Rfam ID	Name	Length	M	M_{eff}	Accuracy	AUC-roc
RF00002	5.8S ribosomal RNA	155	375612	1197.3	1.0	1.0
RF00005	tRNA	73	298470	3533.2	0.999	1.0
RF00010	Bacterial RNase P class A	359	6397	1215.1	0.999	1.0
RF00017	Metazoan signal recognition particle RNA	300	22685	4358.5	0.998	1.0
RF00023	transfer-messenger RNA	366	5983	1558.9	0.999	1.0
RF00050	FMN riboswitch (RFN element)	135	4516	367.1	0.998	1.0
RF00059	TPP riboswitch (THI element)	105	11197	2249.5	0.999	1.0
RF00162	SAM riboswitch (S box leader)	108	4757	585.7	0.995	1.0
RF00167	Purine riboswitch	103	2427	471.1	0.996	1.0
RF00169	Bacterial small signal recognition particle RNA	99	5622	637.2	0.996	1.0
RF00174	Cobalamin riboswitch	188	9056	3190.6	0.997	1.0
RF00234	glmS glucosamine-6-phosphate activated ribozyme	161	842	203.6	1.0	1.0
RF00380	ykoK leader	170	1493	92.2	0.997	1.0
RF00504	Glycine riboswitch	91	6875	895.1	1.0	1.0
RF01734	crcB RNA	65	1267	351.9	0.933	0.992
RF01786	Cyclic di-GMP-II riboswitch	86	237	81.6	0.896	0.95
RF01831	THF riboswitch	100	598	123.1	1.0	1.0
RF01852	Selenocysteine transfer RNA	91	1959	157.2	0.997	1.0
RF02001	Group II catalytic intron D1-D4-3	174	2450	374.0	0.996	1.0

Table 1: the 20 Rfam alignments used for this work with corresponding Length, Depth (M) and Effective Depth (M_{eff}) in addition to the accuracy and AUC-ROC scores of the MLP trained on the alignment. The length corresponds to the number of nucleotides (including gaps) in a sequence after reduction. The depth is the number of total sequences in the alignment. When alignments contain many sequences that are highly related phylogenetically there is less structural information in the alignment than would be suggested by its depth. Thus the effective depth of an alignment corresponds to a weighted depth proportional to how related the sequences are, providing a better indicator of how much structural information the MLP sees during training.

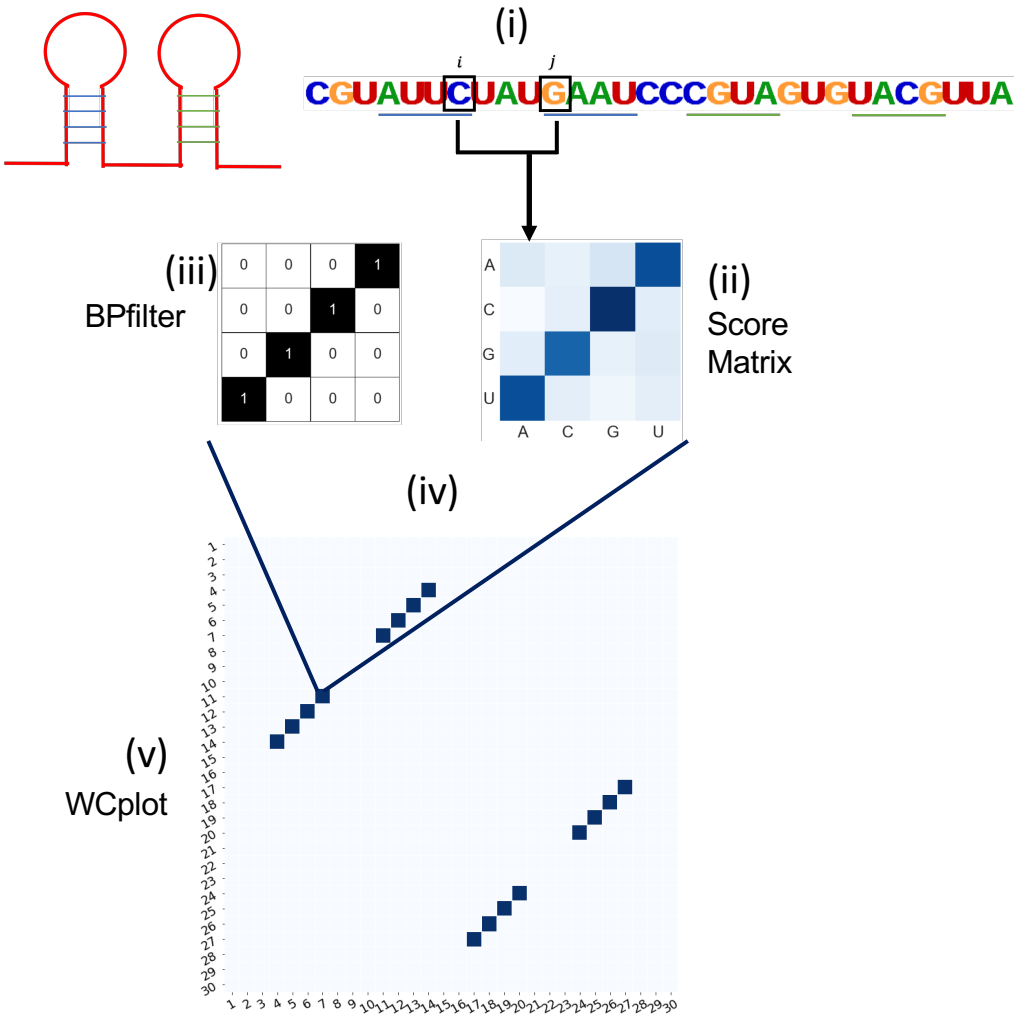

Figure 4: The schematic for second order *in silico* mutagenesis (SoM). (i) For a given sequence of length *L*, we perform saturated pairwise mutagenesis, mutating every possible pair of sites to all pairs of nucleotides. (ii) For a given pair of sites, we produce 15 mutant sequences in addition to the 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.

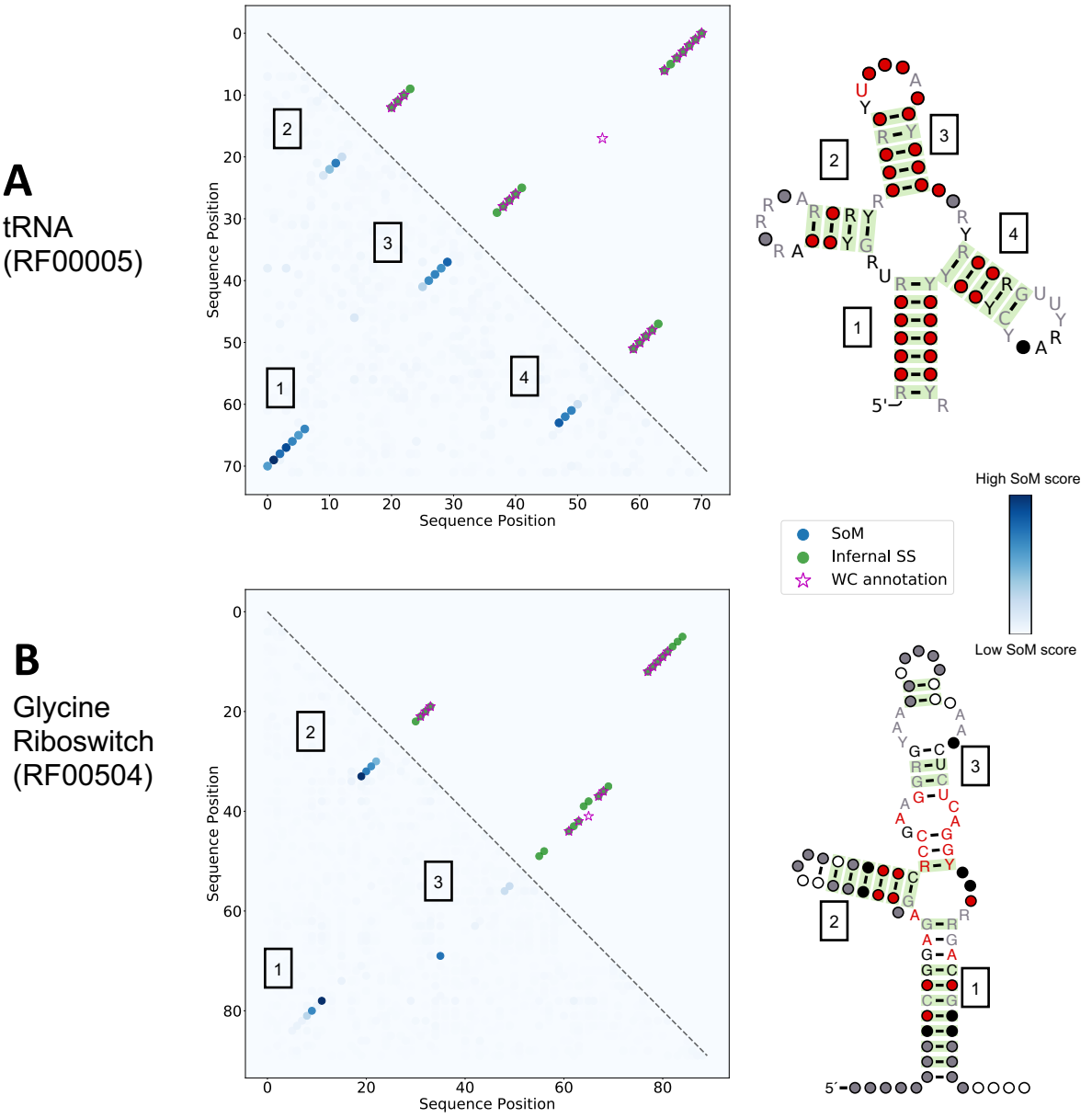

Figure 5: SoM scores (blue) from the MLP trained on (A) tRNA – RF00005 and (B) Glycine Riboswitch – RF00504, compared to the secondary structure (SS) predicted by Infernal (green) and the WC annotation from respective PDB structures, extracted using FR3D (purple stars). Folded structures generated by 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.

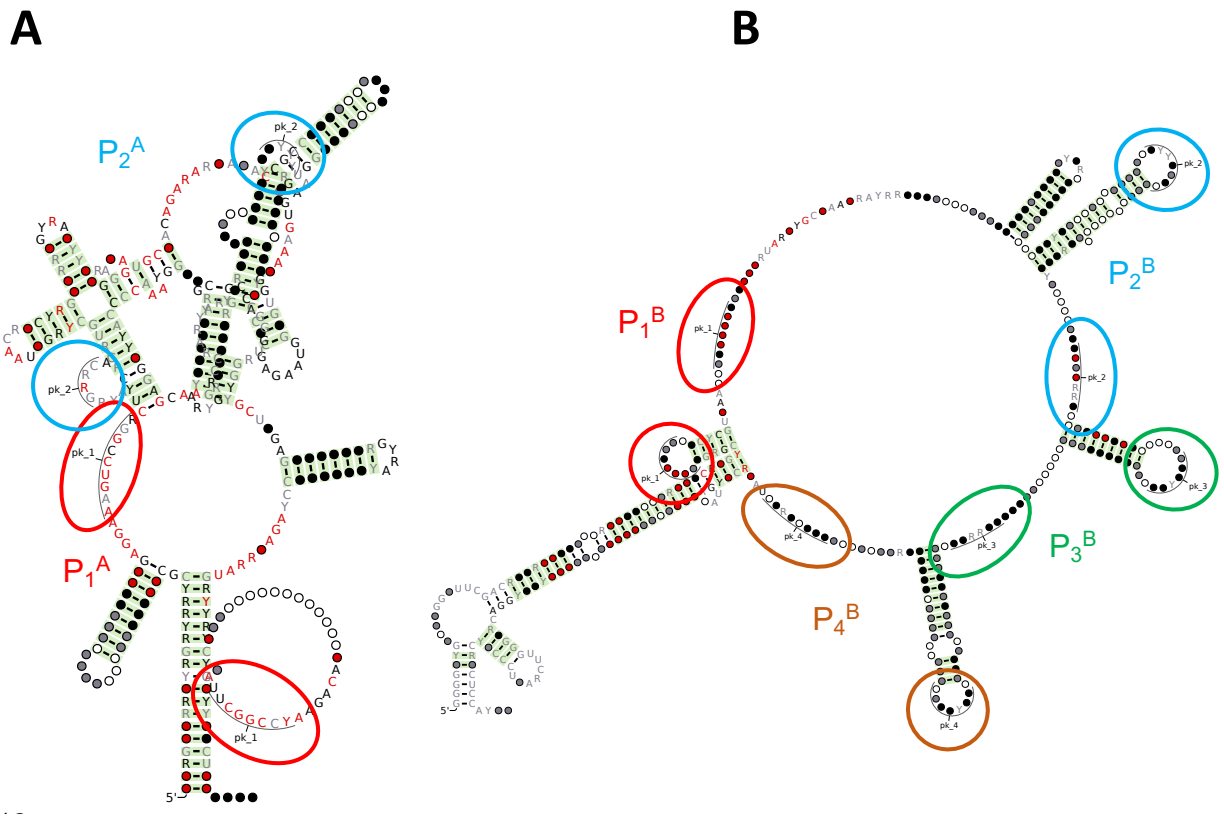

Figure 6: Secondary structures for two pseudoknot containing families: (A) Bacterial RNase P class A – RF00010 and (B) transfer-messenger RNA (RF00023), taken from Rfam. Structures show the folded nested secondary structure, with the regions in the sequence with non-nested pseudoknots annotated by colored 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.

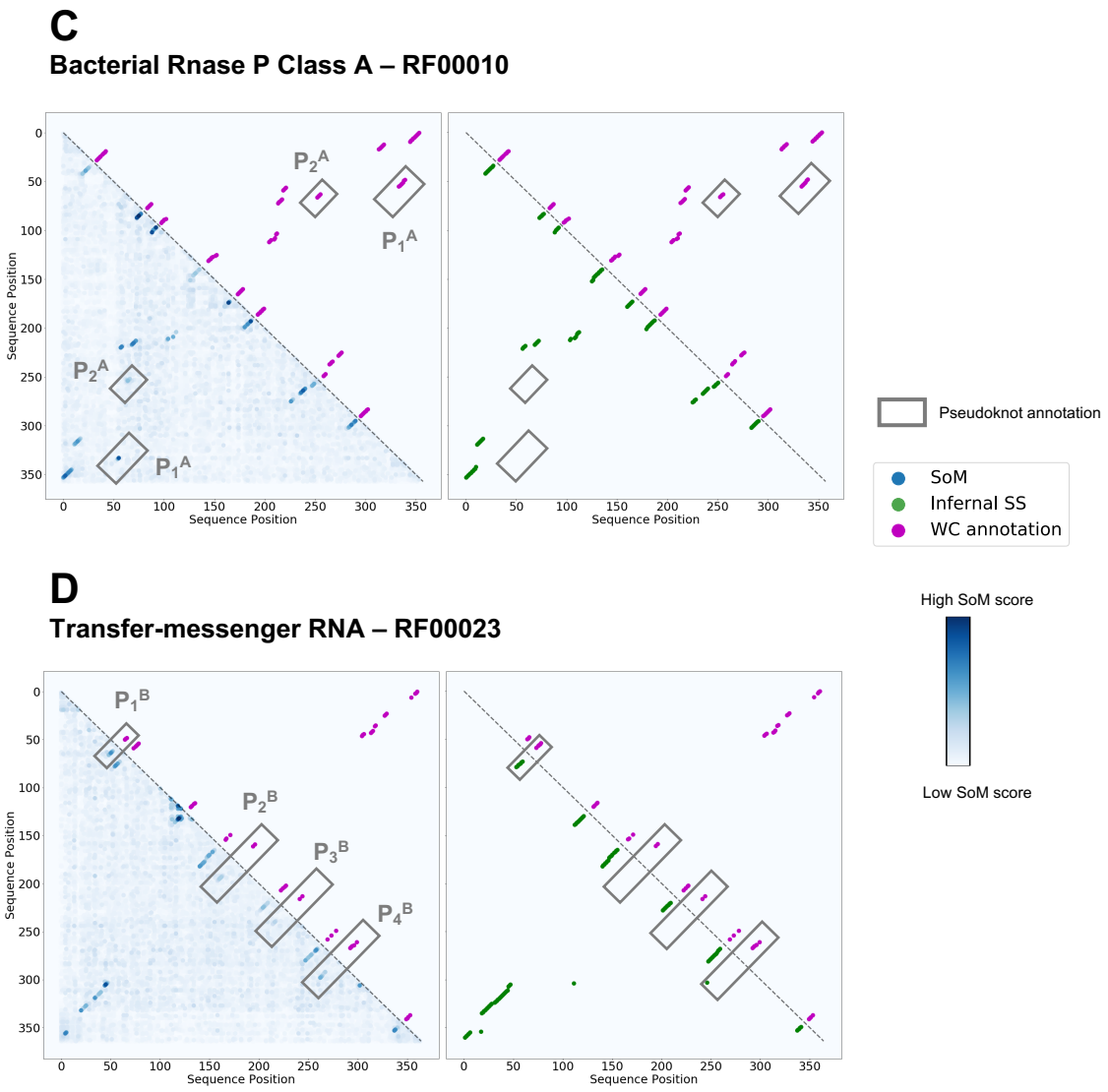

Figure 6 continued: comparing the SoM scores of the trained MLPs with WC annotations and Infernal SS predictions for (C) Bacterial RNase P and (D) transfer-messenger RNA. Plots on the left show the SoM scores (blue) compared against WC annnotations (purple), and plots on the right show Infernal's SS 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.

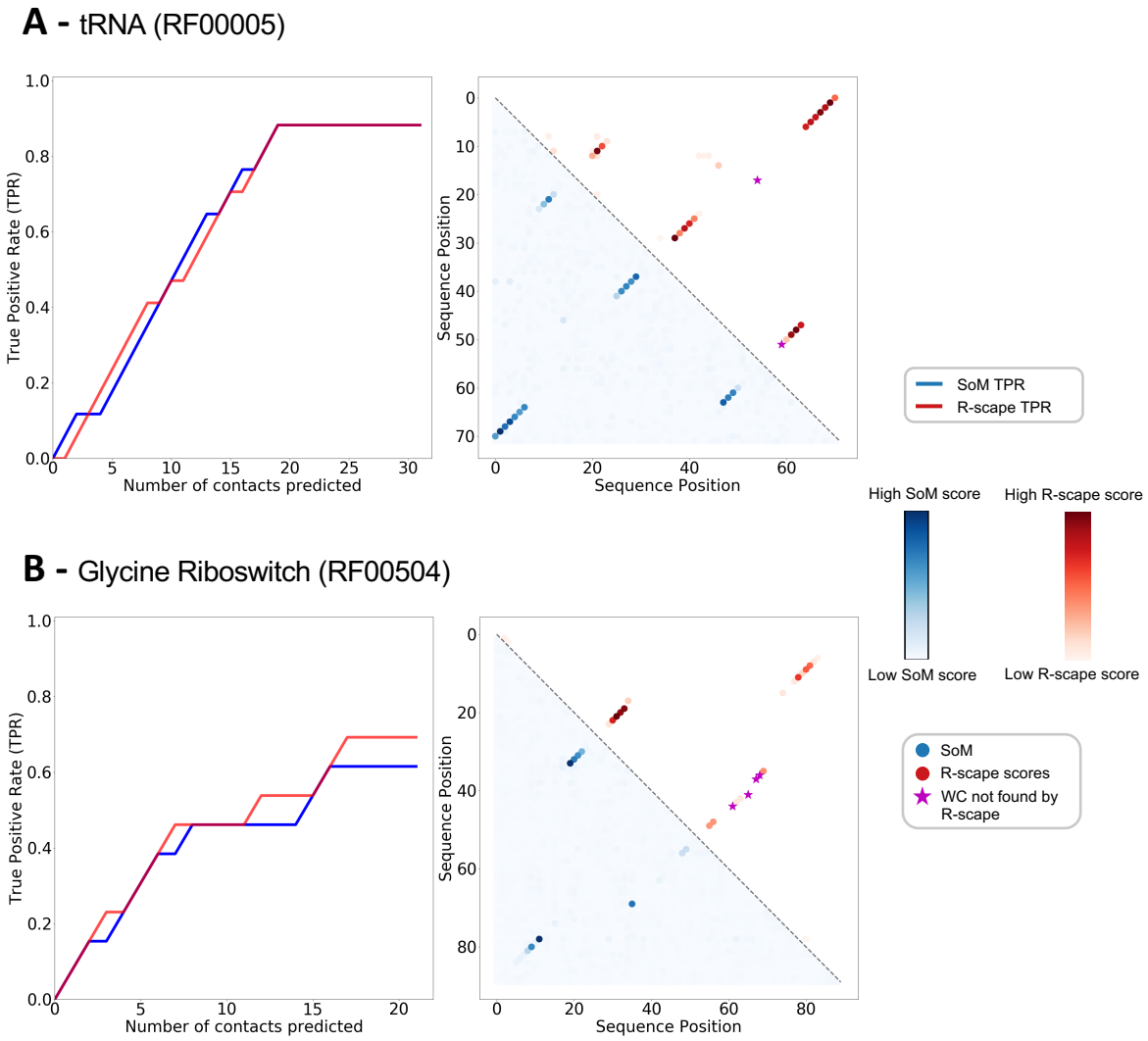

Figure 7: (continued on next page)

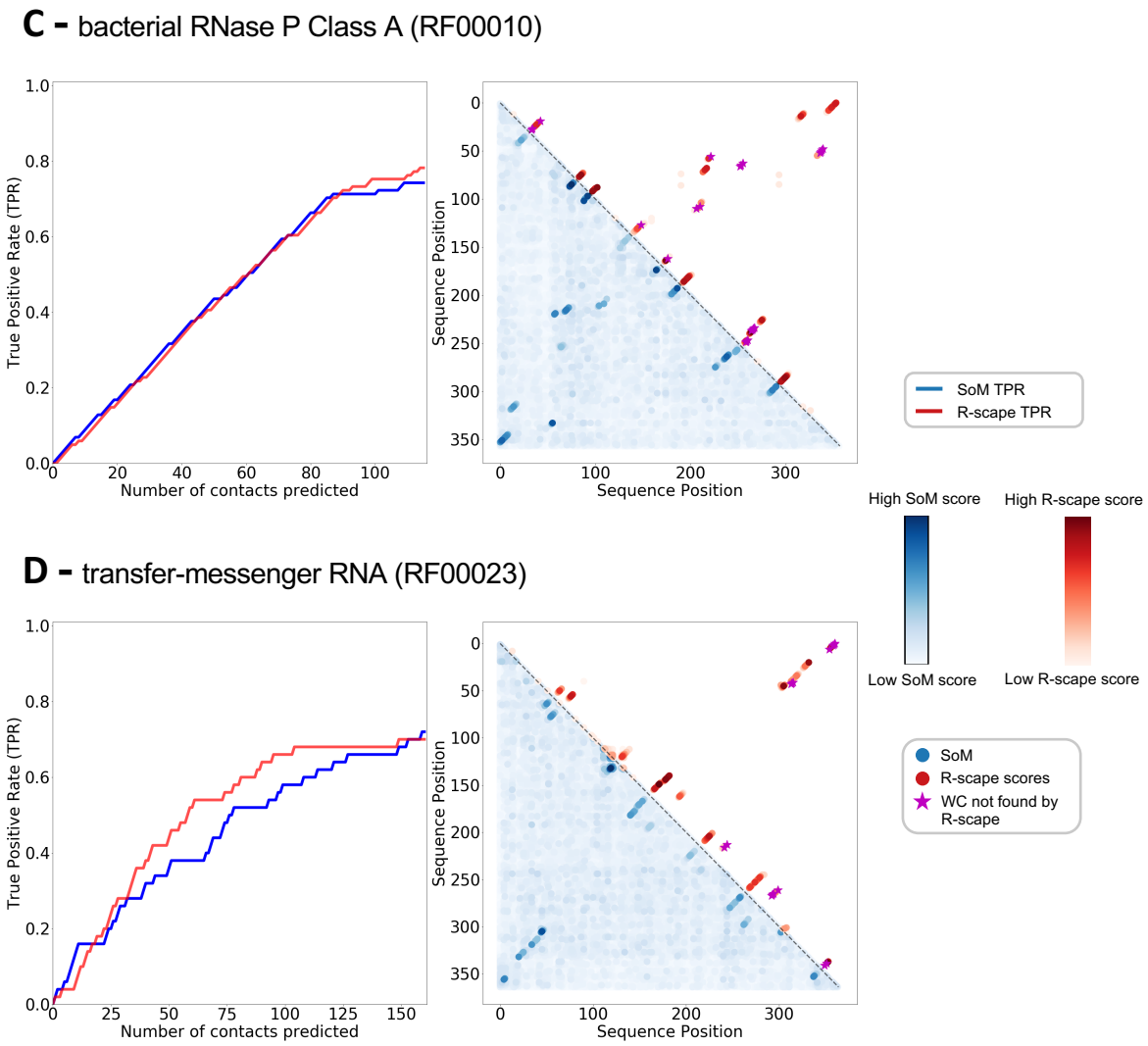

Figure 7: (Left) Graphs showing the ranked TPR the top contacts predicted and (right) SoM scores (blue) compared to R-scape scores (red) with WC annotated base pairs that R-scape does not find also plotted (purple stars). Results are plotted for (A) tRNA – RF00005, (B) Glycine Riboswitch - RF00504, (C) Bacterial RNase P Class A – RF00010 and (D) transfer-messenger RNA (tmRNA) - RF00023. Ranked TPRs are plotted for the top C SoM scores and R-scape scores where C is the total number of base pairs predicted by R-scape. The TPR is a metric of how sensitive the model is to finding WC annotated base pairs. For tRNA and RNase P, the MLP seems to have similar performance to R-scape, learning a similar amount of WC base pairs. For Glycine Riboswitch and tmRNA, R-scape outperforms the MLP slightly.

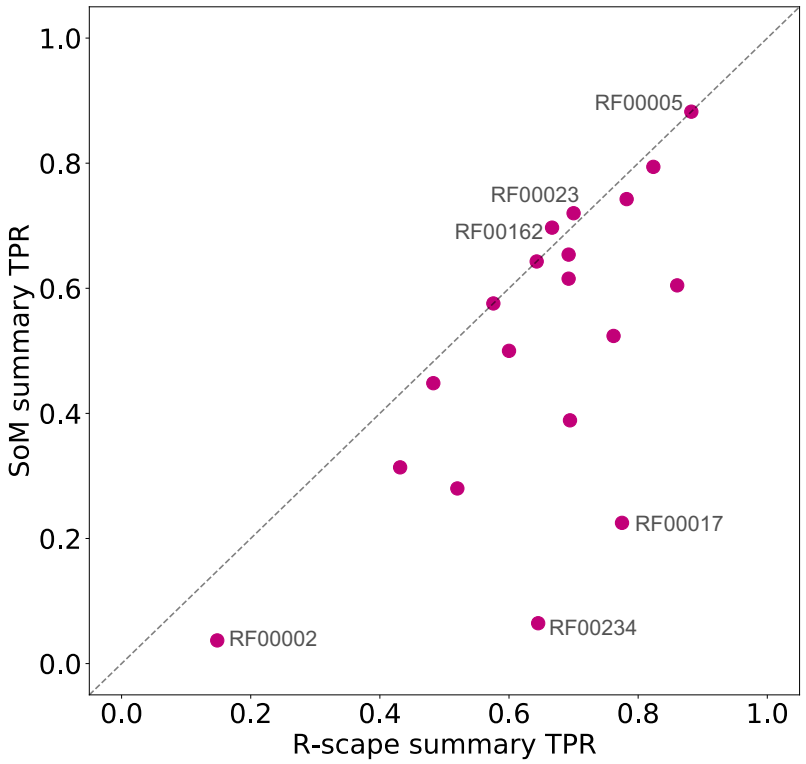

Figure 8: The TPR for the MLPs trained on the Rfam families against the TPR for R-scape shows how the MLP and R-scape's performance compares across 19 of the families used (SoM was performed on only 19 of the 20 families we trained MLPs for; the alignment for Eukaryotic small subunit ribosomal RNA (RF01960) was too large to perform SoM given our computational recourses). Here we plot the TPR for the top C base pairs found by each model, where C is the number of base pairs found by R-scape (i.e. TPR for R-scape using all significant base pairs). Comparing performance across all families shows R-scape outperforms the MLP for all but two of the families (SAM riboswitch (S box leader) – RF0162 and tmRNA - RF00023) for which the MLP has a slightly higher TPR. Shows that R-scape is a much more robust and powerful model than our MLP in its current architecture and training regime. R-scape outperforms the MLP most significantly for glmS activated ribozyme (RF00234) and Metazoan SRP RNA (RF00017). The lowest score for both models was on 5.8S ribosomal RNA (RF00002).

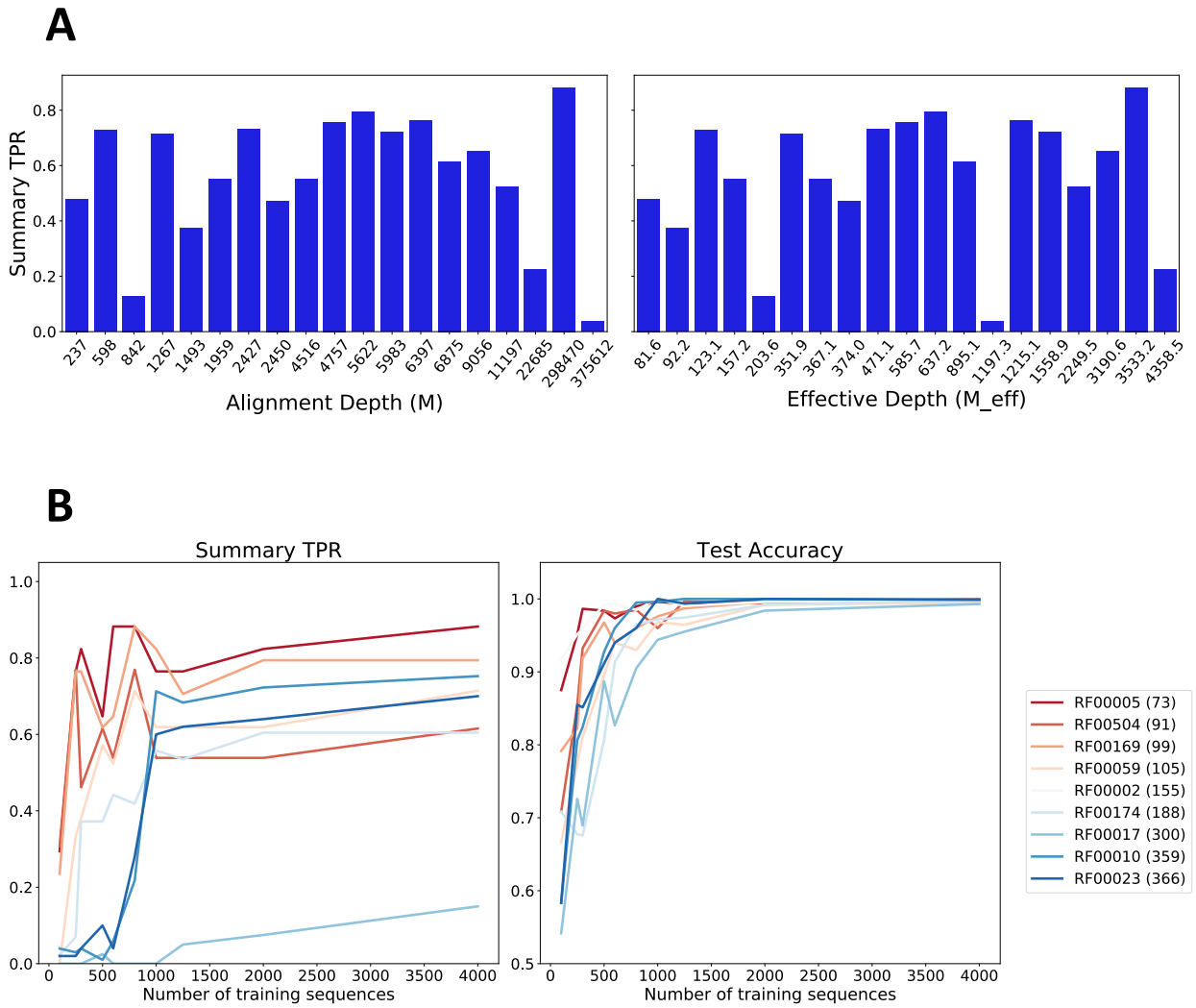

Figure 9: (A) the TPR was calculated for each trained MLP where the top L/2 SoM scores were called positive contacts and L is the length of the respective alignment. These were plotted against the alignment depth (M) and effective depth (M_{eff}) in ascending order. There is no noticeable trend between 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.

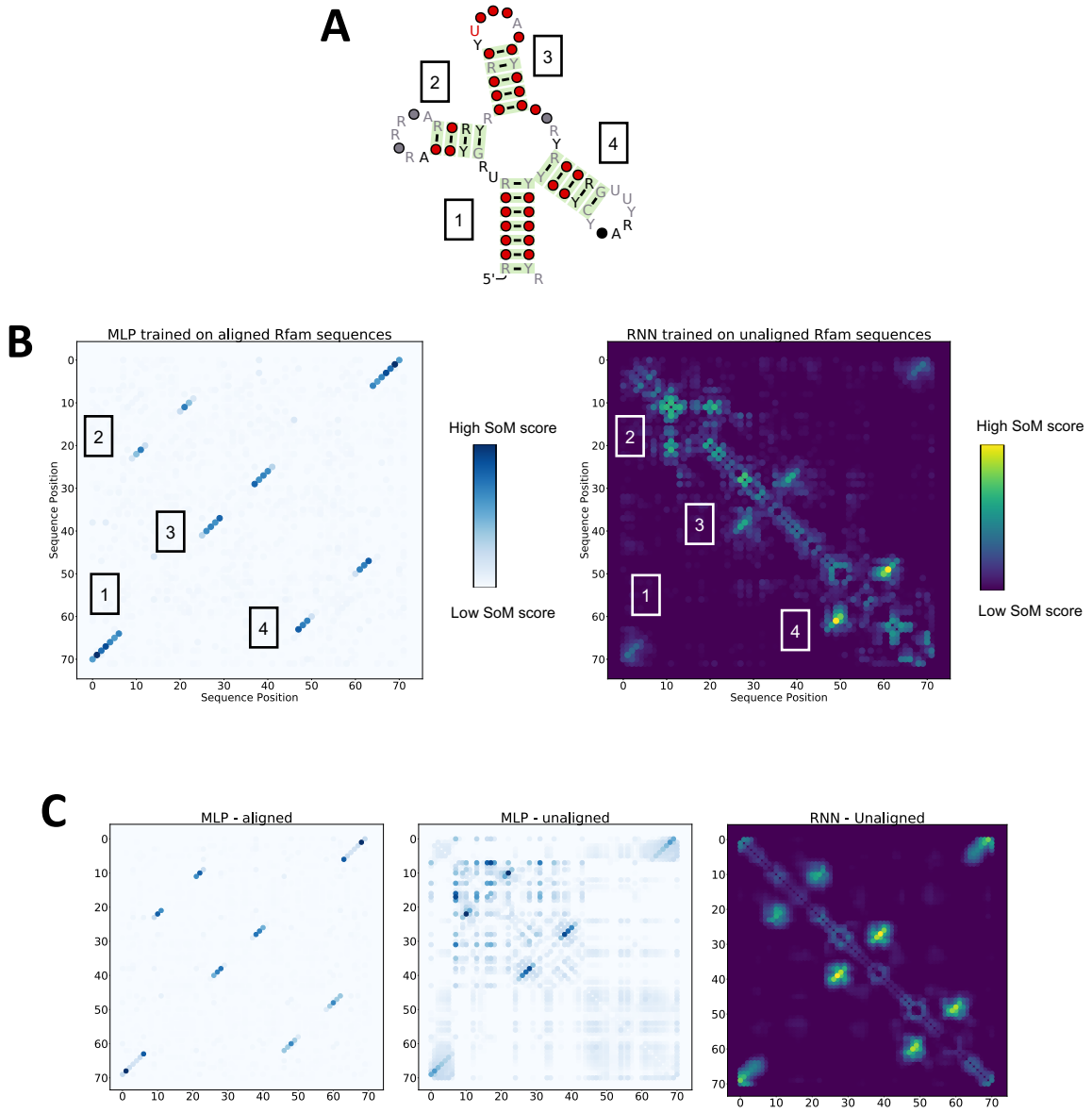

Figure 10: (A) Folded secondary structure of tRNA – RF00005 from Rfam. (B) Results from SoM performed on an MLP trained on the tRNA reduced Rfam alignment (blue) and from an RNN trained on the unaligned Rfam sequences (blue-green). The RNN seems to be learning general regions corresponding to true base 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.