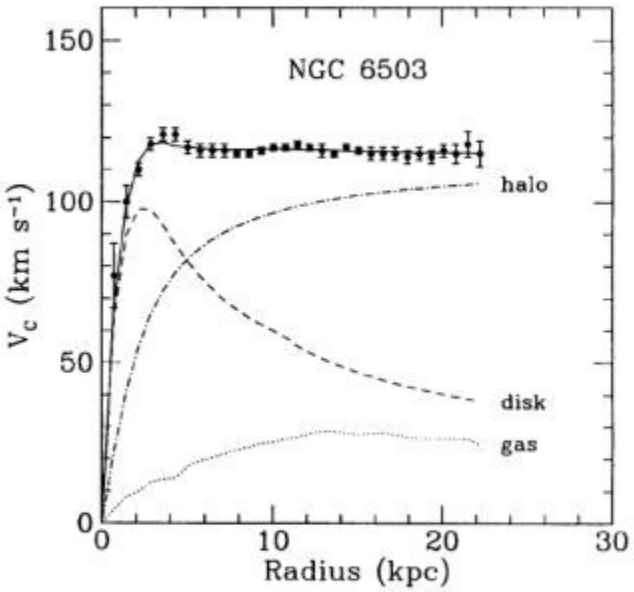
Gaussian Process Regression of the 2016 Invariant Mass Distribution

Aidan Hsu[1], Emrys Peets[1], Tom Eichlersmith[2]

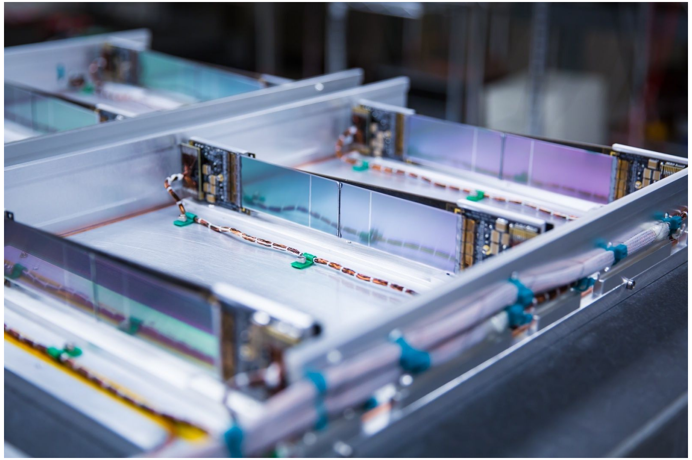
[1]Stanford University
[2]University of Minnesota

Dark Matter



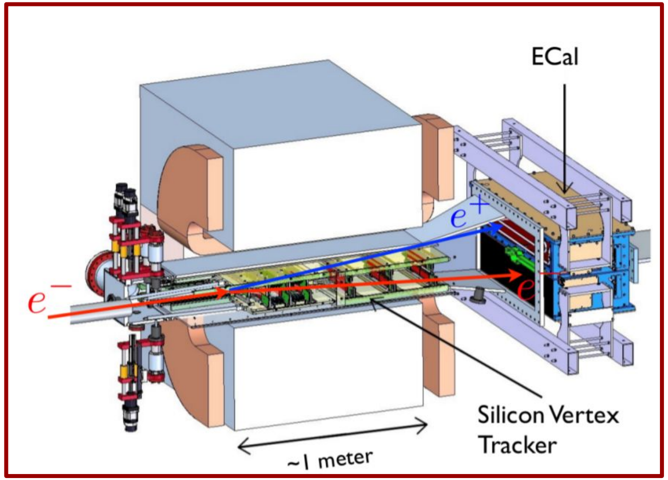
- Speed of rotation of galaxies is flat rather than decreasing with increasing radius.
- Dark Matter "halos" surround galaxies

Heavy Photon Resonance Search

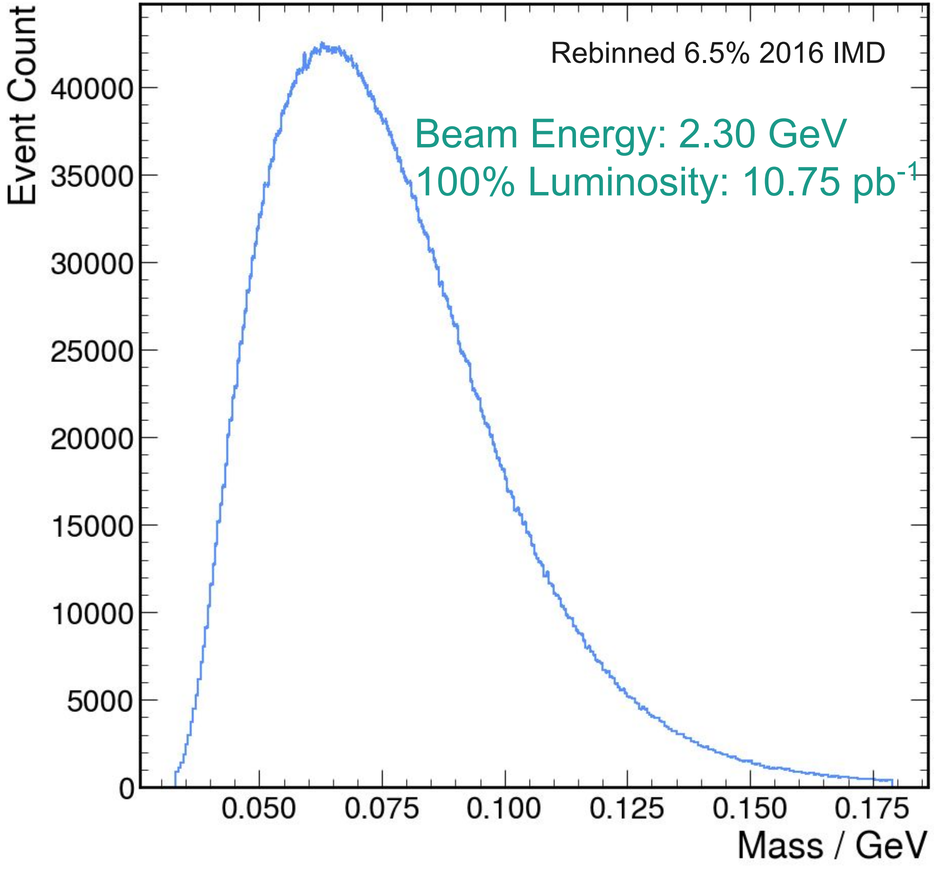


- Search for the “dark” or “heavy” photon via bump in background model
- Data collected by measuring energies of electron-positron pairs produced by bremsstrahlung

- CEBAF particle accelerator



2016 Invariant Mass Distribution



- Distribution of mass derived from energy of reconstructed electron-positron pairs
- A “bump” in the background can mean that the background model is not describing the data well

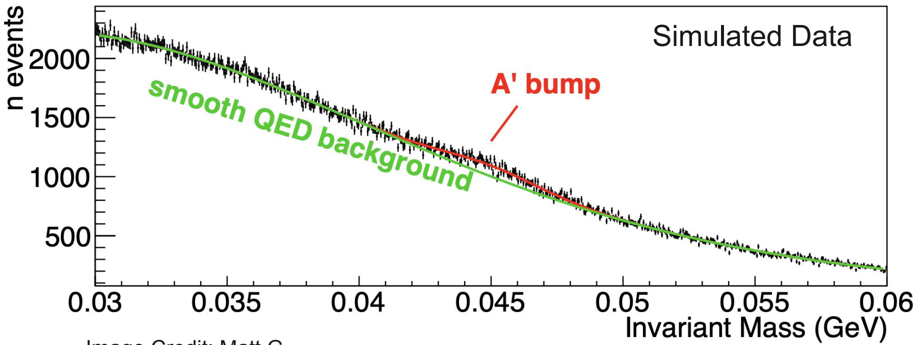
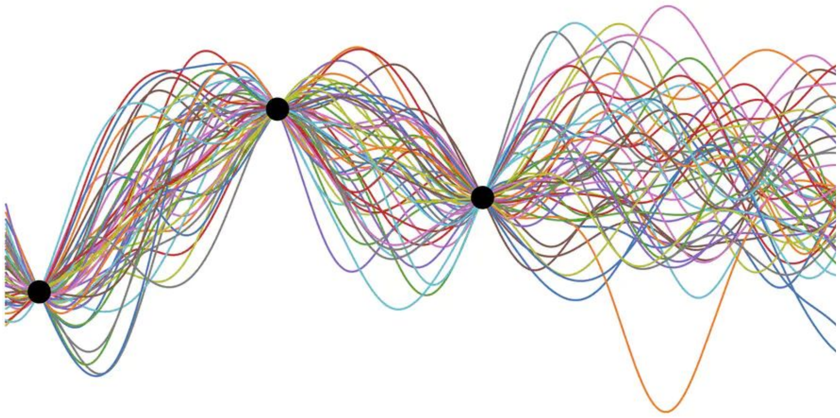
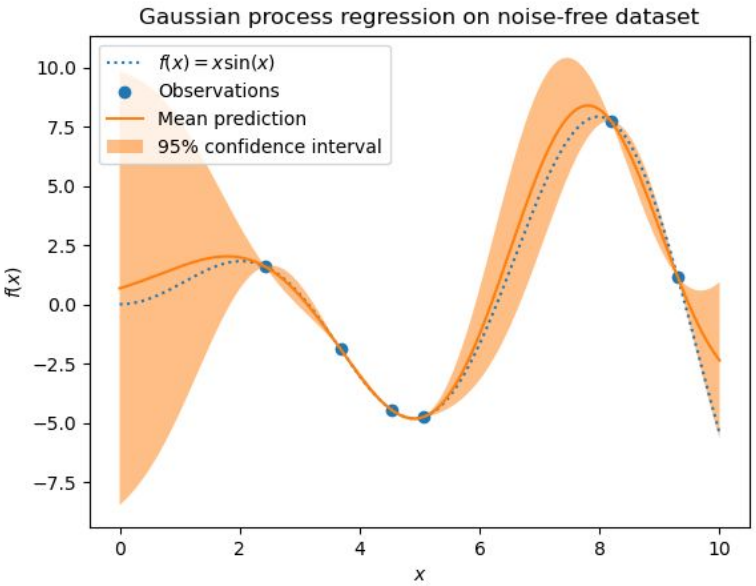


Image Credit: Matt G

Gaussian Process Regression





- Gaussian Process is a probability distribution over functions.
- The GP has mean and covariance (kernel)

Why GPR?

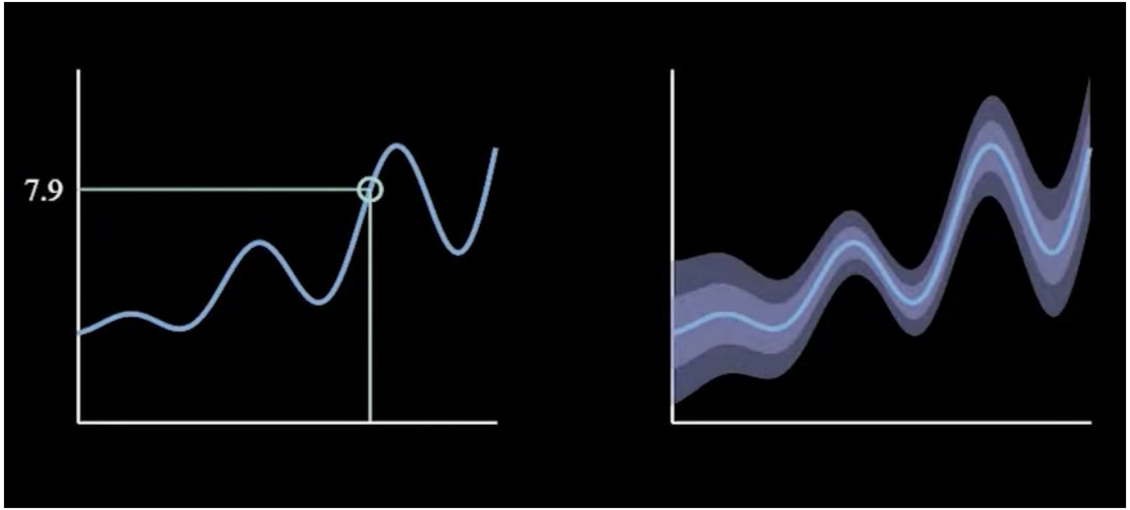
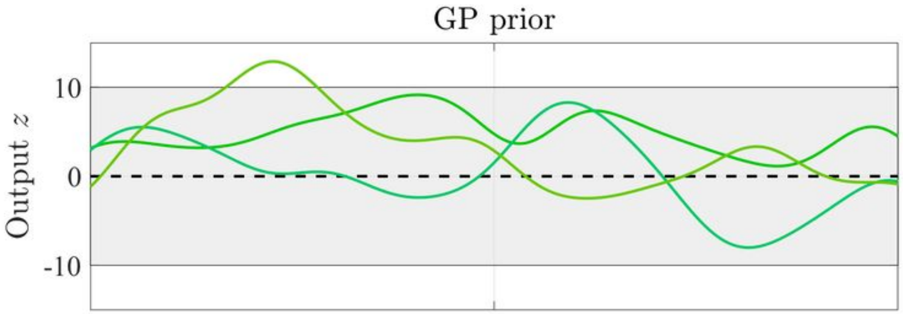


Image Credit

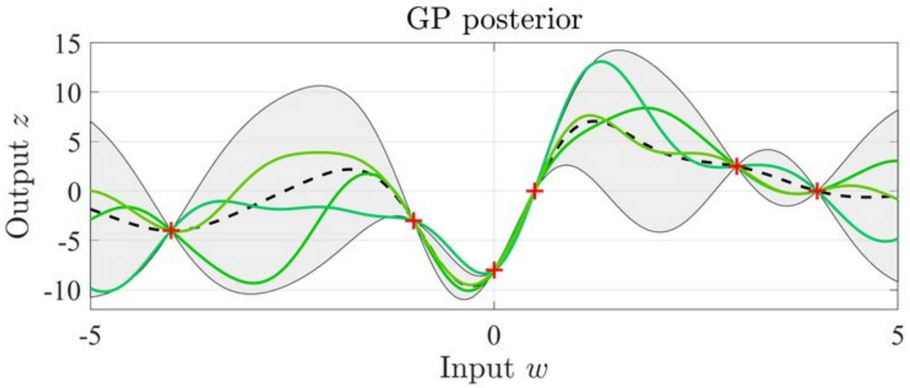
Other machine learning model vs. GPR

- Gives you a distribution instead of a point
- Functions sampled determined by a chosen kernel
- Kernel used for this fit was the rational quadratic.

Training

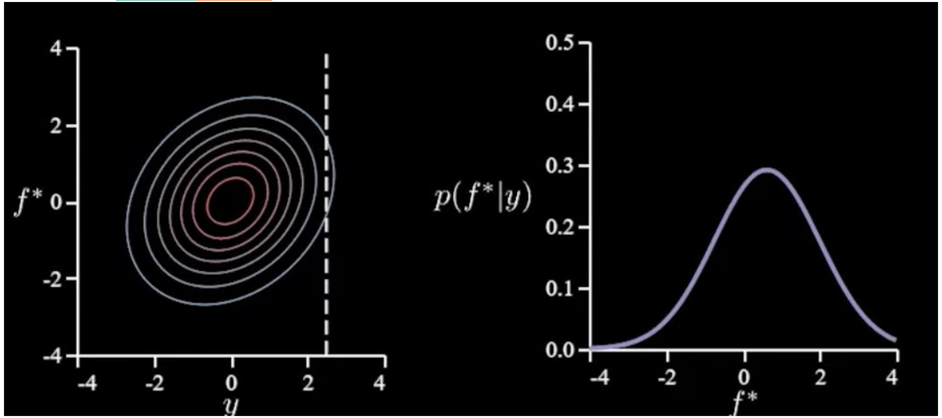


Before training



After training on red points

Predicting



[Image Credit](#)

- Example: you have one observation y which is 2.5
- The distribution of f* (true function) given the observation is the cross-section at y
- Works at higher dimensions

Gaussian Process Regression with HPS

Motivation

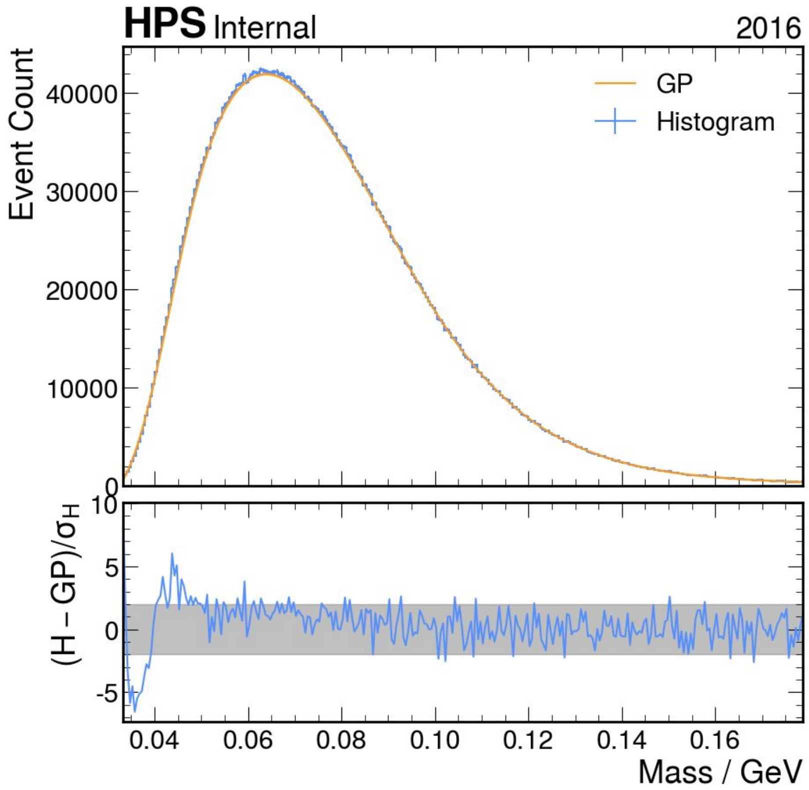
- independent fit methods in progress to compare with global background model

Work done so far

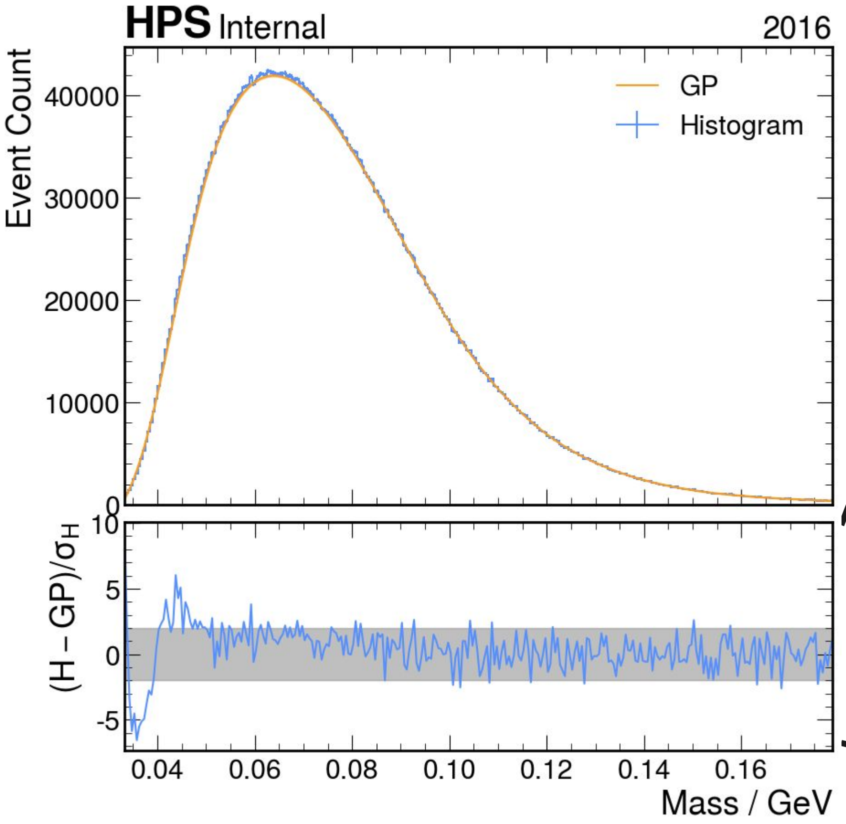
- Aidan, Tom, Emrys have been collaborating on slack
- Github link to code in progress
    - Reach out if you want to look at anything!

Applying GPR to 6.5% 2016 IMD

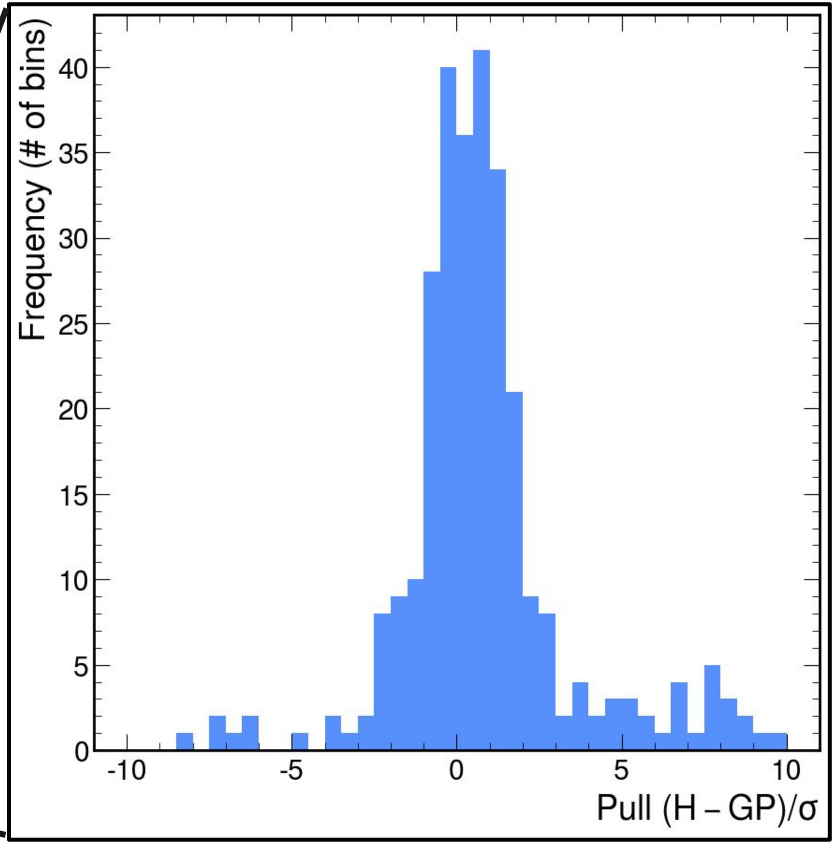
- The top plot is the histogram and the GP prediction. The bottom plot shows a measure of statistical uncertainty at various bins. It is the difference between the histogram and the GP divided by standard deviation.



Key Displays

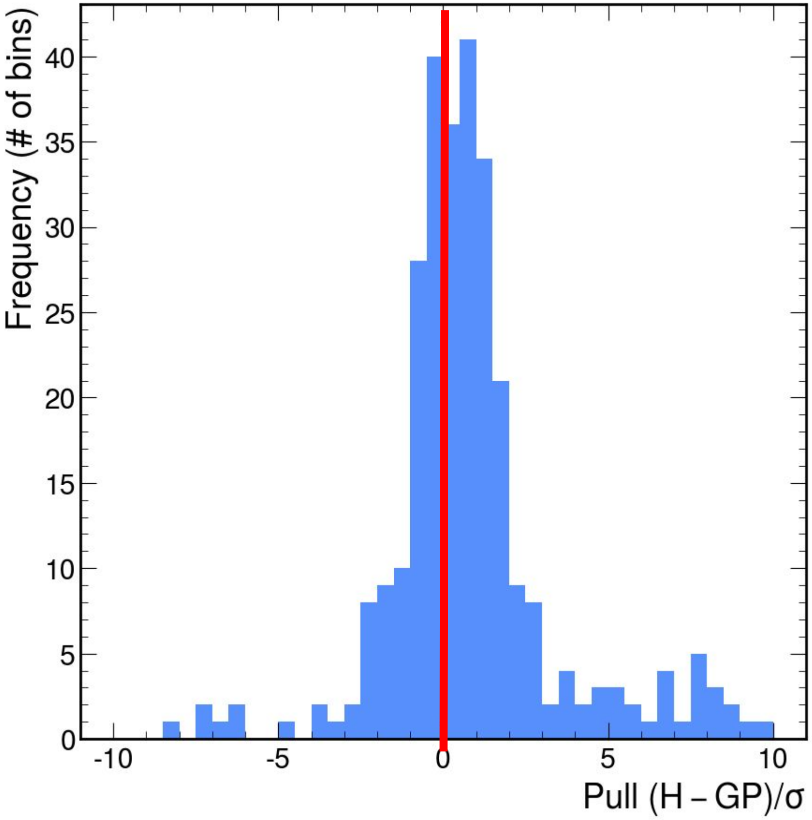


Pull Plot



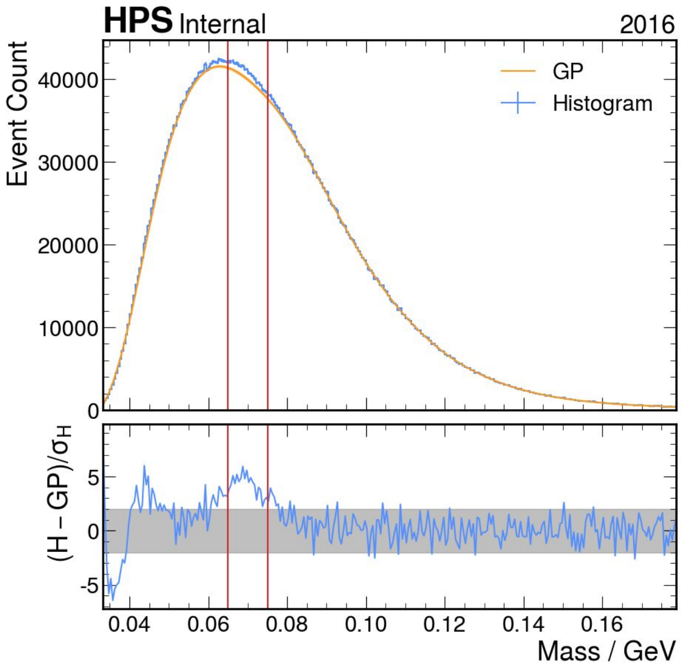
Pull Plot

- Histogram showing amount of times points occurred at various significances.
- Skewed to right



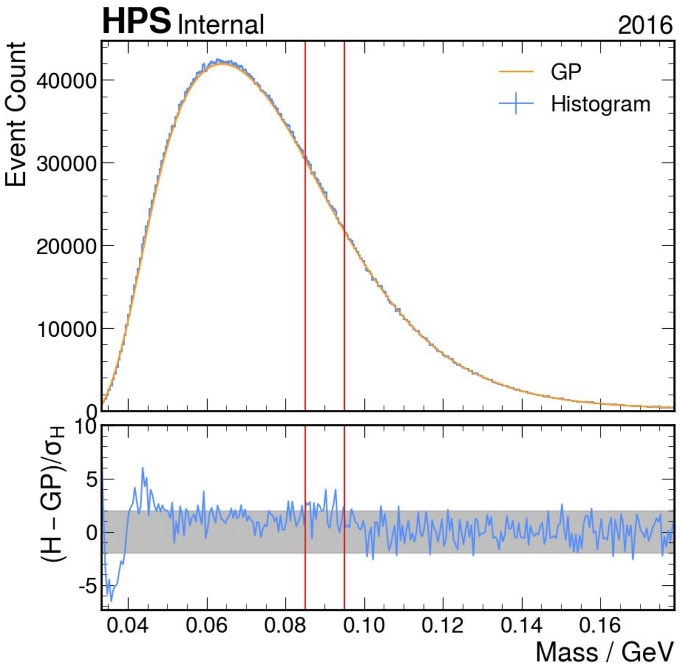
Blinding Study [⅓]

- Range between red lines are blinded to the fitting process
- Blinding at this range makes the fit noticeably lower than the data on either side of the blind

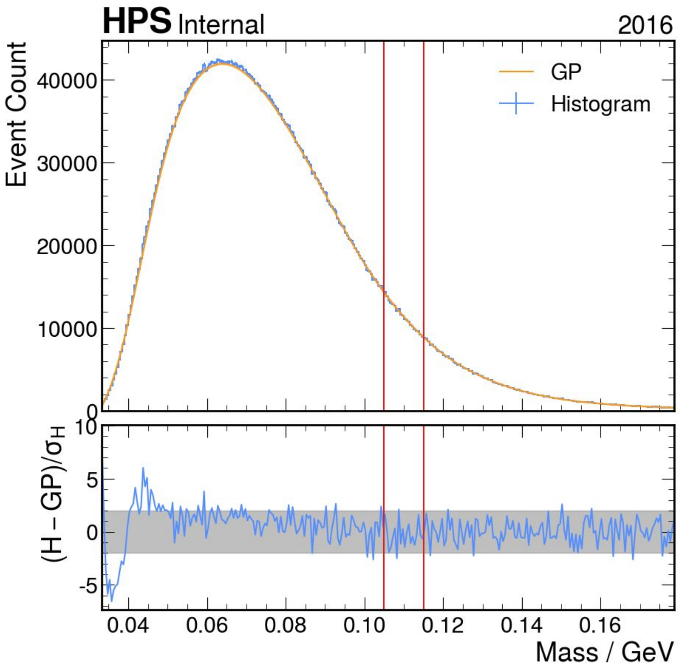


Blinded Region: 65-75 MeV

Blinding Study [⅔]



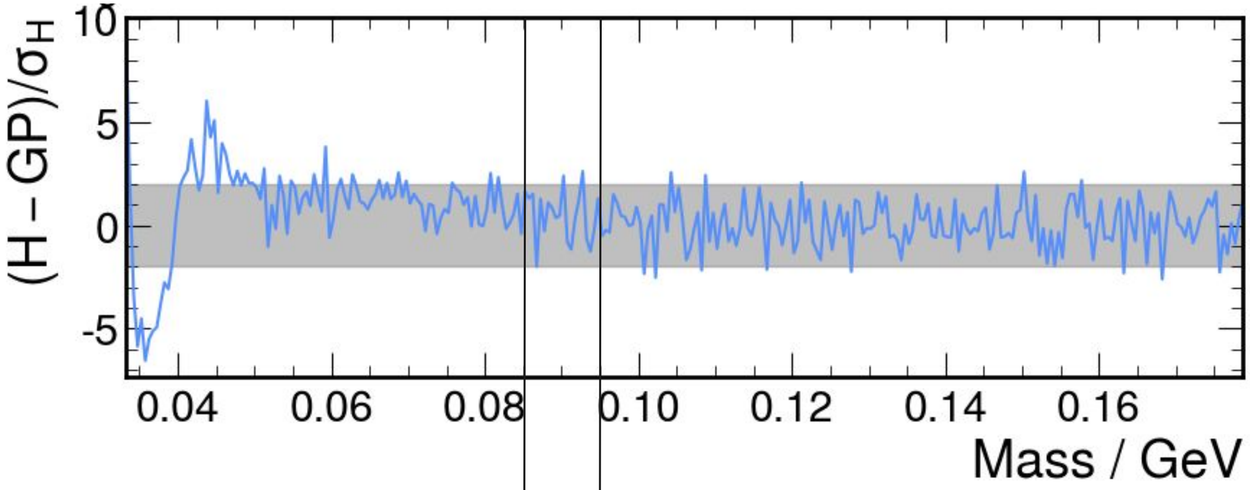
Blinded Region: 85-95 MeV



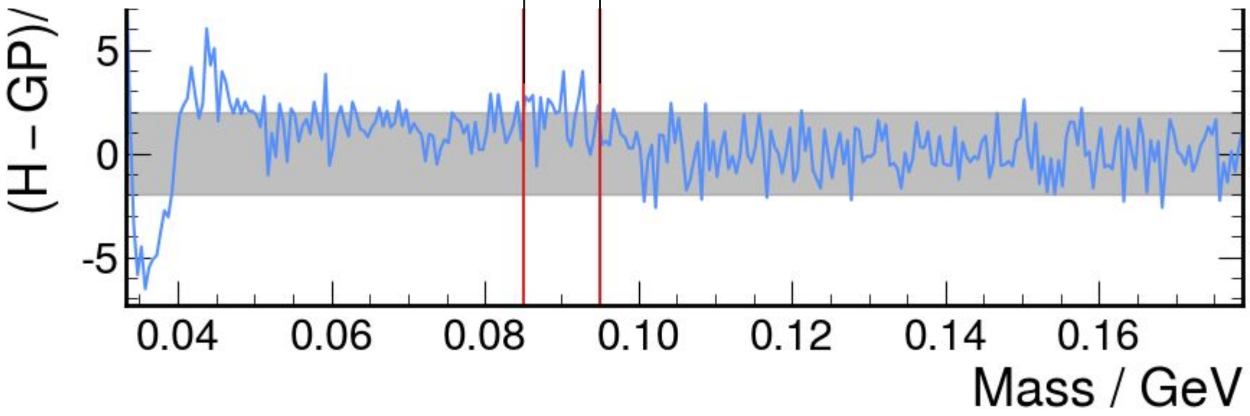
Blinded Region: 105-115 MeV

Blinding Study [⅔]

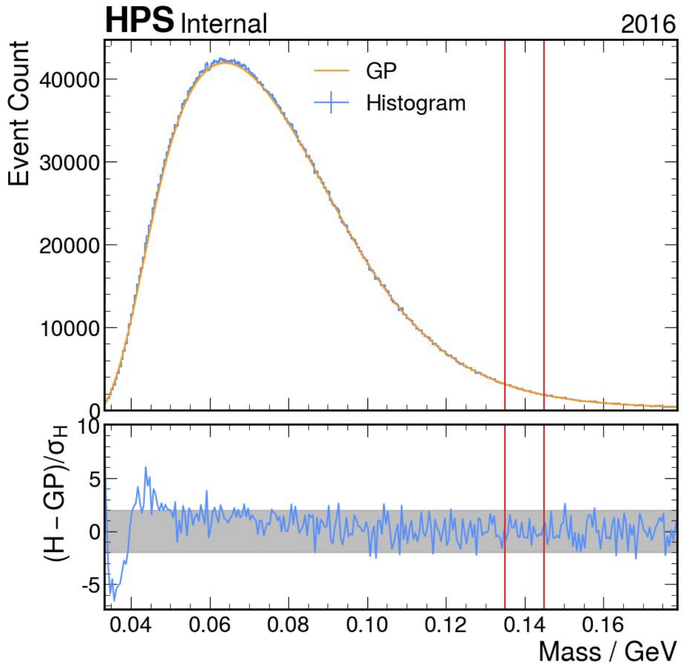
No Blind



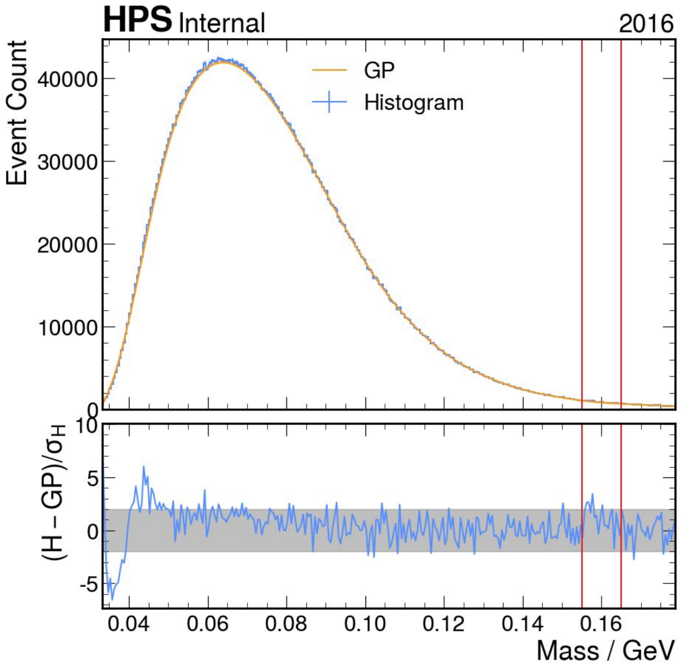
85-95 MeV



Blinding Study [3/3]

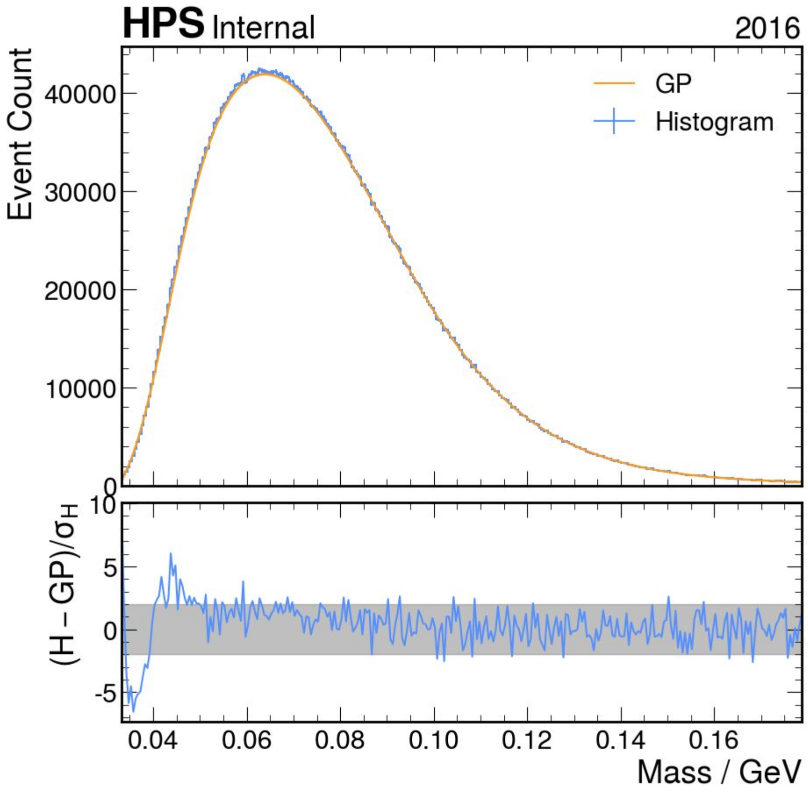


Blinded Region: 135-145 MeV

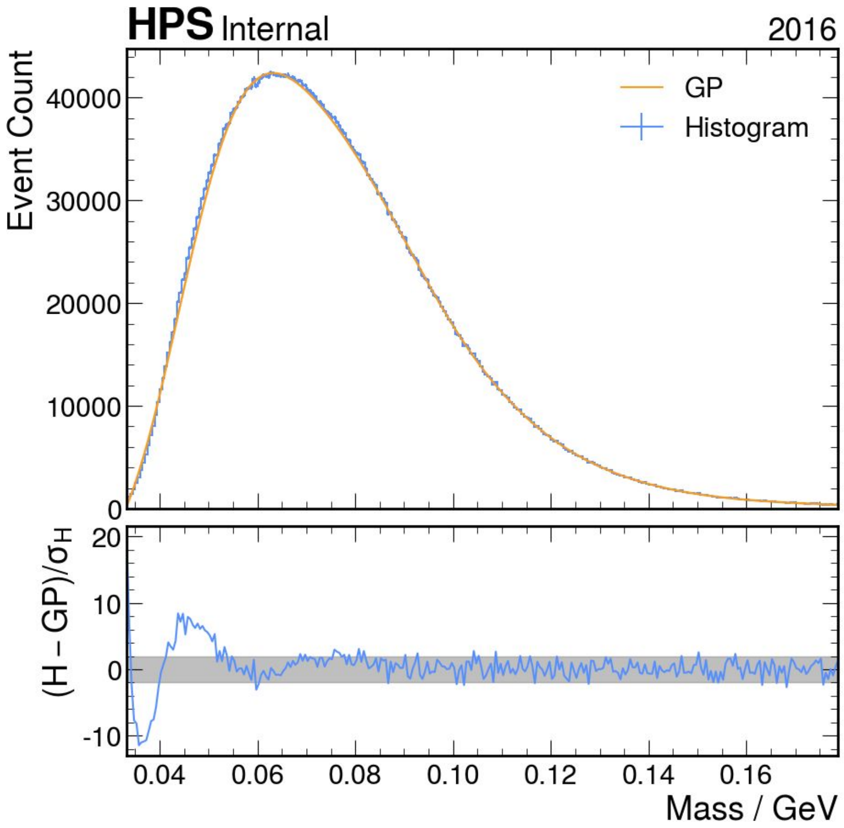


Blinded Region: 155-165 MeV

Testing Kernels

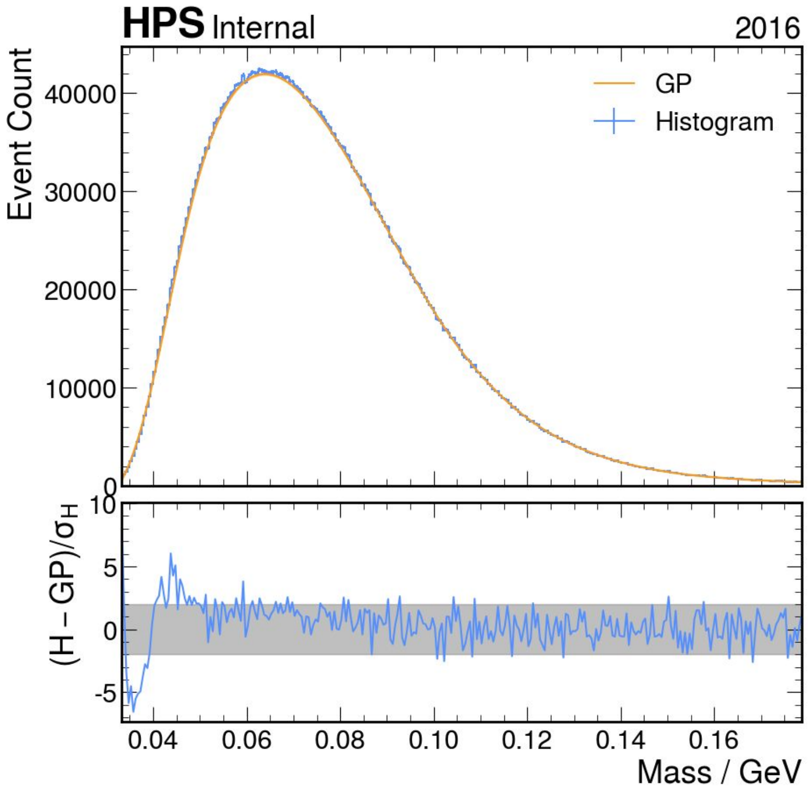


Rational Quadratic

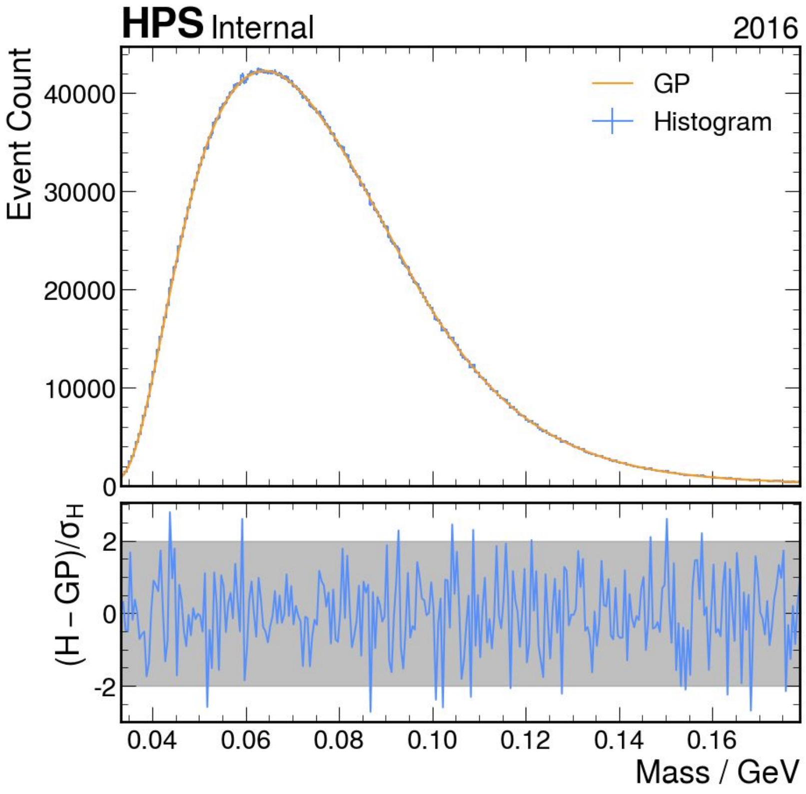


Radial Basis Function

Testing Kernels



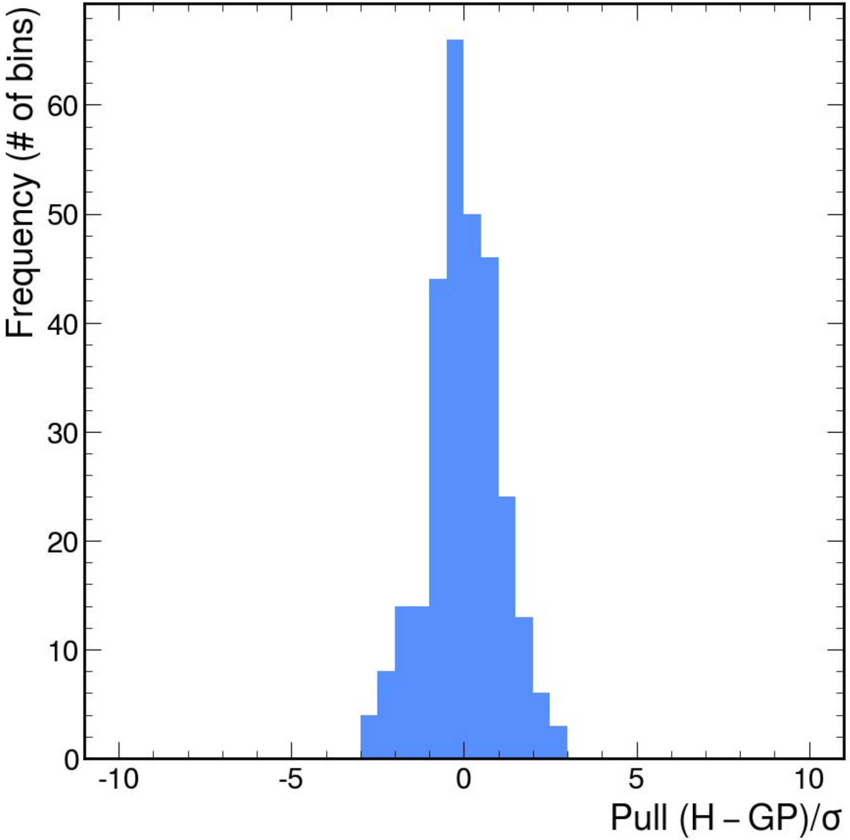
Rational Quadratic



Radial Basis Function * linear kernel

RBF * Linear Pull Plot

- No longer skewed to the right



Conclusions and Next Steps

- GPR yields promising results but for some kernels there are issues in fit at the rising edge of the IMD
    - 5 sigma deviations are suspicious
- Conducted limited blinded window tests
    - Seems to work better in falling edge of distribution
- Experimented with different kernels
    - Linear kernel * RBF kernel seems to provide good fit result!

Next steps [in order]

1. Test “good kernel” with blinding different regions.
    - Determine appropriate window sizes for each mass hypothesis [based on mass resolution]
2. Generate signal histograms to test sensitivity to signal injection at 6.5% level
3. Create 100% 2015 IMDs with signal injected as well
4. Once blinded procedure is determined: go to 100% 2016