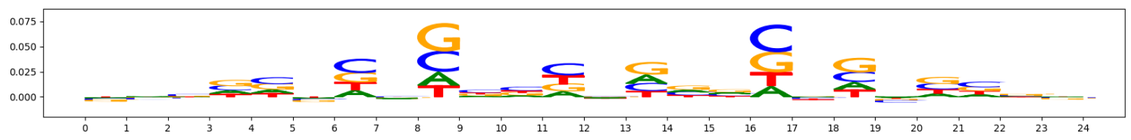
Bias model training and quality check report

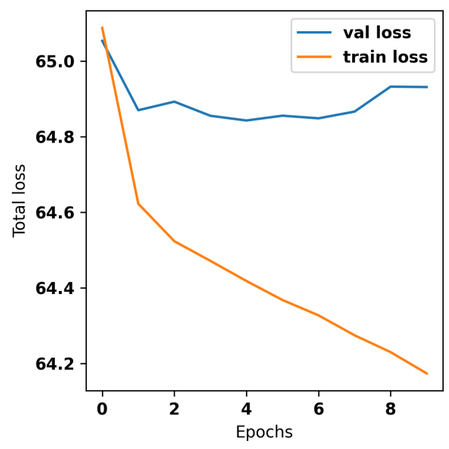
Preprocessing report

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	0.68	0.81	0.39	12.09

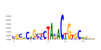
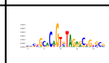
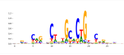
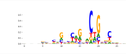
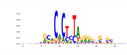
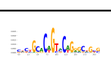
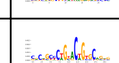
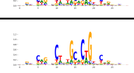
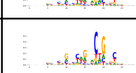
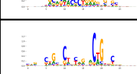
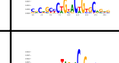
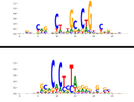
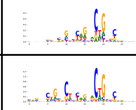
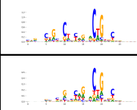
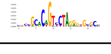
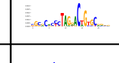
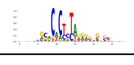
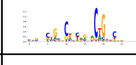
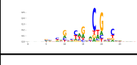
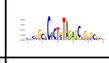
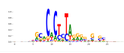
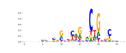
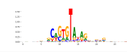
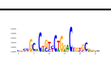
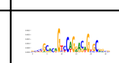
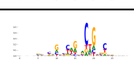
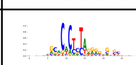
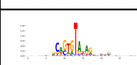
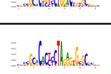
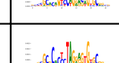
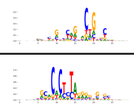
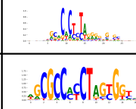
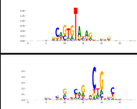
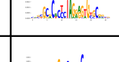
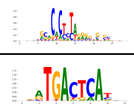
	nonpeaks.median_jsd	nonpeaks.median_norm_jsd	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.78	0.02	0.64	0.08

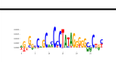
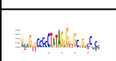
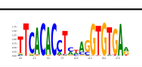
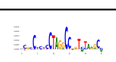
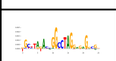
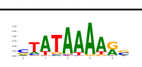
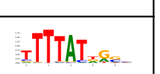
TFModisco motifs learnt from bias model (bias.h5) model

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)

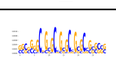
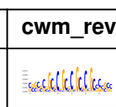
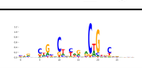
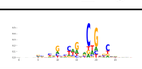
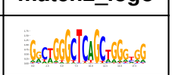
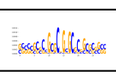
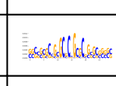
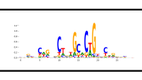
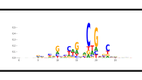
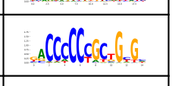
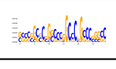
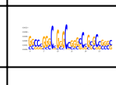
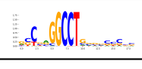
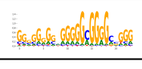
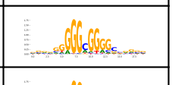
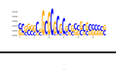
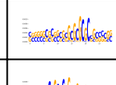
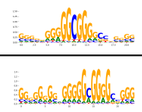
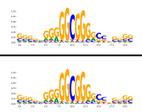
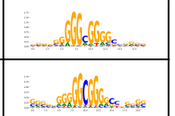
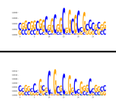
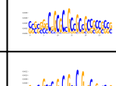
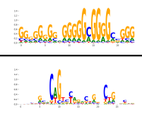
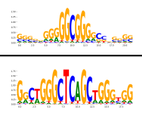
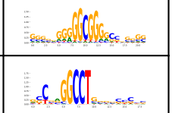
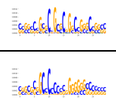
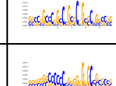
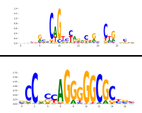
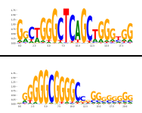
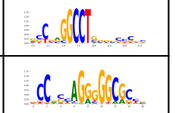
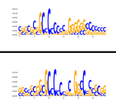
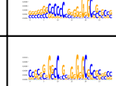
pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos__0	9856			TN5_2	7.010100e-05		TN5_1	1.068740e-04		TN5_3	0.000200	
pos__1	8564			TN5_2	3.801010e-08		TN5_1	4.532620e-08		TN5_7	0.000358	
pos__2	2257			TN5_3	1.767860e-02		TN5_7	4.006500e-02		TN5_1	0.040065	
pos__3	1728			TN5_3	7.637230e-07		TN5_1	2.144860e-03		TN5_4	0.002145	
pos__4	1241			TN5_1	7.391490e-04		TN5_3	3.168080e-02		TN5_4	0.067799	
pos__5	680			TN5_3	7.048660e-10		CTCF_C2H2_1	3.527470e-02		TN5_1	0.035275	
pos__6	606			FOSL1+JUN_MA1128.1	1.000000e+00		JUNB_HUMAN.H11MO.0.A	1.000000e+00		FOS_HUMAN.H11MO.0.A	1.000000	
pos__7	504			TN5_6	4.848480e-04		PAX5_HUMAN.H11MO.0.A	1.668430e-01		ZN322_MOUSE.H11MO.0.B	0.174026	
pos__8	423			TBX21_TBX_6	6.009990e-02		TBX21_TBX_3	6.009990e-02		TN5_3	0.085444	
pos__9	422			TN5_3	7.890980e-03		TN5_1	1.323850e-01		NKX22_MOUSE.H11MO.0.A	0.132385	
pos__10	416			ZNF384_MA1125.1	8.281610e-02		PRDM6_HUMAN.H11MO.0.C	1.142550e-01		STAT1_MOUSE.H11MO.0.A	0.215604	
pos__11	405			TN5_3	2.406420e-04		TBX21_TBX_3	6.125340e-02		TN5_1	0.061253	
pos__12	247			ZNF384_MA1125.1	3.337830e-02		DNASE_2	9.231380e-02		FOXC1_forkhead_1	0.306734	
pos__13	222			FOXC1_forkhead_1	7.527850e-02		ONECUT3_CUT_1	9.203230e-02		ONECUT3_MA0757.1	0.092032	
pos__14	145			FOXB1_forkhead_2	6.522890e-01		FOXB1_MA0845.1	1.000000e+00		FOXB1_forkhead_3	1.000000	
pos__15	141			Tcf12_MA0521.1	1.000000e+00		TFE2_HUMAN.H11MO.0.A	1.000000e+00		TFE2_MOUSE.H11MO.0.A	1.000000	
pos__16	141			DNASE_2	9.148280e-01		CPEB1_RRM_1	9.148280e-01		Arid3b_MA0601.1	1.000000	
pos__17	90			VEZF1_HUMAN.H11MO.0.C	3.152690e-01		ZBT17_HUMAN.H11MO.0.A	3.152690e-01		MAZ_HUMAN.H11MO.0.A	0.315269	
pos__18	62			SOX14_HMG_1	3.361650e-01		TBX21_TBX_3	3.361650e-01		TBX21_TBX_6	0.336165	
pos__19	39			TBX1_TBX_5	5.187510e-02		ZN329_HUMAN.H11MO.0.C	1.000000e+00		T_MA0009.2	1.000000	
pos__20	23			Tcf12_MA0521.1	1.000000e+00		Myog_MA0500.1	1.000000e+00		TFE2_HUMAN.H11MO.0.A	1.000000	

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos__21	23			TBX1_TBX_5	1.000000e+00		NaN	NaN		NaN	NaN	
pos__22	21			TBP_MOUSE.H11MO.0.A	1.000000e+00		TBP_HUMAN.H11MO.0.A	1.000000e+00		HXB13_HUMAN.H11MO.0.A	1.000000	

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	4578			TN5_7	0.021638		TN5_1	0.060915		ZN331_HUMAN.H11MO.0.C	0.074815	
pos__1	4308			TN5_2	0.000396		TN5_1	0.241832		ZIC3_C2H2_1	0.414188	
pos__2	3073			ZFX_MOUSE.H11MO.0.B	0.000507		SP1_MOUSE.H11MO.0.A	0.000843		SP3_HUMAN.H11MO.0.B	0.000843	
pos__3	2780			SP2_HUMAN.H11MO.0.A	0.004209		SP2_MOUSE.H11MO.0.B	0.004209		SP3_HUMAN.H11MO.0.B	0.009132	
pos__4	1996			SP1_MOUSE.H11MO.0.A	0.005426		SP2_HUMAN.H11MO.0.A	0.005426		SP2_MOUSE.H11MO.0.B	0.005426	
pos__5	1822			TN5_8	0.000105		ZN331_HUMAN.H11MO.0.C	0.073823		ZFX_MOUSE.H11MO.0.B	0.164062	
pos__6	749			CTCFL_MOUSE.H11MO.0.A	0.122171		SP1_HUMAN.H11MO.0.A	0.229363		CTCFL_HUMAN.H11MO.0.A	0.280913	
pos__7	393			CTCF_MOUSE.H11MO.0.A	0.000019		CTCF_MA0139.1	0.000019		CTCFL_HUMAN.H11MO.0.A	0.000019	
pos__8	392			TN5_6	0.174220		TN5_8	0.363248		STA5A_MOUSE.H11MO.0.A	0.583670	
pos__9	285			PITX1_homeodomain_3	0.132031		Pitx1_MA0682.1	0.132031		PITX1_homeodomain_2	0.258662	
pos__10	285			Rarb.mouse_nuclearreceptor_2	0.672683		CTCFL_MA1102.1	0.672683		RARG_nuclearreceptor_3	0.672683	
pos__11	277			TN5_6	0.006895		PAX5_HUMAN.H11MO.0.A	0.090715		ZN121_HUMAN.H11MO.0.C	0.162496	
pos__12	123			Tcf12_MA0521.1	1.000000		Myog_MA0500.1	1.000000		TFE2_HUMAN.H11MO.0.A	1.000000	
pos__13	100			BHLHA15_bHLH_1	0.396092		MAX+MYC_MA0059.1	0.396092		OLIG2_MA0678.1	0.396092	
pos__14	61			COT2_HUMAN.H11MO.0.A	0.592170		COT2_MOUSE.H11MO.0.A	0.592170		RARA_MA0730.1	0.592170	
pos__15	50			KLF5_MOUSE.H11MO.0.A	1.000000		GATA2_HUMAN.H11MO.0.A	1.000000		GATA1_HUMAN.H11MO.0.A	1.000000	
pos__16	40			TN5_1	0.397694		TN5_7	0.397694		KLF8_HUMAN.H11MO.0.C	0.750509	
pos__17	38			TN5_7	0.336944		Nfe2l2_MA0150.2	1.000000		KLF5_MOUSE.H11MO.0.A	1.000000	
pos__18	21			NF2L2_HUMAN.H11MO.0.A	0.282132		NFE2_MOUSE.H11MO.0.A	0.282132		ZN667_HUMAN.H11MO.0.C	0.341423	