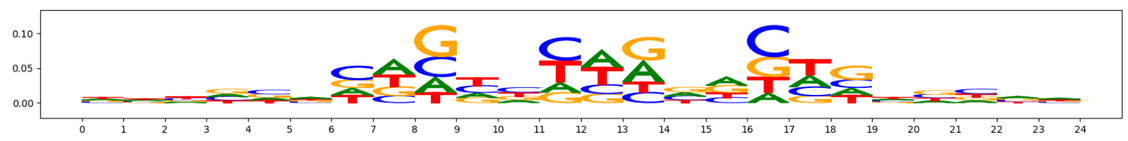
Bias model training and quality check report

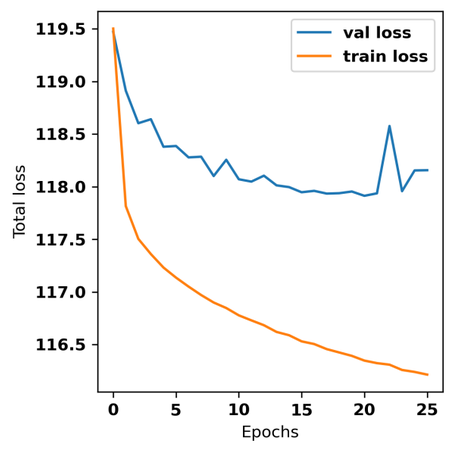
Preprocessing report

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	-0.12	1.47	-0.51	9.48

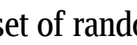
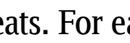
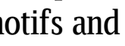
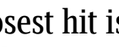
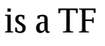
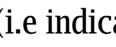
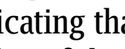
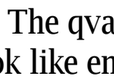
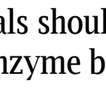
	nonpeaks.median_jsd	nonpeaks.median_norm_jsd	peaks.median_jsd	peaks.median_norm_jsd
profile metrics	0.64	0.15	0.5	0.26

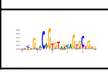
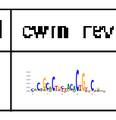
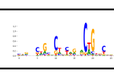
TFModisco motifs learnt from bias model (bias.h5) model

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)

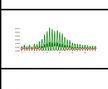
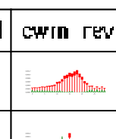
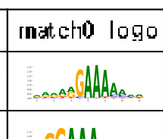
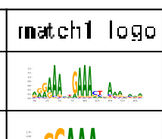
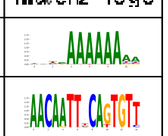
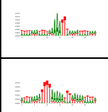
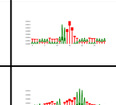
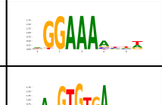
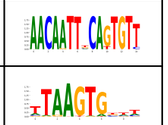
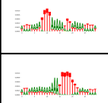
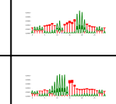
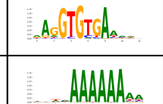
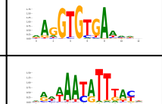
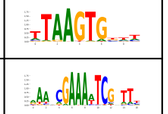
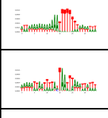
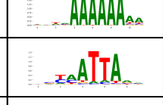
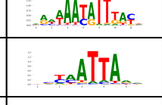
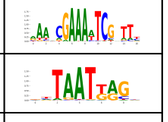
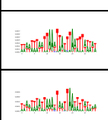
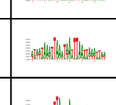
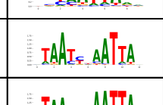
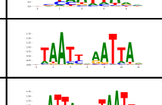
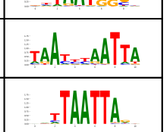
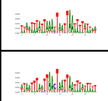
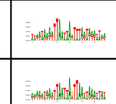
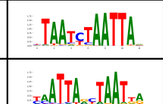
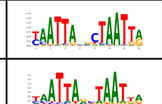
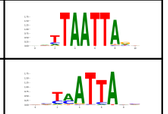
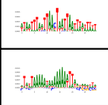
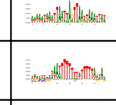
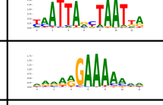
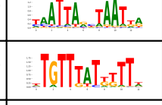
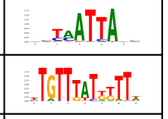
pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos 0	8213			TN5 2	4.359980e-08		TN5 1	7.617080e-08		TN5 7	0.000141	
pos 1	6243			TN5 4	1.210920e-01		TN5 5	1.210920e-01		TN5 8	0.203900	
pos 2	3800			TN5 1	4.750640e-05		TN5 2	7.574880e-05		TN5 3	0.000703	
pos 3	3589			TN5 3	1.189150e-01		KLF4 MA0039.3	1.189150e-01		TN5 4	0.118915	
pos 4	2912			TN5 2	8.901600e-08		TN5 4	2.010450e-02		TN5 5	0.020105	
pos 5	2899			TN5 3	2.700480e-12		TN5 4	5.920940e-02		TN5 5	0.059209	
pos 6	2721			TN5 3	6.873470e-05		TN5 7	2.910220e-04		TN5 1	0.003309	
pos 7	939			TN5 3	3.874750e-05		TN5 1	2.439830e-02		TEX1 TBX 4	0.127420	
pos 8	539			TN5 3	1.220590e-06		TN5 7	6.117800e-03		TN5 4	0.077513	
pos 9	278			RREB1 MA0073.1	1.000000e+00		EGR2 HUMAN.H11MO.0.A	1.000000e+00		EGR1 HUMAN.H11MO.0.A	1.000000	
pos 10	164			ZNF384 MA1125.1	1.030280e-01		PRDM6 HUMAN.H11MO.0.C	1.030280e-01		STAT1 MOUSE.H11MO.0.A	0.178240	
pos 11	154			POU3F4 POU 2	1.000000e+00		NFATC1 NFAT 1	1.000000e+00		TEX1 TBX 1	1.000000	
pos 12	144			DNASE 4	7.987890e-01		POU3F1 MA0786.1	7.987890e-01		POU3F1 POU 1	0.798789	
pos 13	119			FOXG1 forkhead 1	7.020530e-01		DNASE 2	7.020530e-01		ONECUT3 CUT 1	0.702053	
pos 14	115			Hoxc10.mouse homeodomain 2	4.020600e-01		HOXC10 homeodomain 2	4.020600e-01		Hoxd9.mouse homeodomain 2	0.402060	
pos 15	92			ZNF384 MA1125.1	3.181230e-02		PRDM6 HUMAN.H11MO.0.C	4.359010e-02		FOXJ3 HUMAN.H11MO.0.A	0.151227	

pattern	NumSeqs	cwm fwd	cwm rev	match0	qval0	match0 logo	match1	qval1	match1 logo	match2	qval2	match2 logo
pos 10	24			TN5 ?	2.478920e-01		OVO_1 HUMAN.H11MO.0.C	7.535210e-01		OVO_1 MOUSE.H11MO.0.C	0.753521	

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	1419			PRDM6 HUMAN.H11MO.0.C	3.207980e-02		STAT1 MOUSE.H11MO.0.A	3.207980e-02		ZNF384 MA1125.1	4.317780e-02	
pos 1	1326			NFAC4 HUMAN.H11MO.0.C	1.000000e+00		NFAC4 MOUSE.H11MO.0.C	1.000000e+00		Sox10.mouse HMG 1	1.000000e+00	
pos 2	1090			EOMES MA0800.1	1.000000e+00		EOMES TBX 1	1.000000e+00		NKX32 HUMAN.H11MO.0.C	1.000000e+00	
pos 3	787			ZNF384 MA1125.1	5.991180e-01		FOXD2 forkhead 1	5.991180e-01		RFX7 RF 2	5.991180e-01	
pos 4	434			GSX2 MA0893.1	1.227900e-01		GSX2 homeodomain 1	1.227900e-01		LHX2 MOUSE.H11MO.0.A	1.227900e-01	
pos 5	298			DRGX homeodomain 1	1.971080e-01		ARX homeodomain 1	1.971080e-01		PHOX2A MA0713.1	1.971080e-01	
pos 6	219			SIX homeodomain 1	8.192750e-03		LHX2 homeodomain 2	1.023000e-02		mix-a MA0021.1	1.023000e-02	
pos 7	149			EMX1 homeodomain 2	1.150210e-02		EMX2 homeodomain 2	1.150210e-02		Hoxa2.mouse homeodomain 1	1.150210e-02	
pos 8	124			PRDM6 HUMAN.H11MO.0.C	8.051370e-02		FOXJ3 HUMAN.H11MO.0.A	8.051370e-02		FOXJ3 MOUSE.H11MO.0.A	8.051370e-02	
pos 9	124			MAZ HUMAN.H11MO.0.A	1.000000e+00		MAZ MOUSE.H11MO.0.A	1.000000e+00		SIX homeodomain 1	1.000000e+00	
pos 10	123			PHOX2A MA0713.1	1.478320e-01		PHOX2A homeodomain 1	1.478320e-01		PROP1 MOUSE.H11MO.0.C	1.478320e-01	
pos 11	27			RREB1 MA0073.1	1.000000e+00		SP5 MOUSE.H11MO.0.C	1.000000e+00		EGR2 HUMAN.H11MO.0.A	1.000000e+00	