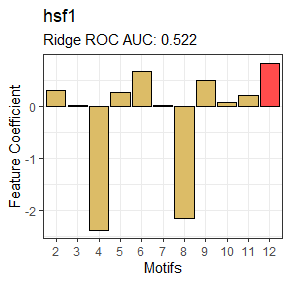
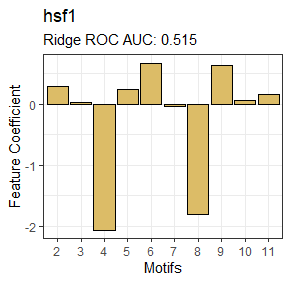


No basic is the single most important feature. Motifs still have some predictive power alone. But this finding is less surprising when you look at the top performing motifs which describe acidic and aromatic regions. Compare motif ML performance to performance of acidic, basic, and aromatic counts as features and you can see that sequence composition is a much better descriptor of functional tads.

|  |  |
| --- | --- |
| Motif | Regex |
| ANAC013 | [DE].{1,2}[YF].{1,4}[DE]L |
| p53 | [ALVIMWYF]..[ALVIMWYF][ALVIMWYF] |
| WxxLF | W..LF |
| AR | F..LF |
| CREBZF | D[VILM][VILM][RKDEQNHSTYC][RKDEQNHSTYC][VILM][VILFWYM] |

Additional data on HSF1 library

Unbalanced training data, I’m going to see what balanced data will do