

# Increasing the Model Agnosticity of Weakly Supervised Anomaly Detection

Thorben Finke<sup>1</sup> Joep Geuskens<sup>1</sup> Marie Hein<sup>1</sup> Gregor Kasieczka<sup>2,3</sup> Michael Krämer<sup>1</sup> Alexander Mück<sup>1</sup> Parada Prangchaikul<sup>2</sup> Tobias Quadfasel<sup>2</sup> David Shih<sup>4</sup> Manuel Sommerhalder<sup>2</sup>

<sup>1</sup>TTK RWTH Aachen University <sup>2</sup>IEP Universität Hamburg <sup>3</sup>CDCS Universität Hamburg <sup>4</sup>NHETC Rutgers University

## Increasing the Model Agnosticity of Weakly Supervised Anomaly Detection

Thorben Finke<sup>1</sup> Joep Geuskens<sup>1</sup> Marie Hein<sup>1</sup> Gregor Kasieczka<sup>2,3</sup> Michael Krämer<sup>1</sup> Alexander Mück<sup>1</sup> Parada Prangchaikul<sup>2</sup> Tobias Quadfasel<sup>2</sup> David Shih<sup>4</sup> Manuel Sommerhalder<sup>2</sup>

**Motivation**

- To find new physics, improve largely model agnostic searches, e.g., resonance searches
- Use pattern recognition capability of machine learning in high dimensional feature space to gain higher sensitivity
- Problem: Currently many papers use only 4 high level features ("baseline" feature set) on one benchmark dataset (LHCO R&D dataset [1])
- For more model agnostic setup need to be able to use more features

Goal: Improve classifier setup for more high level features and low level features

### Weakly supervised anomaly detection

**Classification Without Labels (CWoLa) [2]**

- Classifier between mixed datasets  $p_S(x) = f_1 p_S(x) + (1 - f_1) p_B(x)$  with signal fractions  $f_i$
- $R_{\text{mixed}} = \frac{f_1 R_{\text{signal}}(x) + (1 - f_1) R_{\text{background}}(x)}{f_1 R_{\text{signal}}(x) + (1 - f_1) R_{\text{background}}(x)}$  where  $R_{\text{signal}}(x) = \frac{p_S(x)}{p_B(x)}$
- is the optimal classifier between signal and background distributions  $p_{S/B}$
- Mathematically equivalent as  $R_{\text{mixed}}$  monotonous in  $R_{\text{signal}}$

**Application to resonance searches**

- Divide data into signal region (SR) and sideband (SB), where  $p_S(x) = p_S(x|m \in \text{SR}) + p_B(x|m \in \text{SB})$  and  $p_B(x) = p_B(x|m \in \text{SB})$
- for classification features  $x$ .
- Construct "background template" from SB, ideally with  $p(x) = p_B(x|m \in \text{SB})$
- Here, we use idealized case to study classifier only

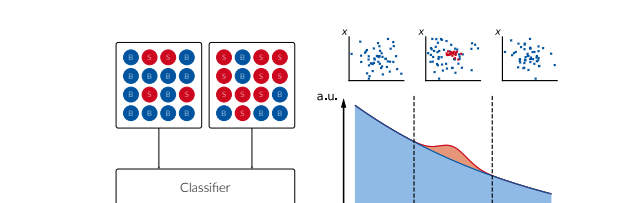


Figure 1: Left: Sketch of weakly supervised classification setup. Right: Division of data into SB and SR for a resonance search.

### BDTs for high level features [3]

**Machine Learning background**

Boosted Decision Trees (BDTs) are known to be very effective on tabular data, especially for small datasets [4].

- Few signal events → small effective dataset
- High level features → tabular data

**Classifier Setup**

- NN: Ensemble of  $N$  fully connected neural networks
- BDT: Ensemble of  $N$  gradient boosted decision trees

**Study: Uninformative features**

We study the classifiers by introducing uninformative features (features drawn from Gaussian noise), which the NN is particularly sensitive to. The BDT's performance is very robust, meaning that we can add more features to an analysis.

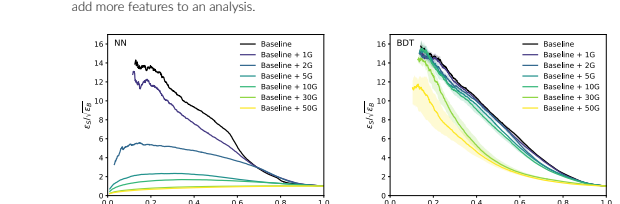


Figure 2: ROC curves of MAD NN/BDT classifiers employing four baseline features and additional Gaussian features. For 30 and 50 Gaussian features, ensembling of BDT increased to  $A_{\text{ROC}} = 38\%$  otherwise  $A_{\text{ROC}} = 36\%$ .

**Study: Additional physics-motivated features**

We study datasets with more subtlety-based features.

- Extended set 1: 10 features (baseline + 6 additional), some largely uninformative
- Extended set 2: 12 features, all slightly informative
- Extended set 3: 56 features, all slightly informative

BDT robustness against uninformative features translates to being well-behaved with additional features. Not present for NN.

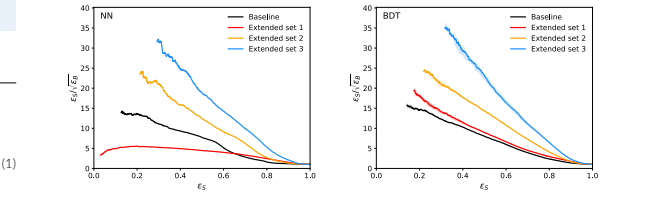


Figure 3: ROC curves of MAD NN/BDT classifier with 4 feature baseline dataset and three extended feature sets.

### Graphs for low level features

**Machine Learning background**

Graph Neural Networks can represent IHEP data in a permutation invariant manner. Architectures can incorporate symmetries directly.

- Very successful on top tagging tasks

**Study: Top tagger on LHCO dataset**

State of the art top taggers were studied on the LHCO R&D dataset.

- Modified LovelaceNet architecture [5] found to result in the best performance.
- Performance drops sooner than observed for high level features.

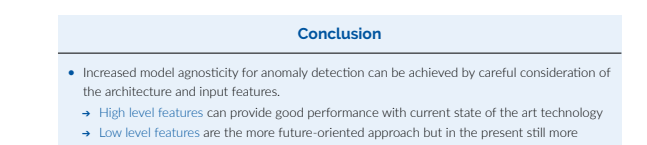


Figure 4: ROC curves for supervised classifier and MAD on low level features.

### Conclusion

- Increased model agnosticity for anomaly detection can be achieved by careful consideration of the architecture and input features.
- High level features can provide good performance with current state of the art technology
- Low level features are the more future-oriented approach but in the present still more difficult to achieve

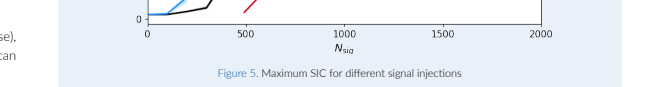


Figure 5: Maximum SIC for different signal injections

### References

[1] G. Kasieczka et al., "The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics," *Phys. Rev. D*, vol. 102, no. 3, p. 034003, 2020.

[2] E. M. Radicevic, B. Kniehm, and J. Thaler, "Classification without labels: Learning from mixed samples in high-energy physics," *JHEP*, vol. 05, p. 174, 2017.

[3] Y. Geng, D. Jiang, J. Jiang, F. Guo, C. Liu, S. Chen, W. Du, Z. M. Ma, and T. Y. Li, "An efficient context-aware graph neural network for jet tagging," *JHEP*, vol. 07, p. 030, 2022.

[4] "Tree-based algorithms for weakly supervised anomaly detection," *Phys. Rev. D*, vol. 102, no. 3, p. 034003, 2020.

[5] L. G. Gerdt, E. Chalkovskiy, and G. W. Moore, "Why do the best models still outperform deep learning on tabular data?"

Weakly supervised anomaly detection can be applied to resonance searches to find BSM physics.

# Increasing the Model Agnosticity of Weakly Supervised Anomaly Detection

## Anomaly Detection

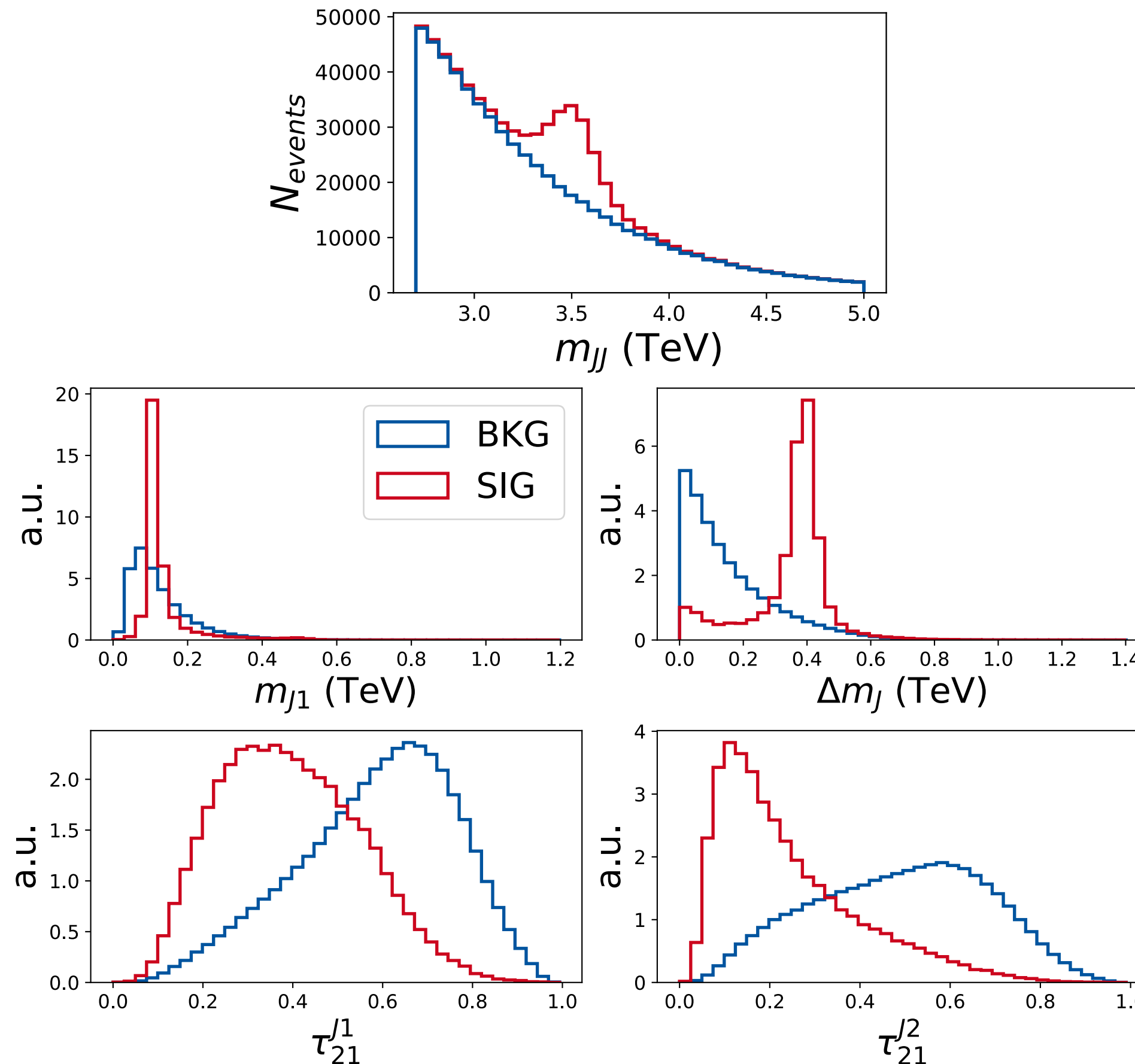
Thorben Finke<sup>1</sup> Joep Geuskens<sup>1</sup> Marie Hein<sup>1</sup> Gregor Kasieczka<sup>2,3</sup> Michael Krämer<sup>1</sup> Alexander Mück<sup>1</sup> Parada Prangchaikul<sup>2</sup> Tobias Quadfasel<sup>2</sup> David Shih<sup>4</sup> Manuel Sommerhalder<sup>2</sup>

<sup>1</sup>TTK RWTH Aachen University <sup>2</sup>IEP Universität Hamburg <sup>3</sup>CDCS Universität Hamburg <sup>4</sup>NHETC Rutgers University

### LHCO R&D dataset

Background: QCD dijets  
Signal:

Usual features for weak supervision papers  
⇒ not very model agnostic



### Increasing the Model Agnosticity of Weakly Supervised Anomaly Detection

Thorben Finke<sup>1</sup> Joep Geuskens<sup>1</sup> Marie Hein<sup>1</sup> Gregor Kasieczka<sup>2,3</sup> Michael Krämer<sup>1</sup> Alexander Mück<sup>1</sup> Parada Prangchaikul<sup>2</sup> Tobias Quadfasel<sup>2</sup> David Shih<sup>4</sup> Manuel Sommerhalder<sup>2</sup>

**Motivation**

- To find new physics, improve largely model agnostic searches.
- Use pattern recognition capability of machine learning in high dimensional feature space to gain higher sensitivity.
- Problem: Currently many papers use only 4 high level features ("baseline" feature set) on one benchmark dataset (LHCO R&D dataset [1]).
- For more model agnostic setup need to be able to use more features.

**Goal: Improve classifier setup for more high level features and low level features**

**Weakly supervised anomaly detection**

Classification Without Labels (CWoLa) [2]

- Classifier between mixed datasets  $p_S(x) = f_1 p_S(x) + (1 - f_1) p_B(x)$  with signal fractions  $f_i$
- $R_{\text{mixed}} = \frac{f_1 R_{\text{optimal}}(x) + (1 - f_1)}{f_1 R_{\text{optimal}}(x) + (1 - f_1)}$  where  $R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$  [2]
- is the optimal classifier between signal and background distributions  $p_S, p_B$ .
- Mathematically equivalent as  $R_{\text{mixed}}$  monotonous in  $R_{\text{optimal}}$ .

**Application to resonance searches**

- Divide data into signal region (SR) and sideband (SB), where
- $p_S(x) = p_S(x|m \in \text{SR}) + p_B(x|m \in \text{SR})$  and  $p_B(x) = p_B(x|m \in \text{SB})$  [2]
- for classification features  $x$ .
- Construct "background template" from SB, ideally with  $p(x) = p_B(x|m \in \text{SB})$
- Here, we use idealized case to study classifier only

**BDTs for high level features [3]**

**Machine Learning background**

Boosted Decision Trees (BDTs) are known to be very effective on tabular data, especially for small datasets [4].

- Few signal events → small effective dataset
- High level features → tabular data

**Classifier Setup**

- NN: Ensemble of  $N$  fully connected neural networks
- BDT: Ensemble of  $N$  gradient boosted decision trees

**Study: Uninformative features**

We study the classifiers by introducing uninformative features (features drawn from Gaussian noise), which the NN is particularly sensitive to. The BDT's performance is very robust, meaning that we can add more features to an analysis.

**Study: Additional physics-motivated features**

We study datasets with more subtlety-based features.

- Extended set 1: 10 features (baseline + 6 additional), some largely uninformative
- Extended set 2: 12 features, all slightly informative
- Extended set 3: 56 features, all slightly informative

BDT robustness against uninformative features translates to being well-behaved with additional features. Not present for NN.

**Graphs for low level features**

**Machine Learning background**

Graph Neural Networks can represent HEP data in a permutation invariant manner. Architectures can incorporate symmetries directly.

- Very successful on top tagging tasks

**Study: Top tagger on LHCO dataset**

State of the art top taggers were studied on the LHCO R&D dataset.

- Modified LovénZiel architecture [5] found to result in the best performance.
- Performance drops sooner than observed for high level features.

**Conclusion**

- Increased model agnosticity for anomaly detection can be achieved by careful consideration of the architecture and input features.
- High level features can provide good performance with current state of the art technology
- Low level features are the more future-oriented approach but in the present still more difficult to achieve

**References**

[1] G. Kasieczka et al., "The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics," Phys. Rev. D, vol. 102, no. 3, p. 034003, 2020.

[2] E. M. Abdalla, B. B. Kniehl, and J. Thaler, "Classification without labels: Learning from mixed samples in high-energy physics," JHEP, vol. 05, p. 174, 2017.

[3] Y. Gao, D. Jiang, J. Jiang, H. Du, C. Li, S. Chen, W. Du, Z. M. Ma, and T. Li, "An efficient context-aware graph neural network for jet tagging," JHEP, vol. 07, p. 030, 2022.

[4] "Tree-based algorithms for weakly supervised anomaly detection," Phys. Rev. D, vol. 107, no. 3, p. 034003, 2023.

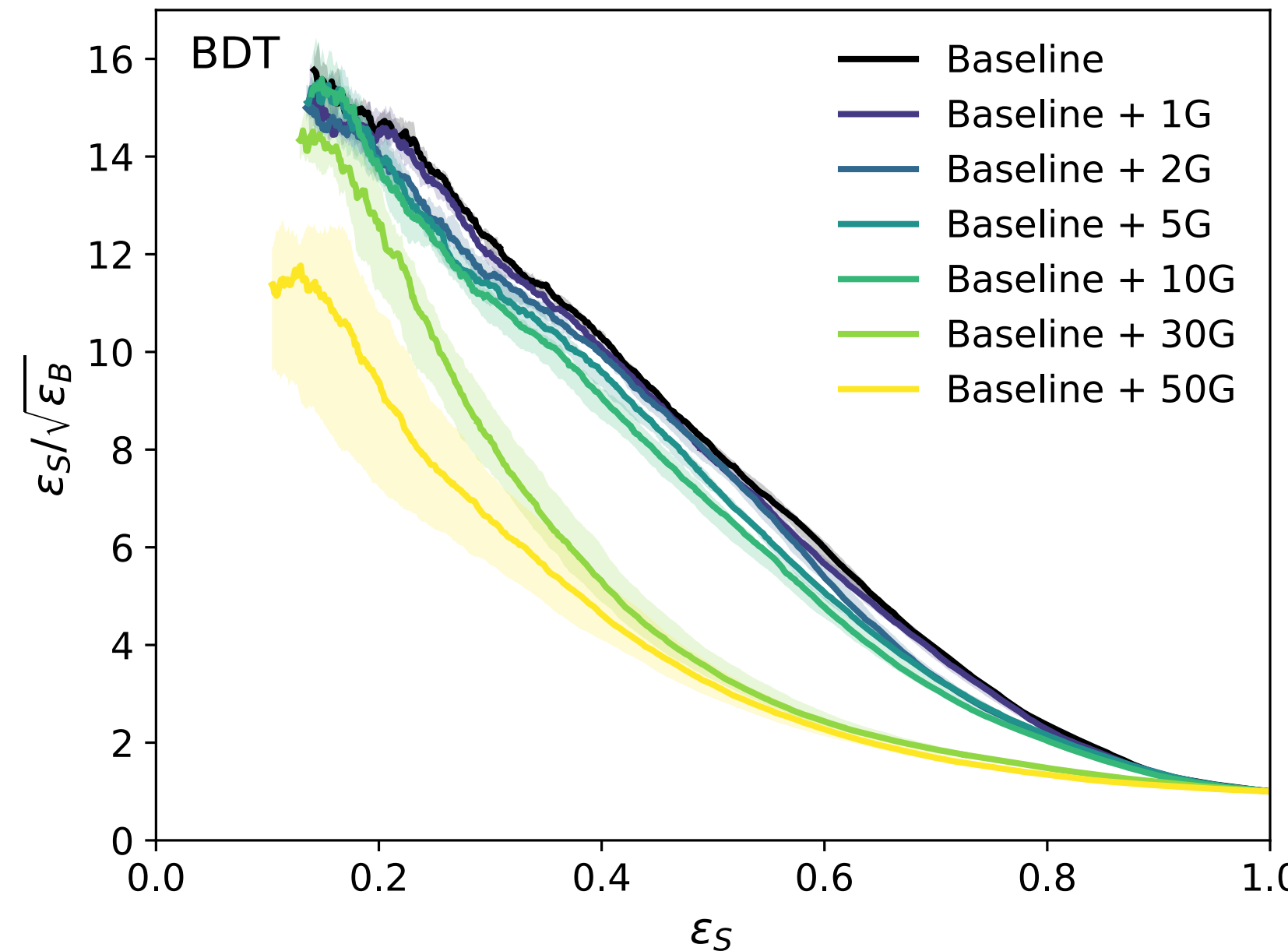
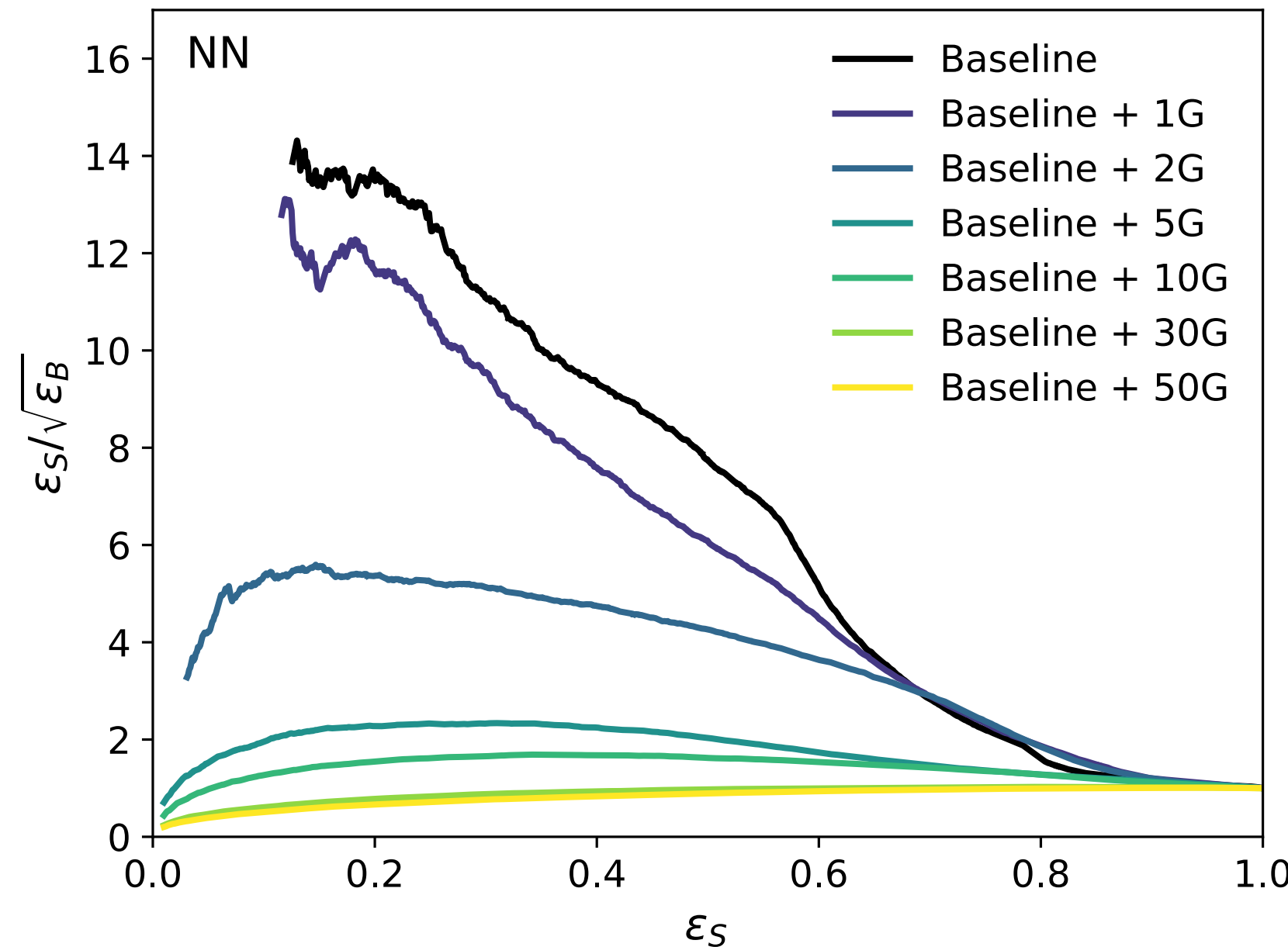
[5] L. Grottel, E. Chalkley, and G. Wenzauer, "Who do the best models still outperform deep learning on tabular data?," arXiv:2305.17427, 2023.

[6] T. Fink, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, and M. Sommerhalder, "The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics," Phys. Rev. D, vol. 102, no. 3, p. 034003, 2020.

# Increasing the Model Agnosticity of Weakly Supervised Anomaly Detection

Thorben Finke<sup>1</sup> Joep Geuskens<sup>1</sup> Marie Hein<sup>1</sup> Gregor Kasieczka<sup>2,3</sup> Michael Krämer<sup>1</sup> Alexander Mück<sup>1</sup> Parada Prangchaikul<sup>2</sup> Tobias Quadfasel<sup>2</sup> David Shih<sup>4</sup> Manuel Sommerhalder<sup>2</sup>

<sup>1</sup>TTK RWTH Aachen University <sup>2</sup>IEP Universität Hamburg <sup>3</sup>CDCS Universität Hamburg <sup>4</sup>NHETC Rutgers University



To include more features, robustness against uninformative features is necessary, which is not present for NNs.

### Increasing the Model Agnosticity of Weakly Supervised Anomaly Detection

Thorben Finke<sup>1</sup> Joep Geuskens<sup>1</sup> Marie Hein<sup>1</sup> Gregor Kasieczka<sup>2,3</sup> Michael Krämer<sup>1</sup> Alexander Mück<sup>1</sup> Parada Prangchaikul<sup>2</sup> Tobias Quadfasel<sup>2</sup> David Shih<sup>4</sup> Manuel Sommerhalder<sup>2</sup>

TTK RWTH Aachen University <sup>1</sup>IEP Universität Hamburg <sup>2</sup>CDCS Universität Hamburg <sup>4</sup>NHETC Rutgers University

**Motivation**

- To find new physics, improve largely model agnostic searches, e.g., resonance searches
- Use pattern recognition capability of machine learning in high dimensional feature space to gain higher sensitivity
- Problem: Currently many papers use only 4 high level features ("baseline" feature set) on one benchmark dataset (LHCO R&D dataset [1])
- For more model agnostic setup need to be able to use more features
- Goal: Improve classifier setup for more high level features and low level features

**Weakly supervised anomaly detection**

Classification Without Labels (CWoLa) [2]

- Classifier between mixed datasets  $p_S(x) = f_1 p_S(x) + (1 - f_1) p_B(x)$  with signal fractions  $f_i$
- $R_{mixed} = \frac{f_1 R_{optimal}(x) + (1 - f_1)}{f_1 R_{optimal}(x) + (1 - f_1)}$  where  $R_{optimal}(x) = \frac{p_S(x)}{p_B(x)}$  [1]
- is the optimal classifier between signal and background distributions  $p_S, p_B$
- Mathematically equivalent as  $R_{mixed}$  monotonous in  $R_{optimal}$

Application to resonance searches

- Divide data into signal region (SR) and sideband (SB), where
- $p_S(x) = p_S(x|m \in SR) + p_B(x|m \in SR)$  and  $p_B(x) = p_B(x|m \in SB)$  [2]
- for classification features  $x$ .
- Construct "background template" from SB, ideally with  $p(x) = p_B(x|m \in SR)$
- Here, we use idealized case to study classifier only

**BDTs for high level features [3]**

Machine Learning background

Boosted Decision Trees (BDTs) are known to be very effective on tabular data, especially for small datasets [4].

- Few signal events → small effective dataset
- High level features → tabular data

Classifier Setup

- NN: Ensemble of  $N$  fully connected neural networks
- BDT: Ensemble of  $N$  gradient boosted decision trees

Study: Uninformative features

We study the classifiers by introducing uninformative features (features drawn from Gaussian noise), which the NN is particularly sensitive to. The BDT's performance is very robust, meaning that we can add more features to an analysis.

**Study: Additional physics-motivated features**

We study datasets with more subjectivity-based features.

- Extended set 1: 10 features (baseline + 6 additional), some largely uninformative
- Extended set 2: 12 features, all slightly informative
- Extended set 3: 56 features, all slightly informative

BDT robustness against uninformative features translates to being well-behaved with additional features. Not present for NN.

**Graphs for low level features**

Machine Learning background

Graph Neural Networks can represent HEP data in a permutation invariant manner. Architectures can incorporate symmetries directly.

- Very successful on top tagging tasks

Study: Top tagger on LHCO dataset

State of the art top taggers were studied on the LHCO R&D dataset:

- Modified LovaszNet architecture [5] found to result in the best performance.
- Performance drops sooner than observed for high level features.

**Conclusion**

- Increased model agnosticity for anomaly detection can be achieved by careful consideration of the architecture and input features.
- High level features can provide good performance with current state of the art technology
- Low level features are the more future-oriented approach but in the present still more difficult to achieve

**References**

[1] G. Kasieczka et al., "The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics," *Phys. Rev. D*, vol. 102, no. 3, p. 034003, 2020.

[2] E. M. Abdellatif, B. B. B. and T. T. "Classifier without labels, learning from mixed samples in high energy physics," *JHEP*, vol. 10, p. 174, 2017.

[3] Y. Gao, D. Jiang, J. Wang, R. Du, C. Li, S. Chen, W. Du, Z. M. Ma, and T. Y. Li, "An efficient context-aware graph neural network for jet tagging," *JHEP*, vol. 07, p. 030, 2022.

[4] "Tree-based algorithms for weakly supervised anomaly detection," *Phys. Rev. D*, vol. 102, no. 3, p. 034003, 2020.

[5] L. G. Oliveira, E. Chinellato, and G. W. Senise, "Who do the best models still outperform deep learning on tabular data?," *arXiv preprint arXiv:2104.00001*, 2021.