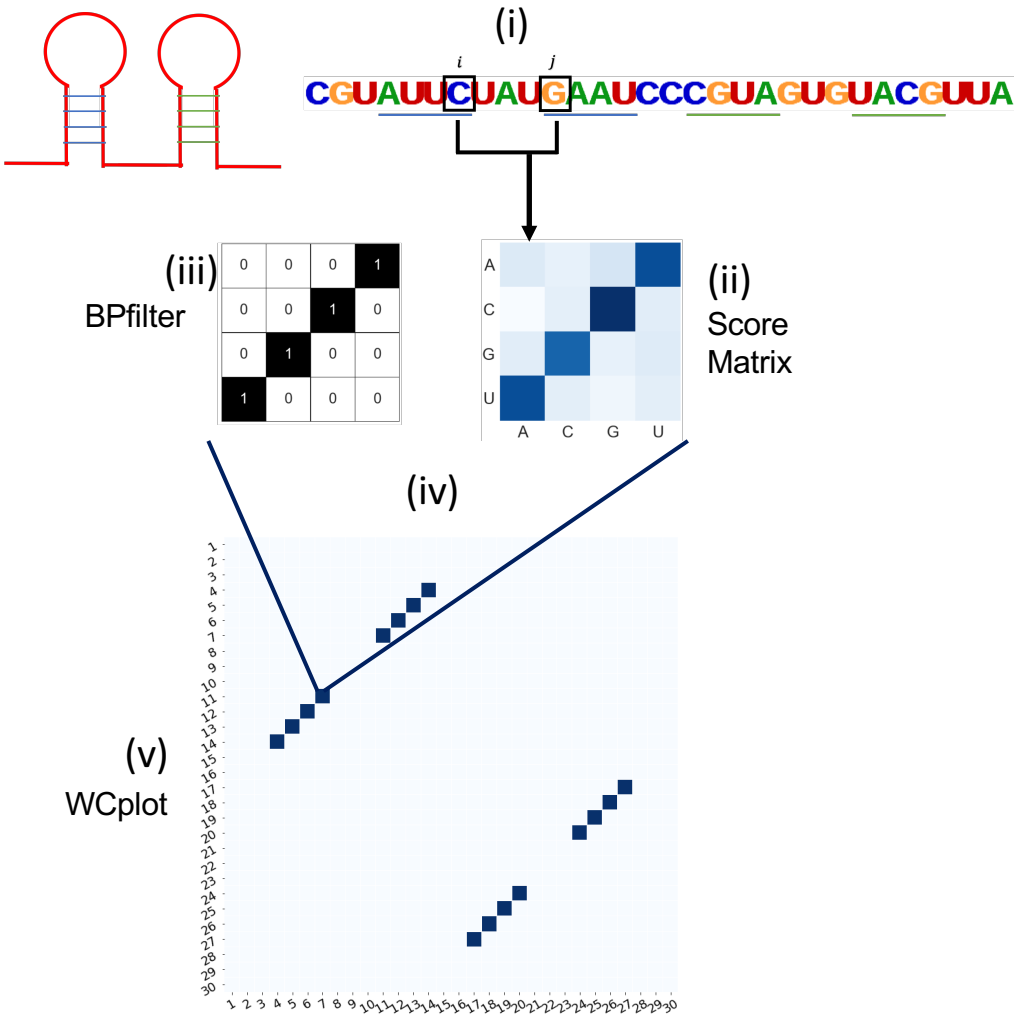

Figure 4: The schematic for second order *in silico* mutagenesis (SoM). (i) For a given sequence of length *L*, we perform saturated pairwise mutagenesis, mutating every possible pair of sites to all pairs of nucleotides. (ii) For a given pair of sites, we produce 15 mutant sequences in addition to the WT and get the scores of each sequence from the NN. The off diagonal of the score matrix corresponds to the complementary nucleotide pairings for watson crick base pairs. If these sites are base paired, we'd expect the NN to output high scores for the sequences with complementary mutations that maintain watson crick base pairs. (iii) To reduce the dimensionality of each score matrix, we employ a BPfilter. (iv) We perform element-wise multiplication and sum the values to retrieve a single SoM score. (v) We denoise the SoM scores by averaging SoM scores across multiple sequences and perform an APC correction. The SoM scores are then plotted according to the positions in the original sequence that were mutated with the colour showing the gradient of SoM scores with darker blue showing a higher SoM score.

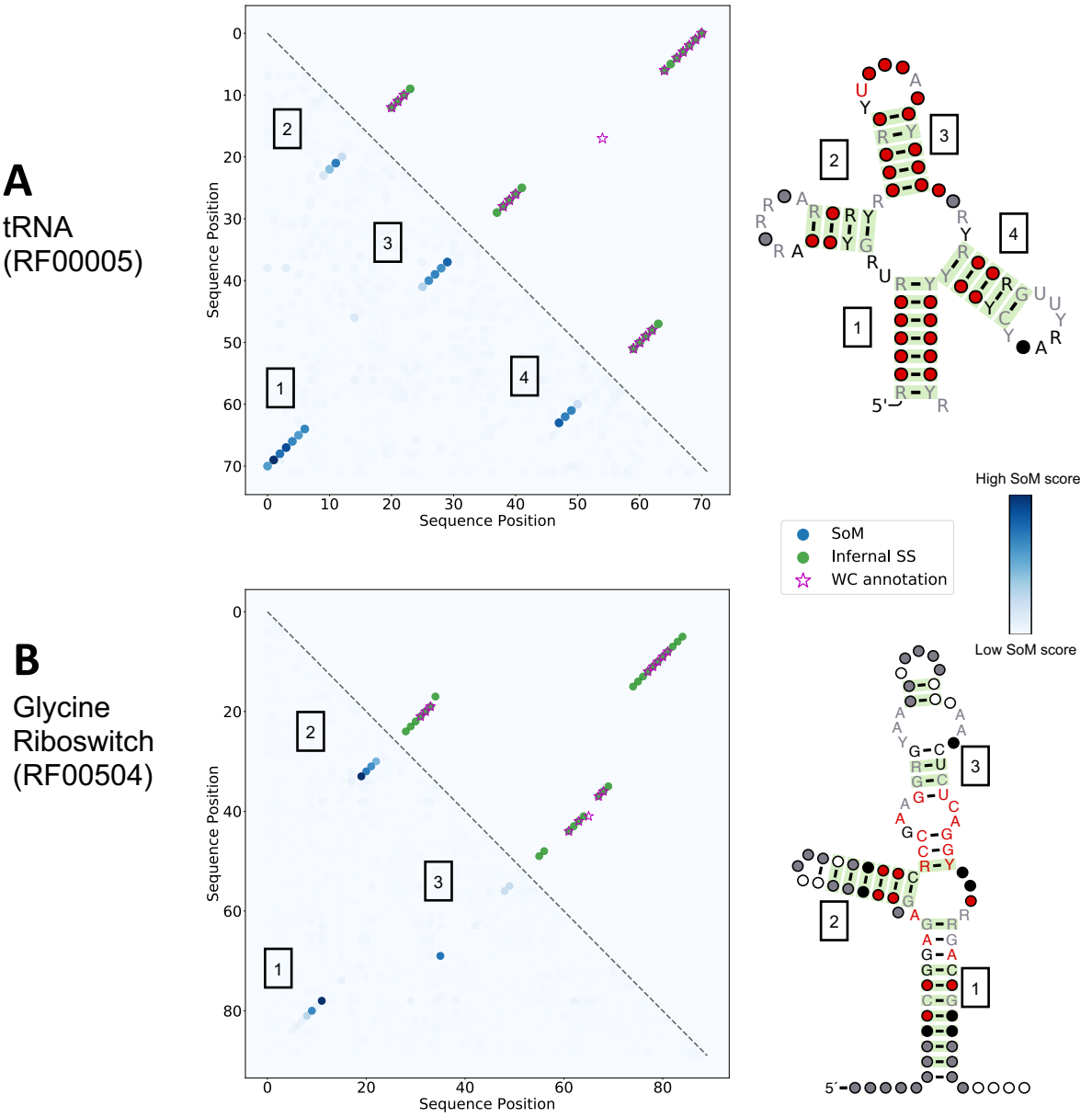

Figure 5: SoM scores (blue) from the MLP trained on (A) tRNA – RF00005 and (B) Glycine Riboswitch – RF00504, compared to the secondary structure (SS) predicted by Infernal (green) and the WC annotation from respective PDB structures, extracted using FR3D (purple stars). Folded structures generated by R2R (Weinberg and Breaker, 2011) using the Infernal SS prediction are next to each plot. Base pairs with statistically significant covariation are shaded in green. Red, black, grey and white dots show nucleotides that are 97%, 90%, 75% and 50% conserved respectively. R and Y correspond to purines (A,G) and pyrimidines (C,U) respectively. The plots show that the MLP is learning covariation corresponding to true base pairs. We expected SoM to mainly elucidate complementary watson crick pairs, but high scores are found where there aren't WC annotations, but are still true covarying pairs as found by Infernal.

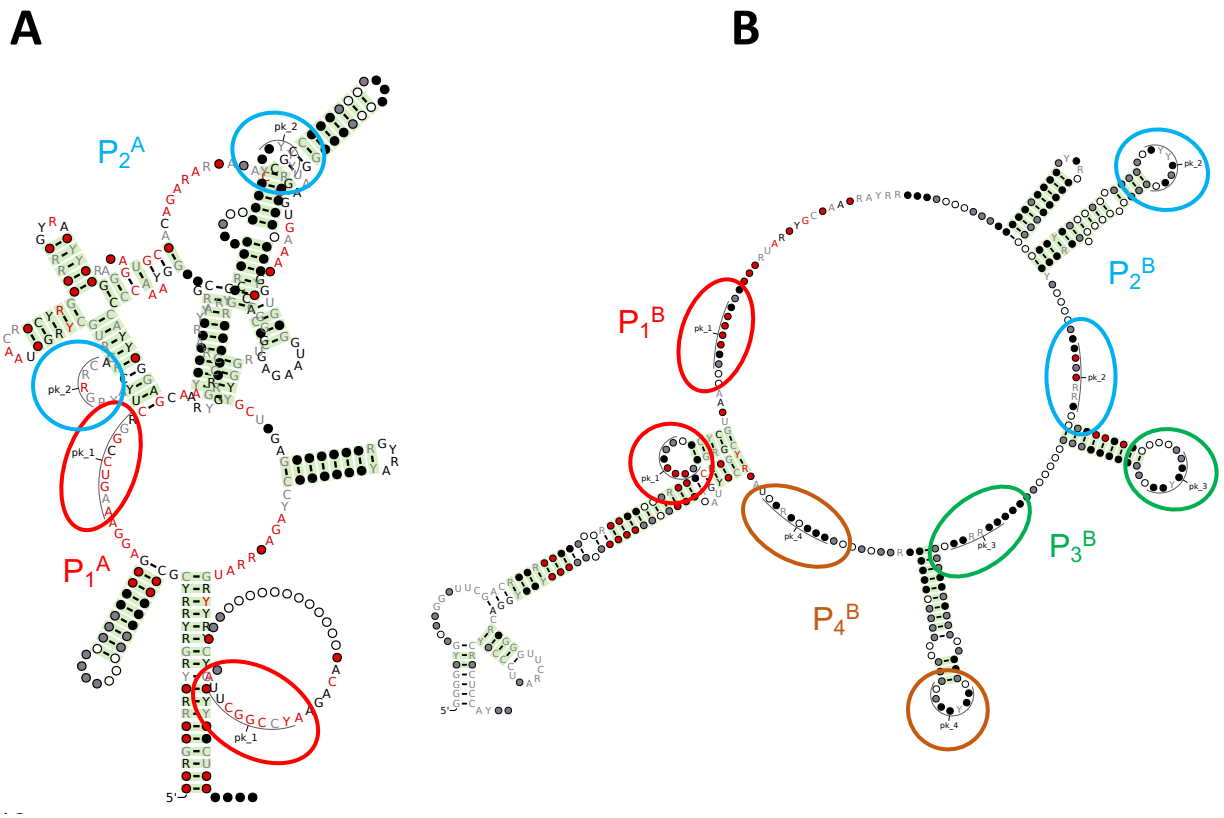

Figure 6: Secondary structures for two pseudoknot containing families: (A) Bacterial RNase P class A – RF00010 and (B) transfer-messenger RNA (RF00023), taken from Rfam. Structures show the folded nested secondary structure, with the regions in the sequence with non-nested pseudoknots annotated by colored rings and labelled P_x^F where *F* refers which to which family, and *x* refers to the pseudoknot number. Folded structure annotations follow same key as Figure 5.

Figure 6 continued: comparing the SoM scores of the trained MLPs with WC annotations and Infernal SS predictions for (C) Bacterial RNase P and (D) transfer-messenger RNA. Plots on the left show the SoM scores (blue) compared against WC annnotations (purple), and plots on the right show Infernal's SS prediction (green) compared to WC annotations. The WC annotations do contain the known pseudoknotted base pairs. Grey boxes are drawn around the pseudoknots with labels corresponding to the colored annotations in A and B. Infernal cannot model pseudoknots, and thus it doesn't have points in the grey boxes. The MLP does learn the pseudo knotted base pairs of both families and shows high SoM scores for the pseudo knotted base pairs that are found in the WC annotations.

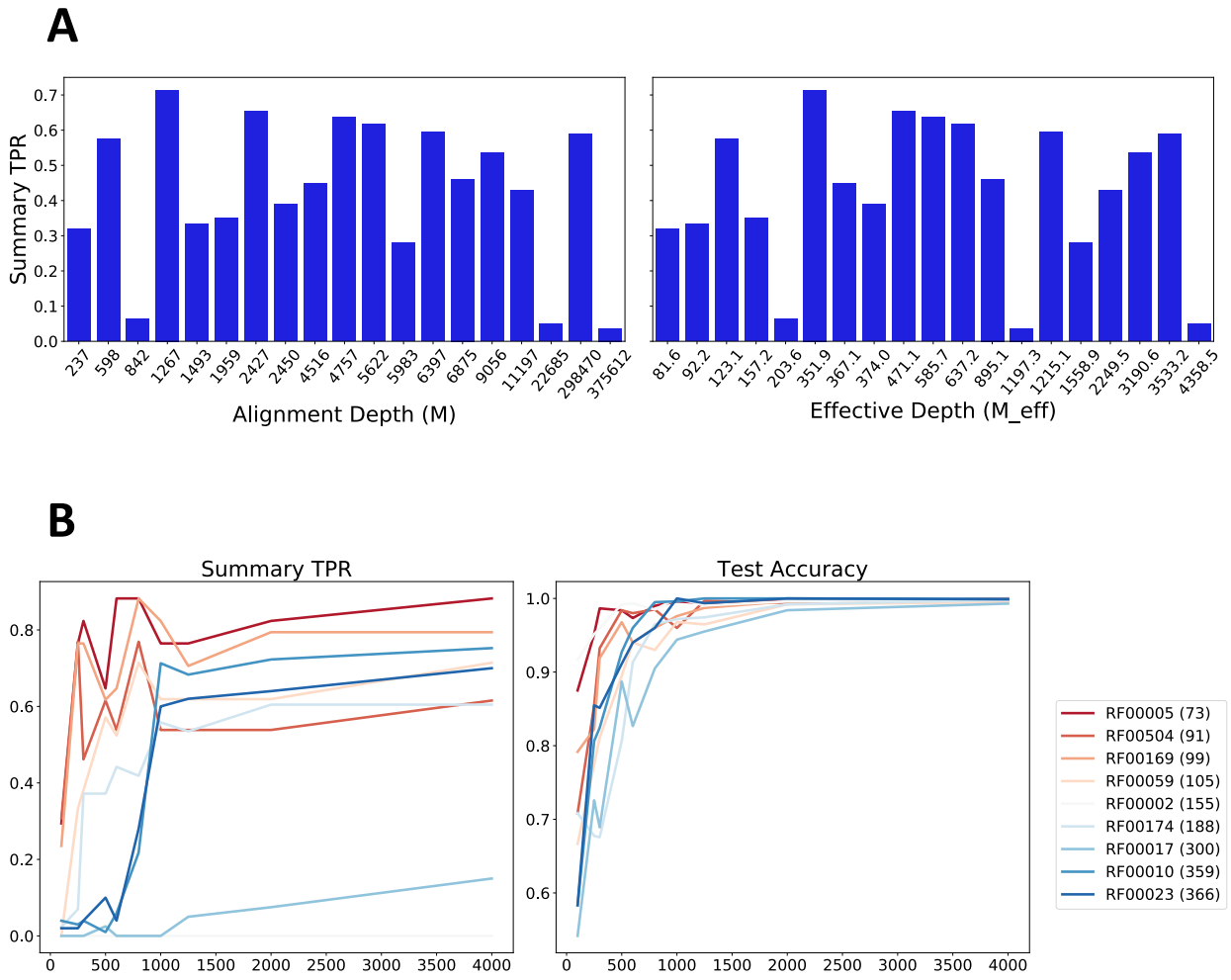

Figure 9: (A) the TPR was calculated for each trained MLP where the top 0.7C SoM scores were called positive contacts. These were plotted against the alignment depth (M) and effective depth (M_{eff}) in ascending order. There is no noticeable trend between MLP performance and alignment depth across all the families. (B) For a given family, the depth of the training alignment was systematically reduced, with an individual MLP trained on each training depth, thus controlling for other confounding factors in the alignment. The summary TPR and the model accuracy (in classifying the test set) of each trained MLP is plotted against the training set depth for 9 different families. This shows a general decreasing trend in classification and structure learning performance as the model has fewer training sequences. Each line corresponds to a different RNA family, and the lines are coloured by the length of the MSA of that family. The MSA length is bracketed next to the Rfam ID in the legend.

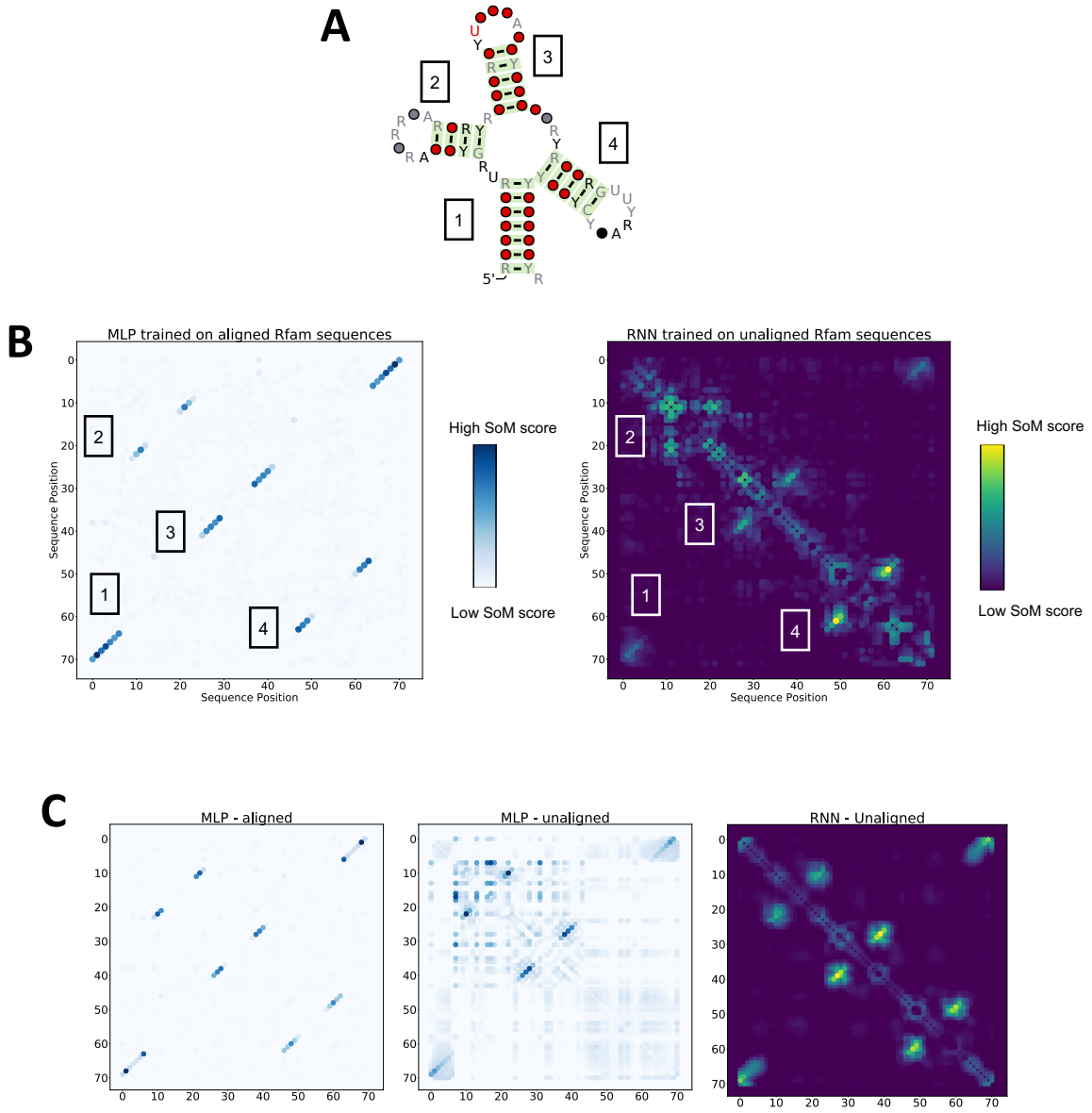

Figure 10: (A) Folded secondary structure of tRNA – RF00005 from Rfam. (B) Results from SoM performed on an MLP trained on the tRNA reduced Rfam alignment (blue) and from an RNN trained on the unaligned Rfam sequences (blue-green). The RNN seems to be learning general regions corresponding to true base pairs in the structure, but the results are visibly much noisier than the MLP trained on the alignment. (B) Results from training the model on a simulated alignment with 100,000 sequences generated by Infernal. The MLP trained on the simulated aligned sequences shows very clear base pair learning, however when training on unaligned sequences the results are much noisier. The RNN trained on unaligned data performs better than the MLP on unaligned data, learning the regions of base pairing from each stem in the structure. The ticks on each WCplot correspond to the sequence position.