**Large-Scale Dimensionality Reduction and Feature Learning for Protein Analysis**

1. **Introduction**
   1. **Motivation**
      1. *Why protein classification is important*
      2. *Why is it big data related*
   2. **Problem Statement**
      1. *Intro to dataset*
      2. *Task (2-3 paragraph)*
         1. *dimensionality Reduction*
         2. *Feature Learning*
         3. *Comparison with downstream classification task*
   3. **State of the Art**
      1. Methods for protein classification
      2. Parallel autoencoders
      3. Project Adam …
      4. Google brain
      5. distributed learning & big data
   4. **Background**
      1. Dimensionality reduction (argue the differences of linear (PCA), downsides of Kernel PCA (because of many data), thus autoencoders)
      2. Autoencoders and its variants (denoising, tied, etc.)
      3. Classification techniques (stacked - KNN - SVM is not fast enough)
2. **Method**
   1. Language environment and API - Why using Spark -
   2. Algorithm (pseudocode) (POORYA)
   3. Synchronous
3. **Results and Evaluation**
   1. Experiments performed - Number of hidden layers - Parameters - Prevented the overflow
   2. Topology of Cluster - Docker - Which Machines
   3. Optimizer
   4. Result pre-processing
   5. Training and Testing
   6. Results of serial
   7. Results of parallel synch
   8. Error measure to the labels (01 loss - accuracy test)
4. **Discussion and Conclusion**
   1. Summary of what we did
   2. What was the observation
   3. Why async is nearly impossible in Spark
   4. Which feature outperformed
5. **References**