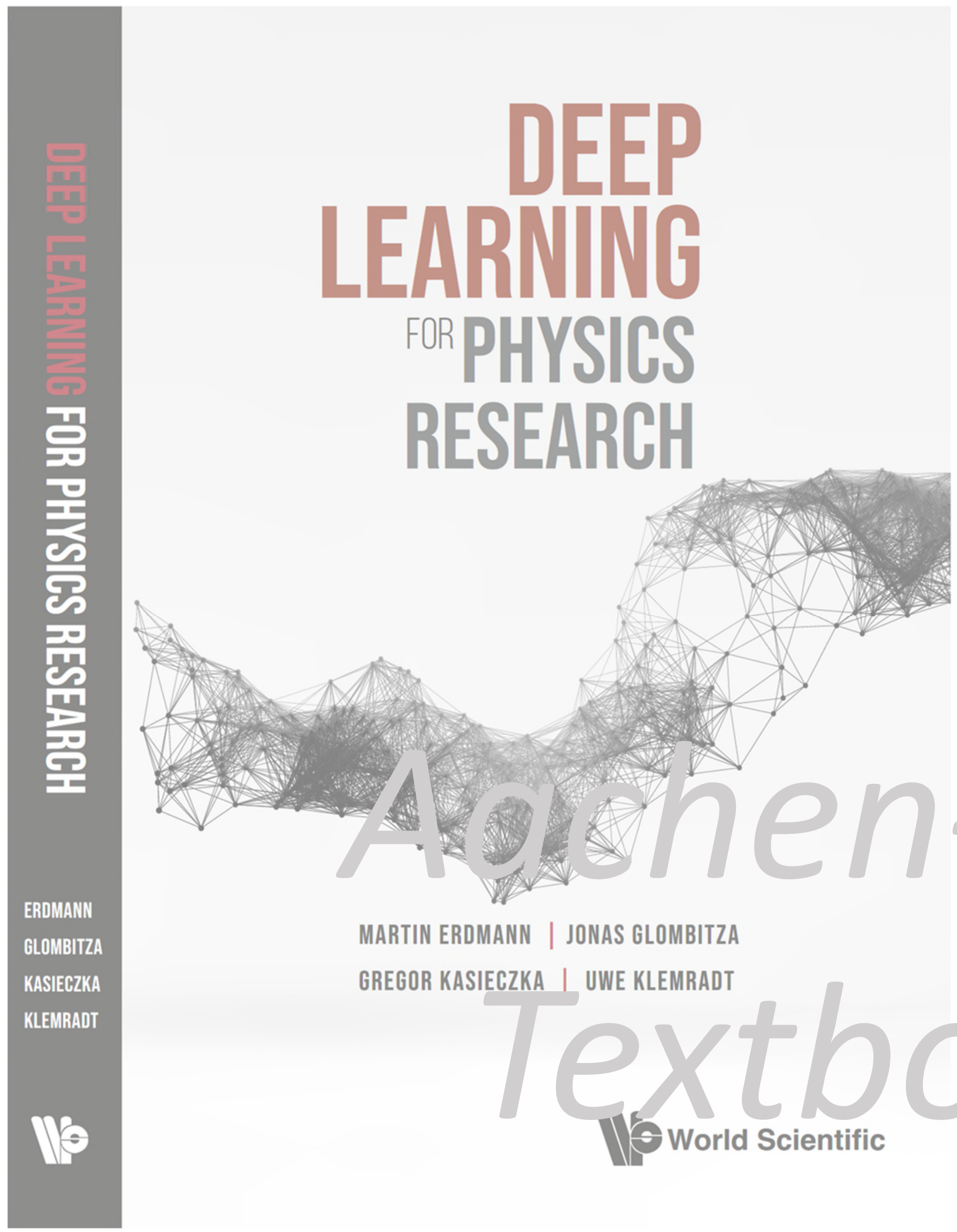
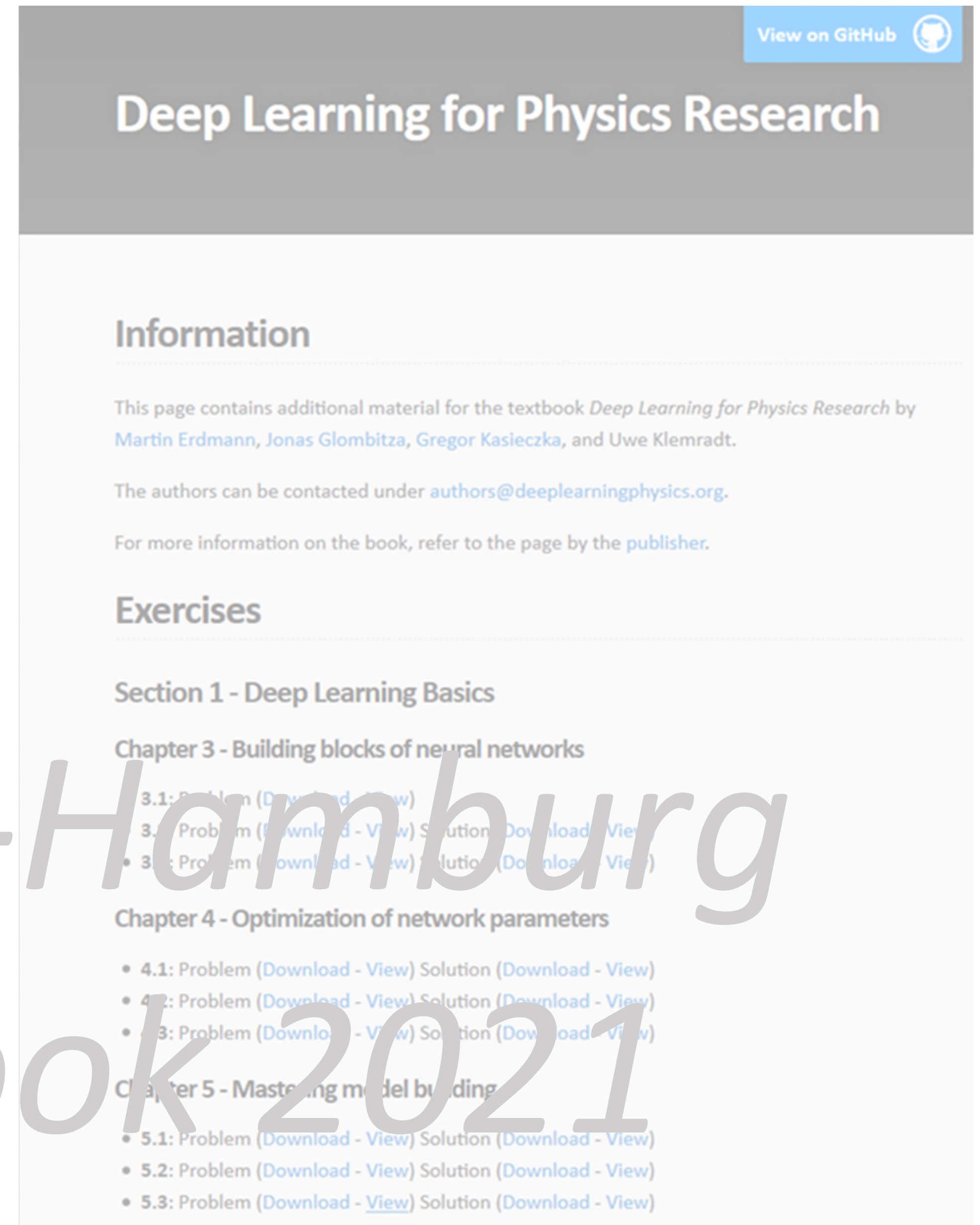
Deep Learning meets Physics





Aachen-Hamburg
Textbook 2021

Prof. Dr. Martin Erdmann, RWTH Aachen University, 14-Sep-2021

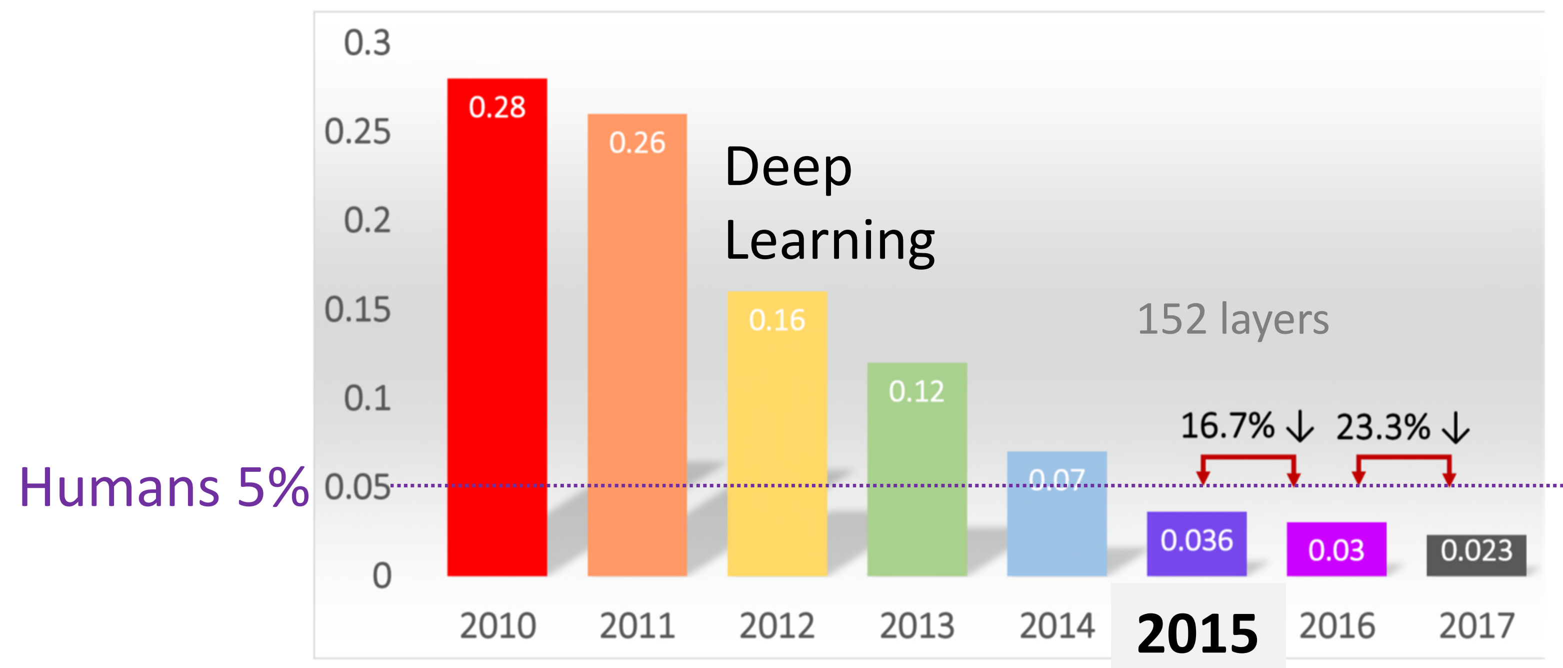
O. Russakovsky et al, arXiv:1409.0575; K. He, X. Zhang, S. Ren, J. Sunar, arXiv:1512.03385
WMW Jie Hu, Li Shen (Oxford), Gang Sun, 2017

Deep Learning spectacular success

Image recognition challenge

ImageNet: 1.2 million images in 1000 categories

Classification error rate



Deep learning errors < human errors

Generative Modeling

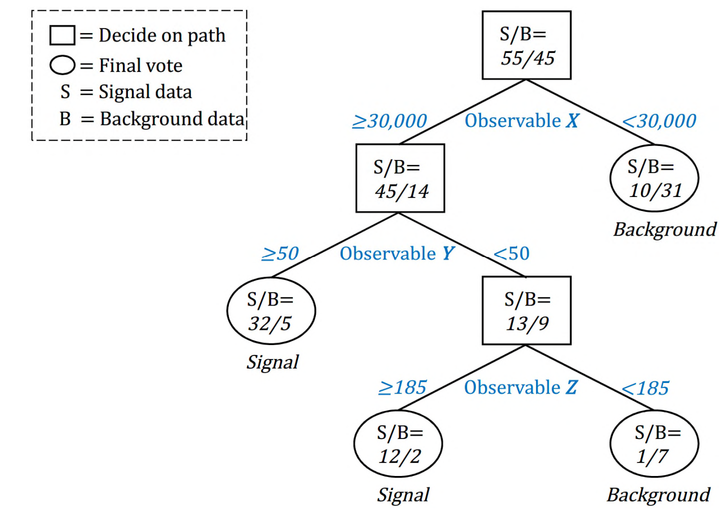
https://thispersondoesnotexist.com

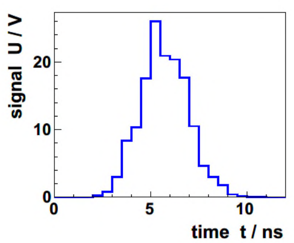
Plan for today

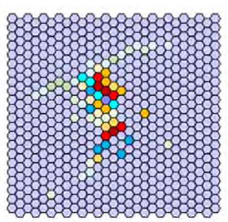
1. What deep learning is precisely: neural networks
2. Deep learning & data symmetries
3. Autonomous model building
4. Experiments´ operation reality and network insight

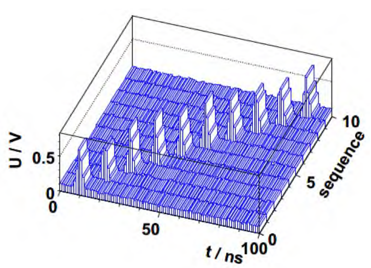
Data analysis → deep learning

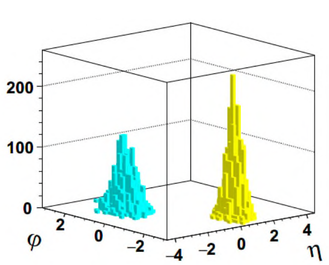
Boosted Decision Tree











Rule based system

input → hand designed program → output

Classic machine learning

learned by machine

input → hand designed variables → mapping → output

Representation learning

input → variables → mapping → output

Deep learning

input → simple variables → more abstract variables → mapping → output

Realized by Neural Networks with ‘many‘ layers

variable = *feature*

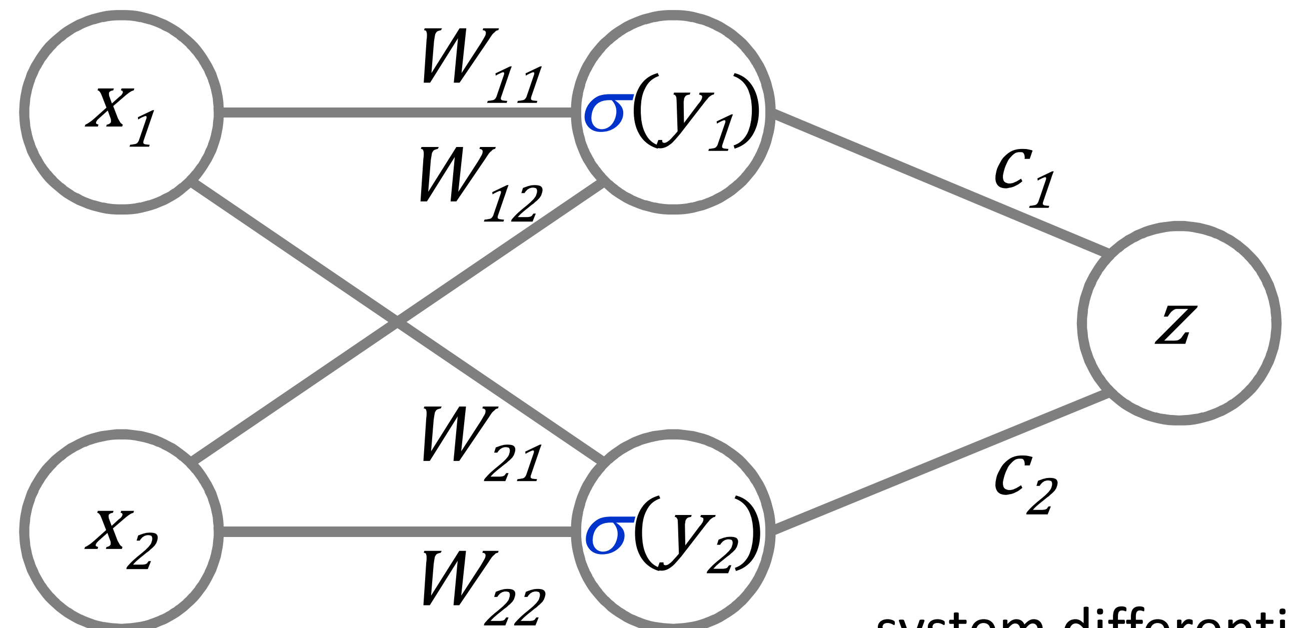
McCulloch, W.S., Pitts, W.: Bulletin of Mathematical Biophysics (1943) 5: 115.
Frank Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

Neural Network Operations

x multi-dimensional input data
W, b to be trained
successively apply 2 operations:

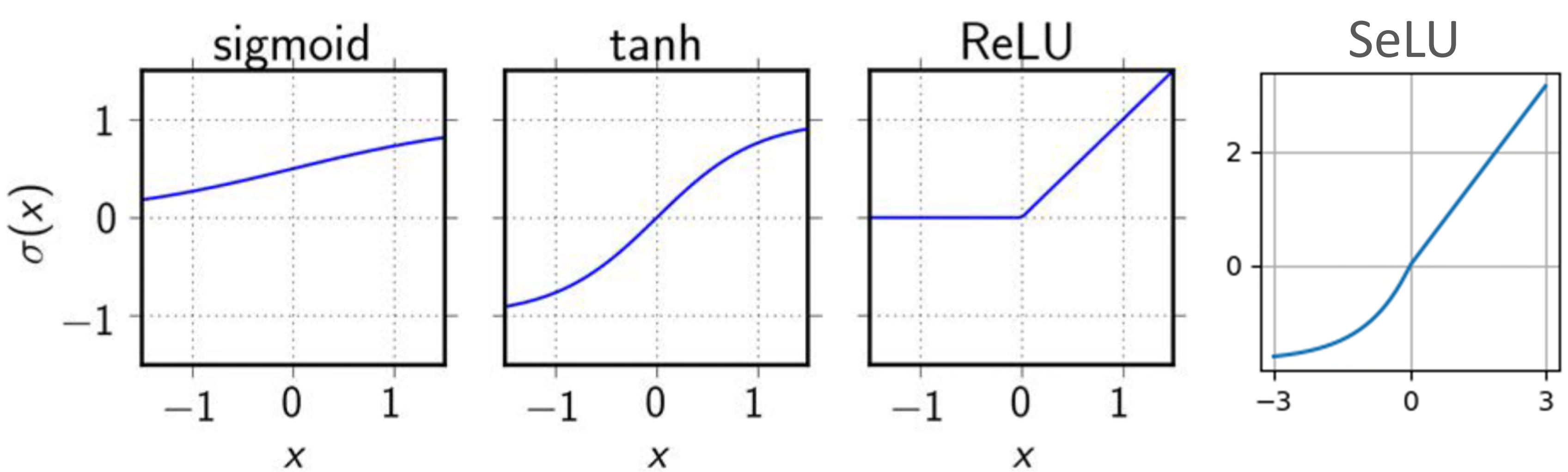
$$y = Wx + b$$

$$h = \sigma(y)$$



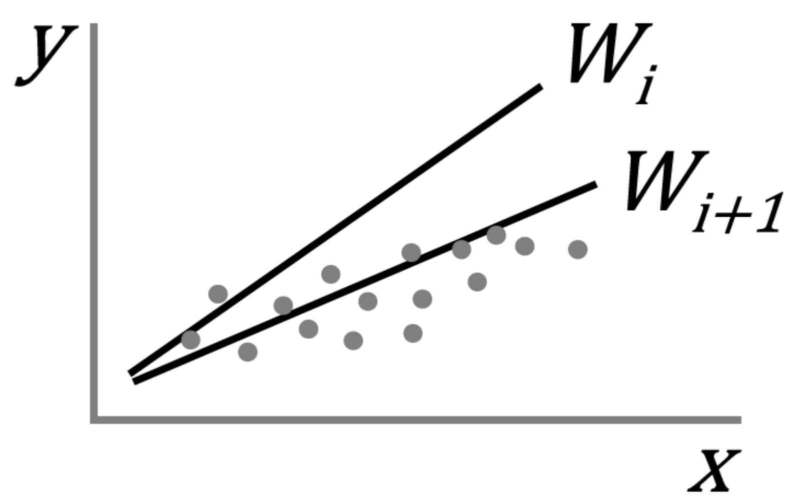
system differentiable

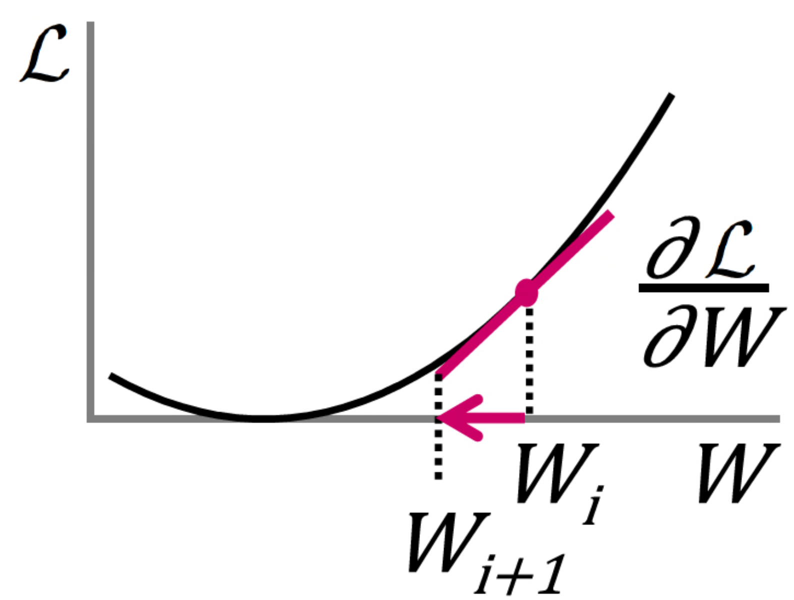
activation function: departure from linear system



Neural Network Training

('supervised')





- **Data** set

  $\{ x_i , y_i \} \quad i = 1, \ldots, N$

- Define **model**

  $y_m(x) = W\, x + b$

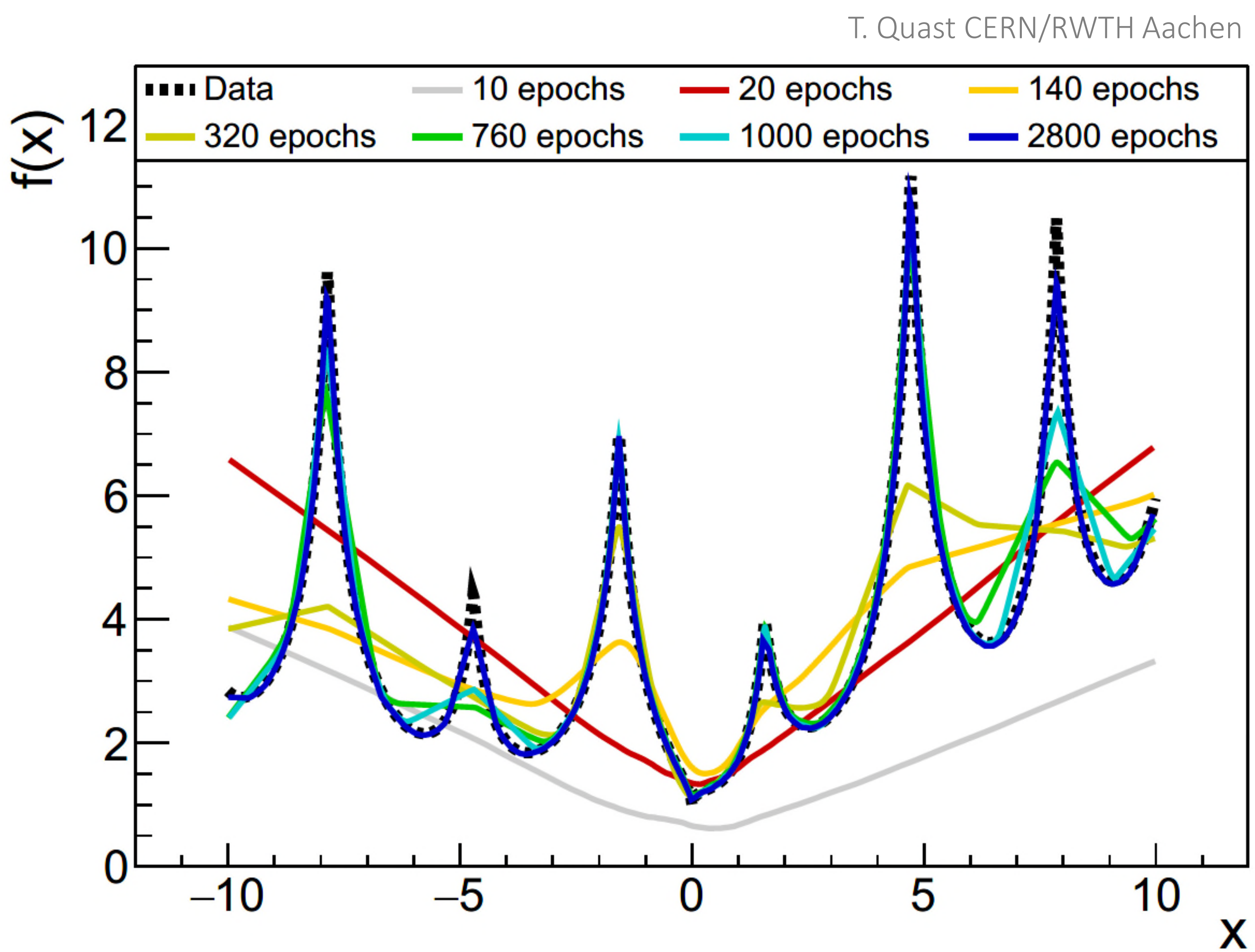
- Define **objective** function (=**loss, cost**)

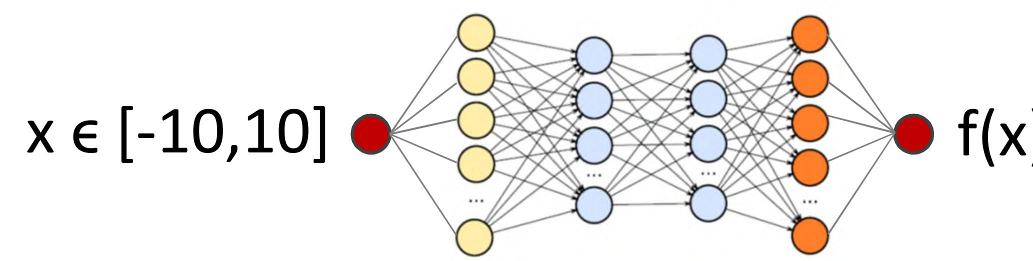
  $$\mathcal{L}(W, b) = \frac{1}{N} \sum_{i=1}^{N} [\, y_m(x_i) - y_i \,]^2$$

- **Train** model by optimizing the parameters

  $(\widehat{W}, \hat{b}) = \arg\min \mathcal{L}(W, b)$

Automated parameterization of arbitrary function





7 hidden layers
200 nodes each
ReLU activation function

original function (black symbols):
fair description after 2800 training steps (purple)

- *Reality: function working in multi-dimensions*

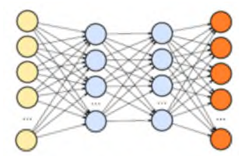
$$\vec{x} \in \mathbb{R}^n \rightarrow \vec{z} \in \mathbb{R}^m$$

- *Function: training is million-parameter fit*

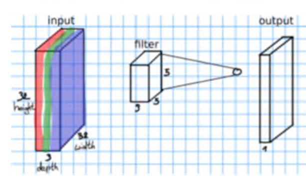
Deep Learning Progress

Concepts

– Fully connected

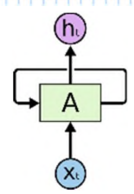


– Convolutional



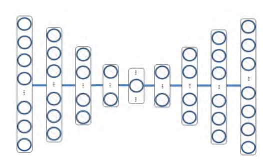
– Graph

– Recurrent

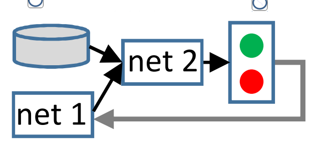


– Lorentz Boost Network

– Autoencoder



– Adversarial



– Reinforcement

– Invertible

Improved set of tools

Train millions of parameters by:

– Data preprocessing

– Normalization

– Regularization

– Short cuts

– Learning strategies

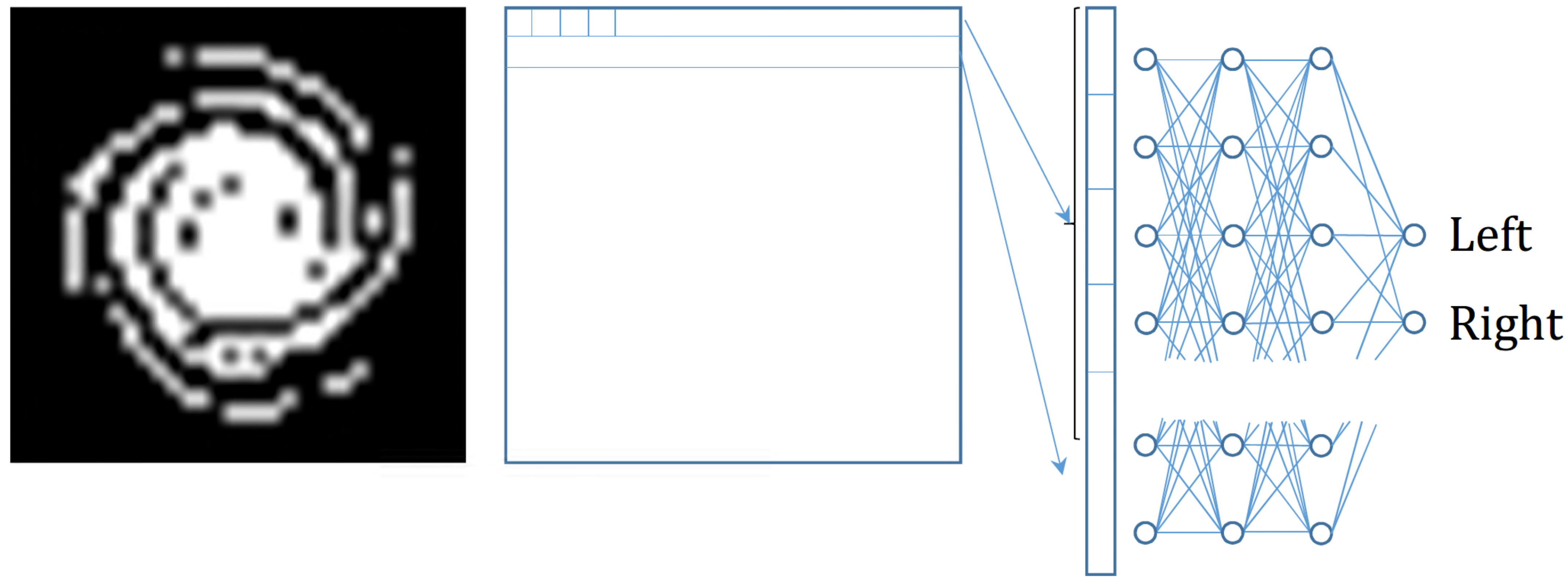
– …

Computing

– Graphics Processing Unit (GPU)

– Software Libraries

  – TensorFlow
  – keras…



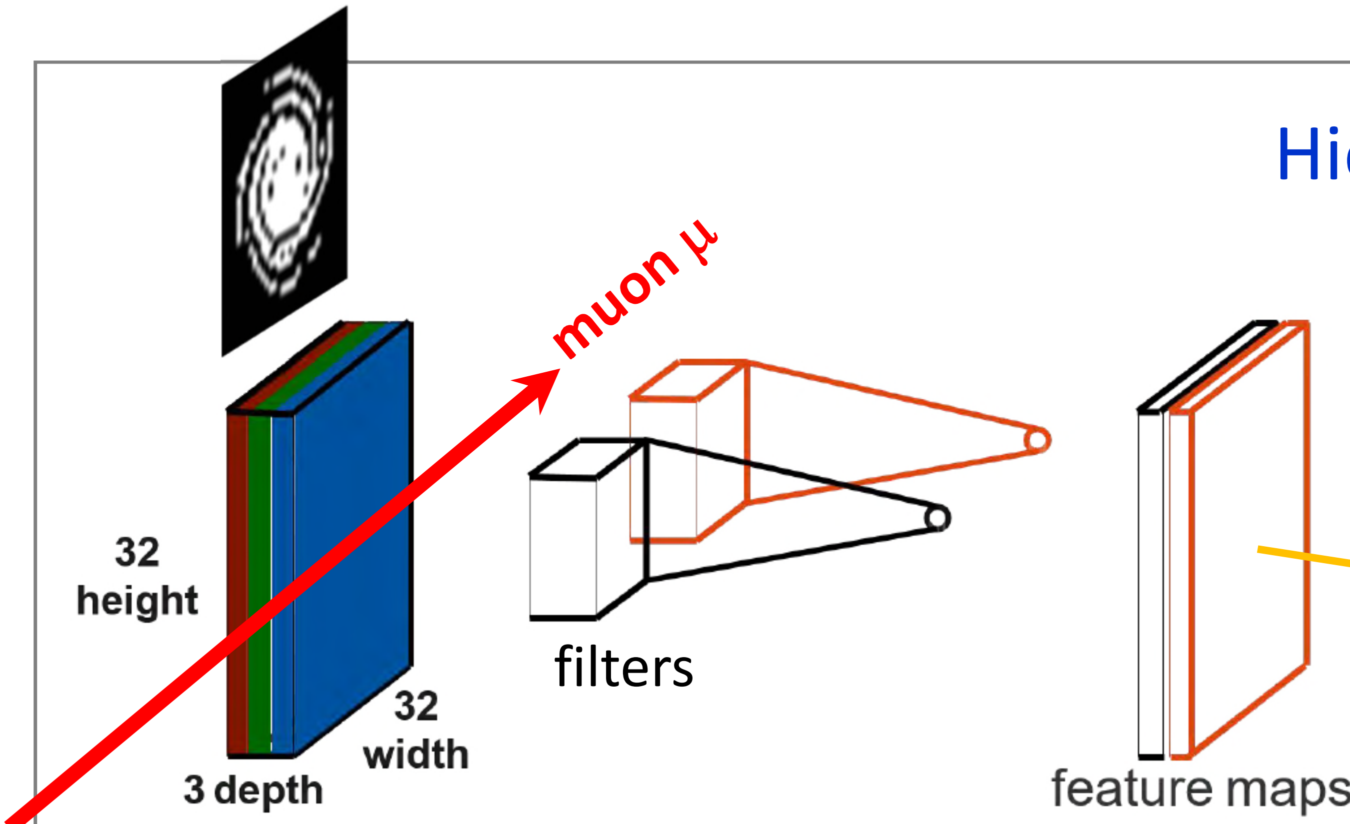
Deep learning & data symmetries

...looking for better ways than 1 pixel = 1 network input node

M. Zeiler, R. Fergus, Stochastic pooling for regularization of deep convolutional neural networks. ICLR, 2013.

Convolutional network to analyse image-like data

Hierarchy of ’features’ with increasing level of complexity



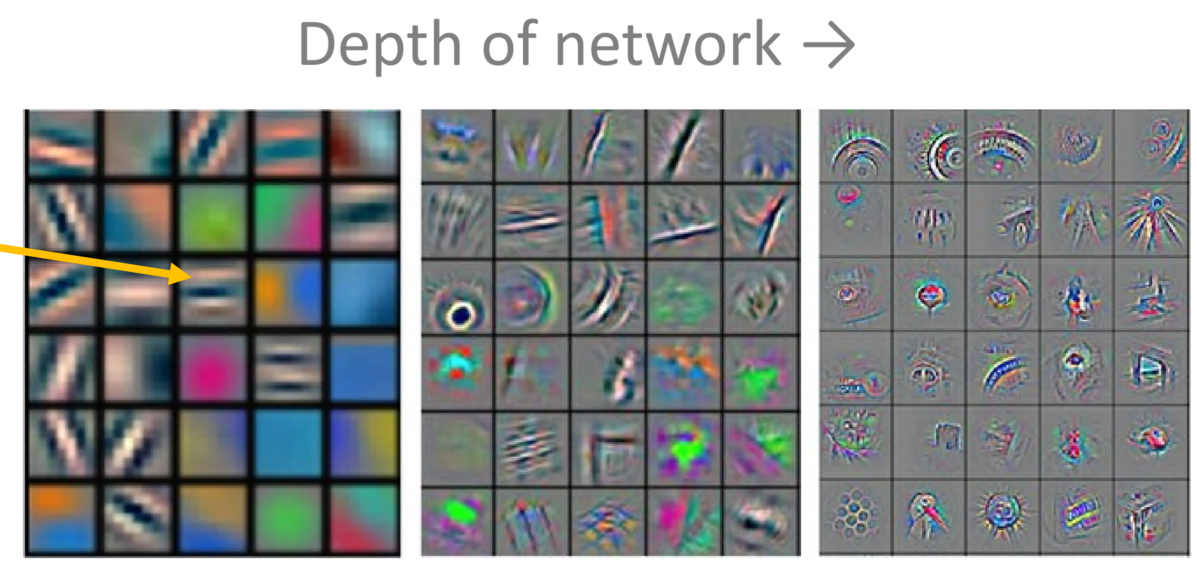
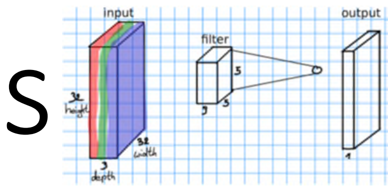


Image recognition

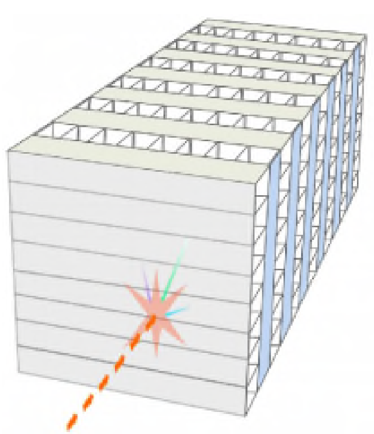
Pixel → edge → texton → motif → part → object

A. Aurisano et al., JINST 11 (2016) P09001

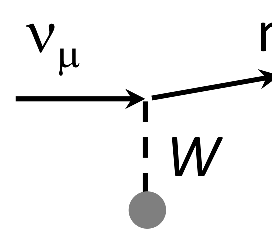
Convolutional network to identify electron neutrinos

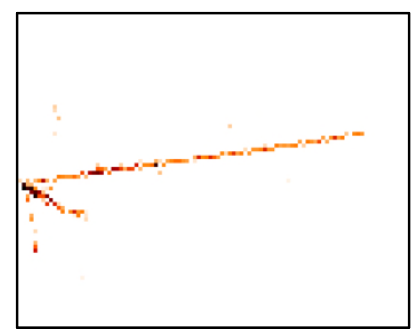


Fermilab (Chicago)
→NOvA experiment 810km



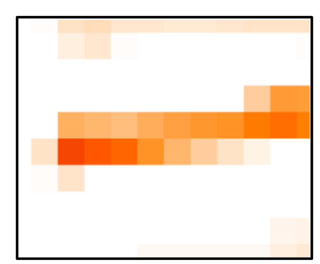
Neural network neutrino event classifier





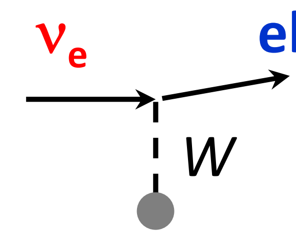
Feature maps

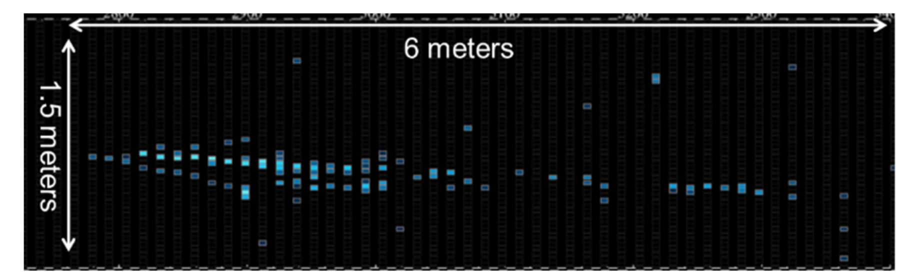
track



hadronic

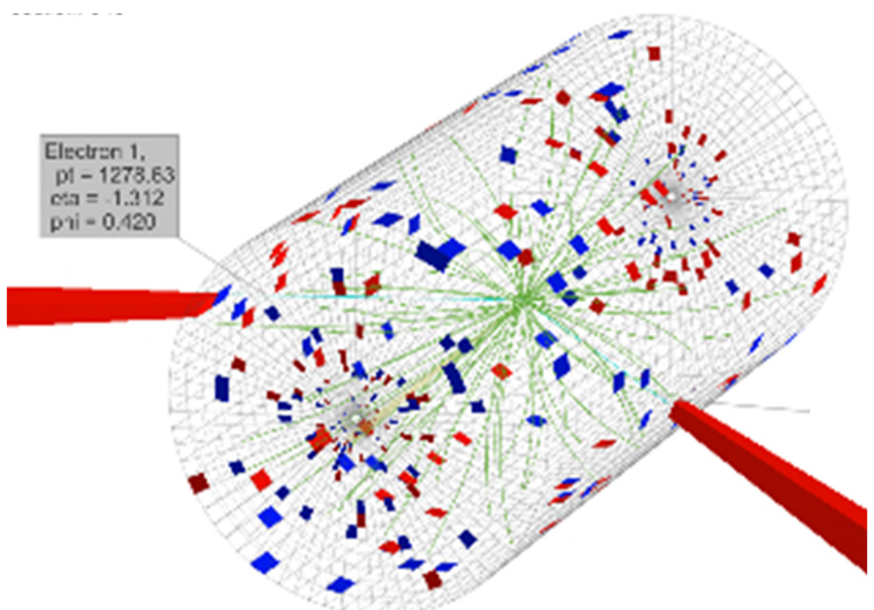
Challenge: electron-neutrinos

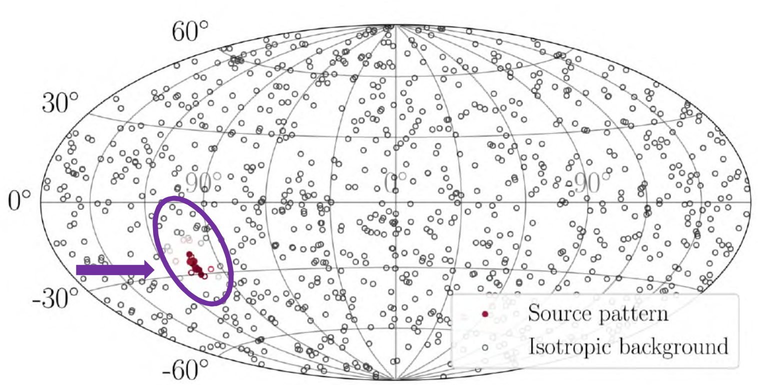




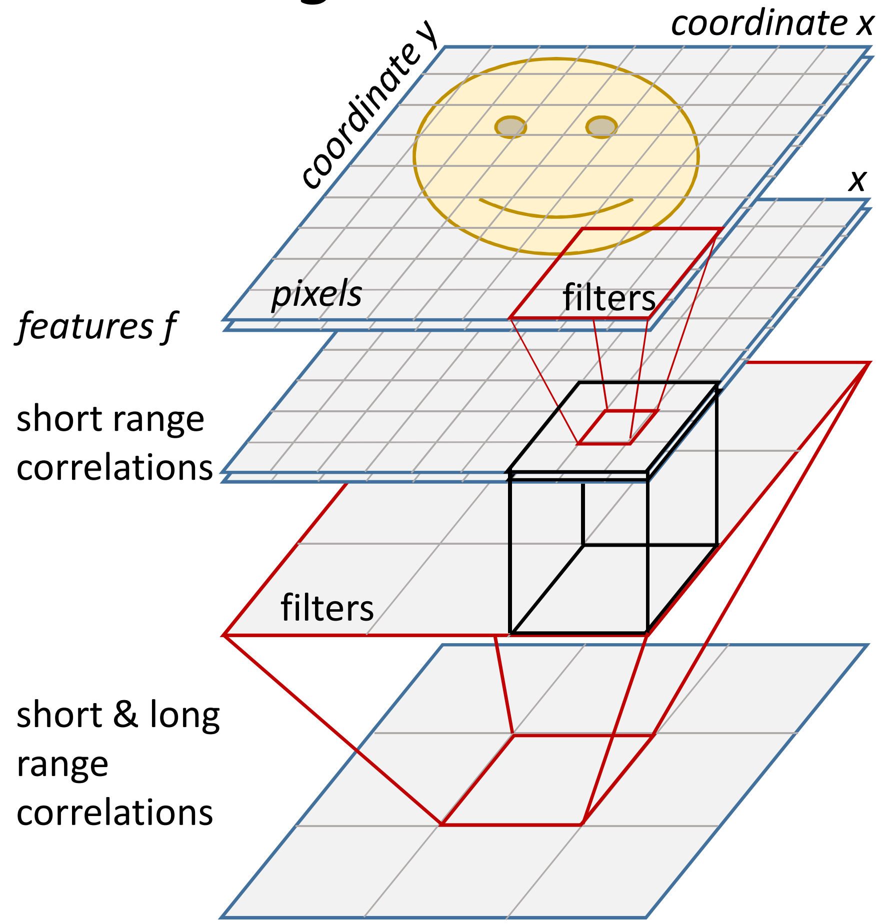
Method	**ν_e efficiency** (same purity)
Physicists algorithm	**35%**
Deep learning neural network	**49%**

Convolution: *Classic* versus *Graph* network



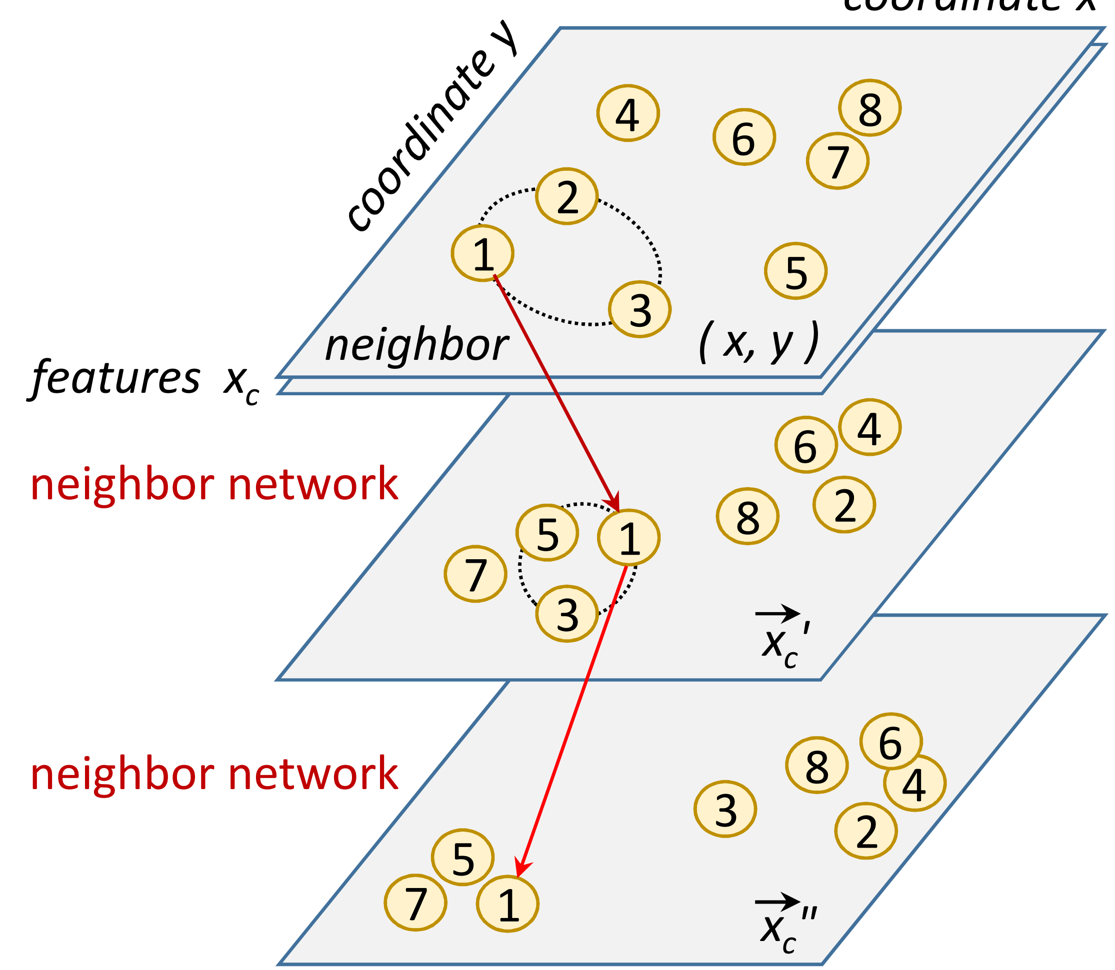


Classic: grid



- dense pixel occupation
- object identification
- analysis: single pixel contributions

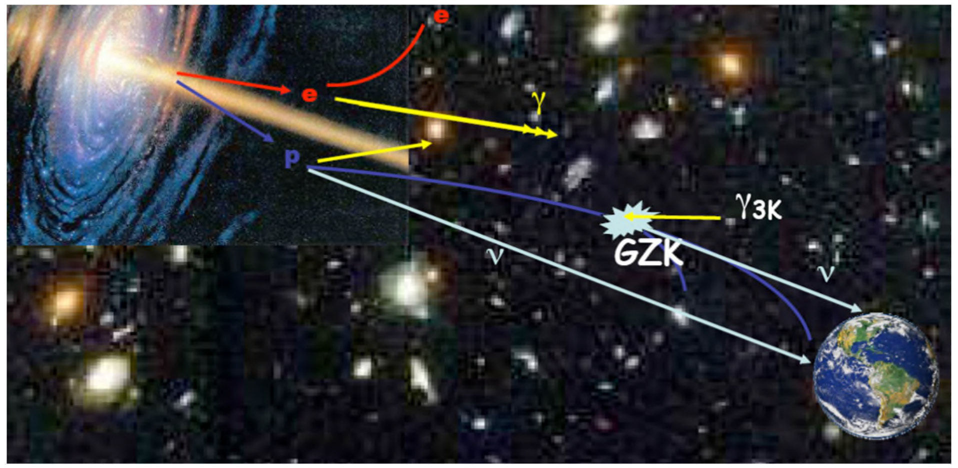
Graph: neighbors



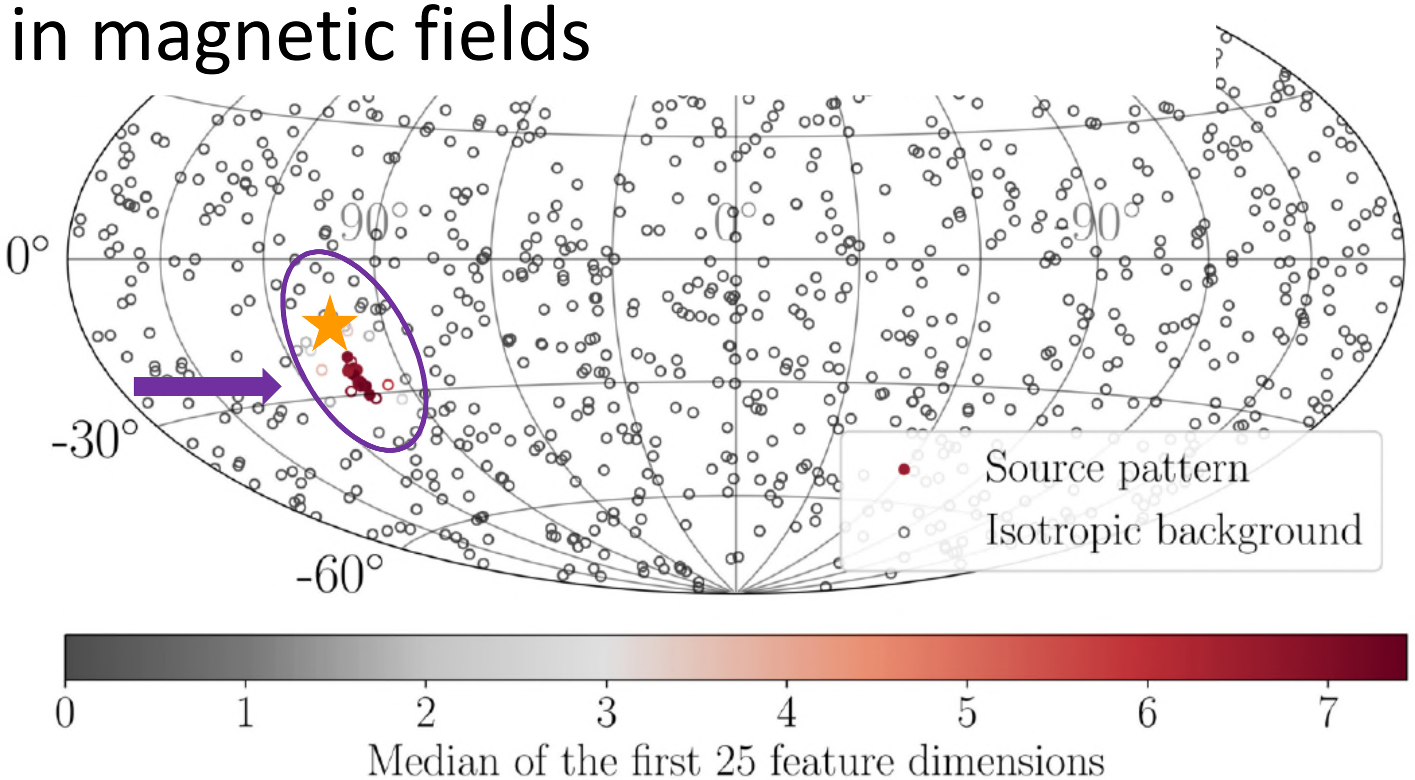
- sparse or dense neighbor occupation
- pattern recognition
- single objects 1 identification

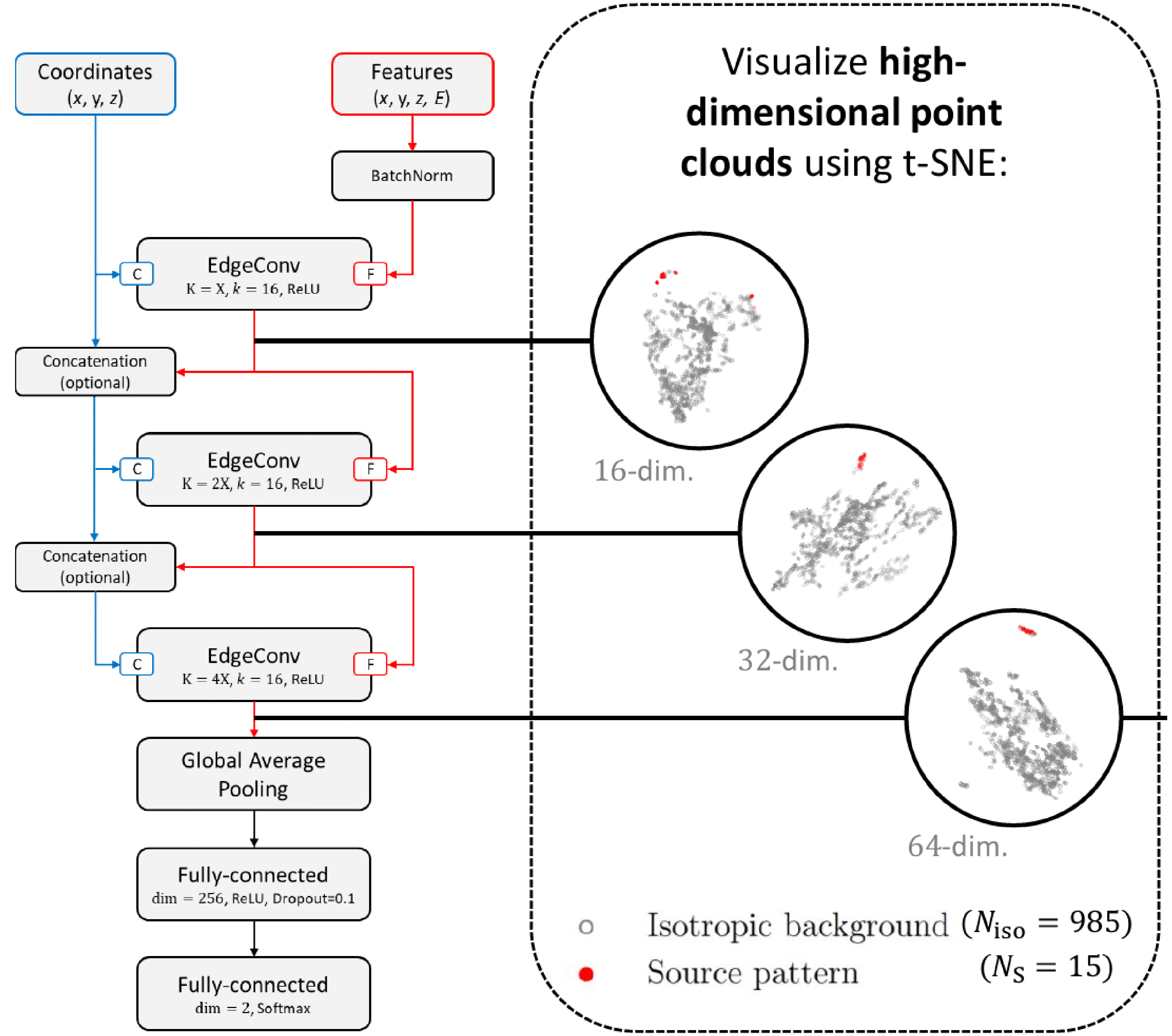
T. Bister, M. Erdmann, J. Glombitza, N. Langner, J. Schulte, M. Wirtz, arXiv:2003.13038

Graph convolutions to detect cosmic magnetic fields

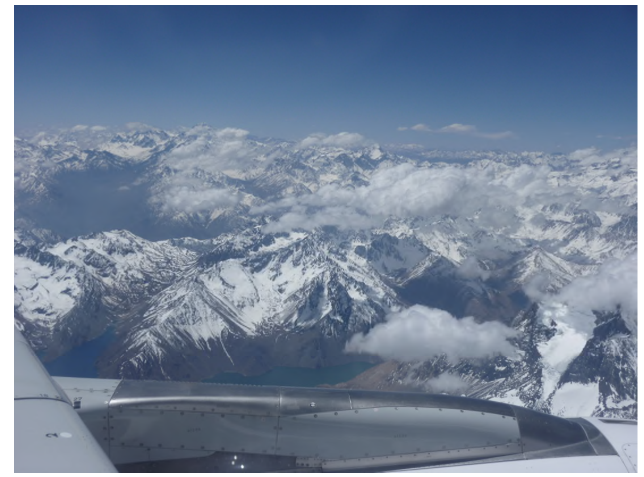


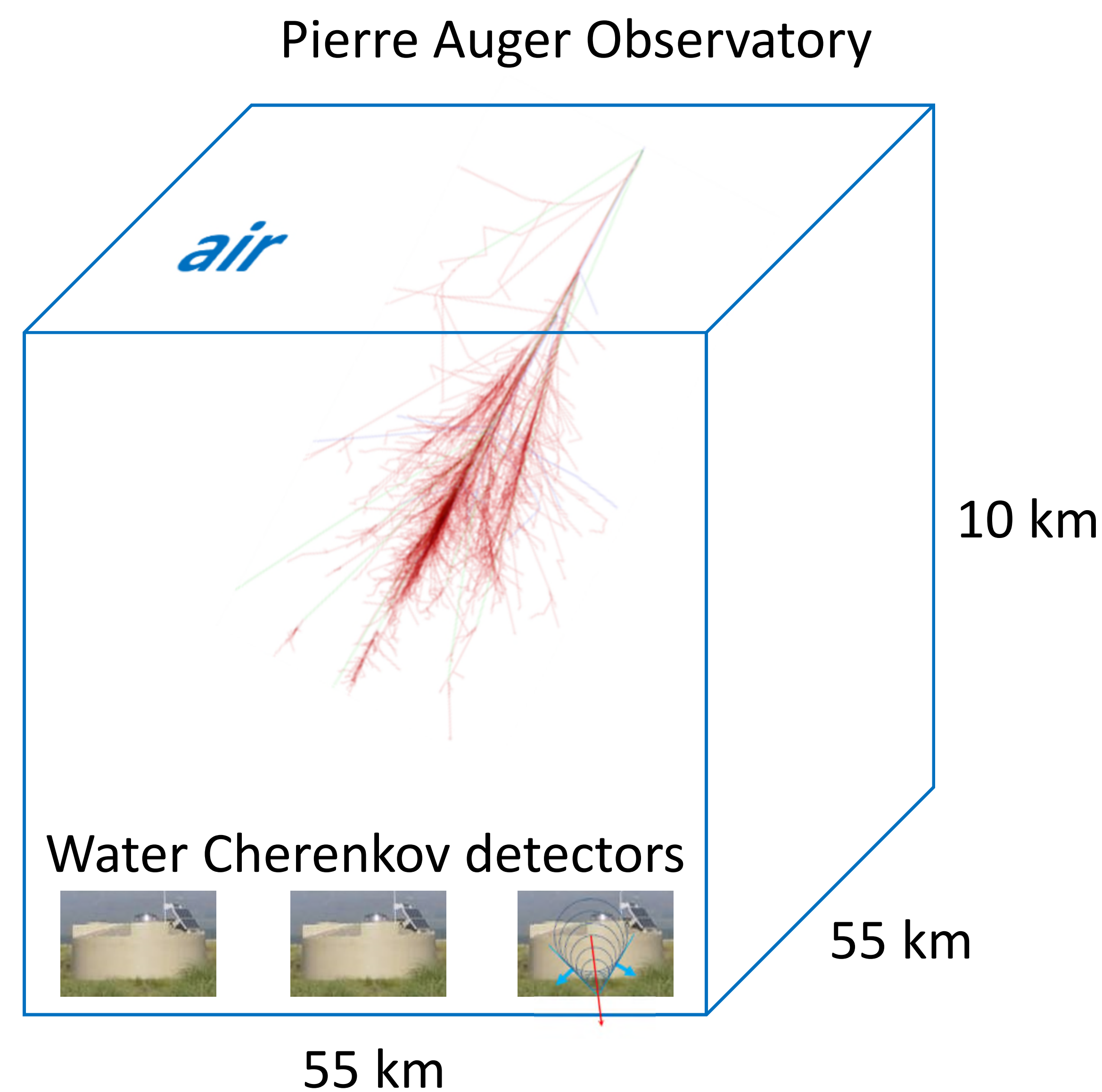
pattern: deflections of cosmic rays
in magnetic fields





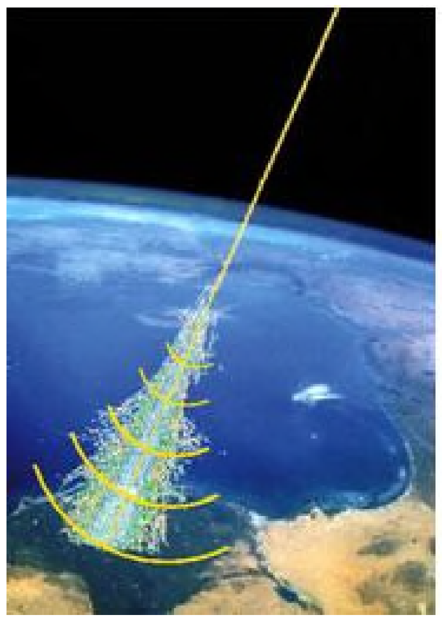
World's largest Calorimeter for Cosmic Rays

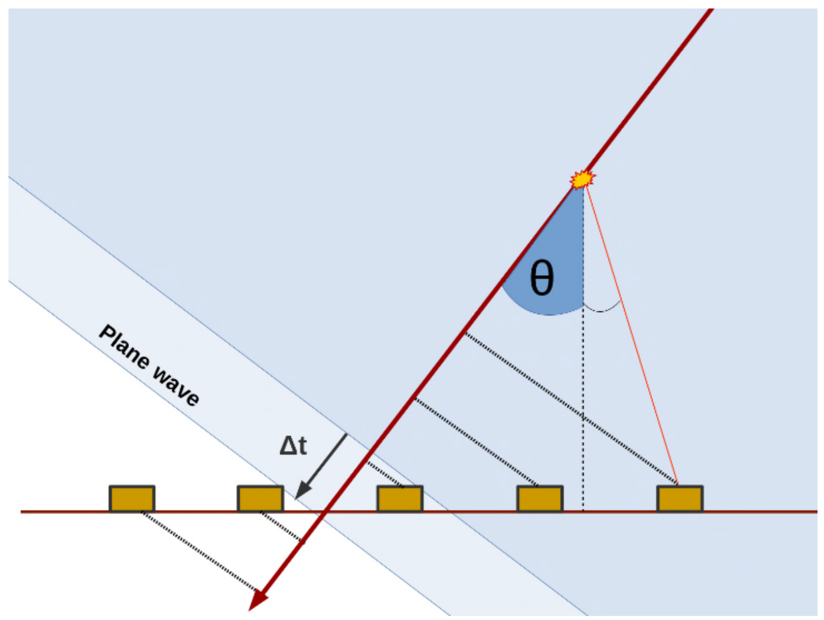


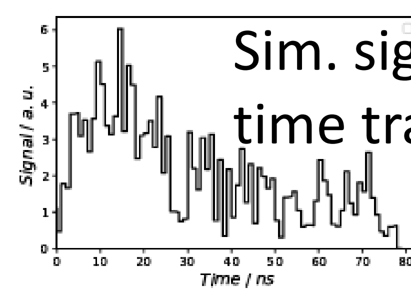


M. Erdmann, Jonas Glombitza, David Walz, 10.1016/j.astropartphys.2017.10.006

Cosmic ray arrival directions by physicist or network





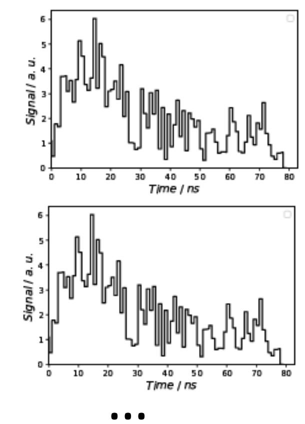


Sim. signal time traces

Educated physicist:

- Time offset between detectors
- Particle velocity
- Detector distance
- Plane fit of time residuals

RAW input data:

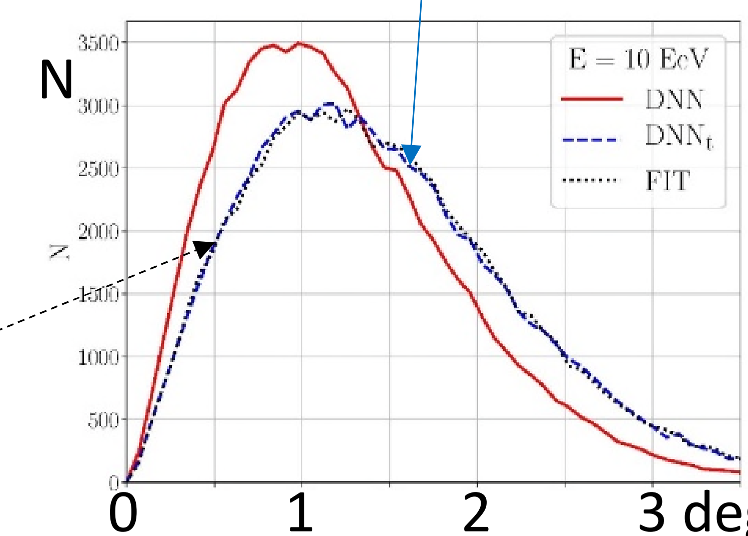


Deep Neural Network
No physics education
No explicit information about
- *locations of detectors*
- *speed of light*

Needs data with true target θ

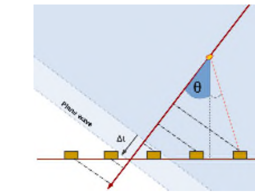
- Time offset only

shower direction angular resolution



Deep Neural Network learns physics from data within 3h

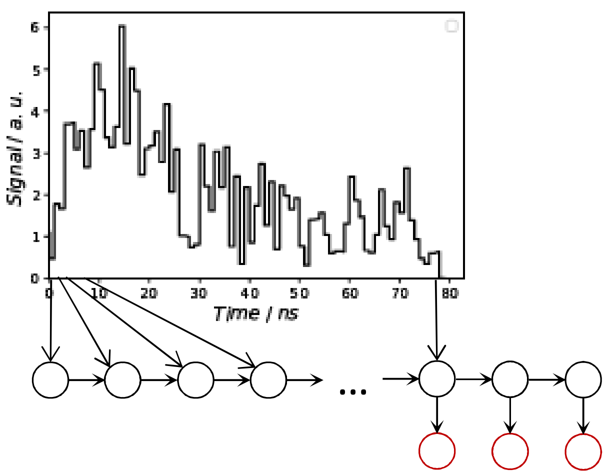
M. Erdmann, Jonas Glombitza, David Walz, 10.1016/j.astropartphys.2017.10.006
Jonas Glombitza for the Pierre Auger Collaboration arXiv:2101.02946



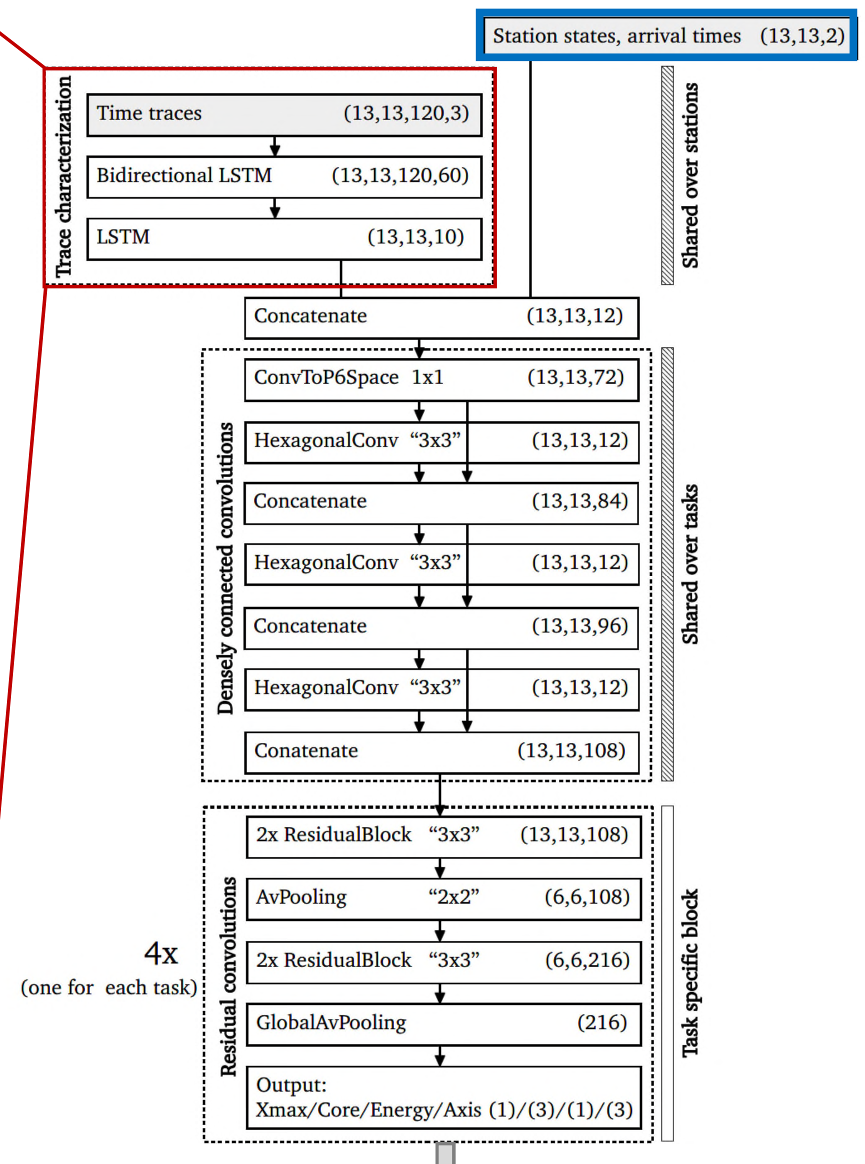
Recurrent network to characterize signal traces

Time offset & total signal

Characterization of signal traces

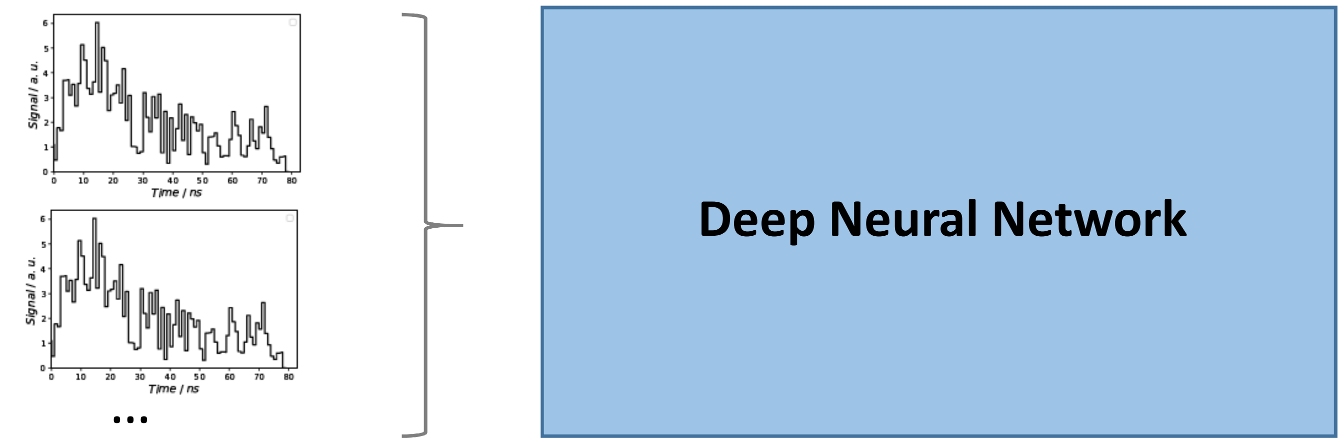


Network extracts from training data *optimized intermediate variables* suited for shower direction



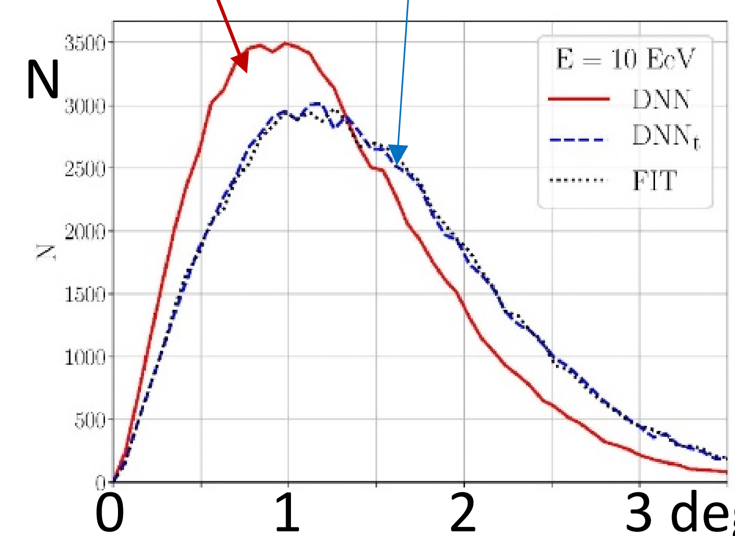
shower direction

RAW input data:



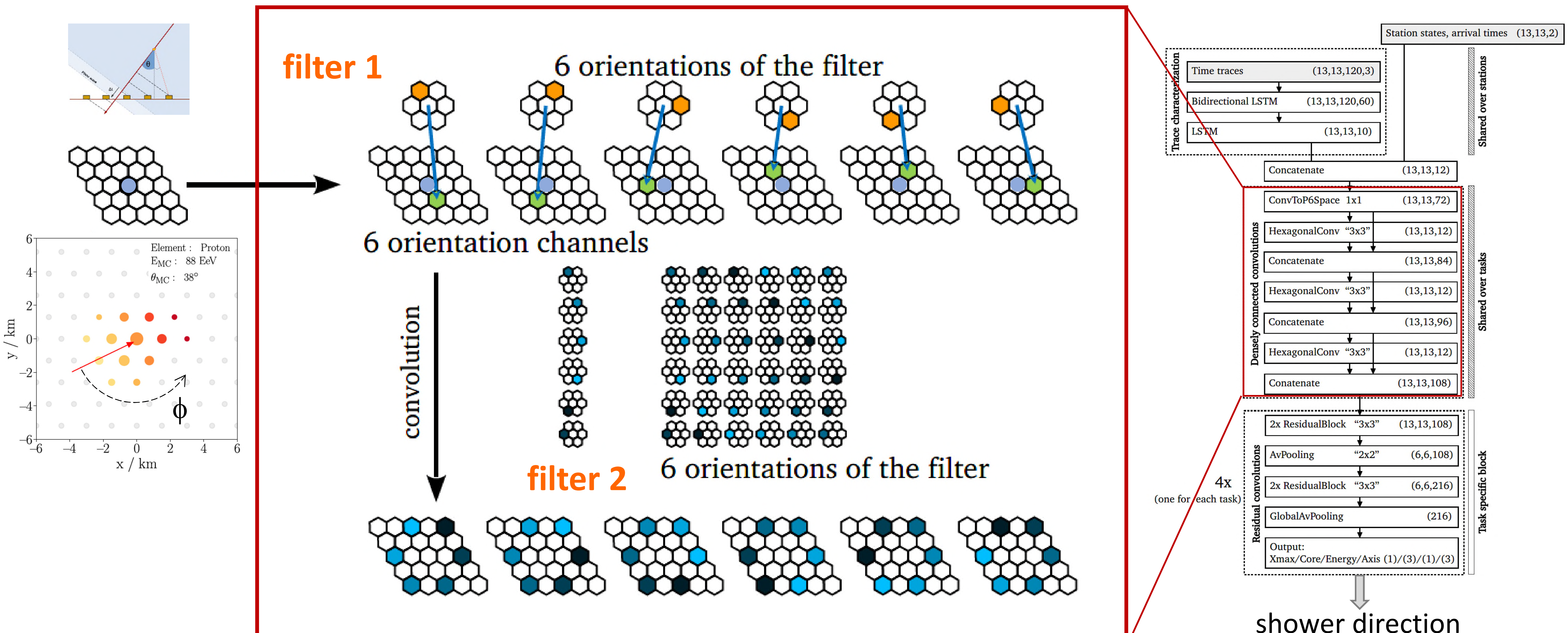
- Time offset only
- Signal traces added

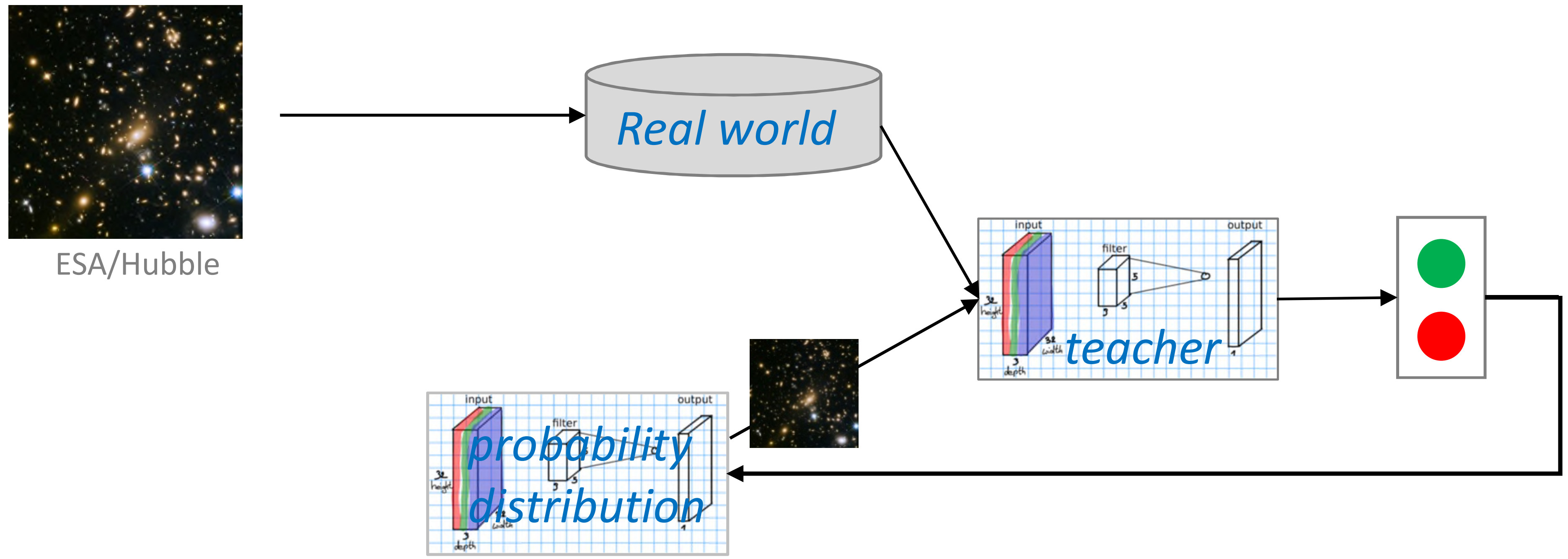
shower direction angular resolution



Emiel Hoogeboom, Jorn W.T. Peters, Taco S. Cohen, Max Welling, arXiv:1803.02108
Jonas Glombitza for the Pierre Auger Collaboration arXiv:2101.02946

Hexagonal convolutions to symmetrize azimuth ϕ





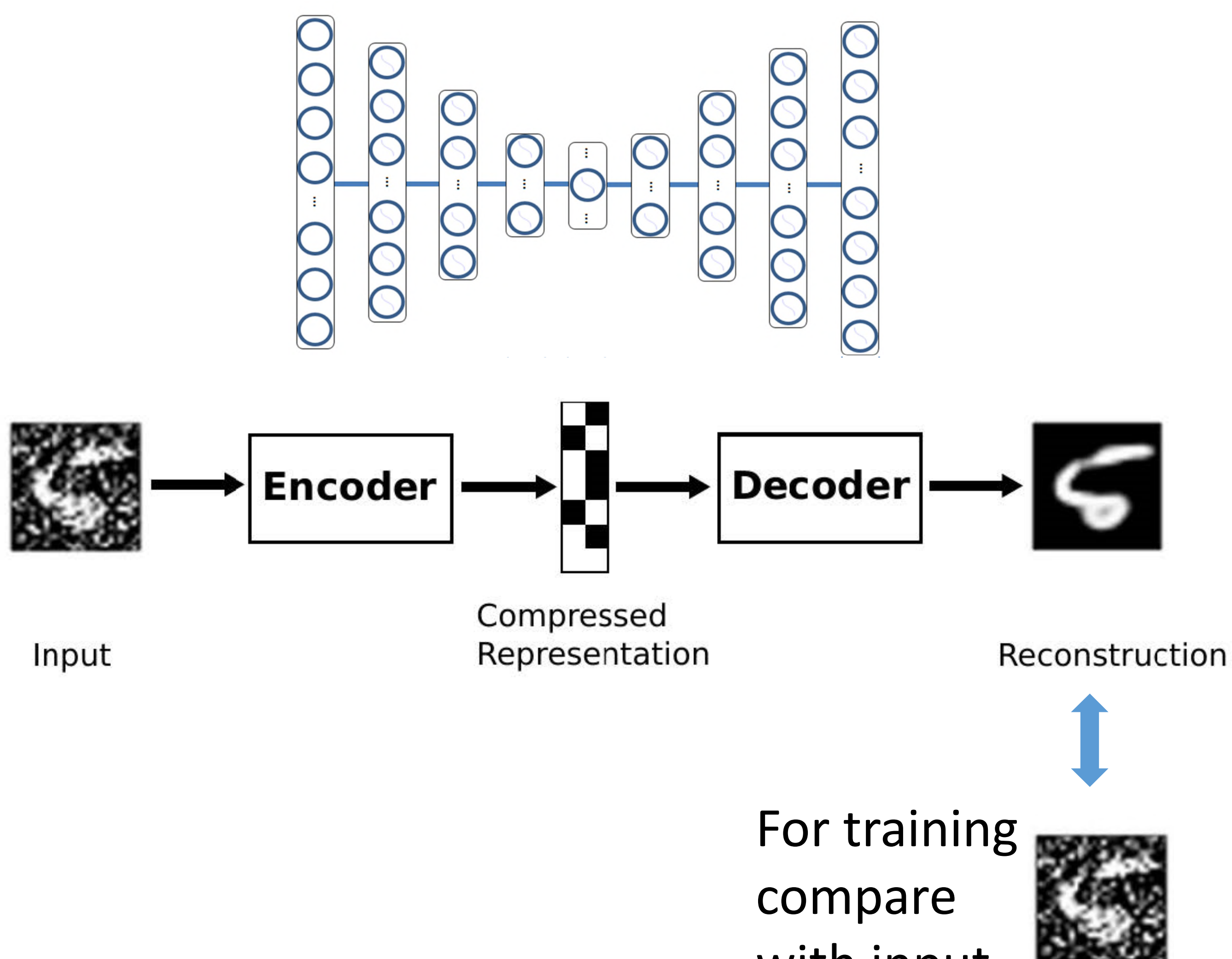
Autonomous model building

Assign functional target → training data optimize network (`unsupervised´)

Autoencoder networks to identify new physics

T. Heimel, G. Kasieczka, T. Plehn, J.M. Thompson, SciPost Phys. 6, 030 (2019)

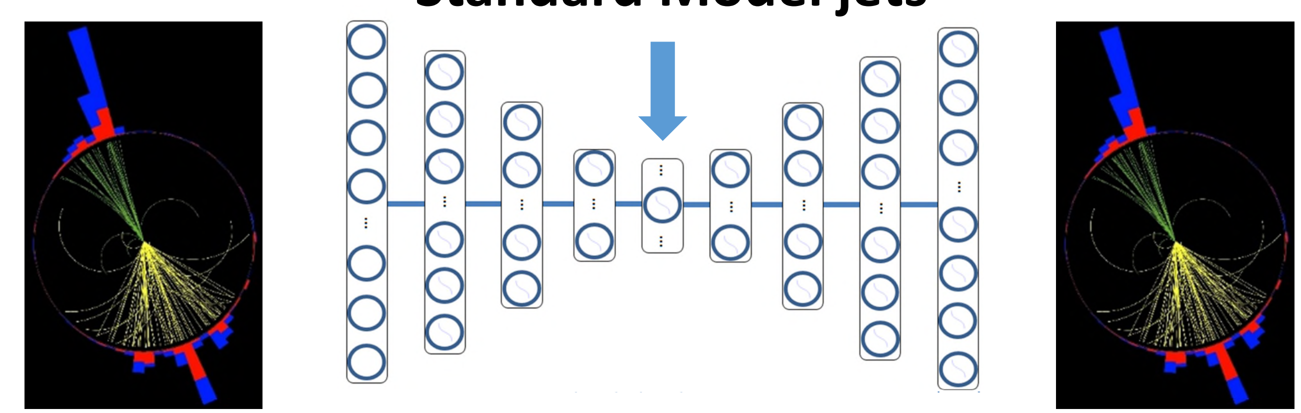
Denoising ('unsupervised')



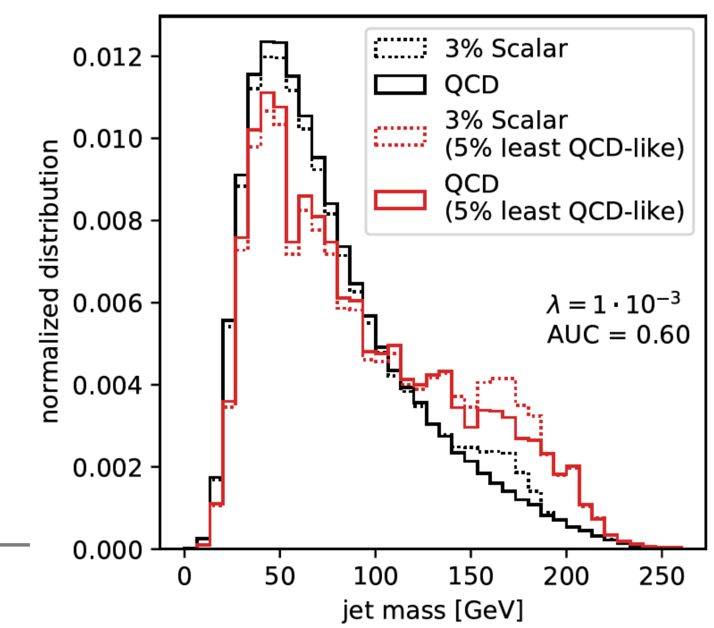
For training compare with input

Search for new physics

Train compressed representation: **Standard Model jets**



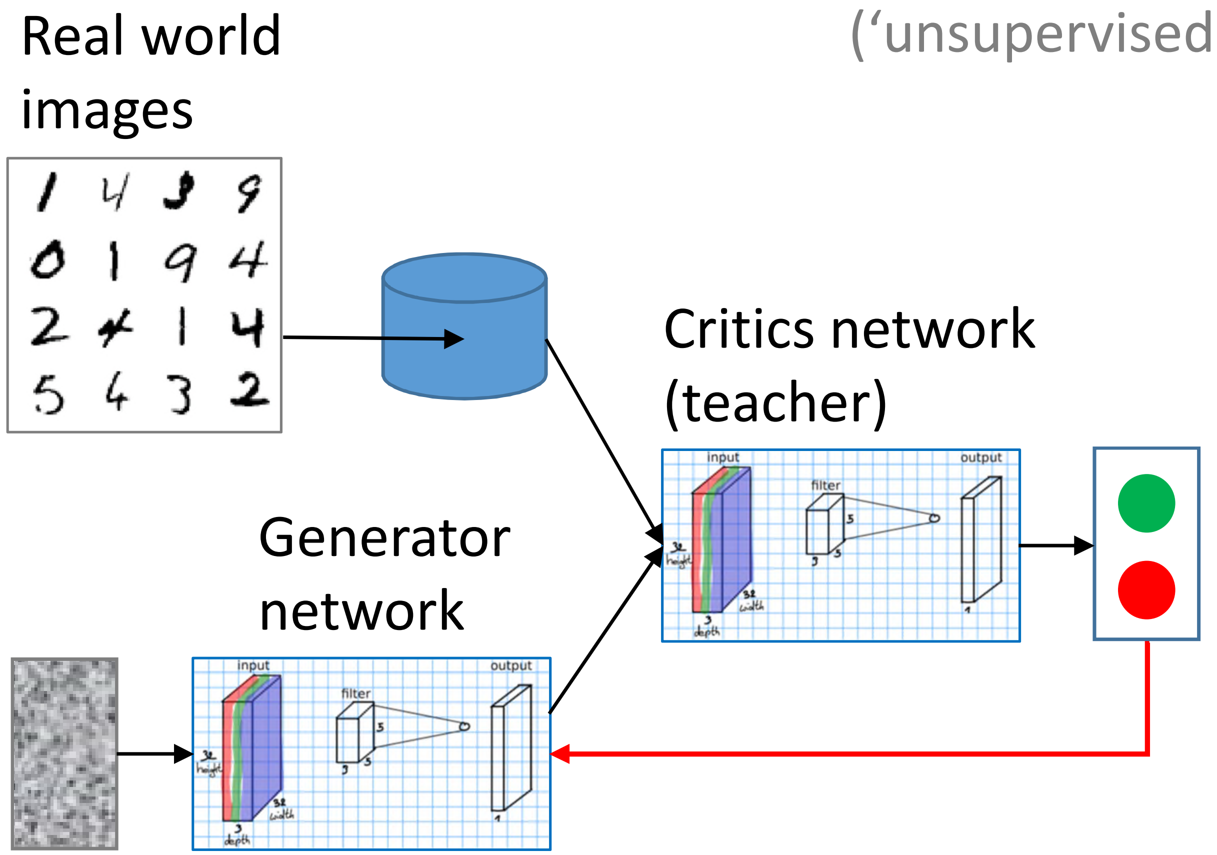
Evaluation of jets: ***Disagreement*** *with input jet indicates* ***new physics***



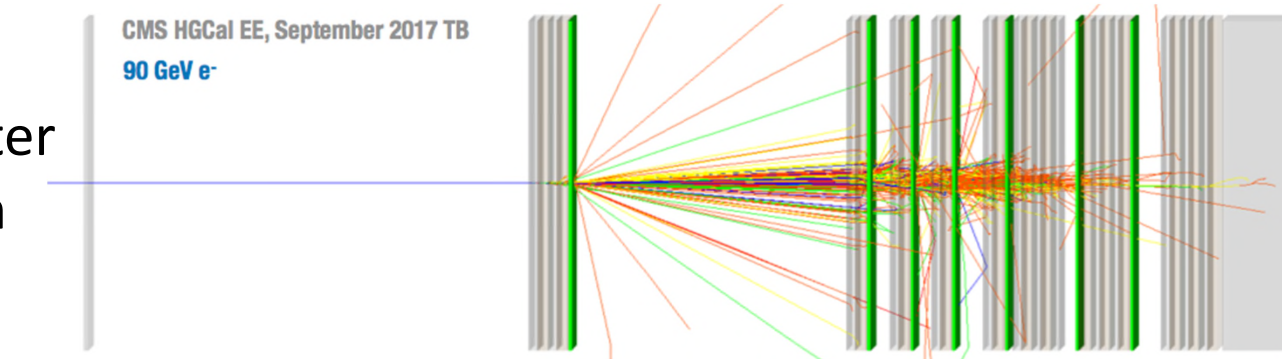
Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua BengioarXiv:1406.2661
M. Erdmann, J. Glombitza, T. Quast, Comput Softw Big Sci (2019) 3: 4

Generative Modeling to simulate particle showers

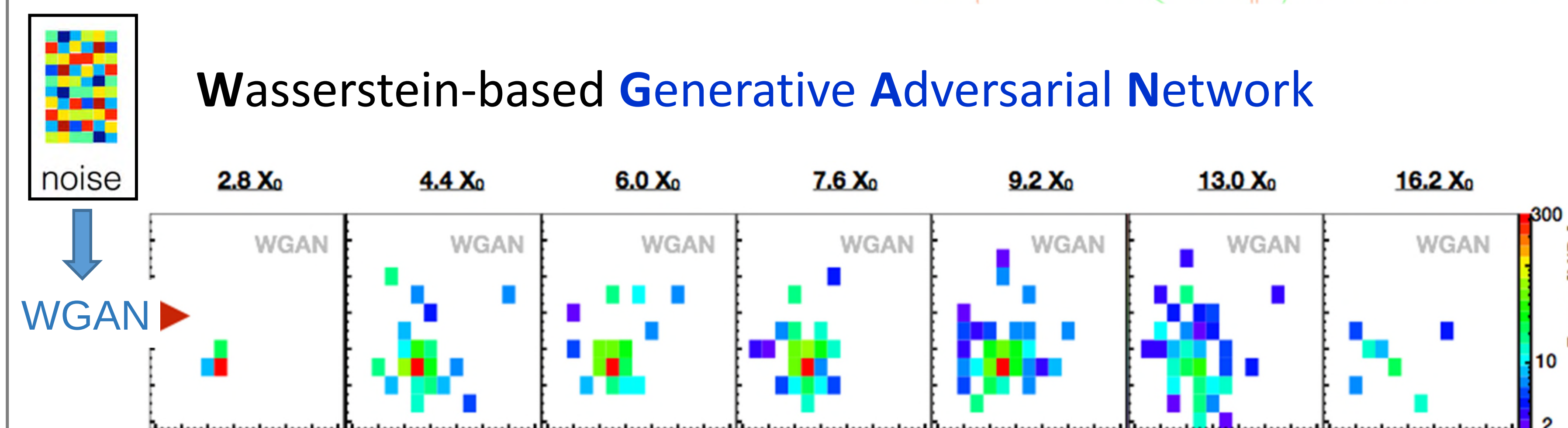
('unsupervised')



Electron calorimeter in CERN test beam



Wasserstein-based **G**enerative **A**dversarial **N**etwork



Generation Method	Hardware	milliseconds/ shower
GEANT4	CPU	**2000**
WGAN	CPU	52
	GPU	**0.3**

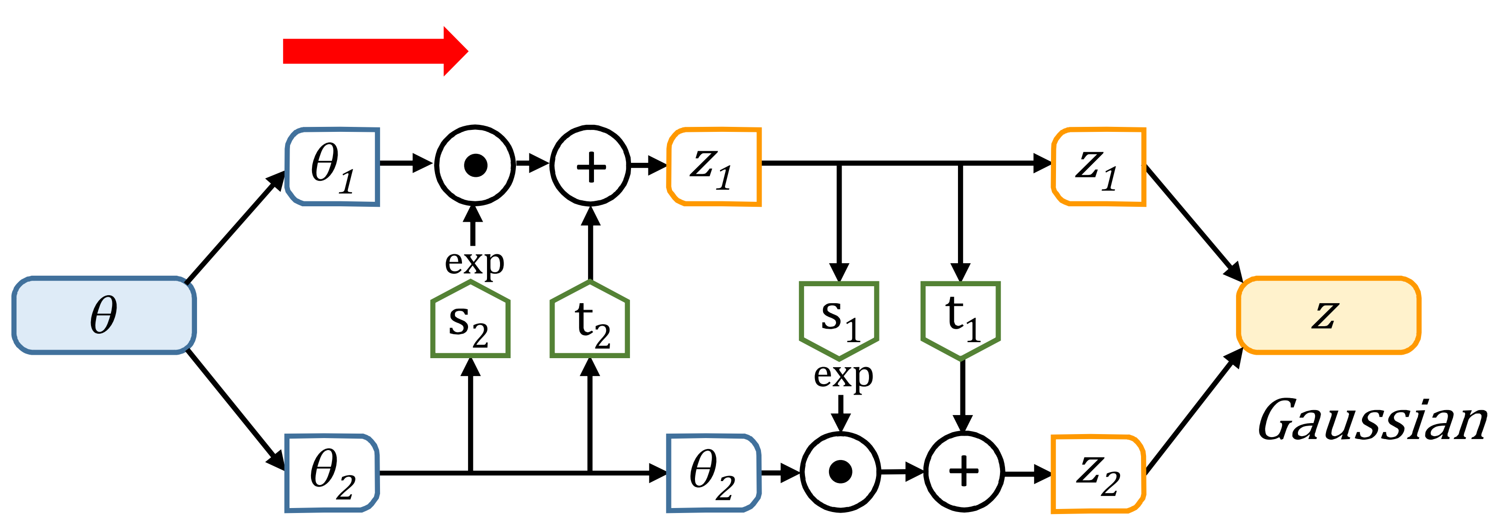
R&D: Ultrafast detector simulations of high quality.
Huge effort ahead for production versions

Invertible Networks to map probability distributions

S. Radev, U. Mertens, A. Voss, L. Ardizzone, U. Köthe, arxiv 2003.06281 J. Schulte, T. Bister, M. Erdmann, RWTH Aachen
M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, U. Köthe, SciPost Phys. 9, 074 (2020)

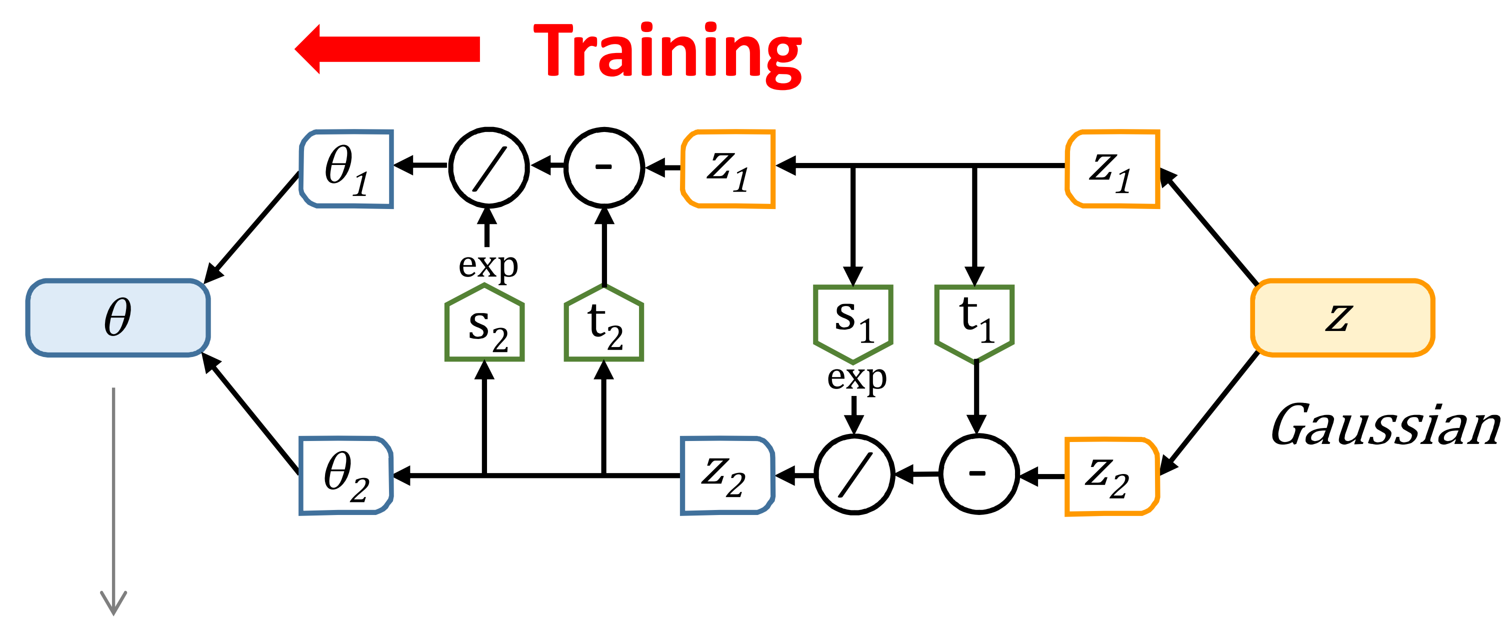
$\boxed{\theta} = (\gamma, R_{\mathrm{cut}})$**@source**

$J_{\mathrm{inj}}(E) \propto E^{-\gamma} \cdot f_{\mathrm{cut}}(E, Z \cdot R_{\mathrm{cut}})$



Probability conservation $\boxed{p_\phi(\theta)\ d\theta} = \boxed{p(z)\ dz}$

Training

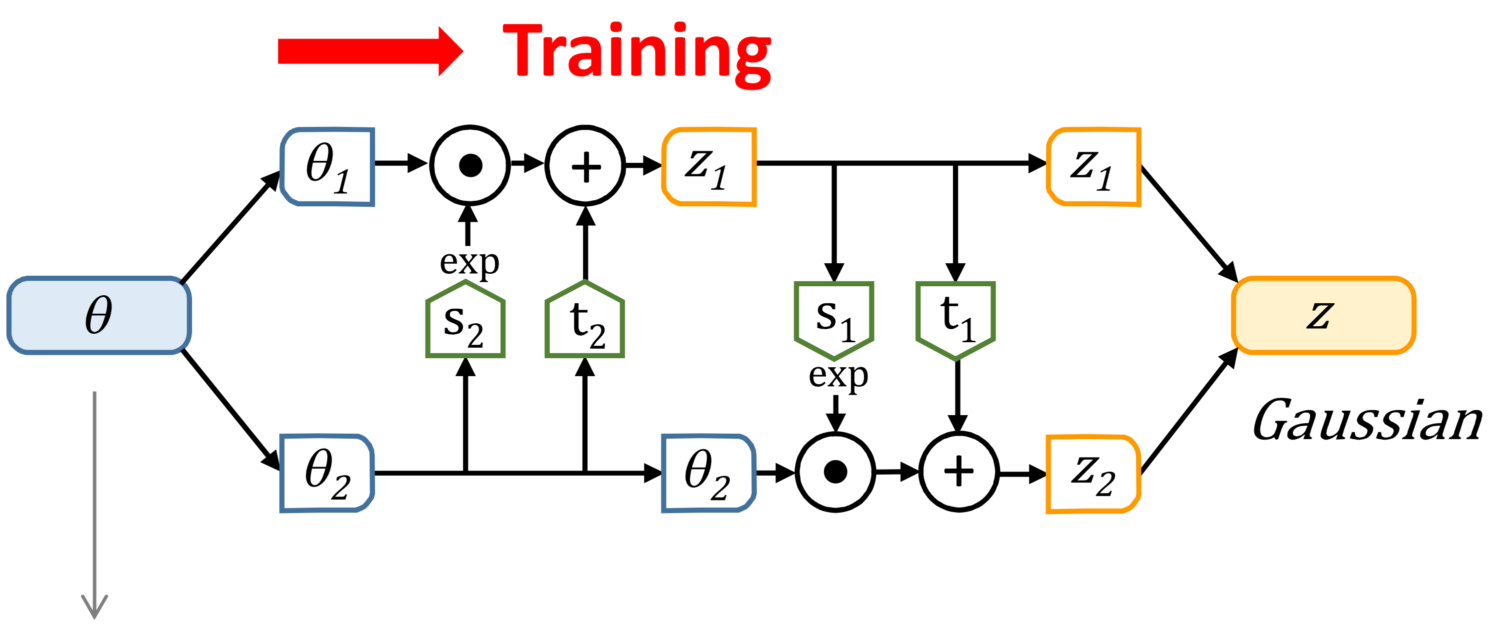


$$\underset{\phi}{\operatorname{argmin}}\ \mathbb{E}_{p(\boldsymbol{\theta}\,|\,\boldsymbol{x})}\left[\log \overbrace{p(\boldsymbol{\theta}\,|\,\boldsymbol{x})}^{\text{sim.data}} - \log \overbrace{\boxed{p_\phi(\boldsymbol{\theta}\,|\,\boldsymbol{x})}}^{\text{network}}\right]$$

Invertible Networks to map probability distributions

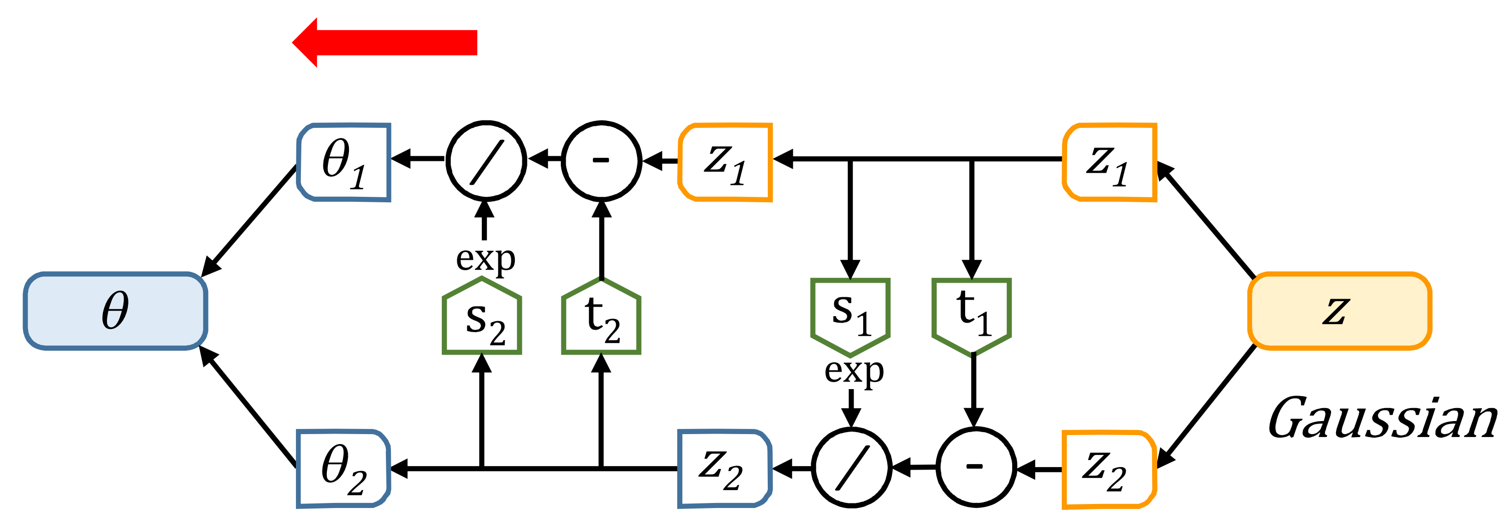
$\boxed{\theta} = (\gamma, R_{\mathrm{cut}})$**@source**

$J_{\mathrm{inj}}(E) \propto E^{-\gamma} \cdot f_{\mathrm{cut}}(E, Z \cdot R_{\mathrm{cut}})$



$$\underset{\phi}{\operatorname{argmin}}\ \mathbb{E}_{p(\boldsymbol{\theta}\,|\,\boldsymbol{x})}\, [\overbrace{\log p(\boldsymbol{\theta}\,|\,\boldsymbol{x})}^{\text{sim.data}} - \log \overbrace{\boxed{p_\phi(\boldsymbol{\theta}\,|\,\boldsymbol{x})}}^{\text{network}}]$$

Probability conservation $\boxed{p_\phi(\theta)\ d\theta} = \boxed{p(z)\ dz}$

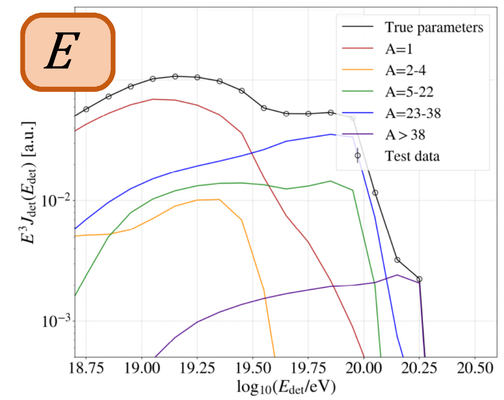


Invertible Networks to unfold observed data

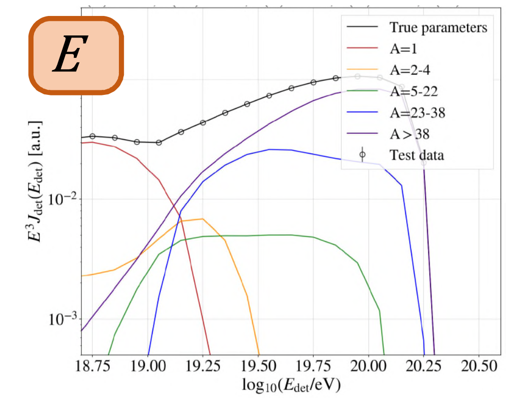
$\theta = (\gamma, R_{\mathrm{cut}})$**@source**

$J_{\mathrm{inj}}(E) \propto E^{-\gamma} \cdot f_{\mathrm{cut}}(E, Z \cdot R_{\mathrm{cut}})$

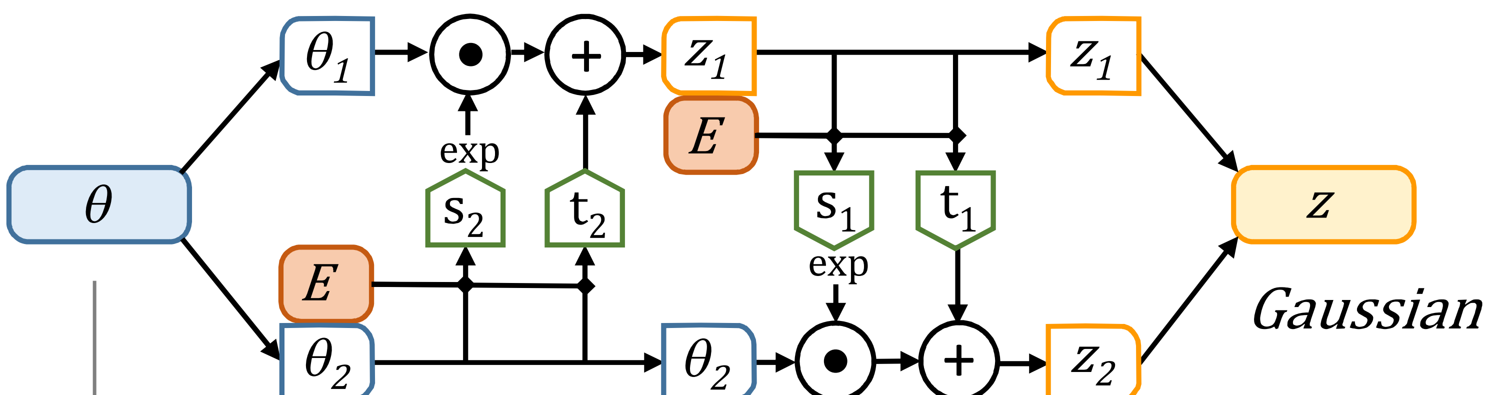
Energy E **@observation**



`measurement´

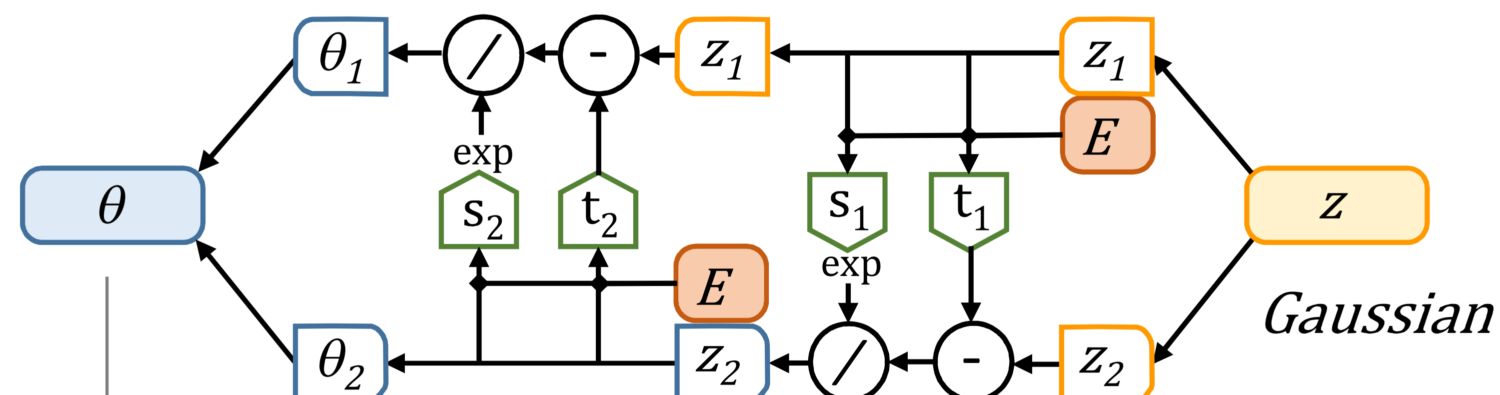


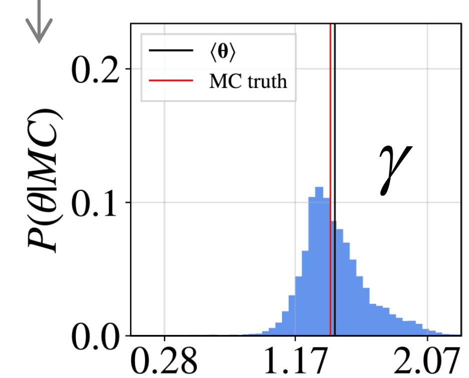
Training



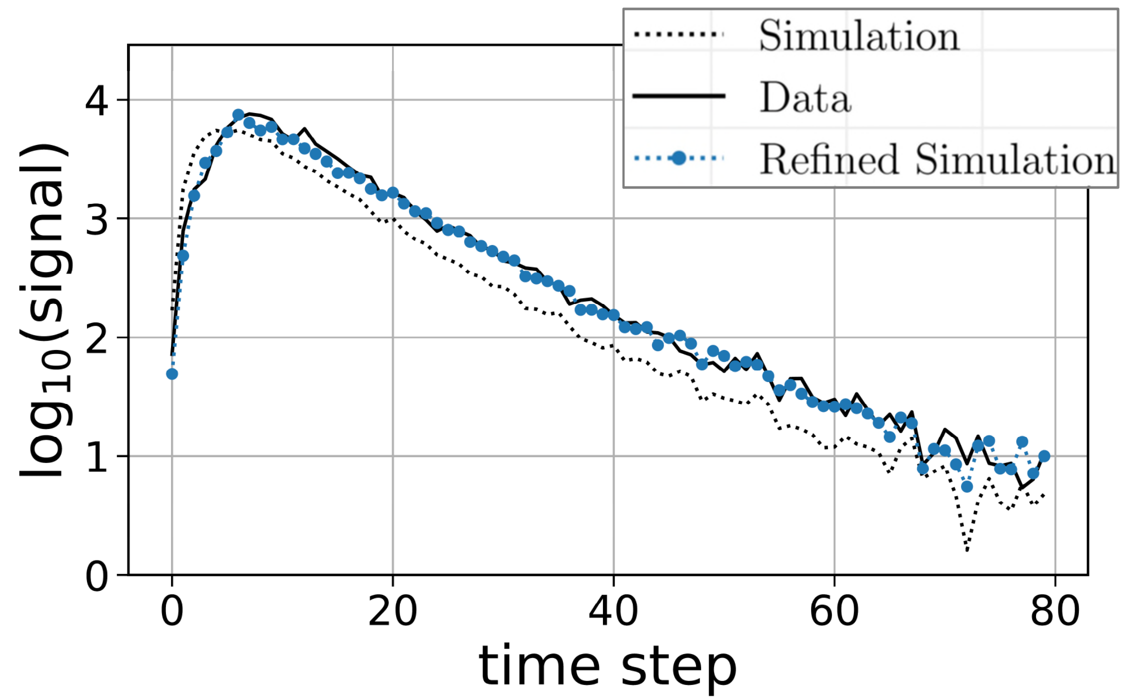
$$\underset{\phi}{\operatorname{argmin}}\ \mathbb{E}_{p(\boldsymbol{\theta}\,|\,\boldsymbol{x})}\,[\overset{\text{sim.data}}{\log p(\boldsymbol{\theta}\,|\,\boldsymbol{x})} - \log \overset{\text{network}}{p_{\phi}(\boldsymbol{\theta}\,|\,\boldsymbol{x})}]$$

Evaluation





Network learns to extract posterior distribution from given data

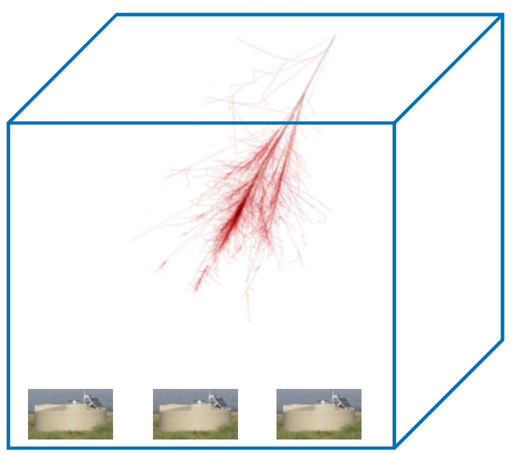


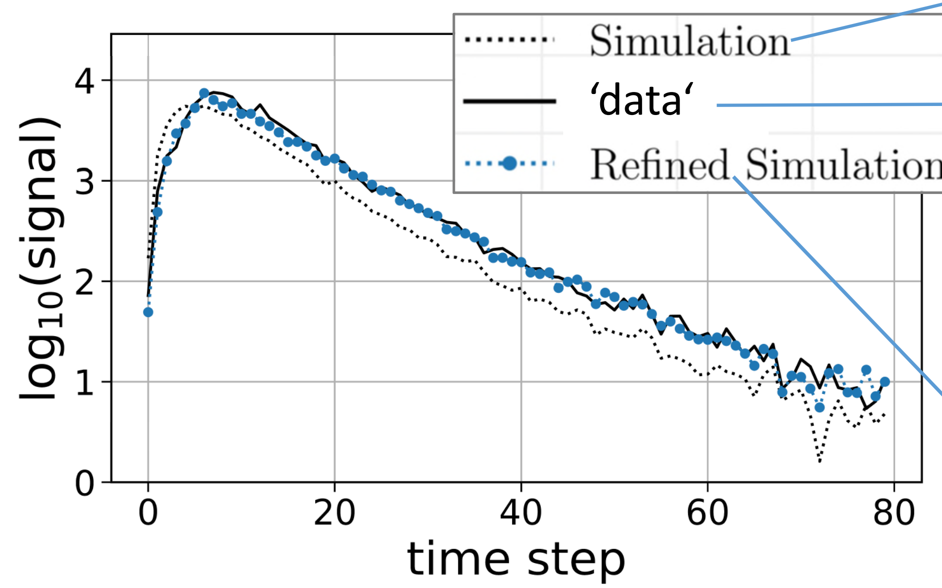
Operation reality & network insight

Solved long-standing machine learning problem data ≠ simulation

Domain adaption to simulate data-like

data ≠ simulation ➡ Refine simulated traces

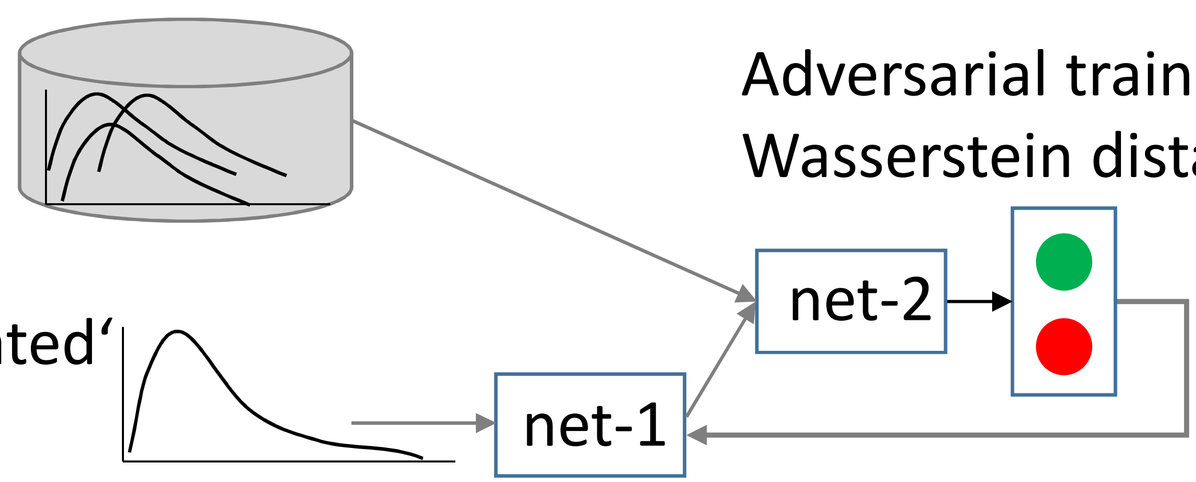




'data' traces
(simulated)

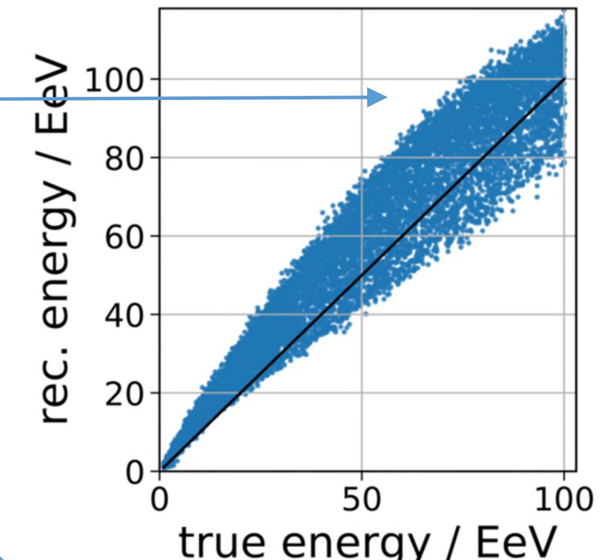
'badly simulated' trace

Adversarial training
Wasserstein distance



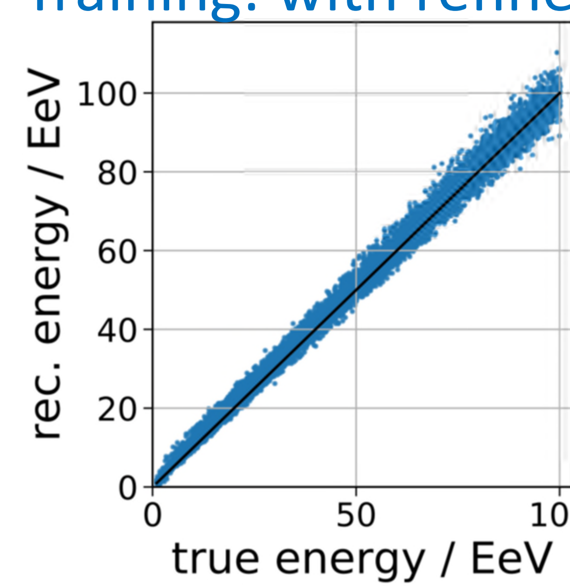
Energy reconstruction

Training: badly simulated traces



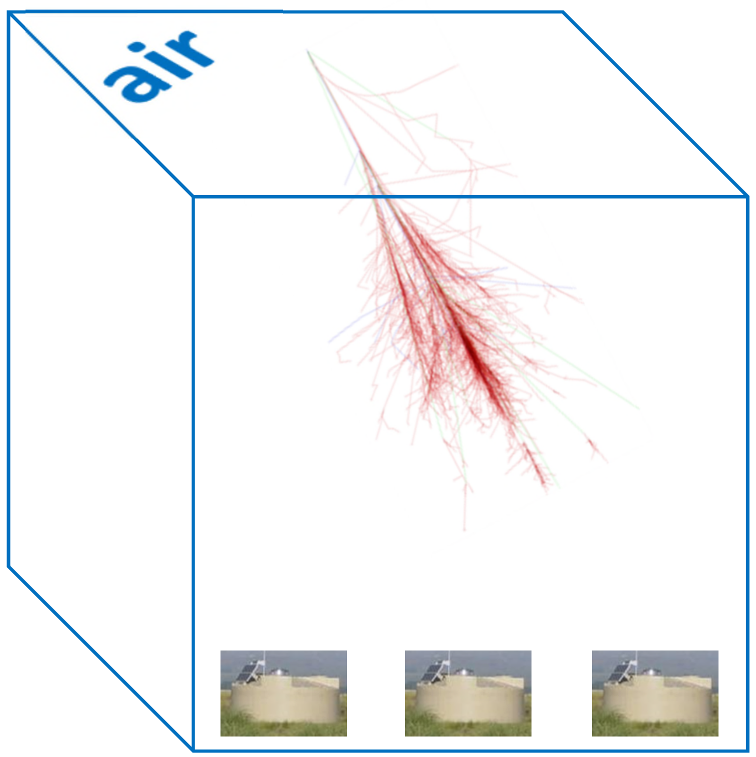
Deep Network net-3: cosmic ray energy from 'data' traces

Training: with refined simulation



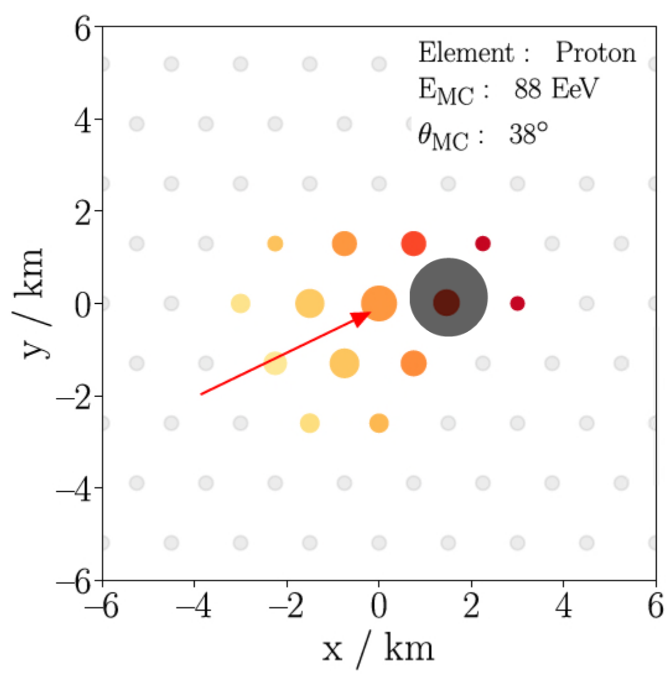
improved energy resolution

Simulation: include operation reality for network training

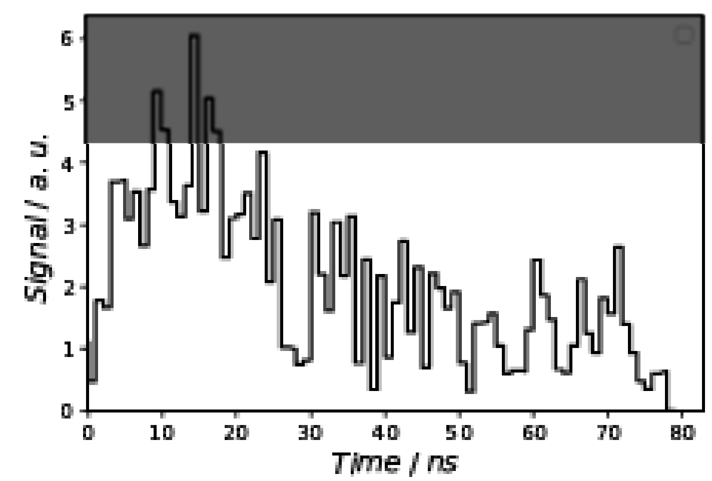


Pierre Auger Observatory

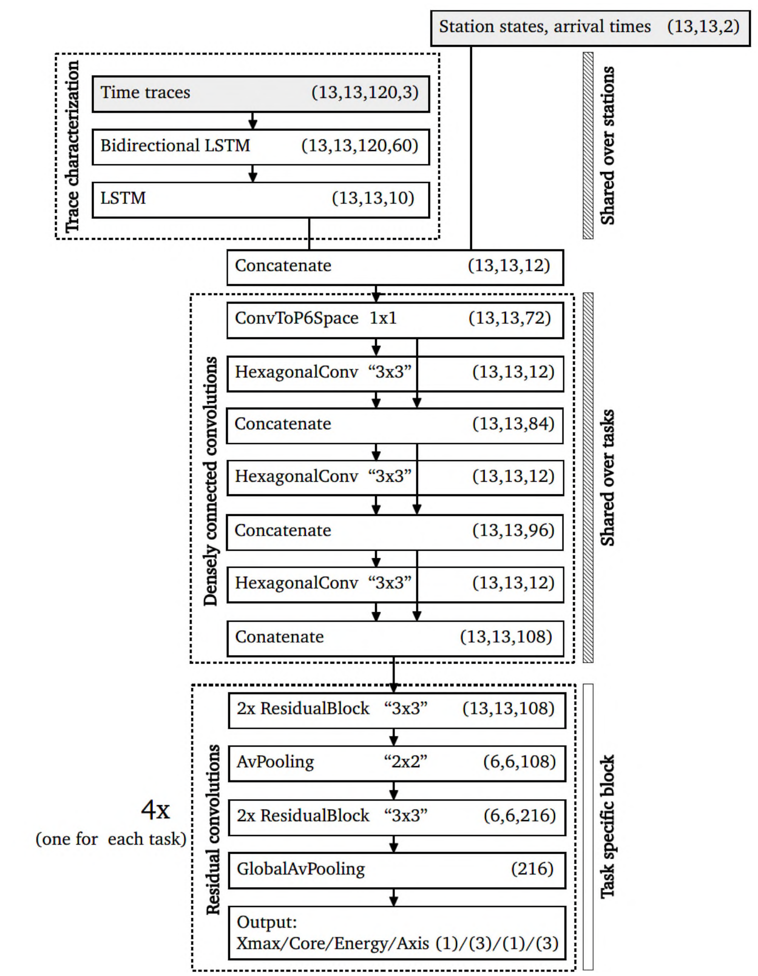
turn off detector



saturate signal trace



Network training with simulated data including defects



→ *Improved generalization capability of trained neural network*

Causality: analysis of network predictions

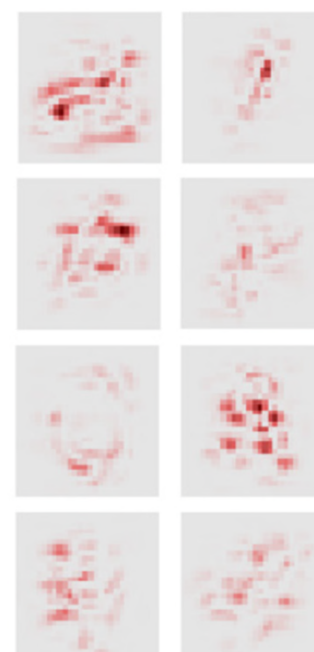
Measure impact: input x_i → overall prediction

G. Montavon, W. Samek, K.-R. Müller, Digital Signal Processing 73 (2018) 1

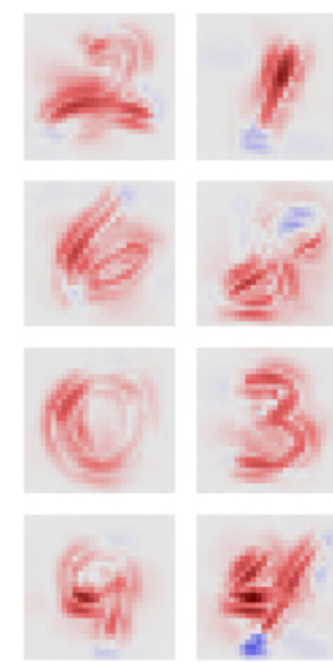
Original

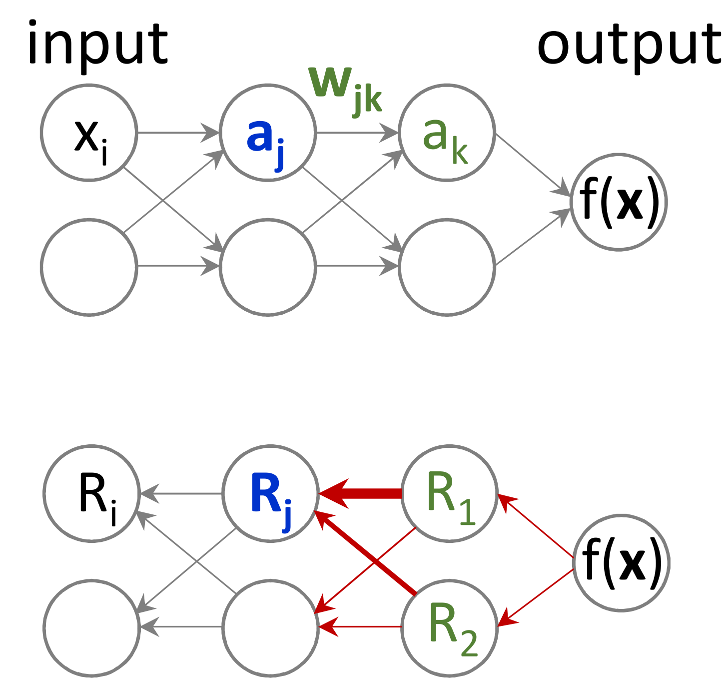
Sensitivity by partial derivative

$$\left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}_o}$$



LRP=Layer-wise Relevance Propagation
separate sums: in favor / against prediction





$$a_k = \sigma\left(\sum_j a_j w_{jk} + b_k\right)$$

$$R_j = \sum_k \left(\alpha \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} - \beta \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-}\right) R_k$$

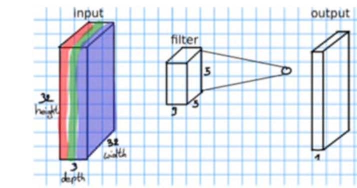
Messages Deep Learning

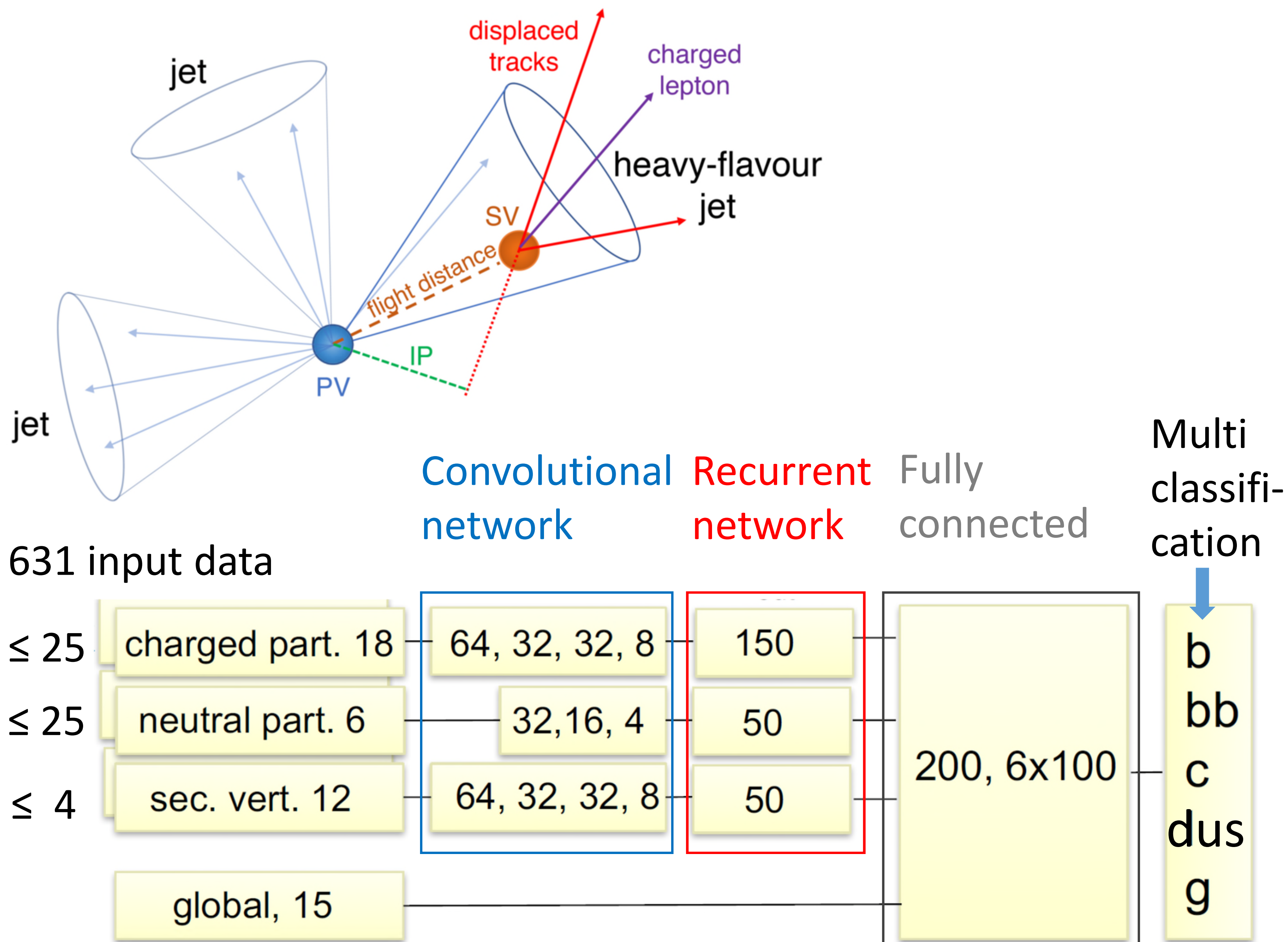
- **Physicists** exploit more information from data: choose network according to data symmetries
- Standard network like fitting functions, except network is **ultra-flexible physics model**
- Advanced concepts assign **functional targets**, training data **autonomously** optimize network
- Tools exist for including experiment‘s **operation reality** and obtaining **insight** into networks

We ought to prepare for fundamental change to include machines in our daily work

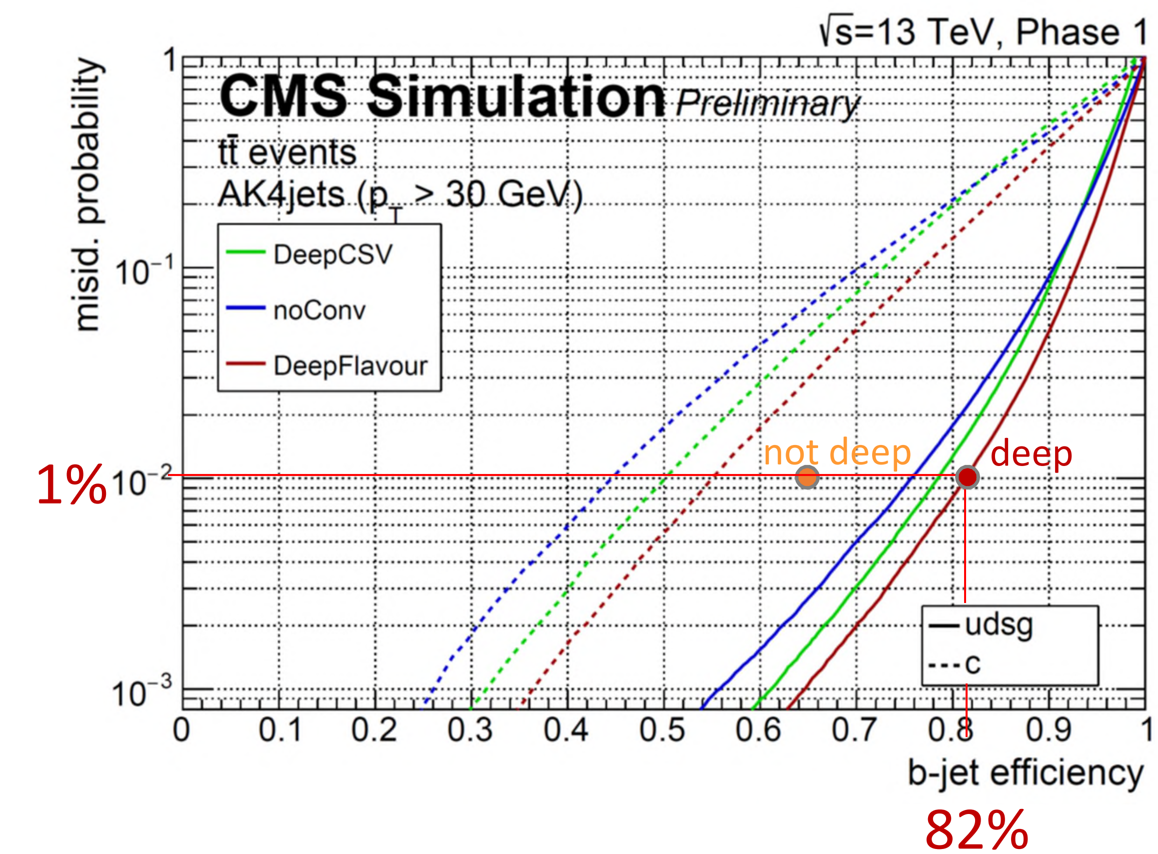
backup

CMS jet flavor tagging





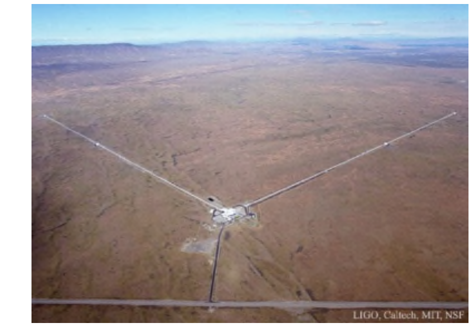
The up to 25 tracks have 18 properties px,py,pz,eta,quality,... the 1-dim Convolution adds up all properties of a track. In the RNN the summed properties of the tracks are merged.



Brilliant improvement with high impact on data analyses

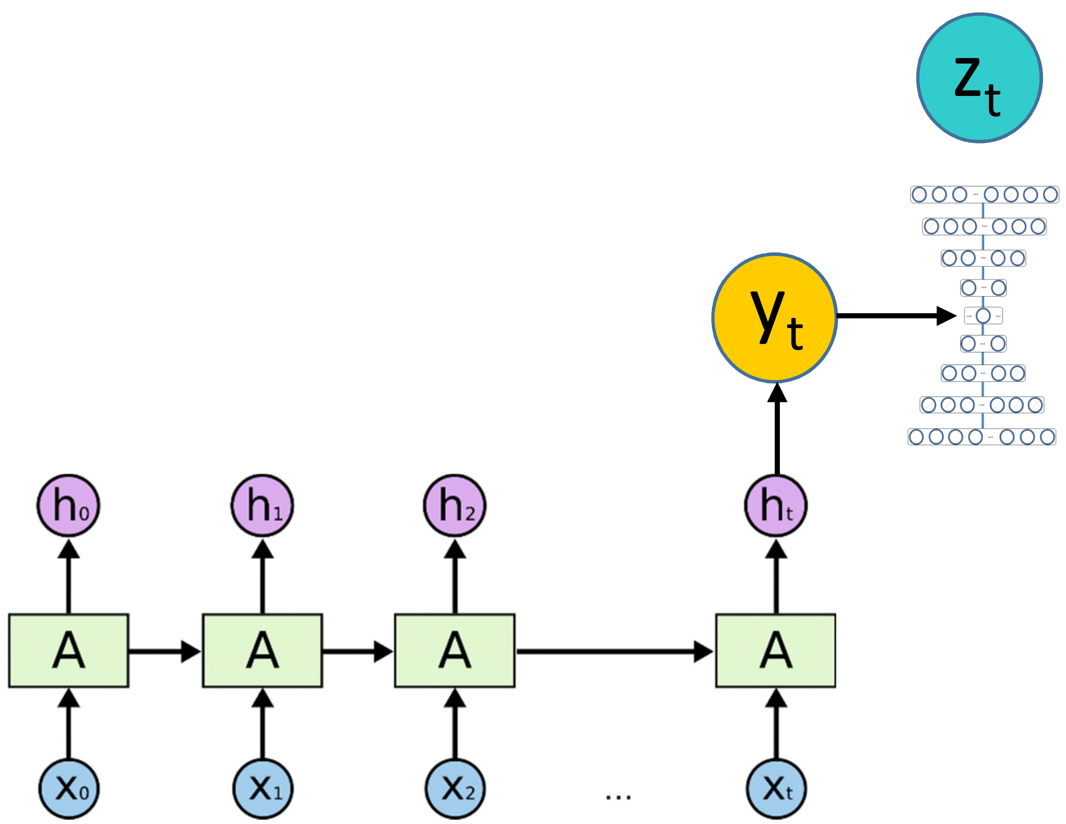
2 Higgs → 4 bottom quarks: $\varepsilon^{4}_{bottom}: 0.18 \rightarrow 0.45$

Denoising Gravitational Waves with Recurrent Denoising Autoencoder



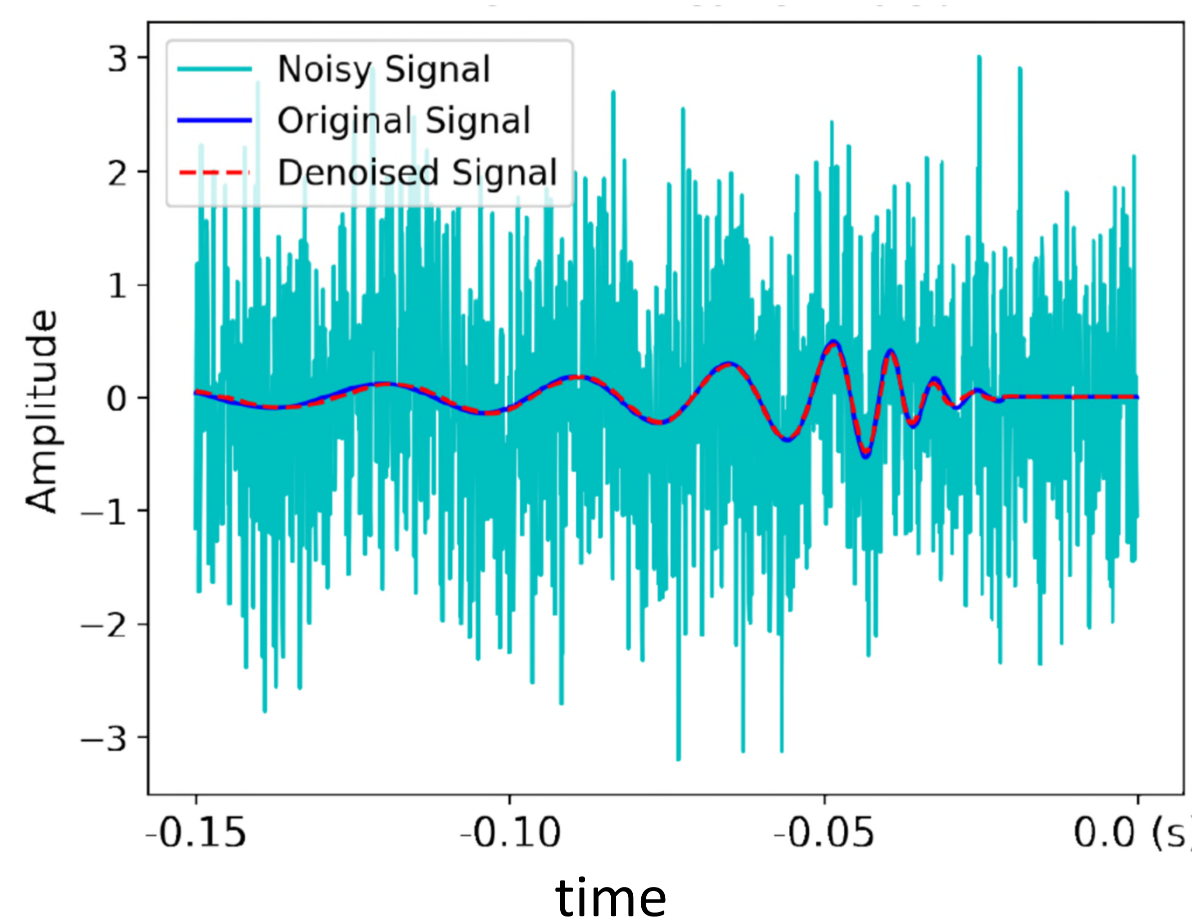
Recurrent Network

Autoencoder Network



many-to-one network

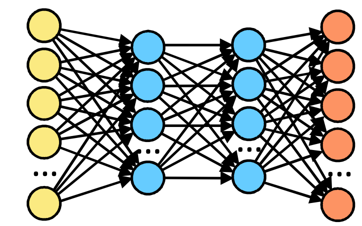
Gravitational wave (simulated)



binary black hole merger

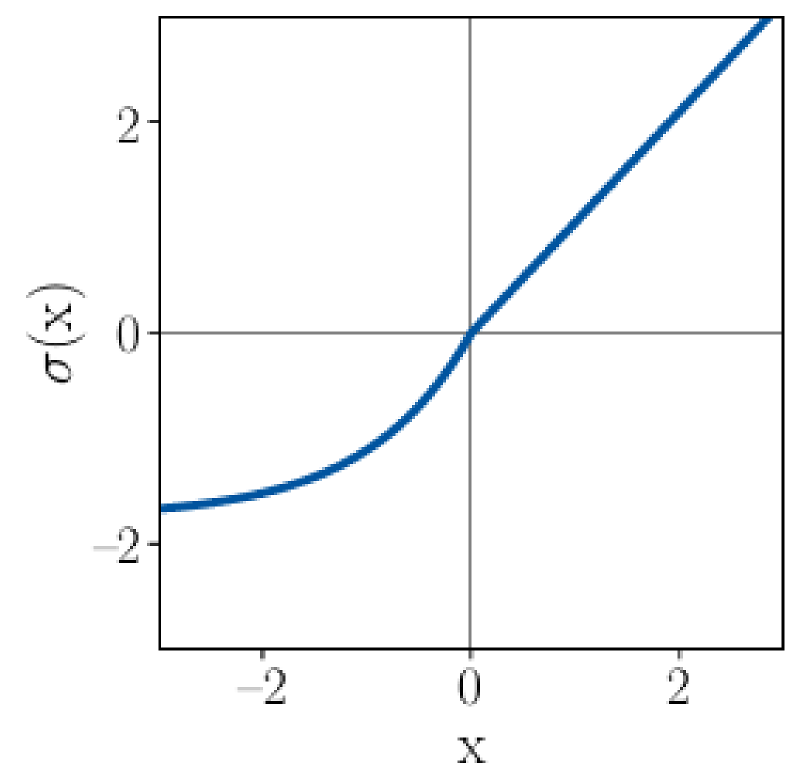
Excellent recovery of original signal

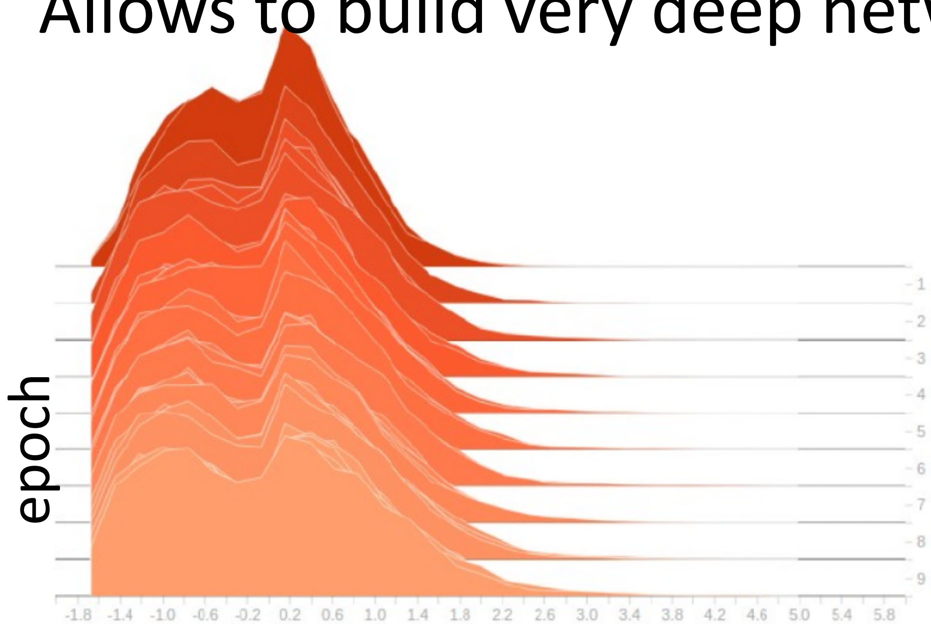
Self Normalizing Networks



$$\mathrm{selu}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leqslant 0 \end{cases}$$

- Batch normalization adds perturbations for training fully connected networks
- Use activation function ***selu*** which ensures standard normalized output: $\mu = 0$, $\sigma = 1$
- Initialization:
  Gauß with $\mu = 0$, $\sigma = 1/n$ with n=nodes in lower layer
- Alpha-dropout
  (insert specified value instead of turning node off)
- Stabilizes the training
- Allows to build very deep networks!





result activation function

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_i$$

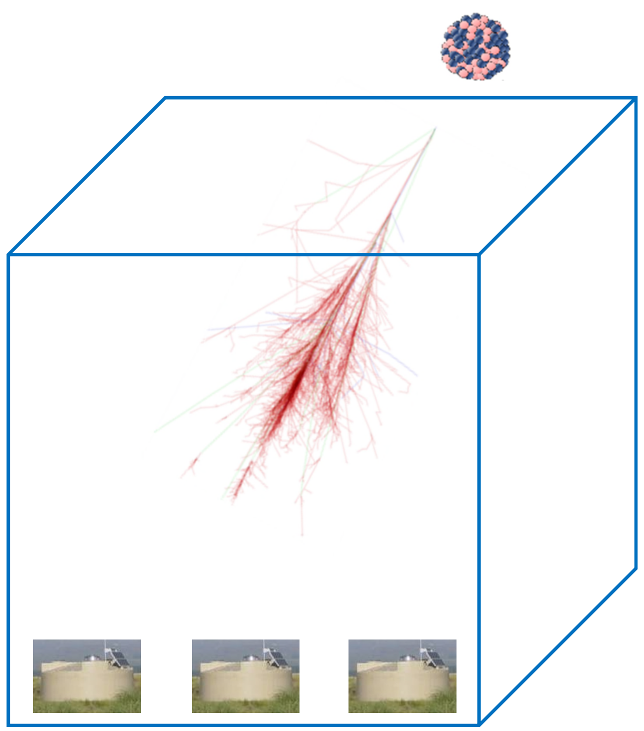
$$\nu_j = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_j)^2$$

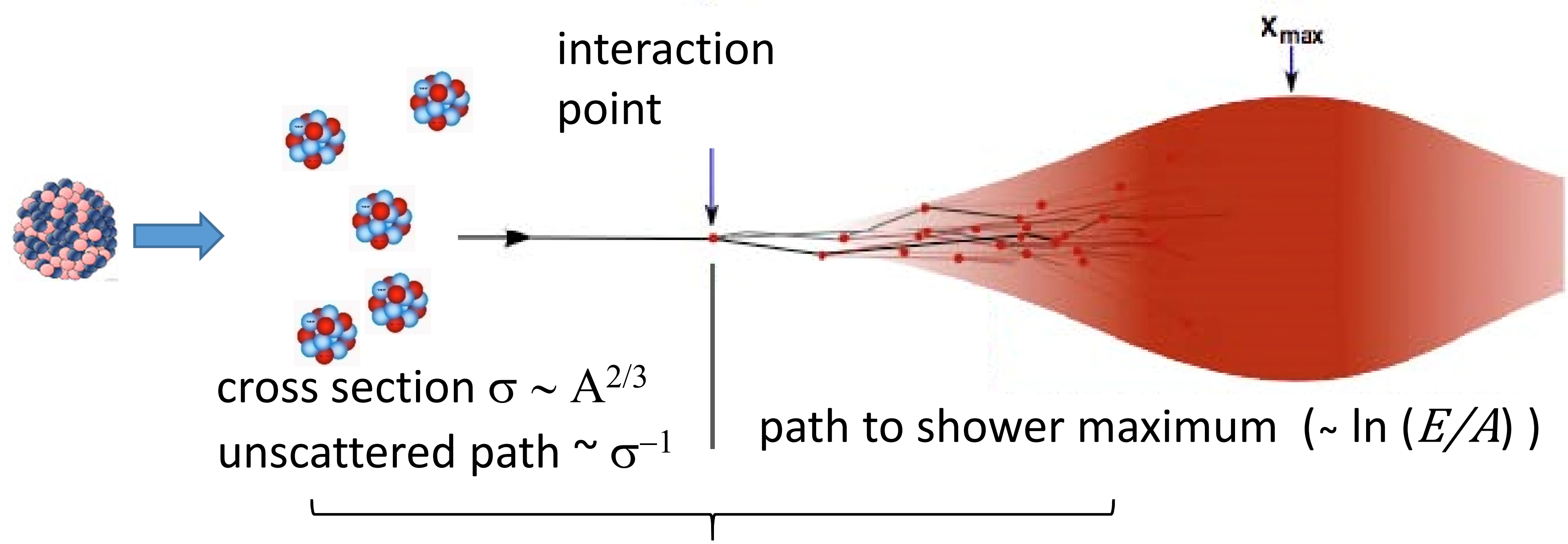
$$\begin{pmatrix} \mu_{j+1} \\ \nu_{j+1} \end{pmatrix} = K \begin{pmatrix} \mu_j \\ \nu_j \end{pmatrix}$$

Requirements:
negative & positive values to control the mean
Slope < 1 for damping the variance
Slope > 1 to rise the variance

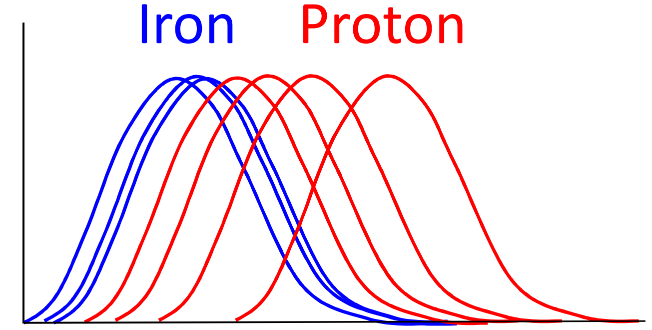
Cosmic ray: shower maximum

$$x_{max} = X_0 \cdot \left(\ln \frac{E/A}{E_c} - 0.5 \right)$$





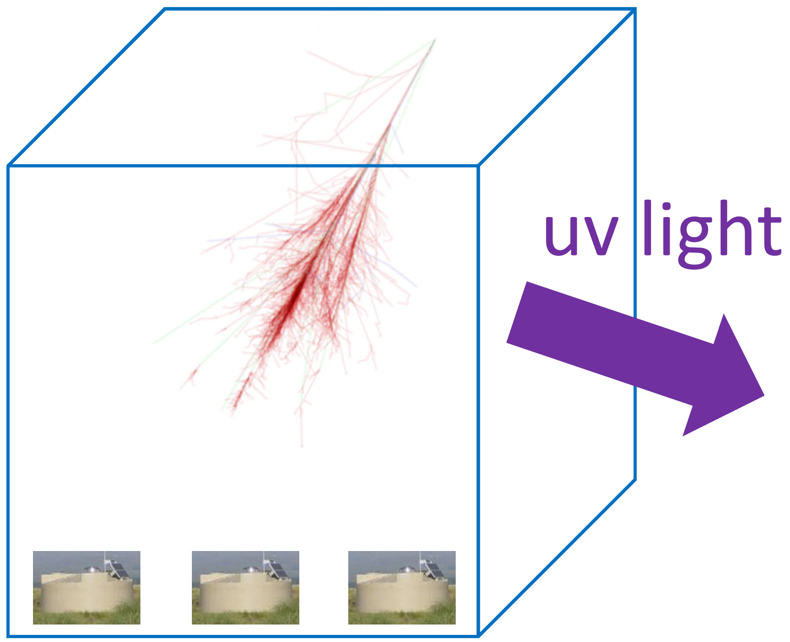
cosmic ray energy & mass

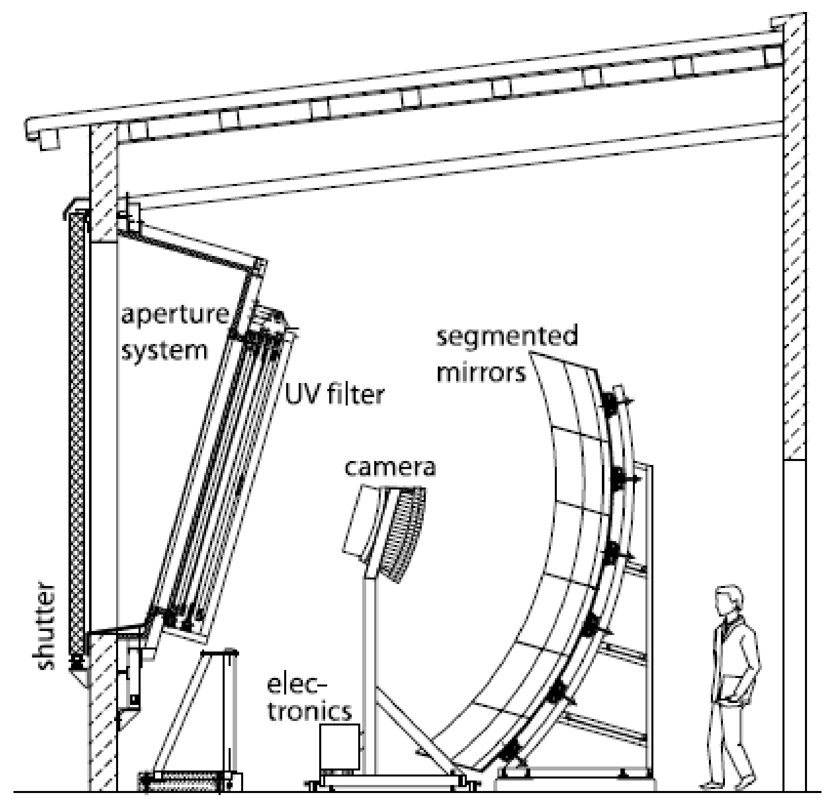


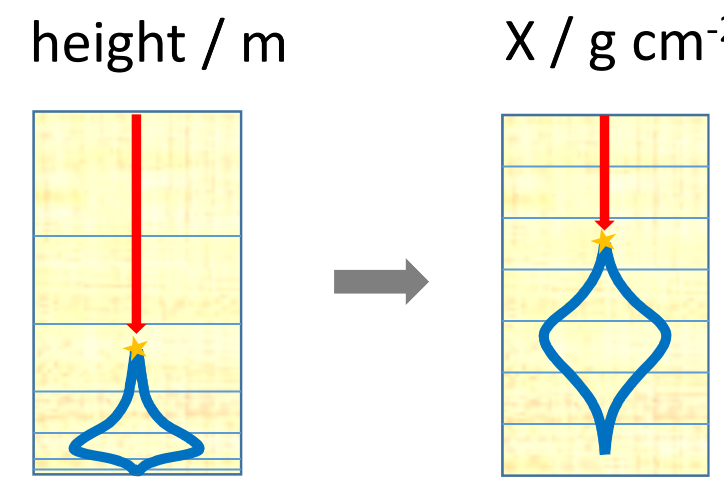
Fluorescence Light Detection

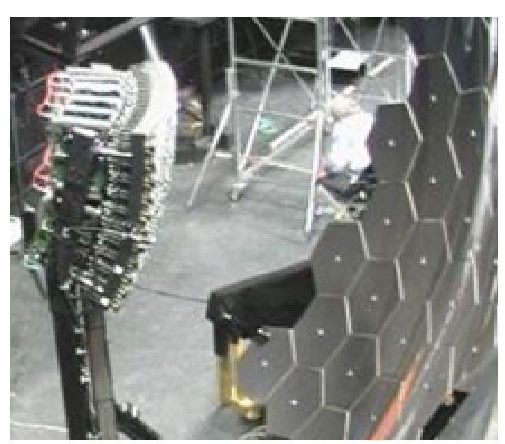
correct for atmosphere

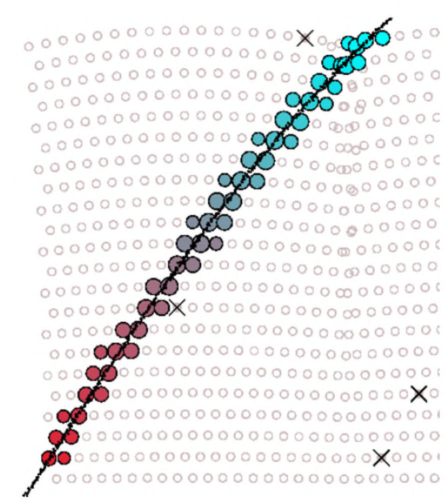
$$\rho\ (h) = \rho_o\ e^{-\frac{\rho_o}{p_o} g h}$$

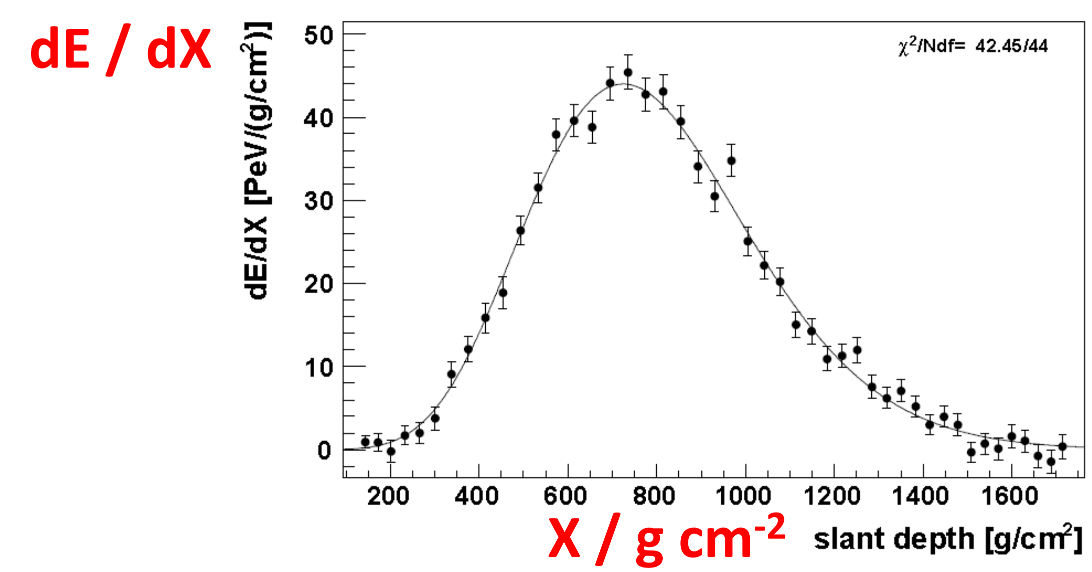






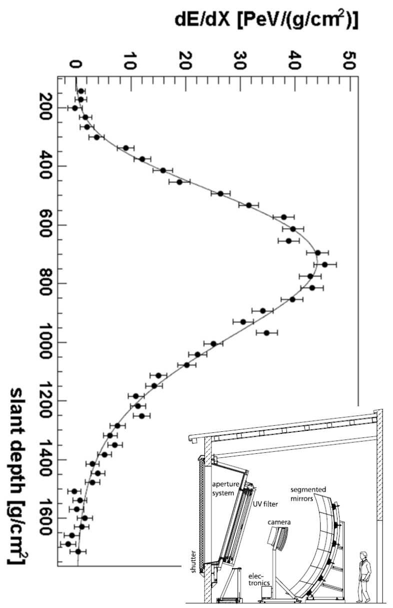




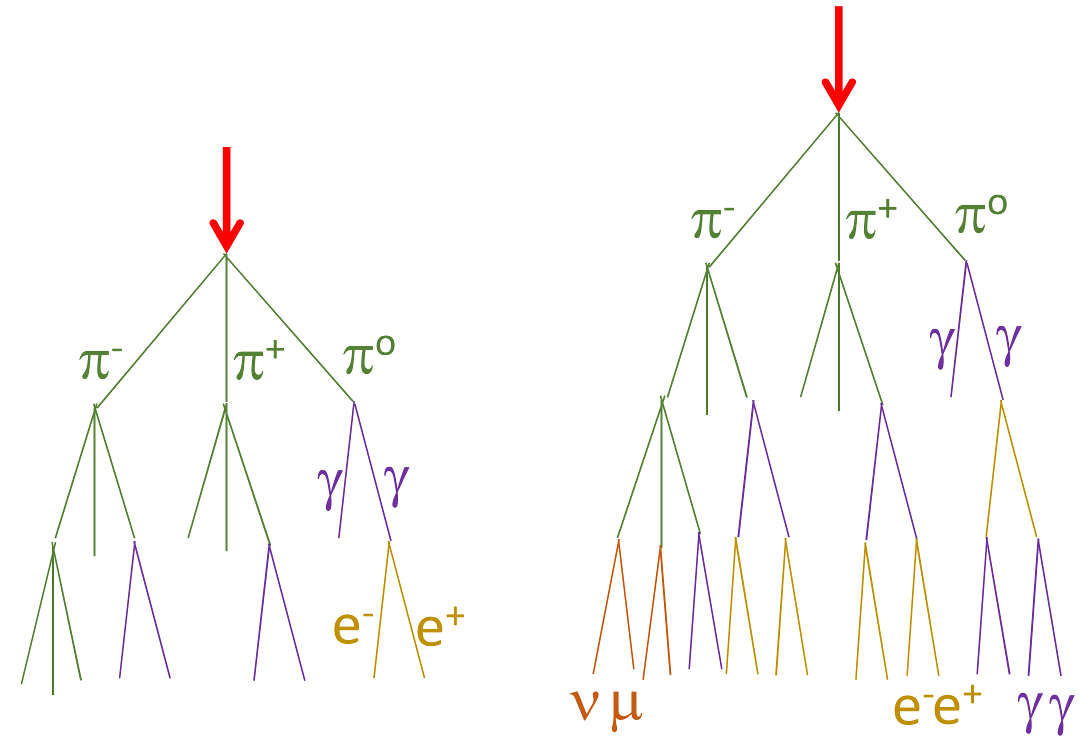


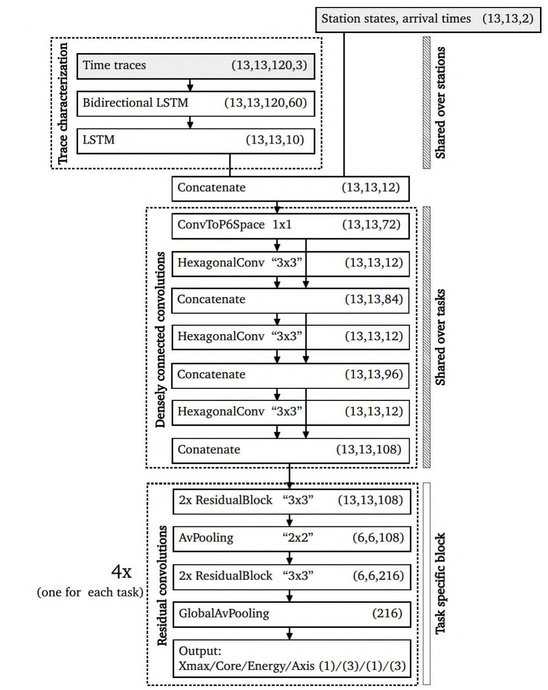
Maximum of shower development directly visible in the camera during *moonless nights*

Air shower maximum with particle detectors on Earth

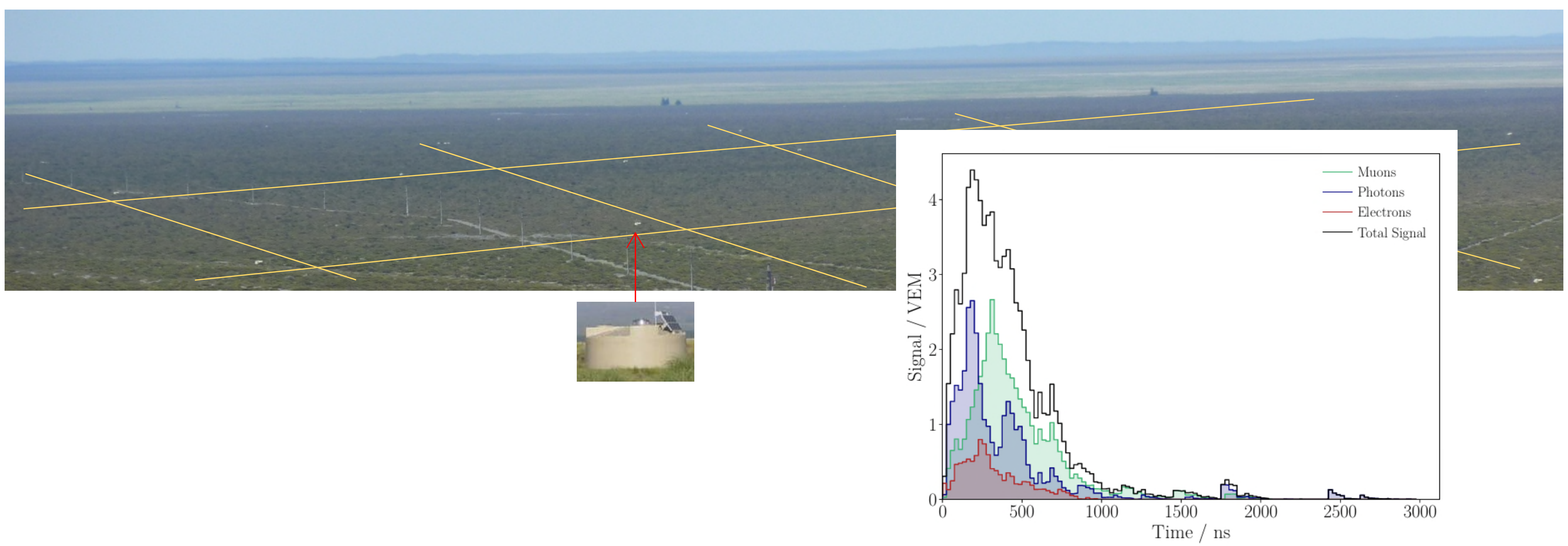


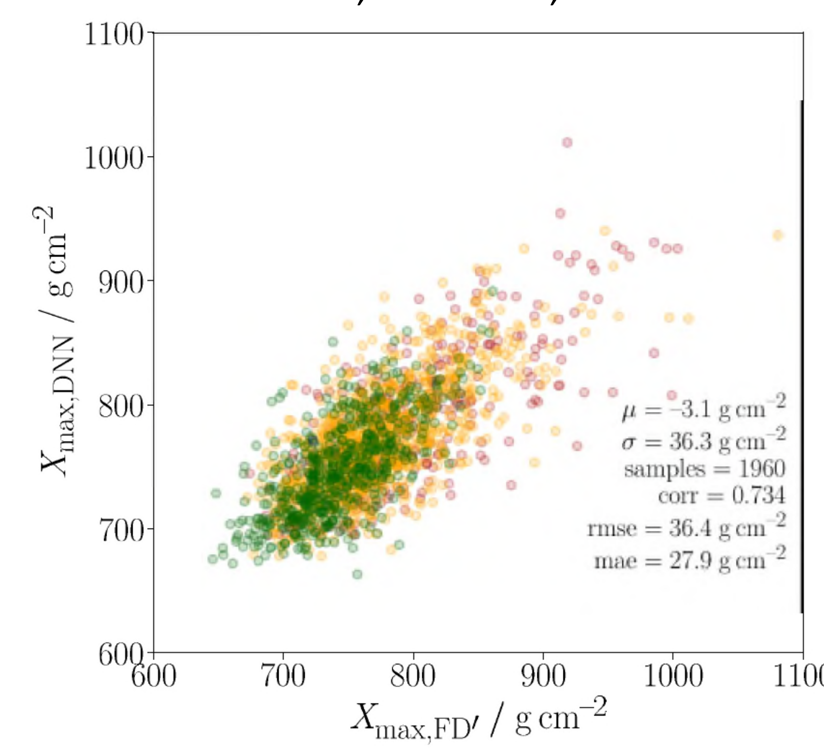
electromagnetic component + muons: depending on height of 1st interaction shower particles differ when reaching earth





Jonas Glombitza, PhD thesis, RWTH Aachen

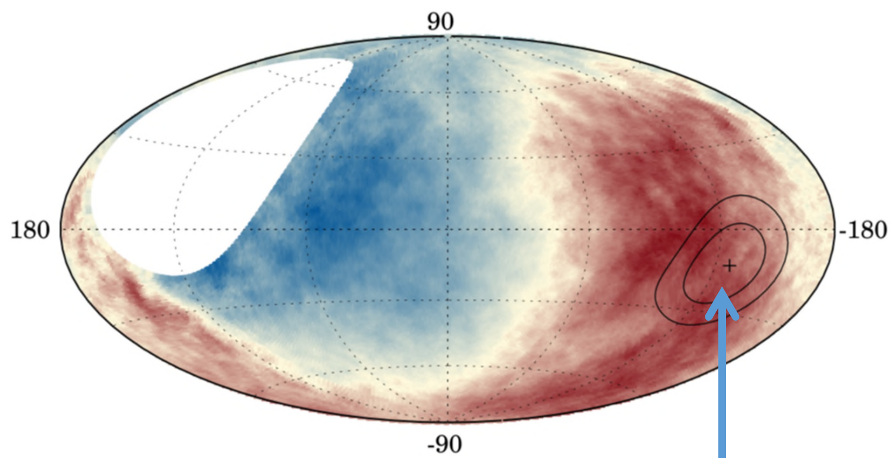




Cosmic Rays: Energies & Nuclei

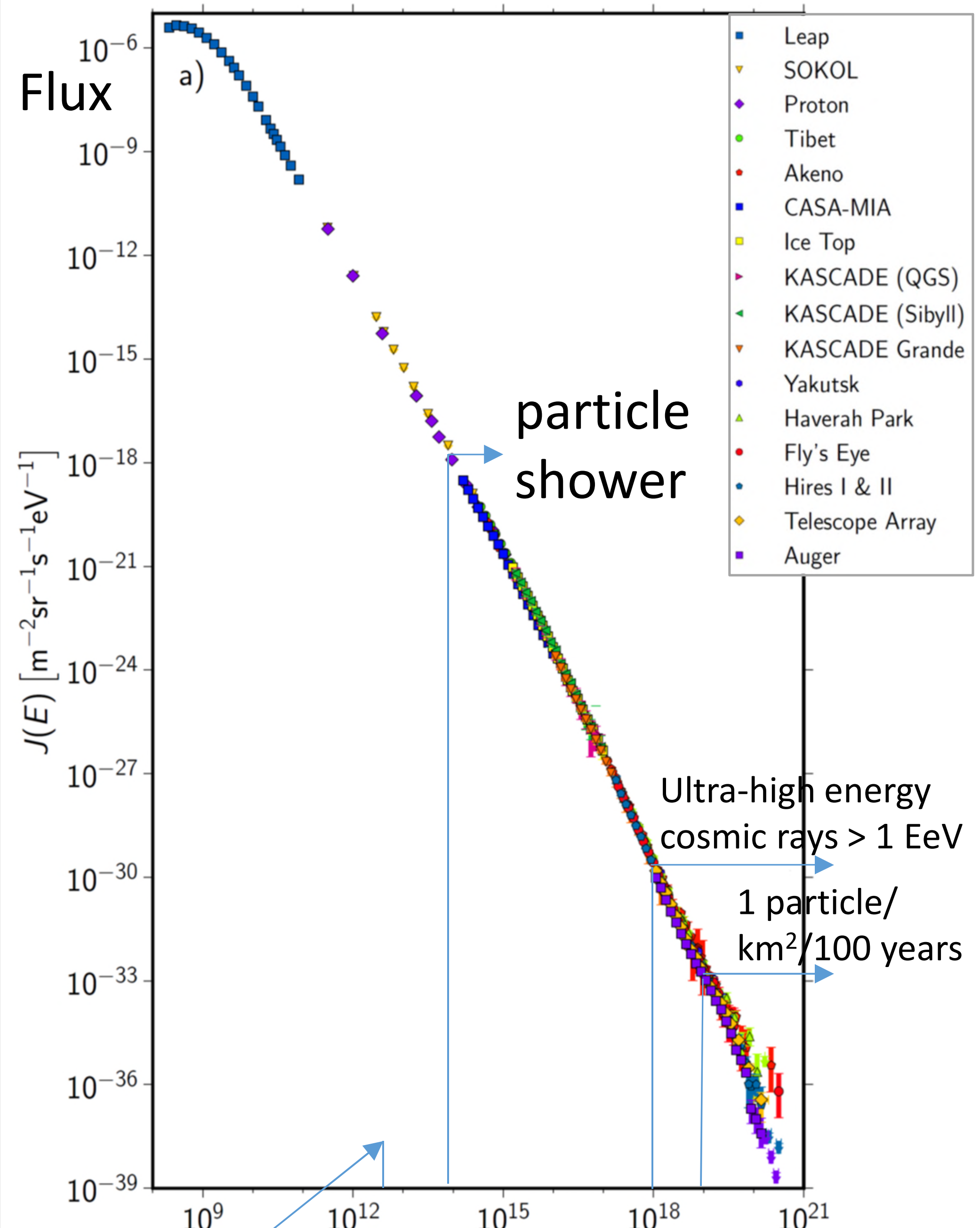
Pierre Auger Observatory

Arrival directions

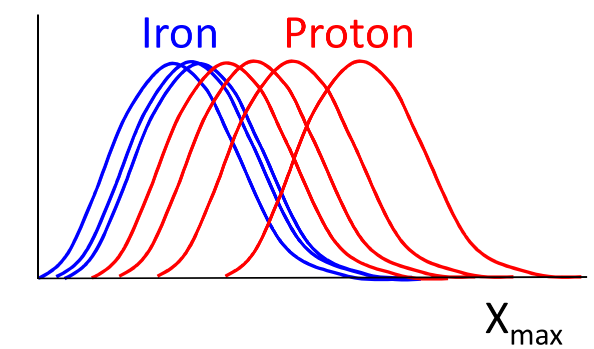


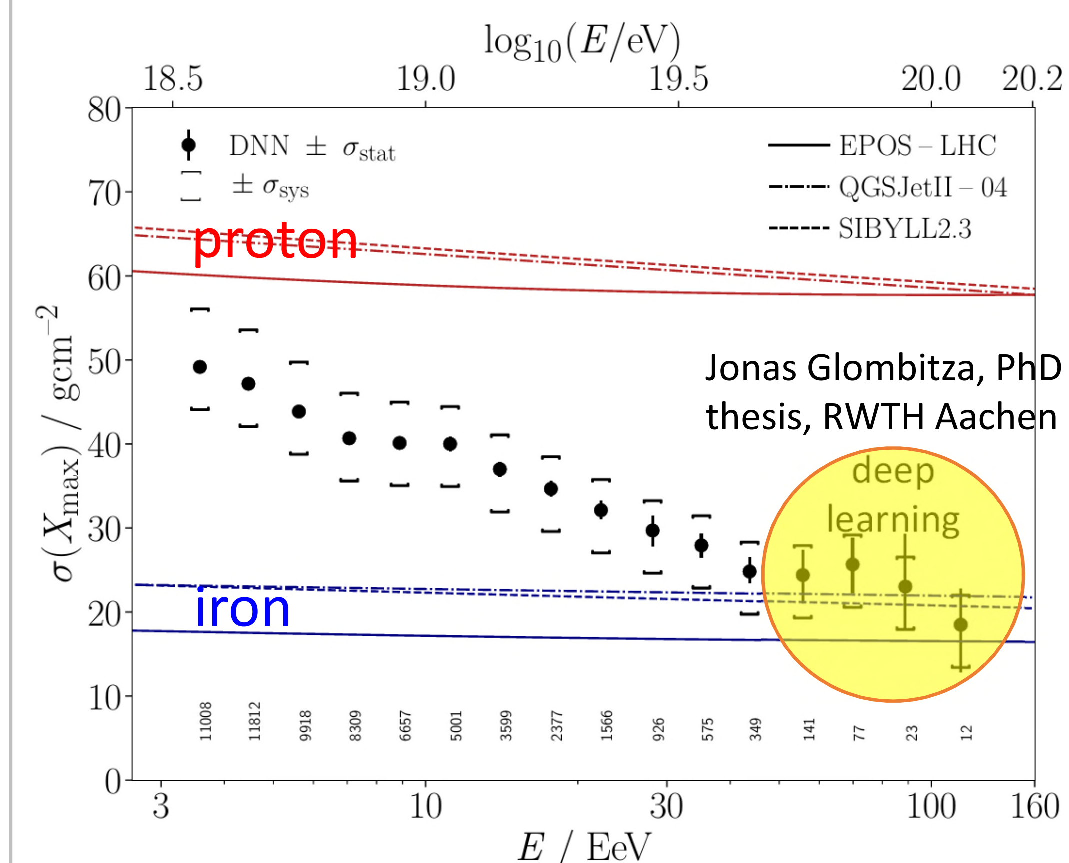
In galactic coordinates: from this direction more cosmic rays than from the other side

Energy



Shower maximum

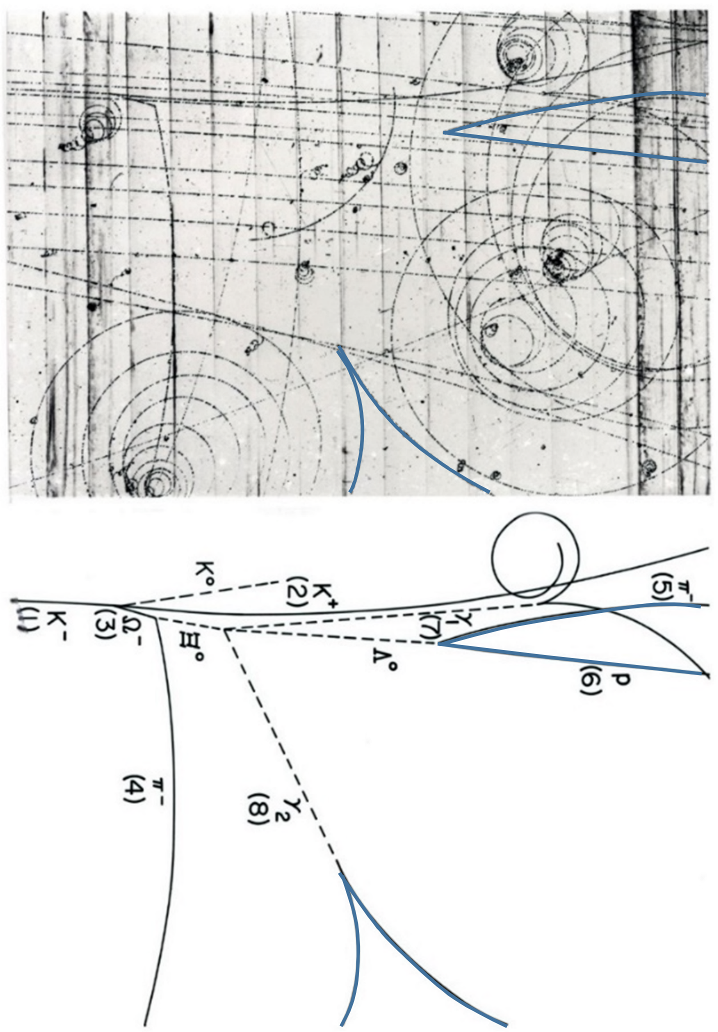




M. Erdmann, E. Geiser, Y. Rath, M. Rieger, JINST 14 (2019) P06006

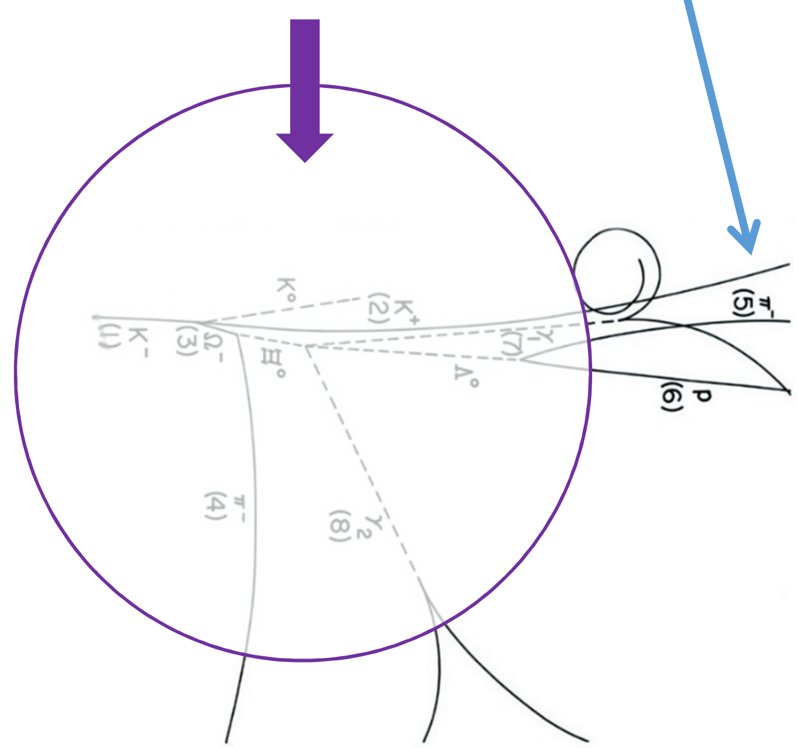
Lorentz Boost Network to recover interaction

In classic bubble chamber experiments most of the interaction was visible



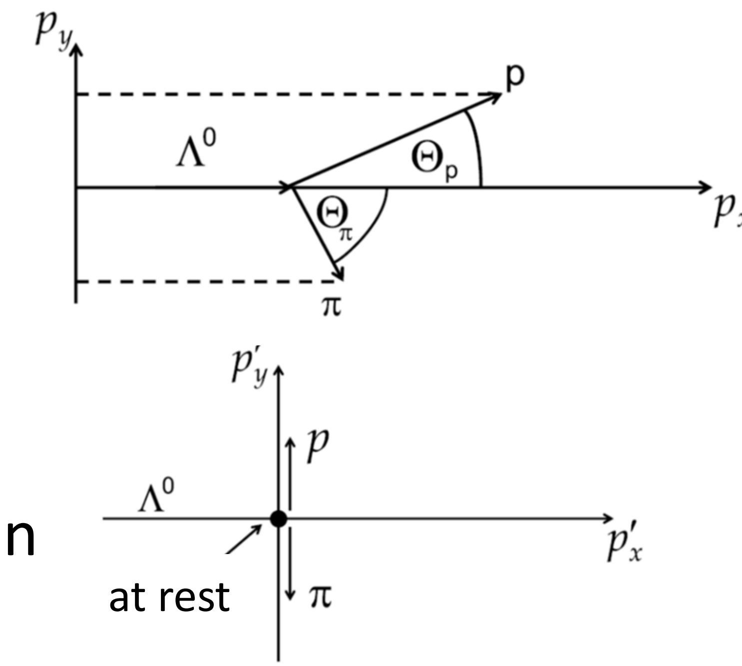
Todays experiments: reconstruct interaction from visible particles

interaction at very small scales < 10^{-15} m remain invisible

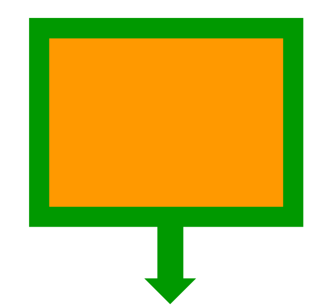


boosted particles

undo boost by Lorentz Transformation



Particle physics deep network architecture

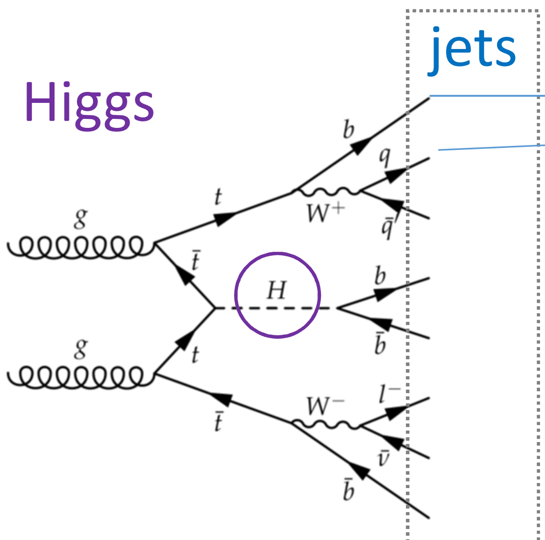


output: `high-level data' more smart data

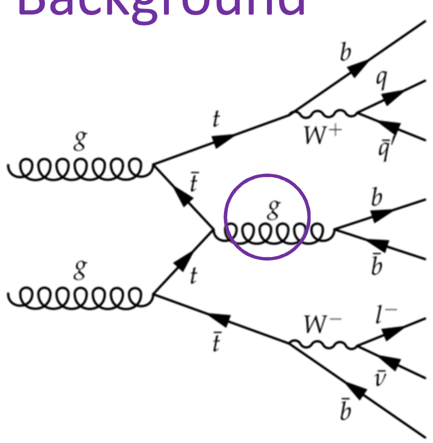
M. Erdmann, E. Geiser, Y. Rath, M. Rieger, JINST 14 (2019) P06006

Autonomous engineering of discriminating variables

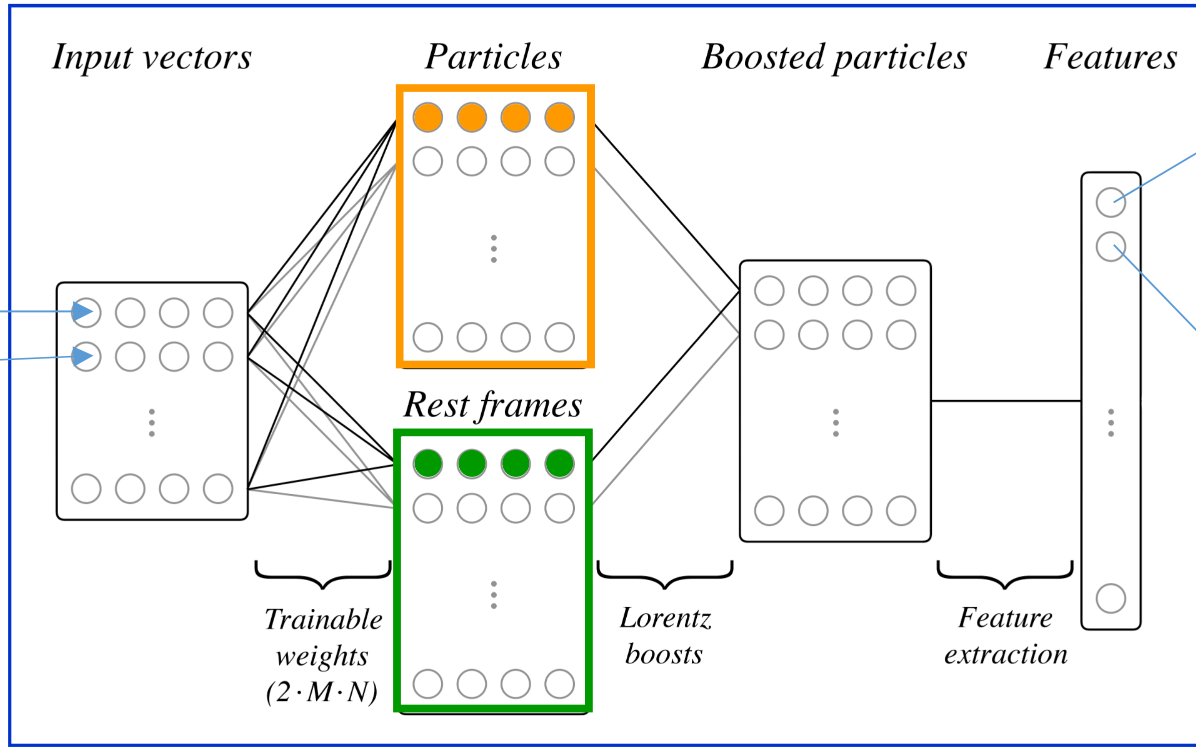
Higgs

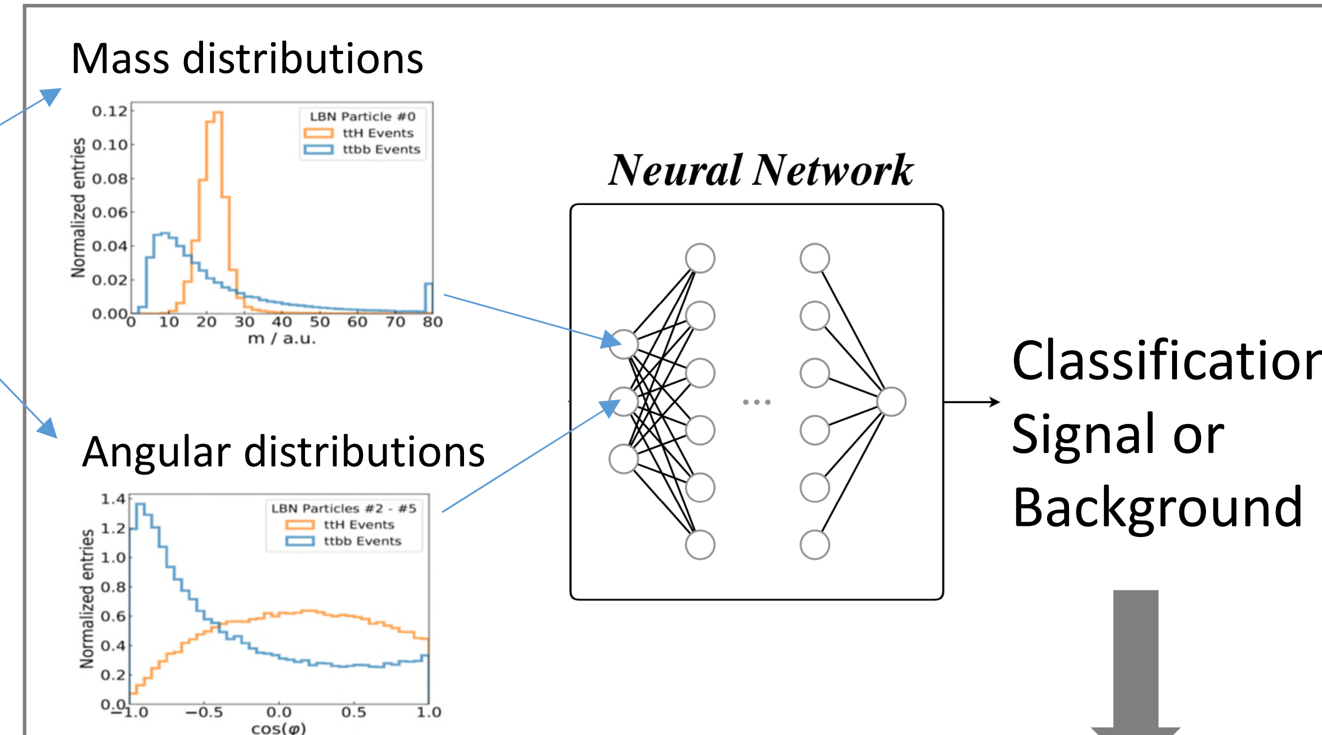


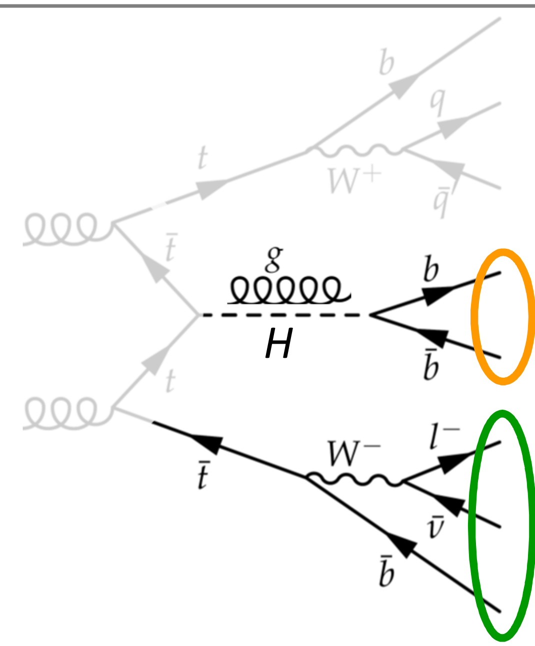
Background



Lorentz Boost Network ('unsupervised')







after training on simulated data:

network builds Higgs/gluon and

boosts it into rest frame of top quark

For data analysis: ***autonomous variables surpass human-made variables***

Classification efficiency & purity