Appendix 1

Instructions for the Reproduction of the Results in the Paper

First, go to https://github.com/millanp95/DeepCluster and clone the repository. Then run the series of commands for each test.

Test #1:

```
1. python build_dp.py --data_path='../data/Vertebrata/Test Files'
2. python get_pairs.py --data_path='../data/Vertebrata/Test Files/train.p'
   --k=6 --modify='mutation' --output='../data/Vertebrata/Test
   Files/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Vertebrata/Test Files'
   --out_dir='../data/Vertebrata/Test Files'
```

Comparison Models:

```
● python TrainComparison.py --data_path='../data/Vertebrata/Test
  Files/testing_data.p'  --method='Supervised' --k=6
● python TrainComparison.py --data_path='../data/Vertebrata/Test
  Files/train.p'  --method='kmeans' --k=6
● python TrainComparison.py --data_path='../data/Vertebrata/Test
  Files/train.p'  --method='GMM' --k=6 --k=6
```

Test #2:

```
1. python build_dp.py --data_path='../data/Fish/Test Files/Actinopterygii'
2. python get_pairs.py --data_path='../data/Fish/Test
   Files/Actinopterygii/train.p' --k=6 --modify='mutation'
   --output=../data/Fish/Test Files/Actinopterygii/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Fish/Test Files/Actinopterygii'
   --out_dir='../data/Fish/Test Files/Actinopterygii'
```

Comparison Models:

```
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Actinopterygii/testing_data.p'   --method='Supervised' --k =6
● python TrainComparison.py --data_path='.../data/Fish/Test
  Files/Actinopterygii/train.p'   --method='k-means' --k=6
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Actinopterygii/train.p'   --method='GMM' --k=6
```

Test #3:

```
1. python build_dp.py --data_path='../data/Fish/Test Files/Neopterygii'
2. python get_pairs.py --data_path='../data/Fish/Test
   Files/Neopterygii/train.p' --k=6 --modify='mutation'
   --output=../data/Fish/Test Files/Neopterygii/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Fish/Test Files/Neopterygii'
   --out_dir='../data/Fish/Test Files/Neopterygii'
```

Comparison Models:

```
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Neopterygii/testing_data.p'  --method='Supervised' --k =6
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Neopterygii/train.p'  --method='k-means' --k=6
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Neopterygii/train.p'  --method='GMM' --k=6
```

Test #4:

```
1. python build_dp.py --data_path='../data/Fish/Test Files/Ostariophysi'
2. python get_pairs.py --data_path='../data/Fish/Test
   Files/Ostariophysi/train.p' --k=6 --modify='mutation'
   --output=../data/Fish/Test Files/Ostariophysi/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Fish/Test Files/Ostariophysi'
   --out_dir='../data/Fish/Test Files/Ostariophysi'
```

Comparison Models:

```
● python TrainComparison.py --data_path='./data/Fish/Test
  Files/Ostariophysi/testing_data.p' --method='Supervised' --k =6
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Ostariophysi/train.p' --method='k-means' --k=6
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Ostariophysi/train.p'  --method='GMM' --k=6
```

Test #5:

```
1. python build_dp.py --data_path='../data/Fish/Test Files/Cypriniformes'
2. python get_pairs.py --data_path='../data/Fish/Test
   Files/Cypriniformes/train.p' --k=6 --modify='mutation'
   --output=../data/Fish/Test Files/Cypriniformes/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Fish/Test Files/Cypriniformes'
   --out_dir='../data/Fish/Test Files/Cypriniformes'
```

Comparison Models:

```
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Cypriniformes/testing_data.p'  --method='Supervised' --k =6
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Cypriniformes/train.p' --method='k-means' --k=6
● python TrainComparison.py --data_path='../data/Fish/Test
  Files/Cypriniformes/train.p' --method='GMM' --k=6
```

Test #6:

```
1. python build_dp.py --data_path='../data/Fish/Test Files/Cyprinidae'
2. python get_pairs.py --data_path='../data/Fish/Test Files/Cyprinidae/train.p'
   --k=6 --modify='mutation' --output=../data/Fish/Test
   Files/Cyprinidae/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Fish/Test Files/Cyprinidae'
   --out_dir='../data/Fish/Test Files/Cyprinidae'
```

Comparison Models:

```
• python TrainComparison.py --data_path='../data/Fish/Test
  Files/Cyprinidae/testing_data.p'   --method='Supervised' --k =6
• python TrainComparison.py --data_path='../data/Fish/Test
  Files/Cyprinidae/train.p'  --method='k-means' --k=6
• python TrainComparison.py --data_path='../data/Fish/Test
  Files/Cyprinidae/train.p' --method='GMM' --k=6
```

Test #7:

```
1. python build_dp.py --data_path='../data/Bacteria/Test Files'
2. python get_pairs.py --data_path='../data/Bacteria/Test Files/train.p' --k=6
   --modify='mutation' --output='../data/Bacteria/Test Files/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Bacteria/Test Files'
   --out_dir='../data/Bacteria/Test Files'
```

Comparison Models:

```
• python TrainComparison.py --data_path='../data/Bacteria/Test
  Files/testing_data.p'  --method='Supervised' --k =6
• python TrainComparison.py --data_path='../data/Bacteria/Test Files/train.p'
  --method='k-means' --k=6
• python TrainComparison.py --data_path='../data/Bacteria/Test Files/train.p'
  --method='GMM' --k=6
```

Test #8:

```
1. python build_dp.py --data_path='../data/ProteoBacteria/Test Files'
2. python get_pairs.py --data_path='../data/ProteoBacteria/Test Files/train.p'
   --k=6 --modify='mutation' --output='../data/ProteoBacteria/Test
   Files/testing_data.p'
3. python TrainCluster.py --data_dir='../data/ProteoBacteria/Test Files'
   --out_dir='../data/ProteoBacteria/Test Files'
```

Comparison Models:

```
• python TrainComparison.py --data_path='../data/ProteoBacteria/Test
  Files/testing_data.p' --method='Supervised' --k =6
• python TrainComparison.py --data_path='../data/ProteoBacteria/Test
  Files/train.p' --method='k-means' --k=6
• python TrainComparison.py --data_path='../data/ProteoBacteria/Test
  Files/train.p' --method='GMM' --k=6
```

Test #9:

```
1. python build_dp.py --data_path='../data/Influenza-A/Test Files'
2. python get_pairs.py --data_path='../data/Influenza-A/Test Files/train.p'
   --k=6 --modify='mutation' --output='../data/Influenza-A/Test
   Files/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Influenza-A/Test Files'
   --out_dir='../data/Influenza-A/Test Files'
```

Comparison Models:

```
● python TrainComparison.py --data_path='../data/Influenza-A/Test
  Files/testing_data.p'  --method='Supervised' --k =6
● python TrainComparison.py --data_path='../data/Influenza-A/Test
  Files/train.p' --method='k-means' --k=6
● python TrainComparison.py --data_path=../data/Influenza-A/Test
  Files/train.p' --method='GMM' --k=6
```

Test #10:

```
1. python build_dp.py --data_path='../data/Dengue/Test Files'
2. python get_pairs.py --data_path='../data/Dengue/Test Files/train.p' --k=6
   --modify='mutation' --output='../data/Dengue/Test Files/testing_data.p'
3. python TrainCluster.py --data_dir='../data/Dengue/Test Files'
   --out_dir='../data/Dengue/Test Files'
```

Comparison Models:

```
● python TrainComparison.py --data_path='../data/Dengue/Test
  Files/testing_data.p' --method='Supervised' --k =6
● python TrainComparison.py --data_path='../data/Dengue/Test Files/train.p'
  --method='k-means' --k=6
● python TrainComparison.py --data_path='../data/Dengue/Test Files/train.p'
  --method='GMM' --k=6
```

Test #11:

```
1. python build_dp.py --data_path='../data/HBV/Test Files'
2. python get_pairs.py --data_path='../data/HBV/Test Files/train.p' --k=6
   --modify='mutation' --output='../data/HBV/Test Files/testing_data.p'
3. python TrainCluster.py --data_dir='../data/HBV/Test Files'
   --out_dir='../data/HBV/Test Files'
```

Comparison Models:

```
1. python TrainComparison.py --data_path='../data/HBV/Test
   Files/testing_data.p'  --method='Supervised' --k =6
2. python TrainComparison.py --data_path='../data/HBV/Test Files/train.p'
   --method='k-means' --k=6
3. python TrainComparison.py --data_path='../data/HBV/Test Files/train.p'
   --method='GMM' --k=6
```

DeLUCS computational pipeline for running your own dataset:

1. Build the dataset:
   - ○ Input: Folders with the sequences in FASTA format
   - ○ Output : file in the form (label,sequence,accession)

   *If the true label is unknown then use the same label for all the sequences.

```
 python build_dp.py --data_path=<PATH_sequence_folder>
```

2. Compute the mimics.
   - ○ Input: file in the form (label, sequence, accession)
   - ○ Output : file in the form of (pairs, x_test, y_test).

```
 python get_pairs.py --data_path=<PATH_pickle_dataset> --k=6 --modify='mutation'
--output=<PATH_output_file>
```

3. Train the model.
   - ○ Input: file in the form of (pairs, x_test, y_test).
   - ○ Output :
     - ● Image with the confusion matrix (provided the "ground truth" is available)
     - ● File with the misclassified sequences in the form (accession, true_label, predicted_label)

```
python TrainCluster.py --data_dir=<PATH_of_computed_mimics> --out_dir=<OUTPURDIR>
```