UNIVERSITÄT BONN

Self-organizing Maps in high energy particle physics

29.11.2021

Kai Habermann, Eckhard von Toerne

Overview

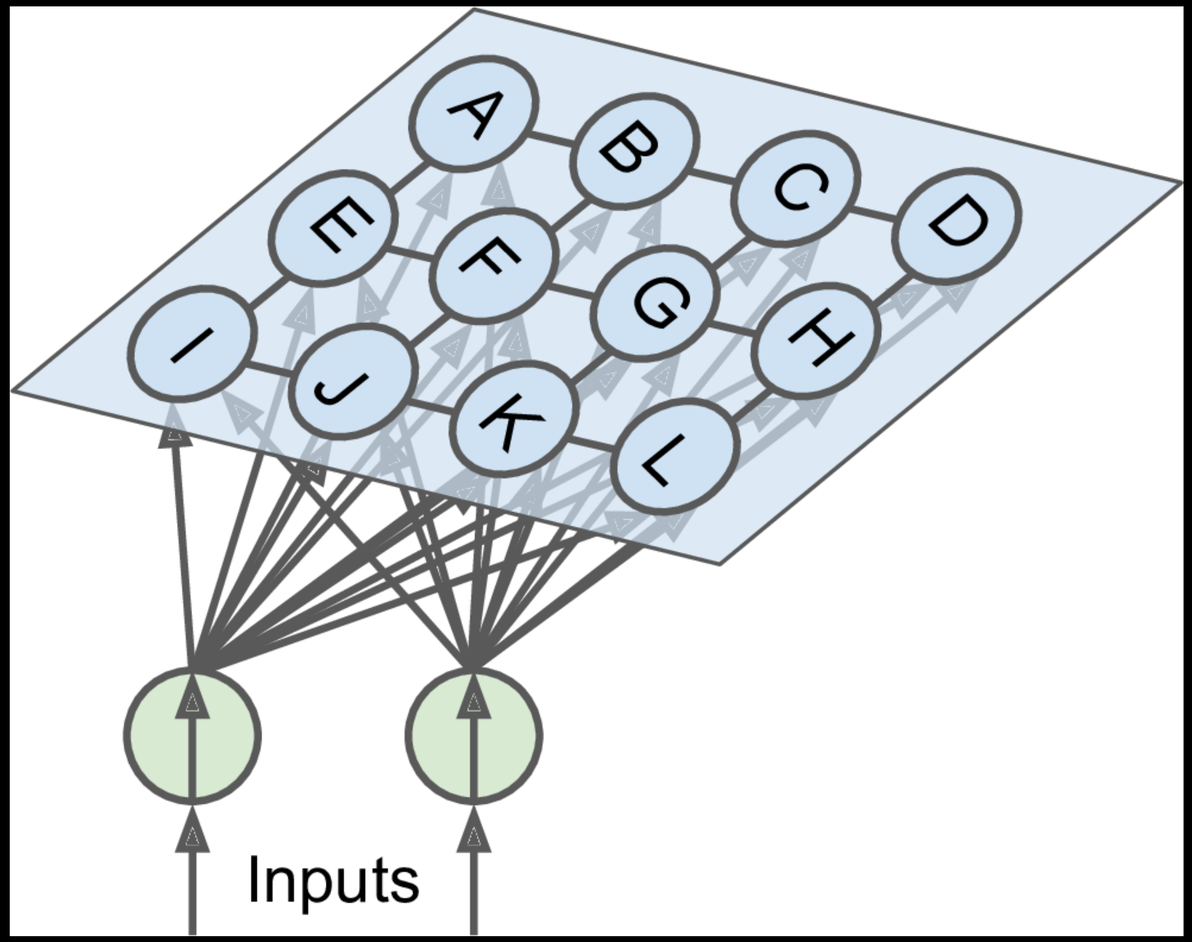
- Motivation
- Self-organizing maps (SOM)
- Training Alogrithms
- Dataset
- Analysis of data composition
- Study of Higgs-enriched region

Unsupervised learning

- SOMs are a common example of unsupervised learning
- Machine learning without the need for pre-classified data
  - Maps data out and provides information on clusters and the structure of the data
  - PCA is an example of unsupervised learning
- We will apply SOM to particle physics data

What is SOM?

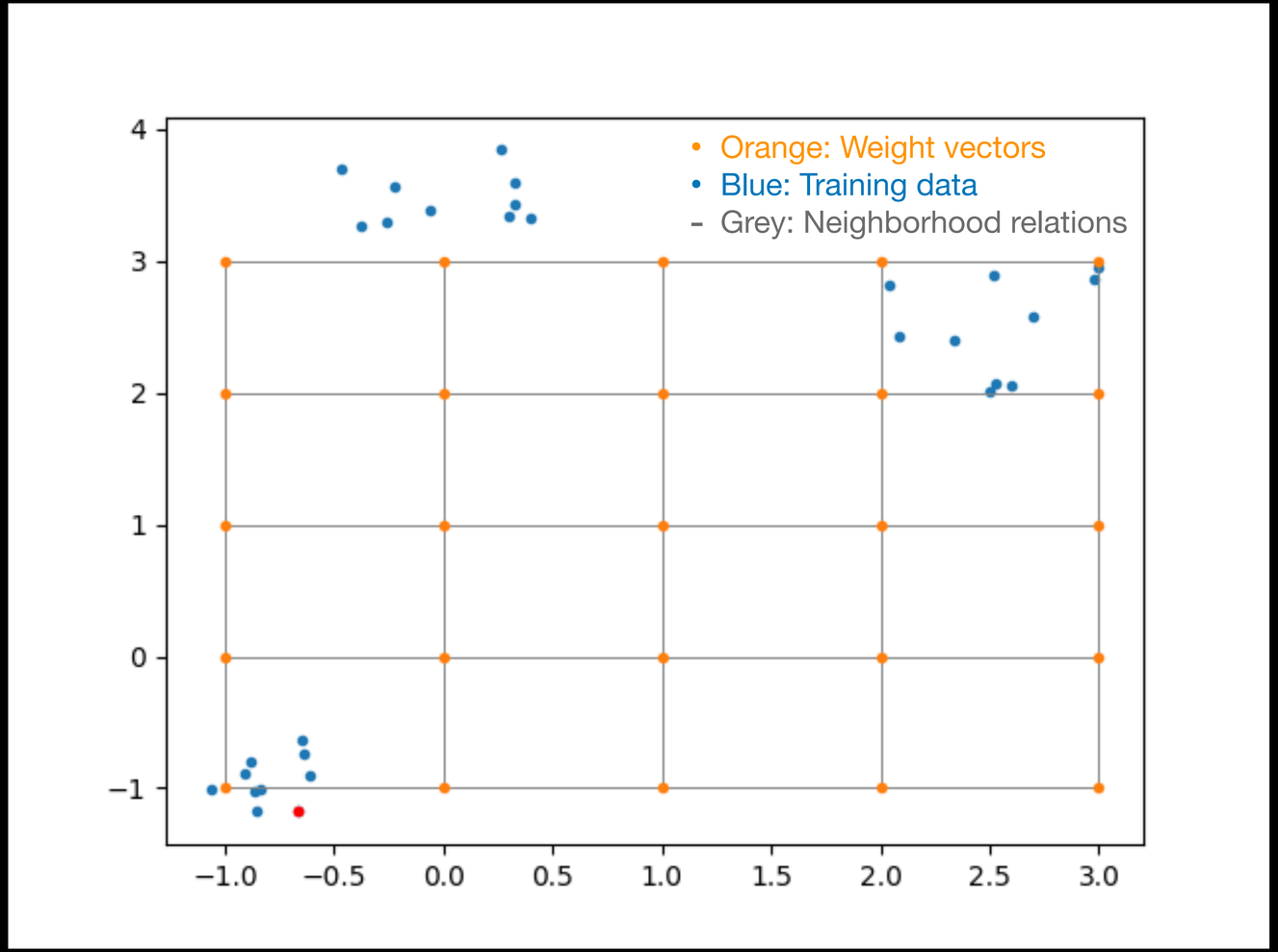
- ANN for discretization and dimension reduction
- Output Layer:
  - Layer of neurons organized in a 2 dimensional grid
- Input Layer:
  - Size of input vectors
- Weight vector the size of input vectors for each neuron of output layer
- Euclidian distance of data vector to weight vectors is used to map input vectors to output layer



„Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to
Build Intelligent Systems“, Aurelien Geron

Training algorithm

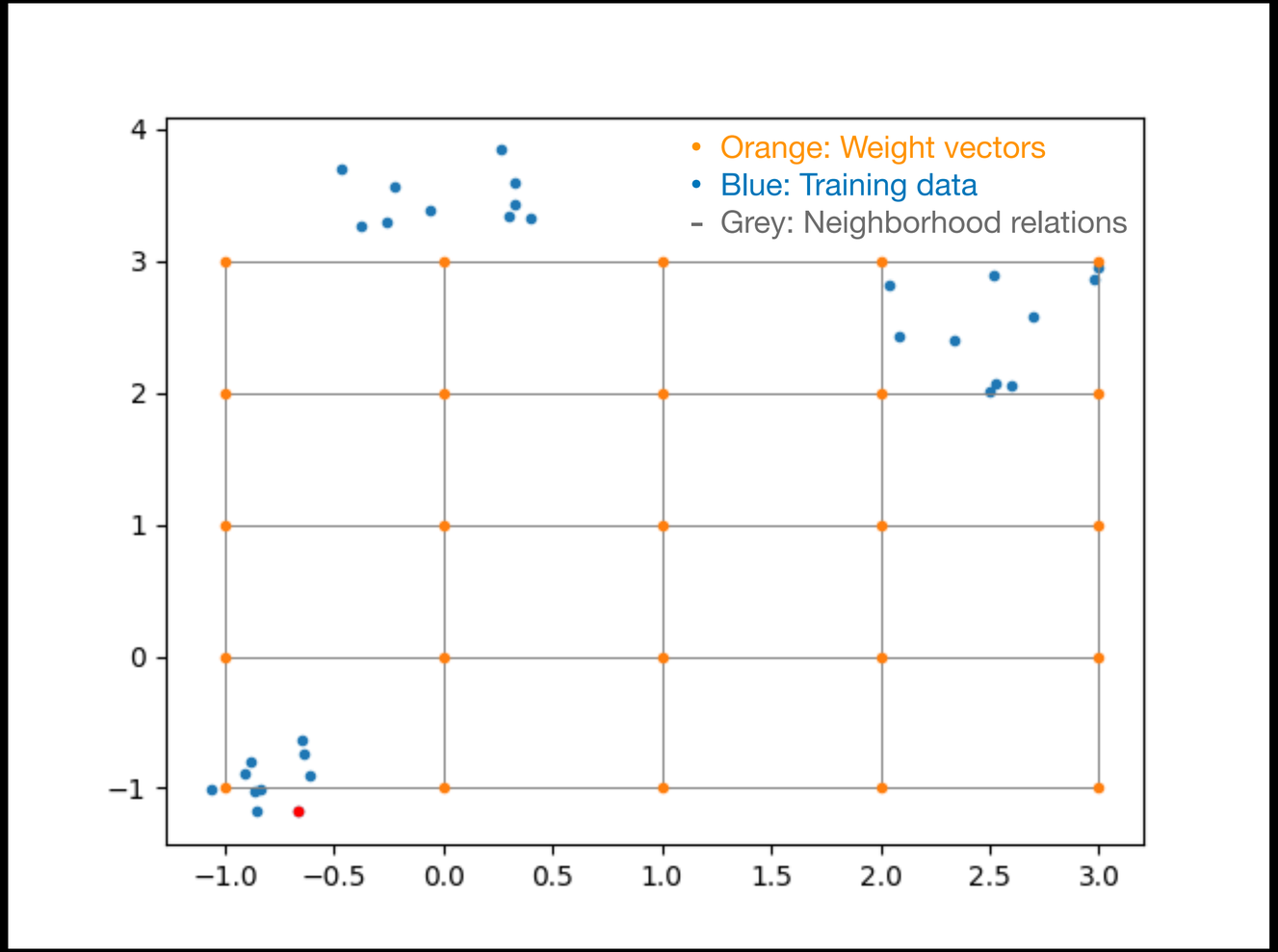
1. Initalize weight vectors.
2. Select one input vector from training sample v_j.
3. Calculate euclidean distance $\| v_j - w_i \|$ to the weight vector of each neuron n_i and choose neuron n_{best} with lowest distance.
4. Renew weights depending on n_{best}.
5. Repeat steps 2-4 until iteration limit is reached.



Training algorithm

$$w_k^{t+1} = w_k^t + (v_j - w_k^t) \cdot \beta(r_k, \sigma) \cdot l$$

- Weights will be moved closer to v_j.
- $\beta(r, \sigma) = F_{\text{gauss}}(r, \sigma)$, r distance to n_{best} in the output space.
- Immediate neighbors of n_{best} will be altered the most.
- Distance relations from input space are mostly conserved.
- Clustering is encouraged.
- Convergence through decreasing l and σ.



Improved training algorithms

- **Batch-SOM:**
  - Present entire dataset at once
  - Weight vectors become β-weighted mean of all input vectors
- **Batch-SOM adjusted:**
  - Present a portion of the data at once
  - Neuron will be updated with the β-weighted mean of all vectors in the batch
  - Convergence again via learning rate

Test Data

ATLAS Open Data[1] with $\sqrt{s} = 13\,\mathrm{TeV}$

- Openly available dataset from ATLAS (meant for educational use)
- We chose electron-muon dilepton final states + $N_{jet} \geq 0$ (500k events)
  - Contributions from $t\bar{t}$, $Z \to \tau\bar{\tau}$, WW and Higgs
- Remove 77 events with energies of more then $13\,\mathrm{TeV}$
- Opposite charge for leptons
- $m_T^{ll} > 70\,\mathrm{GeV}$, Isolation < 0.1
- Quantile transform to pull outliers in

[1] http://opendata.atlas.cern/release/2020/documentation/index.html

Input variables

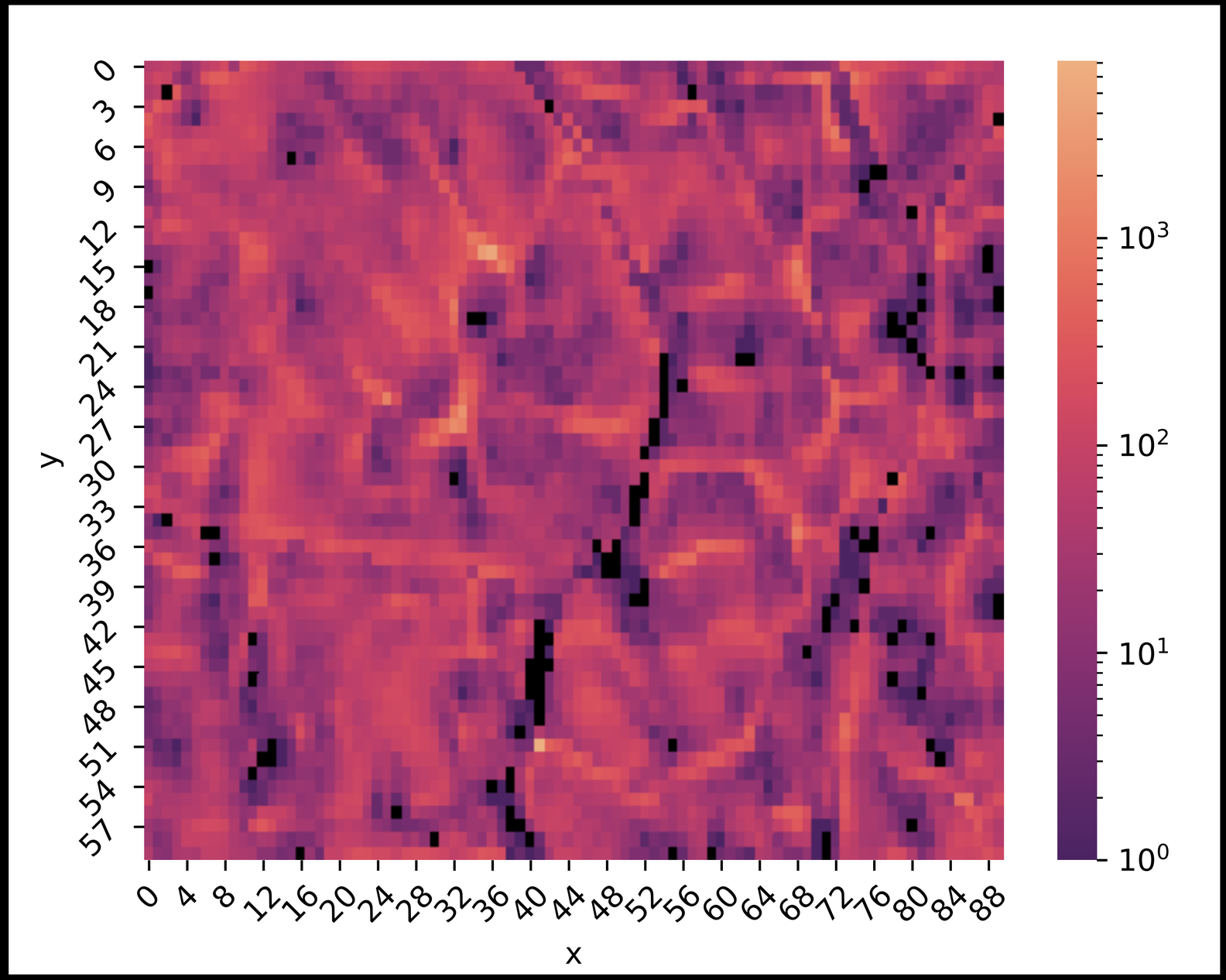
- Aim: cover most of the phase space within the variables

Variables :

- Transverse masses
- Transverse momenta
- Isolation of leptons
- MV2 b-tagging
- N_{jet}
- Invariant masses
- $h = \sum |p_T|$
- ΔR between leptons
- $\Delta\phi$ between leptons

Training result

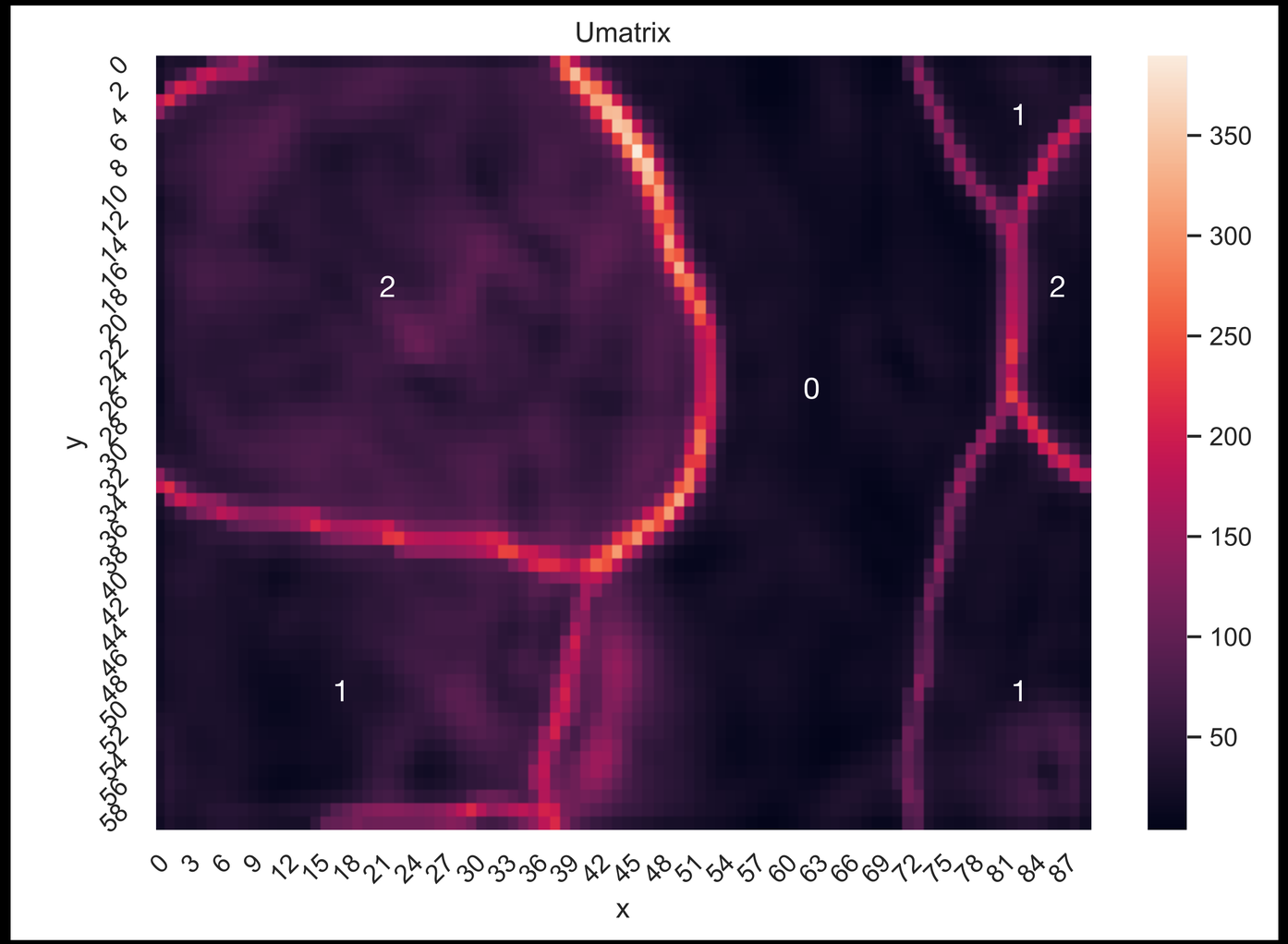
- SOM with 60x90 neurons
- Training with batches of 500 datapoints
- Each pixel represents a single neuron
- z-axis shows events per neuron
- Mapping of events to neurons via euclidian distance



Universal Distance Matrix (U-Matrix)

$$u_{ij} = \sum_{k=i-1}^{i+1} \sum_{m=j-1}^{j+1} |w_{km} - w_{ij}|$$

- w_{ab} are weight vectors of the neurons.
- Clusters of different amounts of jets
- 0: No jets
- 1: one jet
- 2: 2 or more jets

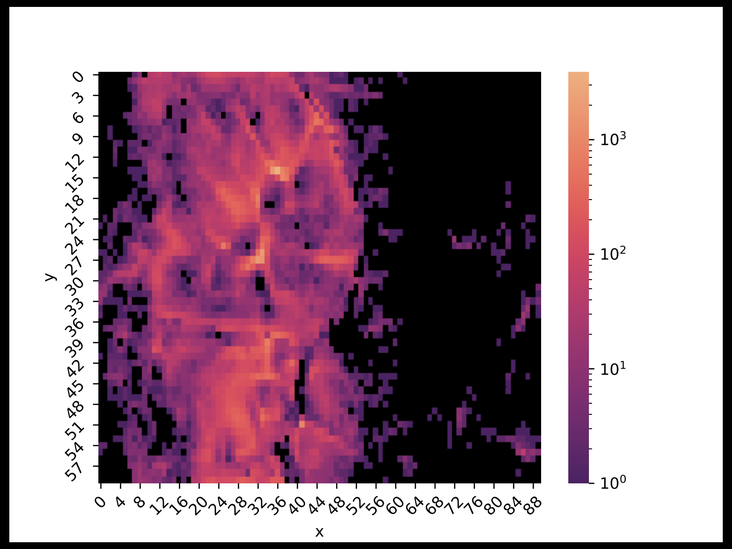


Data composition

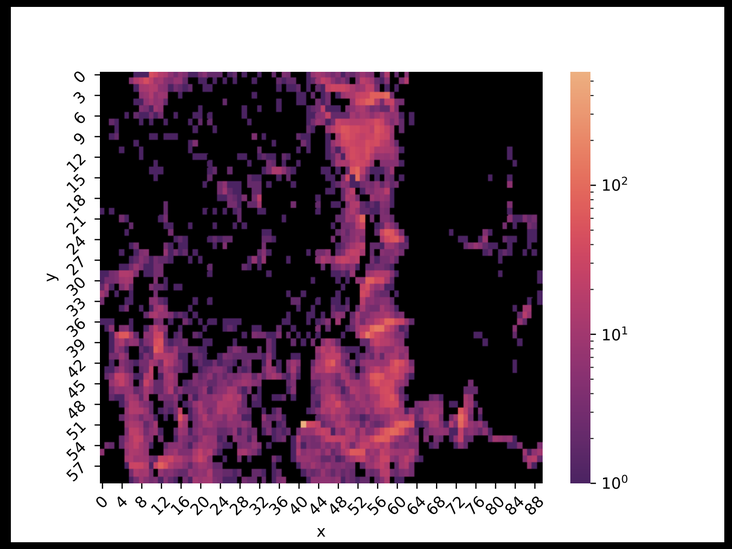
- Perform a fit of MC data to the real data
- For the fit:
  - Map MC data onto SOM trained with real data
  - Try to replicate density of real data on the SOM as sum of MC densities on the SOM
- After fit:
  - Search for regions of well isolated processes

Mapped MC data (m_T^{ll} cut included)

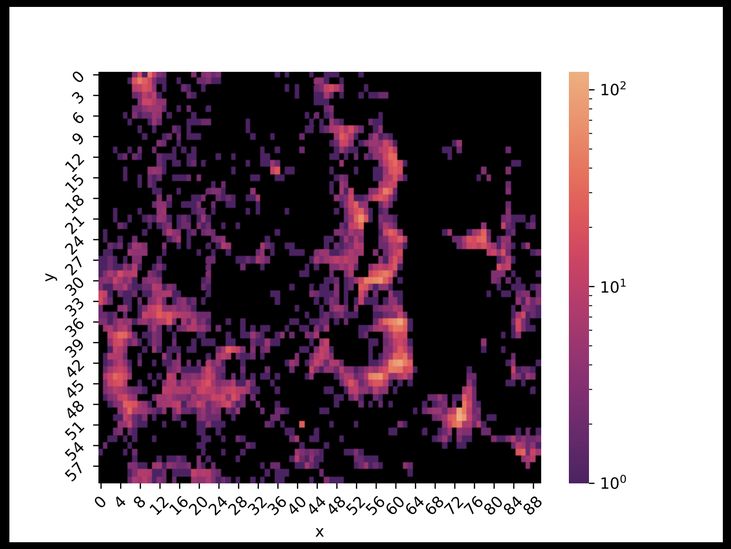
Single Top



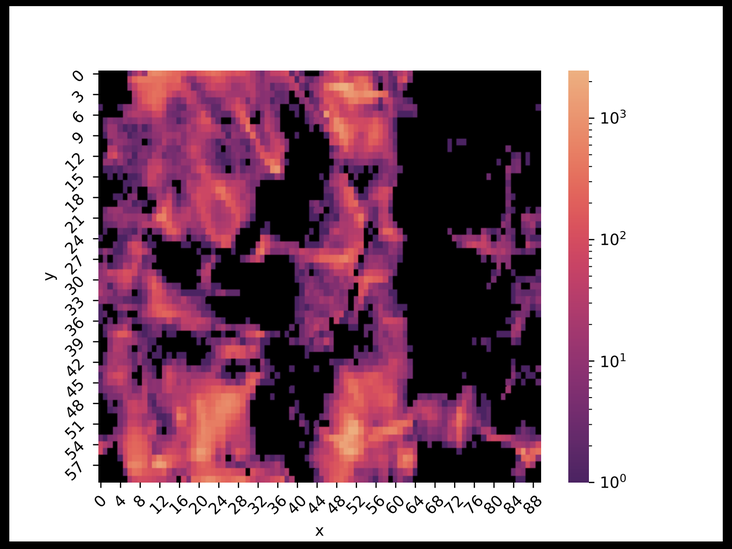
Single W



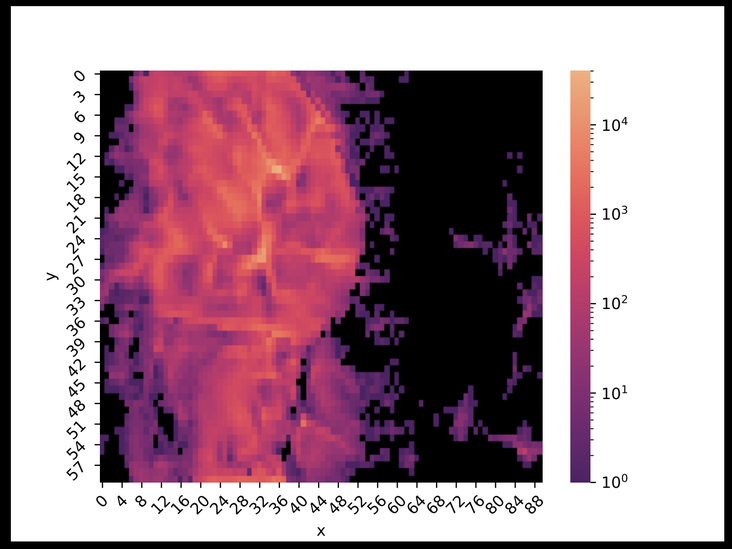
$Z \to \tau\bar{\tau}$



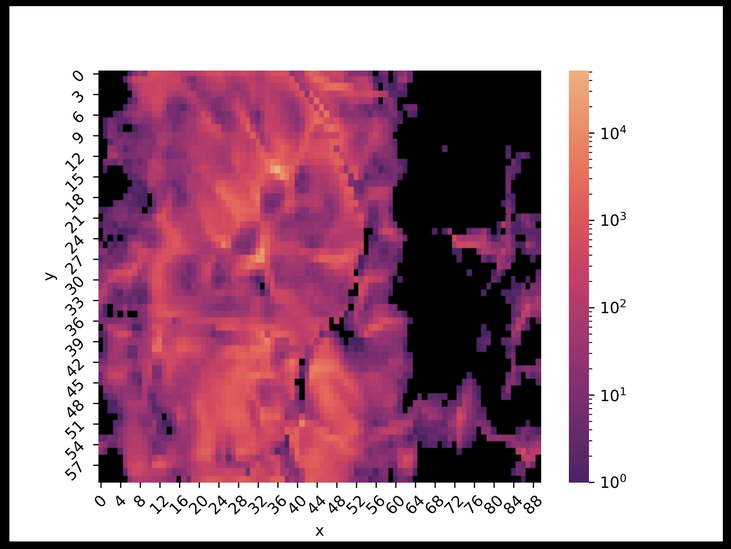
$H \to WW*$



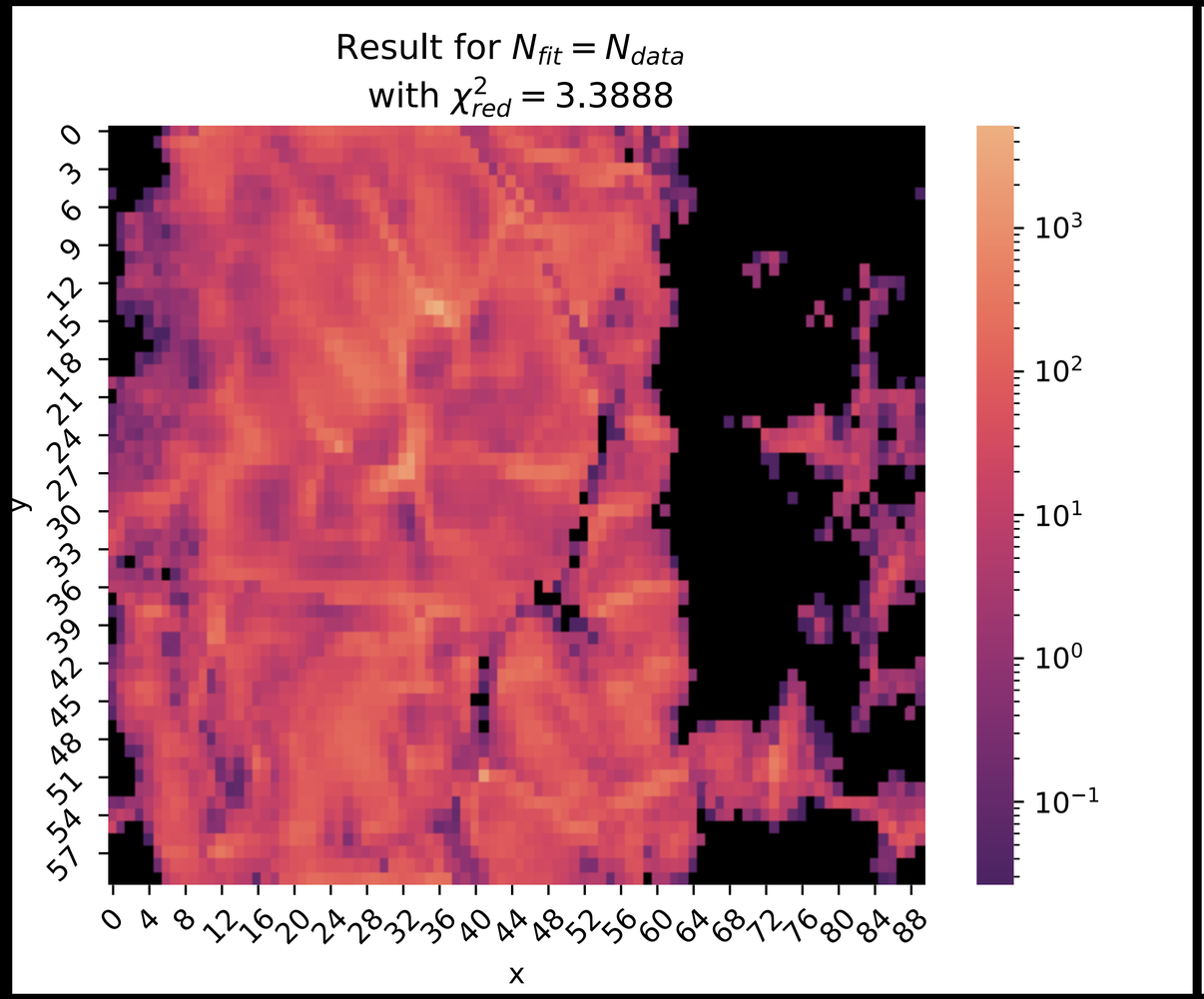
$t\bar{t}$



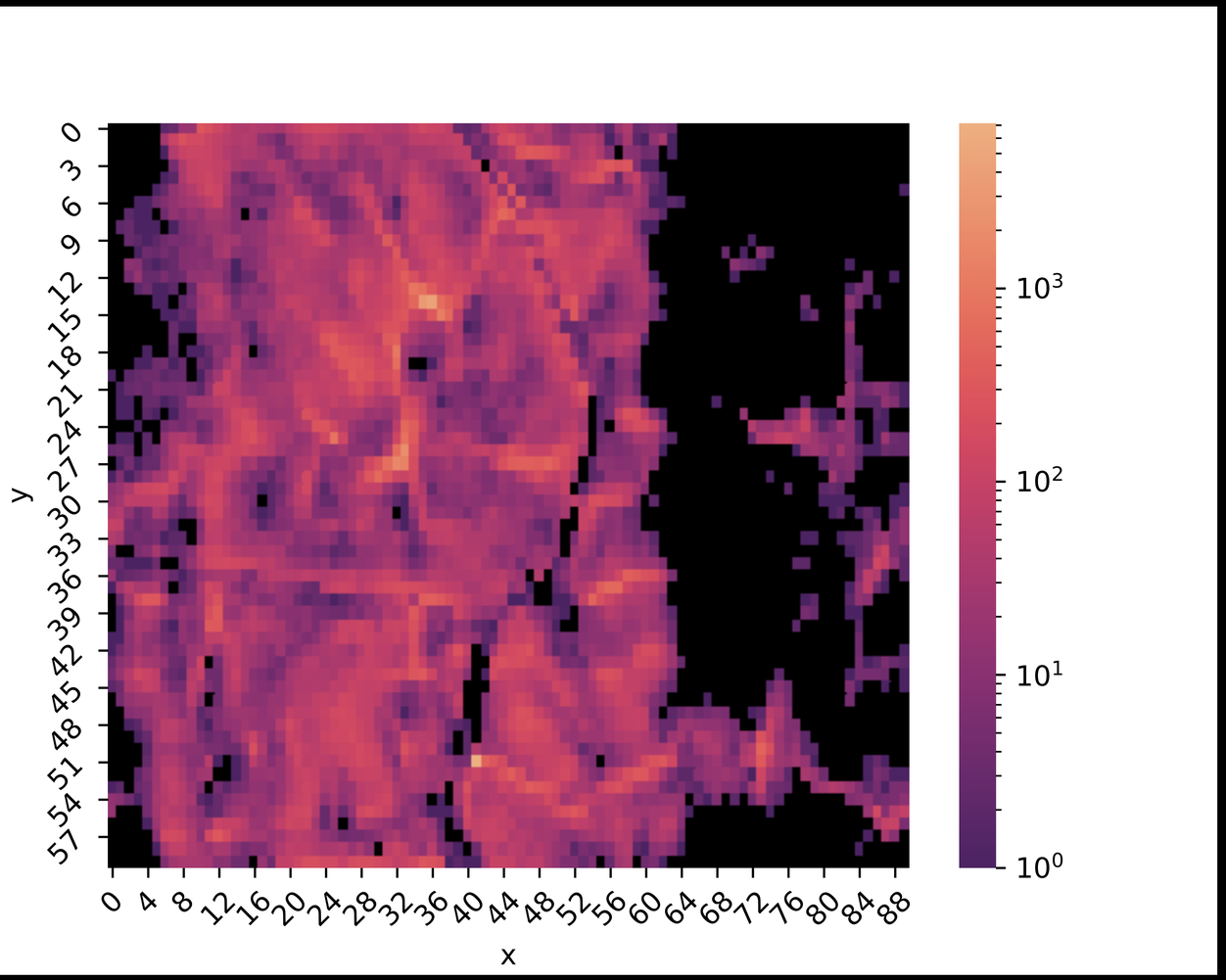
$W\overline{W}$



Fitresult



Real data



Fit of MC-templates to Open Data SOM

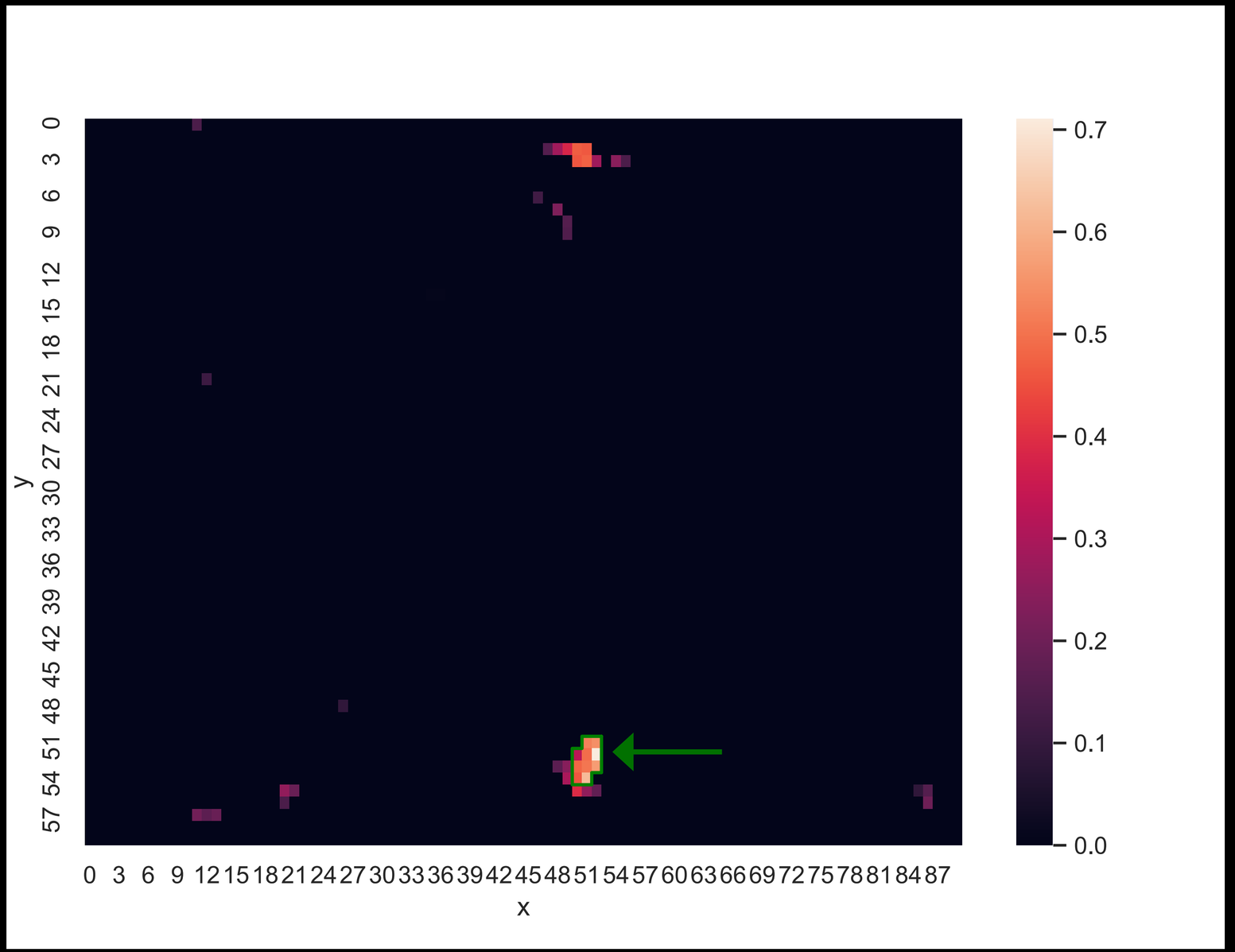
- Isolation has to be smaller than 0.1 to reduce QCD
- $m_T^{ll} > 70\,\mathrm{GeV}$ to reduce Drell Yan $\tau\bar{\tau}$

$$\frac{\chi^2}{N_{df}} = 3.39$$

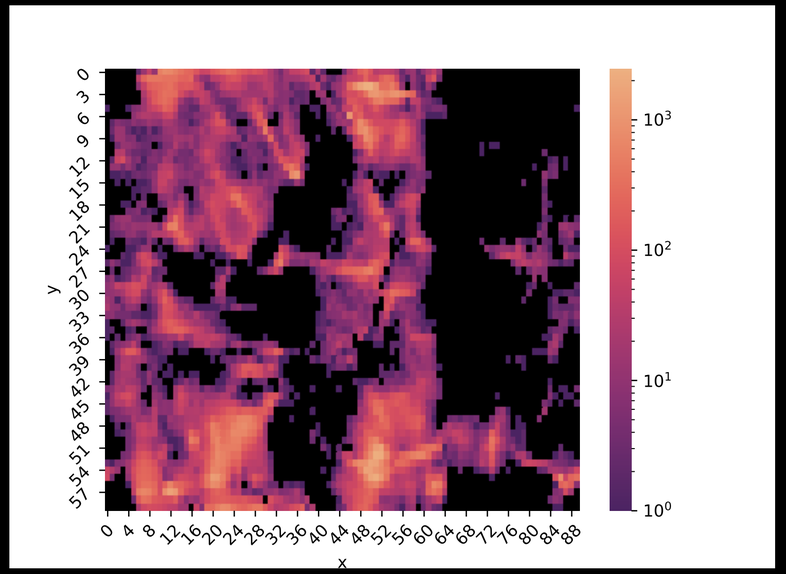
- Bad χ^2 due to missing contribution (probably QCD)

$H \to WW^*$	0.027 ± 0.002
Single top	0.039 ± 0.008
Single W	0.245 ± 0.003
$t\bar{t}$	0.382 ± 0.006
WW	0.189 ± 0.007
$Z \to \tau\bar{\tau}$	0.037 ± 0.002
$\sum$	0.919 ± 0.027

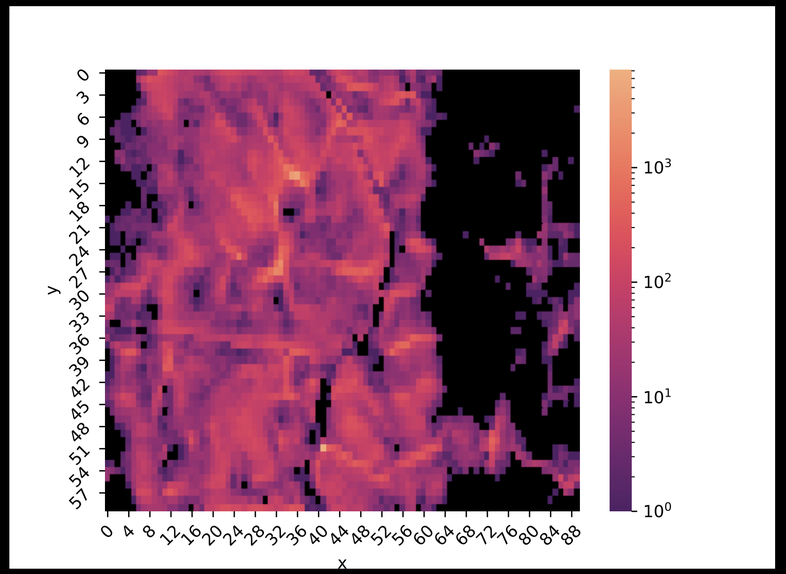
Relative amount of $H \rightarrow WW^*$



$H \rightarrow WW^*$



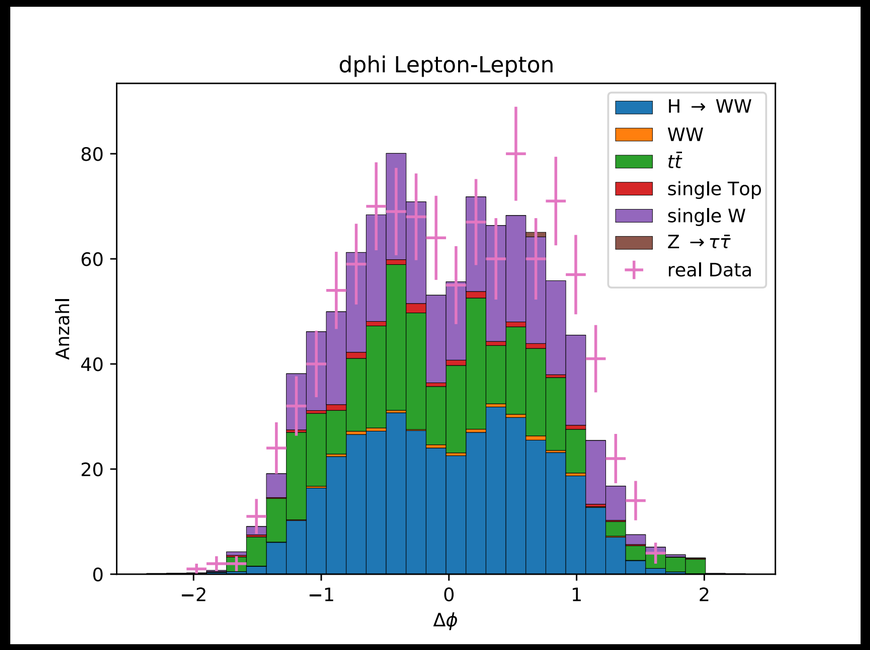
Real data

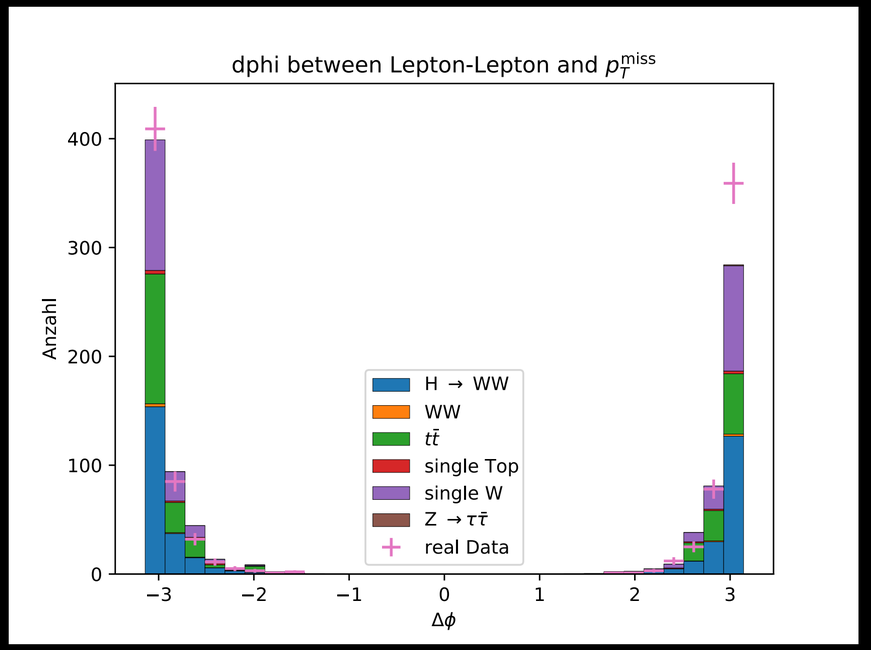


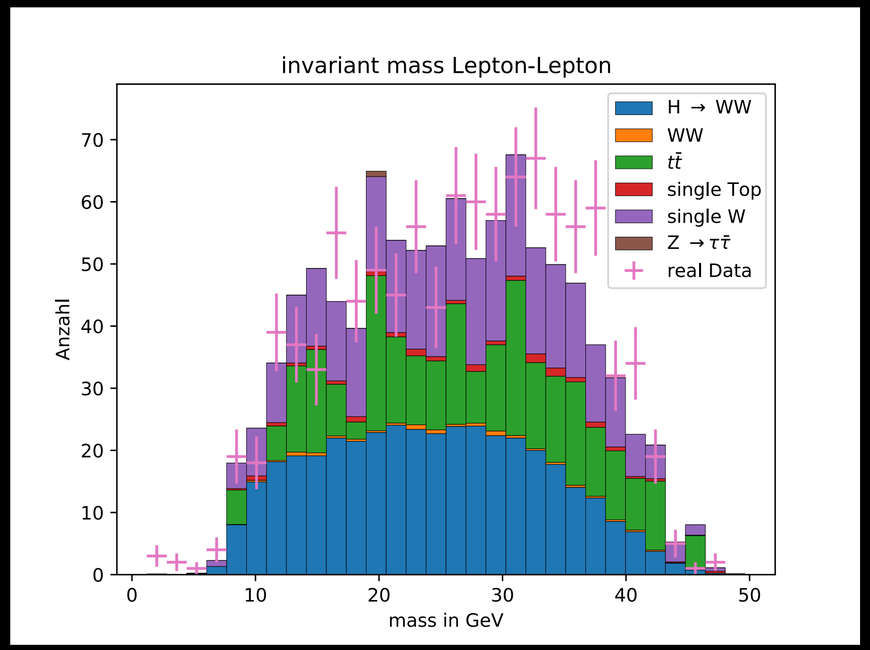
Isolated data

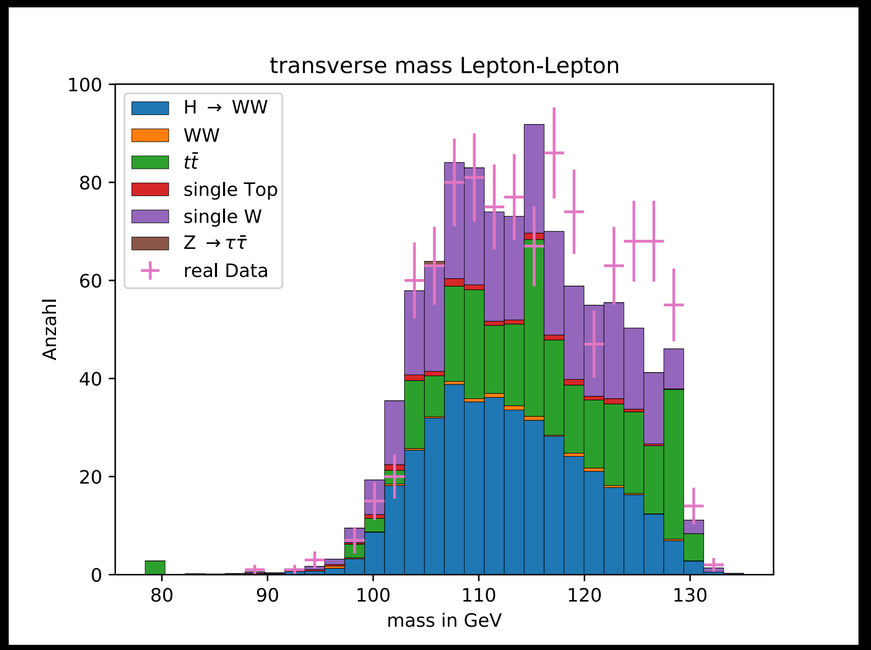
- 6099 $H \rightarrow WW^*$ events in total
- 396 $H \rightarrow WW^*$ events in subdataset
- Purity of 0.39

$$M_T^{ll} = \sqrt{2\, E_T^{ll}\, E_T^{\text{miss}}\, (1 - \cos(\theta_{ll,\text{miss}}))}$$









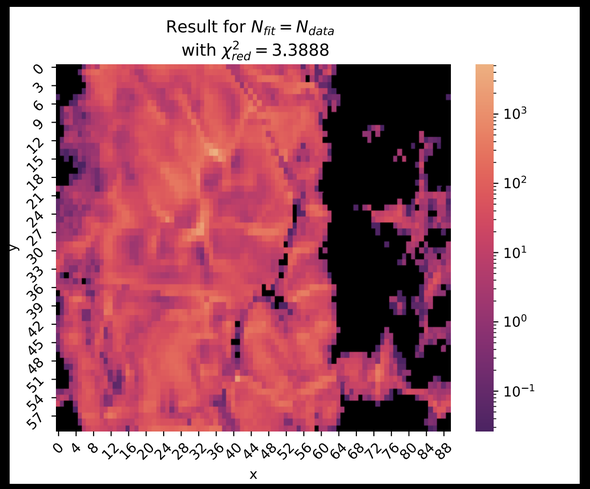
Summary and conclusion

- Method of unsupervised learning applied to $e\mu$-final states
- SOM can be created without need for MC
- Features in map: Regions in map identified as different N_{jet}
- Some regions dominated by a single MC contribution
  - One region with high $H \rightarrow WW^*$ occurrence
  - We can isolate 396 out of 6099 $H \rightarrow WW^*$ processes with a purity of 0.39

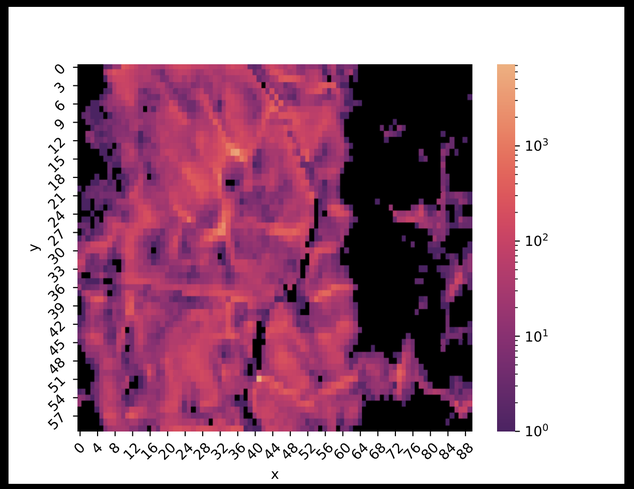
Thank you for listening!

Fit procedure

Fitresult



Real data



- Map out MC-data on pre-trained SOM
- Normalize histograms
- Fit normalized real data histogram as weighted sum of normalized MC-data