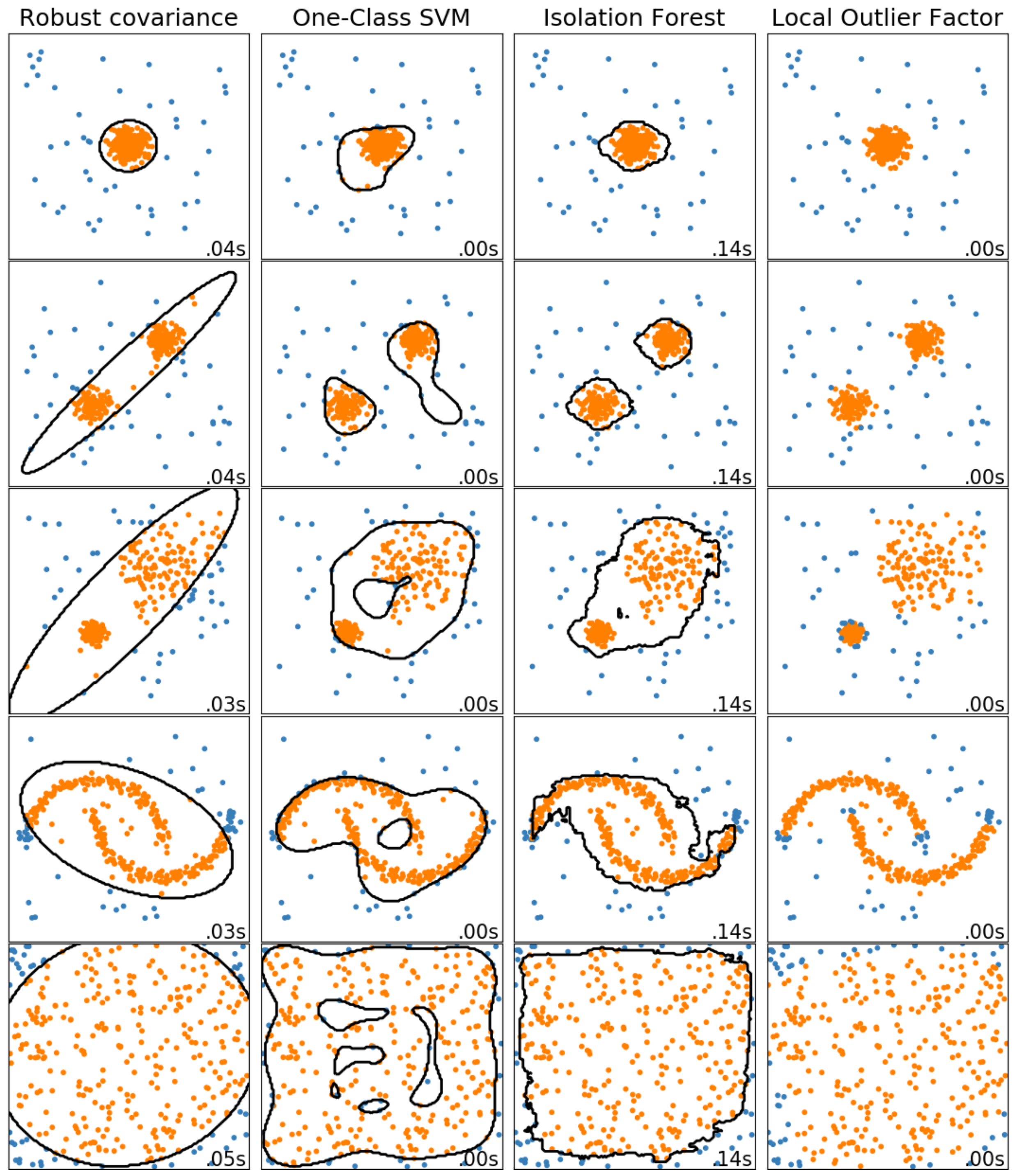
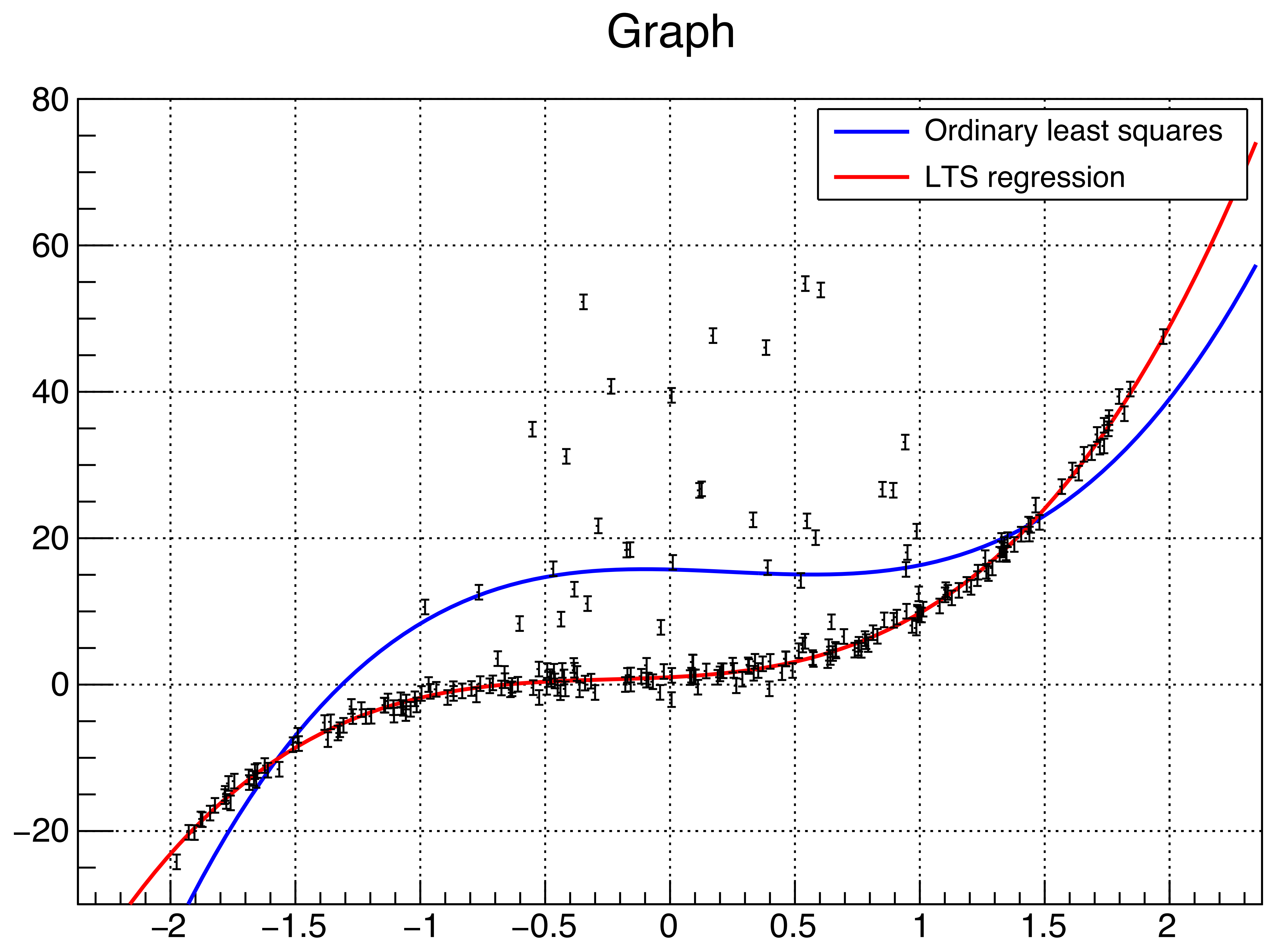
Outlier Detection

- Supervised algorithms
  - train on data with and without anomalies
  - look at difference of data sets
  - Example: 2 sample GoF test
    - compare two histograms (chi-square , KS test)
    - (all available in ROOT)
  - Train a classifier to distinguish the two cases
- Semi-supervised algorithms
  - train only on data without anomaly
  - Example: estimate density and look how anomaly differ from model
    - e.g. 1-sample GoF test (available in ROOT)
- Unsupervised algorithms
  - algorithm identifies anomalies looking at all data
  - can detect unseen anomalies
  - Example: cluster analysis or ML tools (autoencoders)



Robust Fitting

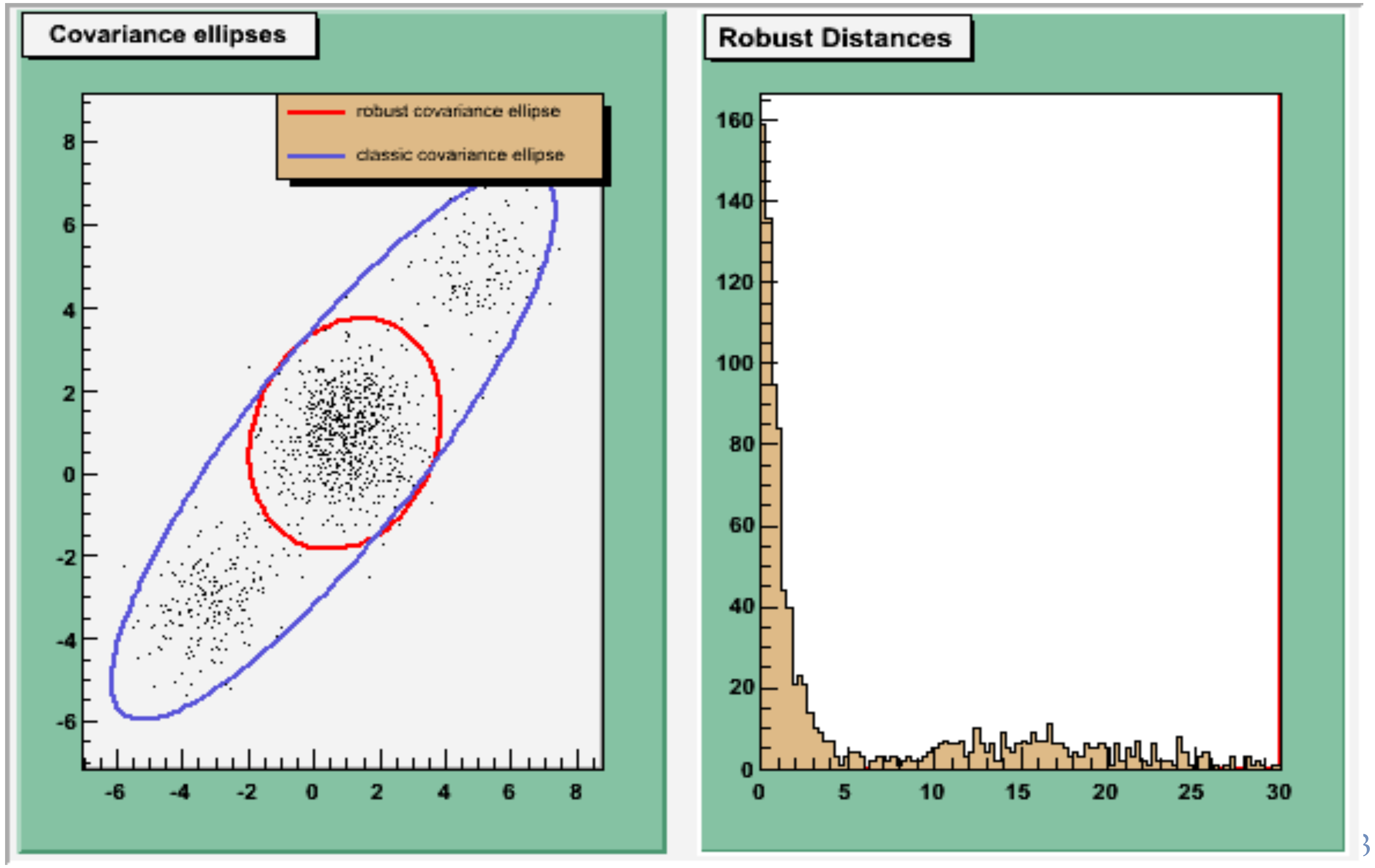
- Robust linear fitting available with Least Trimmed Square (LTS) regression
  - compute chi-square for a fraction h of points (~ 70%)
  - choise points that give the best chi-square



- method in ROOT for performing linear fitting
- no direct method to retrieve outlier points
- work well for not too large limited data points size

Robust Covariance Matrix

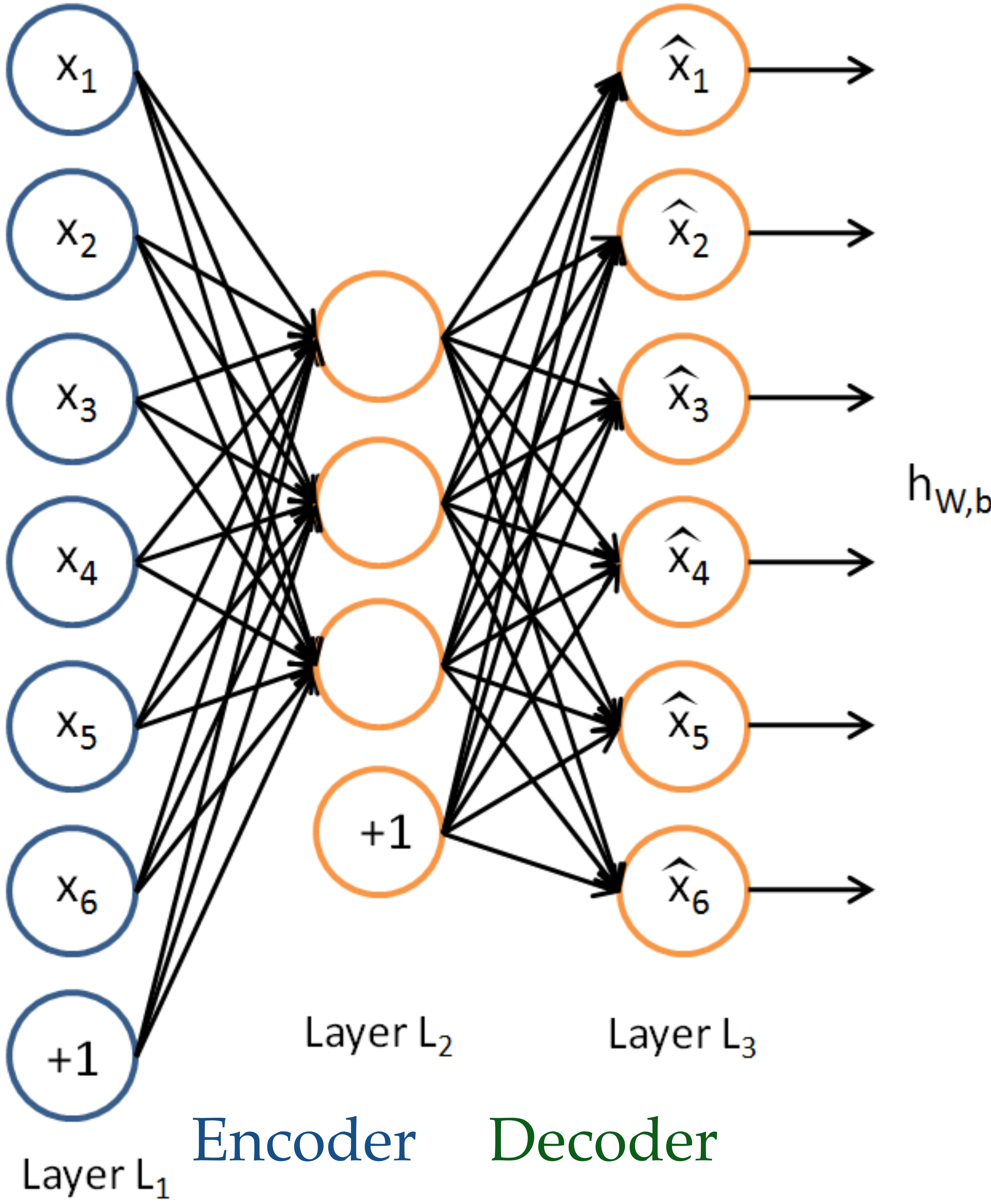
- Robust Covariance matrix estimation from multivariate data
  - Minimum Covariance Determinant (MCD) method
    - find h observations (out of n) whose covariance matrix has the lowest determinant
- TRobustEstimator class in ROOT



Anomaly Detection with ML

- AutoEncoders
  - an unsupervised neural network
  - trained by setting the target values y_i equal to the inputs x_i
  - Detect anomalies by looking at different score values obtained
    - e.g. reconstruction error:

$$\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$



Autoencoders for anomalous events

- Use auto encoder at trigger level (CMS) for potential anomalies
- Train on standard model events
  - identify anomalies by cutting on loss function
  - record anomalous events for further analysis
    - saving to disk ~ 30 evts/day

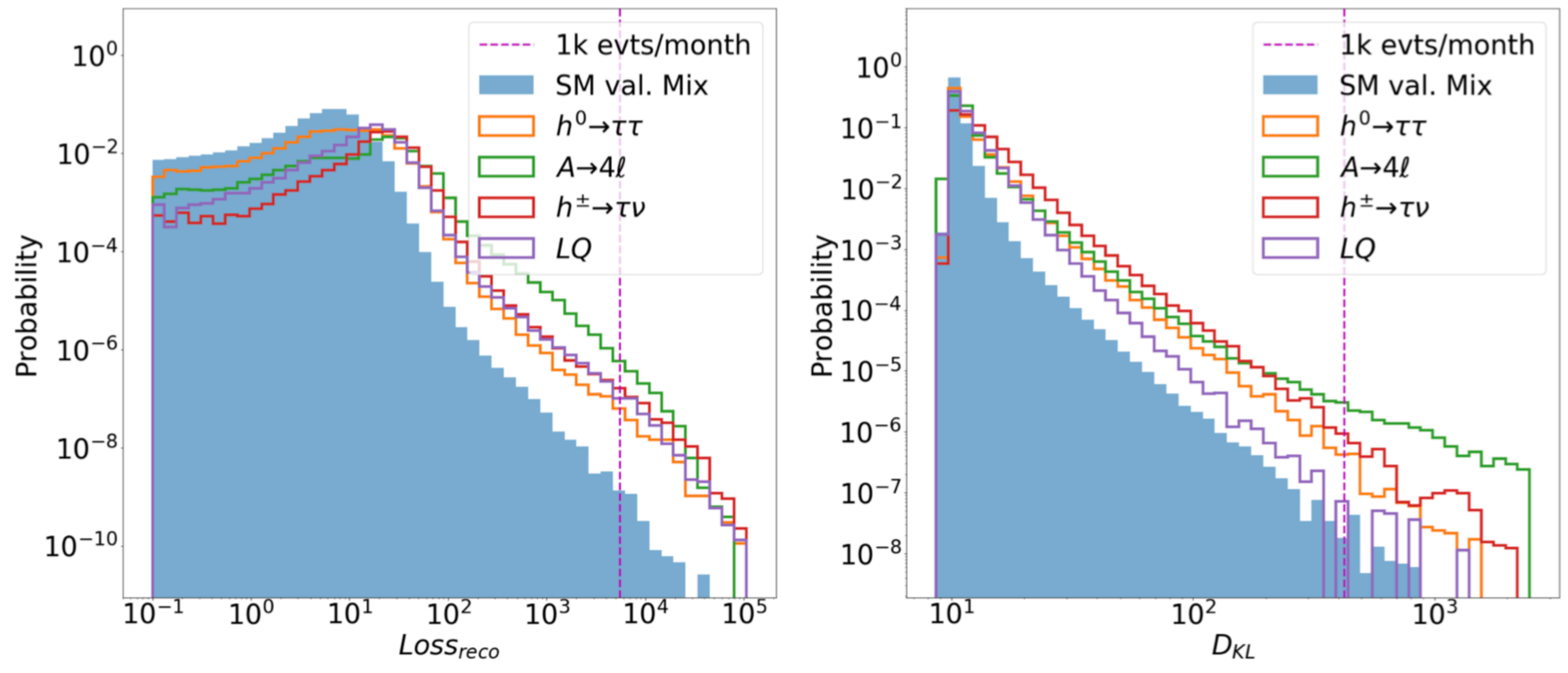
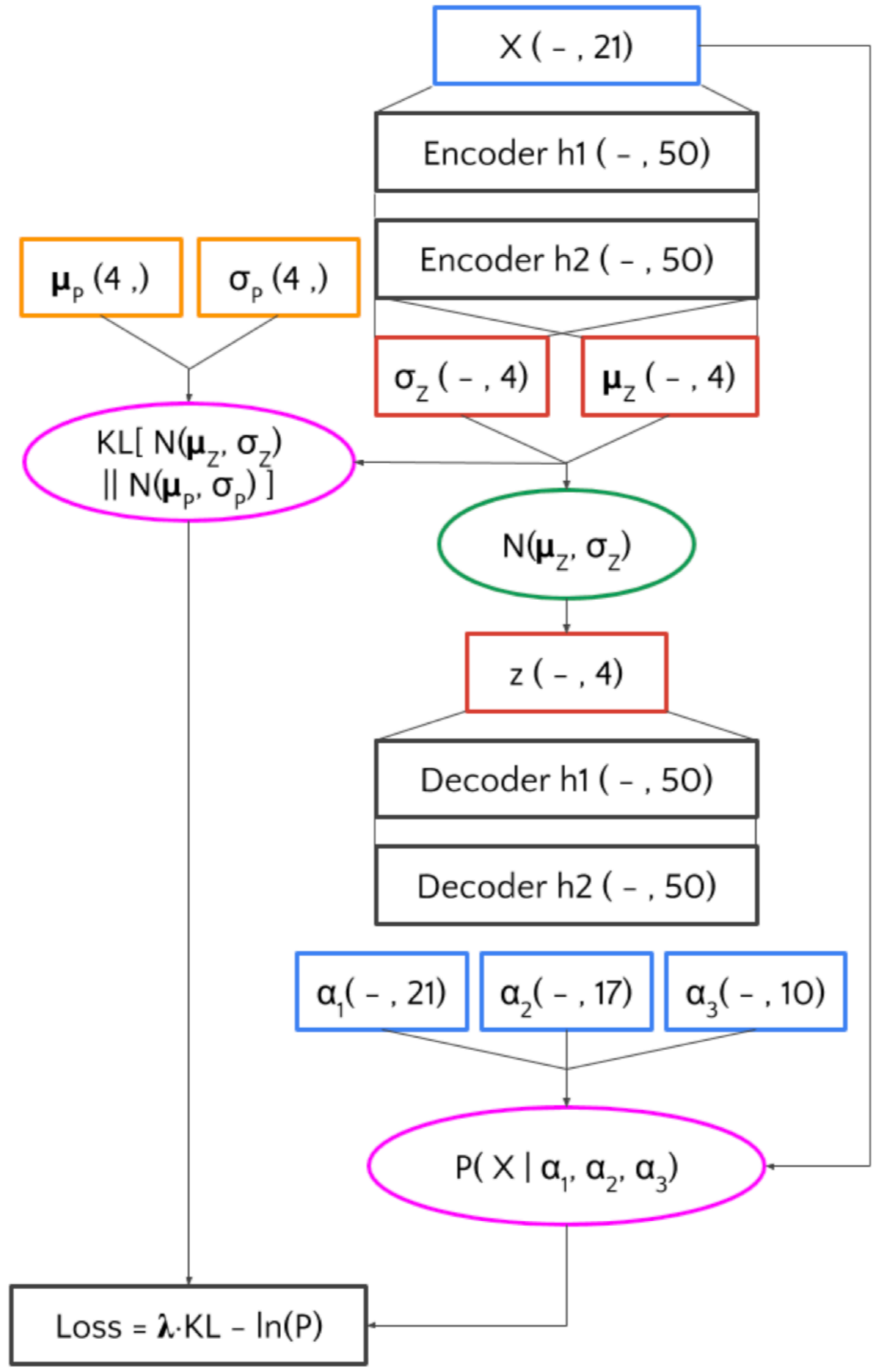


Figure 7: Distribution of the loss components: $Loss_{reco}$ (left) and D_{KL} (right) for the validation dataset. For comparison, the corresponding distribution for the SM processes and the four benchmark BSM models are shown. The vertical line represents a lower threshold such that $5.4 \cdot 10^{-6}$ of the SM events would be retained, equivalent to ~ 1500 expected SM events per month.



O.Cerri et al., arXiv:1811.10276

Data Quality Monitoring

- Unsupervised ML used to spot anomalies

[Pol *et al.*, 2018, arXiv:1808.00911]