Supplementary material

List of intermediate files that can be found at this link https://github.com/LauraVerdesca/LB1_report:

- **best_mcc.py**, this python scrpit takes in input a file containing the evaluation of the performance and returns the best e-value threshold and plots the MCC values corresponding to each threshold
- **bpti_pos_selected.fasta,** this fasta file contains the sequences of the positive dataset used for testing and training
- **clean_kunitz_3d.aln,** the output alignement file of PDBe fold
- **clean_kunitz_3d.hmm,** the ouput file of the HMM model
- **com_set_0123.res**, this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) computed for each threshold in the range 1e-1 - 1e-15 in the training phase of 5-fold cross validation of combined subsets 0, 1, 2, 3.
- **com_set_0124.res,** this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) computed for each threshold in the range 1e-1 - 1e-15 in the training phase of 5-fold cross validation of combined subsets 0, 1, 2, 4.
- **com_set_0134.res,** this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) computed for each threshold in the range 1e-1 - 1e-15 in the training phase of 5-fold cross validation of combined subsets 0, 1, 3, 4.
- **com_set_0234.res,** this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) computed for each threshold in the range 1e-1 - 1e-15 in the training phase of 5-fold cross validation of combined subsets 0, 2, 3, 4.
- **com_set_1234.res,** this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) computed for each threshold in the range 1e-1 - 1e-15 in the training phase of 5-fold cross validation of combined subsets 1, 2, 3, 4.
- **cross_val.sh,** this bash script contains the pipeline to compute 5-fold cross validation
- **out_best,** this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) applied to each testing subset with the best e-value threshold obtained from the training subset, meaning the threshold that gave the best values in the com_set_*.res files.
- **Performance.py,** this python script computes the performance metrics (confusion matrix, accuracy, MCC, F1).
- **rcsb_pdb_custom_report_20240414163443.csv,** this file contains the PDB ids, sequences and chain identifier of each protein structure we selected to build the HMM model
- **select_fasta.py,** the python script reads a sequence file and an ID file as input arguments, and then it prints the sequences corresponding to the IDs listed in the ID file.
- **set_1.res,** this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) computed for each threshold in the range 1e-1 - 1e-15 in the training phase of 2-fold cross validation of subset 1.
- **set_2.res,** this file contains the performance metrics (confusion matrix, accuracy, MCC, F1) computed for each threshold in the range 1e-1 - 1e-15 in the training phase of 2-fold cross validation of subset 2.

5-cross validation

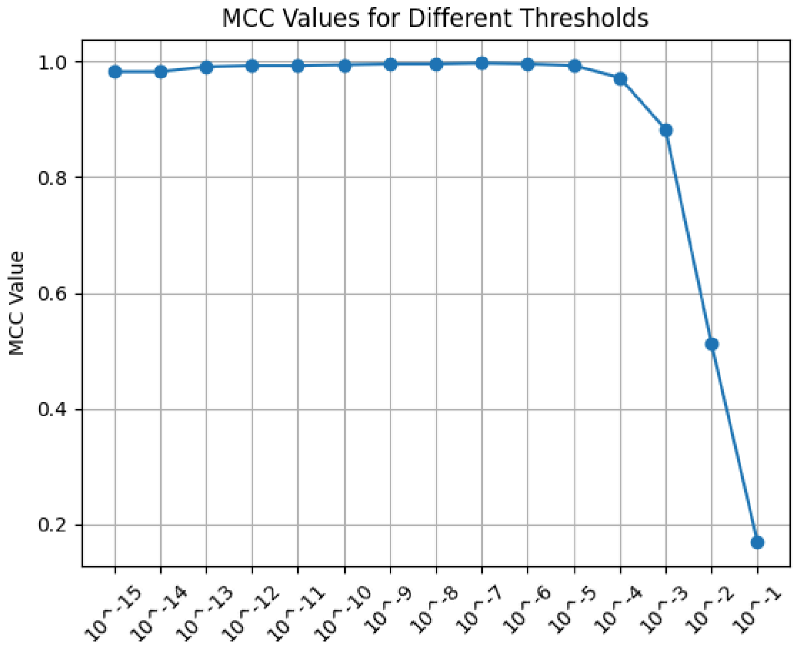

Figure 1 - com_set_0123, this plot shows the levels of MCC values corresponding to each threshold we tested

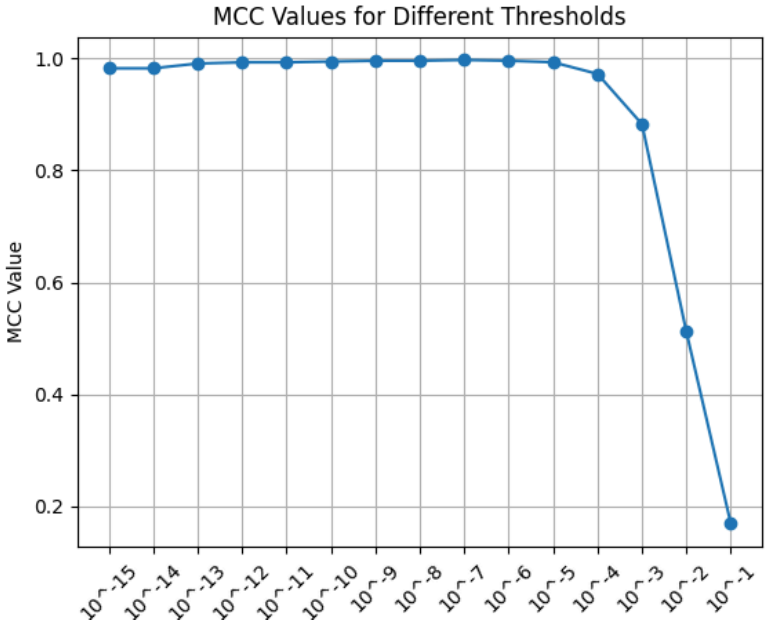

Figure 2 - com_set_0124, this plot shows the levels of MCC values corresponding to each threshold we tested

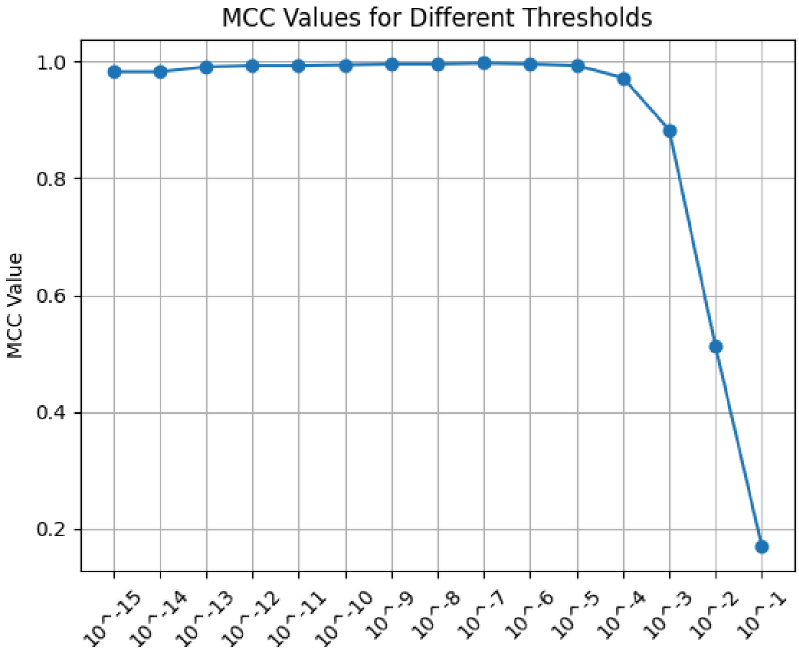

Figure 3 - com_set_0134, this plot shows the levels of MCC values corresponding to each threshold we tested

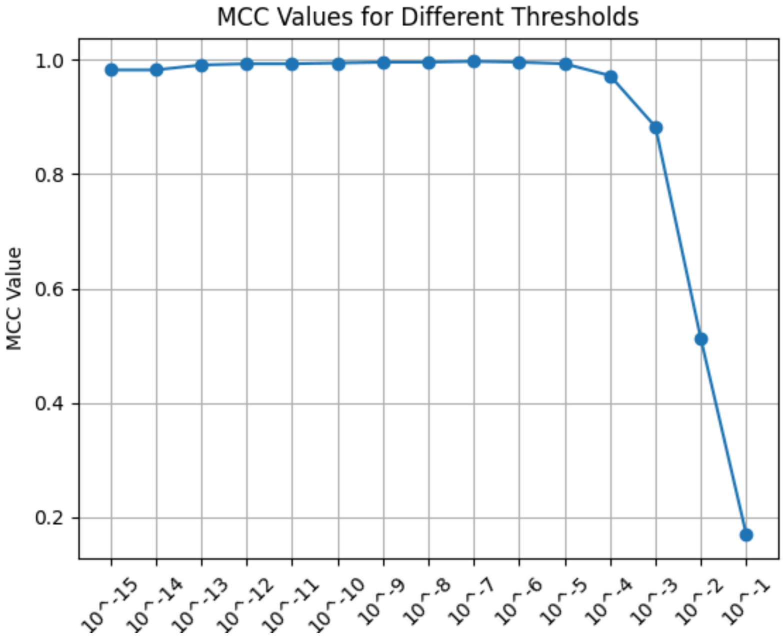

Figure 4 - com_set_0234, this plot shows the levels of MCC values corresponding to each threshold we tested

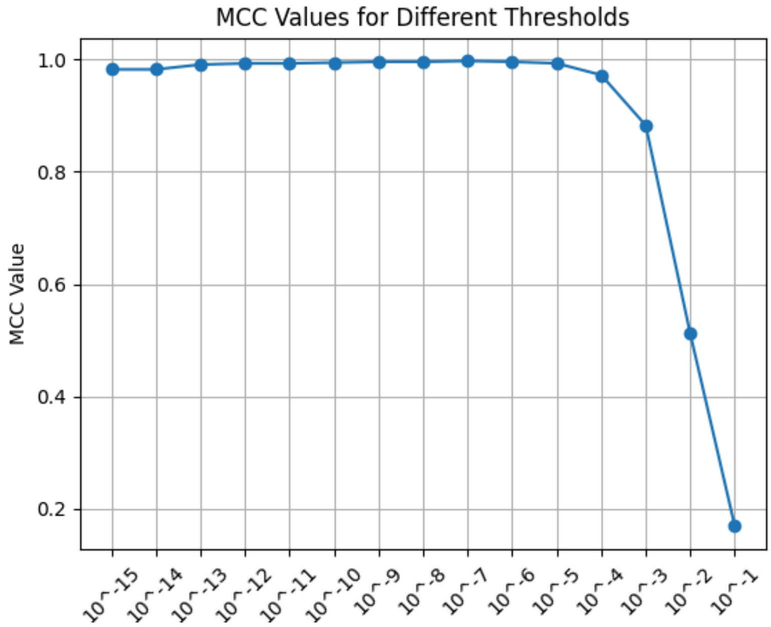

Figure 5 - com_set_1234, this plot shows the levels of MCC values corresponding to each threshold we tested

2-cross validation

	Threshold	Accuracy	MCC
Training set	1e-06	0.99998	0.99171
	1e-07	0.99999	0.99440
	1e-08	0.99999	0.99579
	1e-09	0.99998	0.99579
Testing set	1e-07	0.99999	0.9972

Table 1 - 2-fold cross validation, training on subset 1 and testing on subset 2

	Threshold	Accuracy	MCC
Training set	1e-06	0.99998	0.99171
	1e-07	0.99999	0.9973
	1e-08	0.99998	0.99579
	1e-09	0.99998	0.99579
Testing set	1e-07	0.99999	0.9972

Table 2 - 2-fold cross validation, training on subset 2 and testing on subset 1

	Positives	Negatives

Predicted positives	TP = 357	FP = 1
Predicted negatives	FN = 2	TN =570889

Table 3 - Confusion matrix of the two sets combined when the 1e-7 threshold is applied