Dimensionality reduction for classification using Higgs dataset

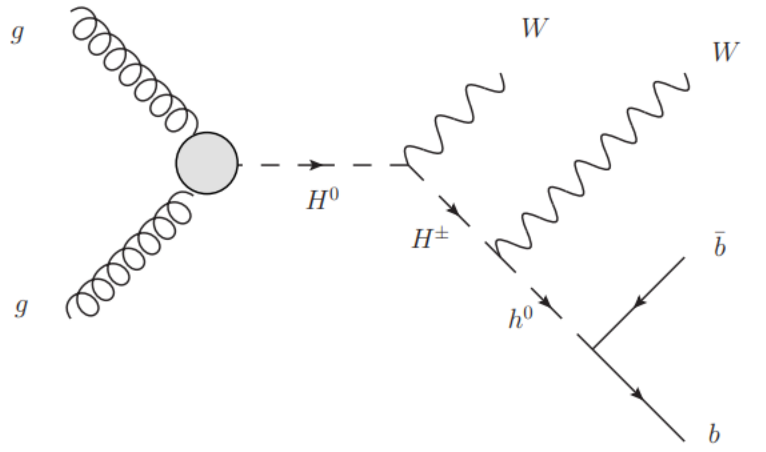
Mentee: Carlos Romero (Elmhurst University)
Mentor: Sergei Gleyzer (University of Alabama)
Ana Maria de Sousa Slivar (University of Alabama)
Eric Reinhardt (University of Alabama)

Overview

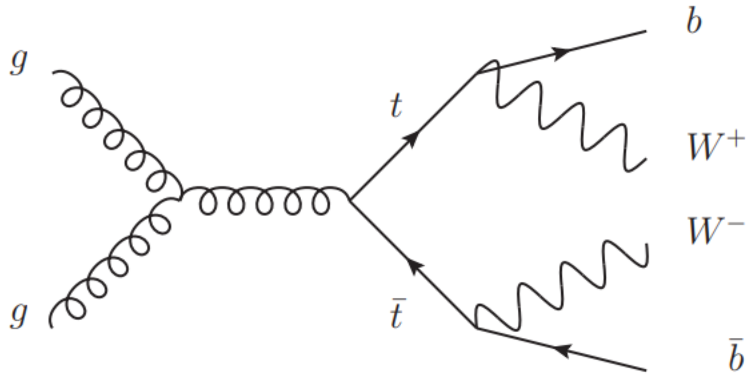
1. Introduction
2. Higgs dataset
3. Neural network design
4. Benchmark cases
5. PCA description
6. AutoEncoder description
7. Comparing performance
8. Selected features removal
9. Conclusions

Introduction

- Signal event (a) in which a Higgs boson decays into a pair of *W* bosons and a pair of bottom quarks.
- Background event (b) in which a pair of top quarks each decays into a *W* boson and a bottom quark.
- We can use low-level and high-level features to distinguish between signal and background events.



(a)



(b)

Higgs Dataset

- The data can be found on the UC Irvine ML repository:

  https://archive.ics.uci.edu/dataset/280/higgs

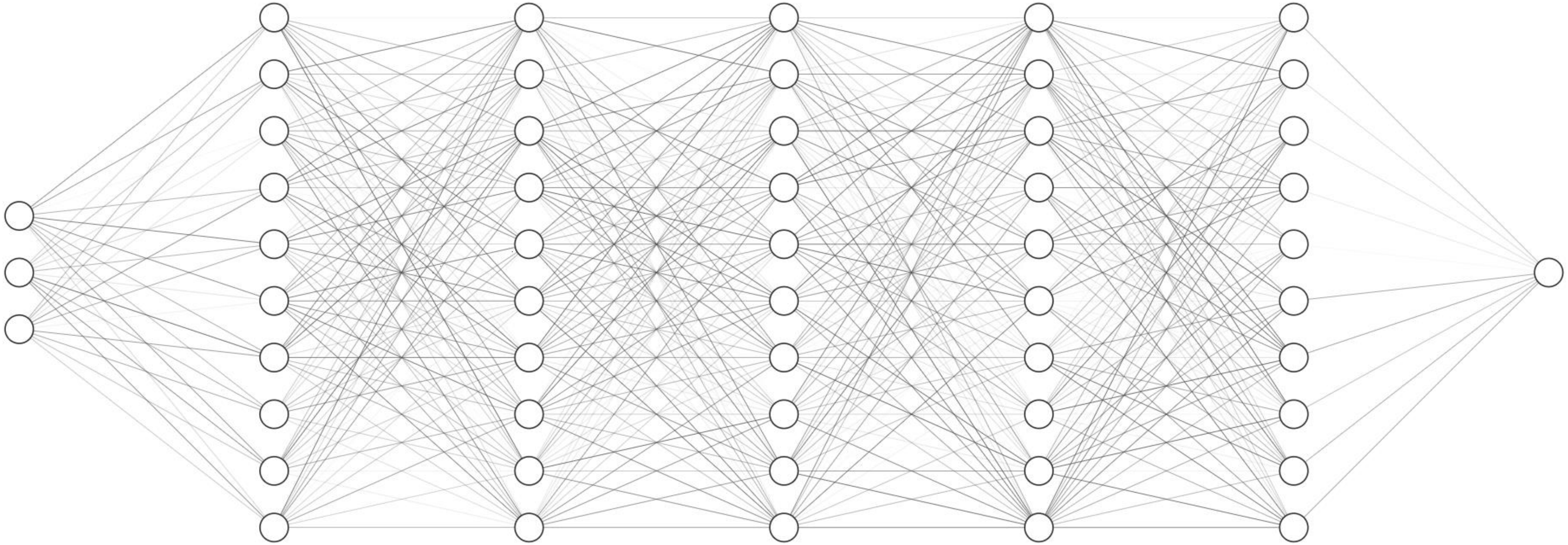
Low-level features

Lepton	Missing Energy	Jet 1	Jet 2	Jet 3	Jet 4
P_T	Magnitude	P_T	P_T	P_T	P_T
η	N/A	η	η	η	η
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ
N/A	N/A	b-tag	b-tag	b-tag	b-tag

High-level features

m_{lv}	m_{jlv}	m_{bb}	m_{wbb}	m_{wwbb}	m_{jj}	m_{jjj}

Neural Network Design



Dataset features 5 300-neurons dense layers Predicition

Final Architecture and Benchmark

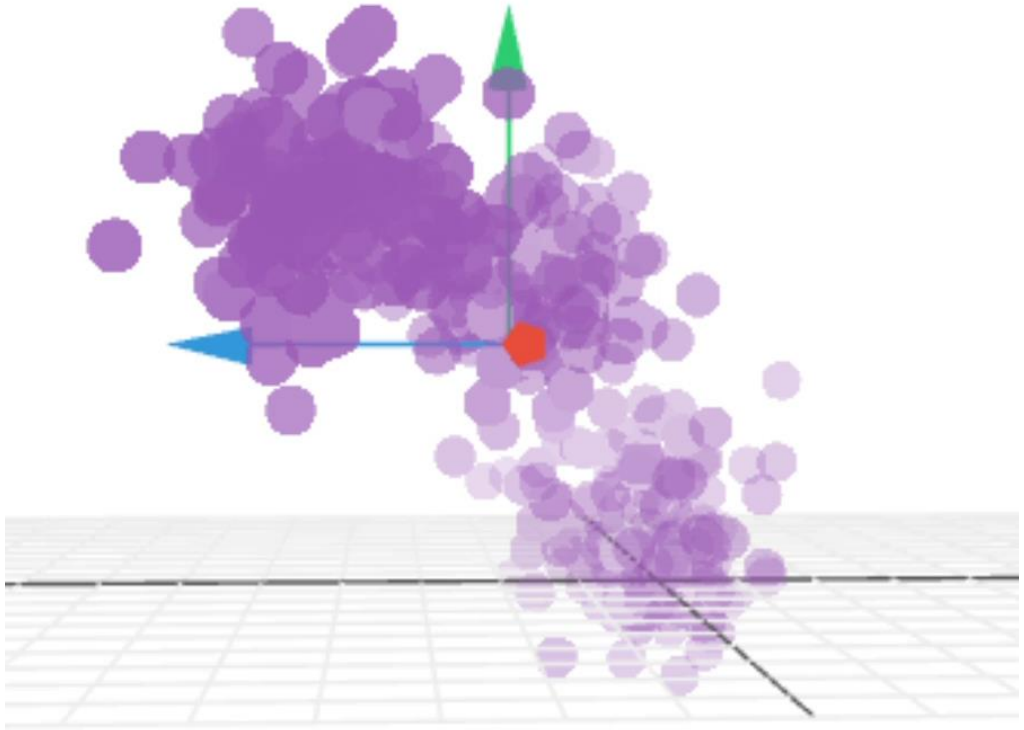
- Leaky ReLU activation function
- LR 0.05 decaying by 0.5 with a patience of 5 epochs.
- 2,500,000 events
- Dropout in every layer of 0.15
- Momentum of 0.9 and Nesterov True
- High and low-level features: ROC-AUC of 0.857
- Low-level features only: ROC-AUC of 0.828

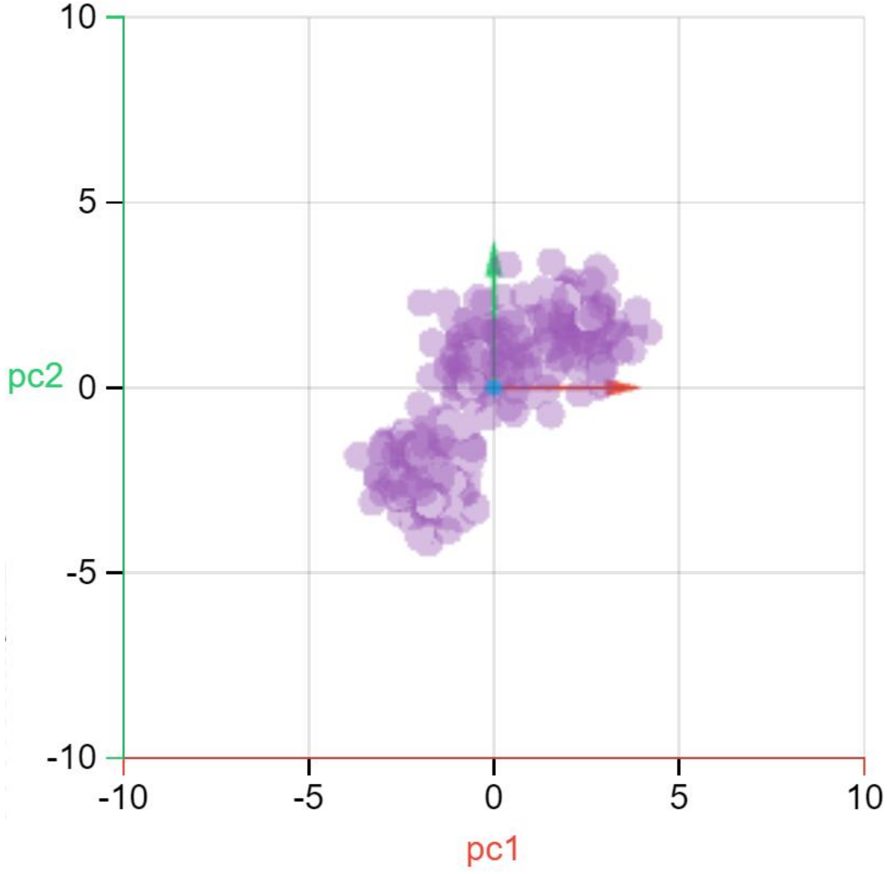
Motivation for Dimensionality Reduction

- CMS experiment produces a huge amount of data.
- The more we can reduce it, the better.
- Faster to perform analysis on events with less features.
- We do not want to lose information.

PCA

- Unsupervised ML technique that transforms high-dimension data into lower-dimensions while retaining as much information as possible.

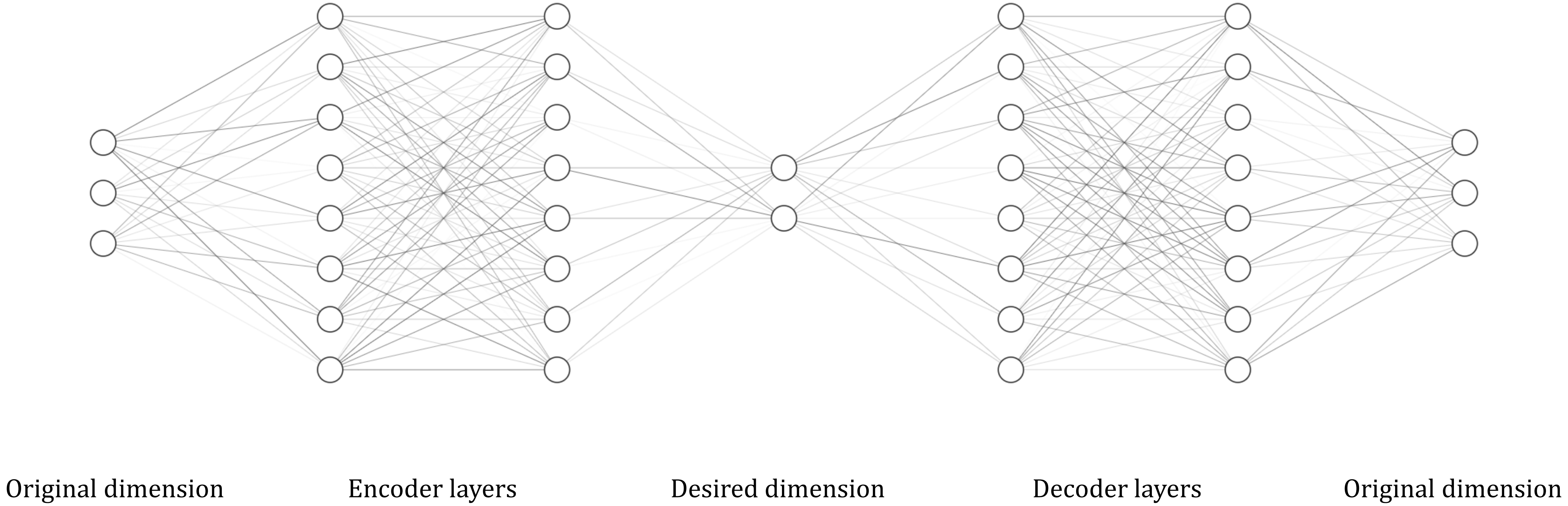




Animation and image from: https://setosa.io/ev/principal-component-analysis/

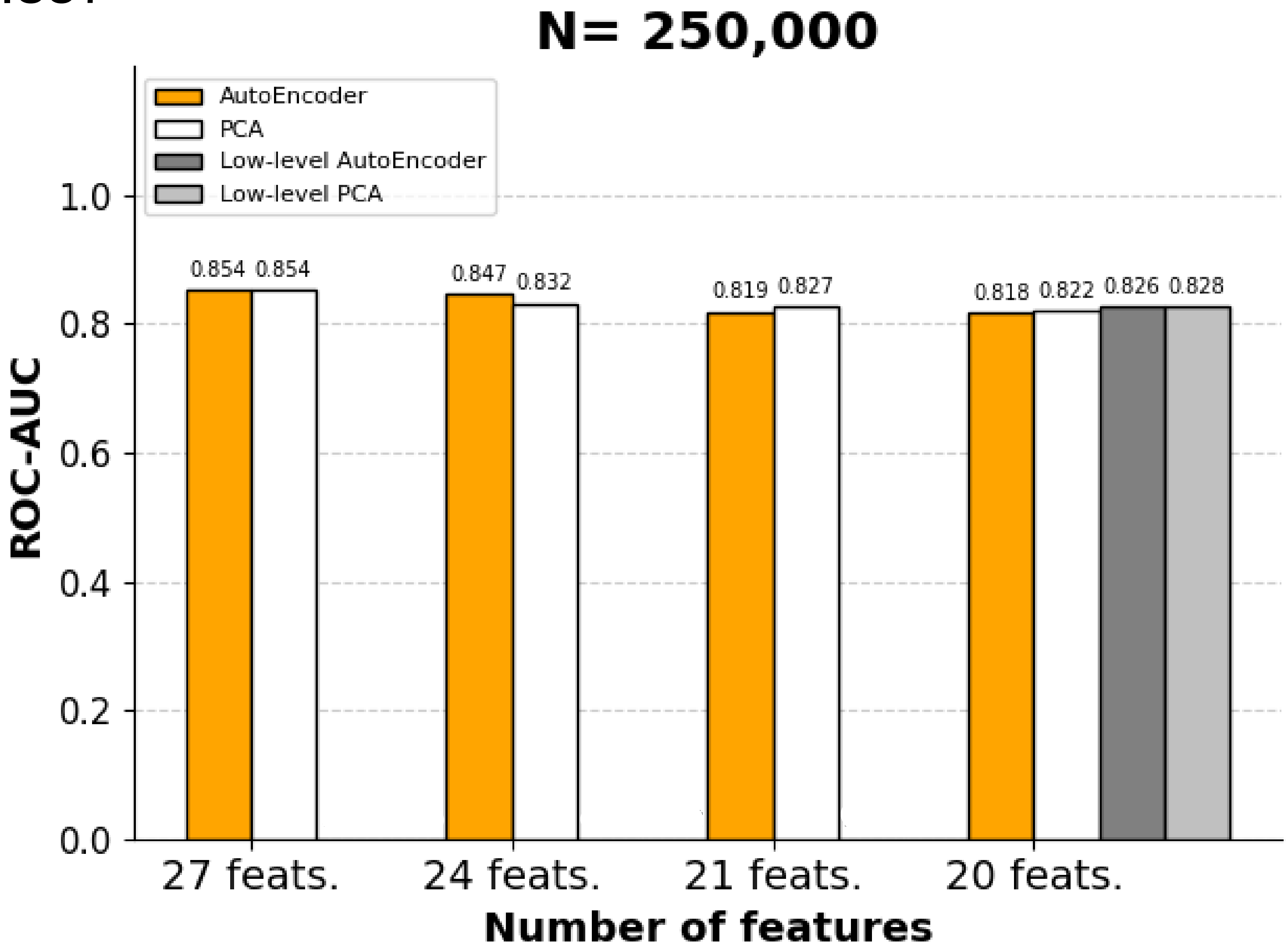
AutoEncoder

- Unsupervised neural network that encodes data by reducing dimensionality while retaining as much information as possible.



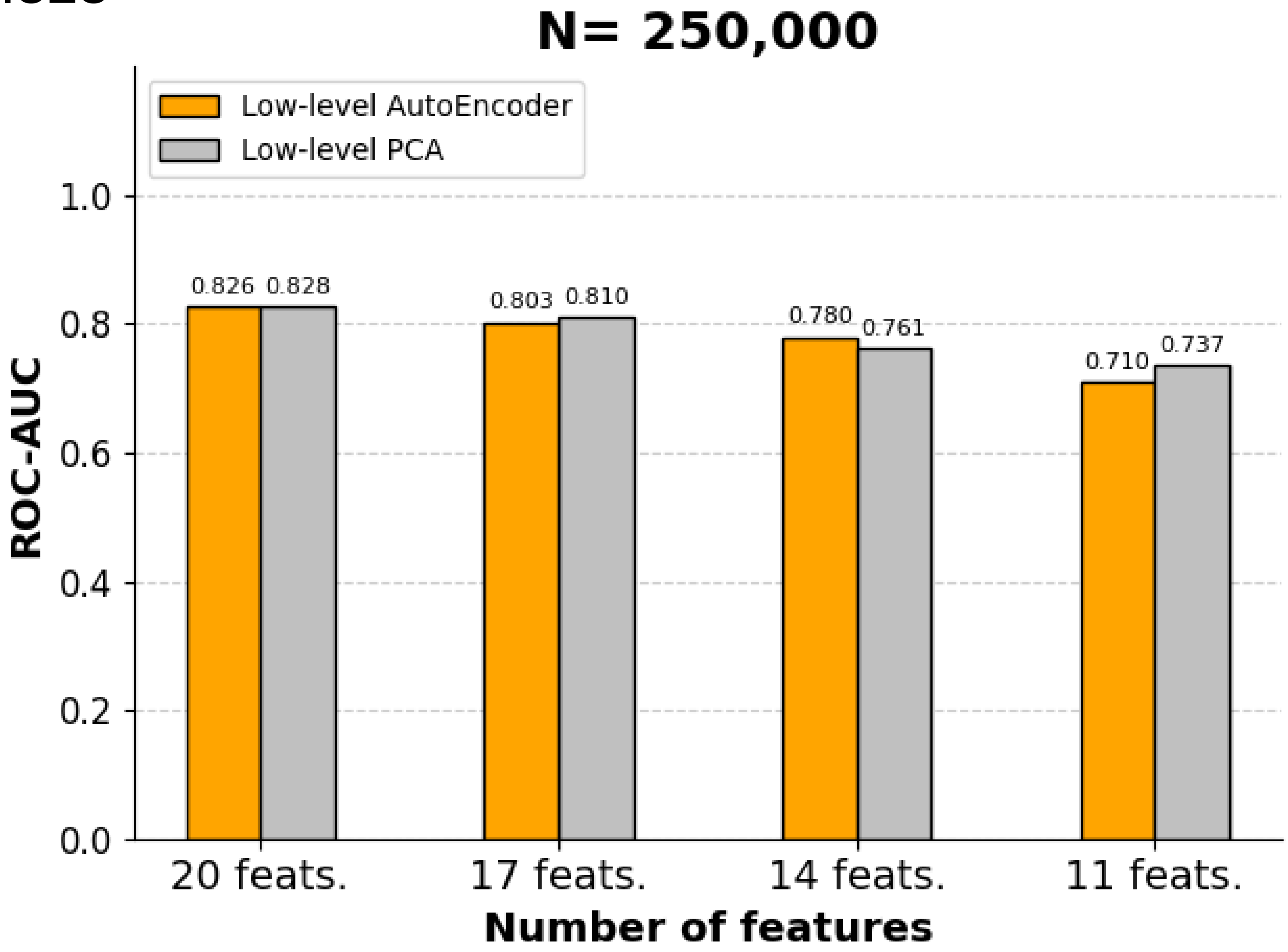
Comparing Performance

28 features: 0.857



Comparing Performance

21 features: 0.828



Selected Features Removal

- No missing energy features (19 features) : 0.815
- No b-tags from the four jets (17 features) : 0.775
- No b-tags from jets 1 and 2 (19 features) : 0.798
- No b-tags from jets 1 and 3 (19 features) : 0.801
- No b-tags from jets 1 and 4 (19 features) : 0.805
- No b-tags from jets 2 and 3 (19 features) : 0.802
- No b-tags from jets 2 and 4 (19 features) : 0.810
- No b-tags from jets 3 and 4 (19 features) : 0.814

Summary of Results

	Benchmark 28 features	Benchmark 21 features	20 features	17 features	14 features	11 features
Best Method	N/A	N/A	PCA	PCA	AutoEncoder	PCA
Best ROC-AUC	0.857	0.828	0.828	0.810	0.780	0.737

- PCA takes less time to implement than an AutoEncoder

Conclusions and Possible Future Steps

- PCA and AutoEncoders can be used to reduce dimensionality of the dataset.
- It is possible to handpick select features without losing too much information.
- Missing energy features are less important for classification.
- Higher transverse momentum jets are more important for classification.
- <u>Possible future steps</u>

  - Fine tune networks trained on reduced dataset.
  - Explore transformer autoencoder.

Thank you!

Considered parameters

- 5 layers of 256 neurons
- Dropout in every layer of 0.2
- First-layer only dropout of 0.5
- Learning rate decay scheduler