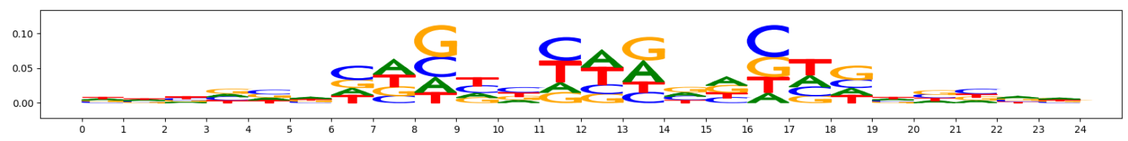
# Bias model training and quality check report

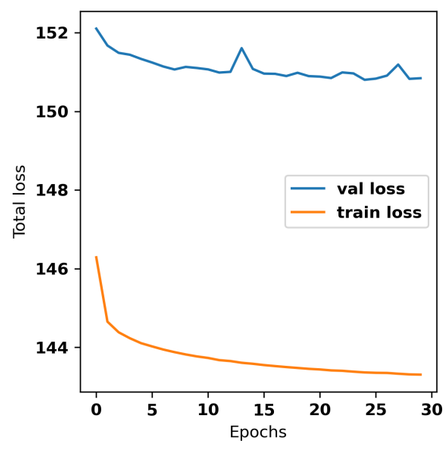
## Preprocessing report

The image below should look closely like a Tn5 or DNase bias enzyme motif.




## Training report

The val loss (validation loss) will decrease and saturate after a few epochs.




## Bias model performance in peaks and non-peaks

**Counts Metrics:** The pearsonr in non-peaks should be greater than 0 (higher the better). The pearsonr in peaks should be greater than -0.3 (otherwise the bias model could potentially be capturing AT bias). MSE (Mean Squared Error) will be high in peaks.

**Profile Metrics:** Median JSD (Jensen Shannon Divergence between observed and predicted) lower the better. Median norm JSD is median of the min-max normalized JSD where min JSD is the worst case JSD i.e JSD of observed with uniform profile and max JSD is the best case JSD i.e 0. Median norm JSD is higher the better. Both JSD and median norm JSD are sensitive to read-depth. Higher read-depth results in better metrics.

**What to do if your pearsonr in peaks is less than -0.3?** In the range of -0.3 to -0.5 please be wary of your chrombpnet_wo_bias.h5 (that wil potentially be trained with this bias model) TFModisco showing lots of GC rich motifs (> 3 in the top-10). If this is not the case you can continue using the chrombpnet_wo_bias.h5. If you end up seeing a lot of GC rich motifs it is likely that bias model has learnt a different GC distribution than your GC-content in peaks. You might benefit from increasing the bias_threshold_factor argument input to the *chrombpnet bias pipeline* or *chrombpnet bias train* command used in training the bias model and retrain a new bias model. For more intuition about this argument refer to the FAQ section in wiki. If the value is less than -0.5 the chrombpnet training will automatically throw an error.

| | nonpeaks.pearsonr | nonpeaks.mse | peaks.pearsonr | peaks.mse |
|---|---|---|---|---|
| counts metrics | -0.0 | 0.68 | -0.49 | 0.75 |

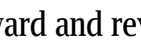
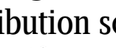
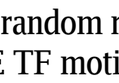
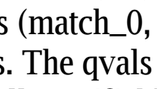
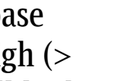
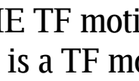
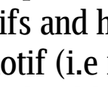
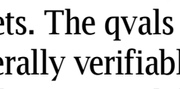
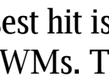
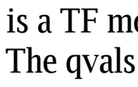
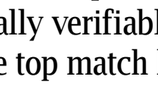
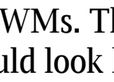
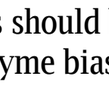
| | nonpeaks.median_jsd | nonpeaks.median_norm_jsd | peaks.median_jsd | peaks.median_norm_jsd |
|---|---|---|---|---|
| profile metrics | 0.64 | 0.15 | 0.5 | 0.26 |

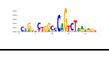
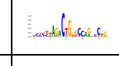
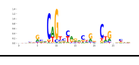
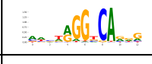
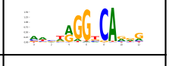
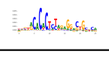
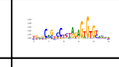
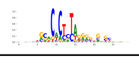
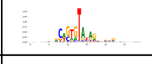
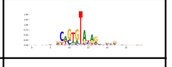
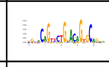
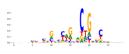
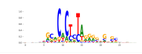
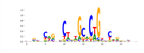
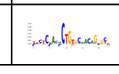
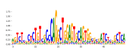
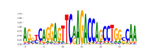
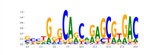
## TFModisco motifs learnt from bias model (bias.h5) model

**TFModisco motifs generated from profile contribution scores of the bias model.** cwm_fwd, cwm_rev are the forward and reverse complemented consolidated motifs from contribution scores in subset of random peaks. These CWM motifs should be free from any Transcription Factor (TF) motifs and should contain either only bias motifs or random repeats. For each of these motifs, we use TOMTOM to find the top-3 closest matches (match_0, match_1, match_2) from a database consisting of both MEME TF motifs and heterogenous enzyme bias motifs that we have repeatedly seen in our datasets. The qvals (qval0,qval1,qval2) should be high (> 0.0001) if the closest hit is a TF motif (i.e indicating that the closest match is not the correct match) - this is also generally verifiable by eye as the closest match will look nothing like the CWMs. The qvals should be low if the closest hit is enzyme bias motif and generally verifiable that the top match looks like the CWM. The first 3-5 motifs in the list below should look like enzyme bias motif.

**What to do if you find an obvious TF motif in the list?**
Do not use this bias model as it will regress the contribution of the TF motifs (along with bias motifs) from the chrombpnet_nobias.h5. Reduce the bias_threshold_factor argument input to the *chrombpnet bias pipeline* or *chrombpnet bias train* command used in training the bias model and retrain a new bias model. For more intuition about this argument refer to the FAQ section in wiki.

**What to do if you are unsure if a given CWM motif is resembling the match_0 logo for example?**
Get marginal footprint on the match_0 motif logo (using the command *chrombpnet footprints* and make sure that the bias models footprint is closer to that of controls with no motif inserted - for examples look at FAQ )

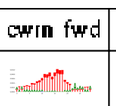
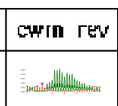
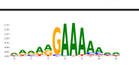
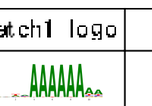
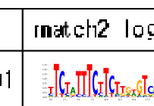
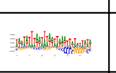
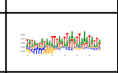
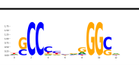
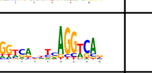
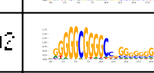
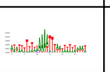
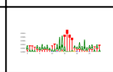
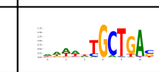
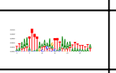
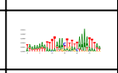
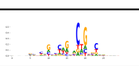
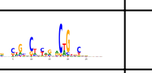
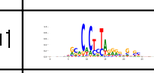
| pattern | NumSeqs | cwm_fwd | cwm_rev | match0 | qval0 | match0_logo | match1 | qval1 | match1_logo | match2 | qval2 | match2_logo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pos_0 | 8519 |  |  | TN5_1 | 2.343550e-09 |  | TN5_2 | 1.714920e-08 |  | TN5_7 | 0.000100 |  |
| pos_1 | 6312 |  |  | TN5_4 | 2.034240e-02 |  | TN5_5 | 2.034240e-02 |  | TN5_8 | 0.027068 |  |
| pos_2 | 4140 |  |  | TN5_1 | 5.084640e-07 |  | TN5_3 | 3.240360e-06 |  | TN5_2 | 0.003034 |  |
| pos_3 | 3654 |  |  | TN5_3 | 9.030770e-15 |  | TN5_4 | 2.155130e-03 |  | TN5_5 | 0.002155 |  |
| pos_4 | 3416 |  |  | TN5_3 | 1.104160e-01 |  | TN5_4 | 1.104160e-01 |  | TN5_5 | 0.110410 |  |
| pos_5 | 2918 |  |  | TN5_3 | 8.388180e-05 |  | TN5_7 | 3.430450e-04 |  | TN5_1 | 0.001952 |  |
| pos_6 | 2794 |  |  | TN5_2 | 3.045640e-17 |  | TN5_4 | 4.788040e-06 |  | TN5_5 | 0.000005 |  |
| pos_7 | 796 |  |  | TN5_3 | 4.789920e-07 |  | TN5_1 | 9.740420e-04 |  | TEX1_TEX_4 | 0.084708 |  |
| pos_8 | 223 |  |  | POU3F4_POU_2 | 1.000000e+00 |  | NFAC2_HUMAN.H11MO.0.B | 1.000000e+00 |  | NFAC2_MOUSE.H11MO.0.C | 1.000000 |  |
| pos_9 | 146 |  |  | ZNF384_MA1125.1 | 5.454130e-02 |  | PRDM6_HUMAN.H11MO.0.C | 6.009230e-02 |  | STAT1_MOUSE.H11MO.0.A | 0.114602 |  |
| pos_10 | 139 |  |  | FOXD2_forkhead_1 | 1.000000e+00 |  | NFATC1_NFAT_1 | 1.000000e+00 |  | CPEB1_RRM_1 | 1.000000 |  |
| pos_11 | 125 |  |  | ZNF384_MA1125.1 | 2.832720e-02 |  | PRDM6_HUMAN.H11MO.0.C | 2.832720e-02 |  | FOXJ3_HUMAN.H11MO.0.A | 0.165850 |  |
| pos_12 | 95 |  |  | DNASE_2 | 2.008540e-01 |  | ZNF384_MA1125.1 | 1.000000e+00 |  | _HX3_HUMAN.H11MO.0.C | 1.000000 |  |
| pos_13 | 85 |  |  | NFATC2_MA0152.1 | 5.288450e-01 |  | NFATC1_NFAT_1 | 5.288450e-01 |  | PRDM6_HUMAN.H11MO.0.C | 0.703003 |  |
| pos_14 | 74 |  |  | _N54_MA0619.1 | 1.000000e+00 |  | ONECUT3_CUT_1 | 1.000000e+00 |  | ONECUT3_MA0757.1 | 1.000000 |  |
| pos_15 | 72 |  |  | EGR2_HUMAN.H11MO.0.A | 1.000000e+00 |  | EGR1_HUMAN.H11MO.0.A | 1.000000e+00 |  | RREB1_MA0073.1 | 1.000000 |  |

| pattern | NumSeqs | cwm fwd | cwm rev | match0 | qval0 | match0 logo | match1 | qval1 | match1 logo | match2 | qval2 | match2 logo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pos 16 | 64 |  |  | TN5 8 | 9.999990e-01 |  | RORA HUMAN.H11MO.0.C | 9.999990e-01 |  | RORA MOUSE.H11MO.0.C | 0.999999 |  |
| pos 17 | 42 |  |  | TN5 3 | 4.137570e-04 |  | TN5 4 | 2.010910e-01 |  | TN5 5 | 0.201091 |  |
| pos 18 | 33 |  |  | TN5 1 | 1.252380e-01 |  | TN5 3 | 1.252380e-01 |  | TN5 2 | 0.125238 |  |
| pos 19 | 20 |  |  | TN5 6 | 4.400370e-08 |  | DNASE 5 | 3.970190e-01 |  | PAX5 HUMAN.H11MO.0.A | 0.419224 |  |

**TFModisco motifs generated from counts contribution scores of the bias model.** cwm_fwd, cwm_rev are the forward and reverse complemented consolidated motifs from contribution scores in subset of random peaks. These motifs should be free from any Transcription Factor (TF) motifs and should contain motifs either weakly related to bias motifs or random repeats. For each of these motifs, we use TOMTOM to find the top-3 closest matches (match_0, match_1, match_2) from a database consisting of both MEME TF motifs and heterogenous enzyme bias motifs that we have repeatedly seen in our datasets. The qvals should be high (> 0.0001) if the closest hit is a TF motif (i.e indicating that the closest match is not the correct match, this is also generally verifiable by eye and making sure the closest match looks nothing like the CWMs).

**What to do if you find an obvious TF motif in the list?**
Do not use this bias model as it will regress the contribution of the TF motifs (along with bias motifs) from the chrombpnet_nobias.h5. Reduce the bias_threshold_factor argument input to the *chrombpnet bias pipeline* or *chrombpnet bias train* command used in training the bias model and retrain a new bias model. For more intuition about this argument refer to the FAQ section in wiki.

**What to do if you are unsure if a given CWM motif is resembling the match_0 logo for example?**
Get marginal footprint on the match_0 motif logo (using the command *chrombpnet footprints* and make sure that the bias models footprint is closer to that of controls with no motif inserted - for examples look at FAQ )

| pattern | NumSeqs | cwm fwd | cwm rev | match0 | qval0 | match0 logo | match1 | qval1 | match1 logo | match2 | qval2 | match2 logo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pos 0 | 673 |  |  | PRDM6 HUMAN.H11MO.0.C | 7.910040e-02 |  | ZNF384 MA1125.1 | 1.981690e-01 |  | ZFP28 HUMAN.H11MO.0.C | 5.013390e-01 |  |
| pos 1 | 204 |  |  | TFAP2A TFAP 3 | 2.722730e-02 |  | THA HUMAN.H11MO.0.C | 2.722730e-02 |  | SP1 HUMAN.H11MO.0.A | 2.722730e-02 |  |
| pos 2 | 219 |  |  | POU3F4 POU 2 | 1.000000e+00 |  | MAFK bZIP 1 | 1.000000e+00 |  | MAFK bZIP 3 | 1.000000e+00 |  |
| pos 3 | 196 |  |  | TN5 1 | 4.490060e-01 |  | TN5 7 | 4.490060e-01 |  | TN5 3 | 5.059470e-01 |  |
| pos 4 | 159 |  |  | ZFX HUMAN.H11MO.0.A | 1.000000e+00 |  | ZBT14 HUMAN.H11MO.0.C | 1.000000e+00 |  | Barhl1_mouse homeodomain 1 | 1.000000e+00 |  |
| pos 5 | 121 |  |  | NRF1 MA0506.1 | 3.822950e-01 |  | MAX1 HUMAN.H11MO.0.A | 3.822950e-01 |  | MAX1 MOUSE.H11MO.0.A | 3.822950e-01 |  |
| pos 6 | 117 |  |  | TBX21 TBX 6 | 5.989840e-01 |  | TN5 7 | 1.000000e+00 |  | TBX21 TBX 3 | 1.000000e+00 |  |
| pos 7 | 114 |  |  | MYC MOUSE.H11MO.0.A | 1.197070e-01 |  | MAX1 HUMAN.H11MO.0.A | 1.197070e-01 |  | MAX1 MOUSE.H11MO.0.A | 1.197070e-01 |  |
| pos 8 | 96 |  |  | TN5 7 | 1.000000e+00 |  | ZNG7 HUMAN.H11MO.0.C | 1.000000e+00 |  | PRDM4 C2H2 1 | 1.000000e+00 |  |
| pos 9 | 85 |  |  | SP2 HUMAN.H11MO.0.A | 1.204860e-04 |  | SP2 MOUSE.H11MO.0.B | 1.204860e-04 |  | ZFX MOUSE.H11MO.0.B | 3.991010e-04 |  |
| pos 10 | 84 |  |  | SREBF2 HUMAN.H11MO.0.B | 7.350690e-01 |  | SREBF2 MOUSE.H11MO.0.B | 7.350690e-01 |  | RARG nuclearreceptor 5 | 1.000000e+00 |  |
| pos 11 | 73 |  |  | SP2 HUMAN.H11MO.0.A | 2.121730e-05 |  | SP2 MOUSE.H11MO.0.B | 2.121730e-05 |  | SP3 HUMAN.H11MO.0.B | 1.744830e-04 |  |
| pos 12 | 46 |  |  | TEAD1 TEA 2 | 2.400170e-01 |  | NR1I3 HUMAN.H11MO.0.C | 4.311910e-01 |  | NR1I3 MOUSE.H11MO.0.C | 4.311910e-01 |  |
| pos 13 | 45 |  |  | SP2 HUMAN.H11MO.0.A | 2.209150e-05 |  | SP2 MOUSE.H11MO.0.B | 2.209150e-05 |  | SP1 HUMAN.H11MO.0.A | 1.149740e-04 |  |
| pos 14 | 31 |  |  | SP2 HUMAN.H11MO.0.A | 1.235400e-04 |  | SP2 MOUSE.H11MO.0.B | 1.235400e-04 |  | SP1 MOUSE.H11MO.0.A | 2.509420e-04 |  |
| pos 15 | 27 |  |  | SP2 HUMAN.H11MO.0.A | 2.015610e-05 |  | SP2 MOUSE.H11MO.0.B | 2.015610e-05 |  | SP1 HUMAN.H11MO.0.A | 3.504480e-05 |  |
| pos 16 | 20 |  |  | Uncx_mouse homeodomain 1 | 1.000000e+00 |  | UNCX homeodomain 1 | 1.000000e+00 |  | Gfi1 MA0038.1 | 1.000000e+00 |  |