# MixMHC2pred

MixMHC2pred is a pan-allele predictor of MHC class II ligands and epitopes. It is described in:
Racle, J., et al., Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *bioRxiv* (2022) (available here).

and

Racle, J., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* 37, 1283–1286 (2019) (available here).

MixMHC2pred is also available as a web application: http://mixmhc2pred.gfellerlab.org.

## Installation

1. Download MixMHC2pred-2.0.zip file (https://github.com/GfellerLab/MixMHC2pred/releases) and move it to a directory of your choice, where you have writing permissions.

2. Unzip MixMHC2pred-2.0.zip in this directory.

3. Above's zip file already contains the human alleles' definition files. If you want to do predictions for MHC-II alleles from other species or if you cloned the git repository instead of downloading this zip file, you will need to download the wished alleles' definition files from the page: http://mixmhc2pred.gfellerlab.org/PWMdef. These additional alleles' PWM definition files need to then be unzipped in the folder of your choice. The option `-f <folders>` is then used to indicate where these files are (see below for details). Alternatively, you can copy/move all downloaded allele definition files into the folder *PWMdef* at the root of the directory where MixMHC2pred is installed, to avoid using the `-f` option.

4. To test your installation, make sure you are in *MixMHC2pred-2.0* directory and run the following command, depending on your operating system:

   - Mac OS: `./MixMHC2pred -i test/testData.txt -o test/out.txt -a DRB1_15_01 DRB5_01_01 DPA1_02_01__DPB1_01_01 DQA1_01_02__DQB1_05_01 DQA1_01_02__DQB1_06_02`

   - Unix: `./MixMHC2pred_unix -i test/testData.txt -o test/out.txt -a DRB1_15_01 DRB5_01_01 DPA1_02_01__DPB1_01_01 DQA1_01_02__DQB1_05_01 DQA1_01_02__DQB1_06_02`

   - Windows: `MixMHC2pred.exe -i test/testData.txt -o test/out.txt -a DRB1_15_01 DRB5_01_01 DPA1_02_01__DPB1_01_01 DQA1_01_02__DQB1_05_01 DQA1_01_02__DQB1_06_02`

   Your file *test/out.txt* should be the same as *test/out_compare.txt*. Running the software takes few seconds or more when testing lots of peptides and alleles.

   The *testData.txt* file corresponds to a subset of the HLA-II peptidomics data obtained from the cell line *DOHH2* in Dheilly et al., *Cancer Cell* (2020), containing some peptides bound to the HLA in the reverse direction.

5. (Optional) To run MixMHC2pred from anywhere on your computer, make an alias of MixMHC2pred executable or add it in your path.

If using a non-standard OS, it is possible to compile MixMHC2pred using the Makefile found in the *bin* folder.

# Running

## Command

```
MixMHC2pred -i input_file -o output_file -a allele1 allele2 [additional
options]
```

- Depending on your operating system, use MixMHC2pred, MixMHC2pred_unix or MixMHC2pred.exe as indicated in the installation instructions.
- Do not use spaces in your file or directory names.
- Do not use other special characters (e.g., *, ?, %, &,...) in file or directory names.

## Required arguments

- `-i <file>` or `--input <file>` (input file name):
  File listing all the peptides. It should contain two columns: the 1st column being the sequence of the peptide and 2nd column beeing its context sequence (12 amino acids long: 3 residues upstream of the peptide, 3 N-terminal residues of the peptide, 3 C-terminal residues of the peptide and 3 residues downstream of the peptide). When the peptide lies near the begin or end of a protein, the corresponding context AAs should be written as "-", i.e. for the protein *ACDEFG...* if the peptide is *CDEFG...* the first 6 AA encoding its context should be written as *--ACDE* (and these 6 AAs should be directly followed by the 6 AAs describing the context near the C-terminal of the peptide). Also, if some AAs from the context of a peptide are not known, the unknown AAs should be written with the letter *X*. See *test/testData.txt* for an example of input file. When using the `no_context` option (see below), then this input file should only contain the list of the peptides, without any 2nd column of the context (an example input file without in such a case is available at *test/testData_noContext.txt*).

- `-o <file>` or `--output <file>` (output file name):
  The name of the output file (including the directory). Peptides are kept in the same order than in the input file.

- `-a <alleles>` or `--alleles <alleles>`:
  List of MHC-II alleles to test. If you want to make predictions with multiple alleles, list the different alleles separated by a space (e.g. `-a DRB1_11_01 DRB3_02_02`).
  Use the nomenclature *DRB1_03_01* for HLA-DRB1*03:01 and *DPA1_01_03__DPB1_04_01* for HLA-DPA1*01:03-DPB1*04:01. The list of alleles available and corresponding nomenclature is given in http://mixmhc2pred.gfellerlab.org/PWMdef (also given in the files *Alleles_list_xxx.txt* in the corresponding *PWMdef* folders). Simply, the names used in MixMHC2pred are obtained by dropping the *HLA-* from human alleles, by replacing all "-", "*" and ":" by "_", and by placing "__" between the alpha and beta

chains forming the heterodimer (the invariant DR-alpha chains are not indicated in the allele names as in standard practice).

## Optional arguments

- `--no_context`:
  In principle, MixMHC2pred includes the peptide context for its predictions (i.e. corresponding to a sequence of 12 AAs in total, including AAs just before and just after the peptide as explained above). It is nevertheless possible to decide to not consider any context information at all, when using this option. It is generally advised to include the context, in order to search for best candidate epitopes. But if analyzing a posteriori pre-cleaved peptide sequences (e.g. in experiments testing specific peptides directly, that therefore did not need to be cleaved by the cell), it may be a good to not consider the context encoding (often multiple overlapping epitopes are observed, so the peptide tested may not correspond to the best peptide based on context but it could still be recognized by the same T cells when given directly). When using this `--no_context` option, then the input file should only contain the list of peptides, without their context. An example input file is available at *test/testData_noContext.txt*; results of running MixMHC2pred without the context on this file, based on the same alleles as the first example above is given in *test/out_noContext_compare.txt*.

- `-f <folders>` or `--allelesFolder <folders>` (folder(s) containing allele definitions):
  If the folder containing the PWM definition files of the alleles is not at its default location ("path_to_exec/PWMdef"), you can give this option indicating where these files are located. It is possible to list multiple folders, separated by a space (when an allele is found in multiple folders, the first definition found for this allele is used). This path can be given as a full path, a path relative to current location, or with the special keyword *exec:* (e.g., `-f exec:PWMdef`) to give the path relative to the root folder of MixMHC2pred executable. An other special keyword *default* can also be used: it represents the default path where these files are located (i.e., it is equivalent to using `exec:PWMdef`).

- `-e` or `--extra_out`: By default, the score returned by MixMHC2pred is the final peptide presentation *%Rank* score and this is the score recommended to use in all analyses/predictions. If you are however interested by other intermediate scores, you can pass the `-e` option. In this case, additional columns are appended to the output file (first columns are the same as when running MixMHC2pred without this option). The new columns give:

  - the *Score_...* corresponding to the raw score returned by the 2nd block of the neural network described in our paper, a value of 1 being the best score and 0 the worst. Note however that the %Rank are obtained from this score by taking into account the expected peptide length distribution. Note also that the neural network is repeated multiple times, with the returned %Rank / Score being the average of the %Rank / Score from the repetitions, respectively, so it is not possible to directly transform this returned Score to the %Rank.
  - *ScorePWM_...* corresponding to the binding scores based on the position weight matrices (eq. (2) from our manuscript). The worst score is 0 and bigger scores are better without having an upper limit. The corresponding *%RankPWM_...* are percentiles computed separately per peptide length.

## Results returned and additional information

- MixMHC2pred is meant for scoring different peptides and prioritising the most likely HLA-II ligands and epitopes. As it is trained on naturally presented peptides, it does not output a predicted affinity value, simply a score.

- Input should consist in a list of peptides, not proteins. Currently, MixMHC2pred is not cutting longer peptides/proteins into shorter fragments: it uses the peptides given in input as is.

- The score is computed for each allele provided in input. Results are returned for each allele in separate columns and additional columns give the results from the best allele for each peptide (columns *BestAllele* and *..._best* in the output file, determined by the allele that had the best score, i.e. the most likely allele by which the peptide would be presented).

- The two first columns of the output file give the *Peptide* and *Context* sequence of the peptide, which were given in the input file. When the option `--no_context` is used, the column *Context* is kept but it is empty.

- The scores returned (columns *%Rank*) correspond to a percentile rank (best score is about 0, worst score is 100). This tells among random peptides, the percent of peptides expected to be better presented by this allele than the given peptide.

- The *CoreP1_...* columns tell what is the most likely binding core position for the given peptide towards the given allele (this tells the position of the first amino acid from the binding core (which has a size of 9 aa in the predictions), starting at a value of 1 (i.e., if binding core corresponds to the 9 first amino acids from the peptide, this *CoreP1 = 1*)).

- For conveniance, the binding core sequence is also indicated for the best allele per peptide (column *Core_best*, for the other alleles this can be obtained from the *CoreP1* as indicated above).

- The *subSpec_...* columns tell in which sub-specificity the given peptide is likely bound toward the given allele. The value *1* corresponds to the main sub-specificity (the only one for most alleles). But for example for *DRB1*08:01* allele a 2nd sub-specificity exists and is indicated by the value *2*. For alleles accomodating reverse binding, a value of *-1* indicates that the given peptide is bound in the reverse orientation.

- Peptides shorter than 12 amino acids, longer than 21 amino acids or containing non-standard amino acids are kept but with a score of "NA".

- The list of alleles available is provided in http://mixmhc2pred.gfellerlab.org/PWMdef (also given in the files *Alleles_list_xxx.txt* in the corresponding *PWMdef* folders). These files show the nomenclature to use when running MixMHC2pred and the standard nomenclature used for example in the IPD-IMGT/HLA database.

## Latest version

Latest version of MixMHC2pred is available at https://github.com/GfellerLab/MixMHC2pred/releases.

Check the file *NEWS* to see the main changes of the given version.

## Web application

MixMHC2pred is also available as a web application at http://mixmhc2pred.gfellerlab.org.

## License

MixMHC2pred can be used freely by academic groups for non-commercial purposes (see license). The product is provided free of charge, and, therefore, on an "as is" basis, without warranty of any kind.

**FOR-PROFIT USERS**: If you plan to use MixMHC2pred (version 2.0) or any data provided with the script in any for-profit application, you are required to obtain a separate license. To do so, please contact Nadette Bulgin (nbulgin@lcr.org) at the Ludwig Institute for Cancer Research Ltd.

## Contact information

For scientific questions, please contact Julien Racle (julien.racle@unil.ch) or David Gfeller (david.gfeller@unil.ch).

For license-related questions, please contact Nadette Bulgin (nbulgin@lcr.org).

## How to cite

To cite MixMHC2pred, please refer to:

Racle, J., et al., Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *bioRxiv* (2022) (available here).

and

Racle, J., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* 37, 1283–1286 (2019) (available here).