APPENDIX VI: UNSUPERVISED- AND SEMI-SUPERVISED FEATURE EXTRACTION, AND OTHER INTERESTING VISUALIZATIONS

Two forms of Unsupervised Feature Extraction were tested for usefulness in generating features that could be useful during particle identification, i.e. Principal Component Analysis (PCA) and T-distributed stochastic neighbour embedding, the discussion of the mathematical mechanics of these methods lies outside the scope of this thesis. While it does seem promising to use these models to effectively cancel-out some of the noise in the signal by only keeping the major decorrelated factors of variation in the dataset, they are computationally expensive and therefore were not employed to this end. They do however serve as interesting ways to visualize data and their results are therefore shown below.

Principal Component Analysis

It is hard to tell from this perspective whether electrons somehow "surround" pions in this two-dimensional space, based on raw pixel data, or whether they were just plotted last. It might have been possible to find a separating soft margin in kernel space using a Support Vector Machine, but this dataset proved much too large for PCA to run, even on a balanced subset of ~500 000 tracklets.

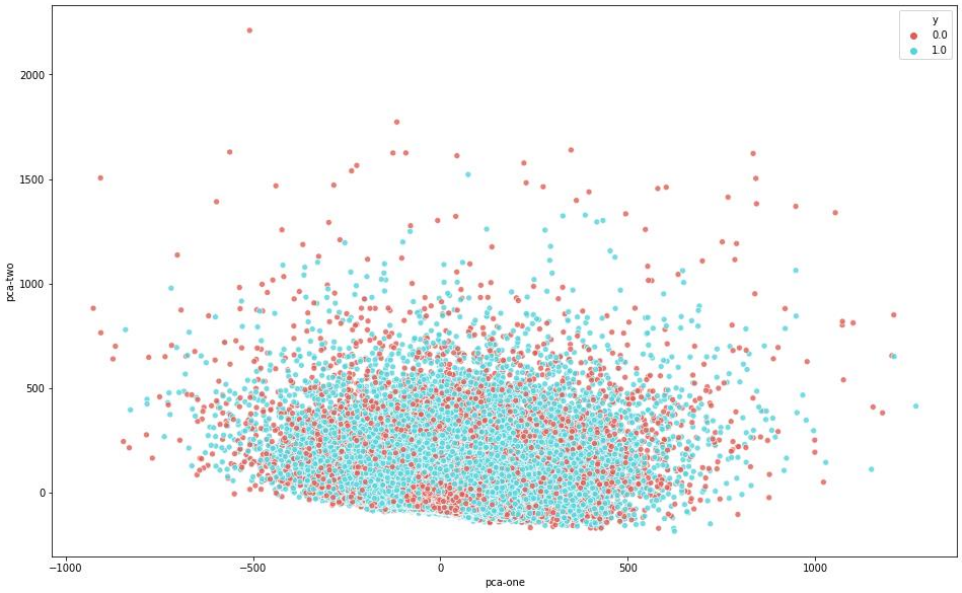

Figure 1: 2D PCA, pion = 0 (red), electron =1 (blue)

T-distributed stochastic neighbour embedding (t-SNE)

While t-SNE seems to separate the data much better, feeding its results into an SVM are equally unfeasible, but makes for an interesting plot with various clusters of points where particles seem to group, at least seemingly on average, according to their particle IDs.

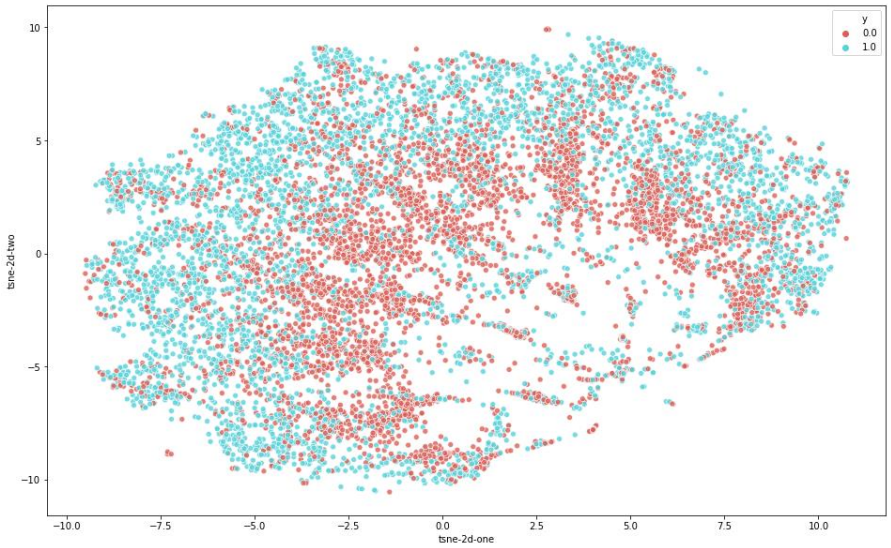

Figure 2: t-SNE

Semi-Supervised Feature Extraction

Autoencoder Dimension Reduction

An autoencoder was built using the H2O framework **Invalid source specified.**. Both the input and target values were the flattened pixel values of ~500 000 tracklet images, with an architecture of 256:128:3:128:256, trained for 600 epochs, the autoencoder has to find a vector of length 3 that explains enough variability in the data to reconstruct it with as low as possible loss. If it does well, it should extract 3 features that explain as much variability in the data as possible and are therefore not highly correlated with one another, an approximation of what happens in PCA. **Figure 3** and **Figure 4** show a 2D and a 3D projection of the weights of the hidden bottleneck layer of width 3. Interestingly, it seems to reverse the encapsulation seen in PCA, i.e. here pions seem to surround electrons, instead of the other way around, but again this could be because of the plotting sequence, since a different plotting library was used.

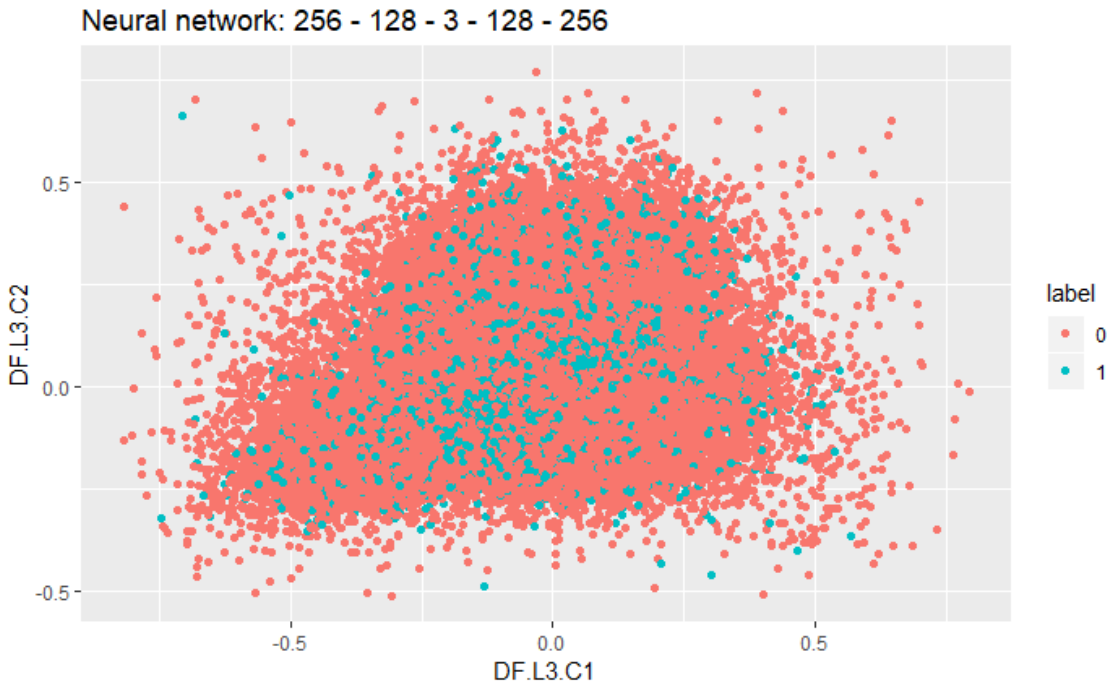

Figure 3: Two Principal Components derived from an AE's latent variables

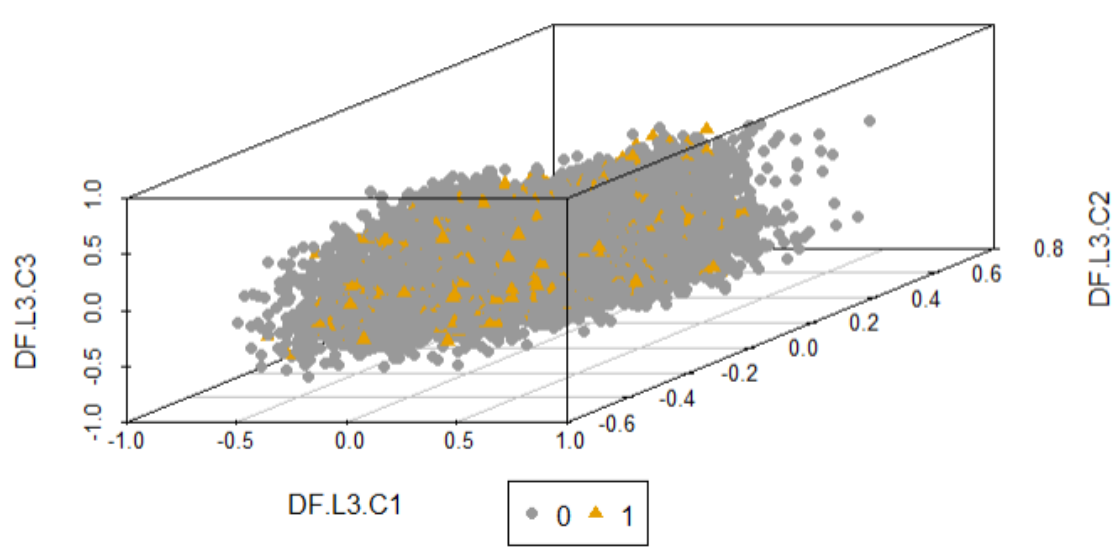

Figure 4: The three Principal Components derived from the same AE's (256:128:3:128:256) latent variables

Alternative ways of visualizing tracklet signals

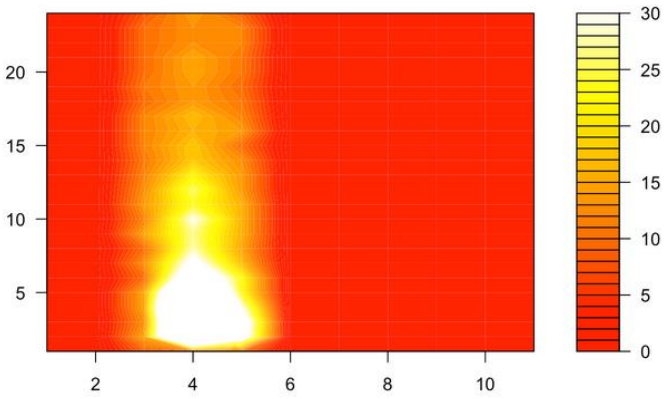

Figure 5: Filled contour map of a single pion tracklet's signal, with pads along the *x* axis (columns) and timebins along the *y* axis (rows)

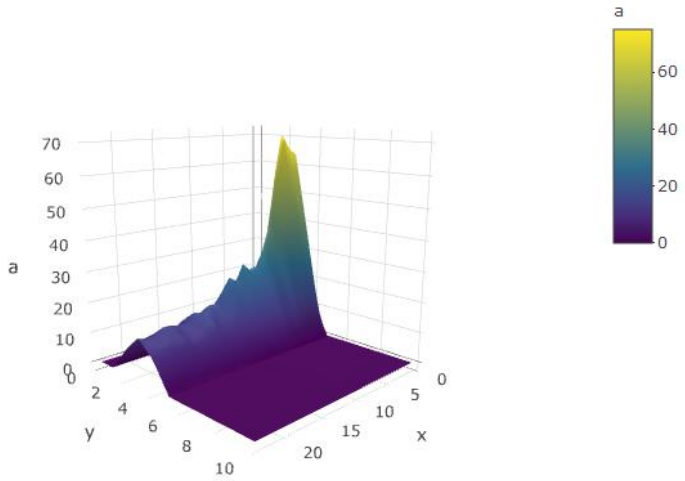

Figure 6: 3D surface plot of the signal of a single pion tracklet's signal, with timebins along *x*, pads along *y* and pulse height along *a*.

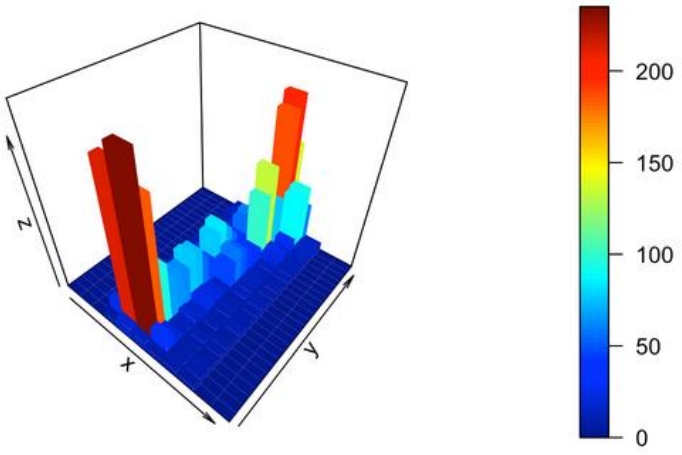

Figure 7: 3D histogram of a single electron tracklet's signal, with pads along *x*, timebins along *y* and pulse height along *z*.