Applying machine learning to CALICE data

Balázs Kégl, Franck Dubard, Roman Poeschl, Naomi Van Der Kolk

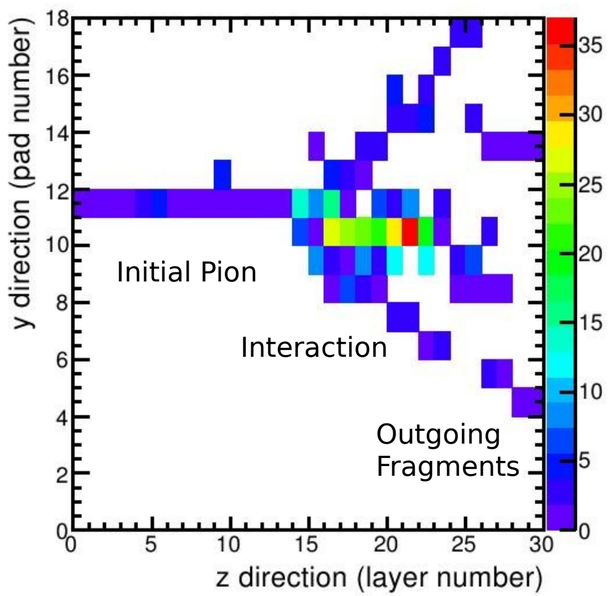
Laboratoire de l'Acélérateur Linéaire,
Université Paris Sud, CNRS/IN2P3, Orsay, France

Objectives

- Improve jet reconstruction in the eCal (hCal)
  - energy resolution
  - assigning each pixel (energy deposit) to an incoming particle: jet separation

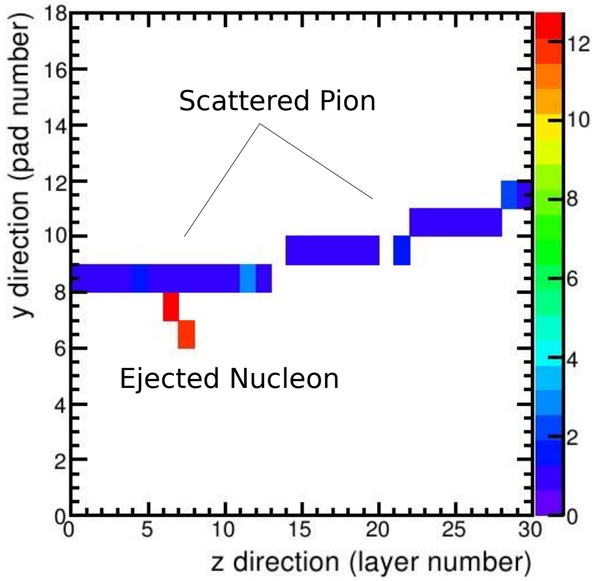
Granularity and hadronic cascades
(Start of) Hadronic showers in the SiW Ecal

Complex and impressive



Inelastic reaction in SiW Ecal

Simple but nice



Short truncated showers

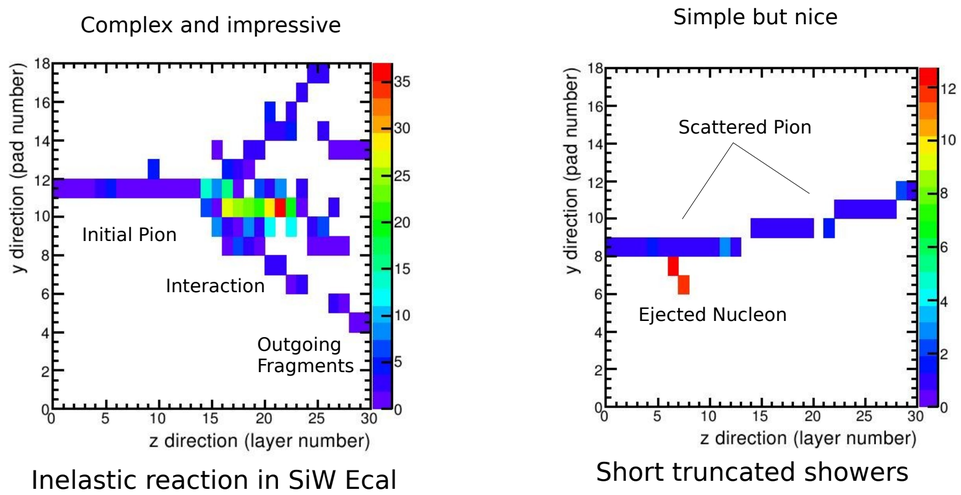
Outline

- Mainly a plan at this point
- The generative approach
  - Parametrizing GEANT4 (model reduction)
  - Fitting the models on calorimeter data
  - EM or MCMC to handle mixtures
- The direct approach
  - Train neural nets/BDTs for the inverse function
  - Engineer features or use deep learning
  - How to handle overlapping jets?

The generative approach

- x: observables: pixels
- y: parameters to infer:
  - e.g: incoming geometry, energy, type
- but also nuisance parameters that can be "read out" from GEANT4

Granularity and hadronic cascades
(Start of) Hadronic showers in the SiW Ecal



$$p(y \mid x) \sim p(x \mid y) \times p(y)$$

posterior likelihood prior

The generative approach

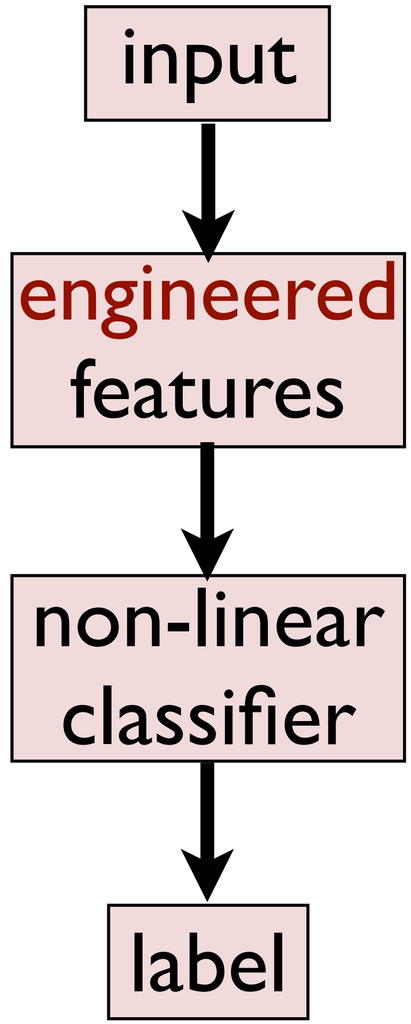
- $p(x \mid y)$ (likelihood)
  - parametrize GEANT4
  - at what granularity? computational and statistical issues
  - existing solutions for mixtures: $p(x \mid y) = \Sigma_i \alpha_i\, p(x \mid y_i)$: EM, MCMC
- natural handling of model dependence, controlled use of simulators
- conceptually neat but computationally heavy: zillions of calls to the likelihood

The direct approach

- model directly the posterior p(y | x)
  - x: pixels, y: elastic/non-elastic label
  - no neat (physical) parametrization: nonparametric approach
  - black box (neural net or boosted decision tree), trained on a sample generated by GEANT4, only end-to-end setup
  - not completely clear how to handle mixtures: similar to image segmentation
  - standard technology
  - computationally fast inference
  - recent advances in deep learning (training multi-layer neural nets) to be harvested

The direct approach

- The classical setup
  - image, speech, text, etc.
  - domain knowledge goes into features
  - not fully automatic but flexible

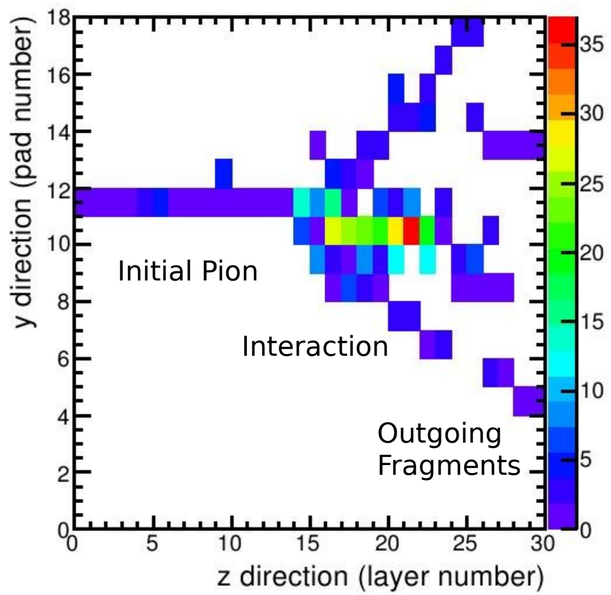


The direct approach

- The features
  - the front image
  - the lateral vector
  - the lateral discrete derivative
  - covariance and correlation structure of the front image
- The classifier
  - AdaBoost + decision trees

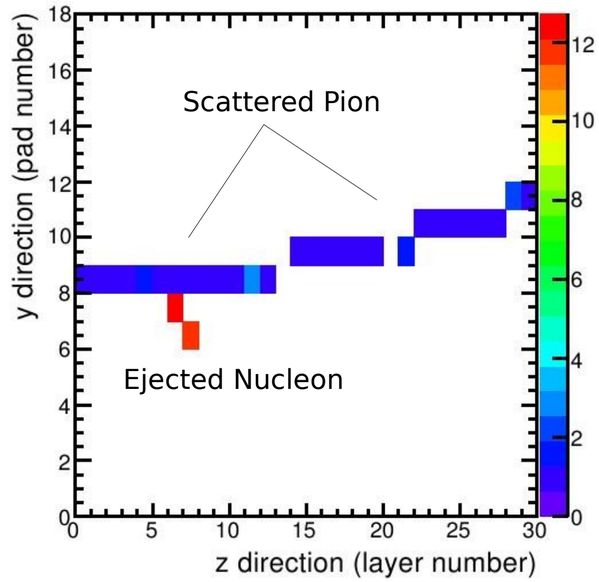
Granularity and hadronic cascades
(Start of) Hadronic showers in the SiW Ecal

Complex and impressive



Inelastic reaction in SiW Ecal

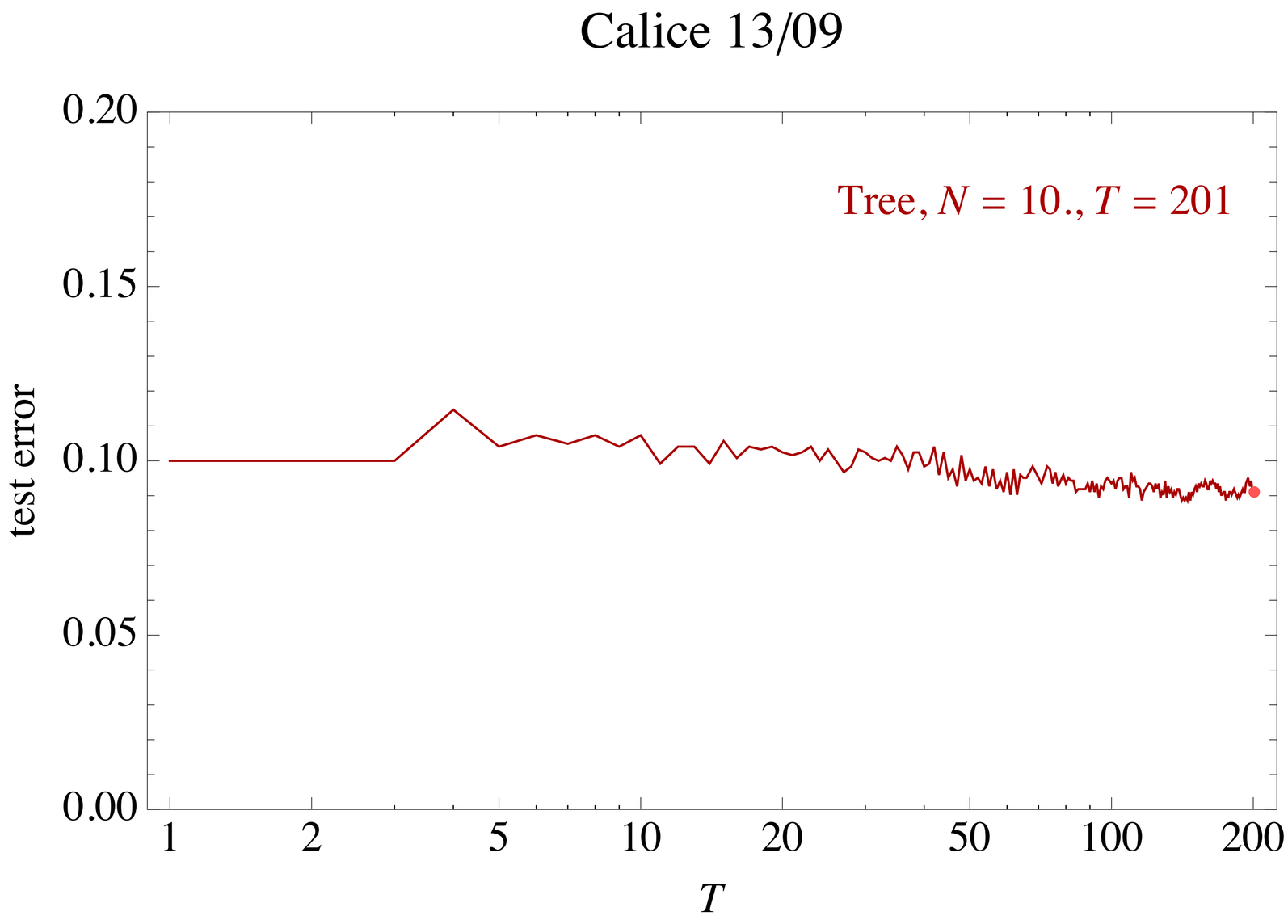
Simple but nice



Short truncated showers

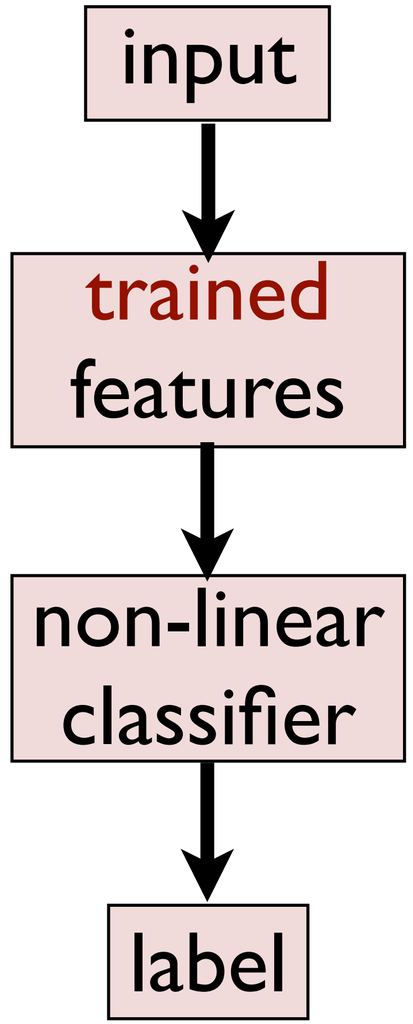
The direct approach

- The learning curve with engineered features



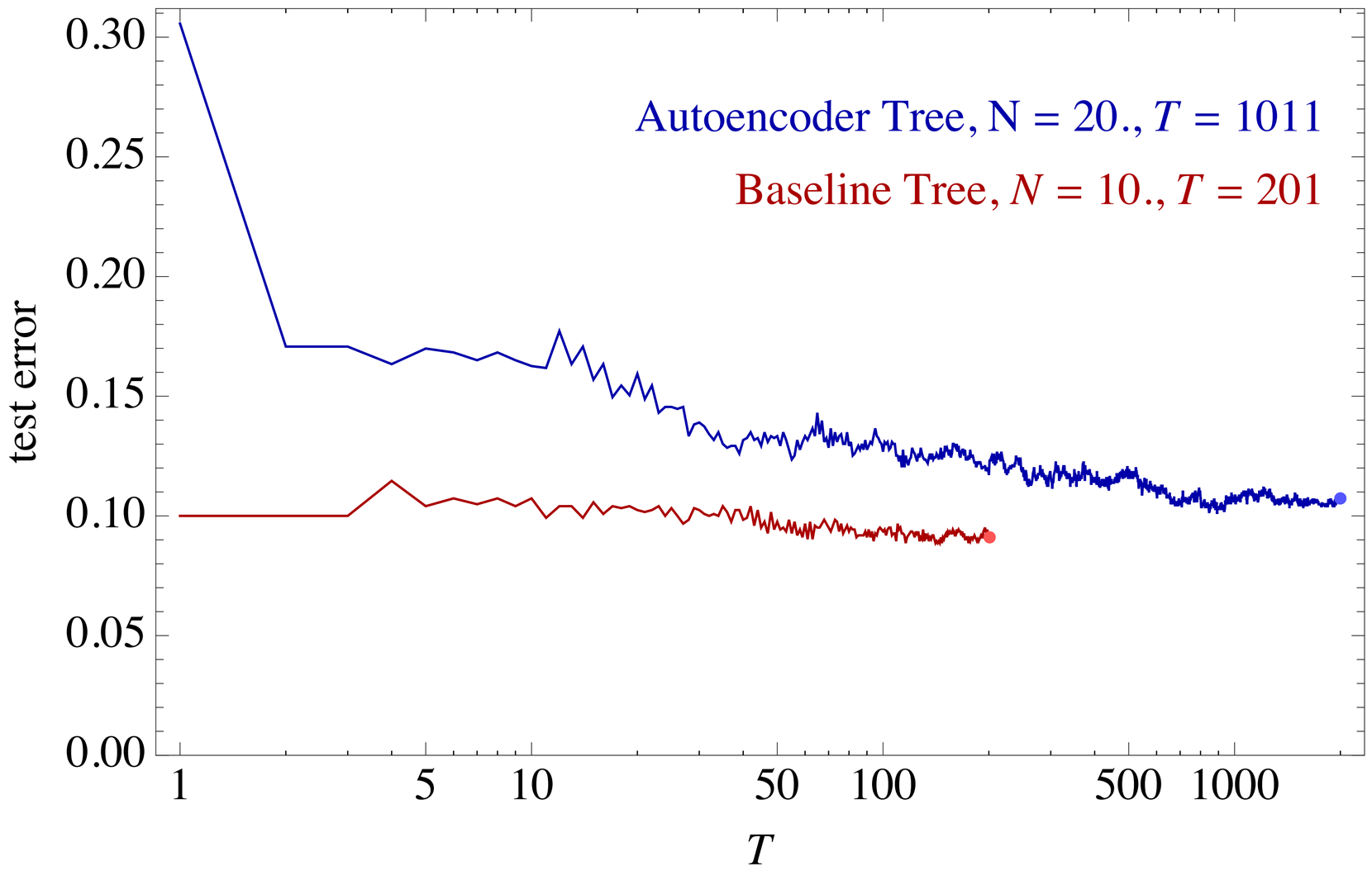
The direct approach

- The deep learning setup
  - unsupervised pre-training (stacked autoencoder)
  - supervised training of a giant network (convolutional net)
- revolutionized speech and image recognition in the last five years



The direct approach

- The learning curve with trained features (autoencoder)



Conclusions

- Feasibility testing of different inference and learning approaches
  - generative or direct?
  - engineered or trained features?
  - how to handle mixtures and pixel assignment?
  - long road ahead