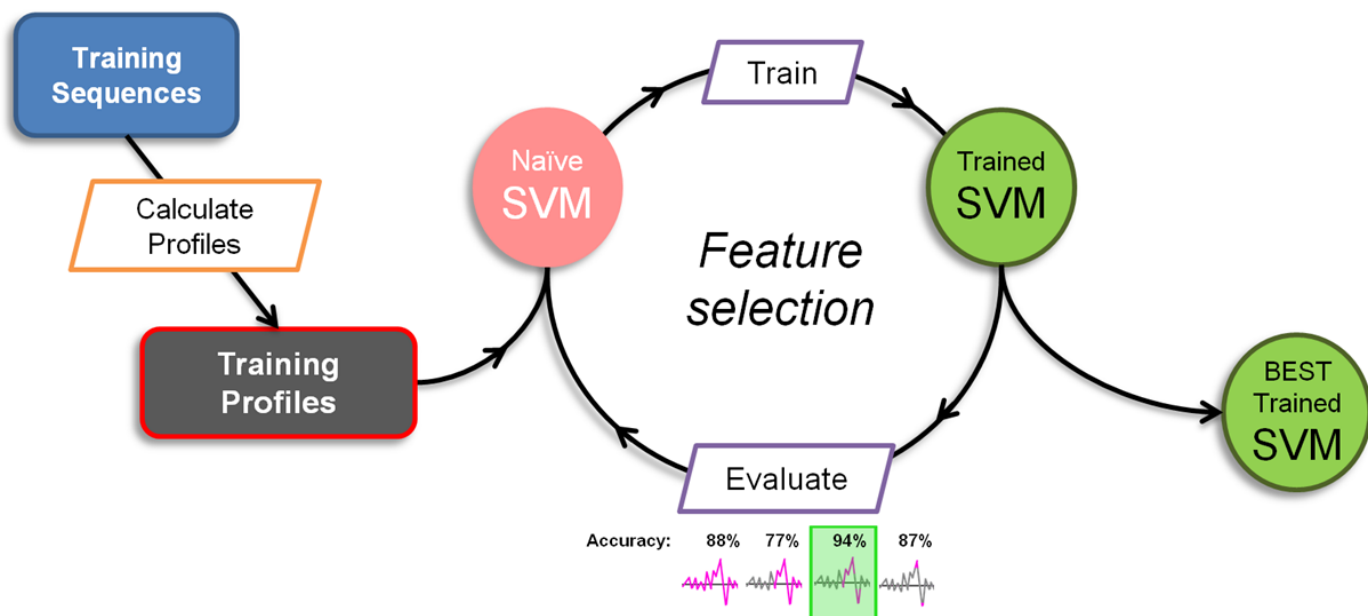
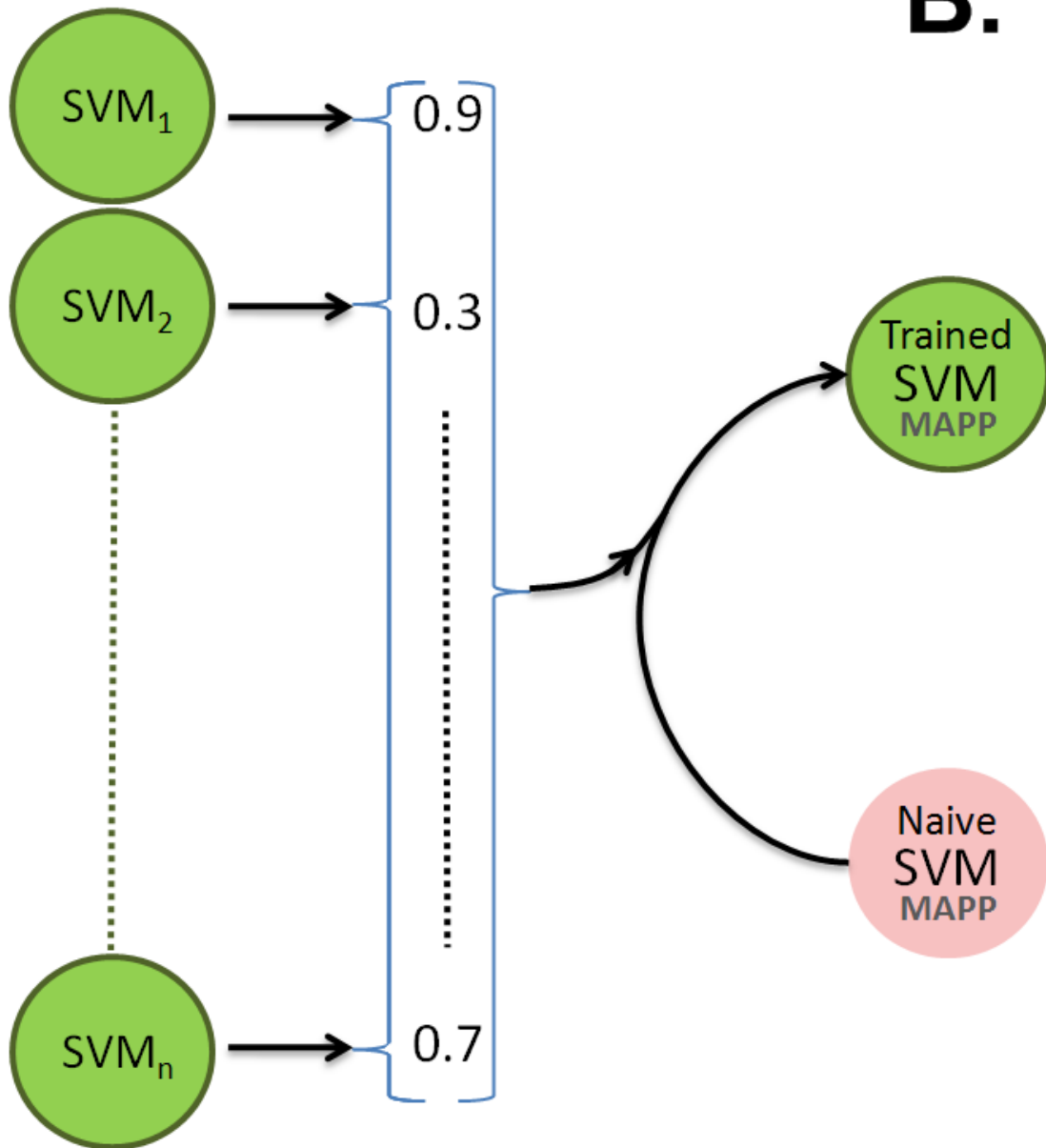


A.



A. The training cycle for each of the single property SVMs.

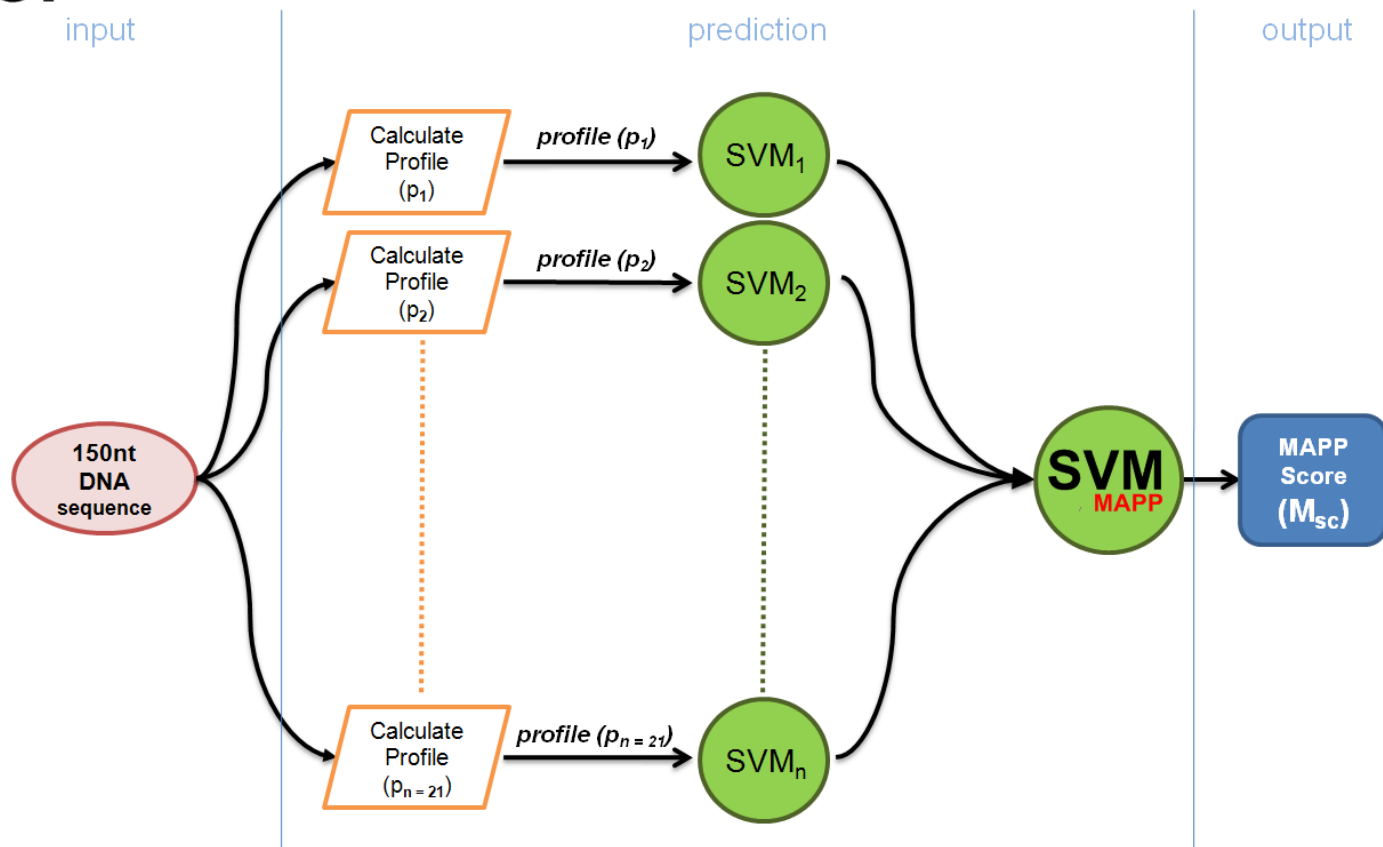
Each DNA sequence is converted into a numerical physicochemical property profile and this dataset of profiles is used to train an SVM. Feature selection is an iterative cycle which trains several SVMs with different combinations of features. The accuracy of the resulting SVMs are assessed and that with the best 5-fold cross validation accuracy is used (in this example, highlighted in green). Such an SVM is trained for each of the 33 non-redundant physicochemical properties.



B. Training the combinatorial SVM

The outputs of each of the n trained SVMs are combined into a single vector which is used to as the input for the final SVM (SVM_{MAPP}). Feature selection is not used for the 5-fold cross validation training process. The output is a likelihood score between 0 and 1 that the nucleotide at position 100 of the input sequence is a TSS.

C.



C. **MAPP prediction schema.**

This is a schema of how MAPP makes a TSS prediction for a given sequence. A bare 150 nt sequence is used as input. This sequence is converted into 21 physicochemical property profiles. Each profile is passed through a trained SVM (see A) for that property. The outputs of these SVMs are then combined and passed through the final SVM (SVM_{MAPP}). The output is a likelihood score that the nucleotide at position 100 of this sequence is a TSS. To generate MAPP predictions for a longer sequence, this procedure is repeated for each 150nt subsequence. Consequently, MAPP scores for the first 100 nt and the final 50 nt of a given sequence cannot be predictions.