Immunogenicity prediction of viral neoantigens in murine hosts

The growing amount of experimental data allows for the creation of better models deciphering mechanisms of the adaptive immune response. Such models may be useful in cancer immunogenomics for personalized vaccine design and to predict the clinical outcome of cancer treatments. While there is a limited human clinical datasets available, the analysis of data obtained from murine studies could provide useful scientific insights.

For this purpose, a curated dataset of T cell (CD8+) immunogenic response to viral peptides presented on murine Major Histocompatibility Complex (MHC) class I molecules, measured using cytotoxicity and IFN γ release assays, will be provided. The task is to build a model predicting the probability of peptide-MHC (pMHC) immunogenicity. The performance of the model will be evaluated on an independent test set.

Dataset for model training is made available in the tab-separated values (.tsv) format:

train.tsv - training dataset
The dataset contains 3 columns:
1. Peptide - 8-11 amino acid sequence, only 20 standard amino acids
2. MHC - one of the 6 most popular murine MHC types
3. Immunogenicity - experimentally measured immunogenicity (mostly using IFN γ, TFNa, IL-2 release and cytotoxicity assays), 1 for immunogenic, 0 for non-immunogenic. This is the taget variable.

Example:

Peptide	MHC	Immunogenicity
AAALSPMEI	H2-Db	0.0
AAASVVGAPV	H2-Db	0.0
AAEEFAFL	H2-Kb	0.0
AAFNLPIIEL	H2-Kb	0.0
AAFTFTKI	H2-Kb	1.0

Note that extra tabs in the example above are inserted for visual clarity only.

Dataset for model evaluation (will be made available 2 hours before the end of the competition):

test.tsv
The evaluation dataset will not have the `Immunogenicity` column, otherwise the format is the same.

Example:

```
Peptide         MHC
AAALSPMEI       H2-Db
AAASVVGAPV      H2-Db
AAEEFAFL        H2-Kb
AAFNLPIIEL      H2-Kb
AAFTFTKI        H2-Kb
```

The participants should prepare a script which accepts a single argument and outputs a predictions file. The participants will not be able to copy the file, it will be provided on a pen drive by mentors.

Example:
$ ./script example_test.tsv > out.tsv

The output file format is the same as for train.tsv, however the Immunogenicity column must contain immunogenicity probability predictions.

Example:

```
Peptide         MHC         Immunogenicity
AAALSPMEI       H2-Db       0.2
AAASVVGAPV      H2-Db       0.3
AAEEFAFL        H2-Kb       0.1
AAFNLPIIEL      H2-Kb       0.45
AAFTFTKI        H2-Kb       0.93
```

Evaluation

Your solution will be evaluated using ROC AUC on the evaluation set. Therefore, make sure you provide the probabilities or the decision function values of your model, not just binary predictions. Higher values must indicate higher immunogenicity.