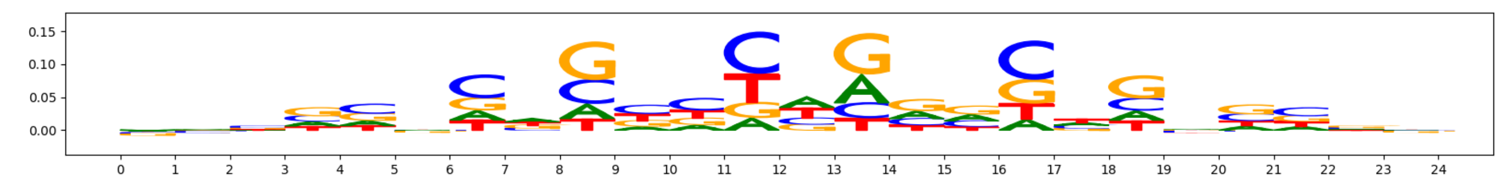
Bias model training and quality check report

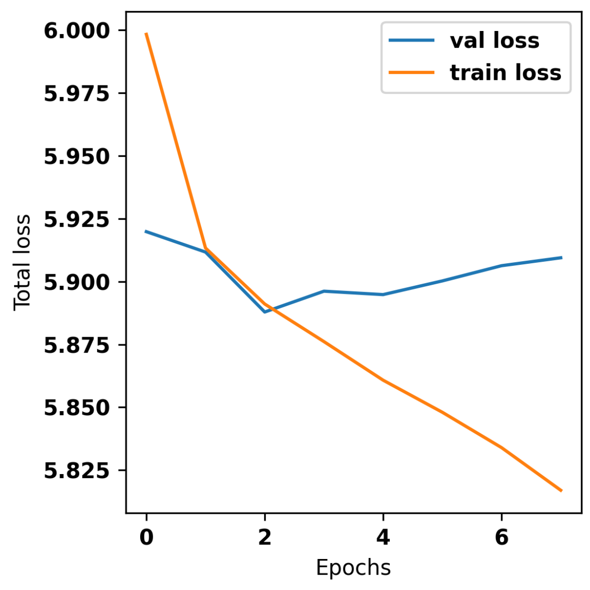
Preprocessing report

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	-0.19	0.87	-0.45	9.21

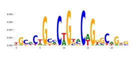
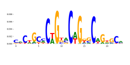
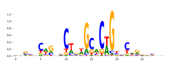
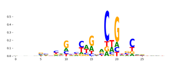
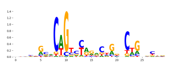
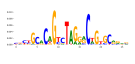
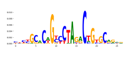
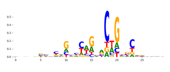
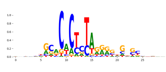
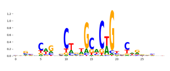
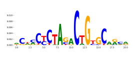
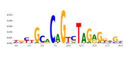
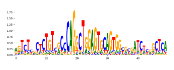
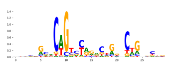
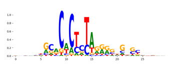
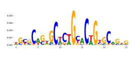
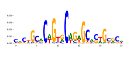
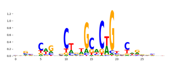
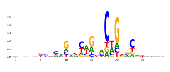
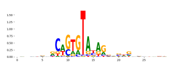
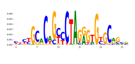
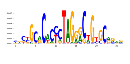
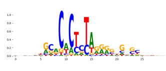
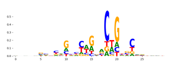
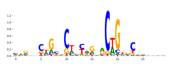
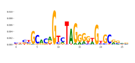
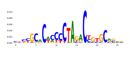
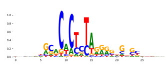
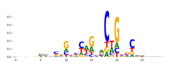
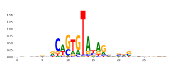
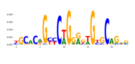
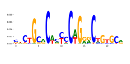
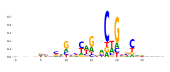
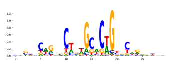
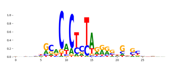
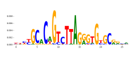
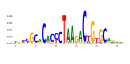
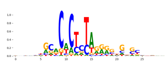
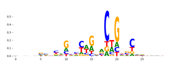
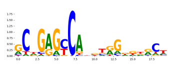
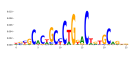
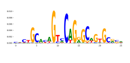
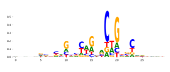
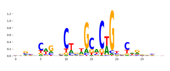
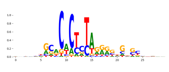
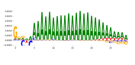
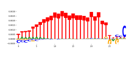
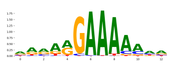
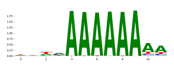
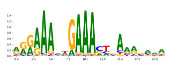
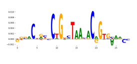
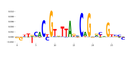
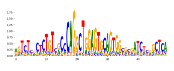
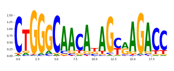
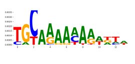
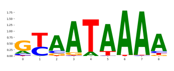
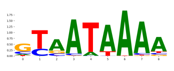
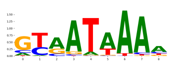
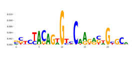
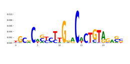
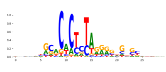
	nonpeaks.median_jsd	nonpeaks.median_norm_jsd	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.81	0.02	0.74	0.08

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)

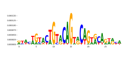
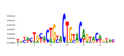
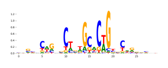
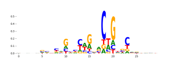
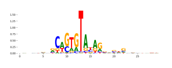
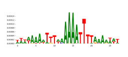
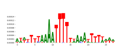
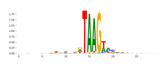
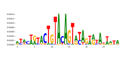
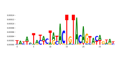
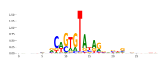
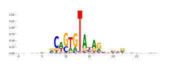
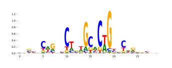
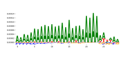
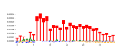
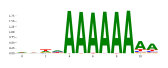
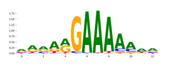
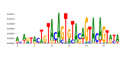
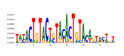
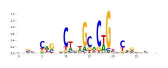
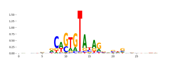
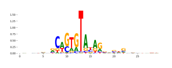
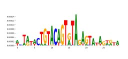
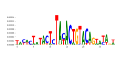
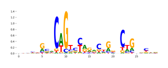
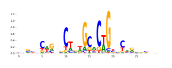
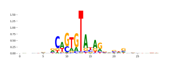
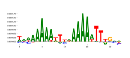
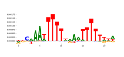
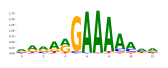
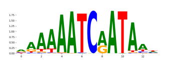
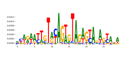
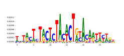
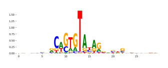
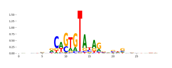
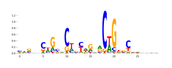
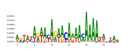
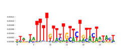
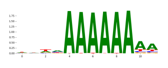
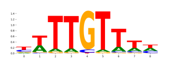
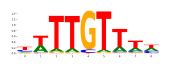
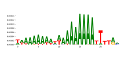
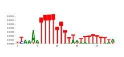
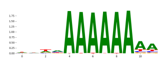
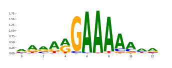
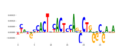
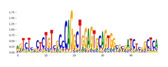
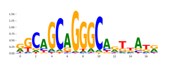
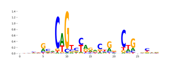
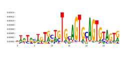
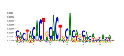
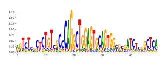
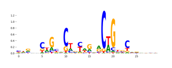
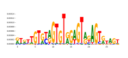
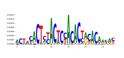
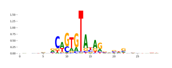
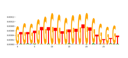
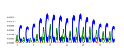
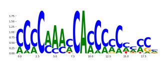
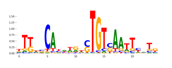
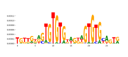
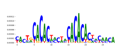
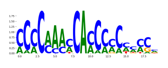
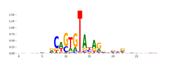
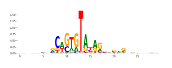
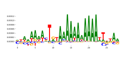
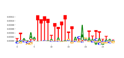
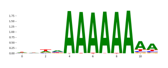
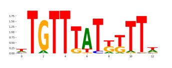
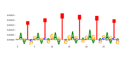
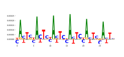
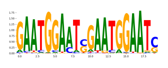
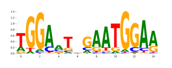
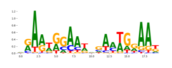
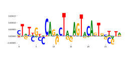
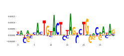
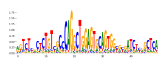
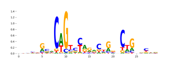
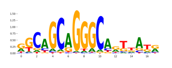
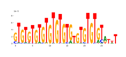
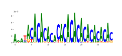
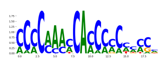
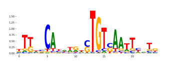
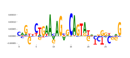
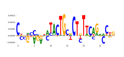
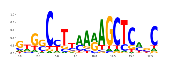
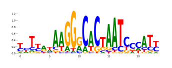
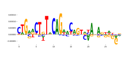
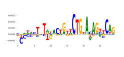
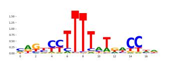
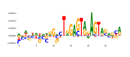
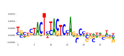
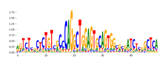
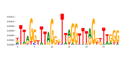
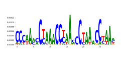
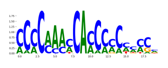
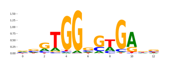
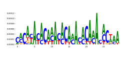
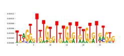
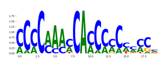
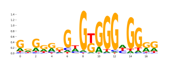
pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos__0	7428			TN5_2	2.288990e-05		TN5_1	0.000088		TN5_8	0.011999	
pos__1	4242			TN5_1	2.355870e-07		TN5_3	0.000004		TN5_2	0.002699	
pos__2	4135			TN5_6	2.403020e-04		TN5_8	0.034761		TN5_3	0.078611	
pos__3	3709			TN5_2	7.359000e-09		TN5_1	0.003010		TN5_4	0.007352	
pos__4	3258			TN5_3	1.298910e-09		TN5_1	0.000113		TN5_7	0.018416	
pos__5	2750			TN5_3	8.151590e-04		TN5_1	0.009594		TN5_4	0.016195	
pos__6	598			TN5_1	6.257520e-04		TN5_2	0.000626		TN5_3	0.001863	
pos__7	515			TN5_3	9.802630e-05		TN5_1	0.000610		ZN554_HUMAN.H11MO.0.C	0.178798	
pos__8	438			TN5_1	9.919950e-03		TN5_2	0.027255		TN5_3	0.048034	
pos__9	304			PRDM6_HUMAN.H11MO.0.C	8.247210e-02		ZNF384_MA1125.1	0.082472		STAT1_MOUSE.H11MO.0.A	0.236510	
pos__10	92			TN5_6	5.732510e-02		FOXB1_forkhead_1	1.000000		ZN121_HUMAN.H11MO.0.C	1.000000	
pos__11	84			HOXC12_homeodomain_1	8.251390e-02		HOXD12_homeodomain_1	0.082514		HOXD12_homeodomain_4	0.082514	
pos__12	34			TN5_3	5.717810e-02		NR2F1_MA0017.2	0.057178		NR2F1_nuclearreceptor_4	0.057178	

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	7907			TN5_2	3.311020e-04		TN5_1	3.551440e-04		TN5_4	1.852220e-02	
pos__1	5058			DNASE_2	1.000000e+00		LHX3_HUMAN.H11MO.0.C	1.000000e+00		None	NaN	
pos__2	3462			TN5_4	1.534940e-02		TN5_5	1.534940e-02		TN5_2	1.427680e-01	
pos__3	1592			ZNF384_MA1125.1	6.524530e-02		PRDM6_HUMAN.H11MO.0.C	6.524530e-02		FOXJ3_HUMAN.H11MO.0.A	2.204240e-01	
pos__4	1017			TN5_2	3.199300e-04		TN5_4	7.584600e-04		TN5_5	7.584600e-04	
pos__5	998			TN5_8	4.318060e-02		TN5_2	4.318060e-02		TN5_4	4.662670e-02	
pos__6	992			PRDM6_HUMAN.H11MO.0.C	1.637360e-01		ONECUT3_CUT_1	1.637360e-01		ONECUT3_MA0757.1	1.637360e-01	
pos__7	951			TN5_4	4.033330e-02		TN5_5	4.033330e-02		TN5_7	2.249750e-01	
pos__8	792			ZNF384_MA1125.1	1.005810e-01		SRY_HUMAN.H11MO.0.B	1.005810e-01		SRY_MOUSE.H11MO.0.B	1.005810e-01	
pos__9	628			ZNF384_MA1125.1	1.505310e-03		PRDM6_HUMAN.H11MO.0.C	4.278660e-01		FOXJ3_HUMAN.H11MO.0.A	4.278660e-01	
pos__10	596			TN5_6	6.803010e-24		ZSC31_HUMAN.H11MO.0.C	9.422140e-01		TN5_8	9.422140e-01	
pos__11	381			TN5_6	4.329560e-04		TN5_7	1.089260e-01		NKX25_MOUSE.H11MO.0.A	6.389360e-01	
pos__12	365			NKX25_MOUSE.H11MO.0.A	2.713330e-01		NKX21_MOUSE.H11MO.0.A	2.713330e-01		TN5_4	2.713330e-01	
pos__13	309			RREB1_MA0073.1	1.000000e+00		BRAC_MOUSE.H11MO.0.B	1.000000e+00		None	NaN	
pos__14	215			RREB1_MA0073.1	1.509250e-01		TN5_4	4.970310e-01		TN5_5	4.970310e-01	
pos__15	174			ZNF384_MA1125.1	3.464990e-02		FOXJ3_HUMAN.H11MO.0.A	3.092960e-01		FOXJ3_MOUSE.H11MO.0.A	3.092960e-01	
pos__16	128			ZN502_HUMAN.H11MO.0.C	2.924840e-07		SMCA5_MOUSE.H11MO.0.C	2.191630e-06		ZN394_HUMAN.H11MO.0.C	9.933130e-06	
pos__17	102			TN5_6	8.559250e-15		TN5_8	9.394500e-01		ZSC31_HUMAN.H11MO.0.C	9.394500e-01	
pos__18	83			RREB1_MA0073.1	1.000000e+00		BRAC_MOUSE.H11MO.0.B	1.000000e+00		None	NaN	
pos__19	70			FEZF1_HUMAN.H11MO.0.C	8.828710e-02		ZN667_HUMAN.H11MO.0.C	1.000000e+00		ZN264_HUMAN.H11MO.0.C	1.000000e+00	
pos__20	51			MEIS2_MEIS_1	1.000000e+00		ZN350_HUMAN.H11MO.0.C	1.000000e+00		MEF2B_MA0660.1	1.000000e+00	
pos__21	29			TN5_6	5.545980e-04		PAX9_MA0781.1	3.572190e-01		PAX9_PAX_1	3.572190e-01	
pos__22	27			RREB1_MA0073.1	1.000000e+00		ZN281_MOUSE.H11MO.0.A	1.000000e+00		SRBP2_HUMAN.H11MO.0.B	1.000000e+00	
pos__23	25			RREB1_MA0073.1	1.708960e-01		EGR2_HUMAN.H11MO.0.A	9.264730e-01		BRAC_MOUSE.H11MO.0.B	1.000000e+00