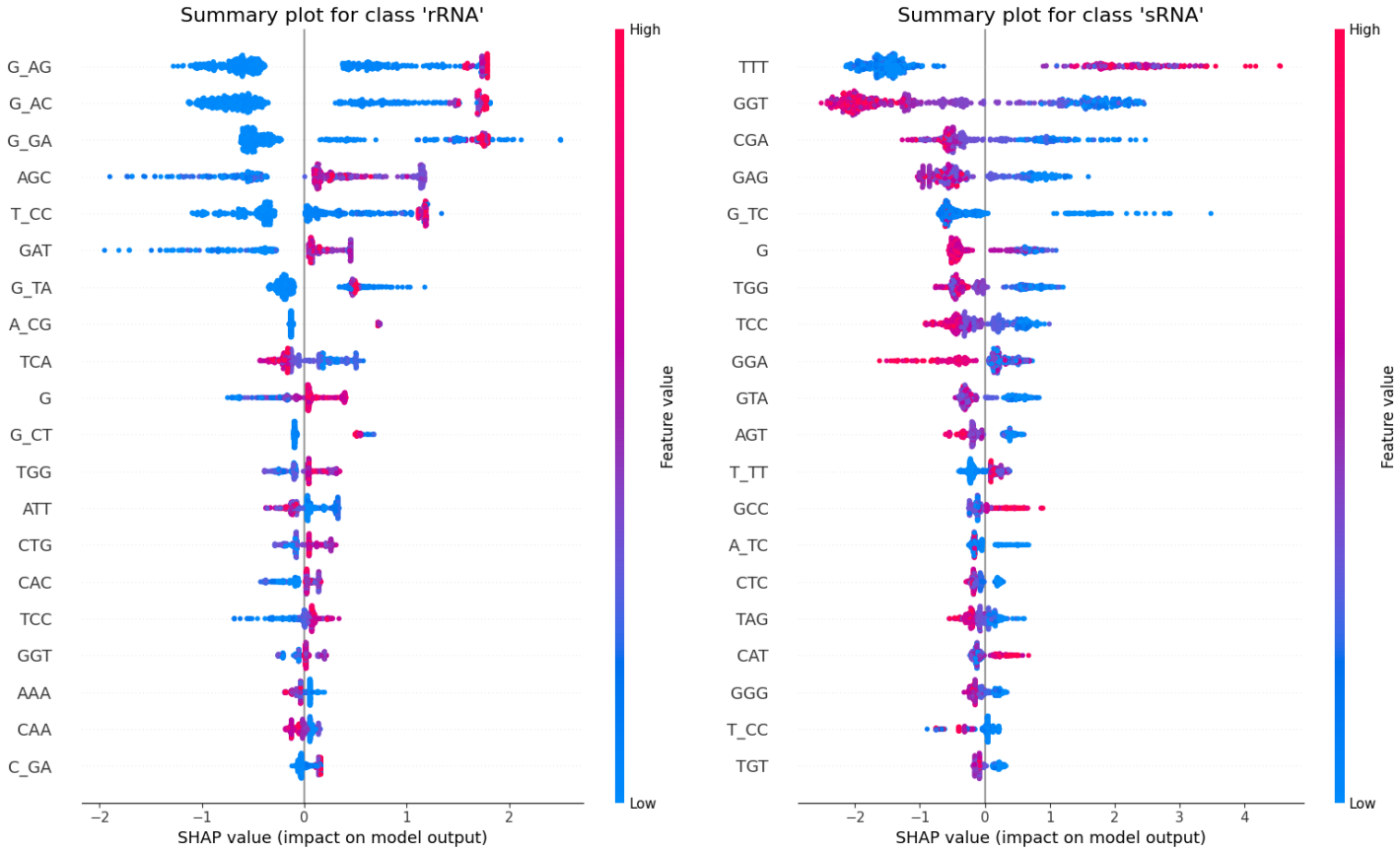
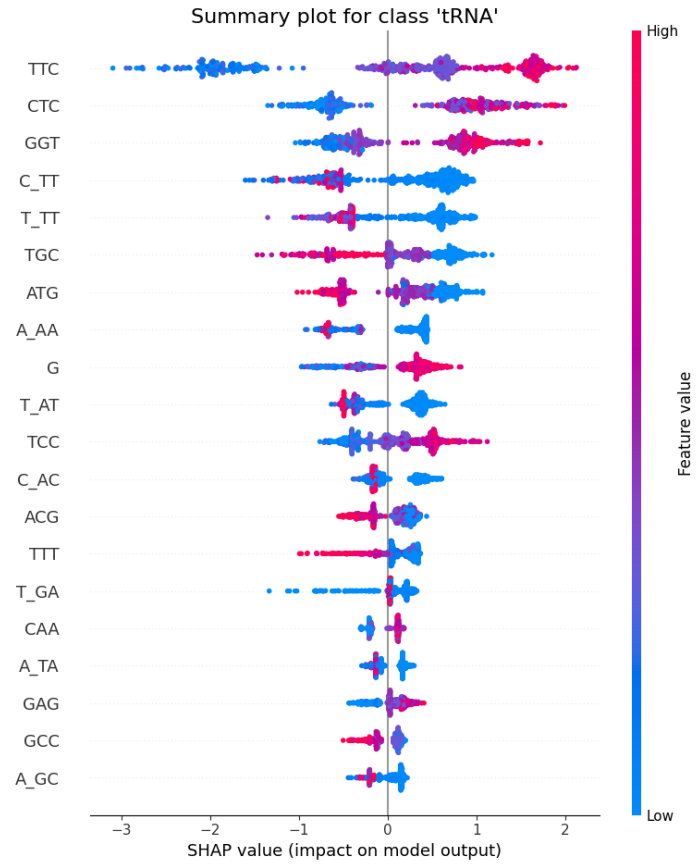
Model Interpretability Report (Multiclass)

BioAutoML
AUTOMATED FEATURE ENGINEERING

This report sustains the idea of being able to interpret and explain how and why the chosen model is classifying each entry as it is. All of Interpretability are based in the SHAP method, in which calculates what's the importance level for each feature in the classification process. It uses Shapley Values, a Game Theory concept, as a descriptive metric to create an hierarquical structure between the features.

Summary Plots





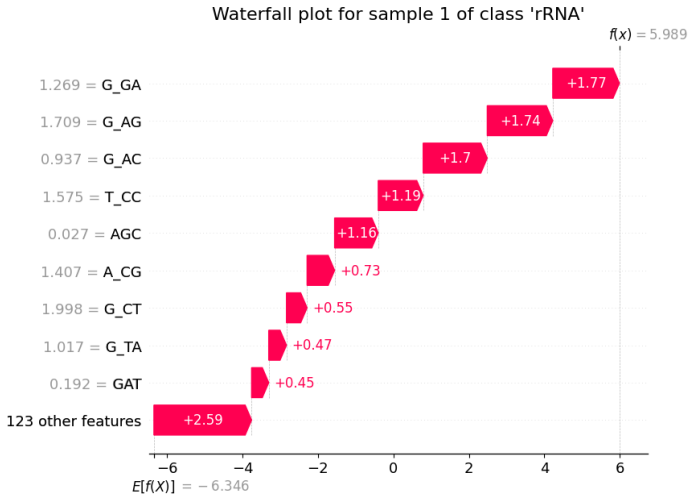
The above plot is called Summary plot and it shows, for each class, how low/high values of each feature contributed for the classification with that class. The features are ranked from most descriptive to least descriptive. This plot is a summarization of all the entries in the test set.

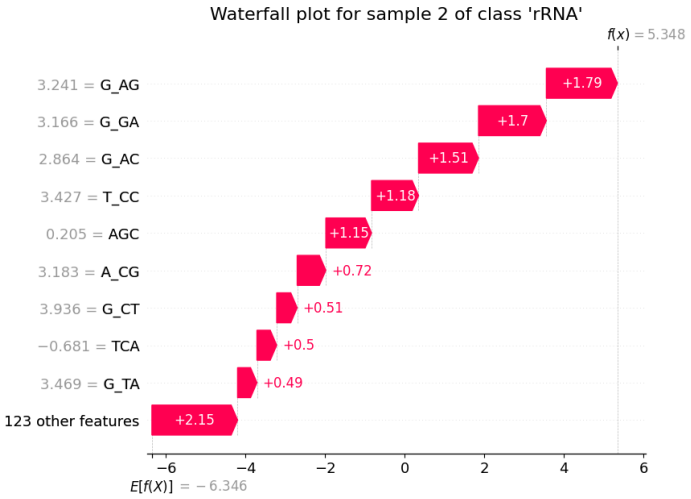
An impact with positive SHAP value means that a high (red dots), medium (purple dots) or low (blue dots) feature value contributes positively of an entry to be classified with that class. The inverse happens with negative SHAP values.

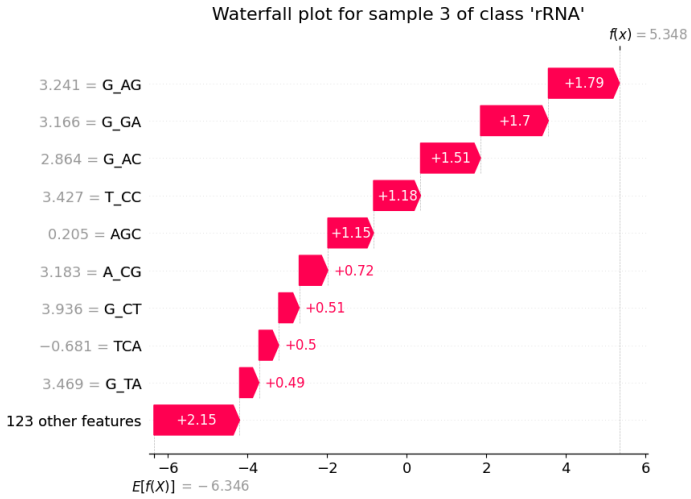
Waterfall Plots

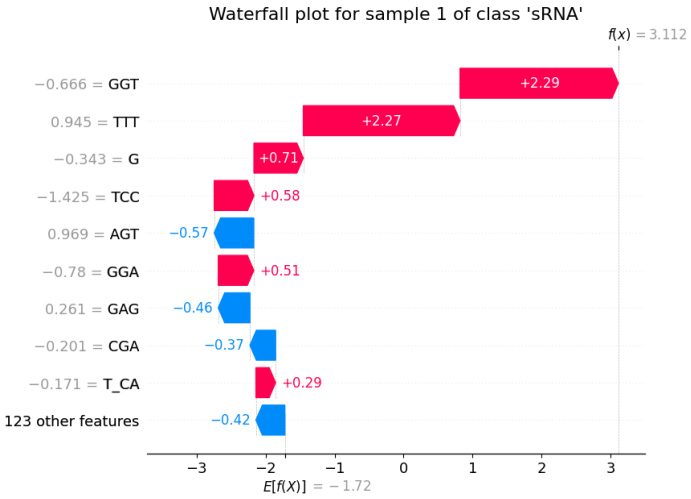
A Waterfall plot shows, for some entries, how each of the features contributed for it to be classified with its classified class. In this case, 3 samples for each class were chosen randomly to be analyzed.

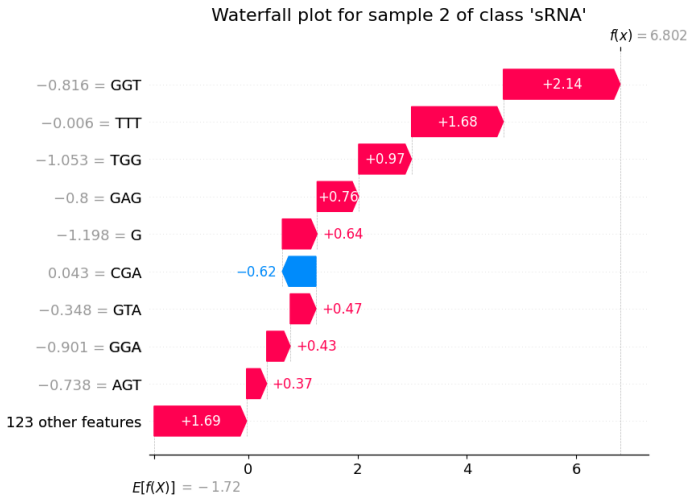
From a base expected value, E[f(x)], each feature contributes positively of negatively towards the entry's given class. At the end, when all the contributions are summed with E[f(x)], we get the final value of f(x) which led to the classification result.

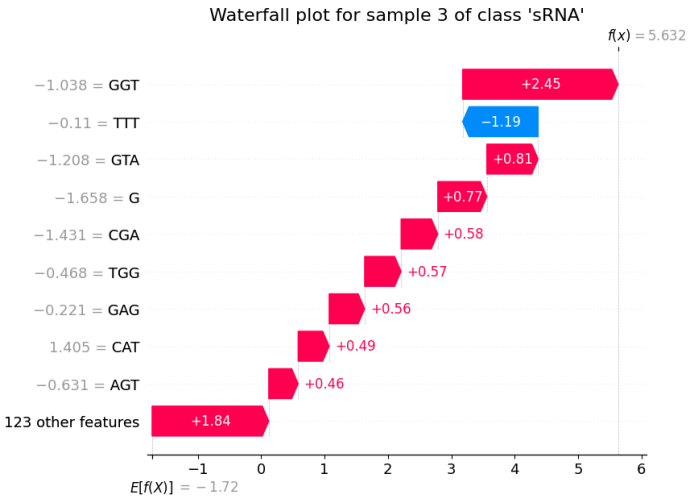


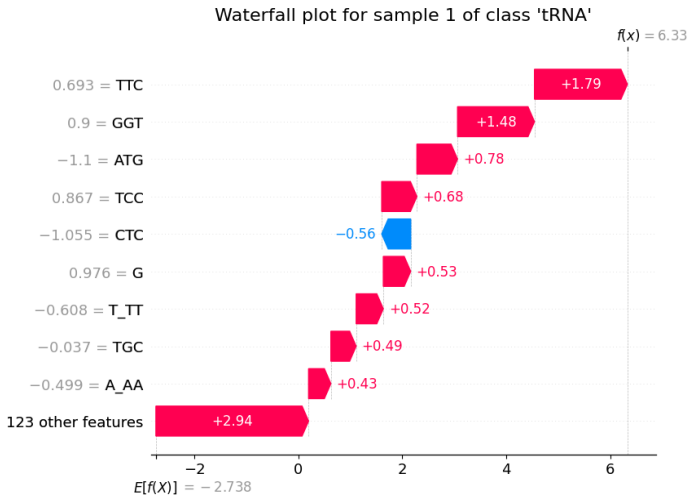


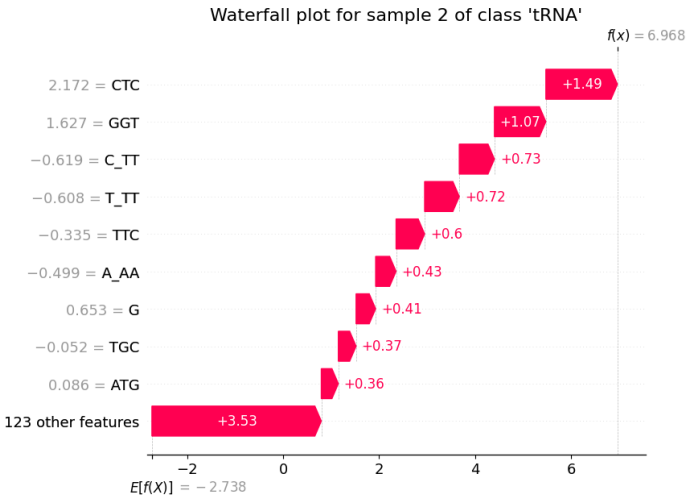

Waterfall plot for sample 3 of class 'rRNA'
f(x) = 5.348
3.241 = G_AG
3.166 = G_GA
2.864 = G_AC
3.427 = T_CC
0.205 = AGC
3.183 = A_CG
3.936 = G_CT
−0.681 = TCA
3.469 = G_TA
123 other features
+1.79
+1.7
+1.51
+1.18
+1.15
+0.72
+0.51
+0.5
+0.49
+2.15
−6 −4 −2 0 2 4 6
E[f(X)] = − 6.346

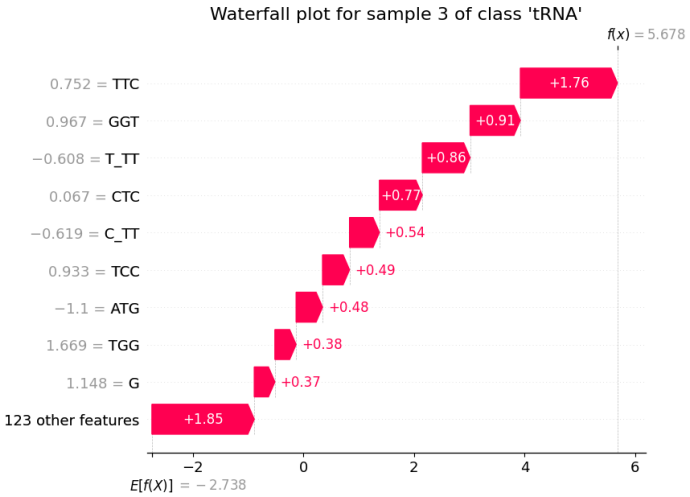

Waterfall plot for sample 1 of class 'sRNA'
f(x) = 3.112
−0.666 = GGT
0.945 = TTT
−0.343 = G
−1.425 = TCC
0.969 = AGT
−0.78 = GGA
0.261 = GAG
−0.201 = CGA
−0.171 = T_CA
123 other features
+2.29
+2.27
+0.71
+0.58
−0.57
+0.51
−0.46
−0.37
+0.29
−0.42
−3 −2 −1 0 1 2 3
E[f(X)] = − 1.72


Waterfall plot for sample 2 of class 'sRNA'
f(x) = 6.802
−0.816 = GGT
−0.006 = TTT
−1.053 = TGG
−0.8 = GAG
−1.198 = G
0.043 = CGA
−0.348 = GTA
−0.901 = GGA
−0.738 = AGT
123 other features
+2.14
+1.68
+0.97
+0.76
+0.64
−0.62
+0.47
+0.43
+0.37
+1.69
0 2 4 6
E[f(X)] = − 1.72


Waterfall plot for sample 3 of class 'sRNA'
f(x) = 5.632
−1.038 = GGT
−0.11 = TTT
−1.208 = GTA
−1.658 = G
−1.431 = CGA
−0.468 = TGG
−0.221 = GAG
1.405 = CAT
−0.631 = AGT
123 other features
+2.45
−1.19
+0.81
+0.77
+0.58
+0.57
+0.56
+0.49
+0.46
+1.84
−1 0 1 2 3 4 5 6
E[f(X)] = − 1.72


Waterfall plot for sample 1 of class 'tRNA'
f(x) = 6.33
0.693 = TTC
0.9 = GGT
−1.1 = ATG
0.867 = TCC
−1.055 = CTC
0.976 = G
−0.608 = T_TT
−0.037 = TGC
−0.499 = A_AA
123 other features
+1.79
+1.48
+0.78
+0.68
−0.56
+0.53
+0.52
+0.49
+0.43
+2.94
−2 0 2 4 6
E[f(X)] = − 2.738


Waterfall plot for sample 2 of class 'tRNA'
f(x) = 6.968
2.172 = CTC
1.627 = GGT
−0.619 = C_TT
−0.608 = T_TT
−0.335 = TTC
−0.499 = A_AA
0.653 = G
−0.052 = TGC
0.086 = ATG
123 other features
+1.49
+1.07
+0.73
+0.72
+0.6
+0.43
+0.41
+0.37
+0.36
+3.53
−2 0 2 4 6
E[f(X)] = − 2.738


Waterfall plot for sample 3 of class 'tRNA'
f(x) = 5.678
0.752 = TTC
0.967 = GGT
−0.608 = T_TT
0.067 = CTC
−0.619 = C_TT
0.933 = TCC
−1.1 = ATG
1.669 = TGG
1.148 = G
123 other features
+1.76
+0.91
+0.86
+0.77
+0.54
+0.49
+0.48
+0.38
+0.37
+1.85
−2
0
2
4
6
E[f(X)] = − 2.738