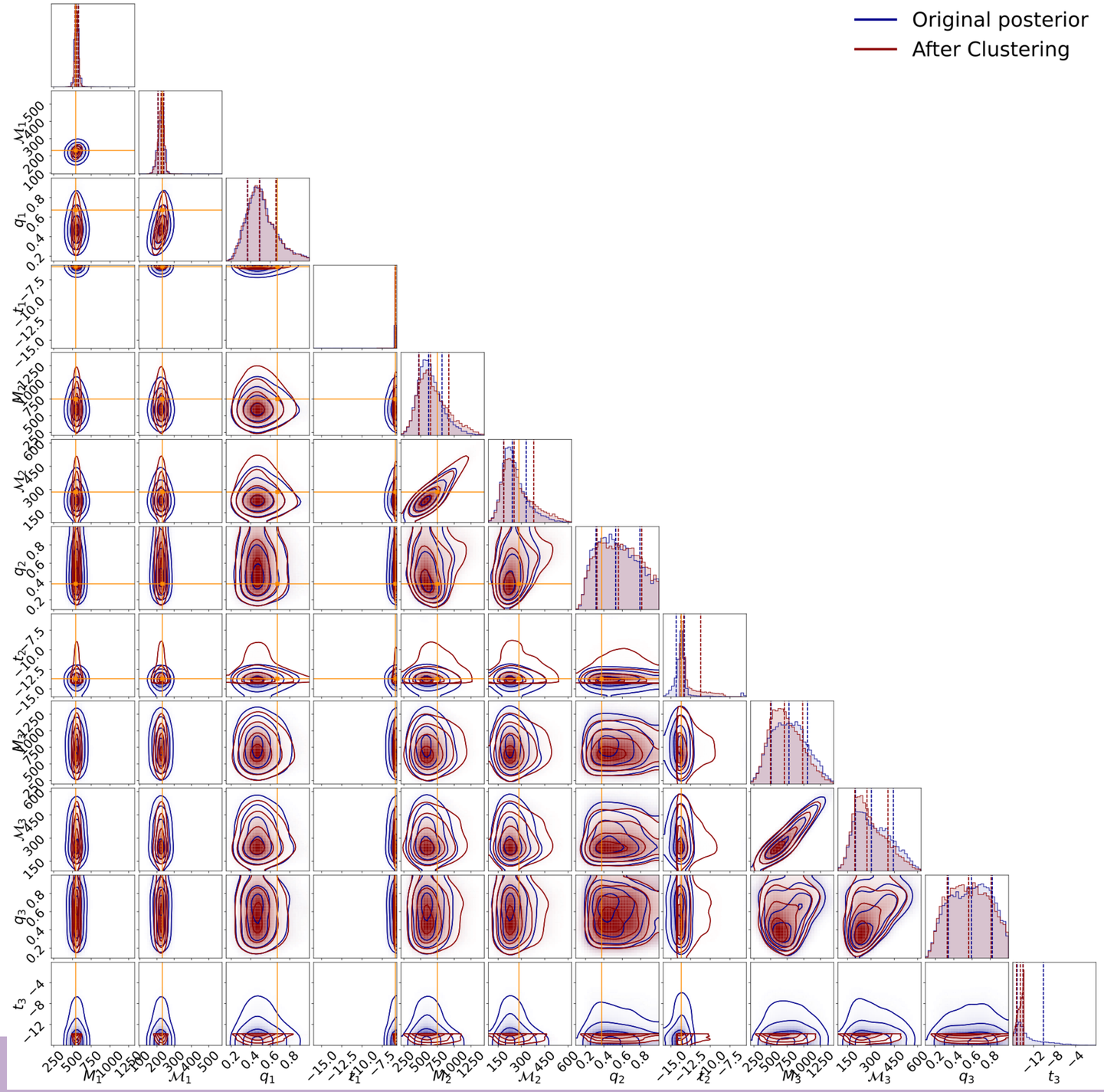


5.4	1.3	6.4	5.2	3.3	1.2	7.5	7.4	9.2	3.6	1.5	3.2
4.5	8.2	2.6	1.8	4.2	7.4	2.4	6.2	2.9	6.4	4.7	4.6
2.7	6.3	7.5	7.3	8.5	2.5	5.3	2.5	8.2	7.3	7.3	2.4
0.4	9.3	3.5	6.4	3.7	6.3	2.2	7.5	5.2	8.1	3.5	1.2
9.3	3.8	3.4	3.7	1.5	3.5	5.4	8.3	1.7	8.2	8.6	9.3
2.6	3.4	8.4	1.5	3.5	7.4	7.9	6.3	5.7	1.6	8.1	7.2

What happens with $N_{BBH} < 3$?

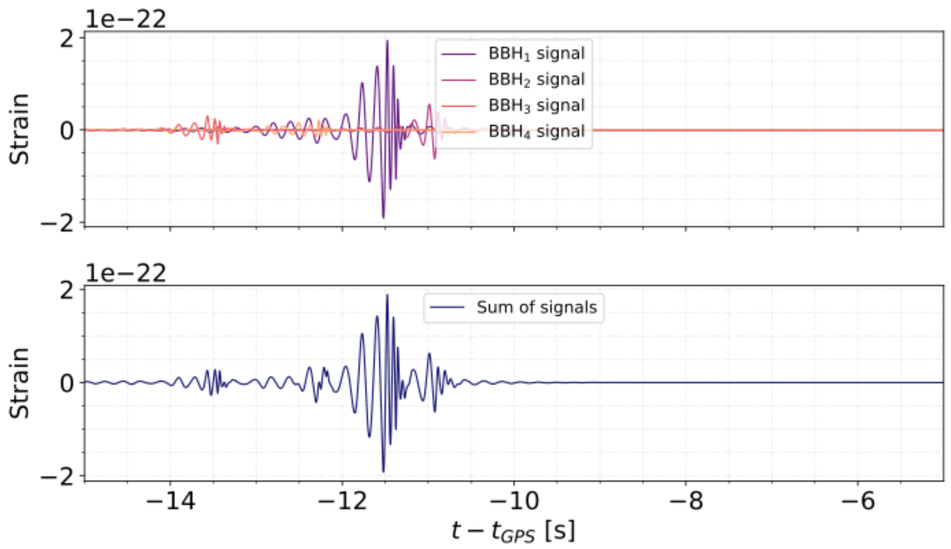


- 2 signal correctly identified
- Very large posteriors in $M, \mathcal{M}, q$ for the 3rd signal
- Take results «carefully»: the model wants to find 3 signals

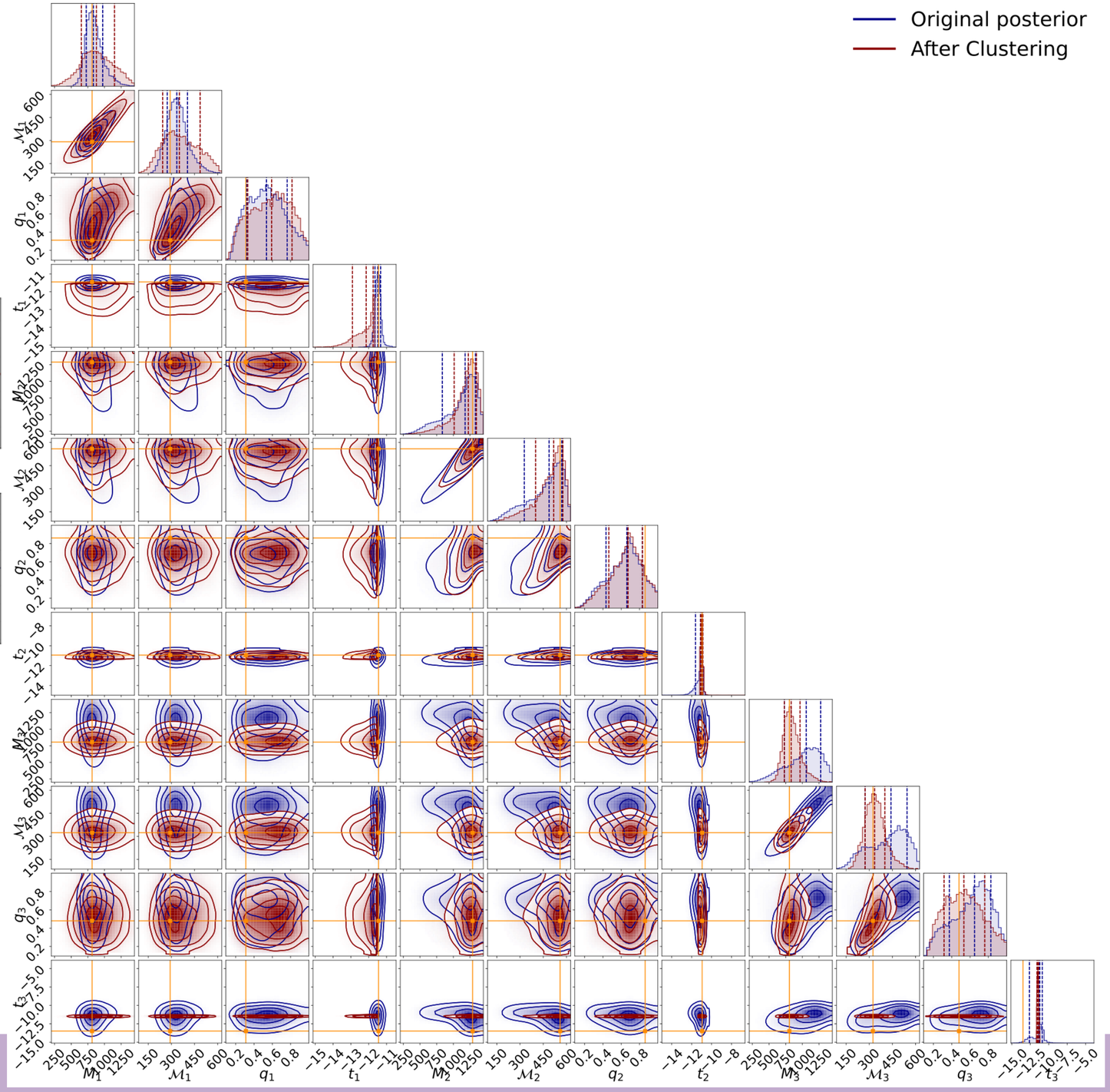


Papalini et al., arXiv:2505.02773, *submitted to CQG*

What happens with $N_{BBH} > 3$?



- The three louder signals are correctly identified !



Papalini et al., arXiv:2505.02773, *submitted to CQG*