Research Training Group Physics of the Heaviest Particles at the LHC | RWTH AACHEN UNIVERSITY

Collaborative Research Center TRR 257

PHI

Particle Physics Phenomenology after the Higgs Discovery

Institute for Theoretical Particle Physics and Cosmology | RWTH AACHEN UNIVERSITY

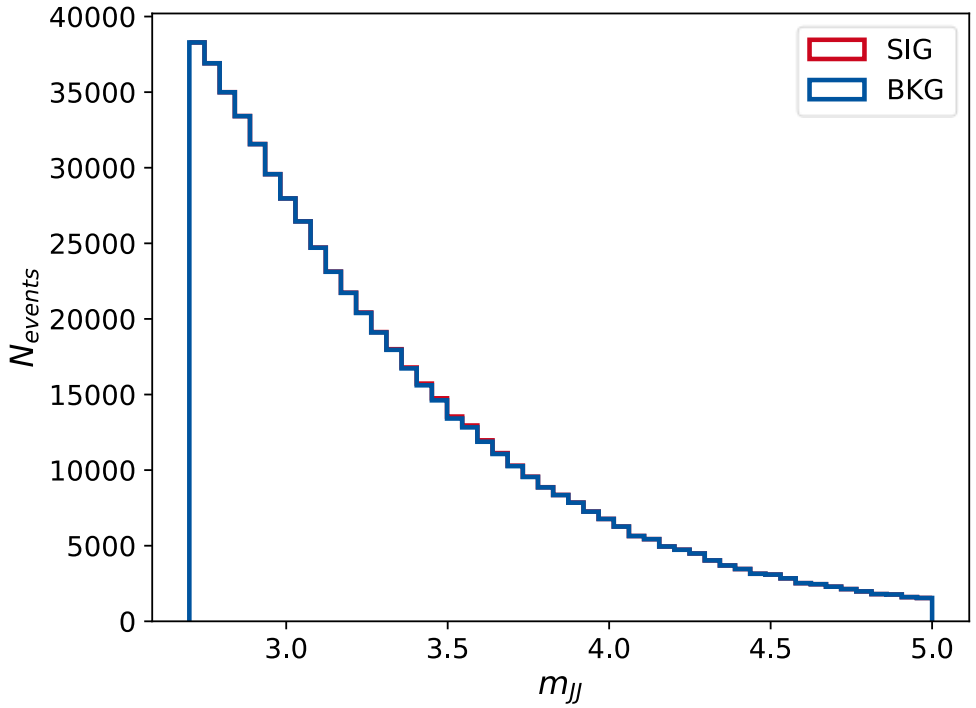
Picking the right setup for anomaly detection

Marie Hein

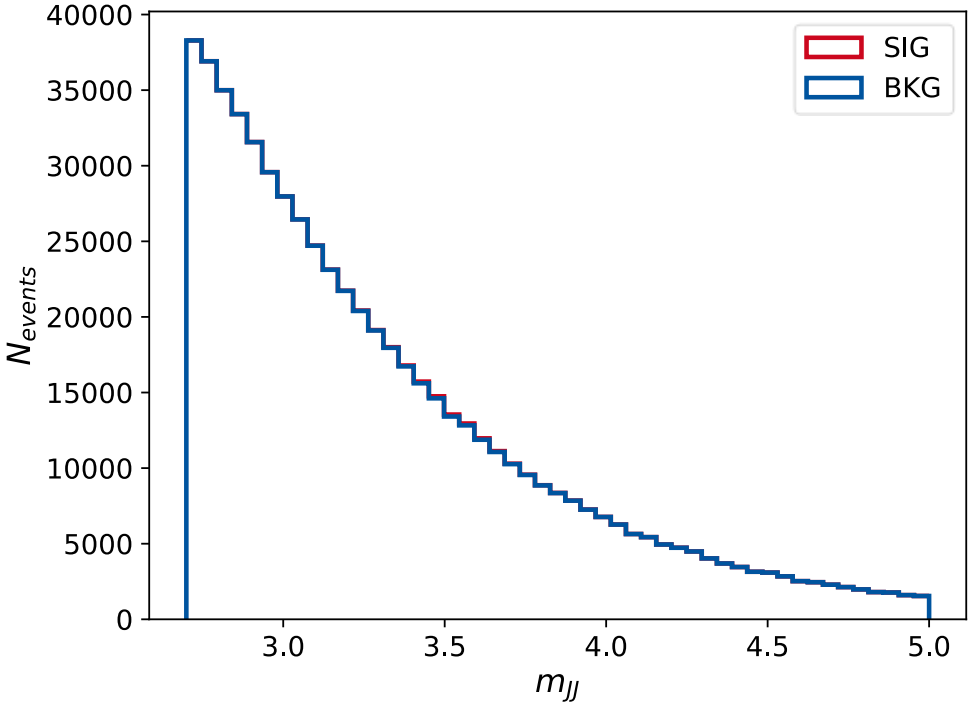
with Gregor Kasieczka, Michael Krämer, Louis Moureaux, Alexander Mück, Tobias Quadfasel and David Shih

CRC Young Scientists Meeting 2025

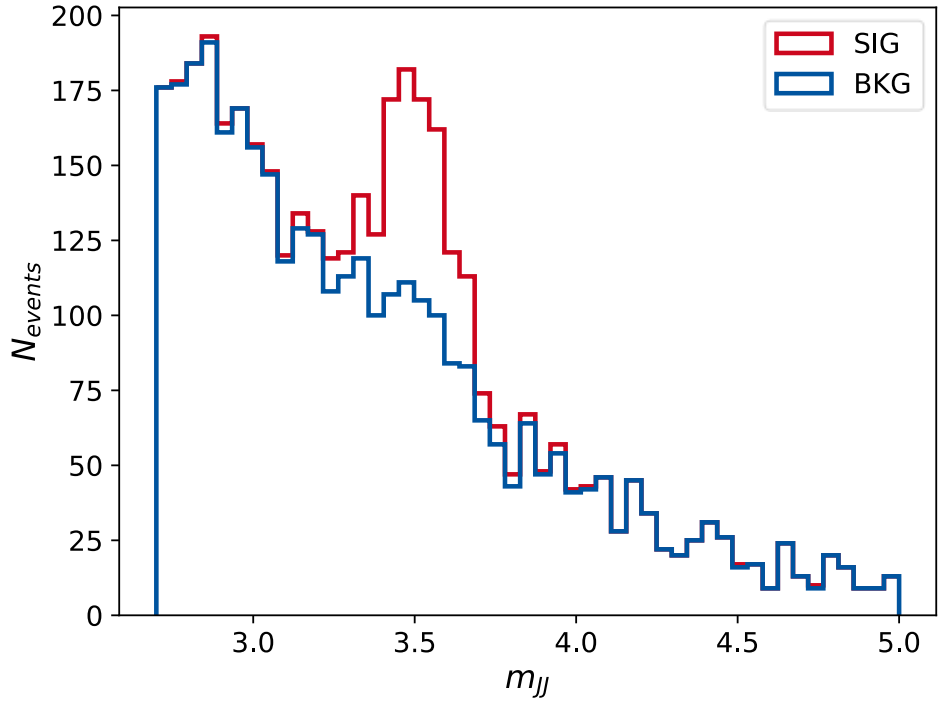
Anomaly Detection



Anomaly Detection



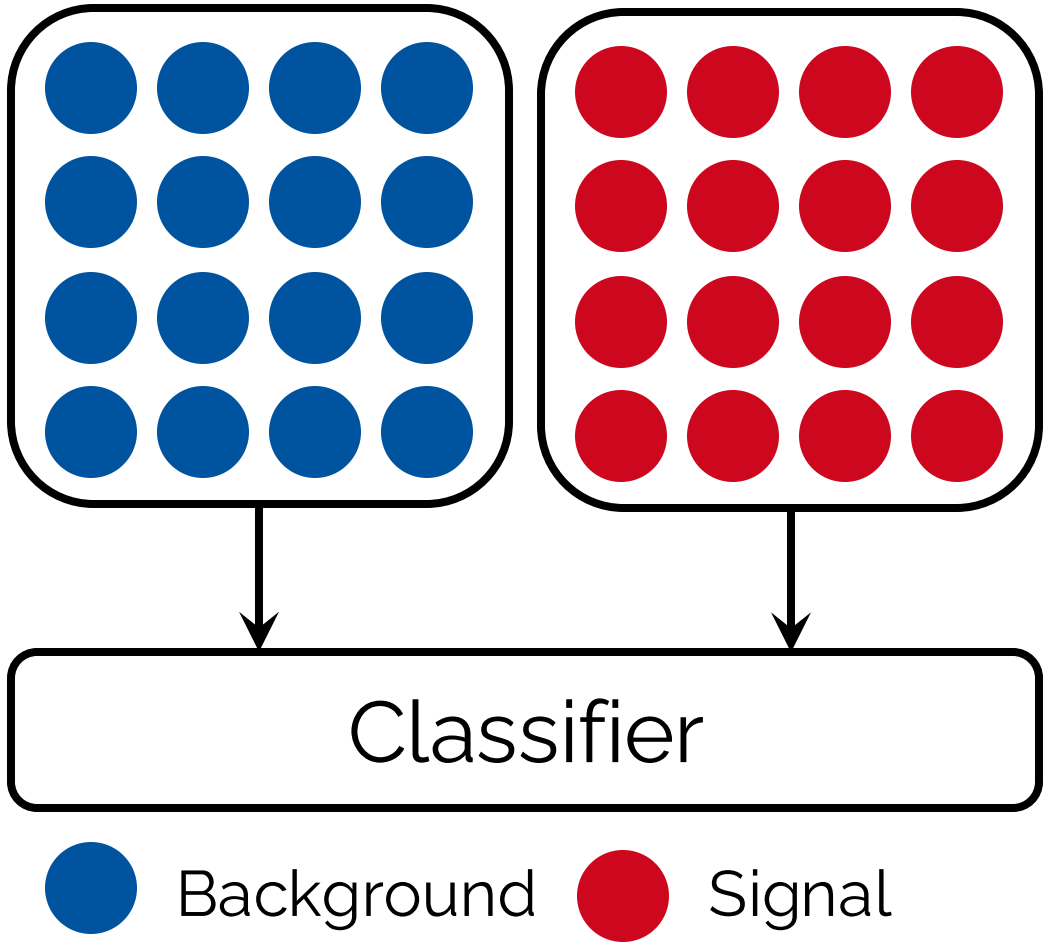
Anomaly Detector



Classification Problem

- Optimal classifier

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$$



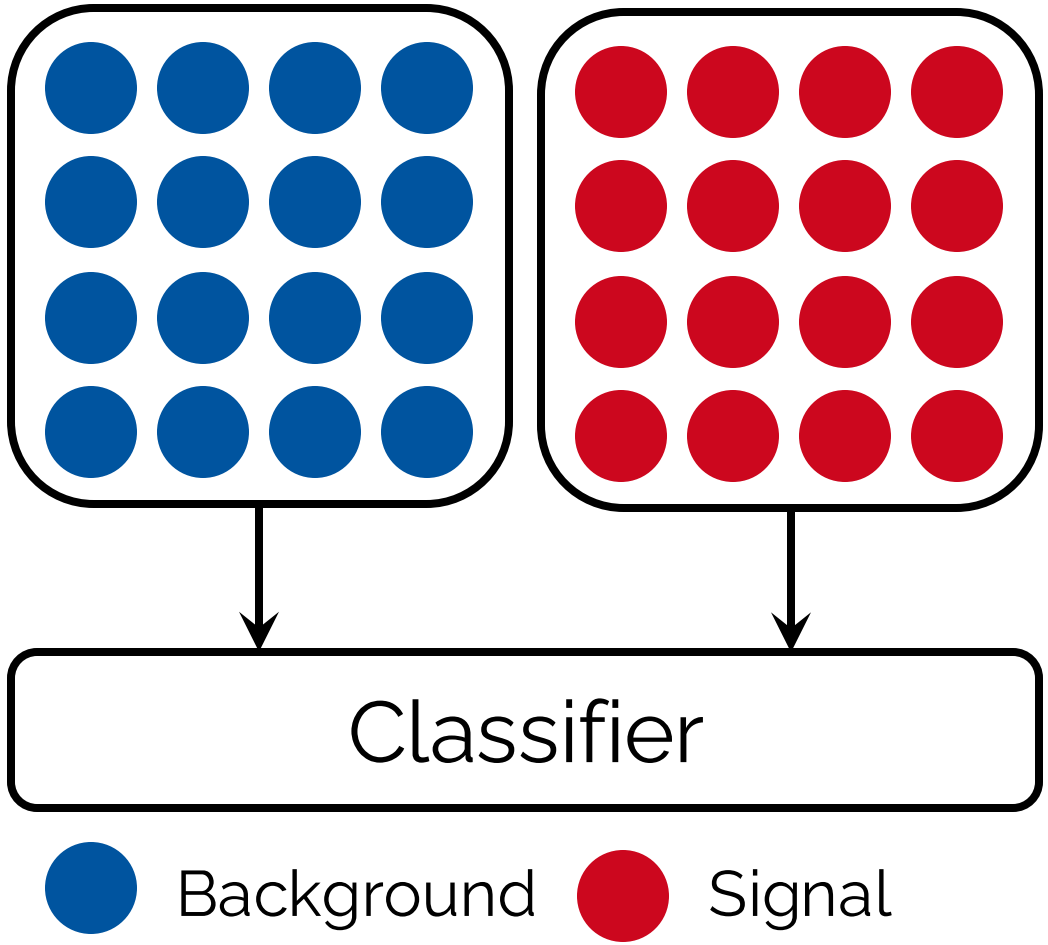
Classification Problem

- Optimal classifier

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$$

- For Machine Learning use binary cross entropy loss

$$BCE = -\log p_{\text{pred, true}}$$



Classification Problem

- Optimal classifier

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$$

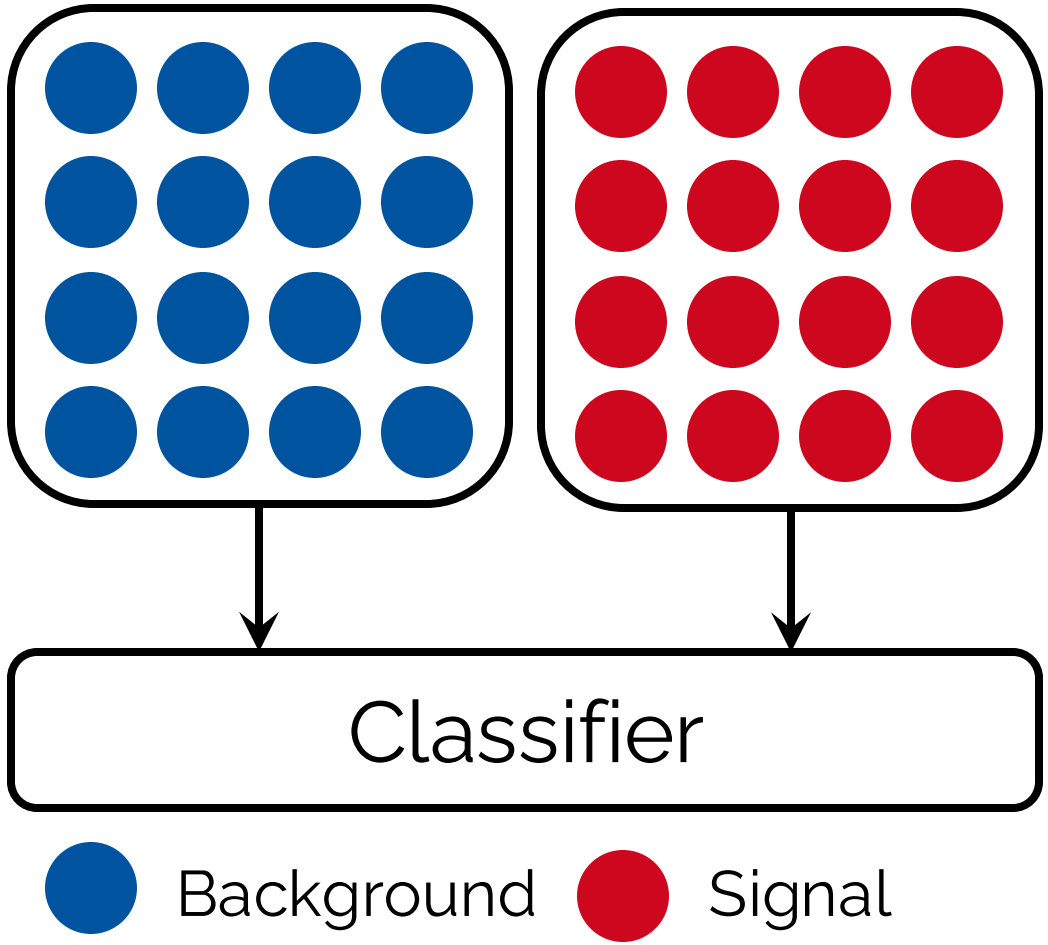
- For Machine Learning use binary cross entropy loss

$$BCE = -\log p_{\text{pred, true}}$$

→Optimal solution function monotonically related to R_{optimal}

$$f(x) = \frac{p_S(x)}{p_B(x) + p_S(x)}$$

→Same decision boundaries



Weakly Supervised Classification

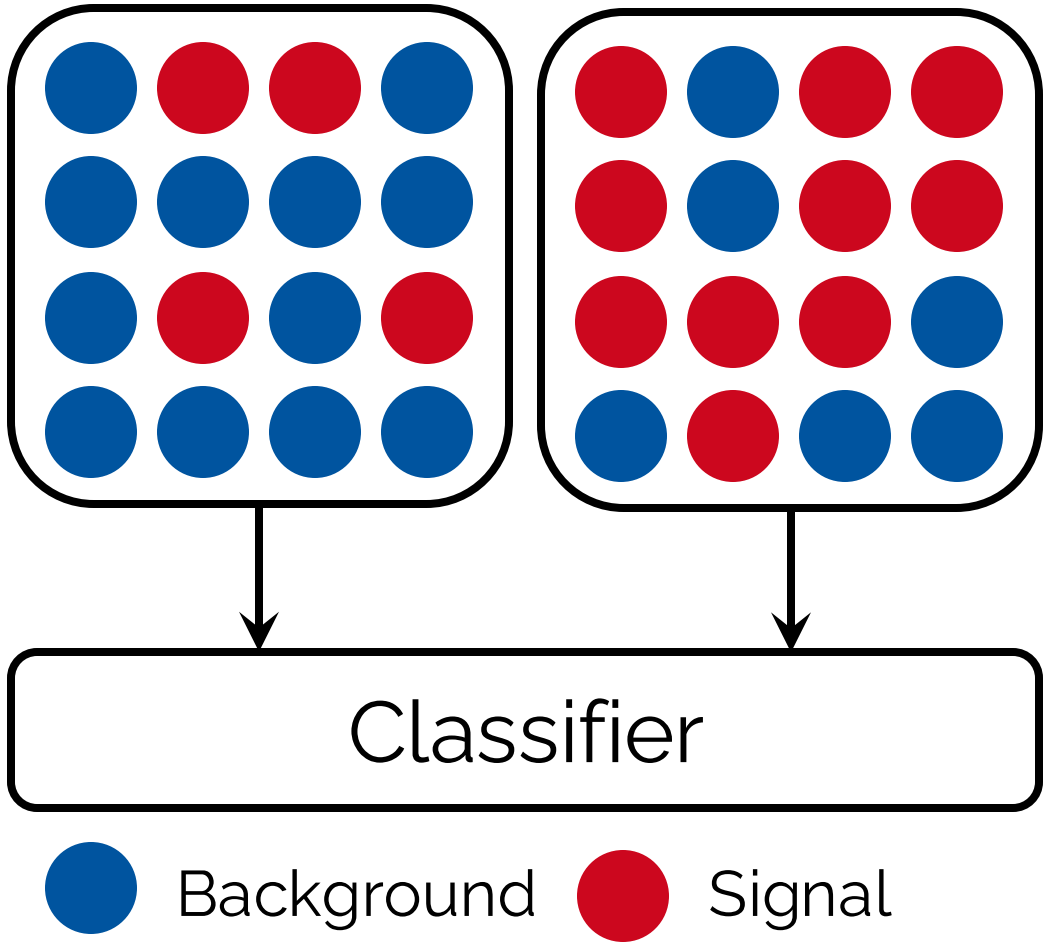
"Classification without labels: Learning from mixed samples in high energy physics" [1709.02949], E. Metodiev, B. Nachman, J. Thaler

- Optimal classifier

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$$

- For mixed datasets with signal fractions f_i

$$R_{\text{mixed}}(x) = \frac{f_1 R_{\text{optimal}}(x) + (1 - f_1)}{f_2 R_{\text{optimal}}(x) + (1 - f_2)}$$



Weakly Supervised Classification

"Classification without labels: Learning from mixed samples in high energy physics" [1709.02949], E. Metodiev, B. Nachman, J. Thaler

- Optimal classifier

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$$

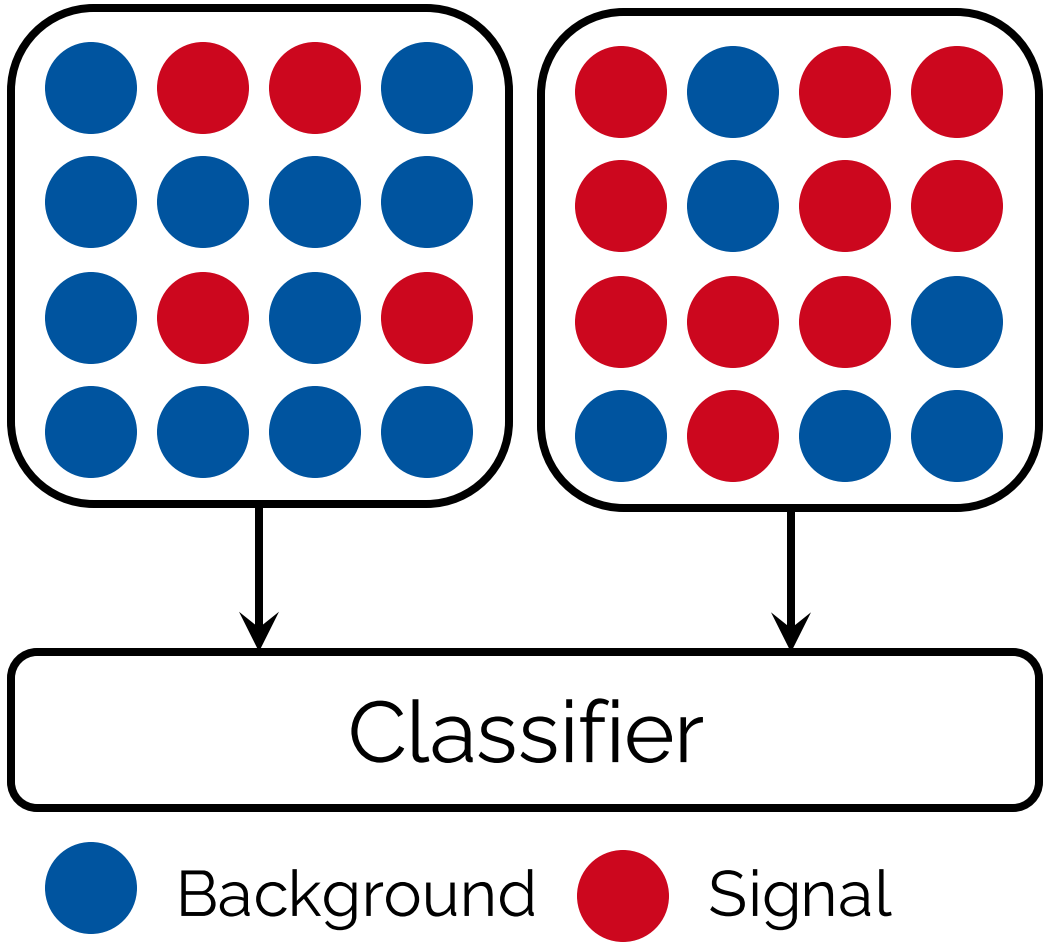
- For mixed datasets with signal fractions f_i

$$R_{\text{mixed}}(x) = \frac{f_1 R_{\text{optimal}}(x) + (1 - f_1)}{f_2 R_{\text{optimal}}(x) + (1 - f_2)}$$

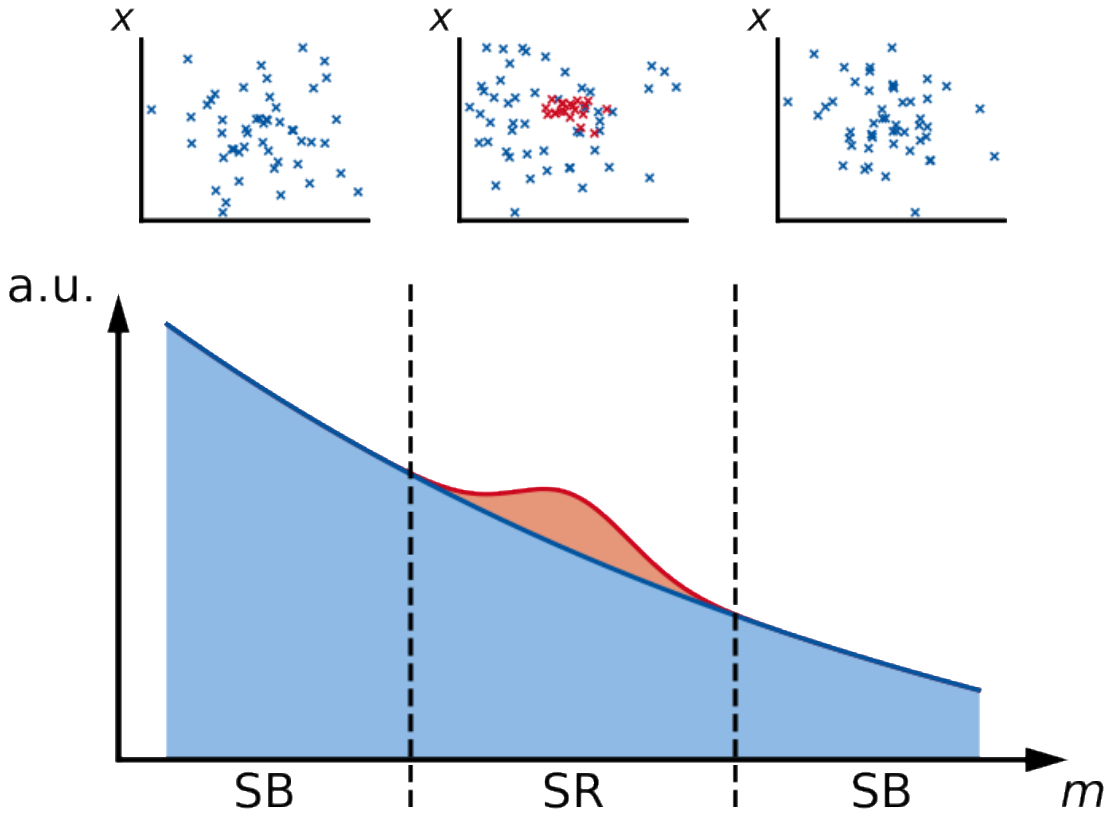
→Monotonically increasing function of

$R_{\text{optimal}}(x)$ as long as $f_1 > f_2$

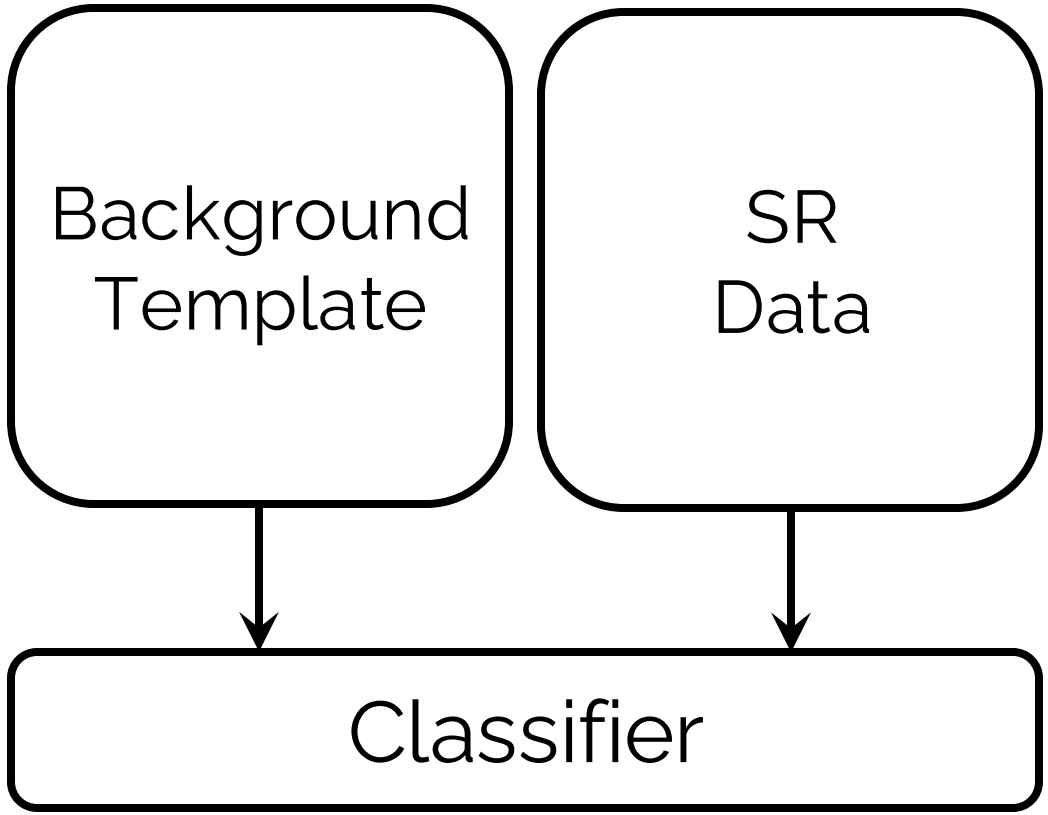
→Same decision boundaries



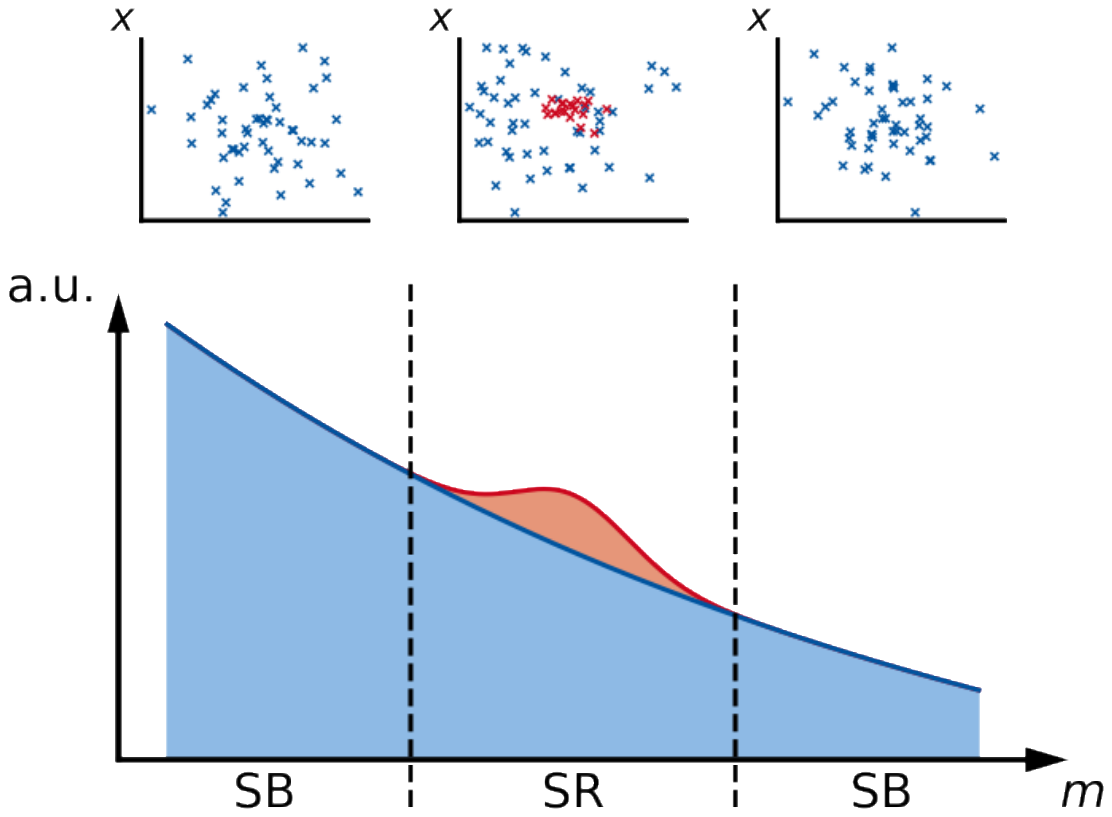
Application to resonance searches



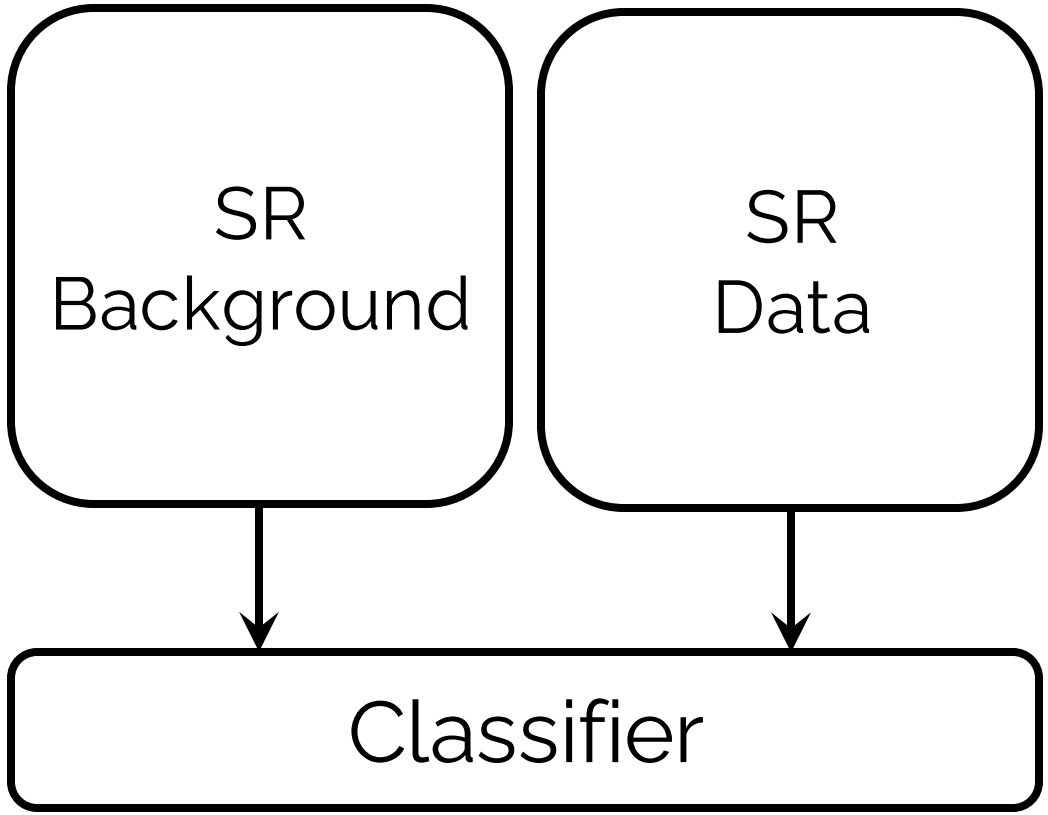
Recreated from [2109.00546]



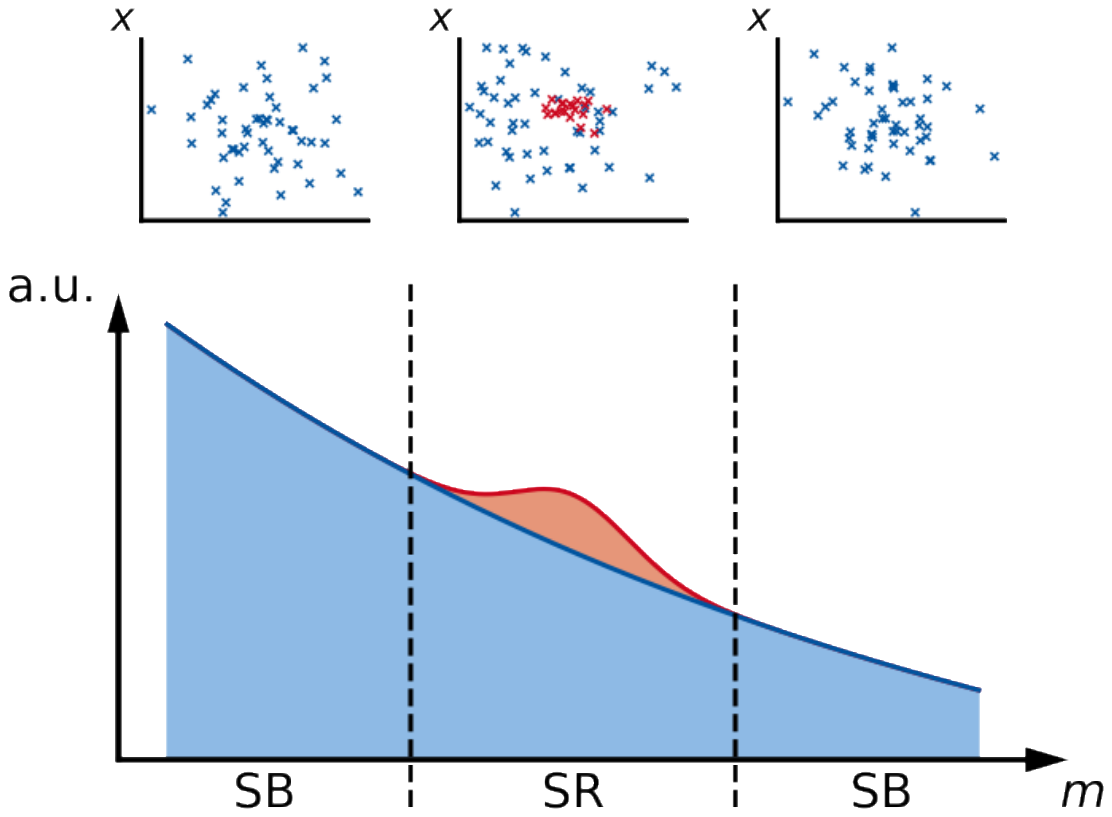
Idealized Anomaly Detector



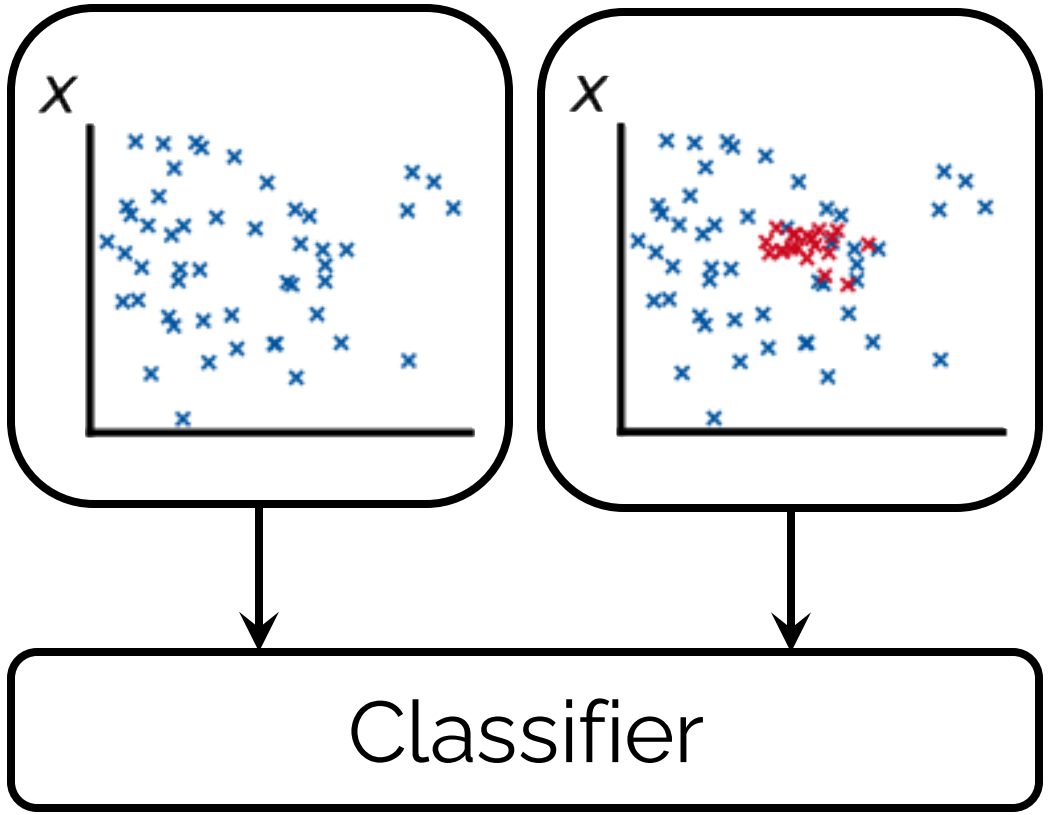
Recreated from [2109.00546]



Application to resonance searches



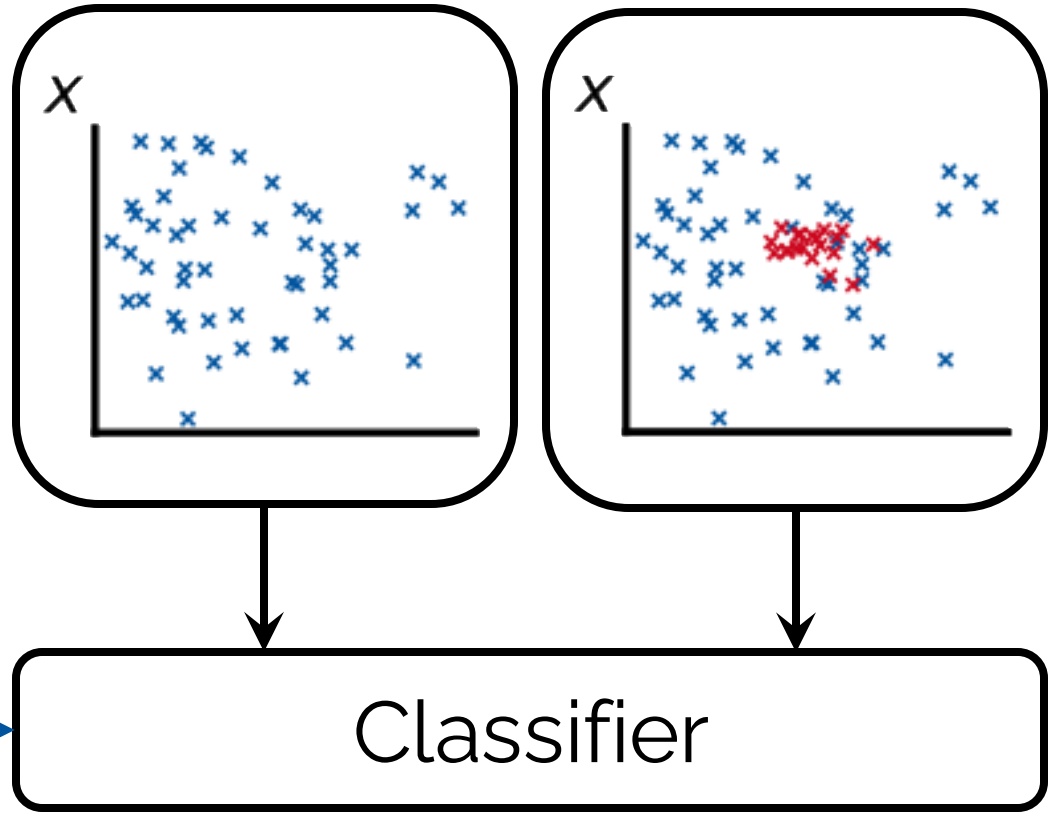
Recreated from [2109.00546]



Application to resonance searches

How do we pick the best classifier architecture?

Recreated from [2109.00546]



Picking the classifier architecture

Pick on simulations

Pick on data

Picking the classifier architecture

Pick on simulations

- Advantages: metrics that directly access background & signal labels

Pick on data

Picking the classifier architecture

Pick on simulations

- Advantages: metrics that directly access background & signal labels
- Disadvantages: less model agnostic, dependent on specific simulation & chosen signals

Pick on data

Picking the classifier architecture

Pick on simulations

- Advantages: metrics that directly access background & signal labels
- Disadvantages: less model agnostic, dependent on specific simulation & chosen signals

Pick on data

- Advantages: more model agnostic

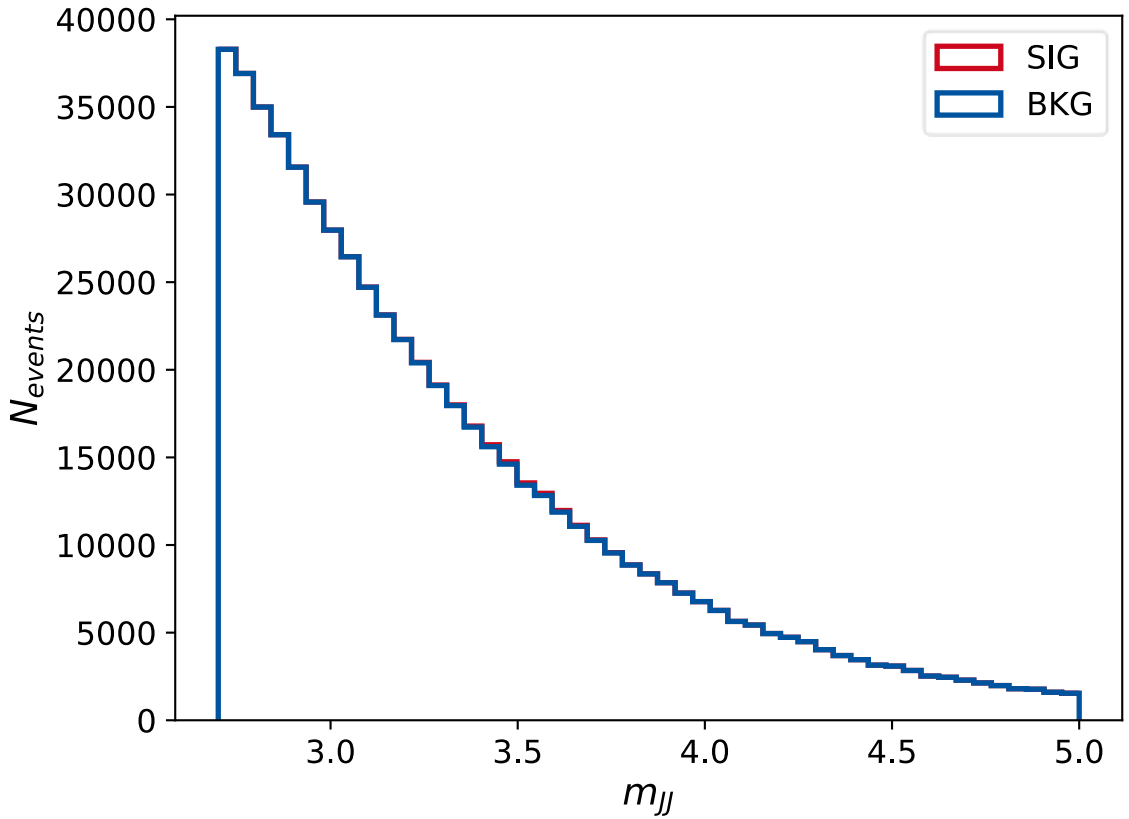
Picking the classifier architecture

Pick on simulations

- Advantages: metrics that directly access background & signal labels
- Disadvantages: less model agnostic, dependent on specific simulation & chosen signals

Pick on data

- Advantages: more model agnostic
- Disadvantages: limited number of signal events results in noisy metrics



Picking the classifier architecture

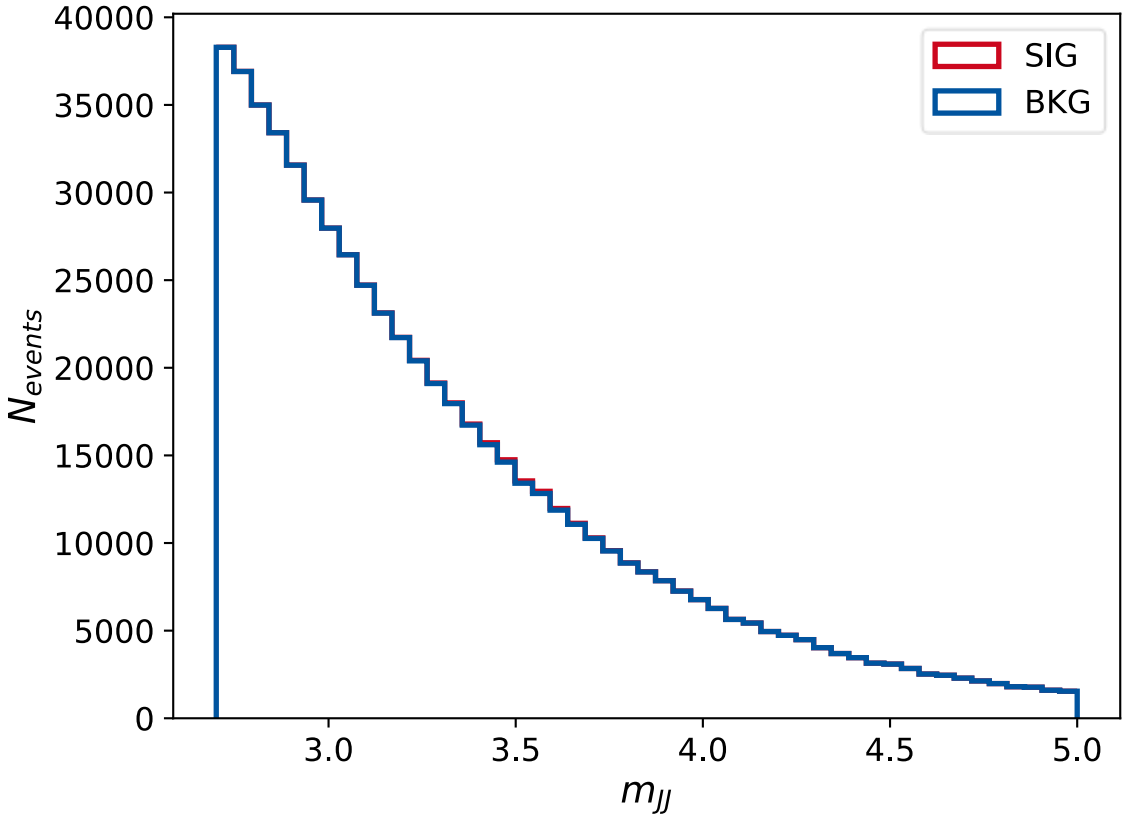
Pick on simulations

- Advantages: metrics that directly access background & signal labels
- Disadvantages: less model agnostic, dependent on specific simulation & chosen signals

Pick on data

- Advantages: more model agnostic
- Disadvantages: limited number of signal events results in noisy metrics

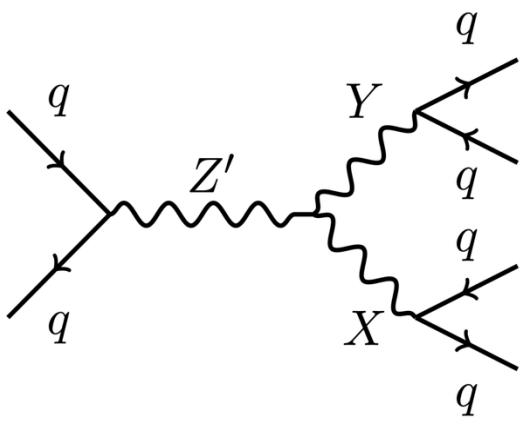
  →Investigate how problematic this noise is

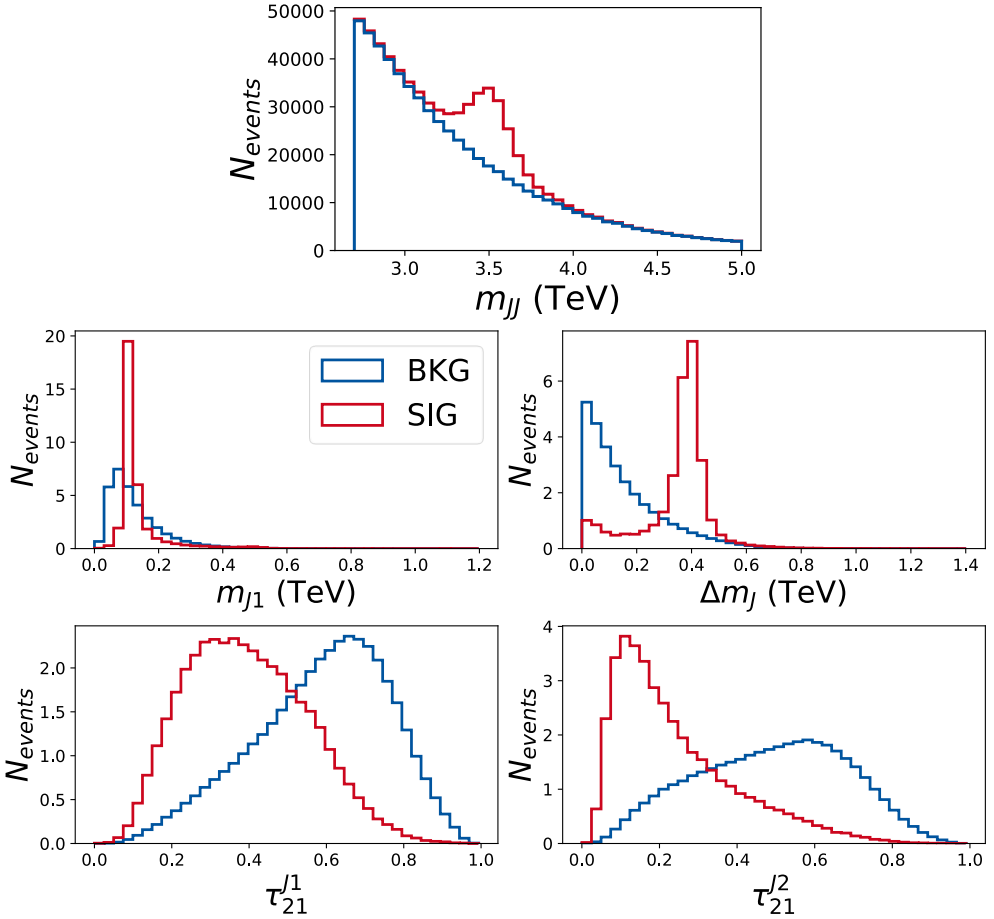


LHCO R&D dataset

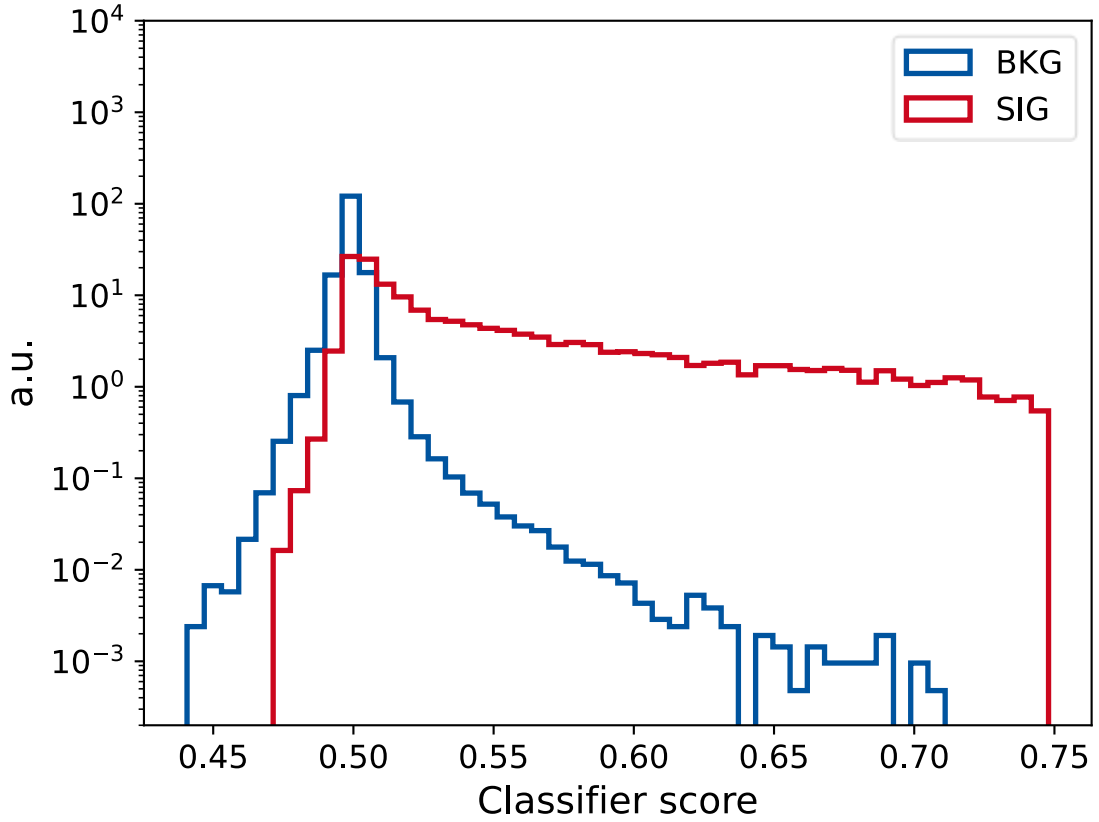
"The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics" [2101.08320], G. Kasieczka, B. Nachman, D. Shih et. al.

- Benchmark dataset for anomaly detection
- QCD dijet background (1M events)
- Signal (N_{sig} events)

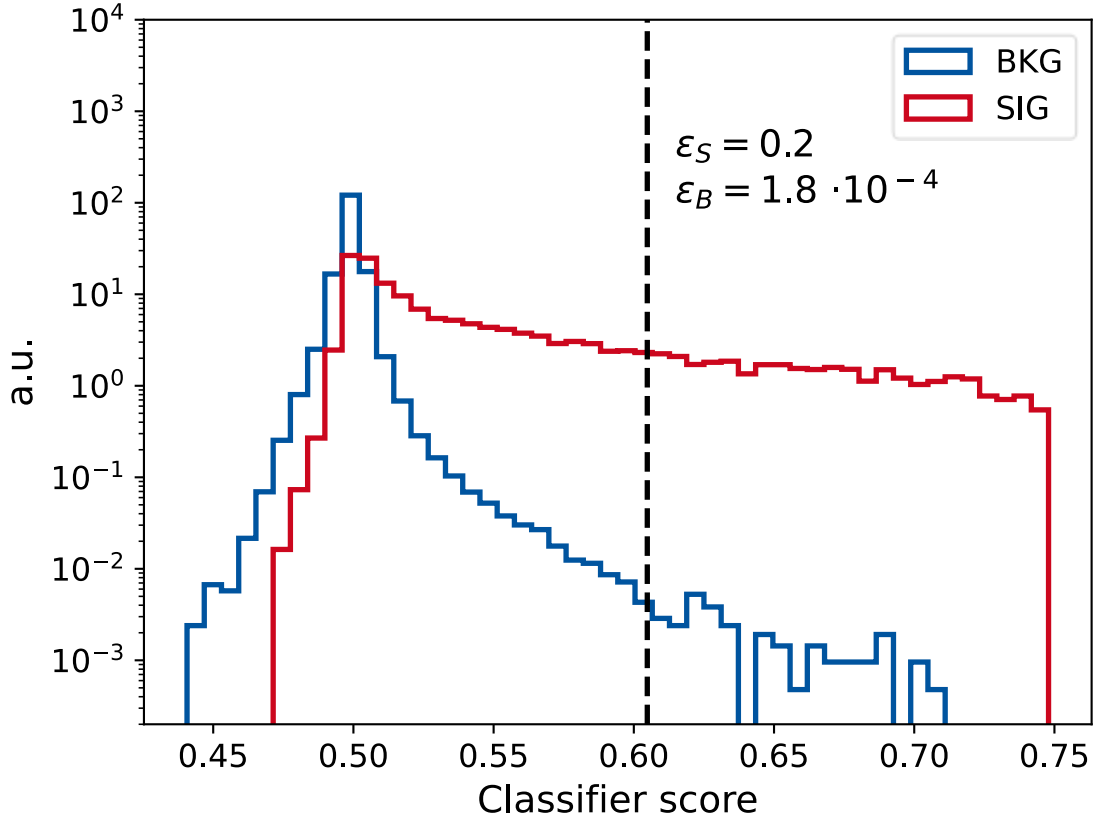


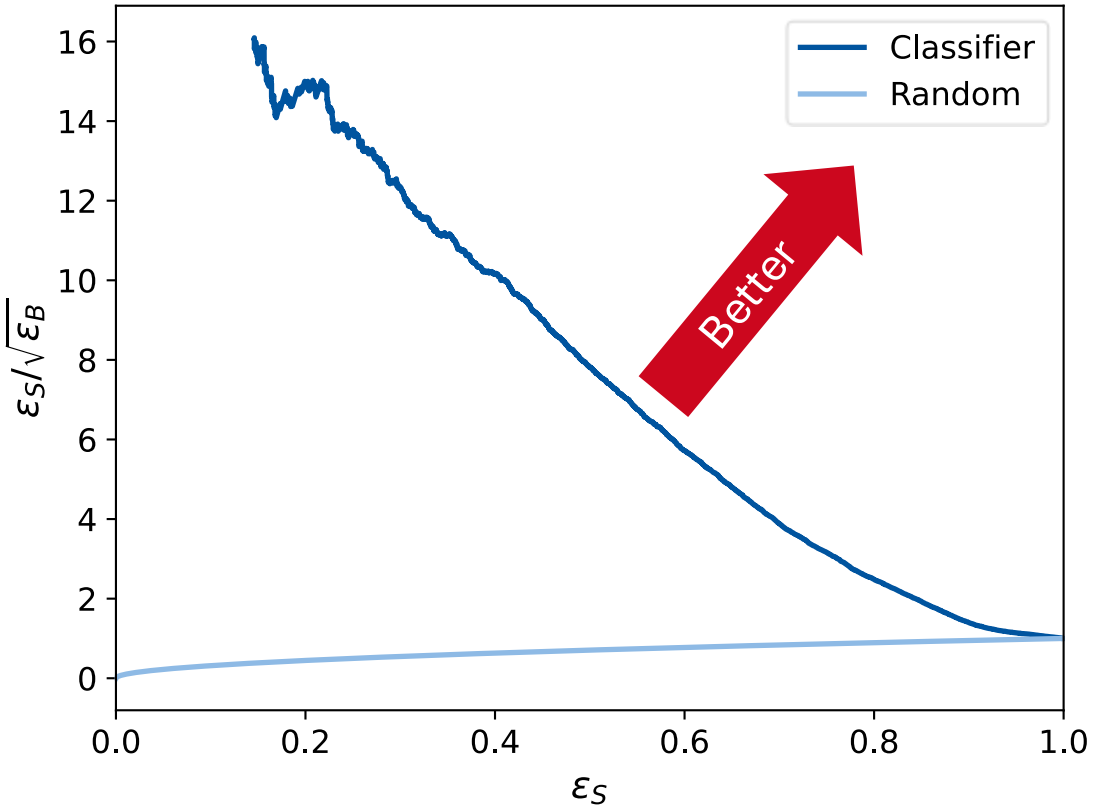


Supervised metric: Max SIC

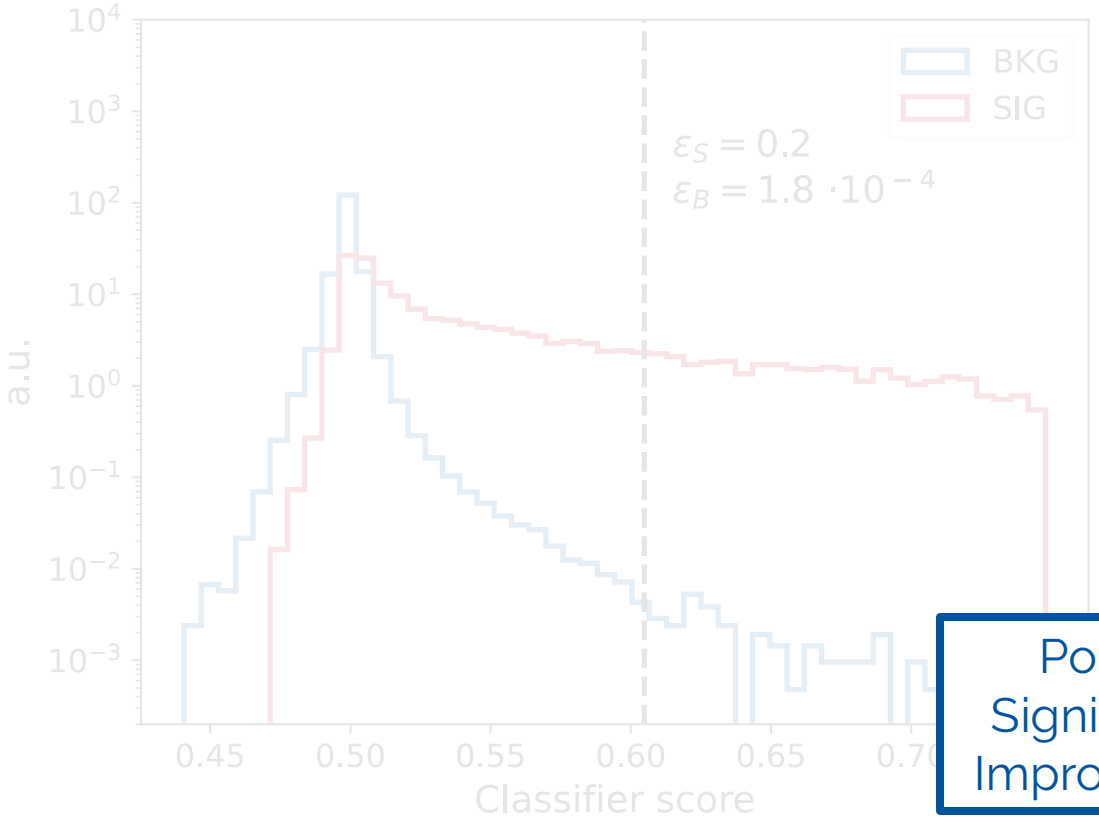


Supervised metric: Max SIC

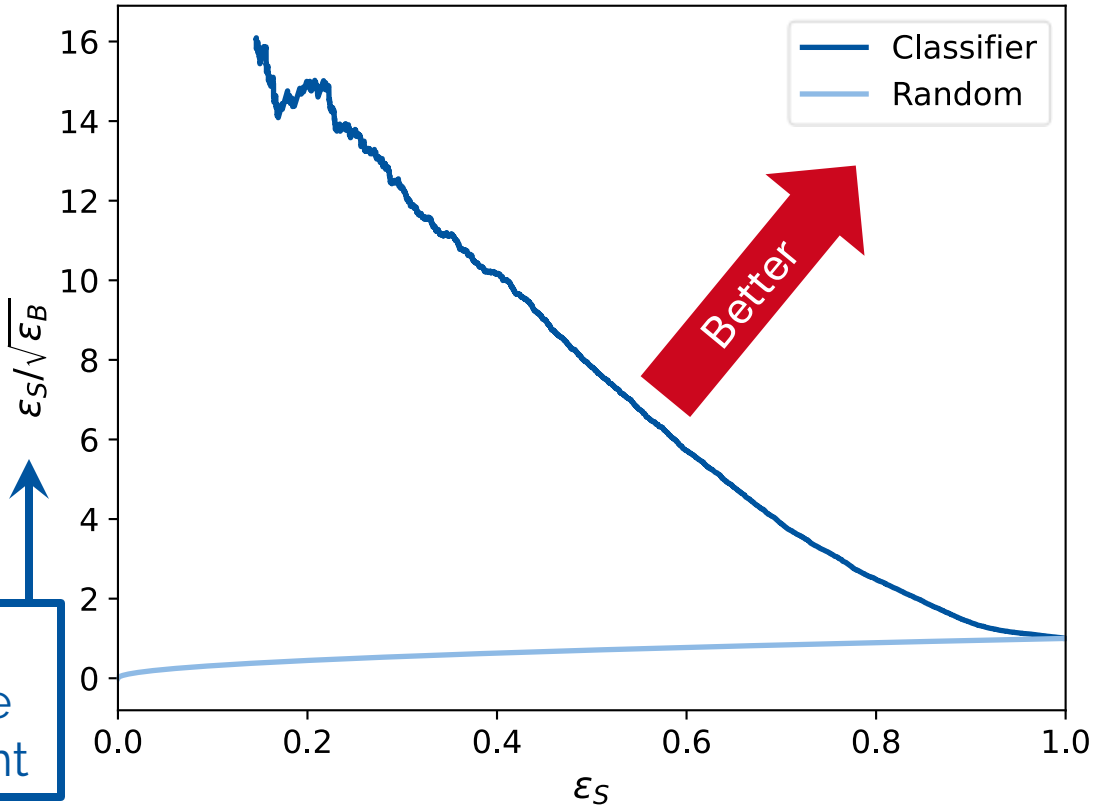




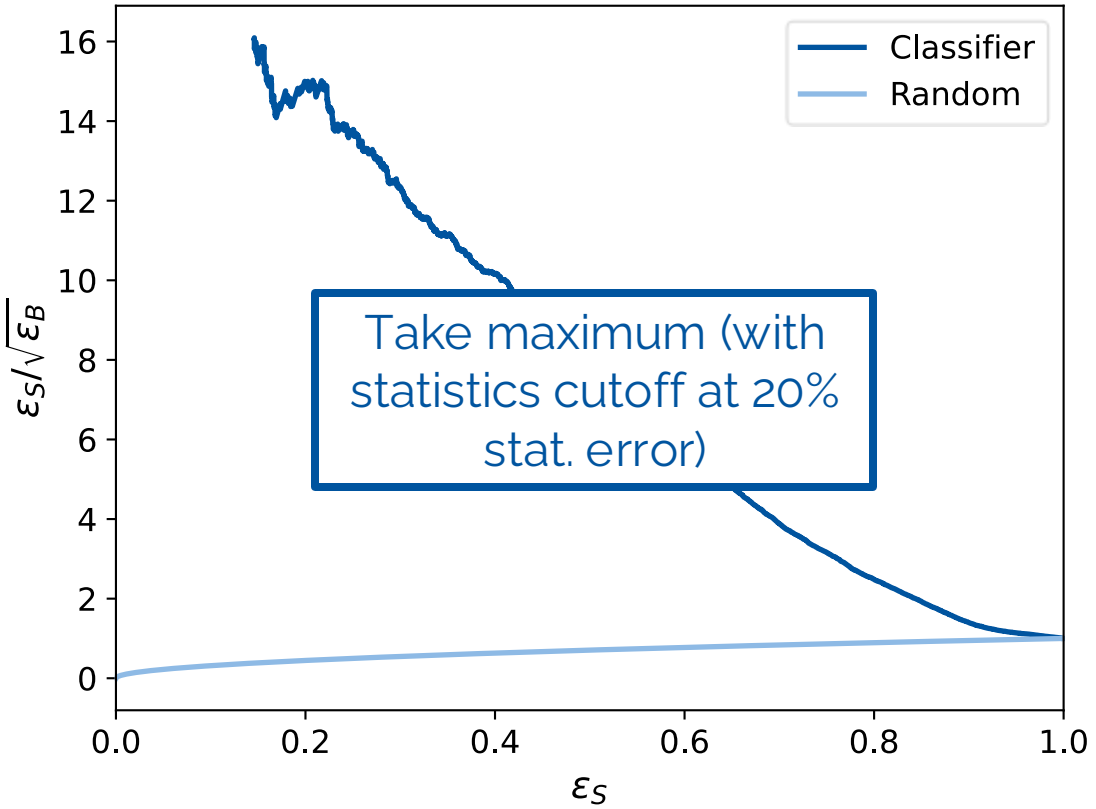
Supervised metric: Max SIC



Poisson Significance Improvement



Supervised metric: Max SIC



Data-driven metric: Val loss

$$BCE = -\log p_{\text{pred, true}}$$

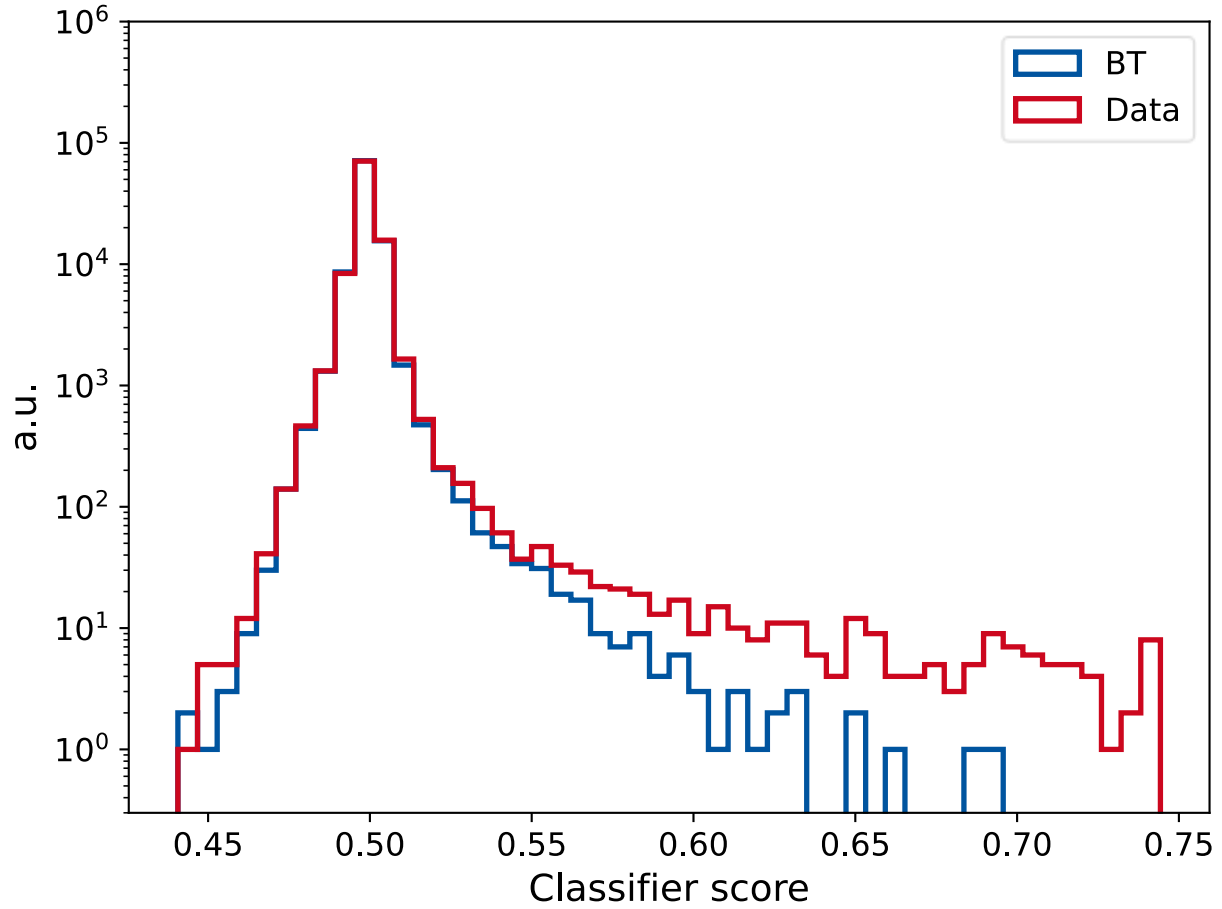
- Random scores: all events $p_{\text{pred}} = 0.5$

$$BCE_{\text{random}} = \ln 2 \approx 0.6931$$

- Optimal scores: $p_{\text{pred},B} = 0.5$, $p_{\text{pred},S} = 1$, e.g.

$$BCE_{\text{opt}} = \frac{10^5 - 1000}{10^5} \ln 2 + \frac{1000}{10^5} \ln 1 \approx 0.6924$$

→Dominated by background (noisy)



Data-driven metric: Val loss

$$BCE = -\log p_{\text{pred, true}}$$

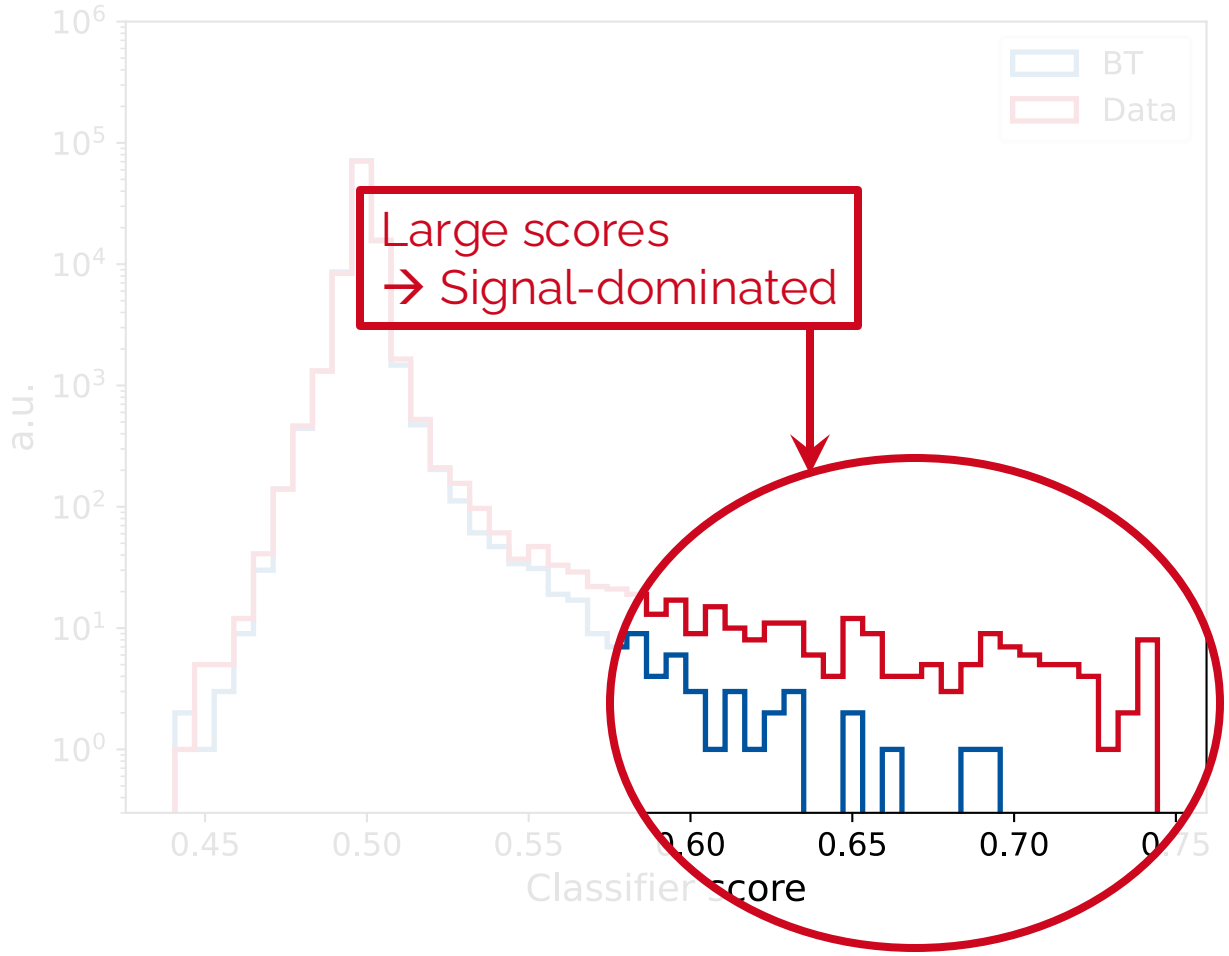
- Random scores: all events $p_{\text{pred}} = 0.5$

$$BCE_{\text{random}} = \ln 2 \approx 0.6931$$

- Optimal scores: $p_{\text{pred},B} = 0.5$, $p_{\text{pred},S} = 1$, e.g.

$$BCE_{\text{opt}} = \frac{10^5 - 1000}{10^5} \ln 2 + \frac{1000}{10^5} \ln 1 \approx 0.6924$$

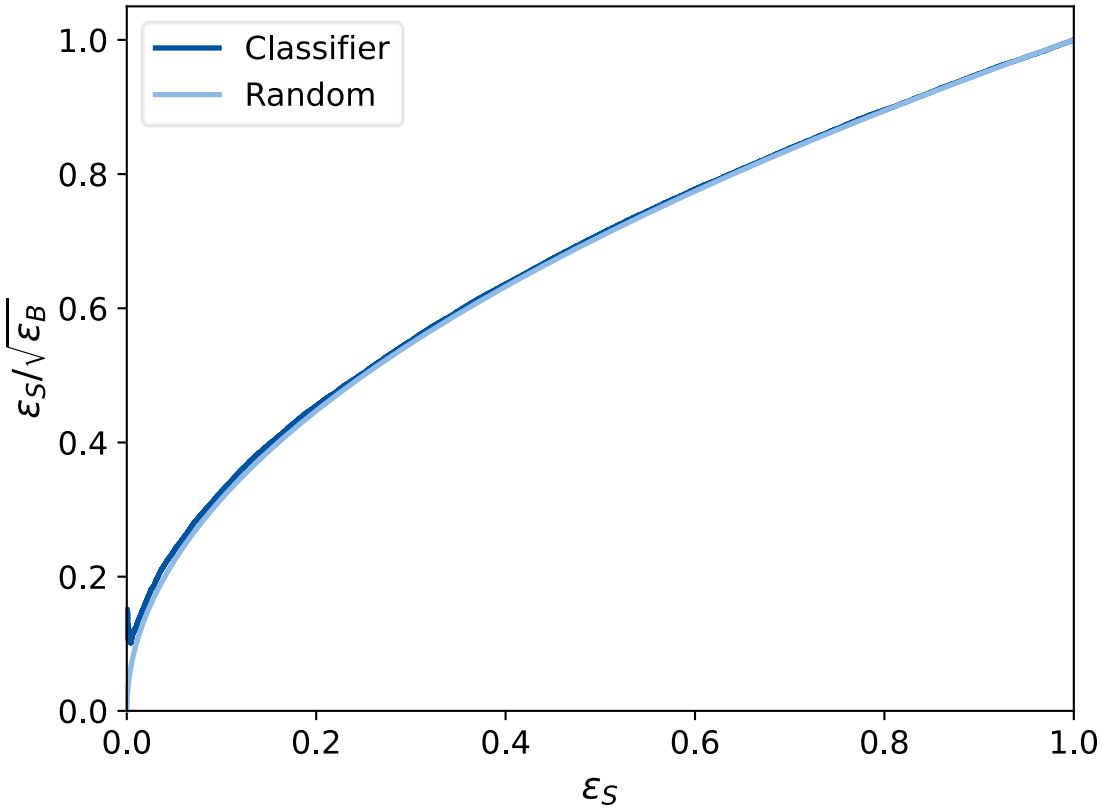
→Dominated by background (noisy)



Data-driven metric: Alternatives

- SIC curve calculated at all classifier scores

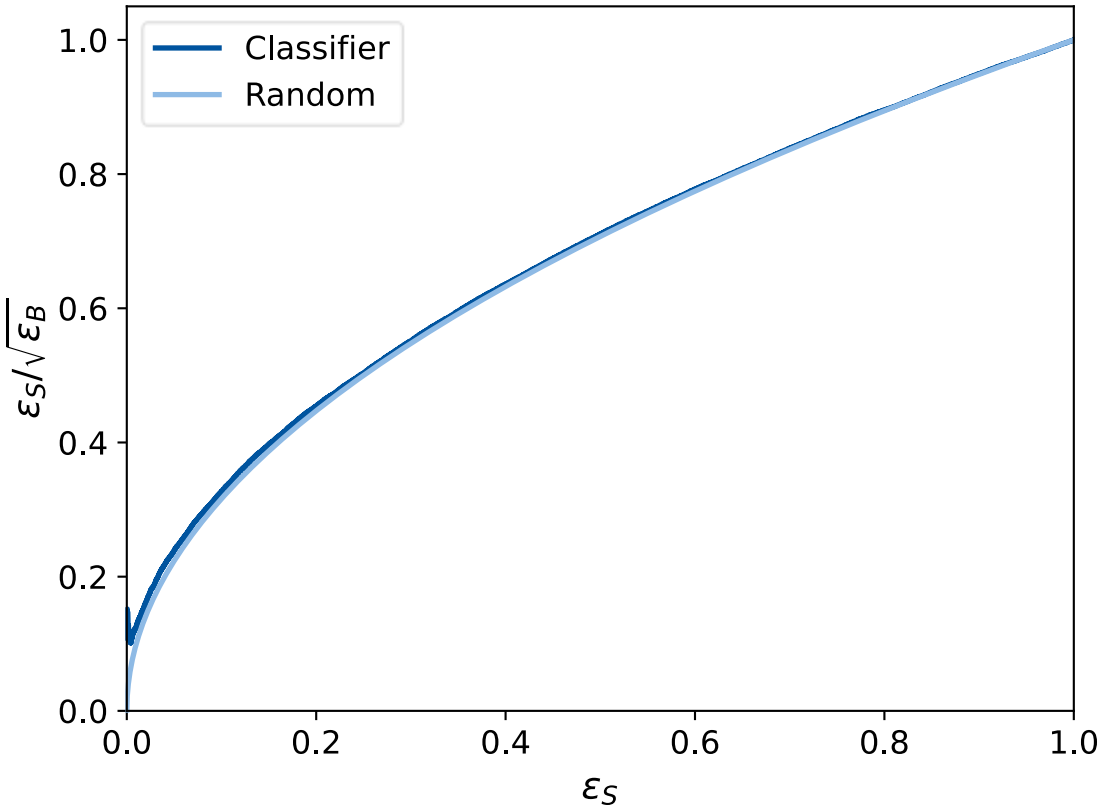
  →Can pick large threshold



Data-driven metric: Alternatives

- SIC curve calculated at all classifier scores

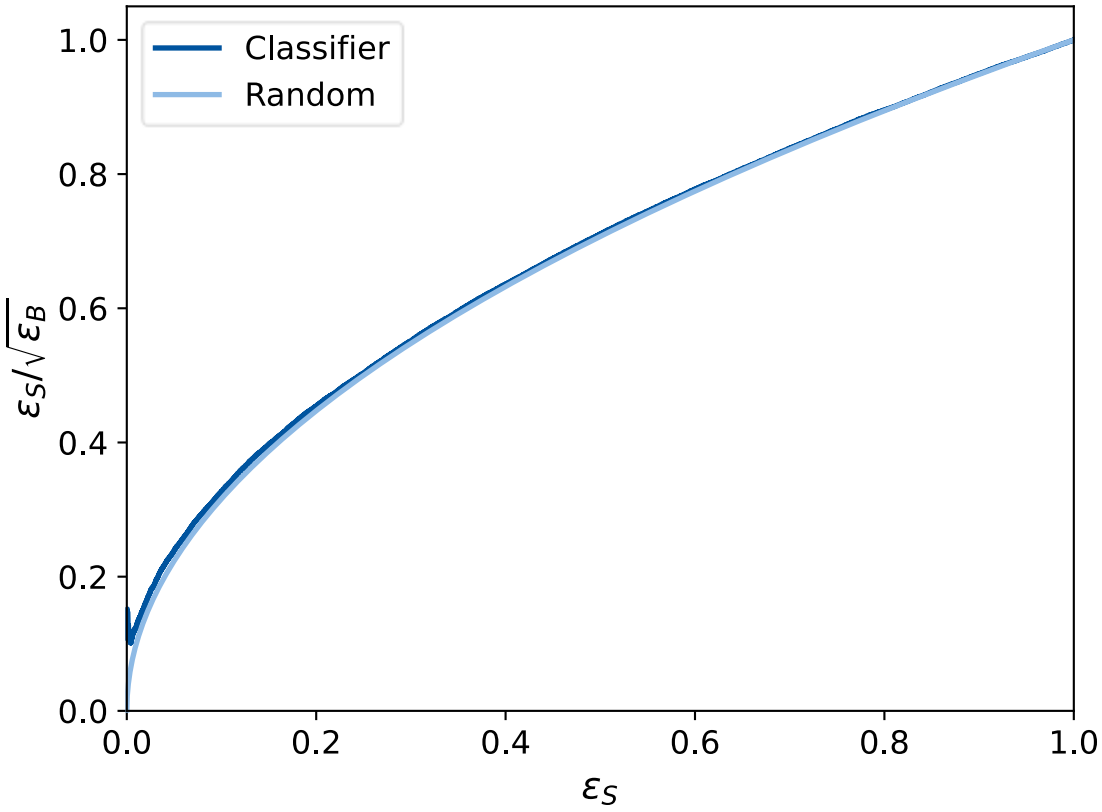
  →Can pick large threshold



Data-driven metric: Alternatives

- SIC curve calculated at all classifier scores

  →Can pick large threshold



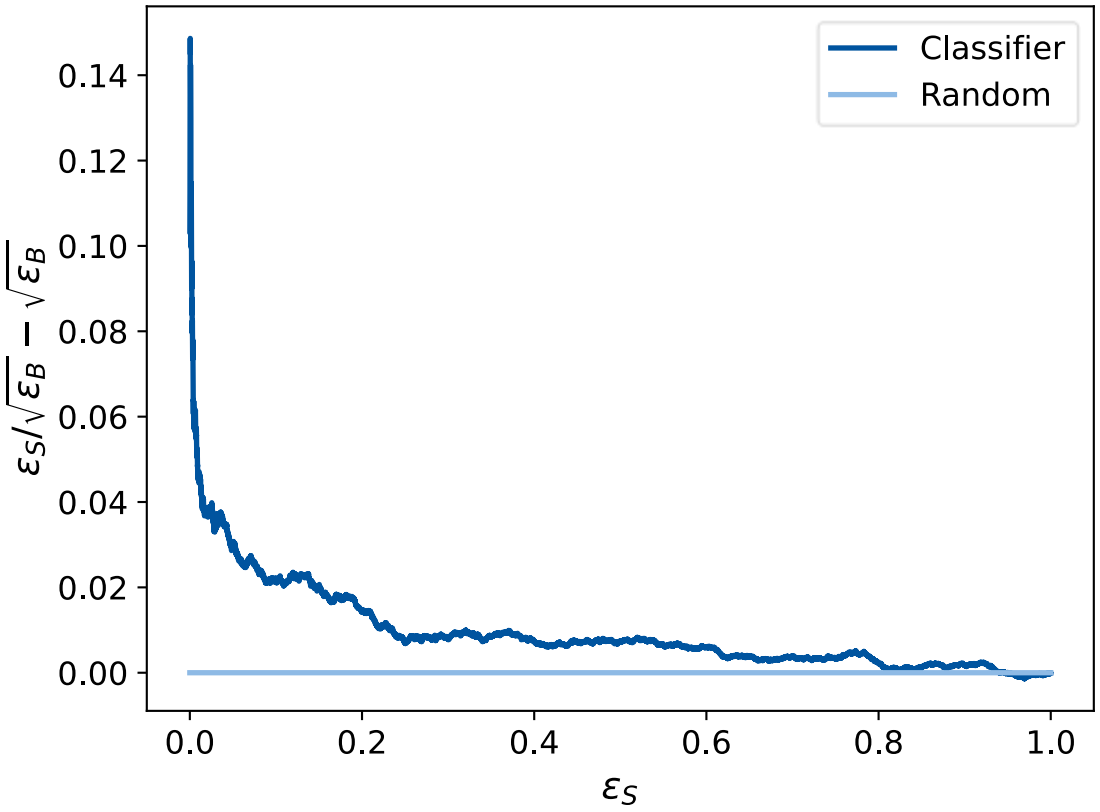
Data-driven metric: Alternatives

- SIC curve calculated at all classifier scores

  →Can pick large threshold

- Val SIC very close to random

  →Subtract random



Data-driven metric: Alternatives

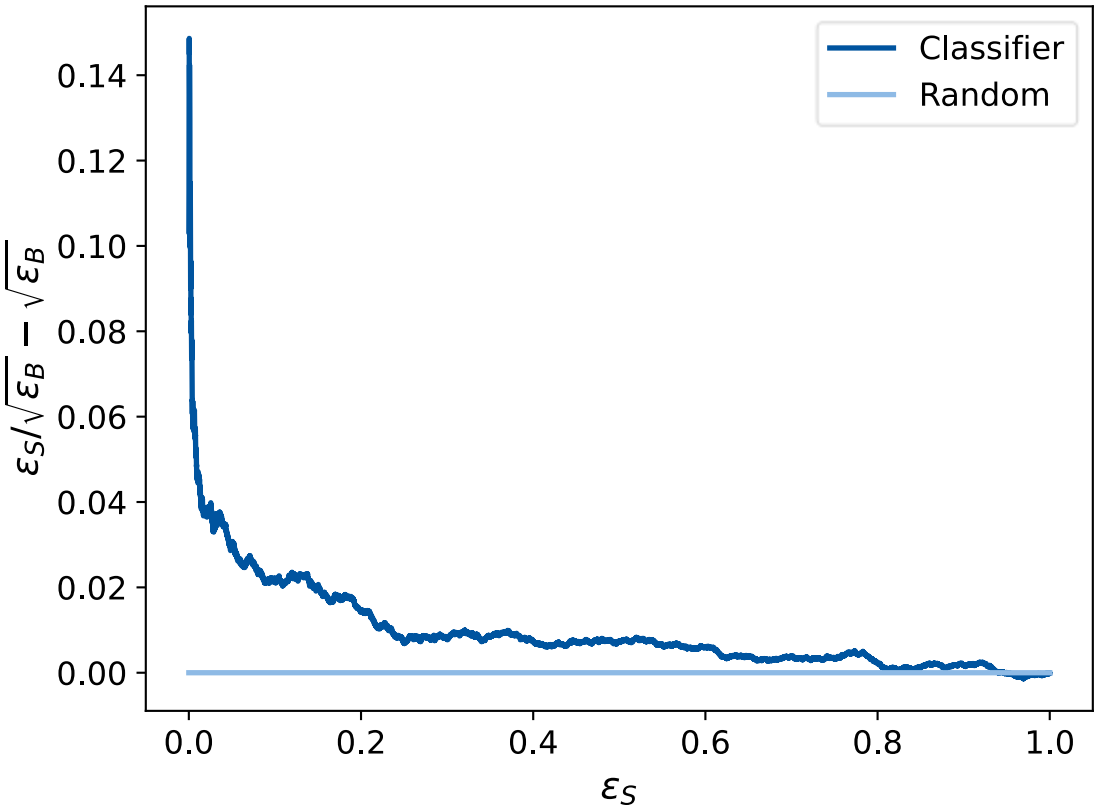
- SIC curve calculated at all classifier scores

  →Can pick large threshold

- Val SIC very close to random

  →Subtract random

- For most signal-like events, pick maximum



Data-driven metric: Alternatives

- SIC curve calculated at all classifier scores

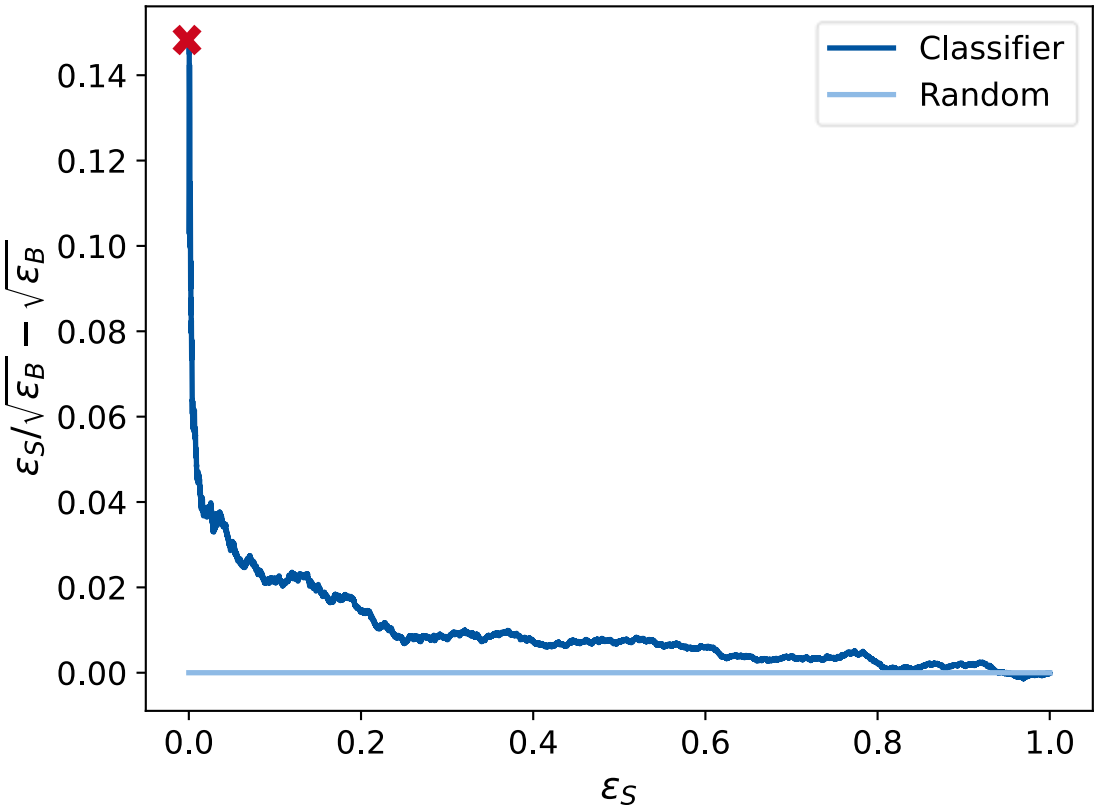
  →Can pick large threshold

- Val SIC very close to random

  →Subtract random

- For most signal-like events, pick maximum

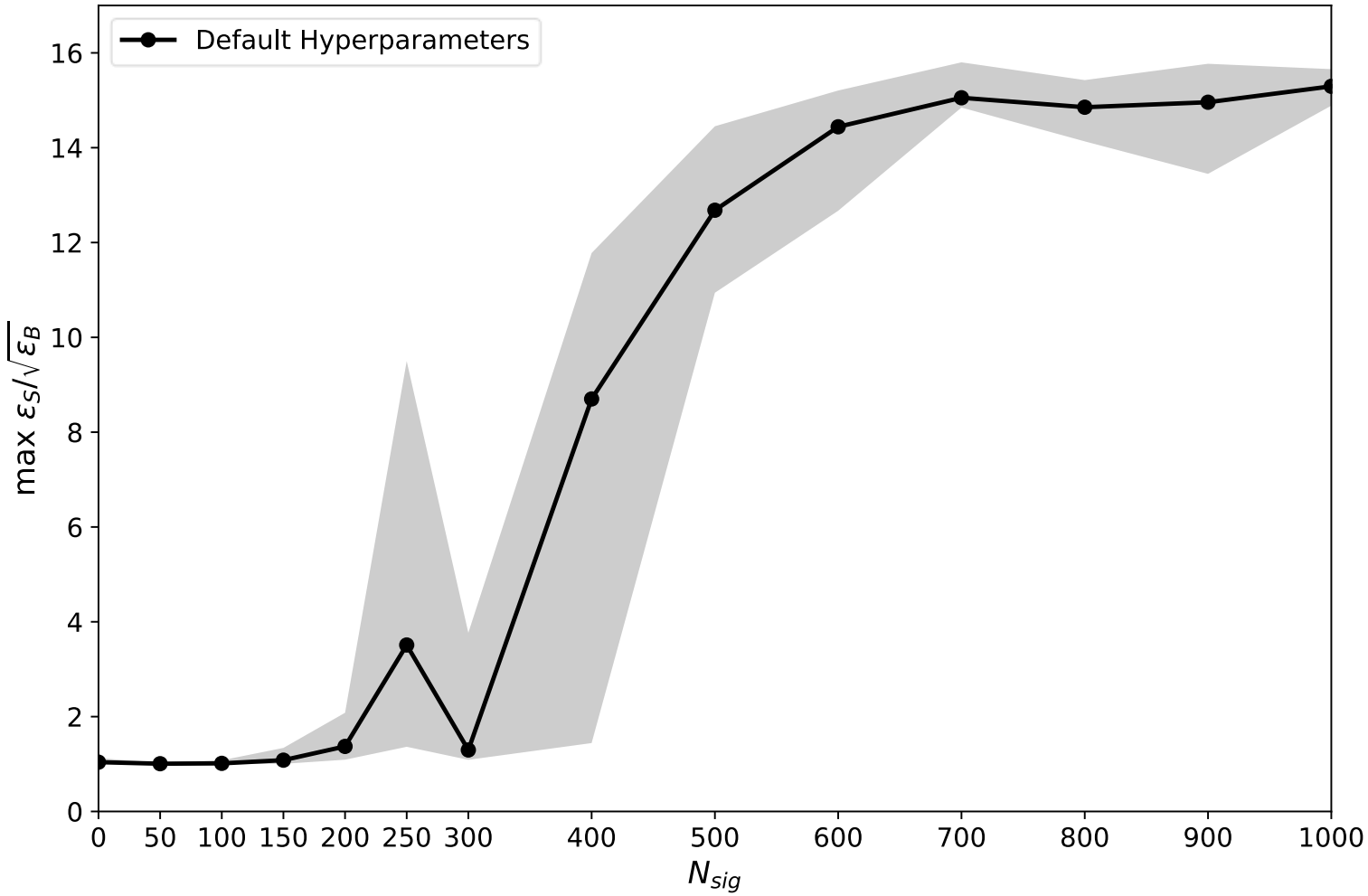
→Will refer to this as Val SIC



Optimizing Hyperparameters

Idea:

- Test HP configurations and pick best based on each metric



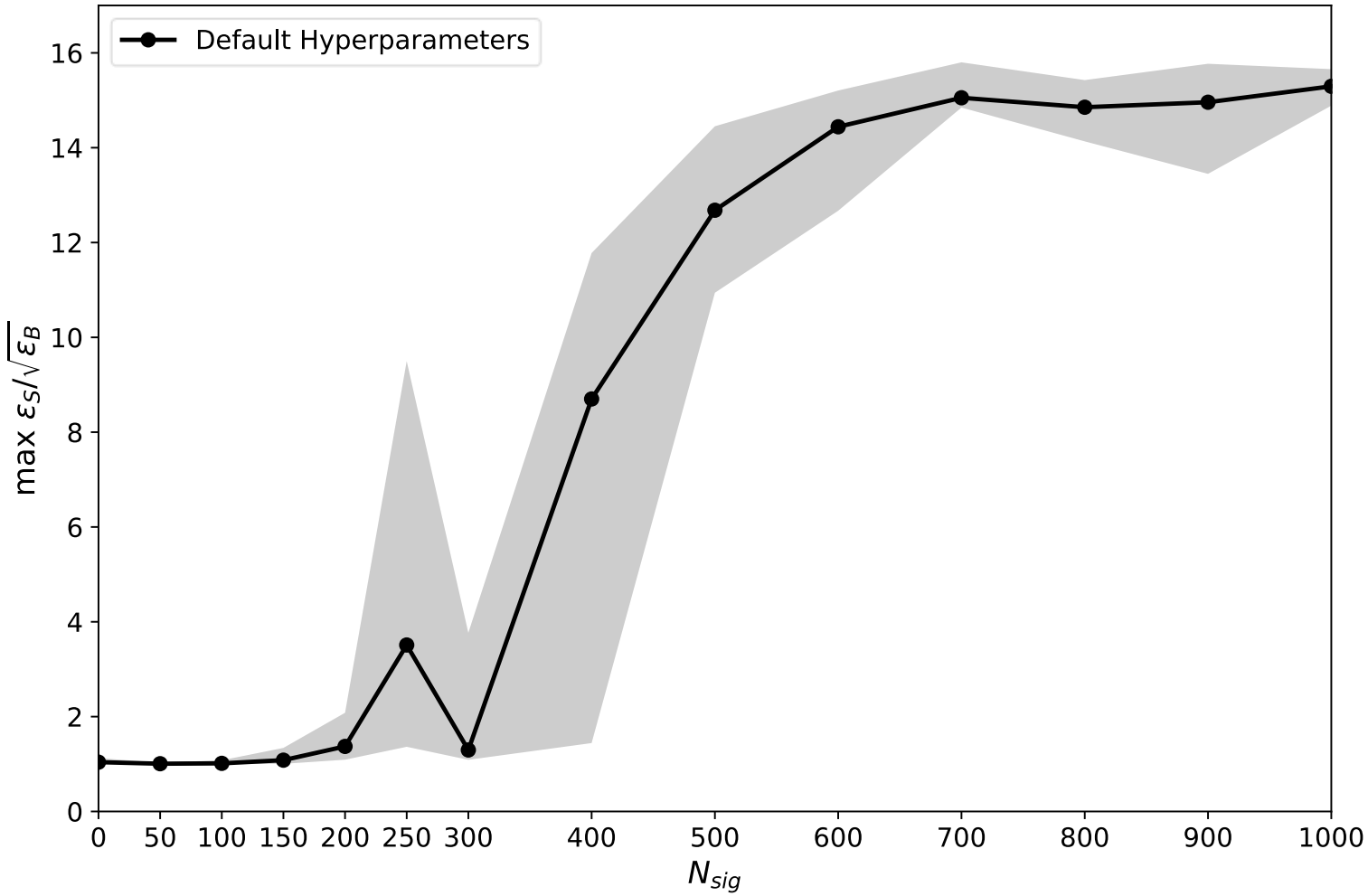
Optimizing Hyperparameters

Idea:

- Test HP configurations and pick best based on each metric

Results:

- Benchmark: Default HP optimized for this setup



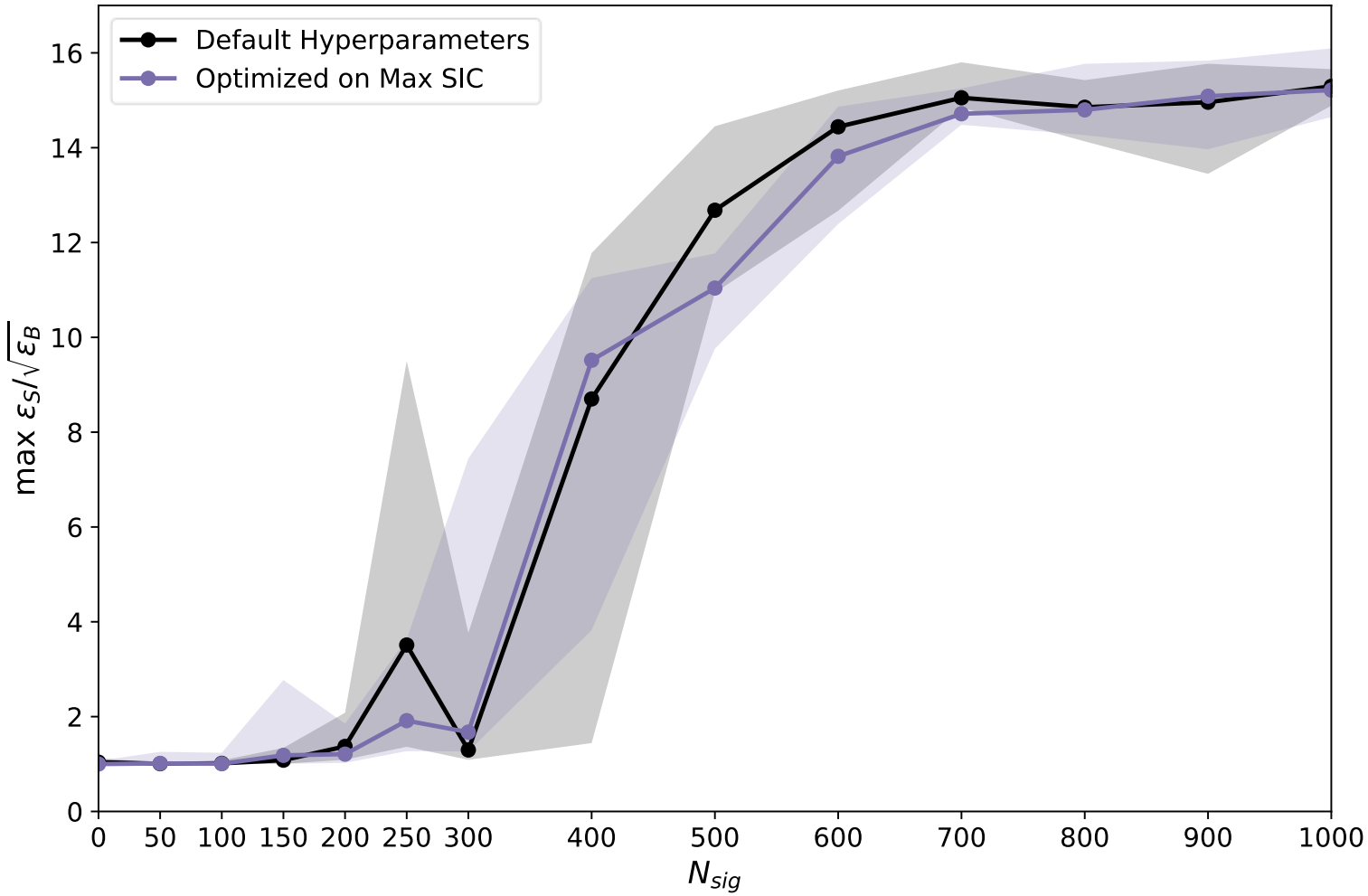
Optimizing Hyperparameters

Idea:

- Test HP configurations and pick best based on each metric

Results:

- Benchmark: Default HP optimized for this setup
- Max SIC: Performance comparable to benchmark



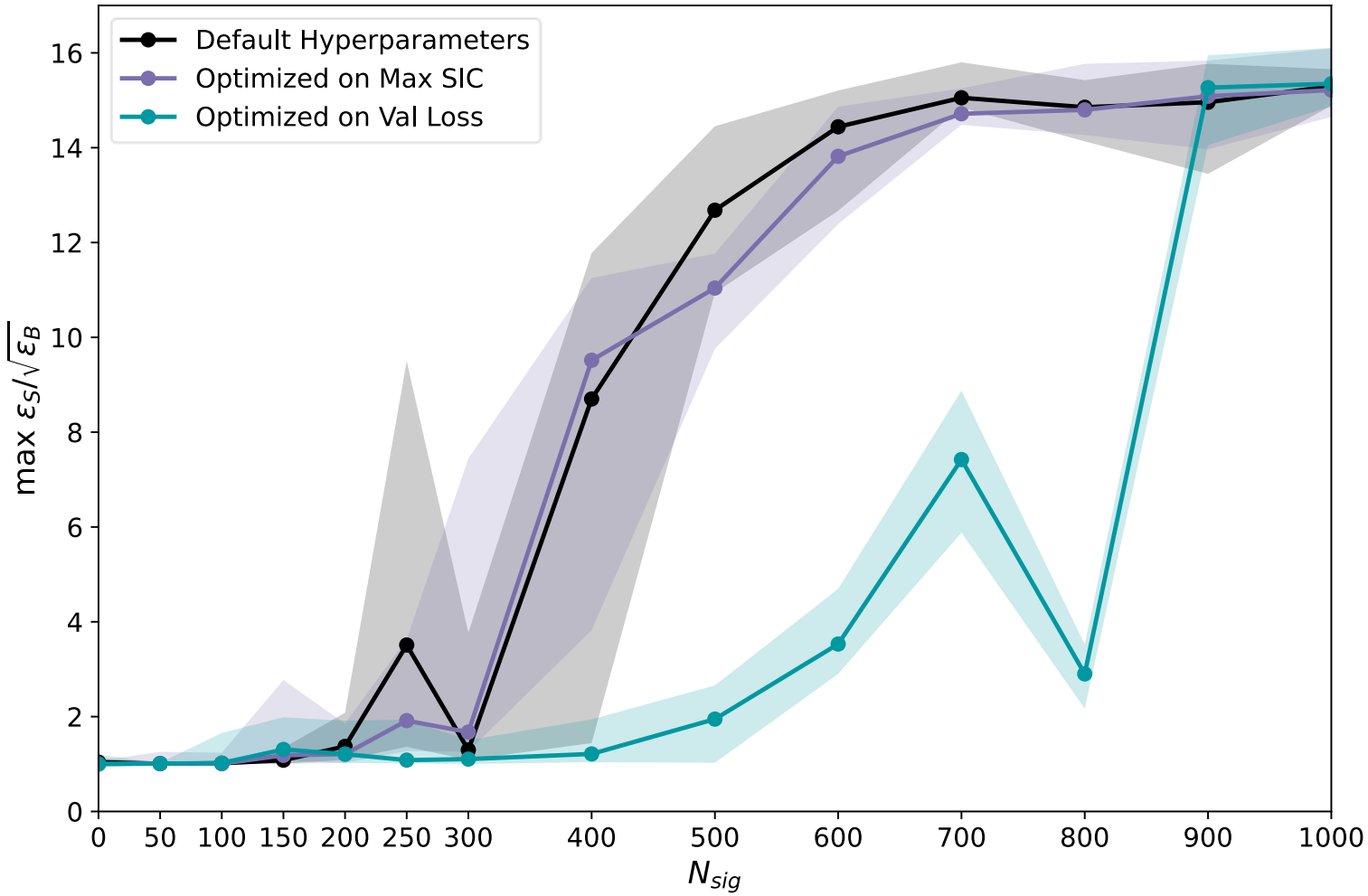
Optimizing Hyperparameters

Idea:

- Test HP configurations and pick best based on each metric

Results:

- Benchmark: Default HP optimized for this setup
- Max SIC: Performance comparable to benchmark
- Val loss: fails at low N_{sig}



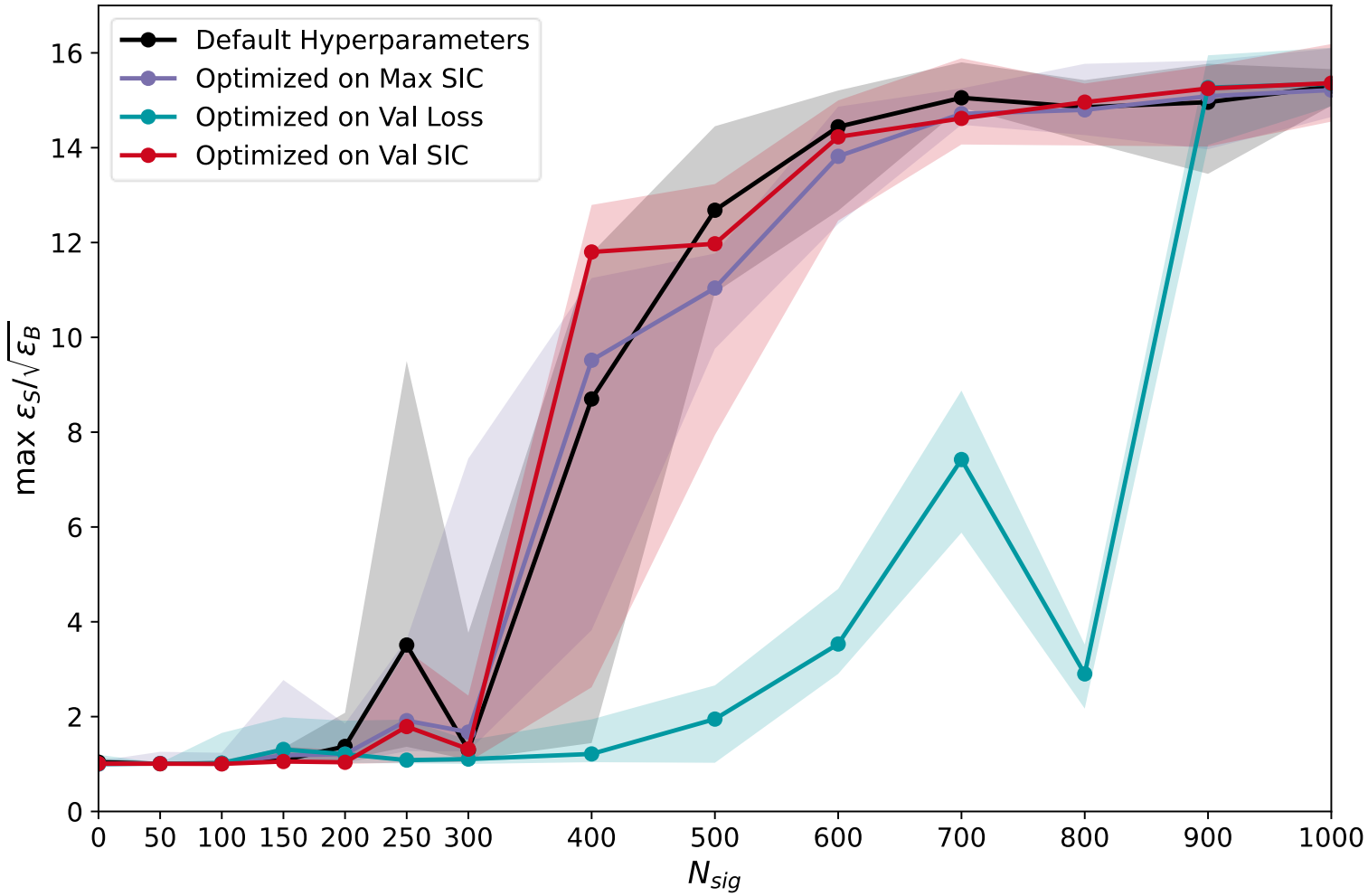
Optimizing Hyperparameters

Idea:

- Test HP configurations and pick best based on each metric

Results:

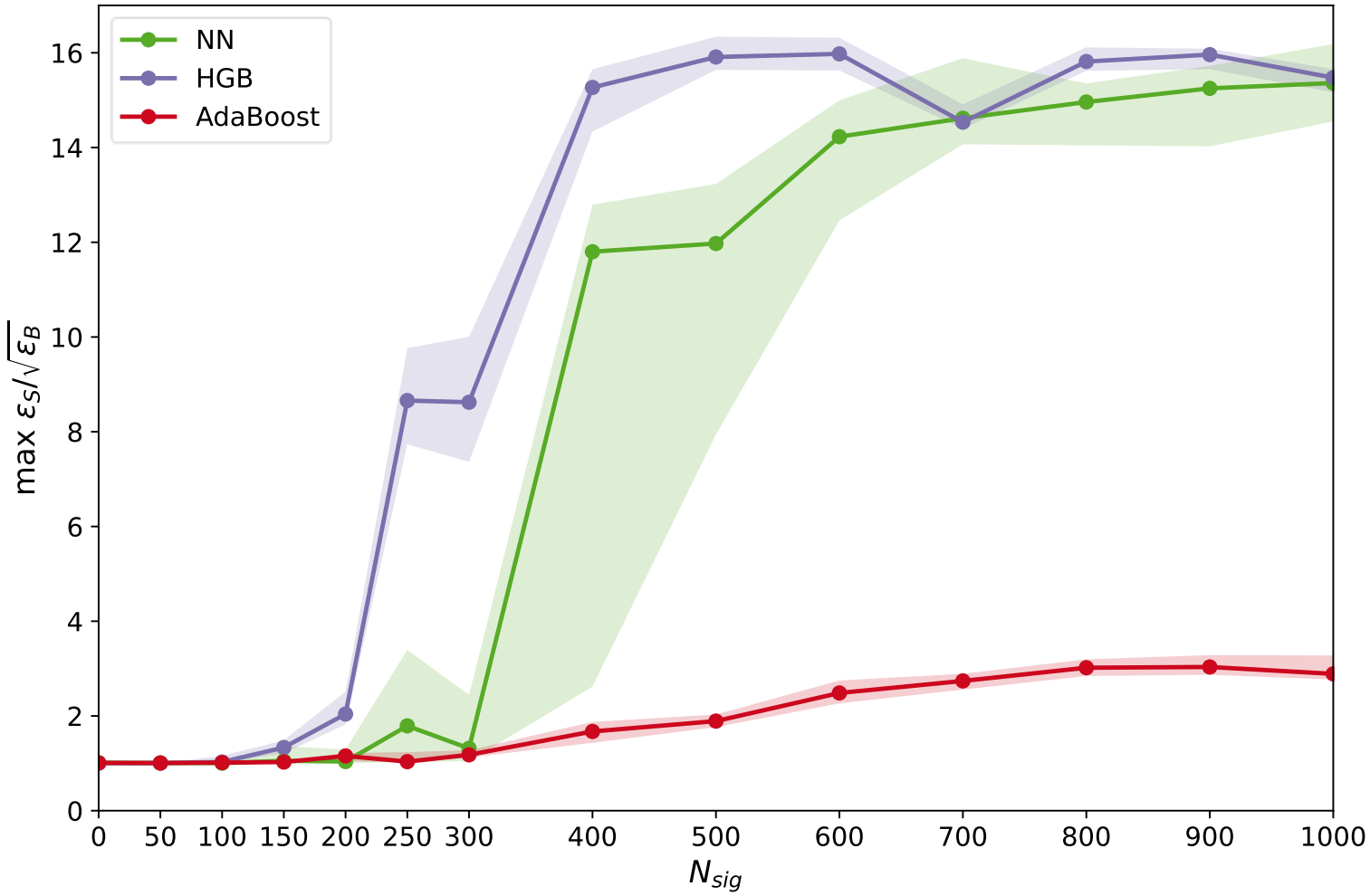
- Benchmark: Default HP optimized for this setup
- Max SIC: Performance comparable to benchmark
- Val loss: fails at low N_{sig}
- Val SIC: Performance close to max SIC & benchmark



Picking the setup

Procedure:

- Pick several architectures
- Optimize HP for each
- Train each setup on ½ data, evaluate metrics on ½ data
- Pick best setup based on metric



Picking the setup

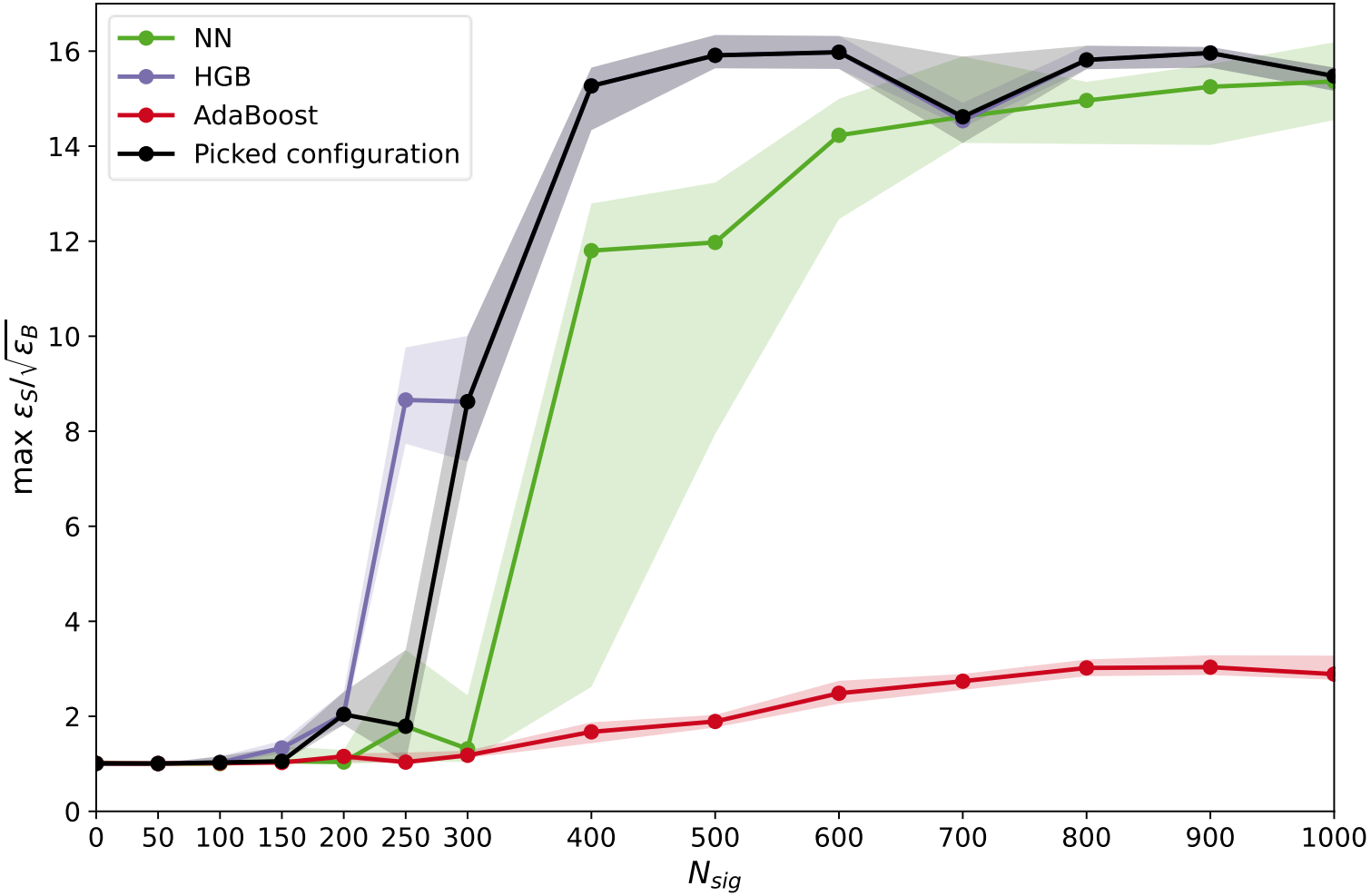
Procedure:

- Pick several architectures
- Optimize HP for each
- Train each setup on ½ data, evaluate metrics on ½ data
- Pick best setup based on metric

Result:

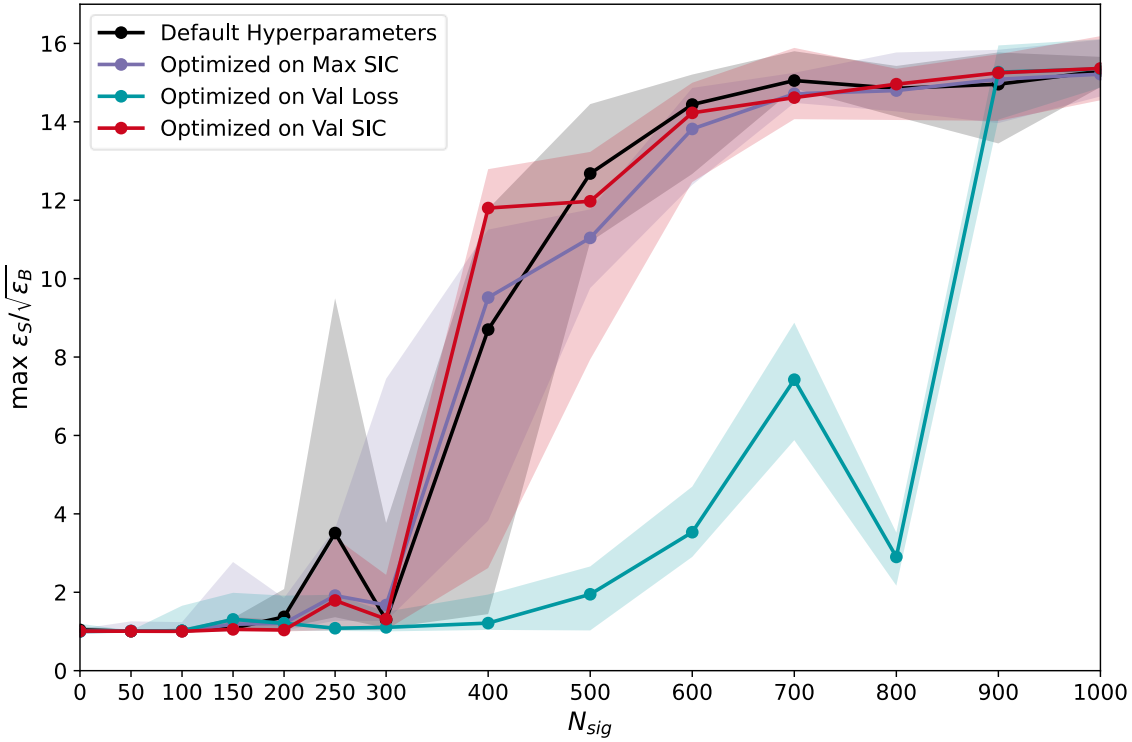
Val SIC picks best model almost everywhere

→Fails at very low signal injections (expected)



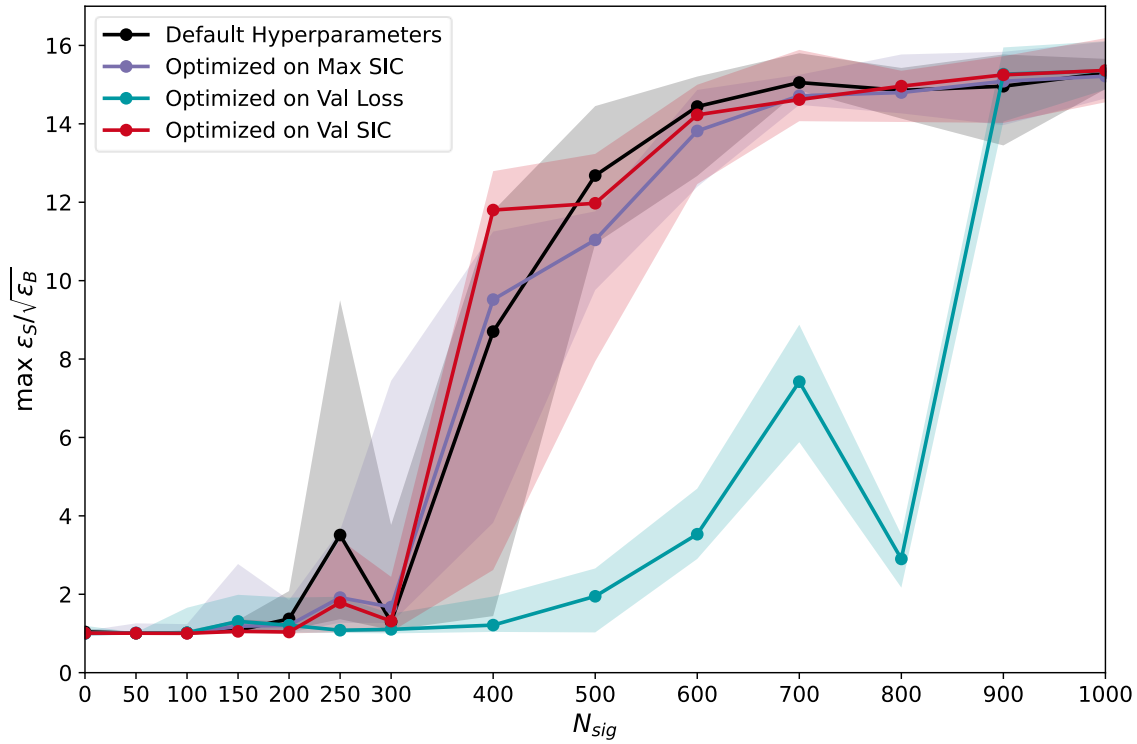
Conclusion

- Investigated two data-driven metrics for model agnostic setup optimization



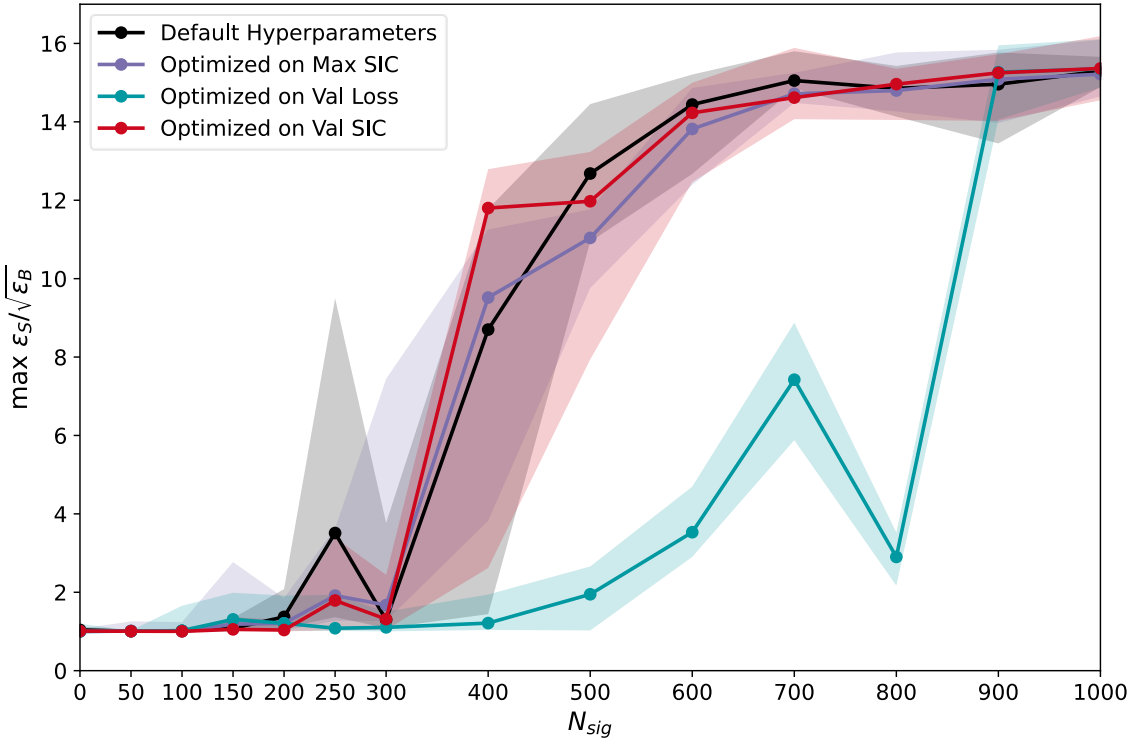
Conclusion

- Investigated two data-driven metrics for model agnostic setup optimization
  1. Val loss: too sensitive to background distribution



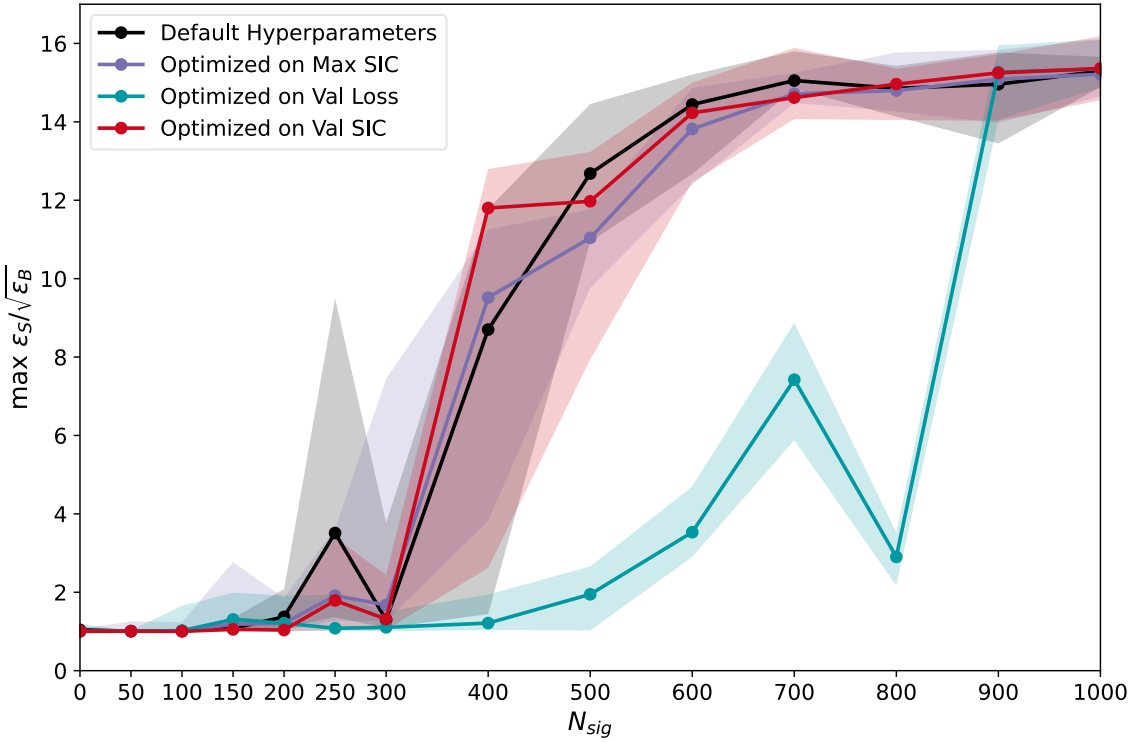
Conclusion

- Investigated two data-driven metrics for model agnostic setup optimization
  1. Val loss: too sensitive to background distribution
  2. Val SIC: highly correlated with max SIC by focusing on most signal-like events



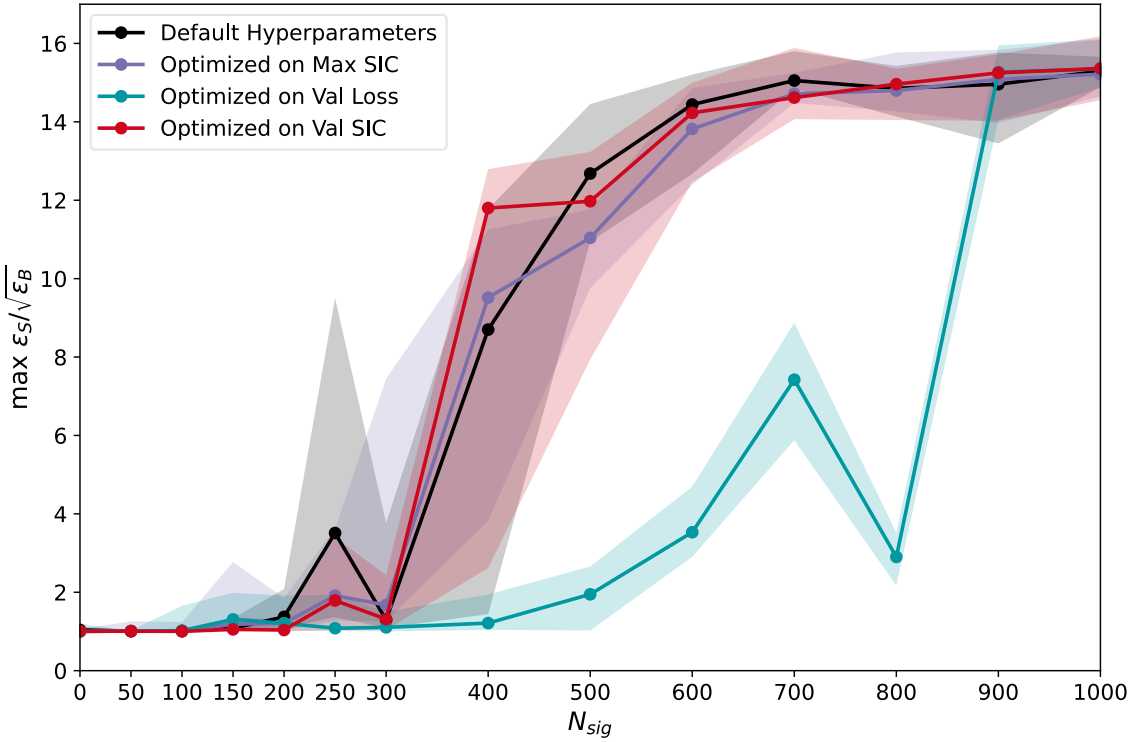
Conclusion

- Investigated two data-driven metrics for model agnostic setup optimization
  1. Val loss: too sensitive to background distribution
  2. Val SIC: highly correlated with max SIC by focusing on most signal-like events
- Optimizing on val SIC leads to excellent anomaly detection performance



Conclusion

- Investigated two data-driven metrics for model agnostic setup optimization
  1. Val loss: too sensitive to background distribution
  2. Val SIC: highly correlated with max SIC by focusing on most signal-like events
- Optimizing on val SIC leads to excellent anomaly detection performance
- Behind the scenes: Seen comparable results for CWoLa Hunting and CATHODE



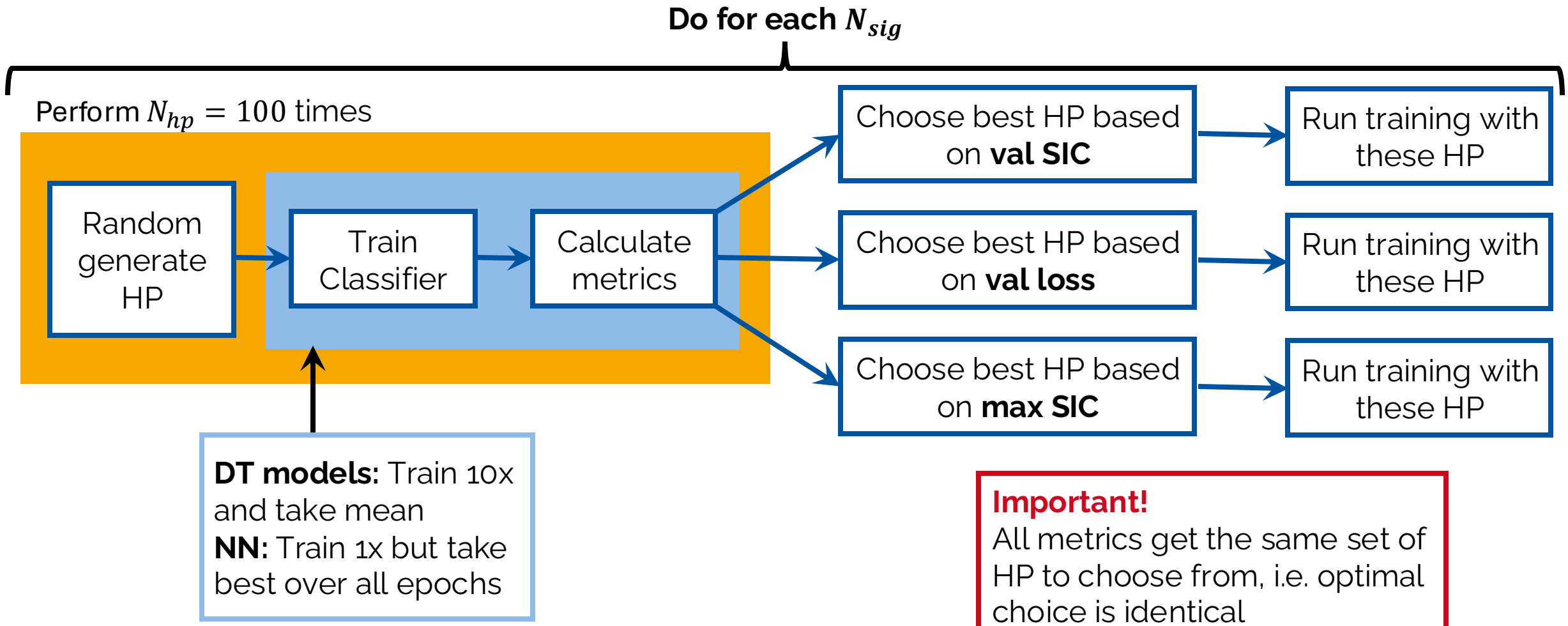
Backup

Optimized hyperparameters

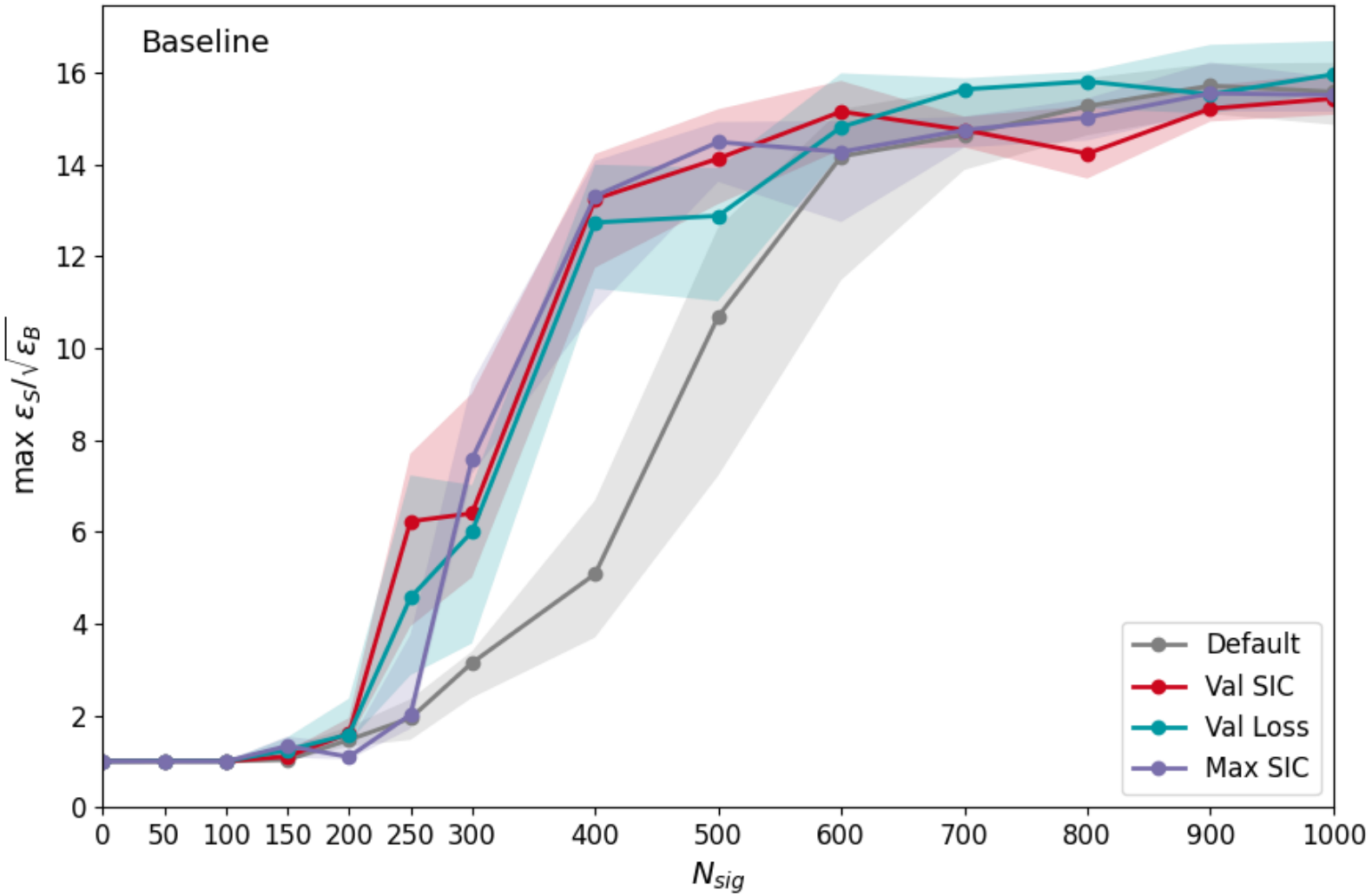
Institute for Theoretical Particle Physics and Cosmology | RWTH AACHEN UNIVERSITY

```
params = {}
params["lr"] = float(np.random.choice([0.01, 0.005, 0.001, 0.0005, 0.0001]))
params["batch_size"] = int(np.random.choice([64, 128, 256, 512, 1024, 2048, 5096]))
params["layers"] = [64,64,64]#layers
params["epochs"] = int(30)
params["dropout"] = float(np.random.choice([0, 0.1, 0.2, 0.3, 0.4, 0.5]))
params["weight_decay"] = float(np.random.choice([0,1e-4, 1e-3, 1e-2, 1e-5]))
params["momentum"] = float(np.random.choice([0.9, 0.99, 0.8, 0.95]))
```

HP optimization: General setting



HP optimization: HGB Baseline



- Different metrics show very similar performance
- While performance at $N_{sig} = 1000$ is optimal with default hyperparameters, optimizing for lower signal injections can result in slight performance gain
- Trend generally holds up for extended sets but performance gain decreases

Optimization as a multi-step process

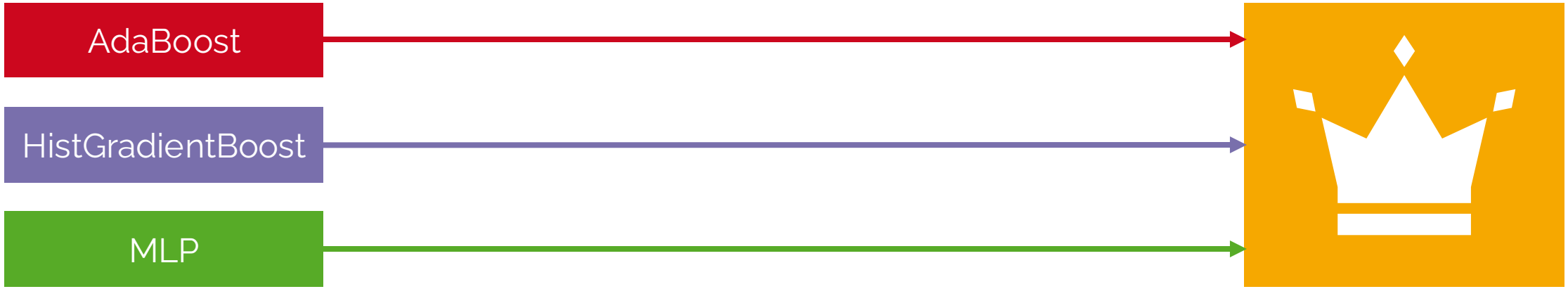
Think up general setup options:
- Classifier choice
- Ensemble strategy

Optimize HP for each option

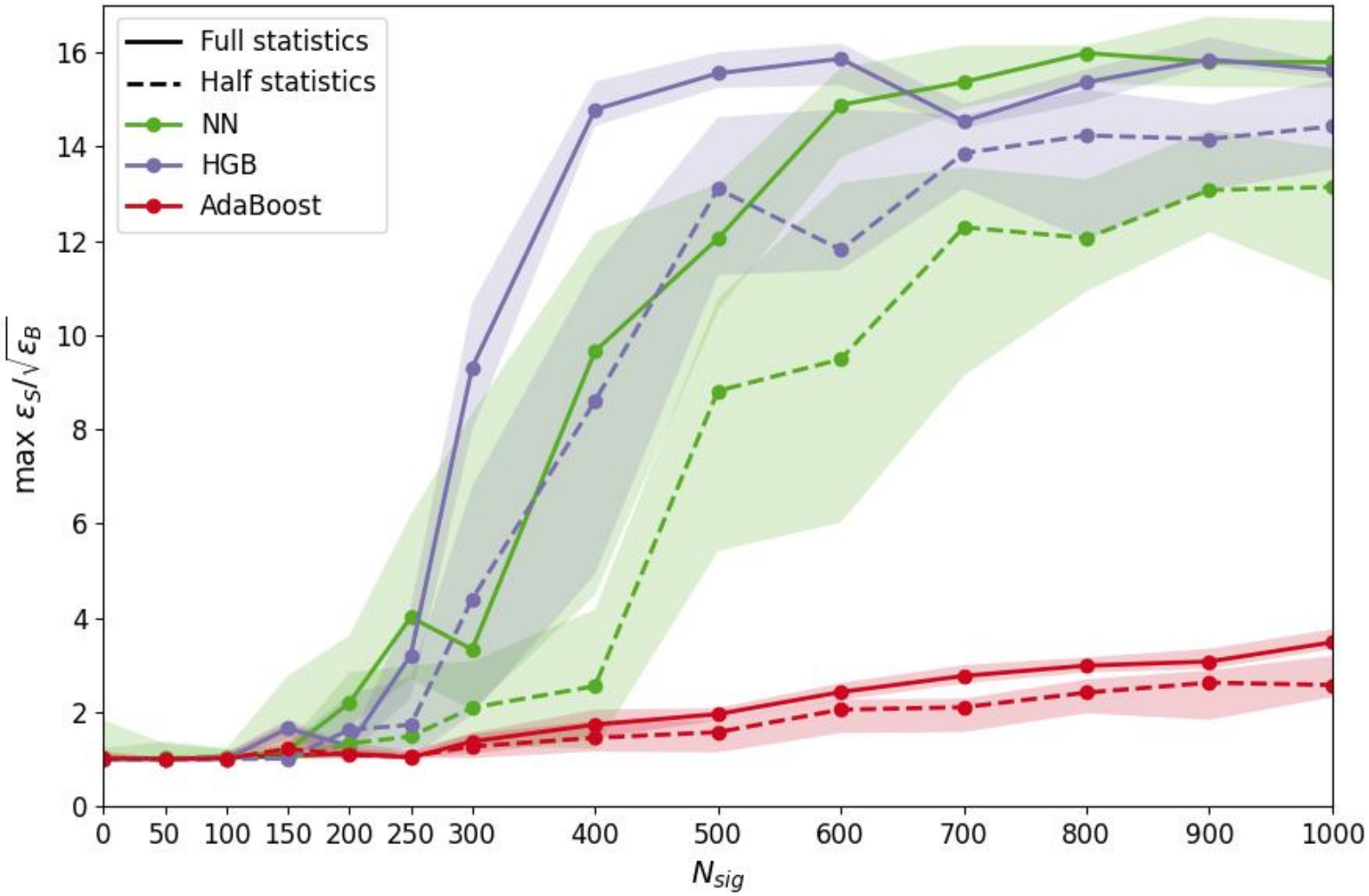
Train optimized setup on ½ data, calculate metric on ½ data

Pick final setup based on metric

Retrain setup on full data

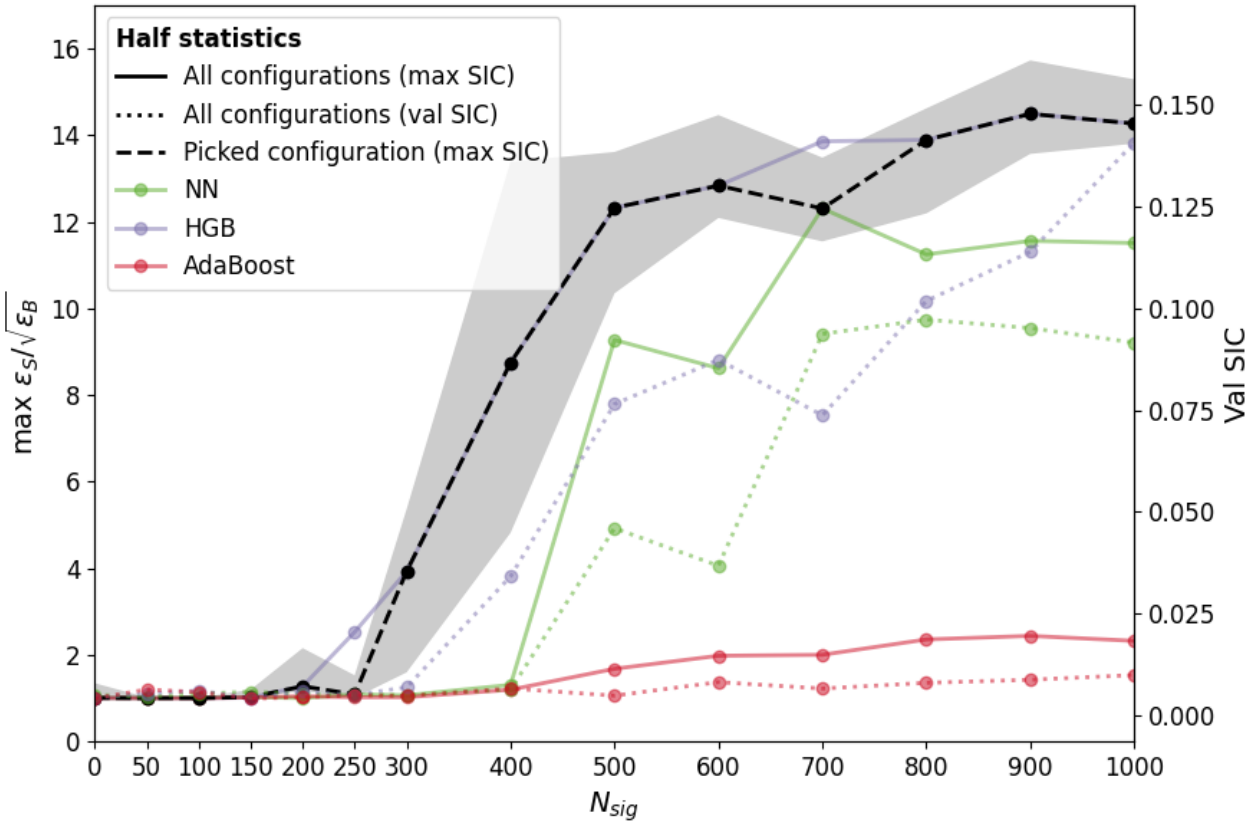


Performance on half statistics



Picking option based on val SIC

What we pick



Final performance