Explanation for running AutoEncoder code

Installations:

1. Install workspace (anaconda) or use requirements file
2. Install the package tensorflow-gpu version 2.1.0

How to run the code for your data set:

Download all files and folder "weights_Default" from GitHub.

Add the folder of your data set (details under section: "Input format (of the data set)" in next section), change the parameters in the JSON file, as will be explained in this document, and run it.

Input format (of the data set):

The data set have to be in the following format:

Folder with the name of the data set, including csv files.

Each file arranged in the following order:

Column includes the amino acid sequence of CDR3, with a header (the parameter "CDR3_S" in the JSON file must get the value of this header)

Optional:

Column includes the clone frequency, with a header (the parameter "F_S" in the JSON file must get the value of this header).

Notice: the frequency can be represented by 'int' or 'float'. If it is 'float', you must put "False" in "RUN_KDE" and "RUN_MDS"

Column includes the v gene, with a header (the parameter "V_S" in the JSON file must get the value of this header)

Notice: the parameter "DEL_S" in the JSON file must get the value of the delimiter between the columns in the csv file

Output:

After running the code, a folder is created, named "data set name_results". It contains the files which received from the running (projections, accuracy, visualizations). There is a sub-folder for each encoder that was run. The projections are in sub-folder named "projections_to_user"

Change parameters to specific need in file "parameters.json"

In order to adjust the code to the desired data, the following parameters must be changed in JSON file:

- **root**: the path to data set (folder name of your data set)
- **DEL_S**: which delimiter to use to separate headers
- **CDR3_S**: header of amino acid sequence of CDR3
- **V_S**: header of V gene, "None" for data set with no resolved V
- **F_S**: header of clone frequency, "None" for data set with no frequency

The following parameters have default values (but they can be changed):

- **EPOCHS**: number of epochs for each model (default: 100)
- **P_LOAD**: number of data to load for training, float or int (default: 1.0)
- **ENCODING_DIM**: number of dimensions in the embedded space (default: 30)
- **N_PROJECTIONS**: number of projections to use in further analysis (default: 100)
- **N_FOR_DISTANCES**: number of projections to use in the ED and KL (default: 100)
- **"encoder_dir":** the sub-folder that will contain the projections received by the encoder AE (default: "encoder_projections")
- **"v_encoder_dir":** the sub-folder that will contain the projections received by the encoder V_AE (default: " v_encoder_projections ")
- **"embedding_dir" :** the sub-folder that will contain the projections received by the encoder EMB (default: " embedding_projections ")

- **"v_embedding_dir" :** the sub-folder that will contain the projections received by the encoder V_EMB (default: " v_embedding_projections ")
- **"tsne_props_encoder":** the folder of the projection that TSNE and PROPS use. (They are explained in the section "More Tools").
  Must be one from the 4 parameters above (encoder_projections/ v_encoder_projections/ embedding_projections/ v_embedding_projections). Default: "encoder_projections"**.**

Run parameters – change the analysis you desire to run for True, otherwise – False

Auto-Encoders:

- **RUN_AE:** the basic auto-encoder
- **RUN_V_AE:** the auto-encoder with V
- **RUN_EMB:** the auto-encoder with respect to the original distances
- **RUN_V_EMB:** the auto-encoder with respect to the original distances + V representation

Notice: if your data set does not contain genes (v), put "False" on the following auto-encoders: RUN_V_AE, RUN_V_EMB

Default Auto-Encoders: (Use existing auto-encoder. Recommended in case of minimal data, where there is not enough for ML to run)

- **RUN_AE_DEFAULT**
- **RUN_EMB_DEFAULT**

More Tools:

- **RUN_SAVE**: save all projections per sample for each of the trained models
- **RUN_TSNE**: visualize all samples using TSNE
- **RUN_KDE**: run KDE to find all pairwise distances between samples. Run for each of the trained models twice, with and without frequencies
- **RUN_PROPS**: check repertoire features within the embedded space
- **RUN_KL**: KL distances

- **RUN_ED**: ED distances
- **RUN_MDS**: run MDS for each KDE matrix

Notice: if you put less than 3 sequences in the data set, you need to change the following parameters to False: RUN_TSNE, RUN_KDE, RUN_PROPS, RUN_KL, RUN_ED, RUN_MDS. You will get the projections only.