Quantum-inspired event reconstruction with Tensor Networks

Jack Y. Araz
Institute For Particle Physics Phenomenology
Durham University

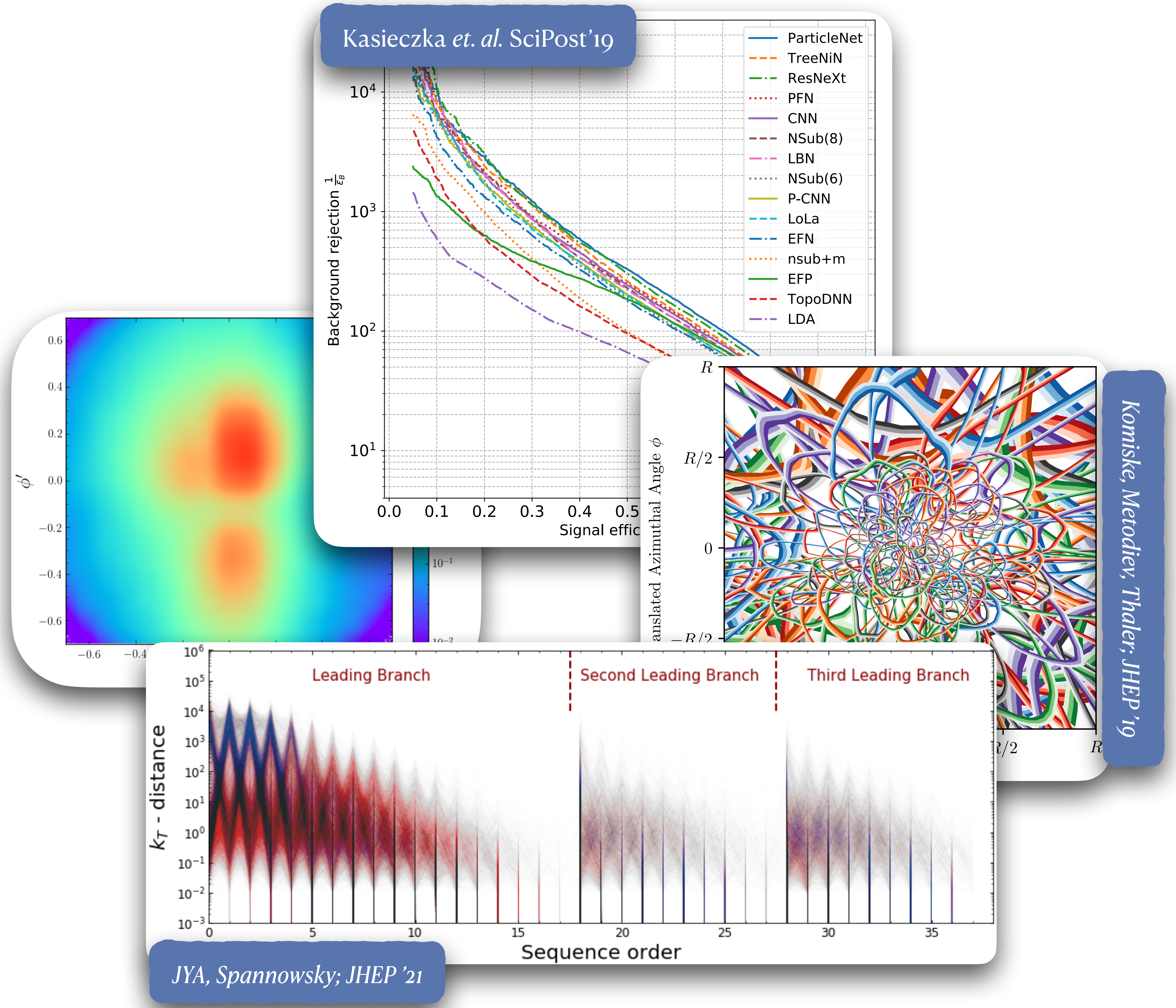
Based on JHEP 08 (2021) 112; arXiv: 2106.08334 [hep-ph]
with Michael Spannowsky

2nd symposium on Artificial Intelligence for Science, Industry, and Society

October 11th - 15th, 2021

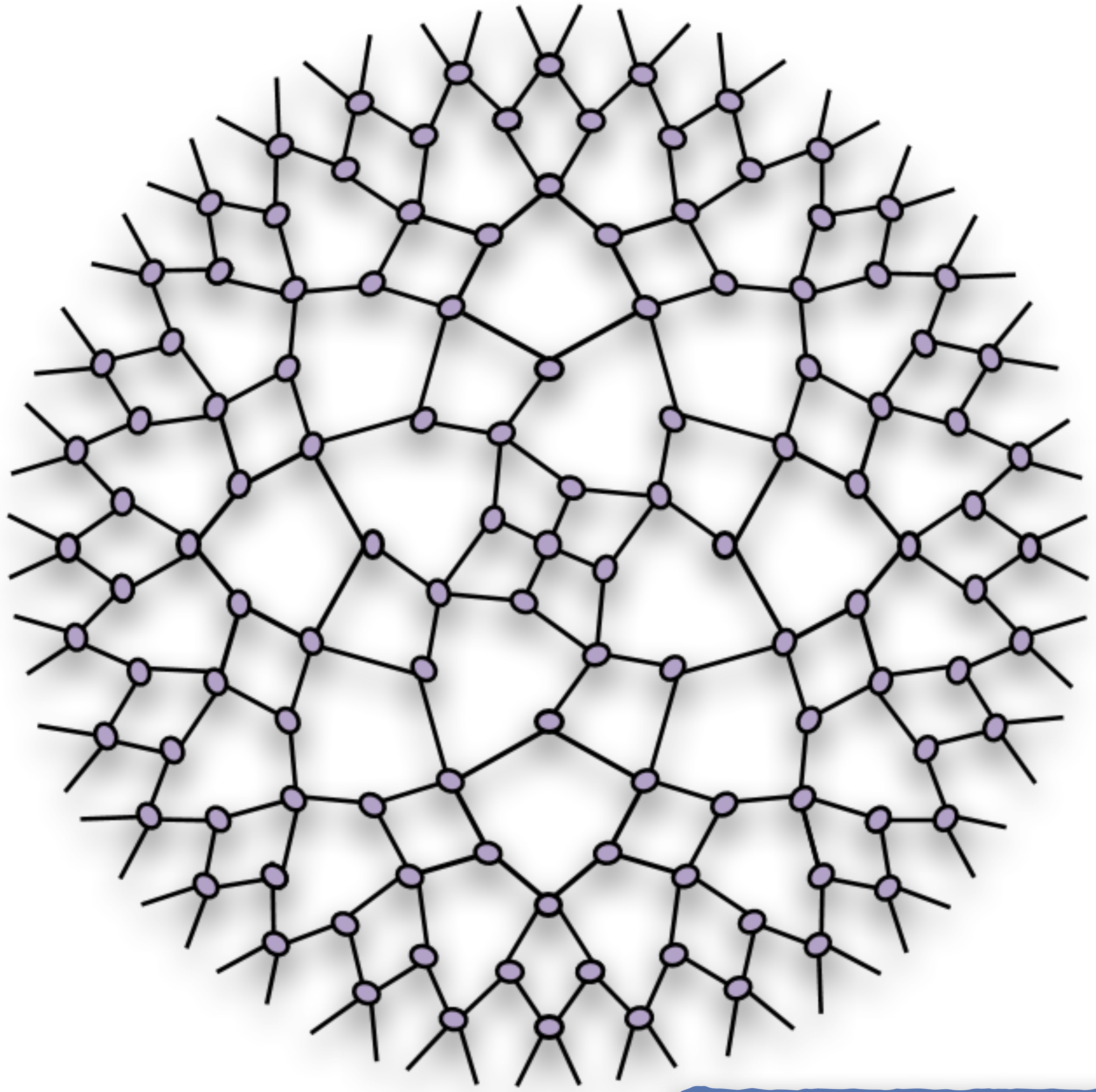
Sales pitch of the talk!

- We more or less know how to get well performing Neural Network to classify jets, LHC events, even cats and dogs…
- What we don’t know is what this network learns.
- Can we use Quantum Mechanics to have more insight about the learning process?
  - ✦ What has a model learned?
  - ✦ What is learning?
  - ✦ How to develop “insightful” algorithms?



Outline

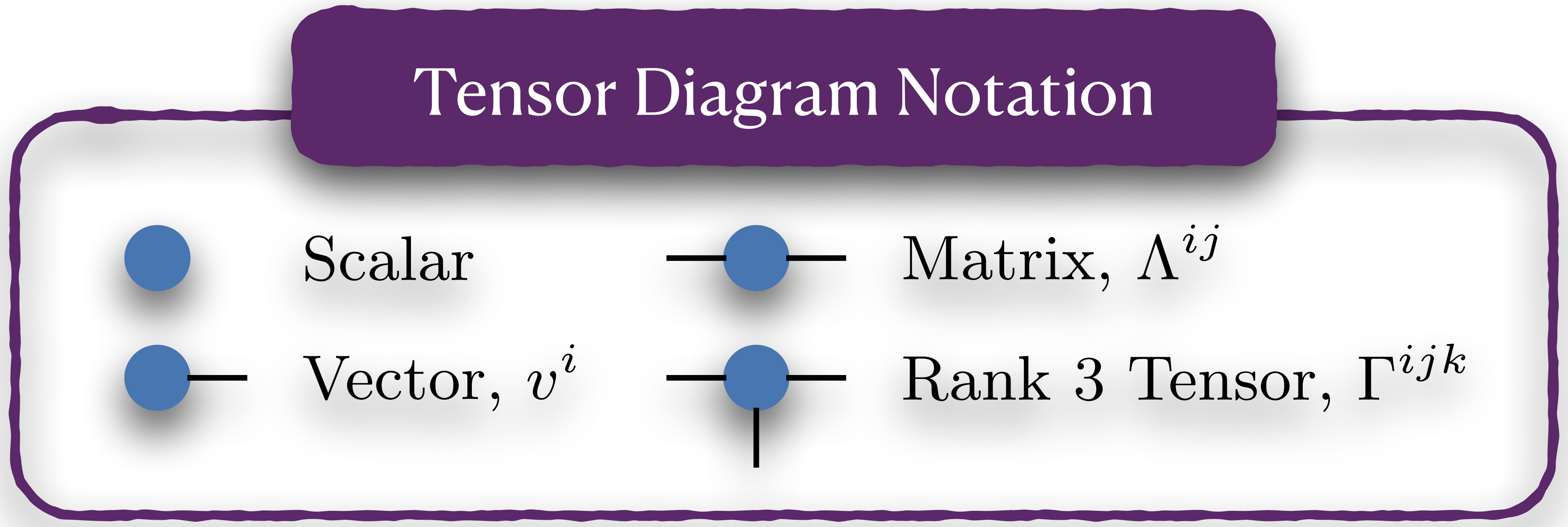
- Introduction
  - What are Tensor Networks and how to play with them?
- Tensor Networks for Machine Learning
  - Top tagging through Tensor Networks
- Conclusion

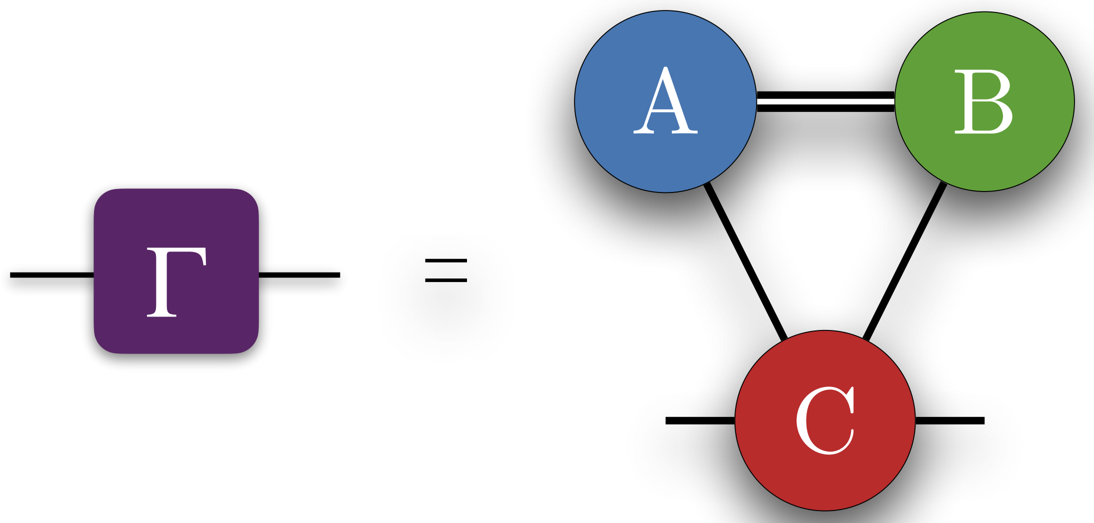


Evenbly, Vidal; J Stat Phys 2011

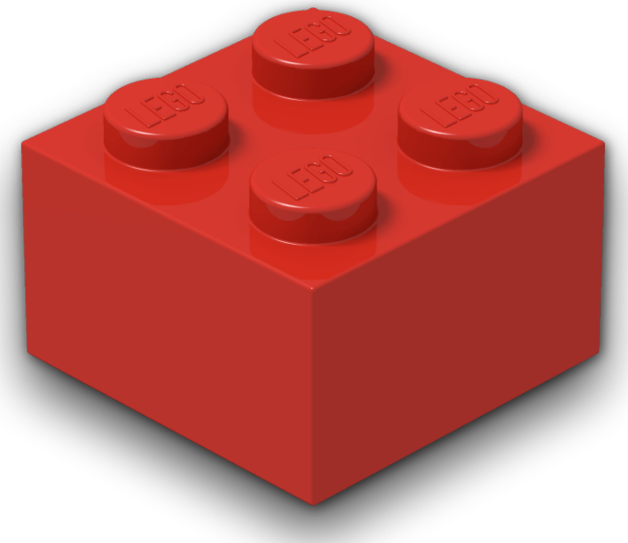
Introduction

Tensor Networks: Origins





$$\Gamma^{ij} = A^{klm} B^{n}{}_{lm} C^{ij}{}_{kn}$$

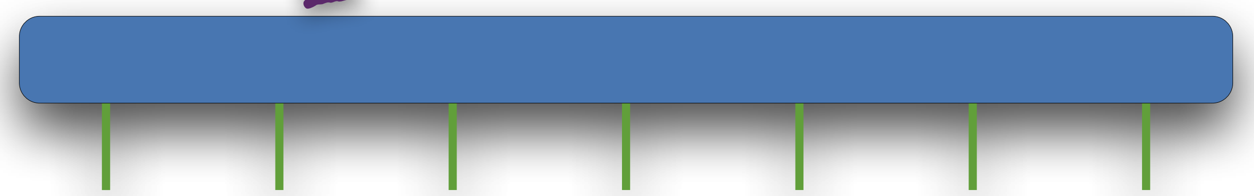


$$= L^{ijkl}{}_{\mu\nu\rho\theta}$$

Amplitude of the wave-function

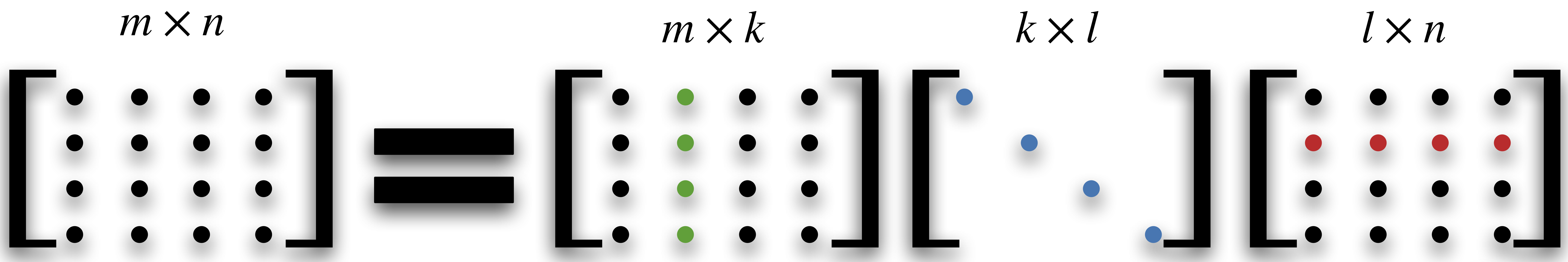
$$|\Psi\rangle = \sum_{p_1,\dots,p_n=0} \mathscr{W}_{p_1\dots p_n} |p_1\rangle \otimes |p_2\rangle \otimes \dots \otimes |p_n\rangle$$

$$\forall |p_i\rangle \in \mathscr{H}^{\otimes N} \quad \rightarrow \quad |p_i\rangle \in \{|\uparrow\rangle, |\downarrow\rangle\}$$



Computational cost is $\mathcal{O}(d^N)$!!!

Singular Value Decomposition



$$\mathbf{M} = \mathbf{U} \; \mathbf{S} \; \mathbf{V}^\dagger$$

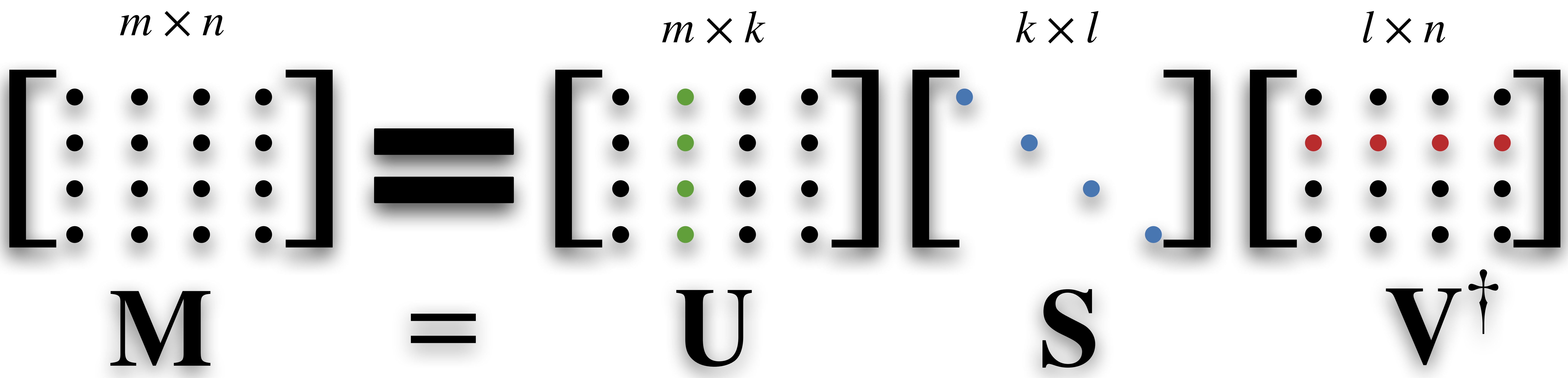
Orthogonal singular column vectors

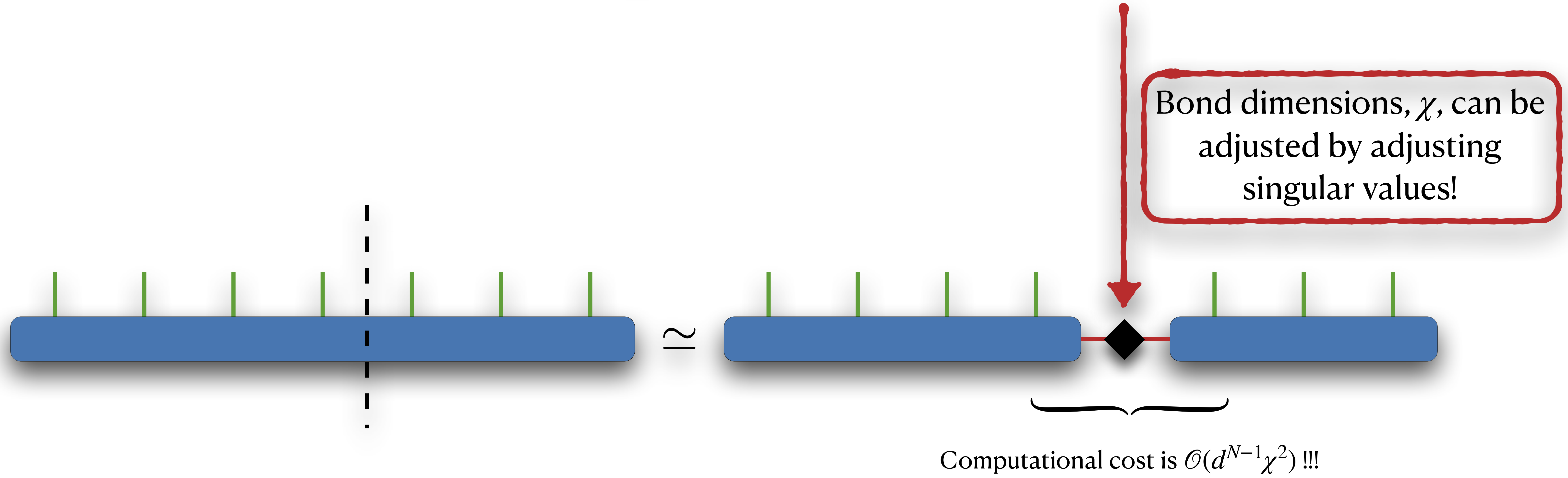
Positive definite singular values in descending order

Orthogonal singular row vectors

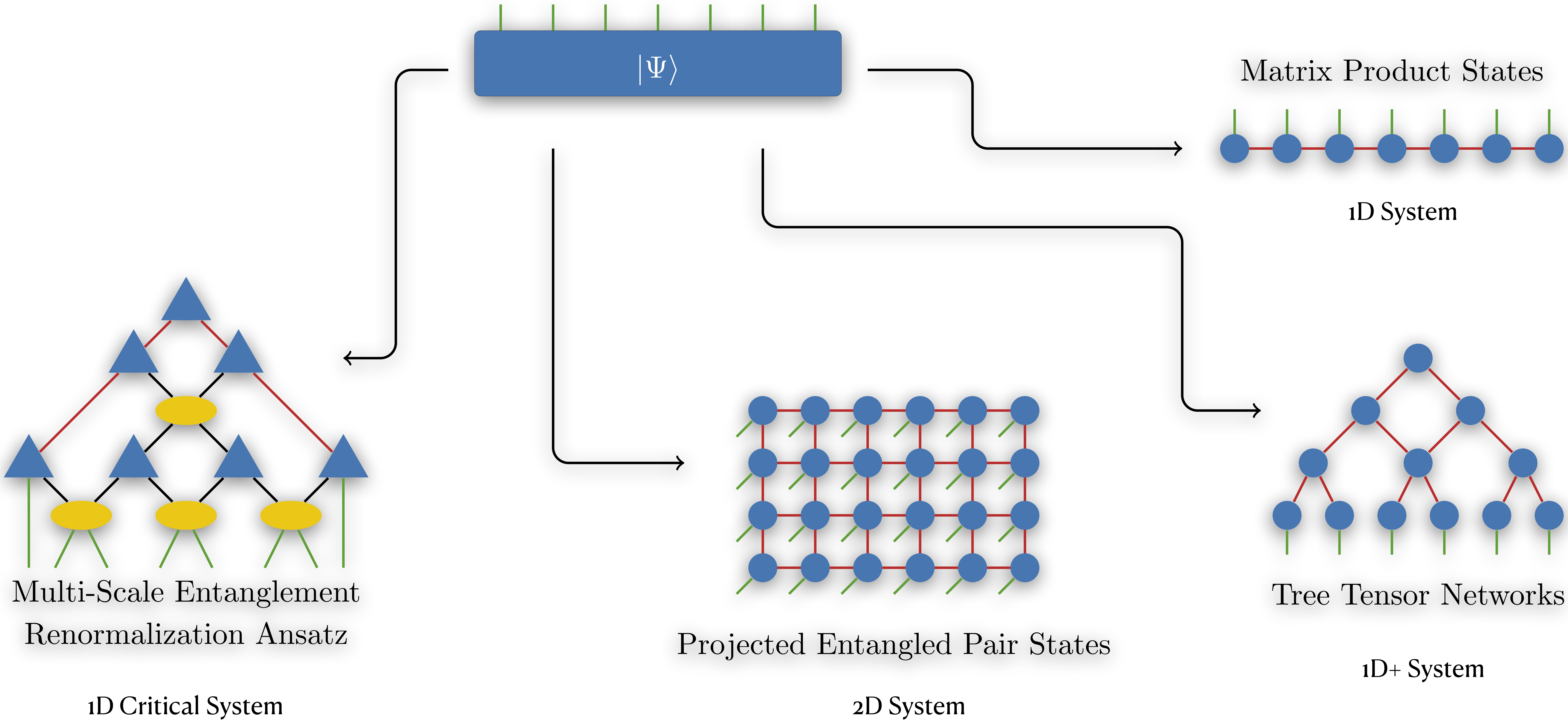
By changing the number of singular values one can change the accuracy of the decomposition!!!

Singular Value Decomposition

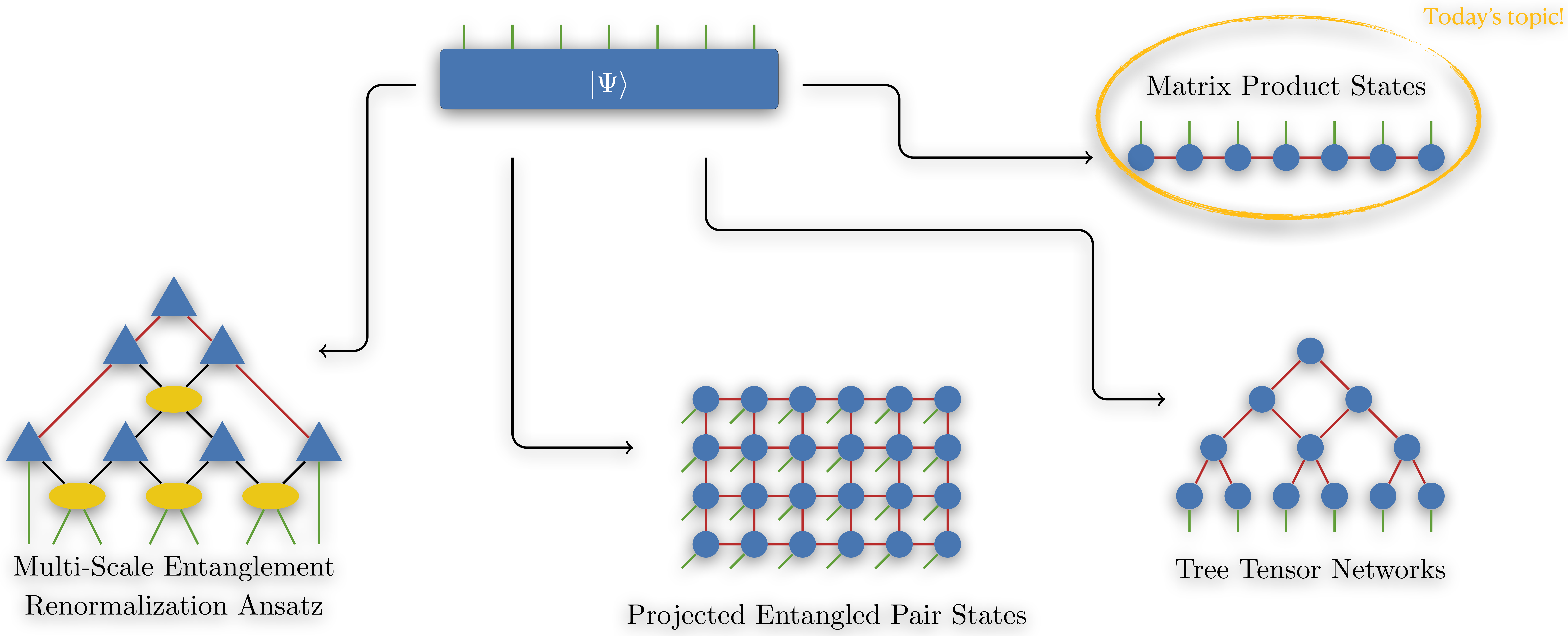




Types of Tensor Networks (some of them)



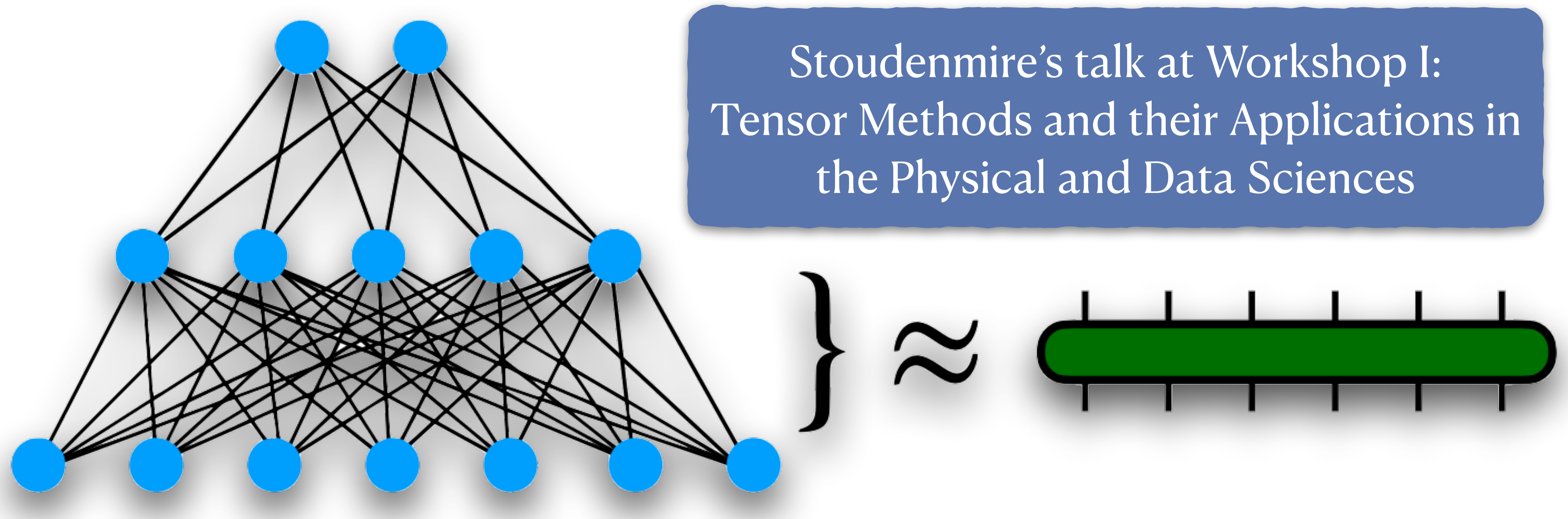
Types of Tensor Networks (some of them)



Why TNs “might” perform well in classification tasks?

Garipov, Podoprikhin, Novikov, Vetrov
arXiv:1611.03214

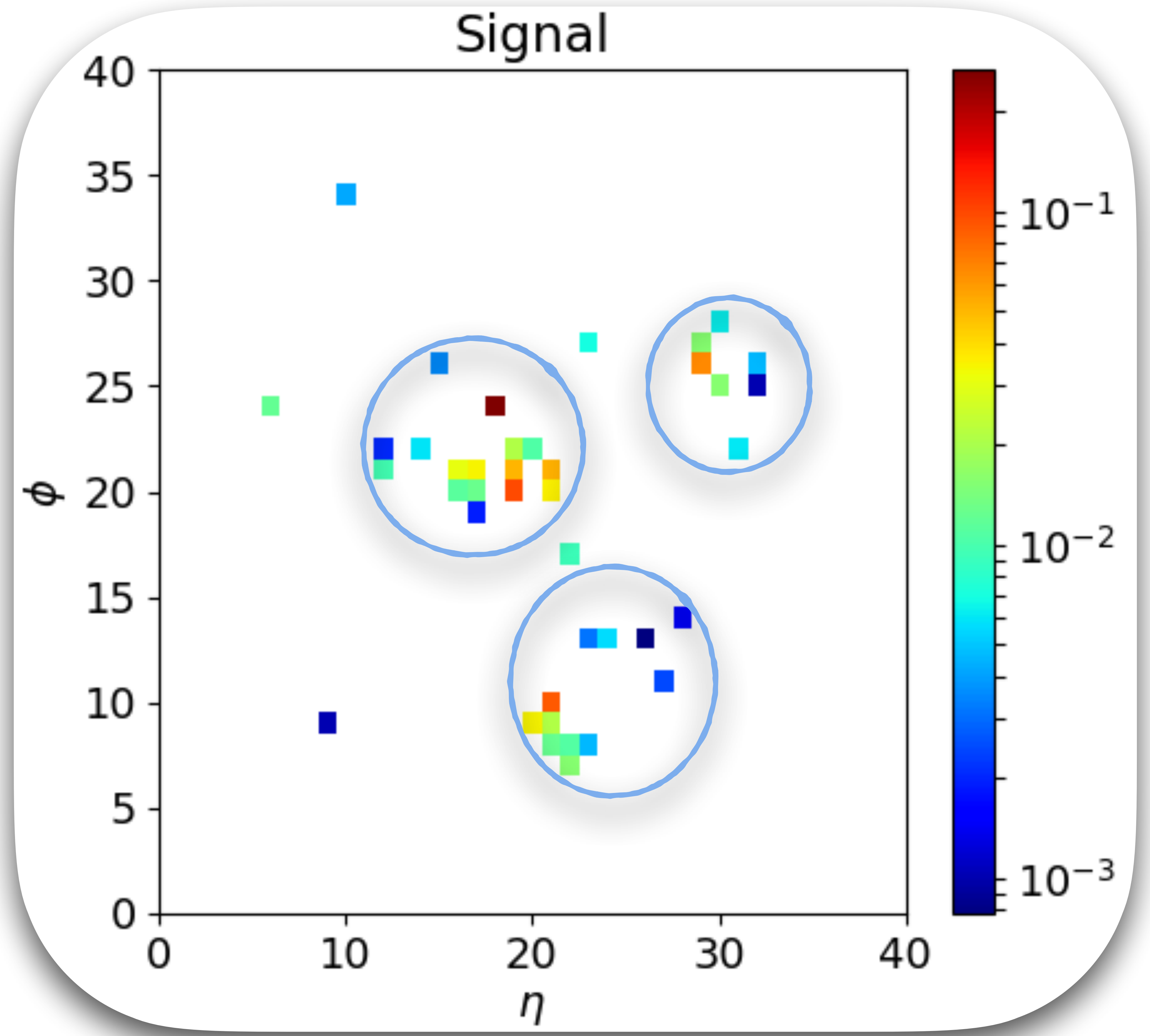
Stoudenmire’s talk at Workshop I: Tensor Methods and their Applications in the Physical and Data Sciences



Not in this talk!

- The range of a node in a Tensor Network bounded by its bond dimension.
- Tensor Networks can capture local “anomalies”.
- Jets can produce localized clusters!!

Kasieczka *et. al.* SciPost’19



Tensor Networks for Machine Learning

Matrix Product States for Classification

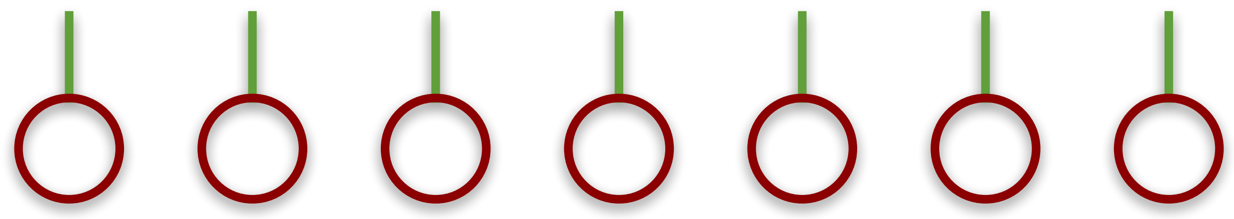
Sub-Outline

- How to embed the data?
- How to form a network?
- How to train the network?

Data Embedding

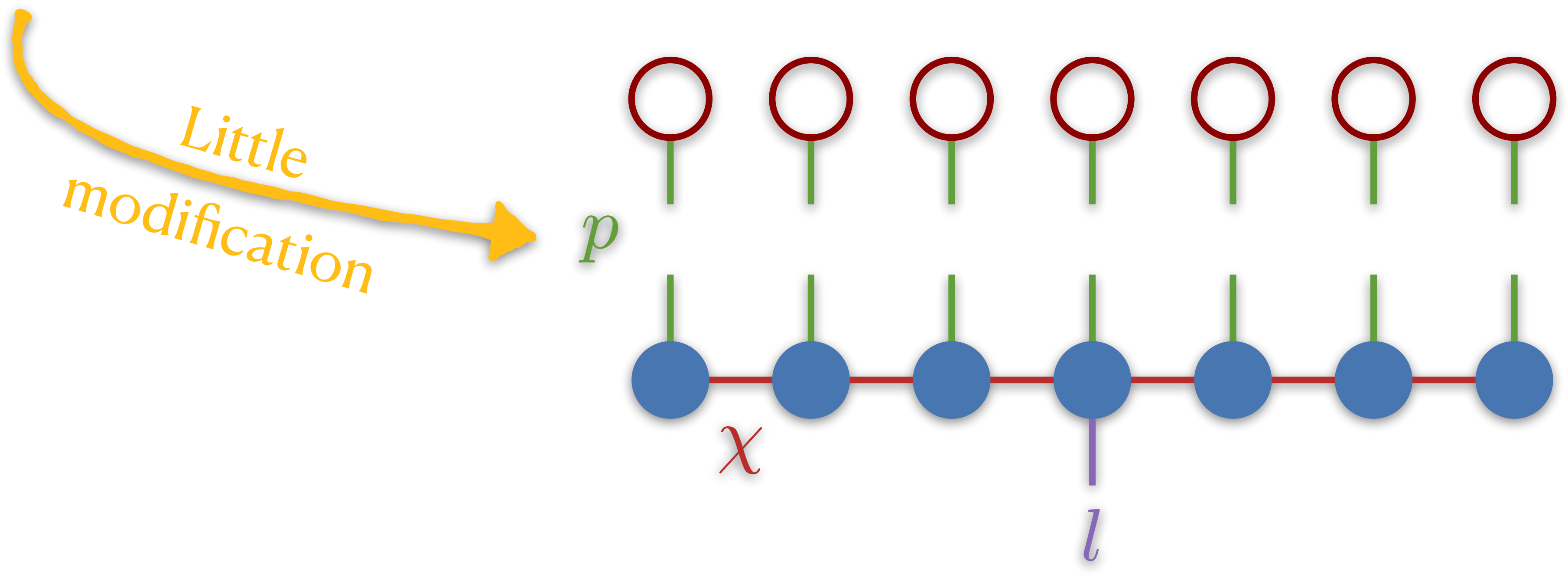
$$\Phi^{p_1\cdots p_n}(\mathbf{x}) = \phi^{p_1}(x_1) \otimes \phi^{p_2}(x_2) \otimes \cdots \otimes \phi^{p_n}(x_n)$$

$$\phi^{p_i}(x_i) = \begin{bmatrix} \cos(x_i\ \pi/2) \\ \sin(x_i\ \pi/2) \end{bmatrix} \quad \text{or} \quad \phi^{p_i}(x_i) = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \end{bmatrix} \quad \text{or} \quad \cdots$$



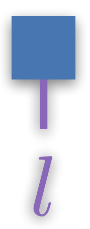
$= \Phi^{p_1\cdots p_n}(\mathbf{x})$

$$|\Psi\rangle = \sum_{p_1,\ldots,p_n=0} \mathscr{W}_{p_1\ldots p_n} |p_1\rangle \otimes |p_2\rangle \otimes \ldots \otimes |p_n\rangle$$



$= \Phi^{p_1\cdots p_n}(\mathbf{x})$

$= \mathcal{W}^l{}_{p_1\cdots p_n}$



$:= f^l(\mathbf{x}) = \mathcal{W}^l{}_{p_1\cdots p_n}\ \Phi^{p_1\cdots p_n}(\mathbf{x})$

Matrix Product States for Classification

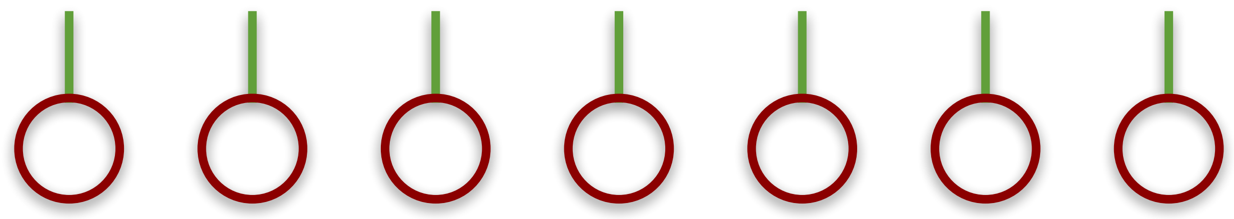
Sub-Outline

- How to embed the data?
- How to form a network?
- How to train the network?

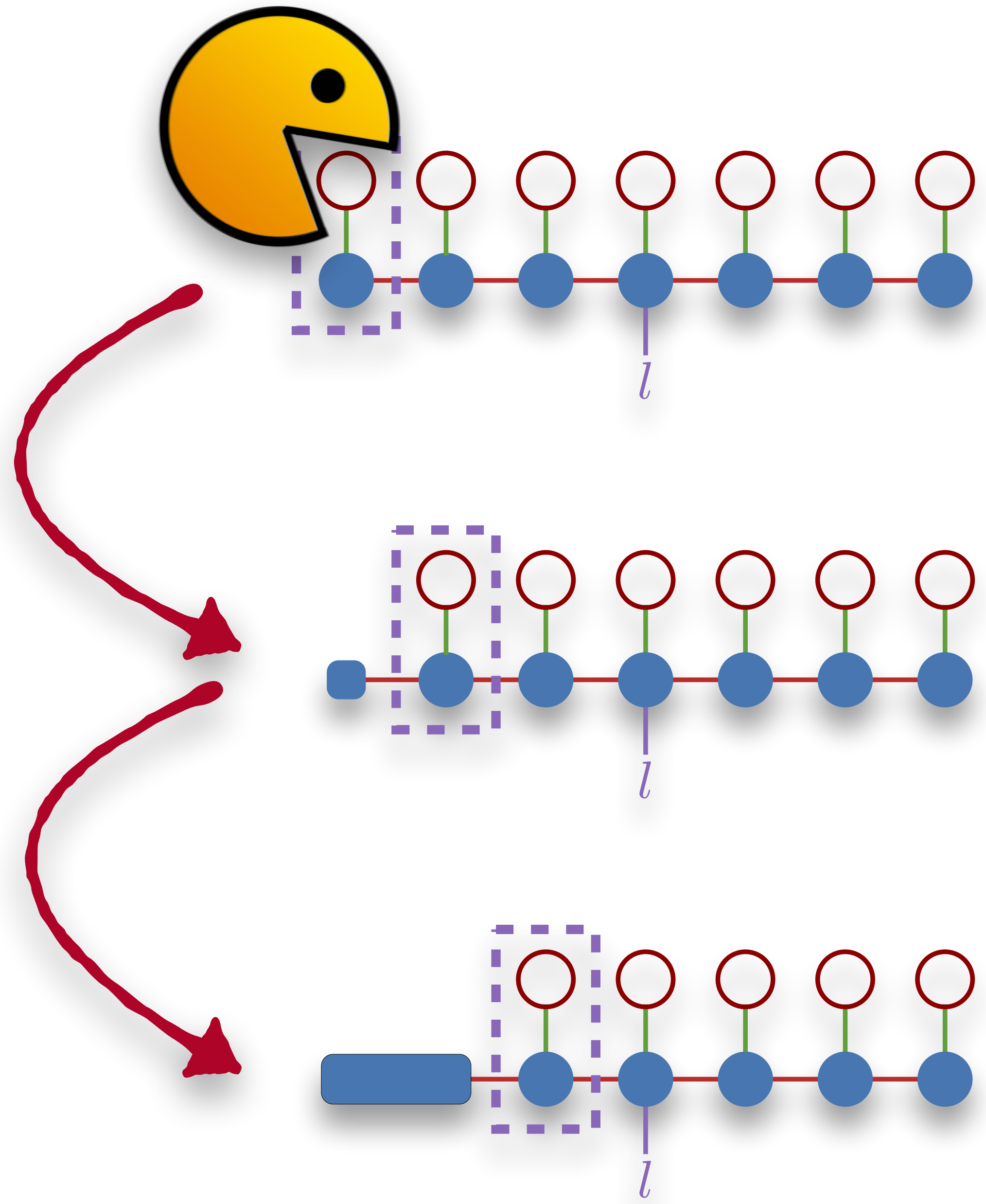
Data Embedding

$$\Phi^{p_1\cdots p_n}(\mathbf{x}) = \phi^{p_1}(x_1) \otimes \phi^{p_2}(x_2) \otimes \cdots \otimes \phi^{p_n}(x_n)$$

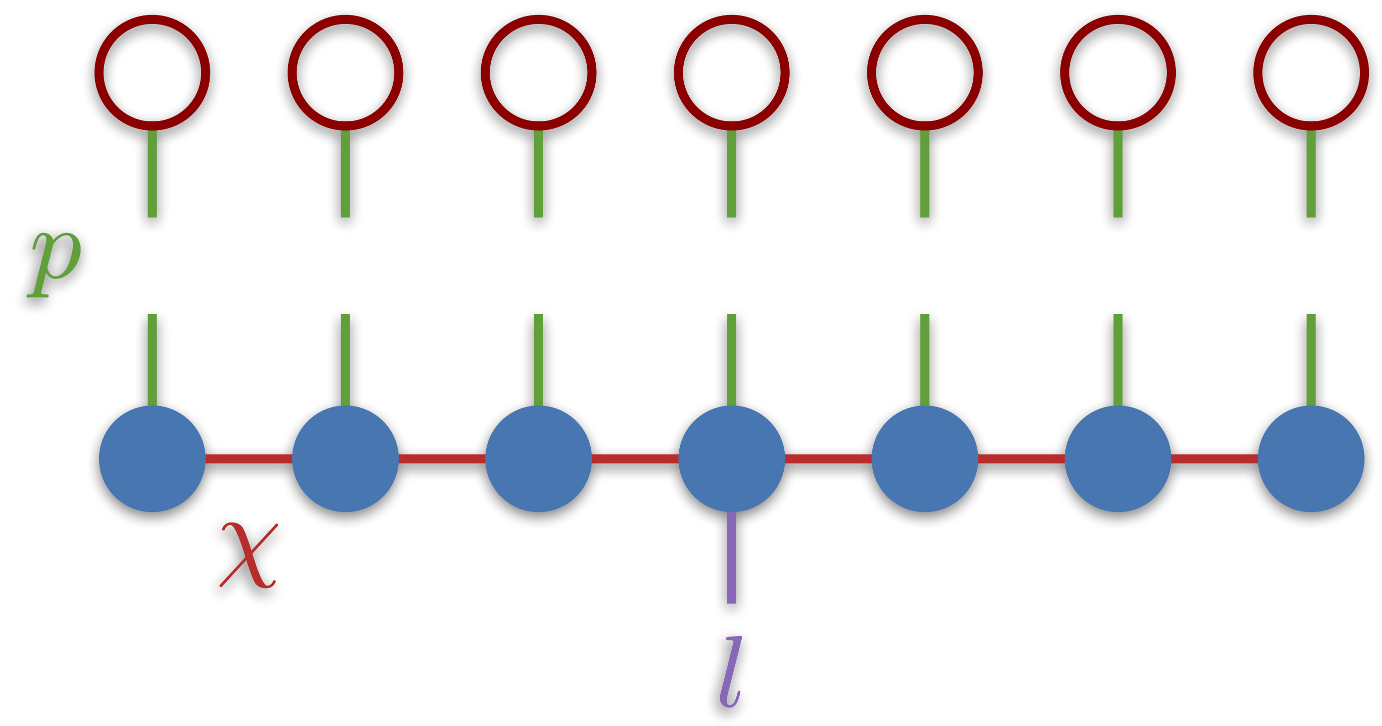
$$\phi^{p_i}(x_i) = \begin{bmatrix} \cos(x_i\ \pi/2) \\ \sin(x_i\ \pi/2) \end{bmatrix} \quad \text{or} \quad \phi^{p_i}(x_i) = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \end{bmatrix} \quad \text{or} \quad \cdots$$



$= \Phi^{p_1\cdots p_n}(\mathbf{x})$

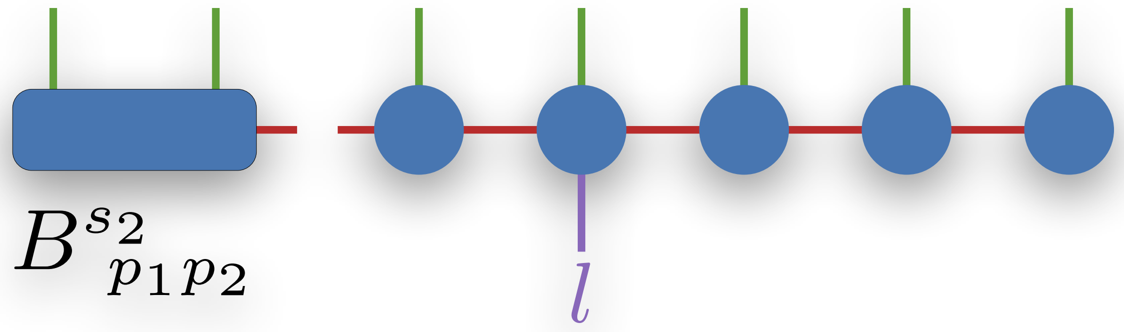


An efficient contraction algorithm is essential for training!

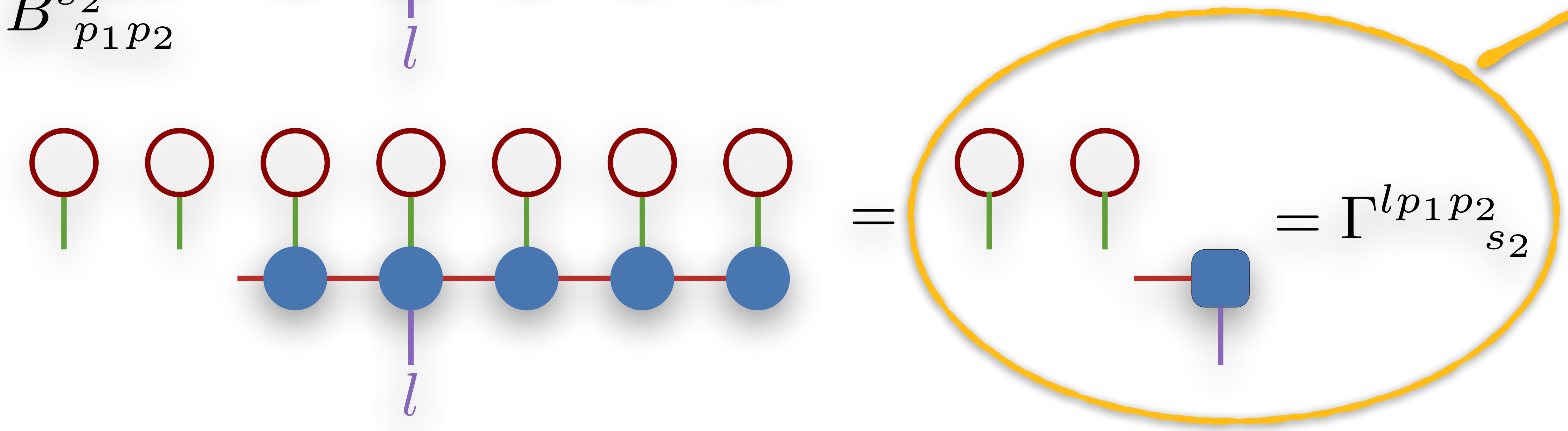


Density Matrix Renormalization Group Algorithm

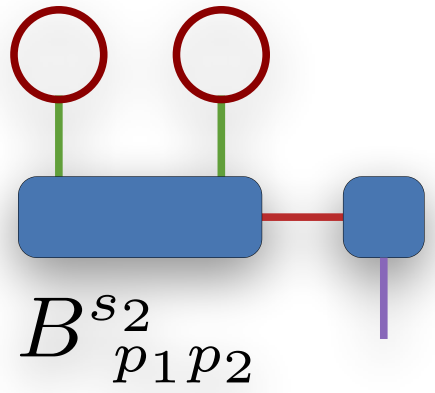
$I.$



$II.$



$III.$



$$= B^{s_2}{}_{p_1 p_2} \Gamma^{l p_1 p_2}{}_{s_2} = f^l(\mathbf{x}^{(\mathbf{n})}) \to p(l, \mathbf{x}) = \left| f^l(\mathbf{x}^{(\mathbf{n})}) \right|^2$$

$$\mathcal{L} = L\left(p(l, \mathbf{x}),\ l^{truth}\right)$$

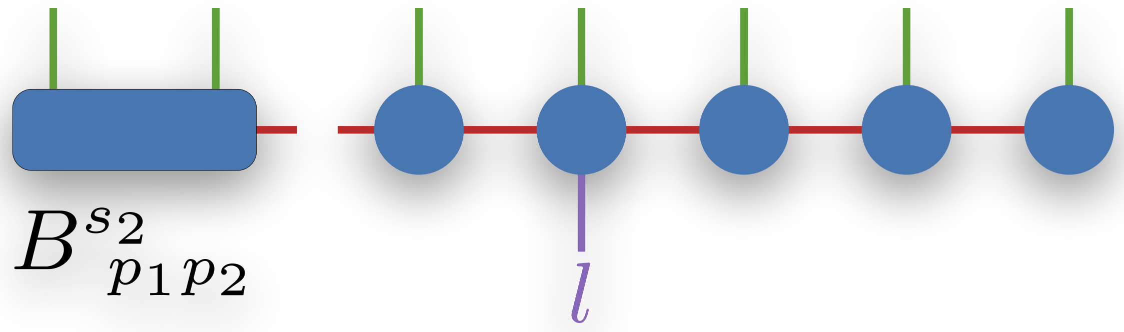
$IV.$

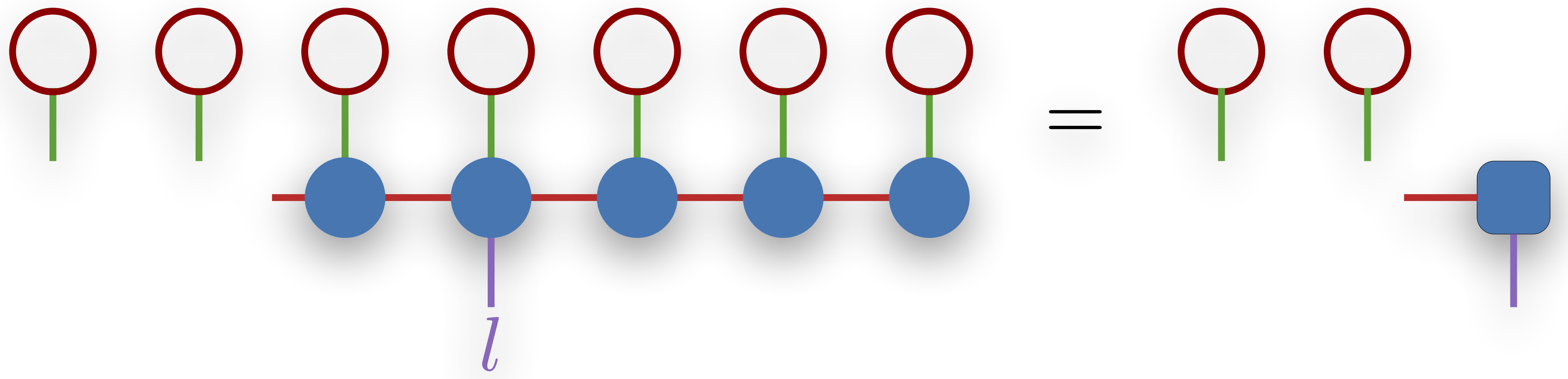
$$\tilde{B}^{s_2}{}_{p_1 p_2} = B^{s_2}{}_{p_1 p_2} - \eta \frac{\partial \mathcal{L}}{\partial B^{s_2}{}_{p_1 p_2}}$$

$$\frac{\partial f^l(\mathbf{x}^{(i)})}{\partial B^{s_2}{}_{p_1 p_2}} = \Gamma^{l p_1 p_2}{}_{s_2}$$

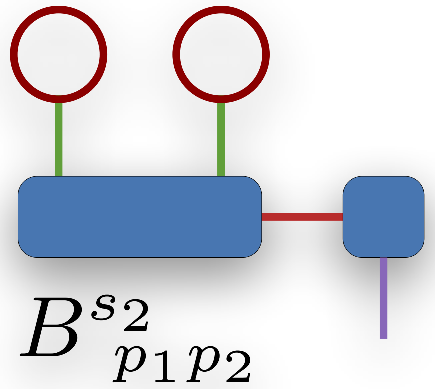
$$L = \frac{1}{N} \sum_{x \in \mathbf{x}^N} l^{truth} \log\left(p(l, x)\right)$$

Density Matrix Renormalization Group Algorithm

I. 

II. 

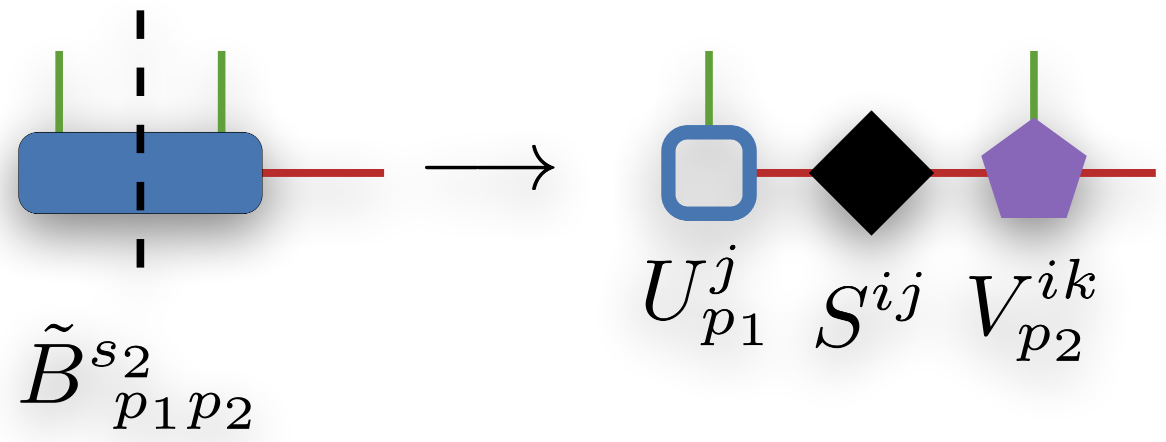
$= \Gamma^{l p_1 p_2}{}_{s_2}$

III. 

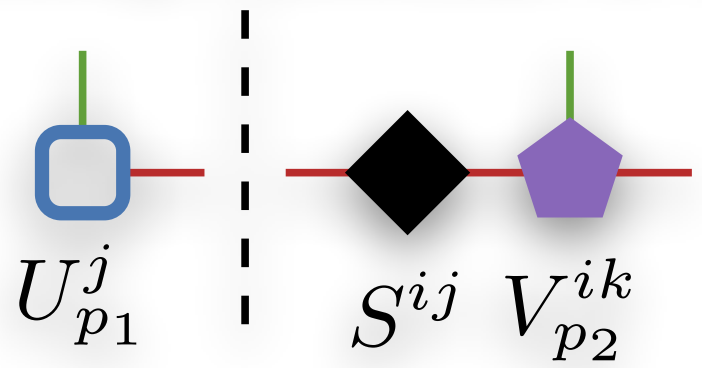
$$= B^{s_2}{}_{p_1 p_2} \Gamma^{l p_1 p_2}{}_{s_2} = f^l(\mathbf{x}^{(\mathbf{n})}) \to p(l, \mathbf{x}) = \left| f^l(\mathbf{x}^{(\mathbf{n})}) \right|^2$$

$$\mathcal{L} = L\left(p(l, \mathbf{x}),\ l^{truth}\right)$$

IV. $$\tilde{B}^{s_2}{}_{p_1 p_2} = B^{s_2}{}_{p_1 p_2} - \eta \frac{\partial \mathcal{L}}{\partial B^{s_2}{}_{p_1 p_2}}$$

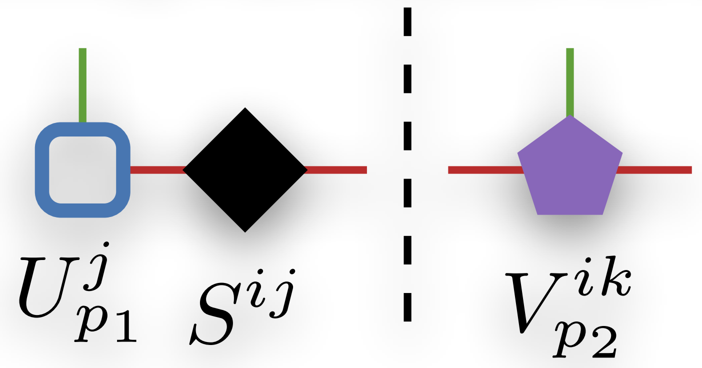
V. 

VI. Sweep direction to the right:



Node I Node II

Sweep direction to the left:



Node I Node II

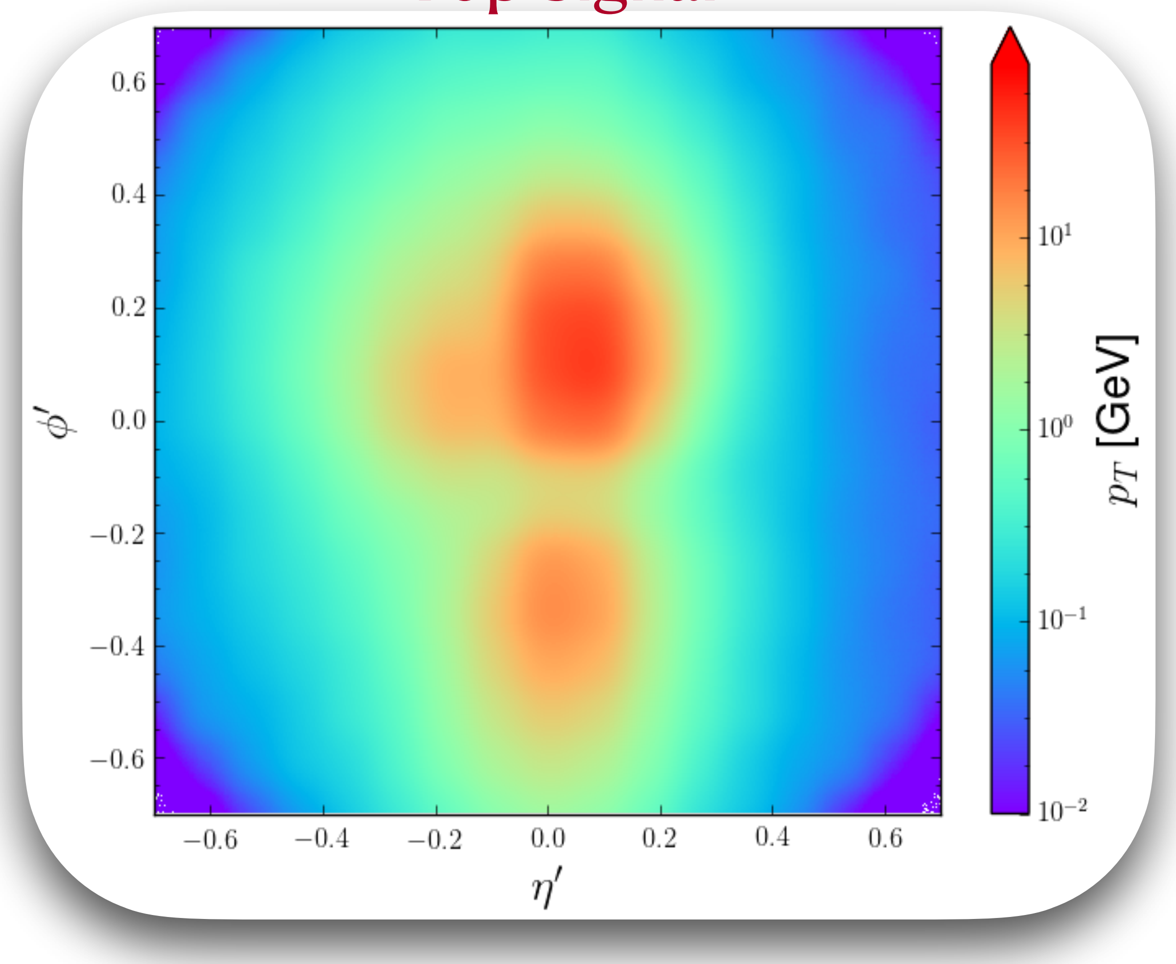
Adjust the bond dimension via SVD!

Initially proposed in: Stoudenmire, Schwab; arXiv:1605.05775

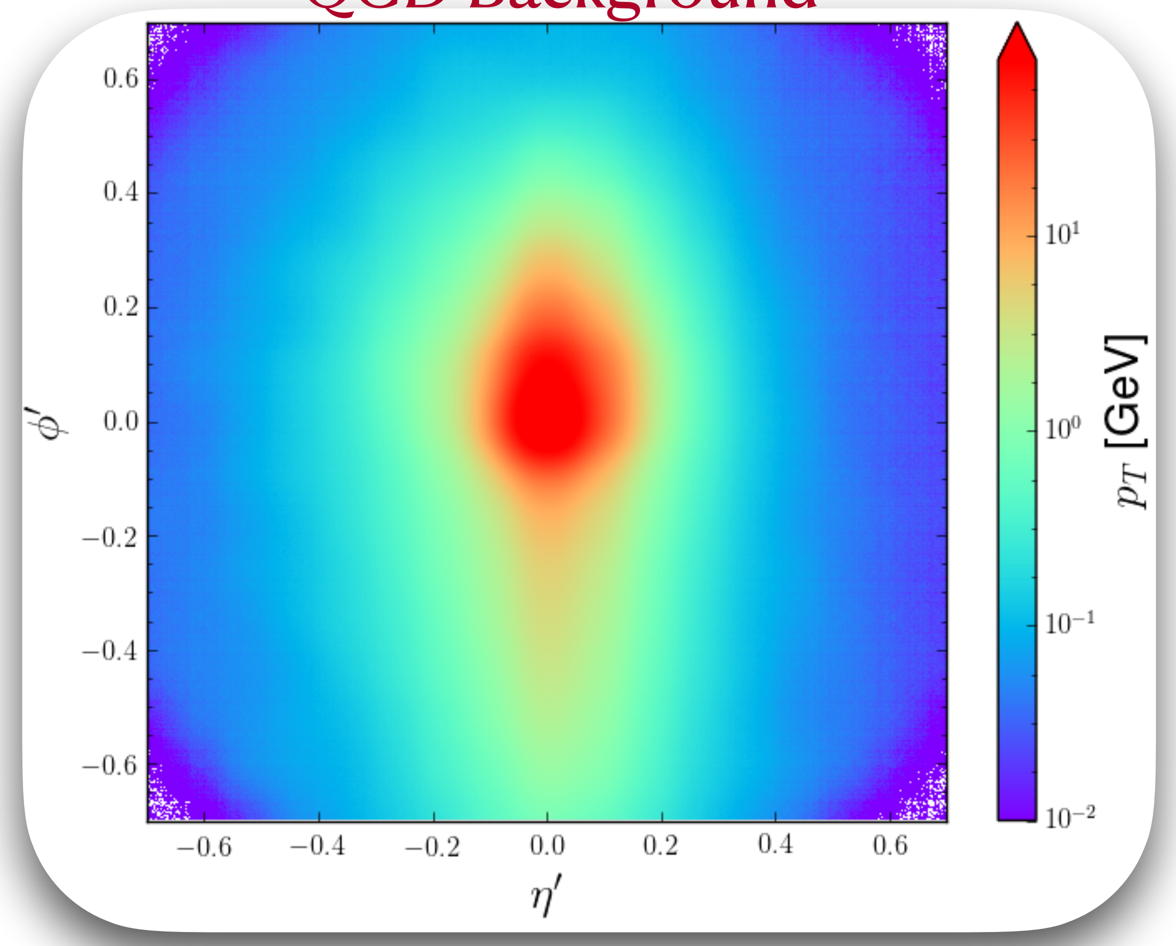
Top Tagging through Matrix Product States

Top Tagging through MPS

Top Signal



QCD Background



Data from: Kasieczka *et. al.* SciPost'19

- Leading FatJet Definition: anti-k_T algorithm with $R = 0.8, p_T \in [550,\ 650]$ GeV, $|\eta| < 2$
- Parton matching with $\Delta R(j, t_{truth}) < 0.8$
- Jets are centred with respect to p_T weighted centroid where jet vector is at $(\phi, \eta) = (0,0)$
- Principal axis has been rotated to $+\eta$ direction
- Energy deposits has been divided into 37×37 pixels which corresponds to η & $\phi \in [-1.5,\ 1.5]$.
- Image has been flipped to place the most energetic quadrant to the top right corner.

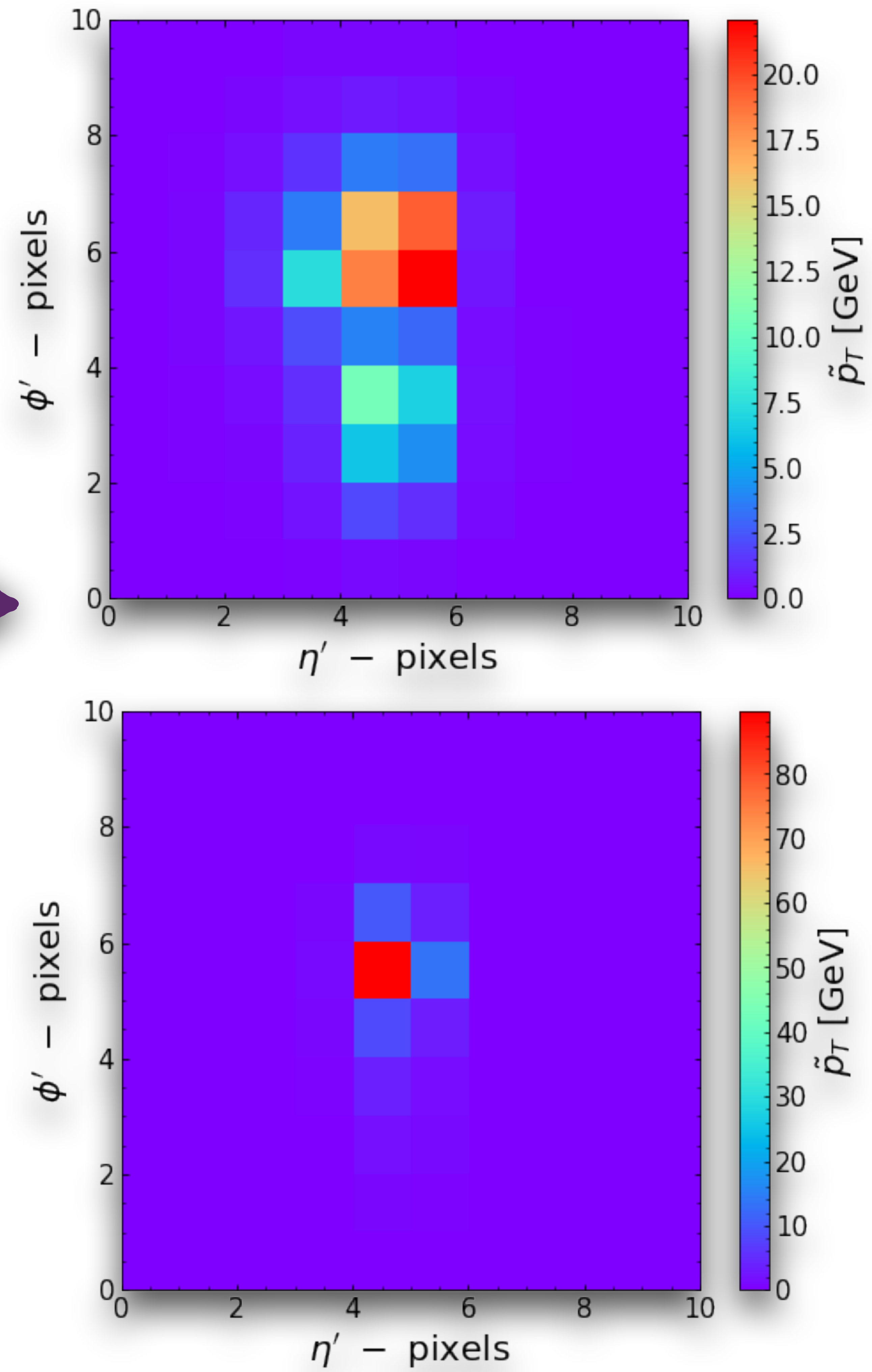
Similar preprocess, based on CNN:

Kasieczka, Plehn, Russell, Schell; JHEP '17

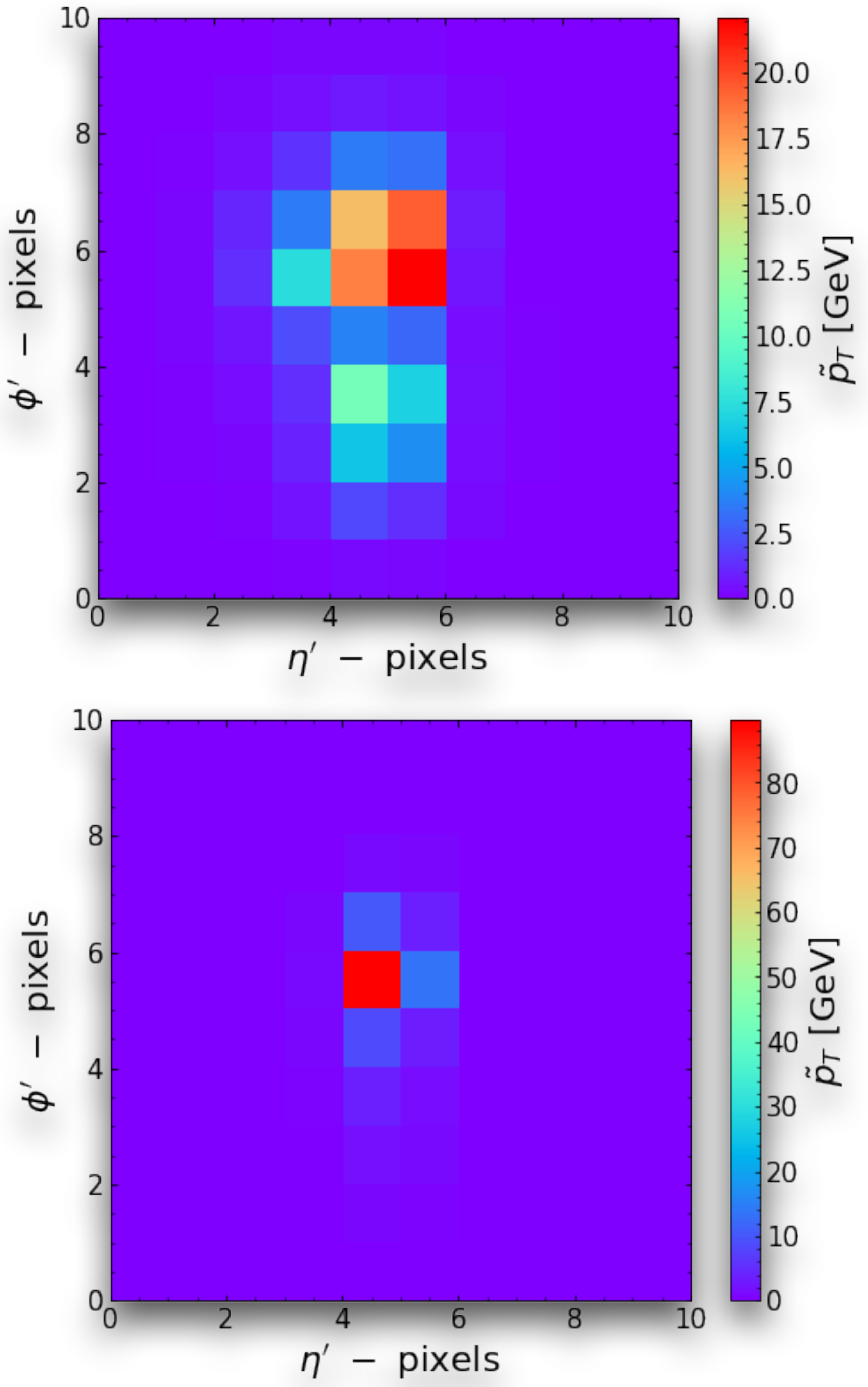
Macaluso, Shih; JHEP '18

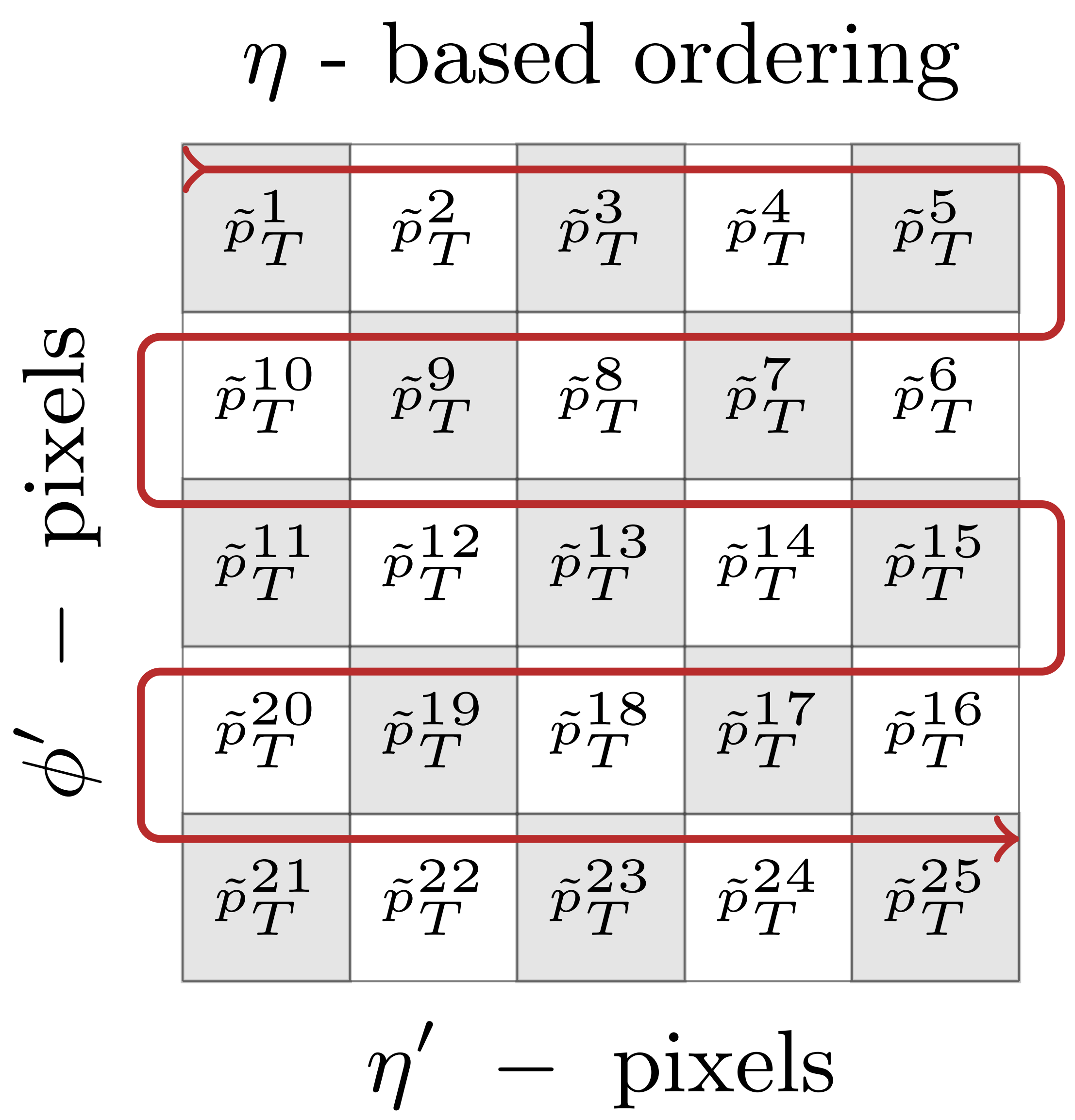
JYA, Spannowsky; JHEP '21

Crop & downsample



Top Tagging through MPS

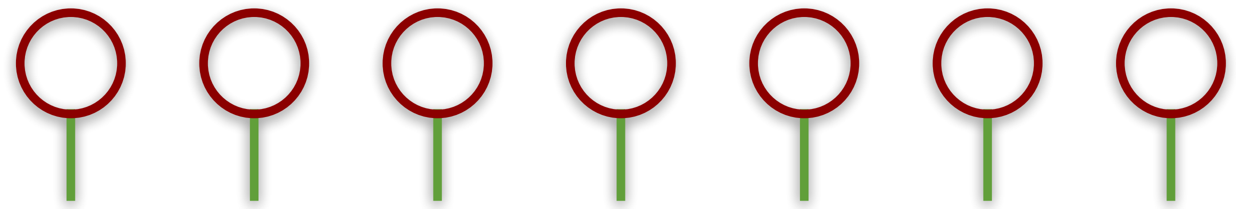




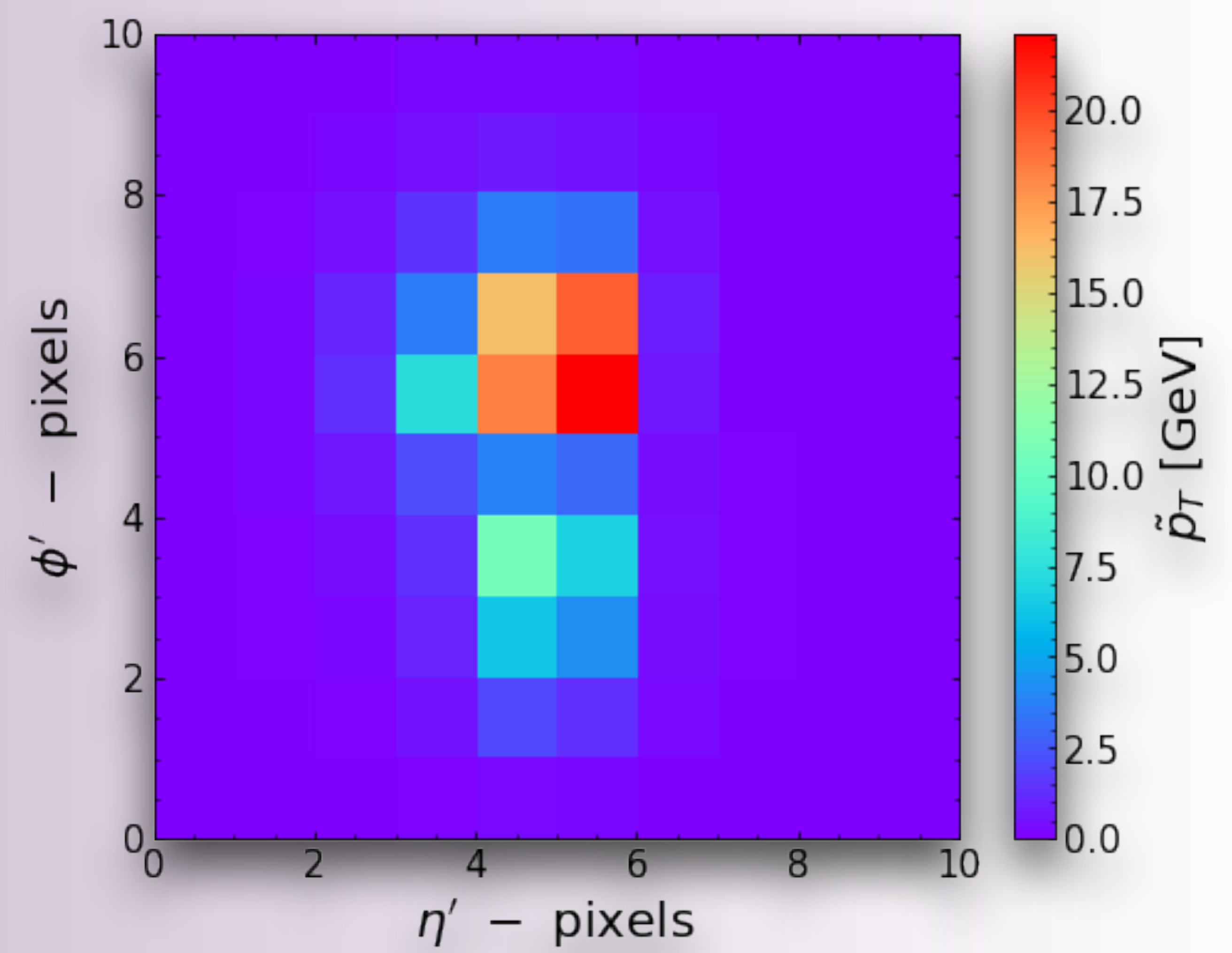
Data Embedding

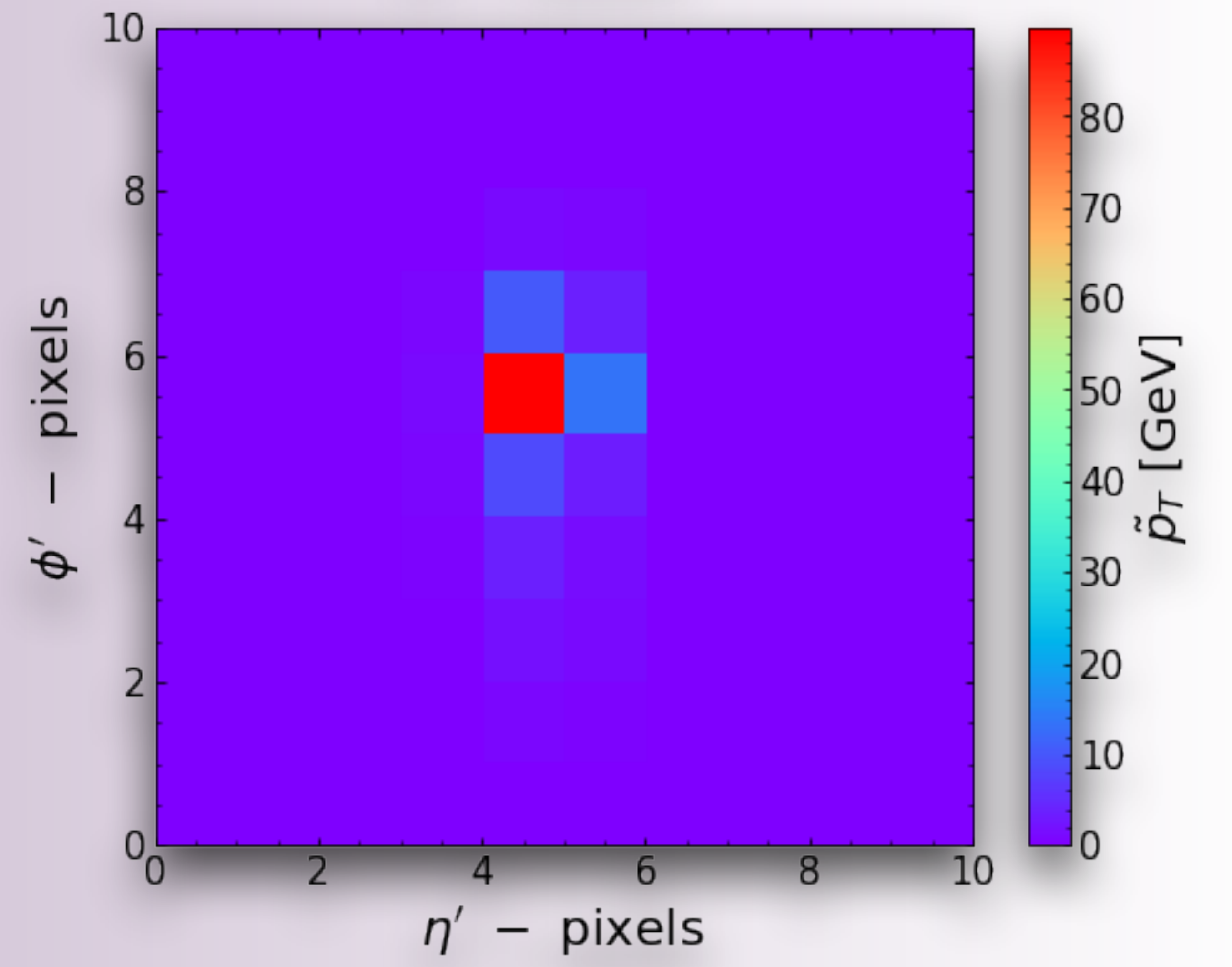
$$\Phi^{p_1\cdots p_n}(\mathbf{x}) = \phi^{p_1}(x_1) \otimes \phi^{p_2}(x_2) \otimes \cdots \otimes \phi^{p_n}(x_n)$$

$$\phi^{p_i}(\tilde{p}_T^i) = \sqrt{\binom{D-1}{p_i-1}} \cos^{D-p_i}\left(\tilde{p}_T^i \frac{\pi}{2}\right) \sin^{p_i-1}\left(\tilde{p}_T^i \frac{\pi}{2}\right)$$

 $= \Phi^{p_1\cdots p_n}(\mathbf{x})$

Top Tagging through MPS



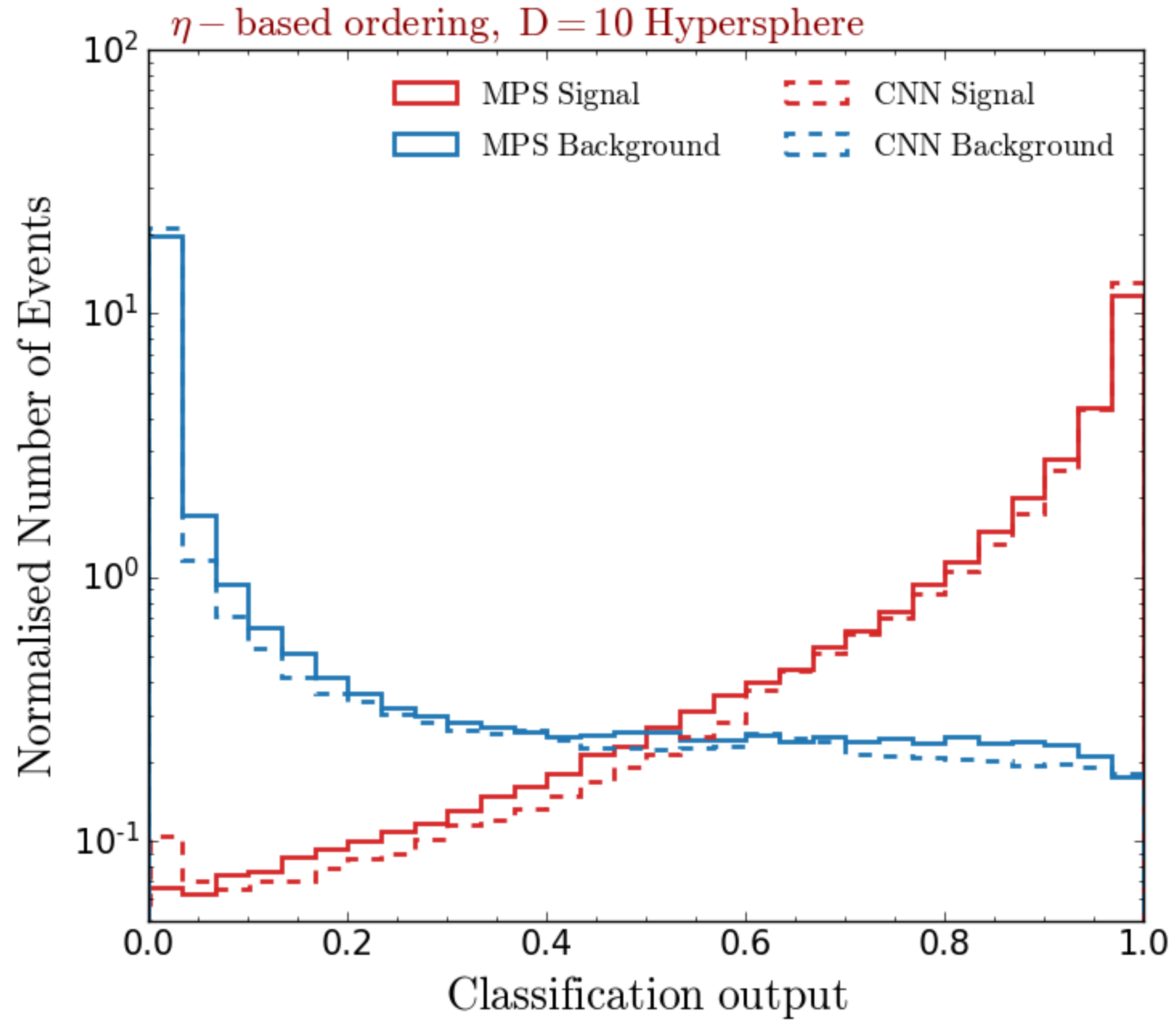


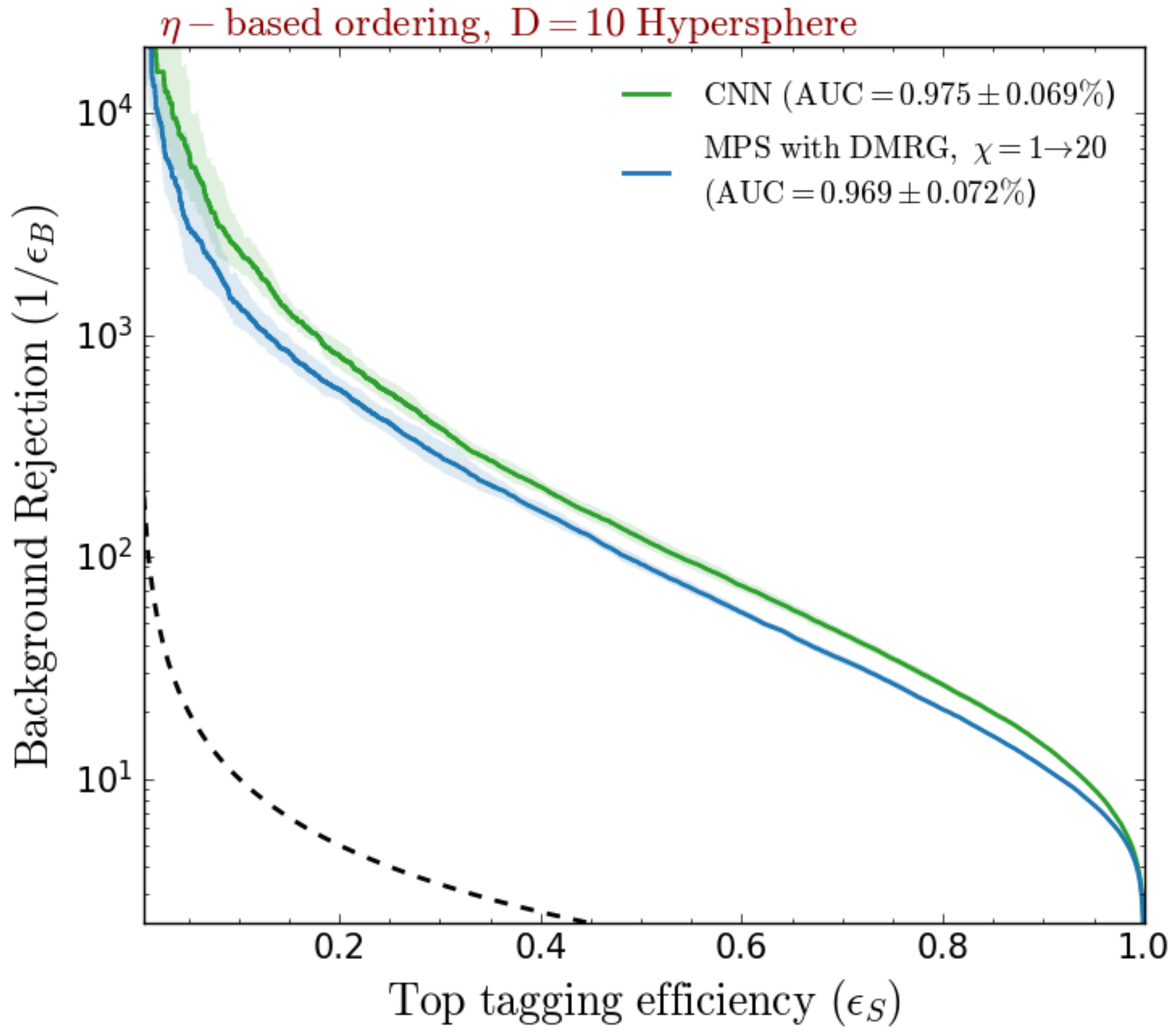
Assumptions & Requirements

- No prior entanglement/correlation between pixels, $\chi_{\text{init}} = 1$
- Network is a Born Machine → square of the wave-function gives the probability of the classification.
- Maximum bond dimension that network can get is 20.

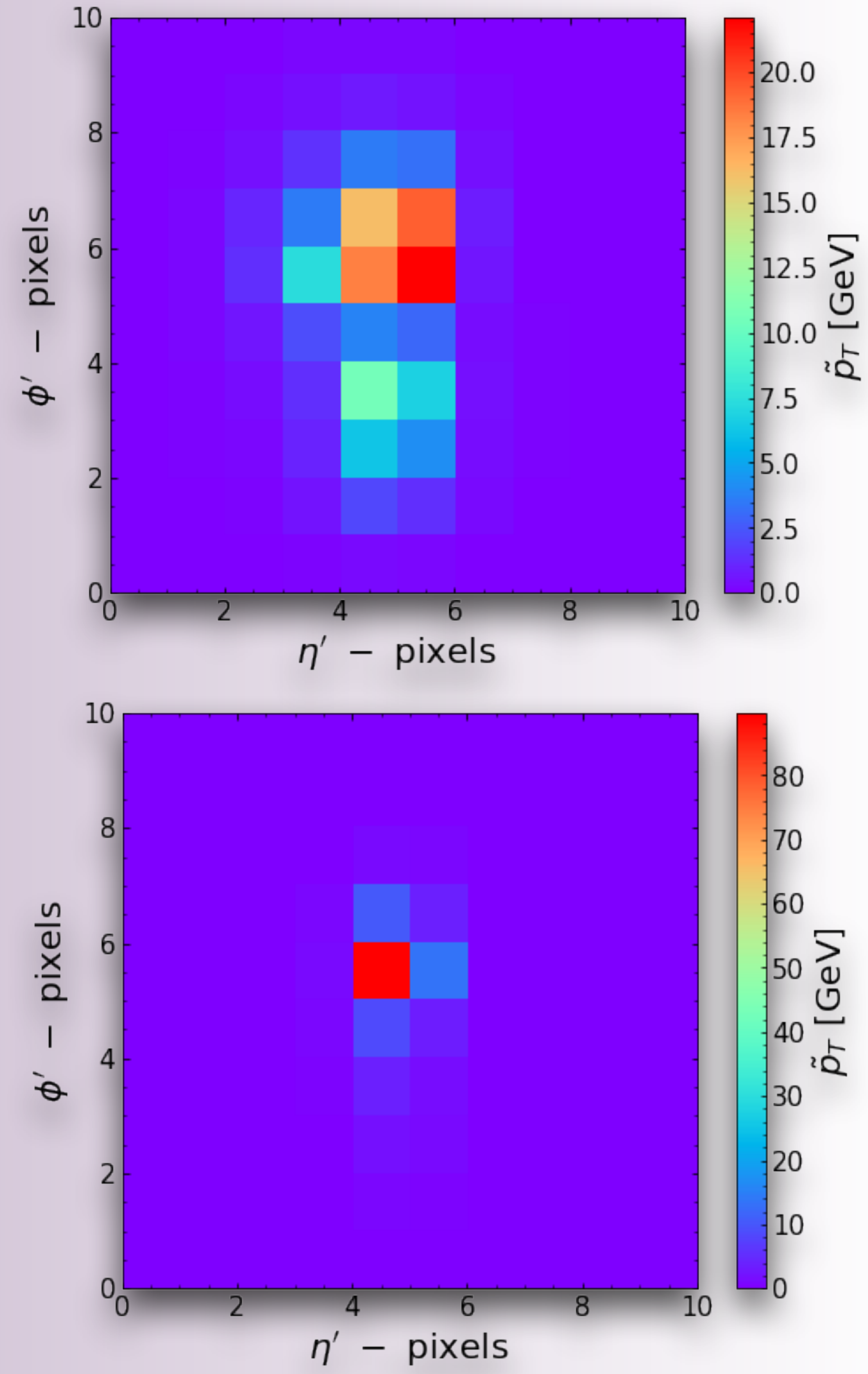
CNN architecture from:

JYA, Spannowsky; JHEP '21

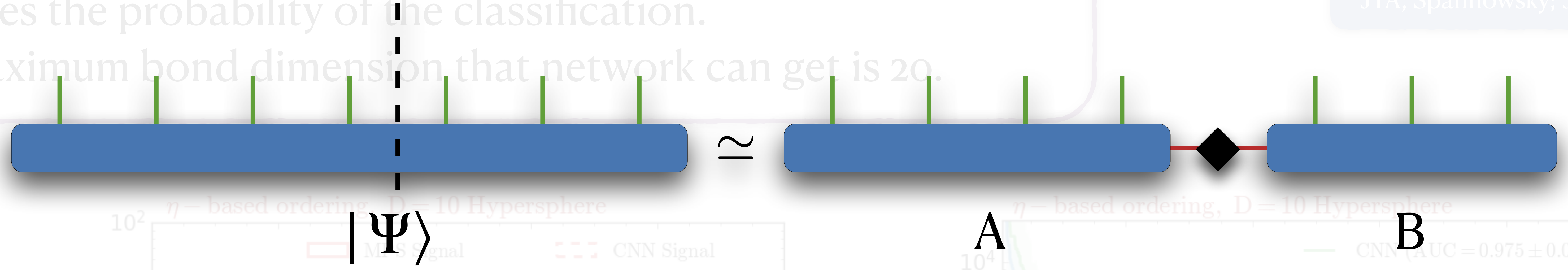




Top Tagging through MPS



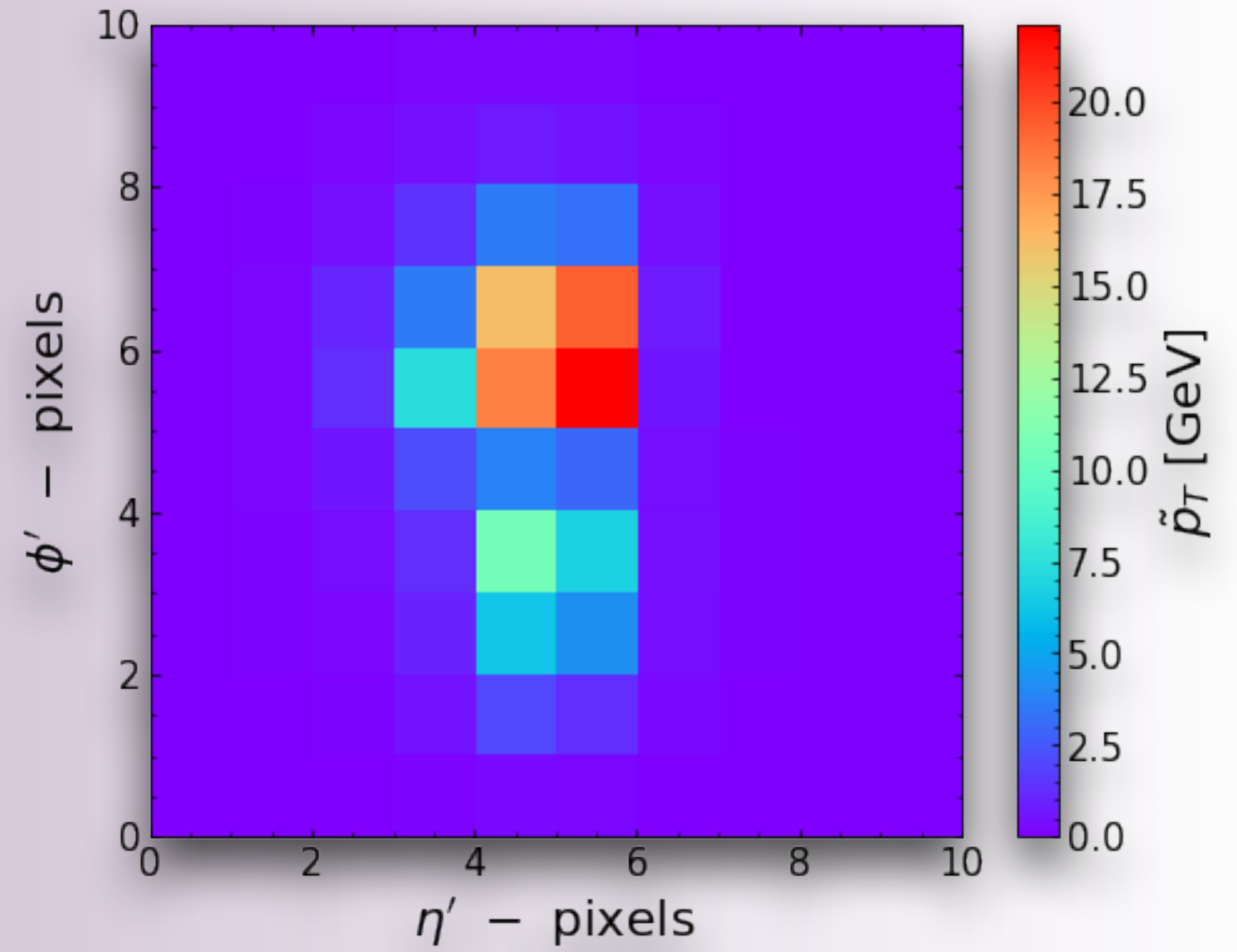
Entanglement Entropy & Schmidt Decompostion

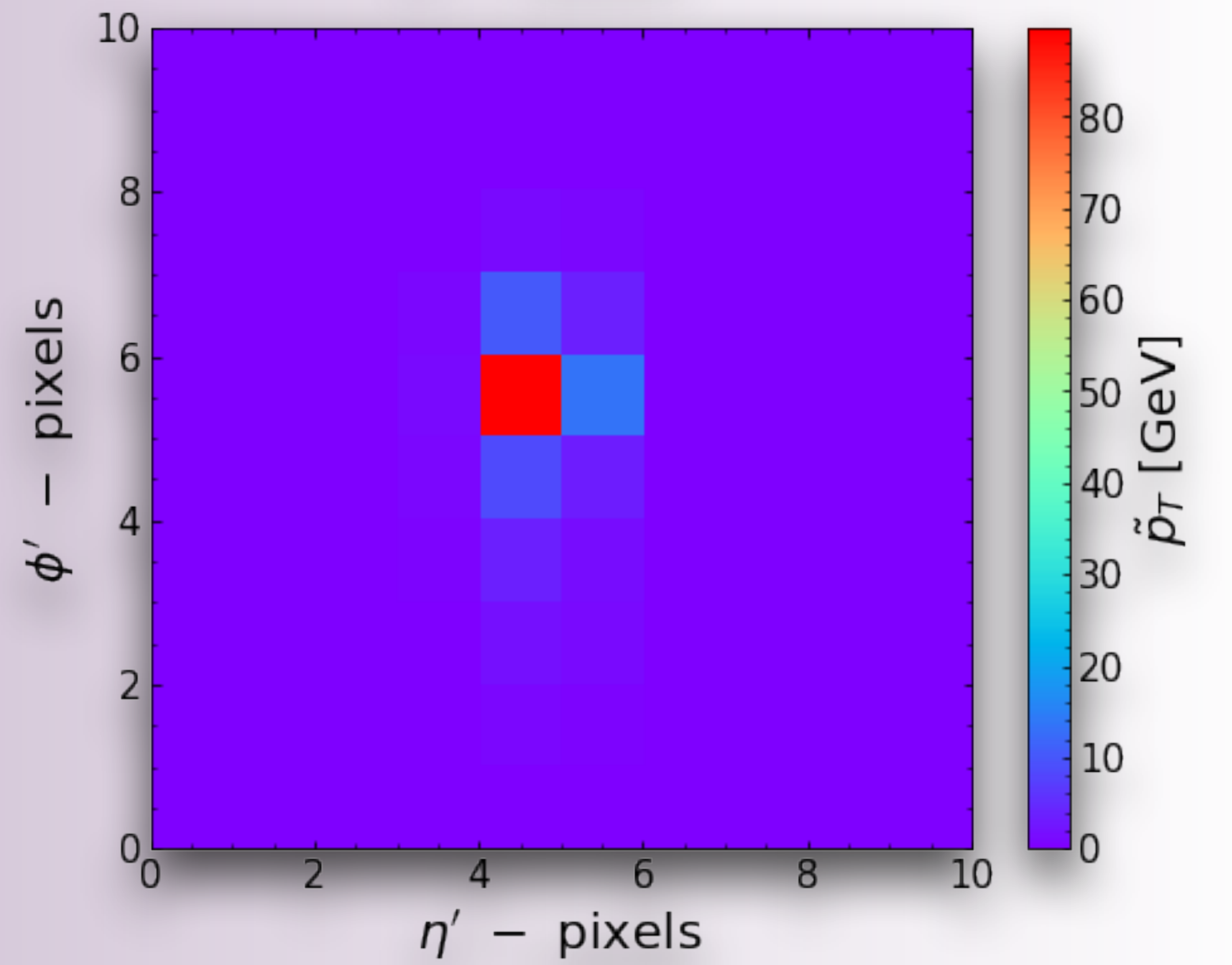


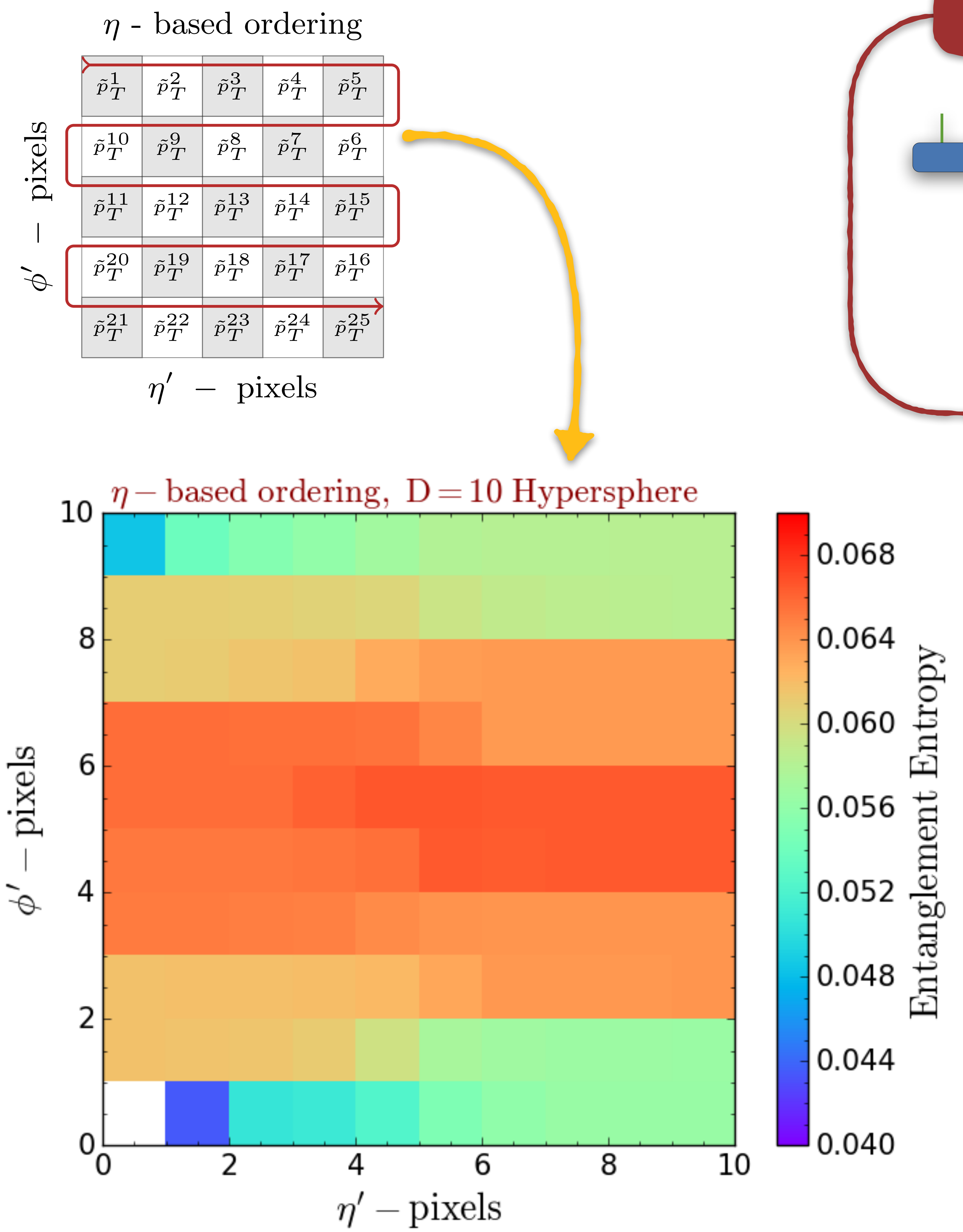
$$|\Psi\rangle_{AB} = \sum_i^{\chi} \lambda_\alpha |\alpha\rangle_A |\alpha\rangle_B$$

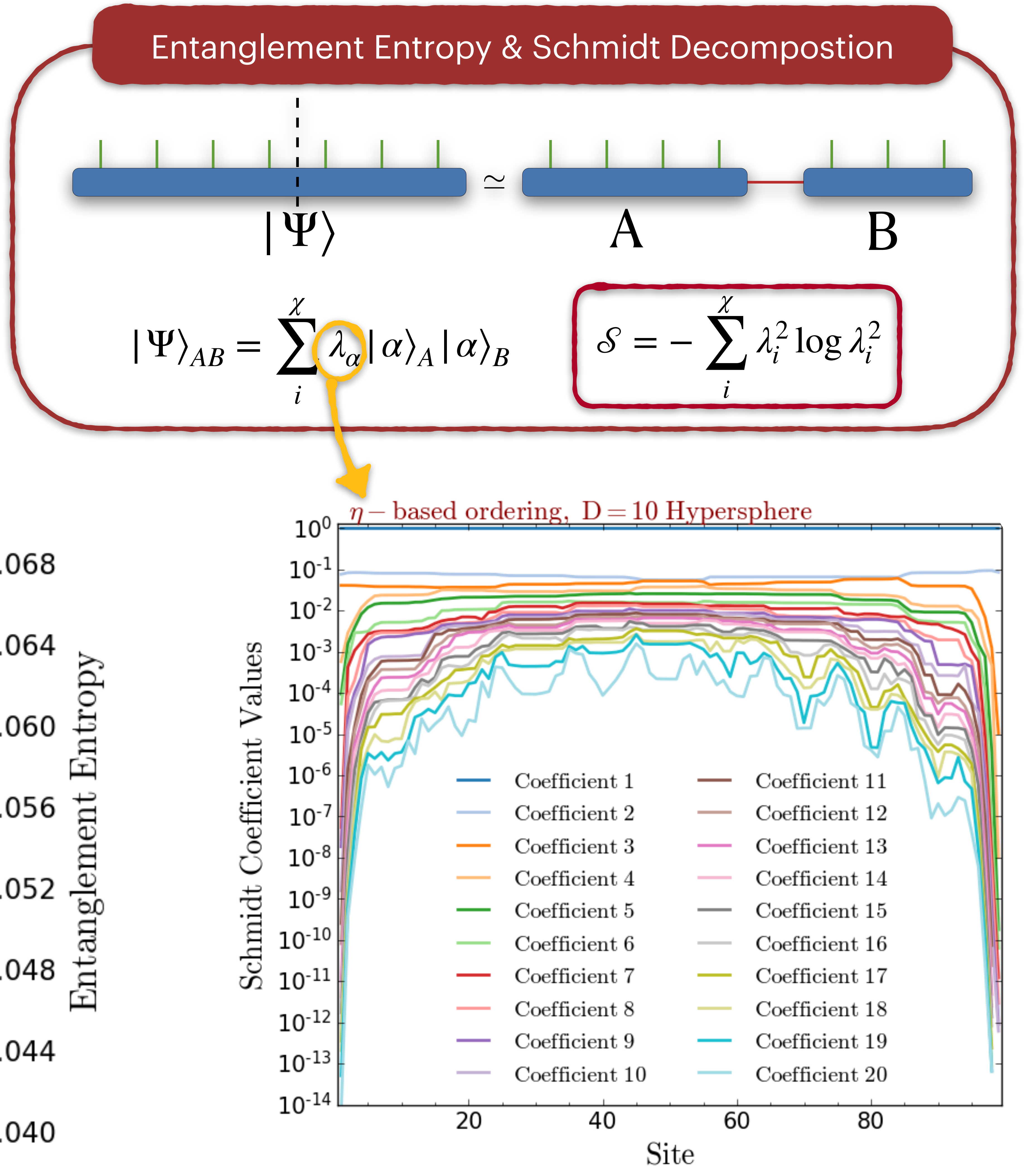
$$\mathcal{S} = -\sum_i^{\chi} \lambda_i^2 \log \lambda_i^2$$

Top Tagging through MPS

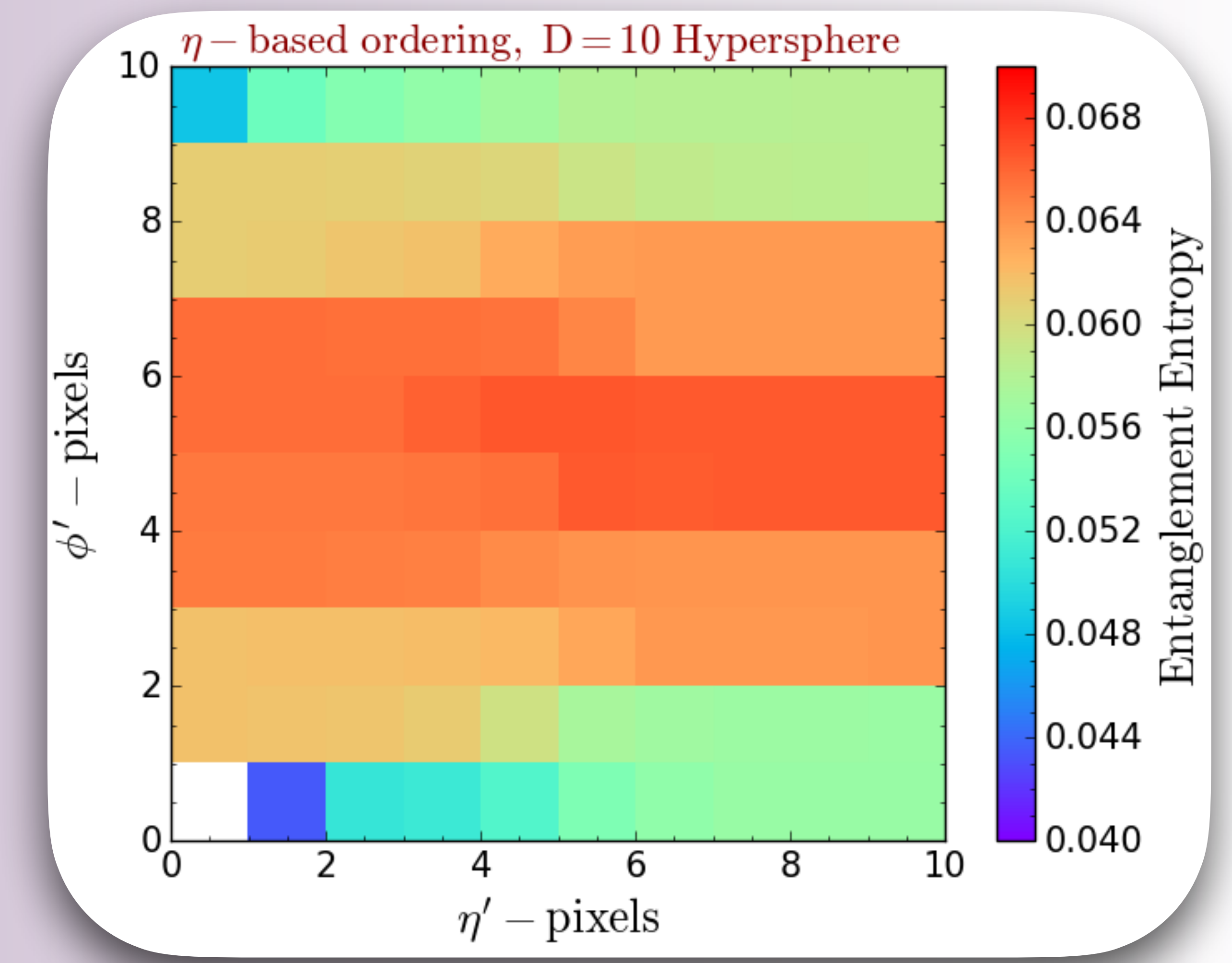


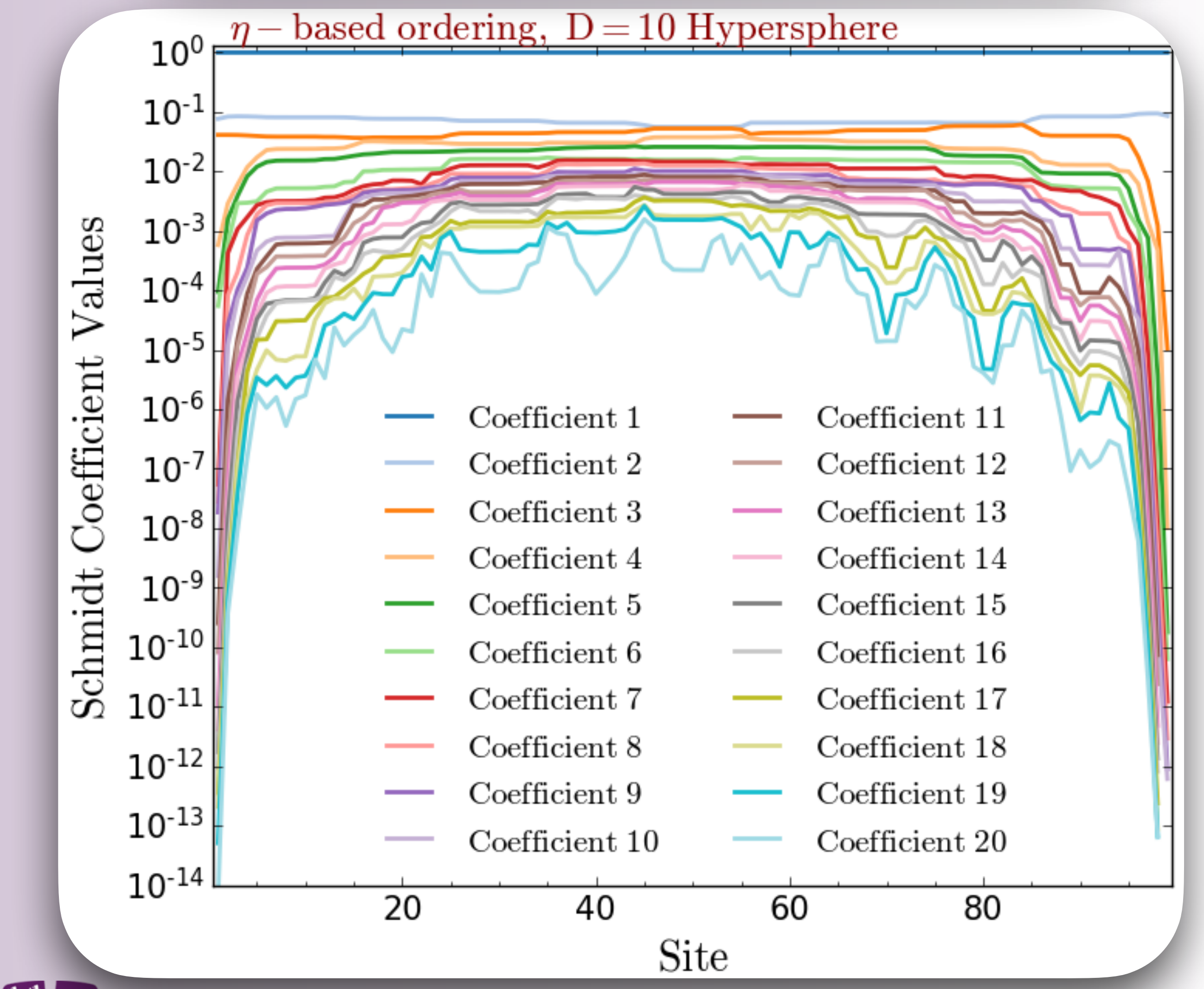


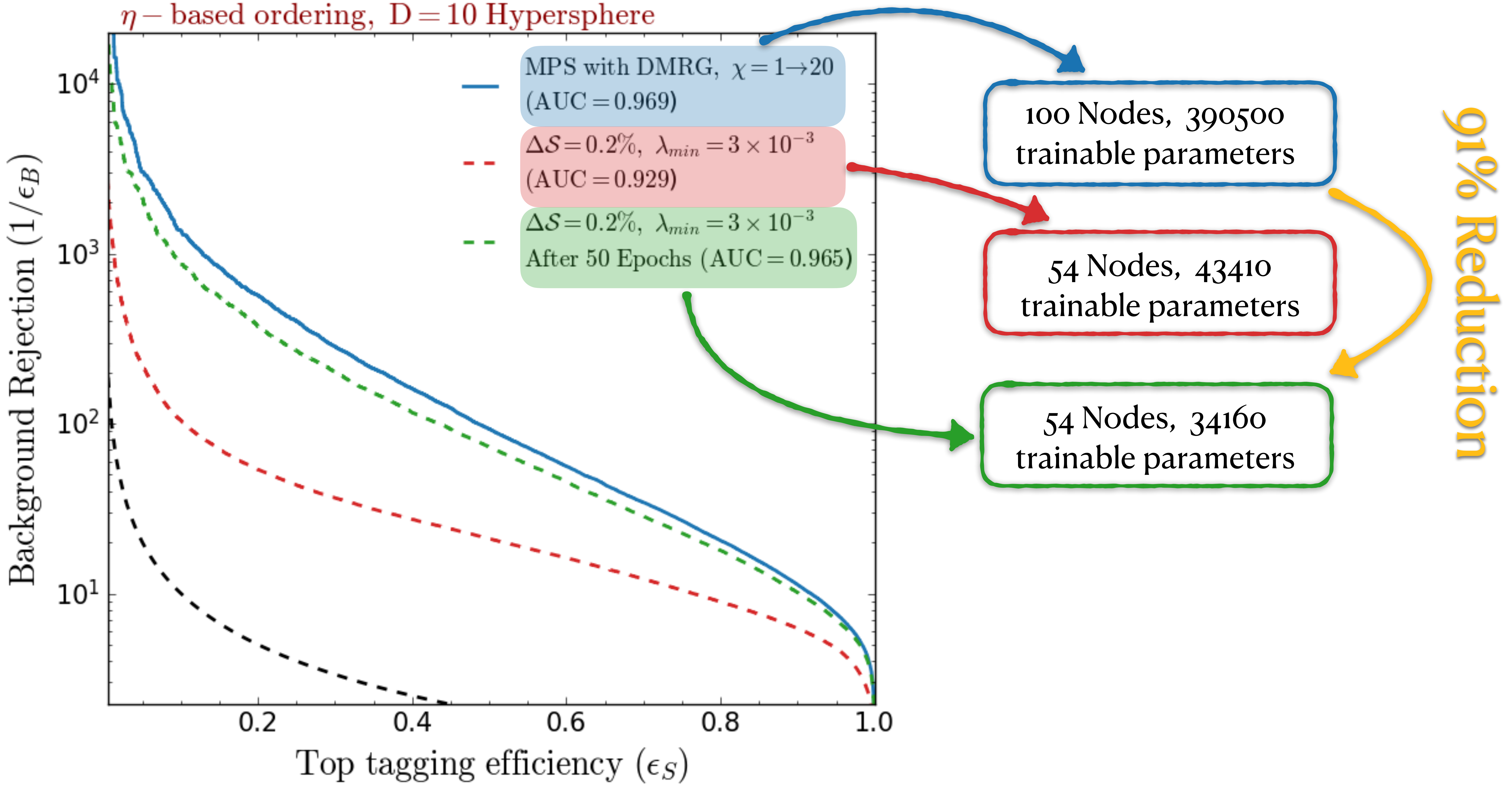




Top Tagging through MPS

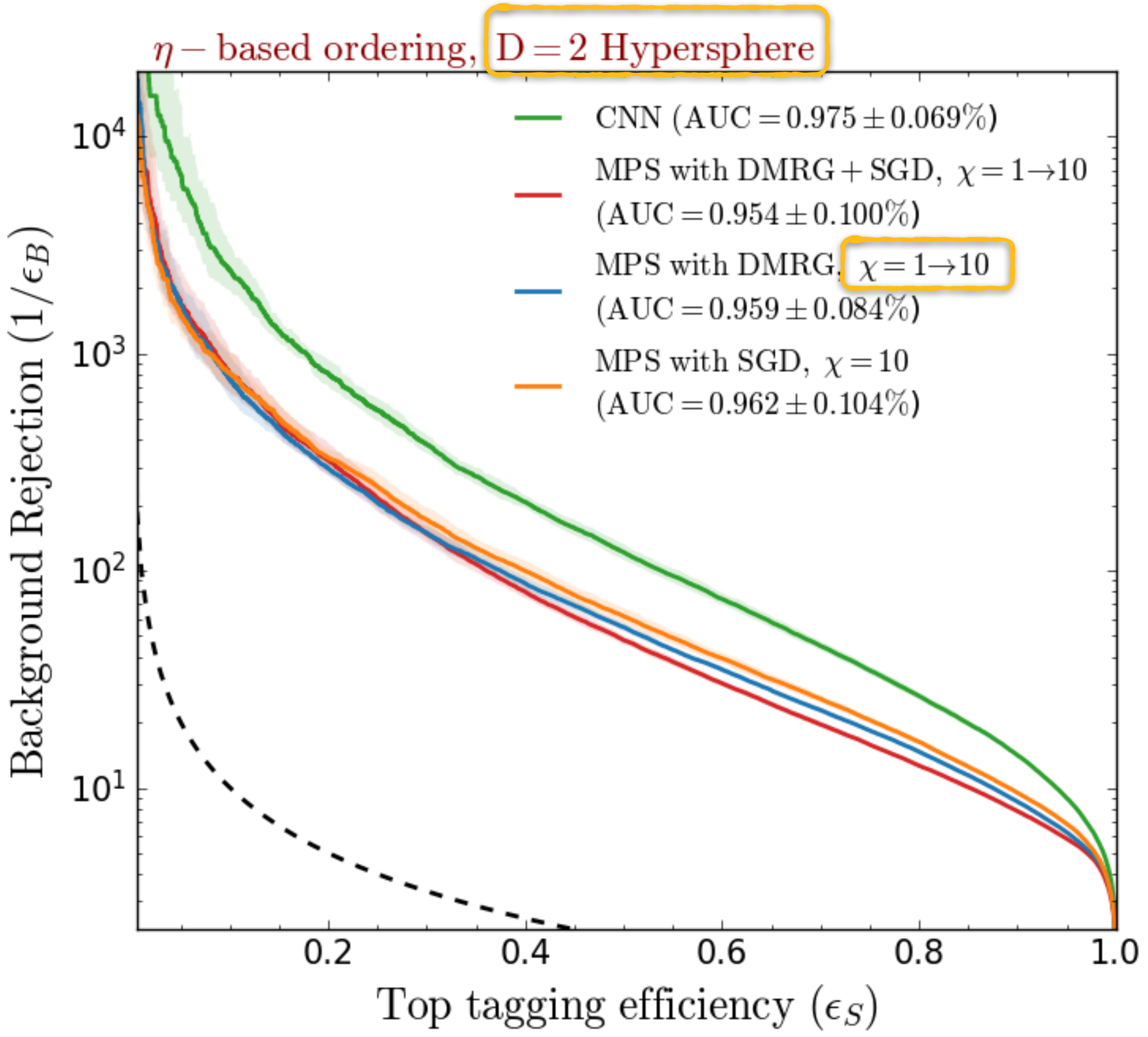






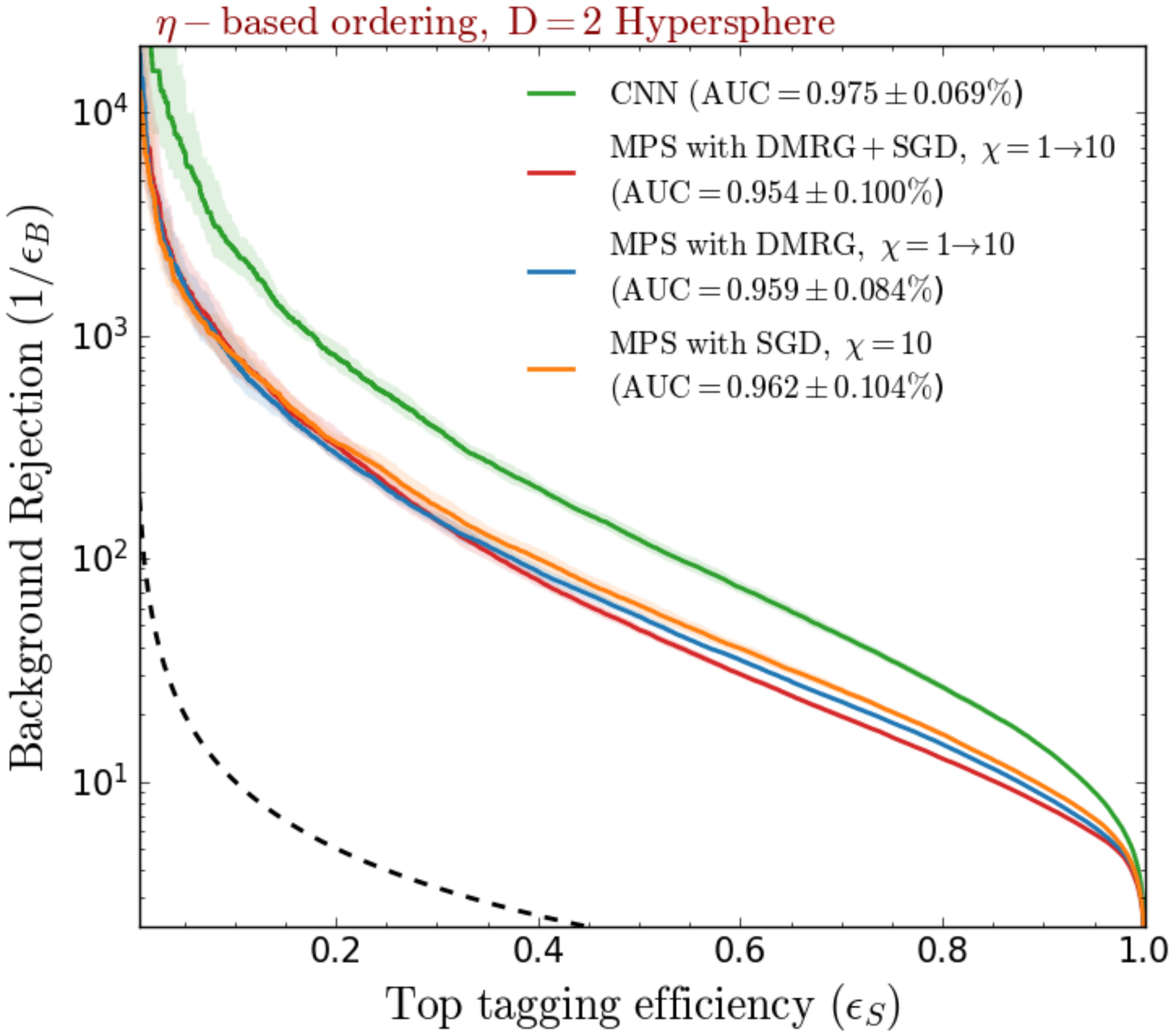
Durham University

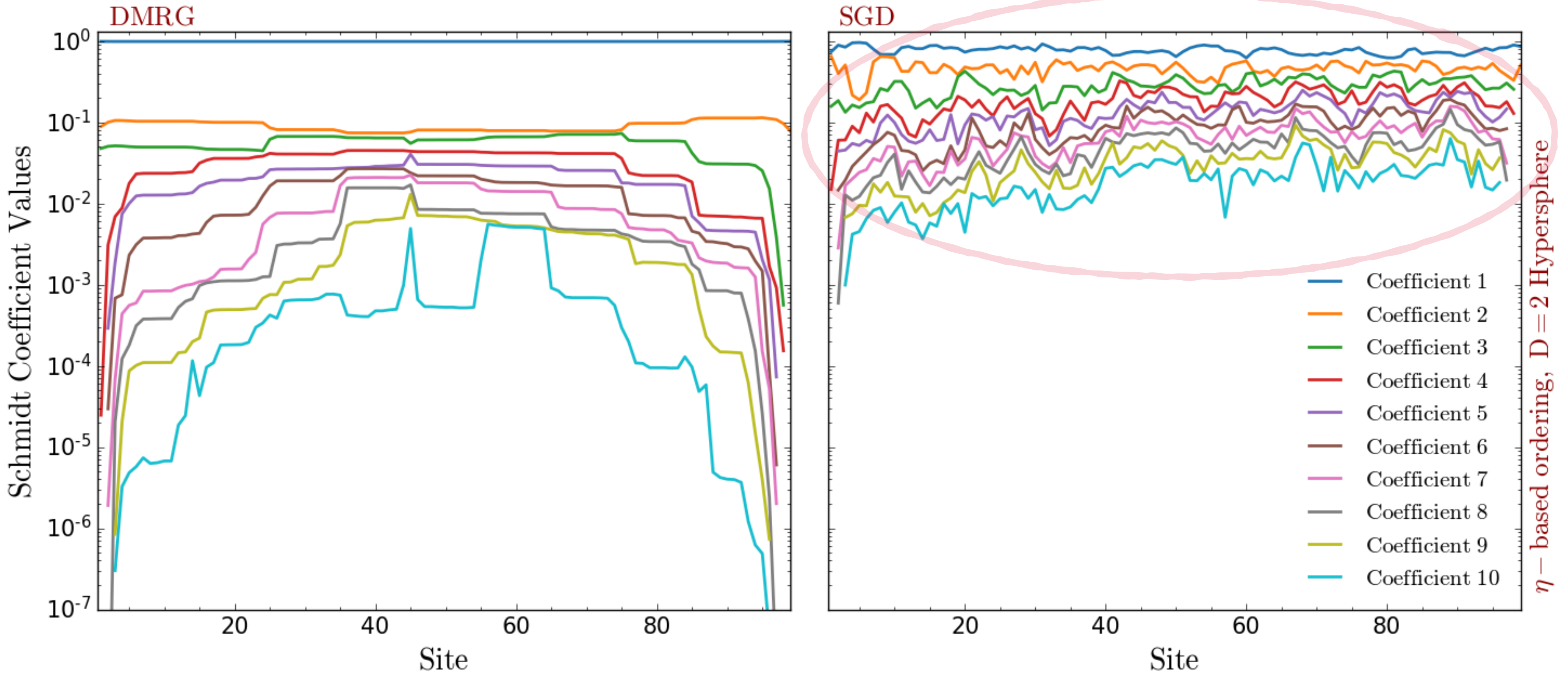
How about good old SGD?

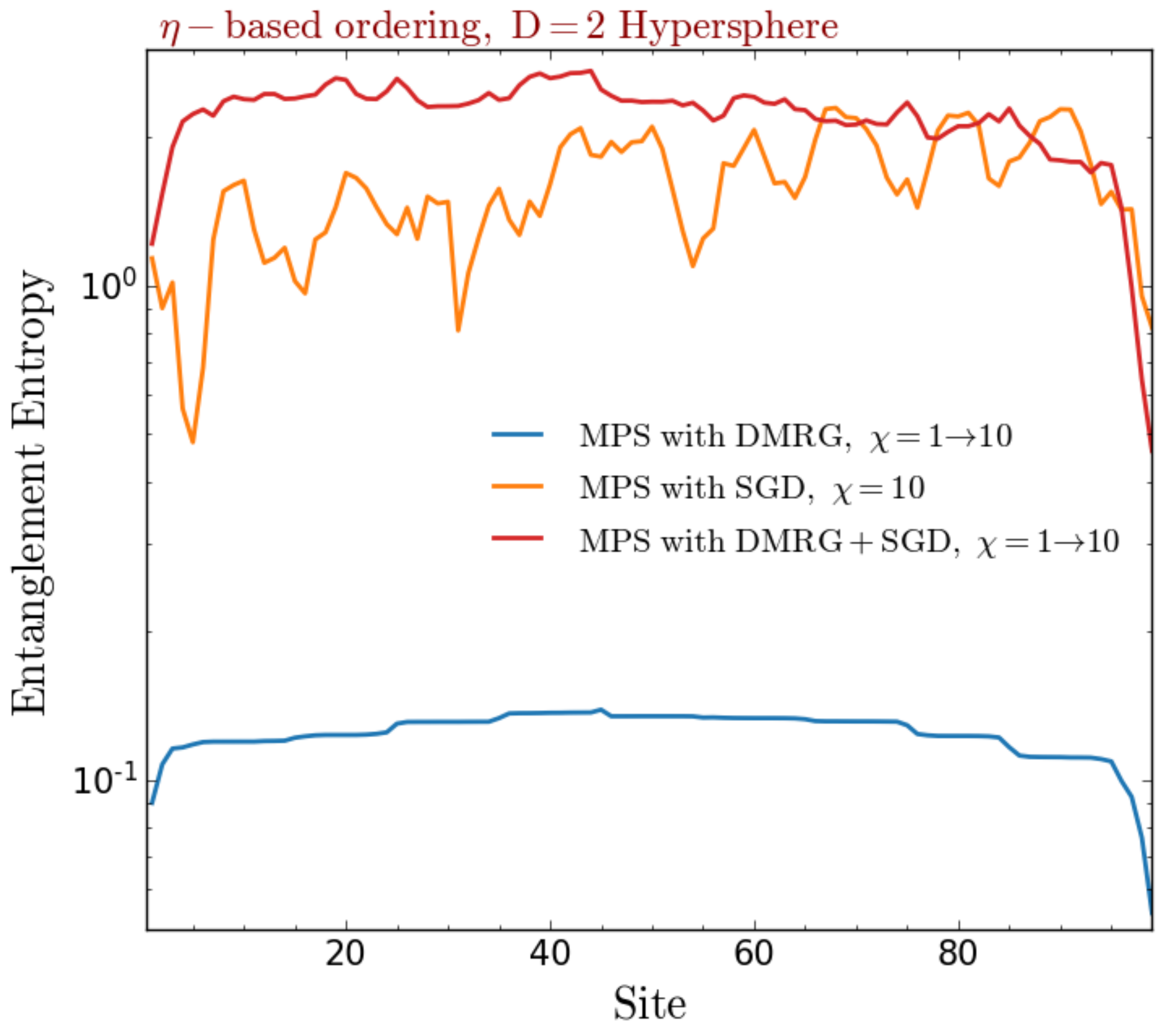


- SGD: As in usual NN back-propagation all tensors are updated simultaneously. All gradient tensors are normalized before the update!
- DMRG+SGD: Each epoch started with 3 DMRG sweeps on first batch and the rest of the epoch trained by standard SGD.

How about good old SGD?







Conclusion

Conclusion

- Tensor Networks opens up the entire world of techniques developed for Quantum Mechanics to the Machine Learning applications.
- A linear network allows easier interpretation.
- Perfect tool to do linear algebra in higher-dimensional spaces.

Main Drawbacks

- Cost to train can be high
- Choice of architecture is still a research area.

Advantages

☑ Interpretability
☑ Understanding
☑ Theory

Conclusion

- Tensor Networks opens up the entire world of techniques developed for Quantum Mechanics to the Machine Learning applications.
- A linear network allows easier interpretation.
- Perfect tool to do linear algebra in higher-dimensional spaces.

Next Steps

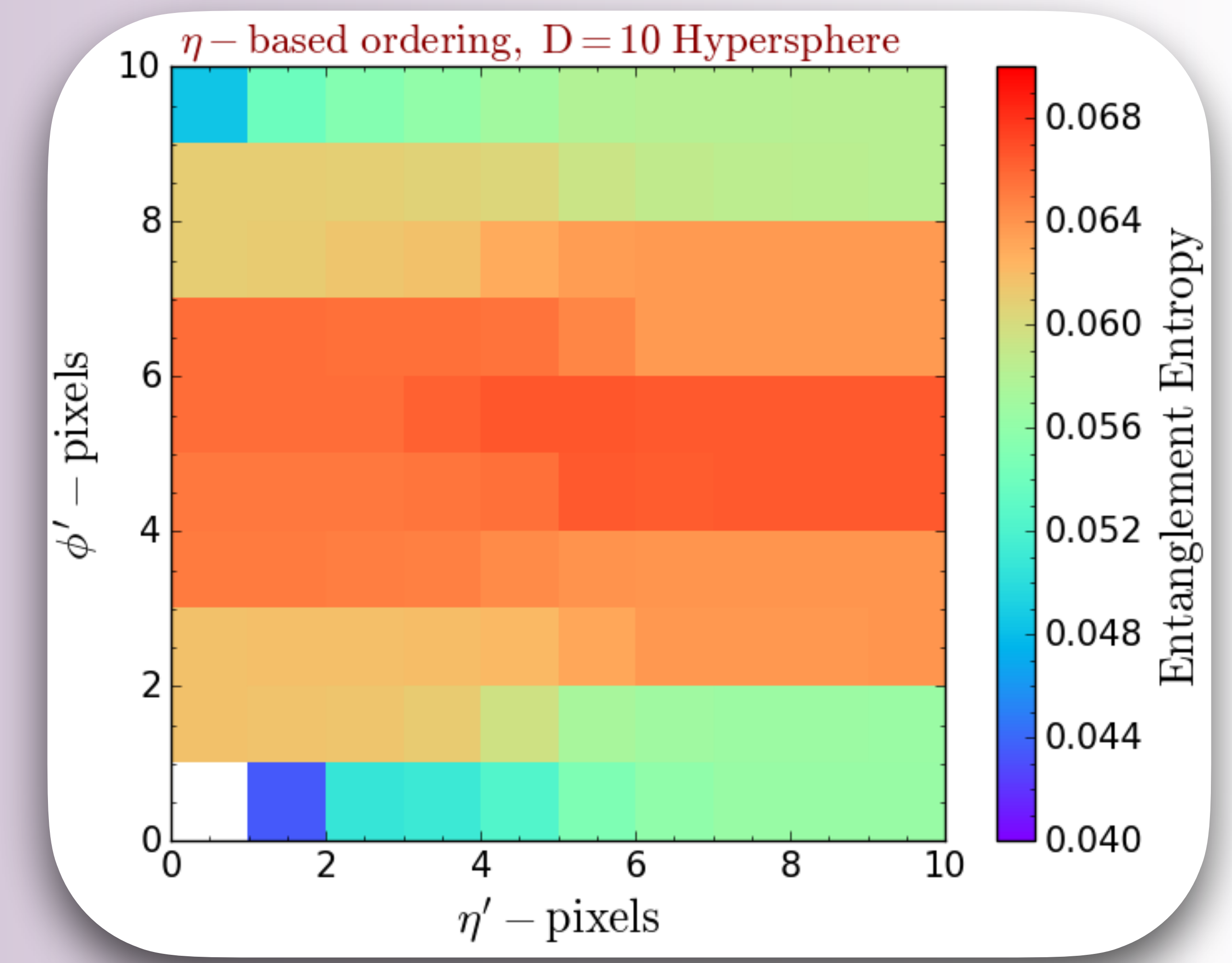
- PEPS: Classification with 2D systems. Some major progress only recently released!

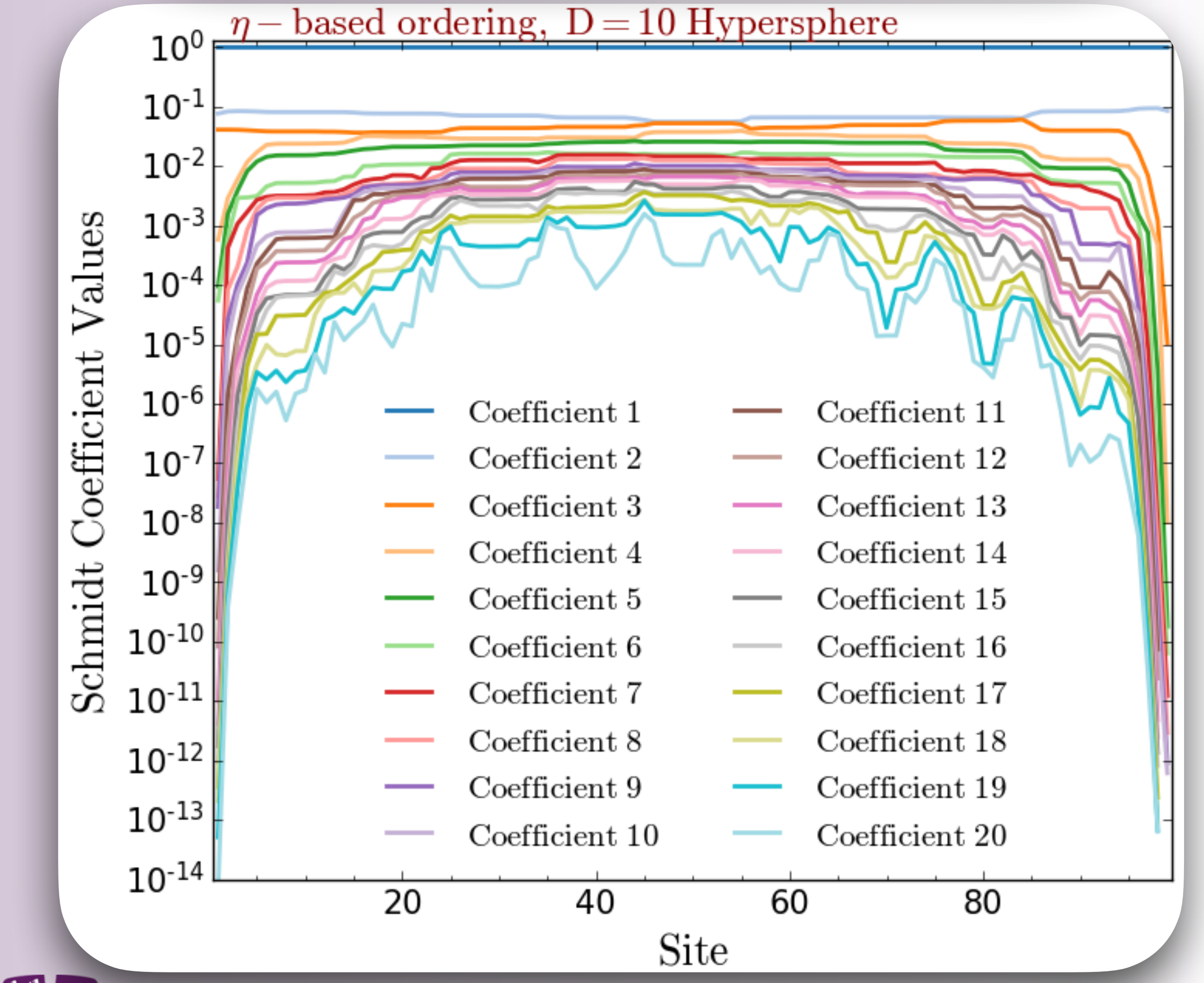
Rakhshan, Rabusseau arXiv: 2003.05101 Zaletel, Pollmann PRL '20

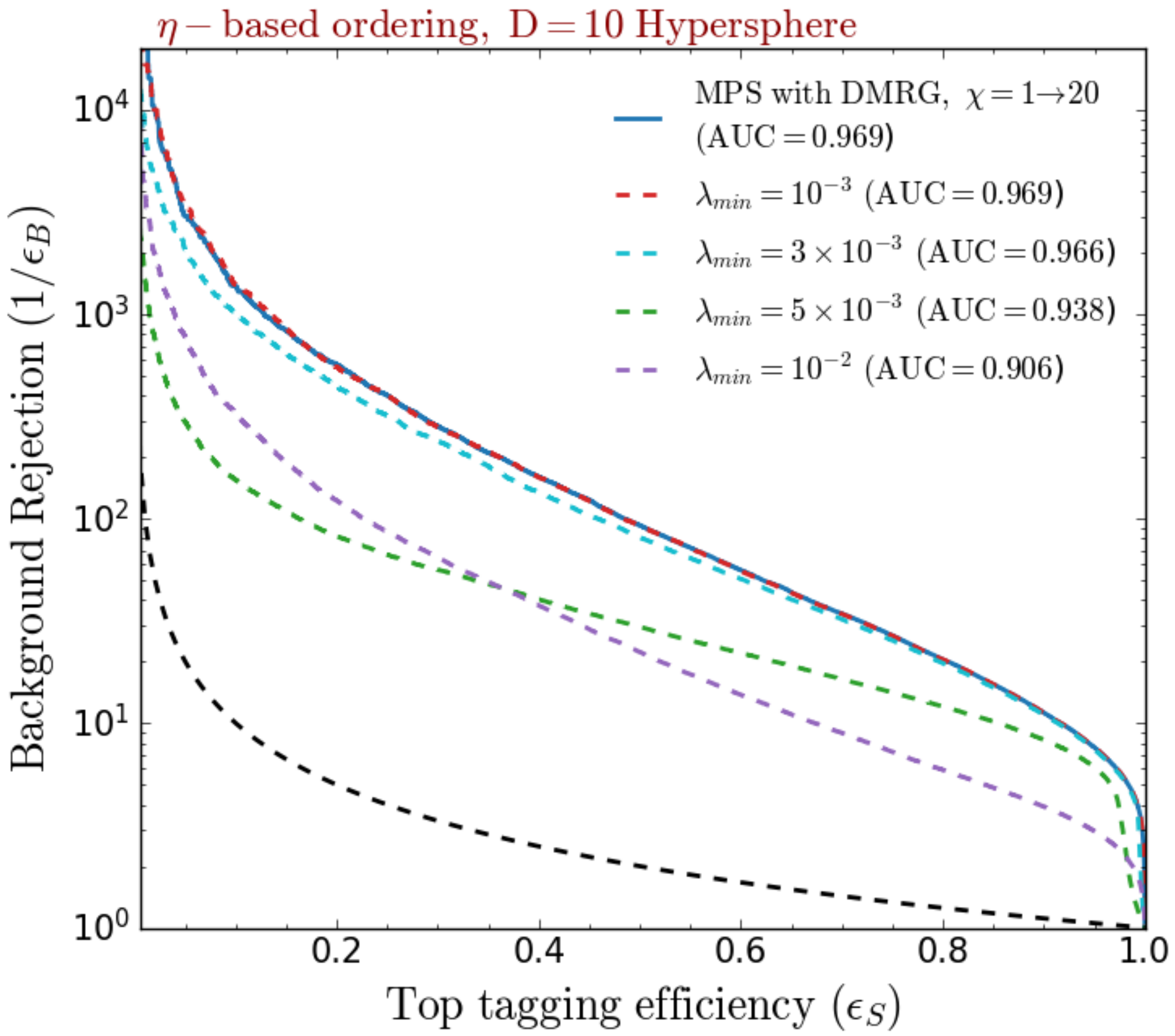
- Many different MPS-based architecture can be explored.
- Specialized algorithm for understanding data better!

BACKUP

Top Tagging through MPS

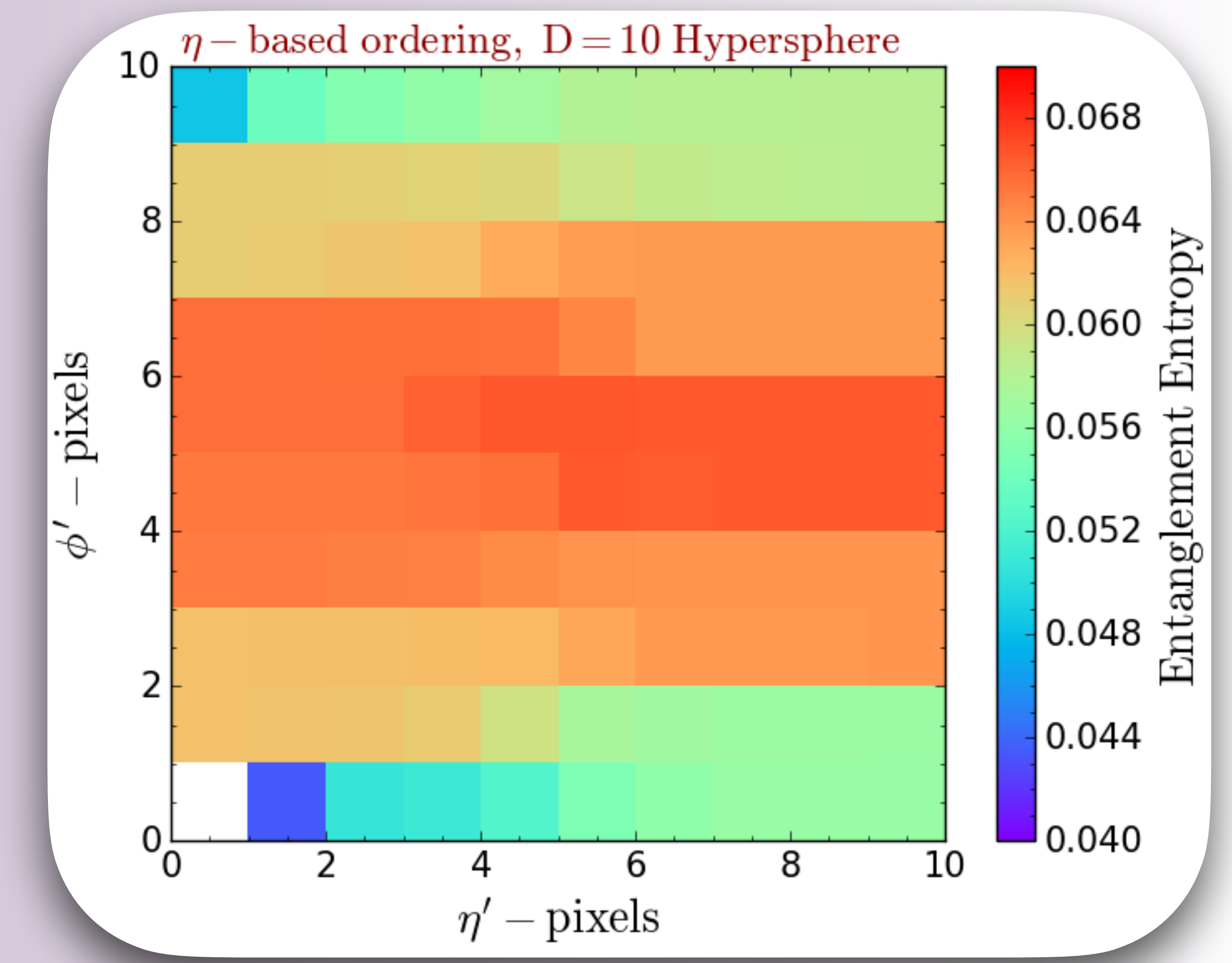


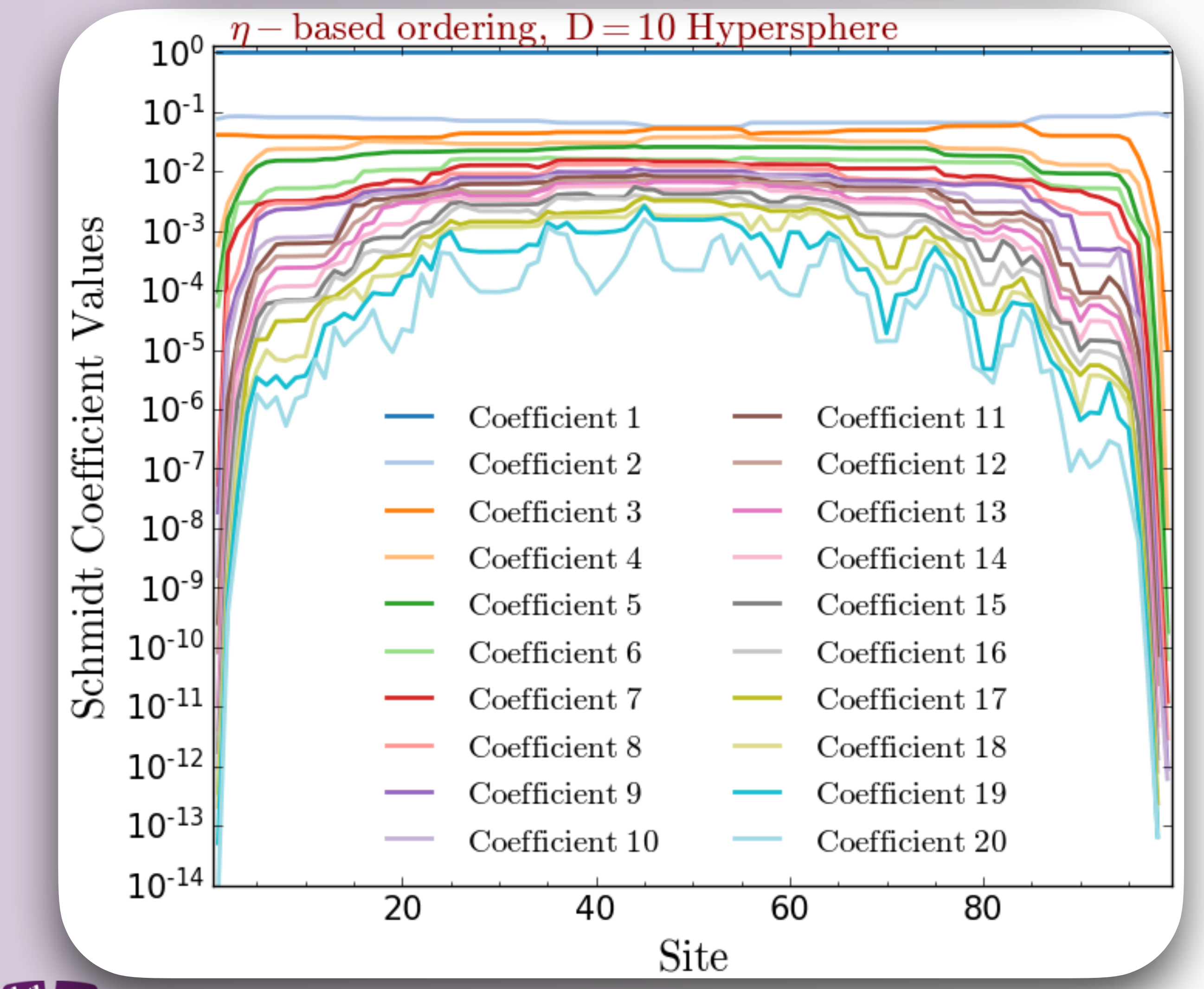


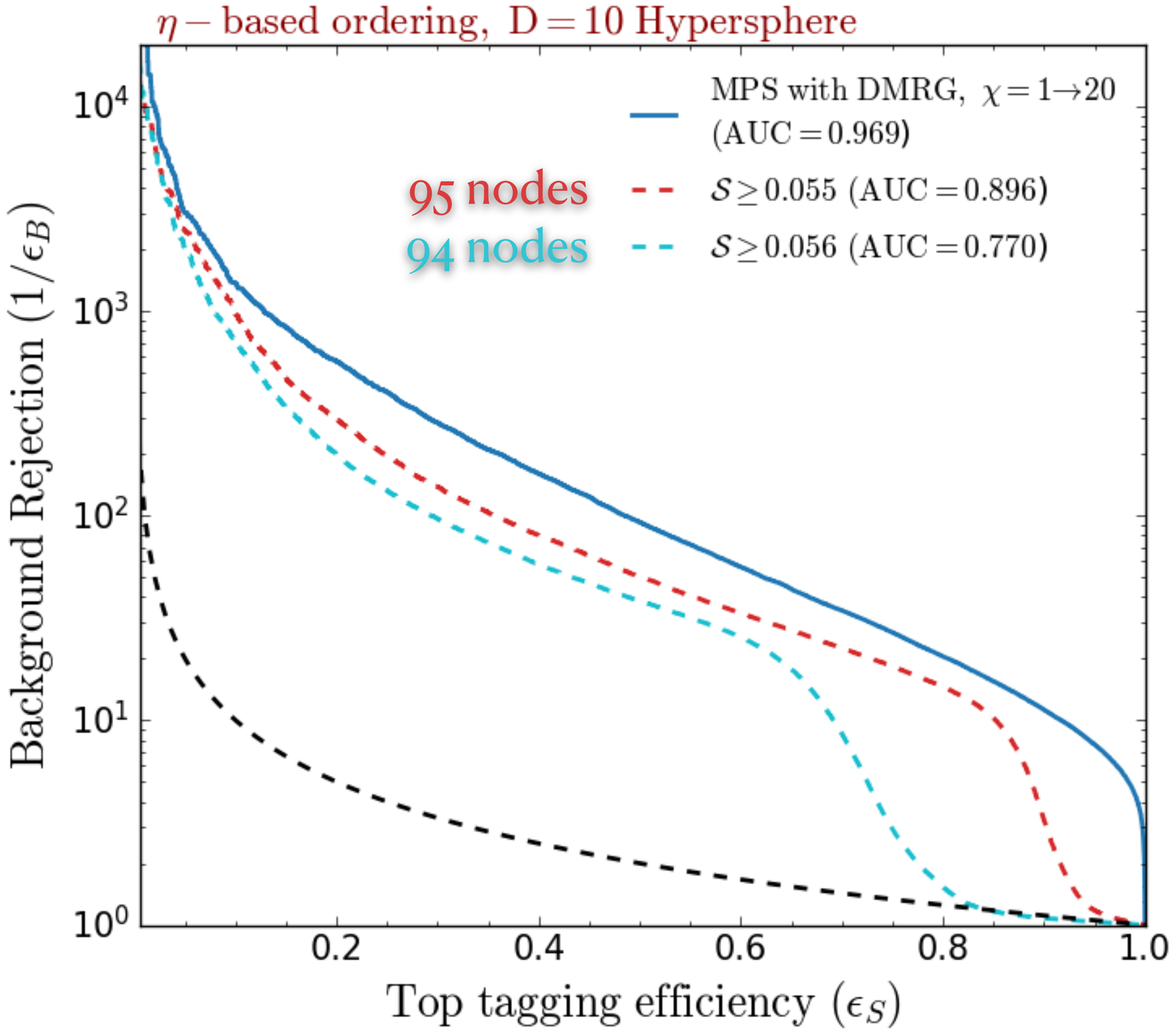


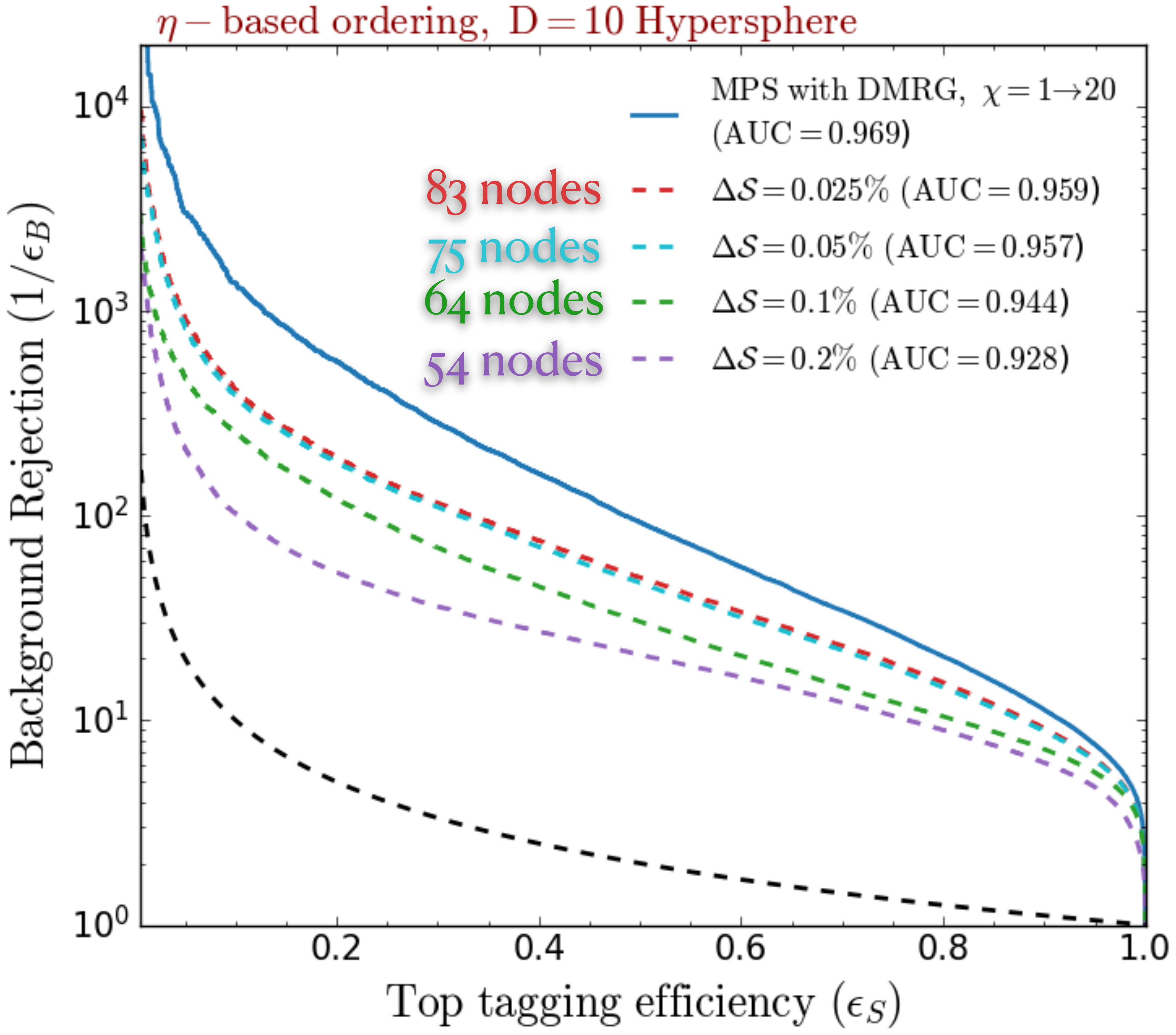
Model	Number of trainable parameters
Original MPS	390500
$\lambda \geq 10^{-3}$	204310
$\lambda \geq 3 \times 10^{-3}$	91690
$\lambda \geq 5 \times 10^{-3}$	32990
$\lambda \geq 10^{-2}$	18020

Top Tagging through MPS

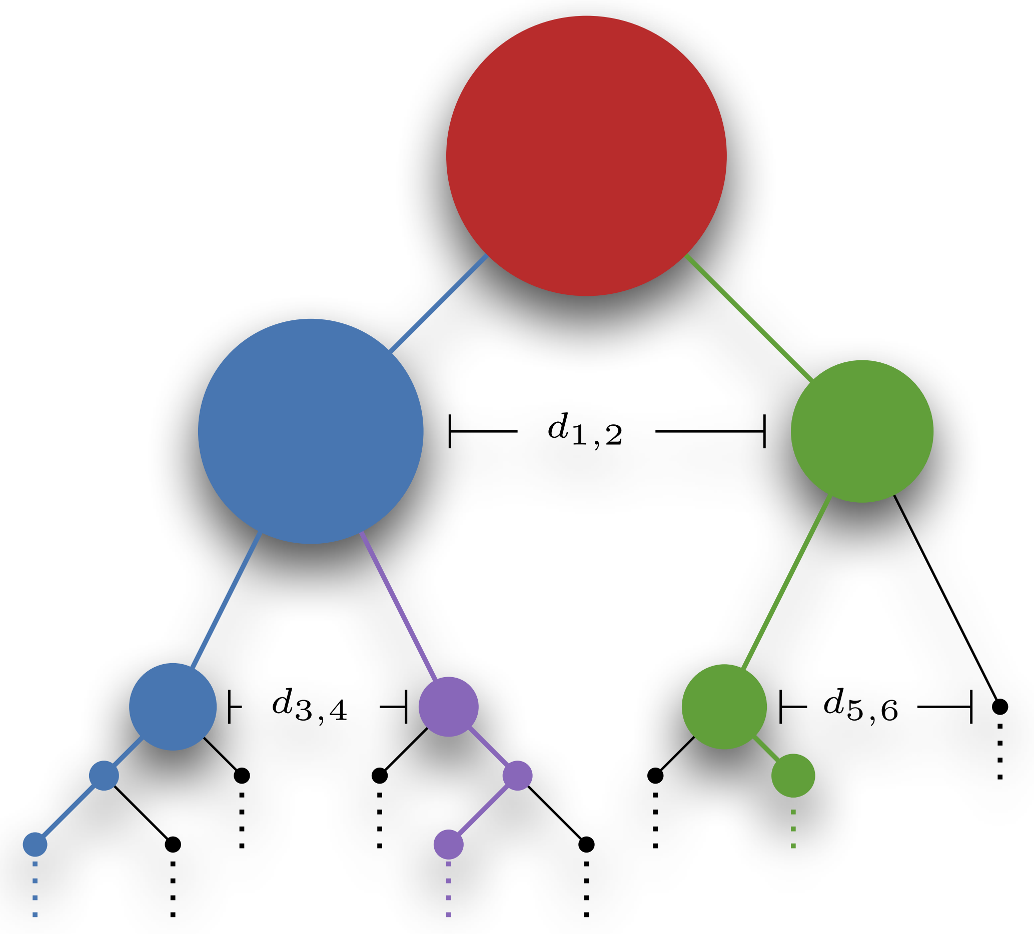








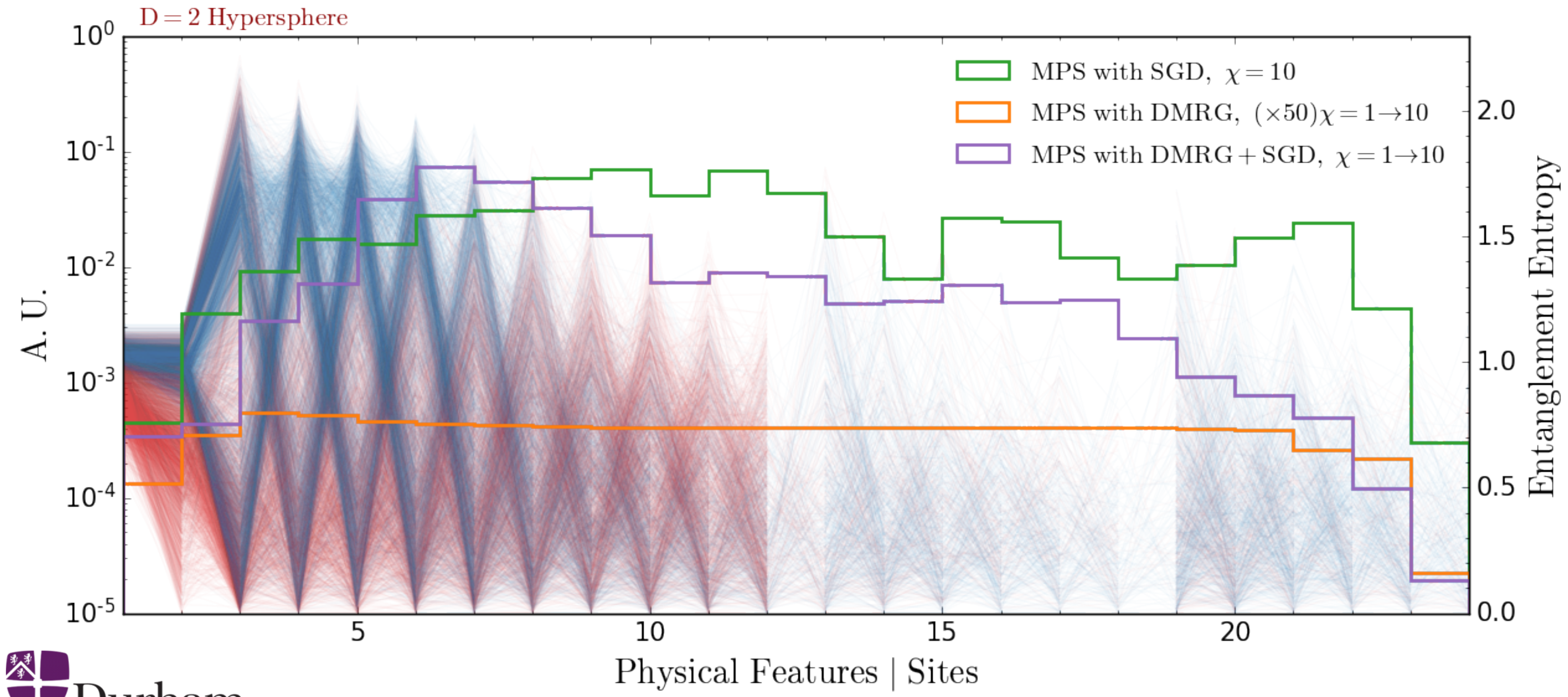
Test with cluster history sequence

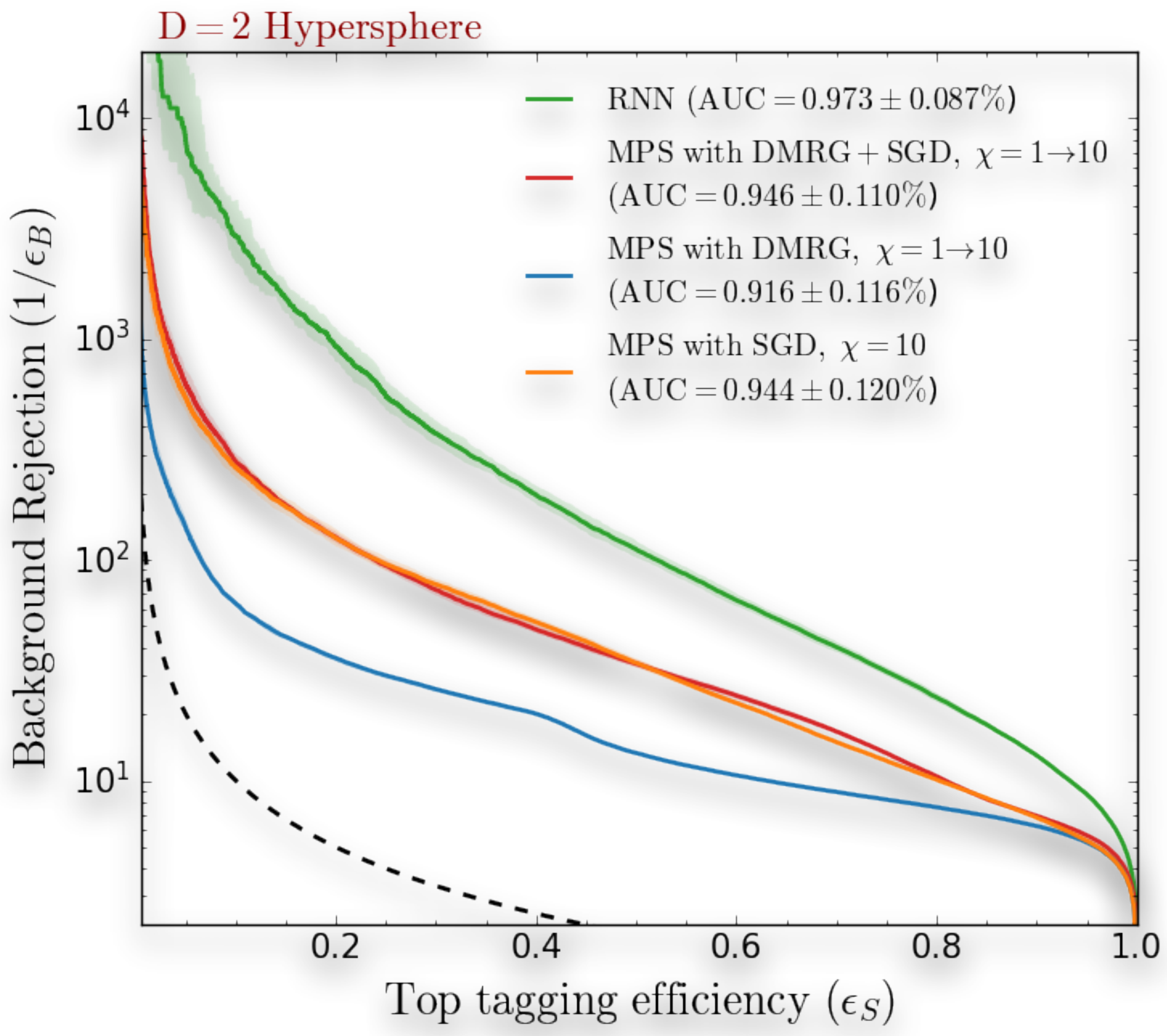


JYA, Spannowsky; JHEP '21

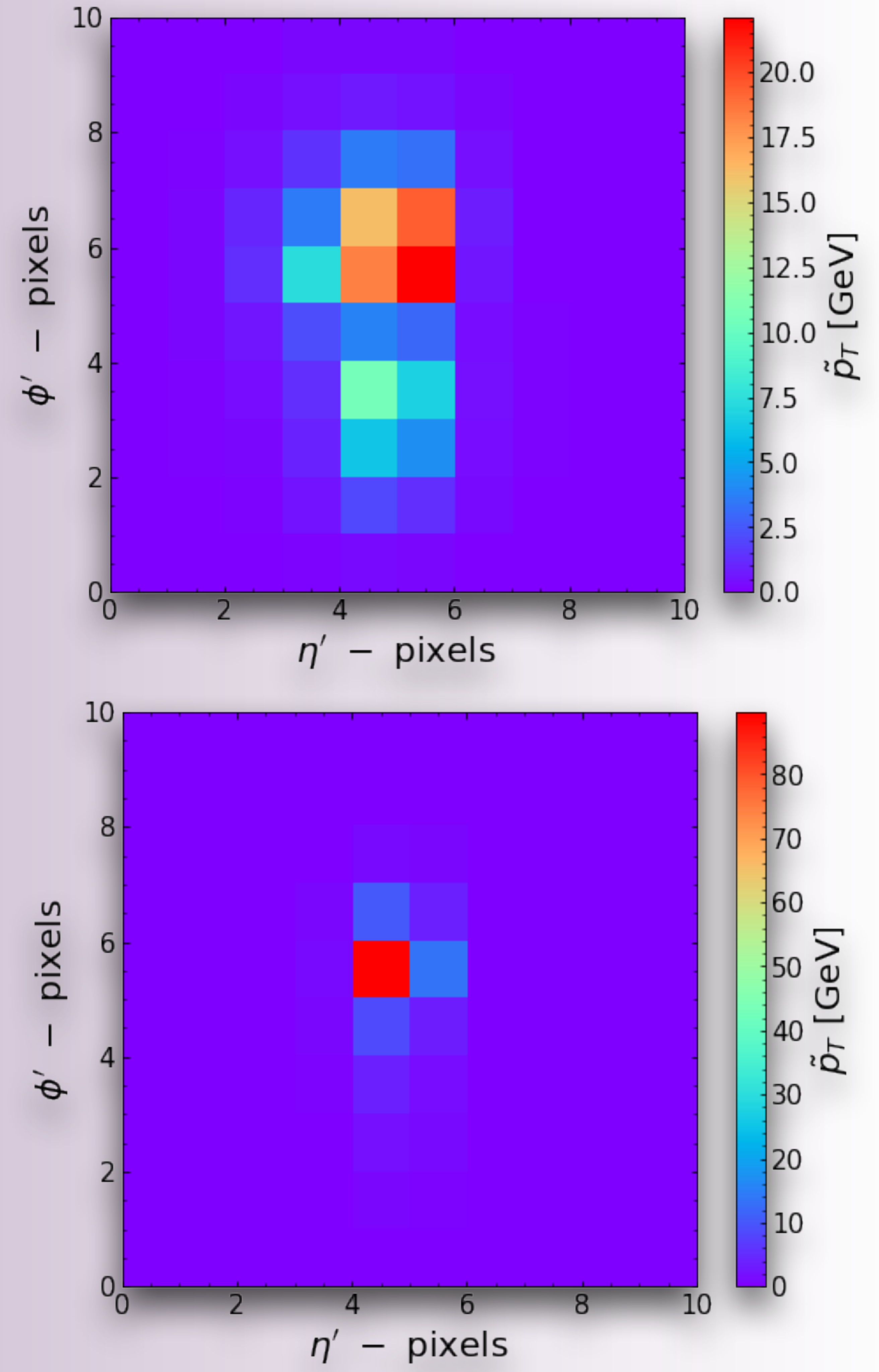
- Clustered jet
- Parent subjets
- Leading branch
- Second leading branch
- Third leading branch
- Discarded branches

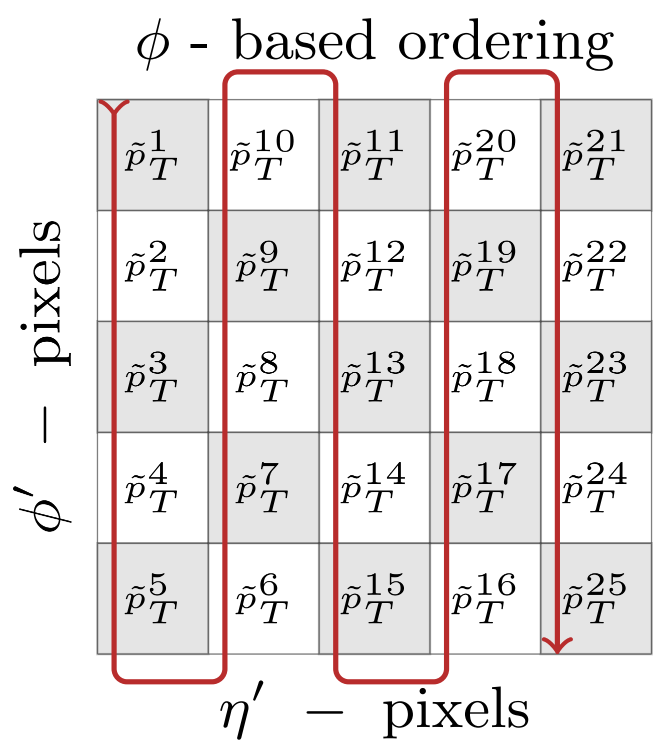
$$k_T - \text{distance} := min(p_{T,i}^2, p_{T,j}^2)\ \frac{\Delta R^2}{R}$$

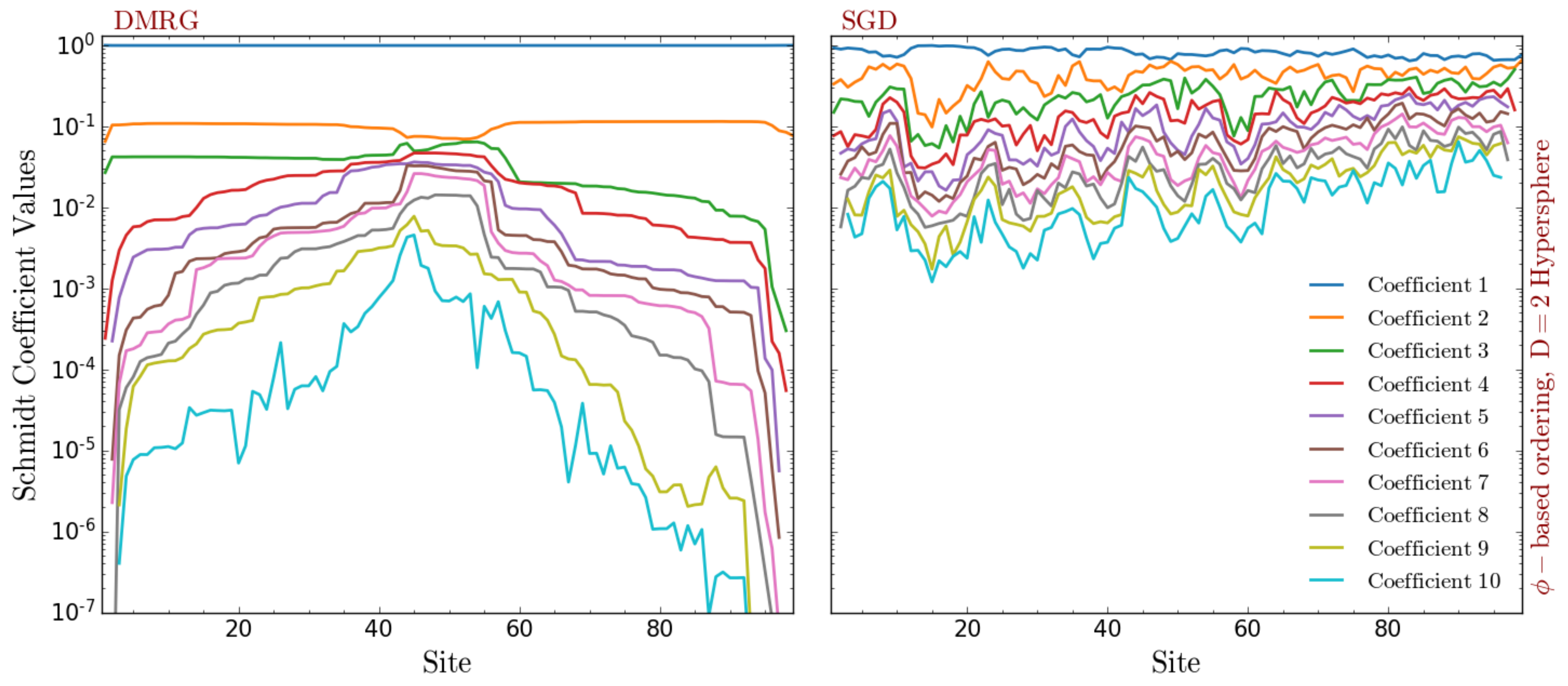


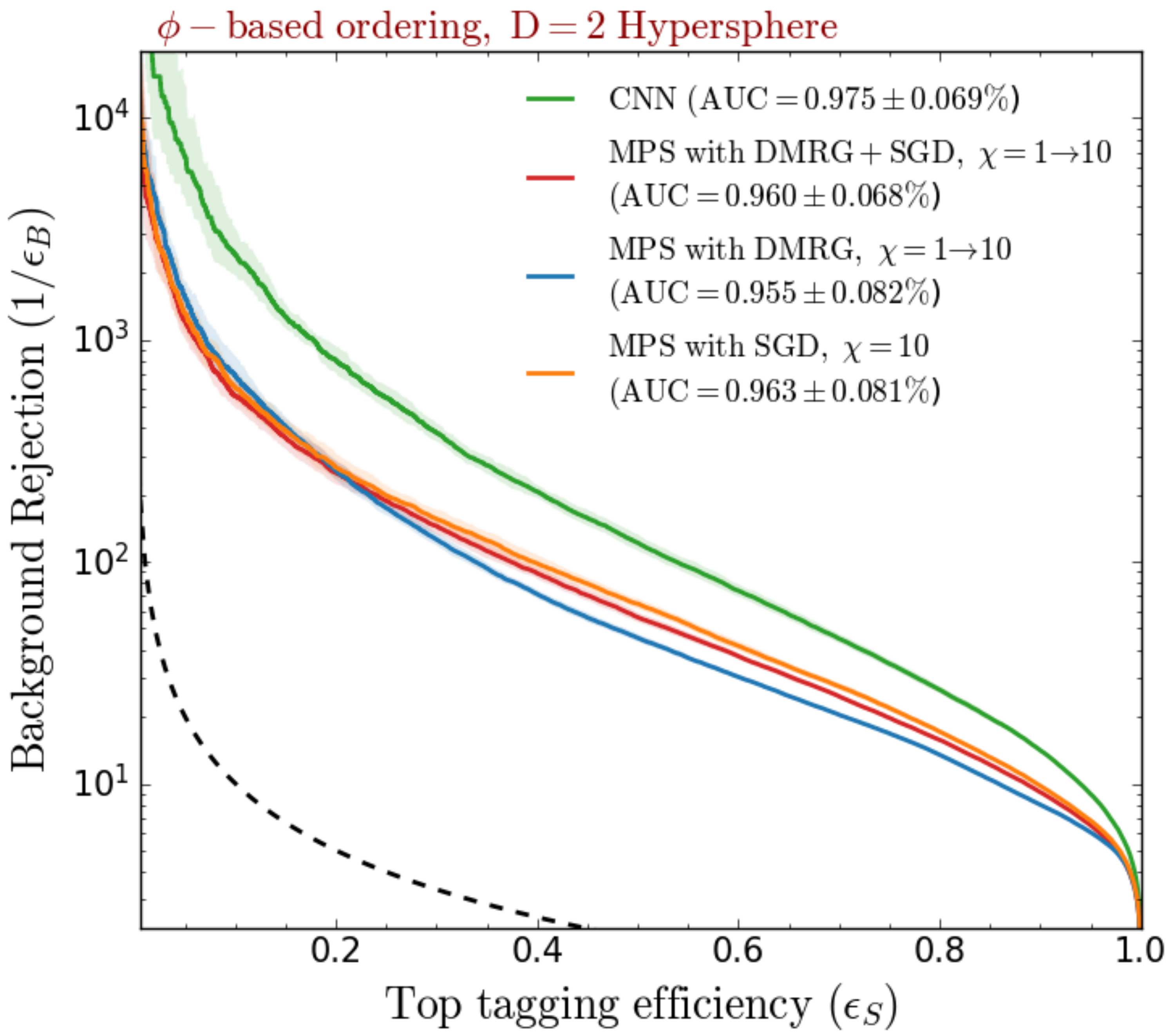


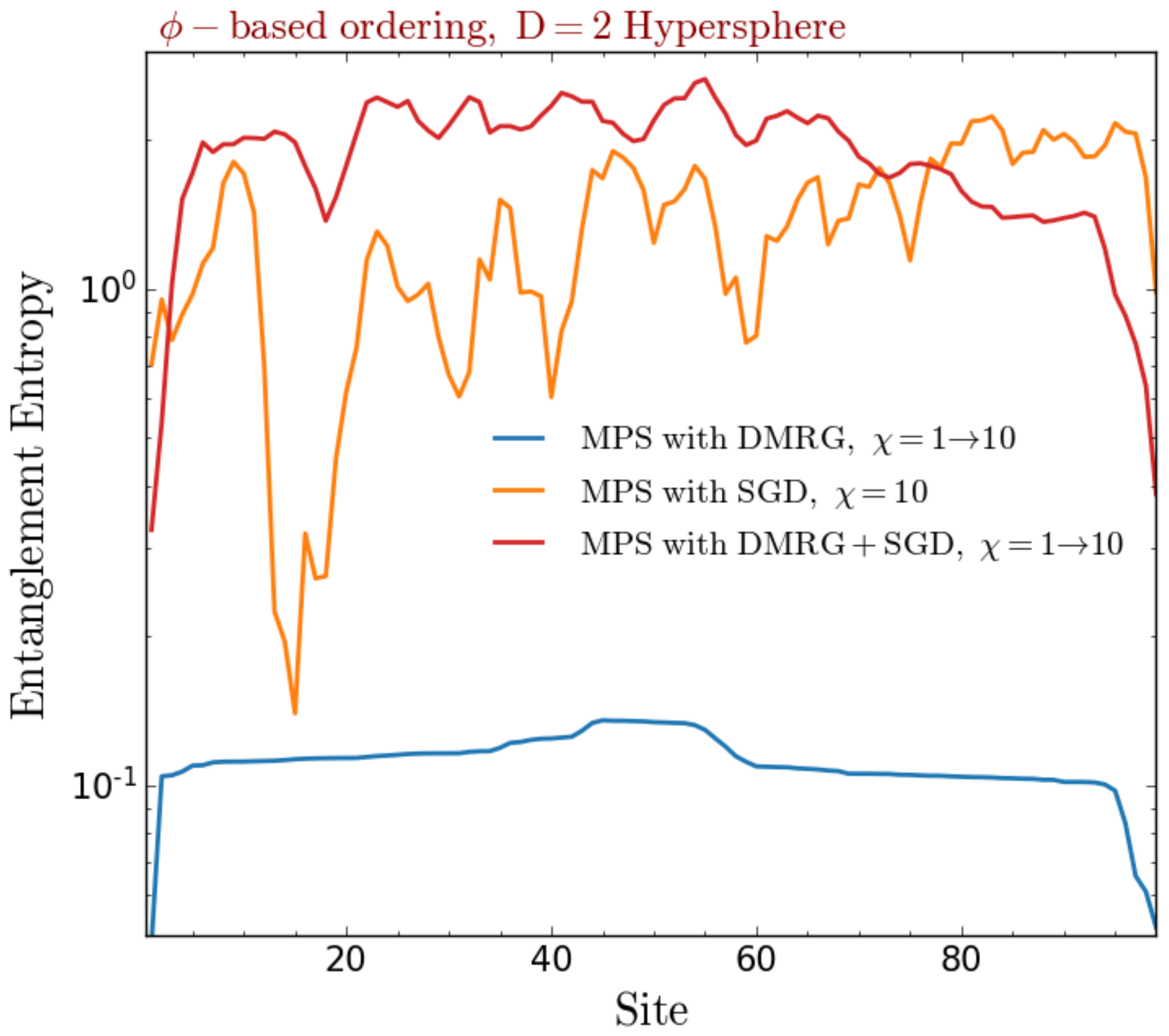
Results for ϕ-based ordering











Results for ϕ-based ordering