## Scripts

Table 4: Overview all python scripts used predicting cros- contamination between samples in a batch with the script name, a small description of functionality script, a step name and the number of the step in figure 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Script** | **Small description** | **Step name** | **Step number figure 1** |
| Getting\_all\_contig\_seqs\_from\_clc.py | Retrieves all contig sequences of samples from CLC database | Real world sample data | 2 |
| Read\_krona\_files.py | Retrieves for every sample virus names, rank and contigs | Real world sample data | 2 |
| Get\_krona\_info.py | Retrieves from files with virus name, rank and contig the sample id, host name, virus names and contigs | Real world sample data | 2 |
| Get\_contaminant\_virus.py | Gets the potential contaminants by getting all viruses per sample  and grouping them per batch. If 2 or more samples in a batch have a virus save  with their sample name, hostname, viruses and the count of viruses. | Getting potential contaminants | 2 |
| Make\_contamination\_files.py | Subsamples files and puts them in other files to get files with contamination. | In silico contamination | 1 |
| Get\_host\_virus\_correlation.py | Get correlation between all hosts and viruses | Get host virus correlation | 3 |
| Calculate\_ani\_score.py | Calculate average nucleotide identity | Get identity | 4 |
| Calculate\_fragmentation.py | Calculate fragmentation of virus in sample compared to same virus in batch | Get fragmentation | 5 |
| Count\_contigs.py | Calculates total length all contigs and count of contigs per virus in a sample | Get fragmentation | 5 |
| Get\_labels\_from\_excel.py | Retrieve label if contamination occurred from excel files from molbio server | Get labels | 6 |
| Make\_train\_and\_test\_file.py | Makes training and testing file for the machine learning model | Prepare data | 7 |
| Train\_model.py | Trains … machine learning models and evaluates them | Model for crosscontamination | 8 |

A diagram of a flowchart

AI-generated content may be incorrect.

Figure 2: Pipeline for training cross-contamination prediction model. First in silico data is made and real world sample data is gathered. This data is processed and host-virus correlation, identity and fragmentation are calculated from this data. These features are prepared into test, validation and train set and different models are trained.