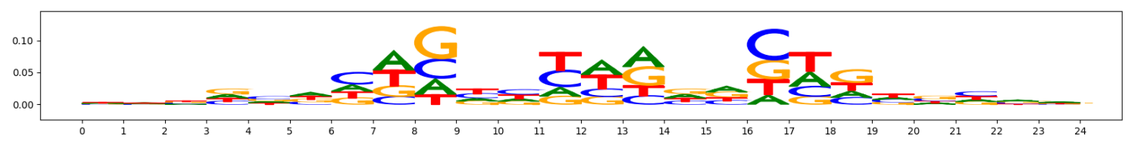
Bias factorized ChromBPNet training and quality check report

Preprocessing report

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



Bias model performance in peaks

Counts Metrics: 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.

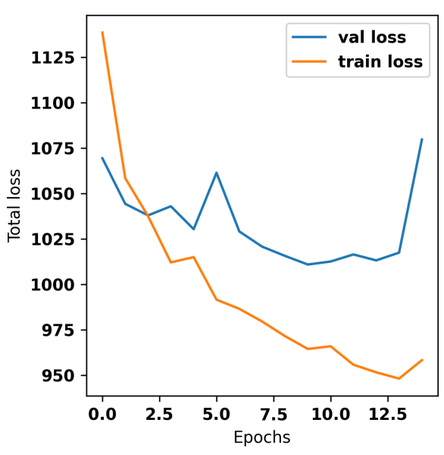
What to do if your pearsonr in peaks is less than -0.3? In the range of -0.3 to -0.5 please be wary of your chrombpnet_wo_bias.h5 TFModisco results showing lots of GC rich motifs (> 3 in the top-10). If this is not the case you can continue using the chrombpnet_wo_bias.h5. If you end up seeing a lot of GC rich motifs it is likely that bias model has learnt a different GC distribution than your GC-content in peaks. If you are transferring a bias model from a different sample you can consider using a different bias model or training a bias model for this sample. If you have trained a bias model for this sample and encounter this you might have to increase 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 pipline will automatically throw an error.

	peaks.pearsonr	peaks.mse
counts_metrics	-0.165711	7.295054

	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.497744	0.243677

Training report

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



ChromBPNet model performance in peaks

Counts Metrics: The pearsonr in peaks should be greater than 0.5 (higher the better). MSE (Mean Squared Error) will be low 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.

	peaks.pearsonr	peaks.mse
counts_metrics	0.65824	0.54448

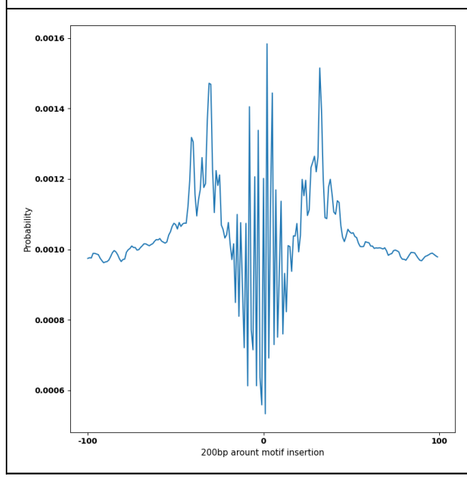
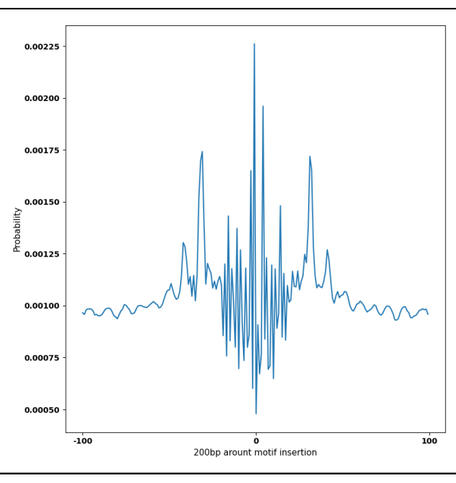
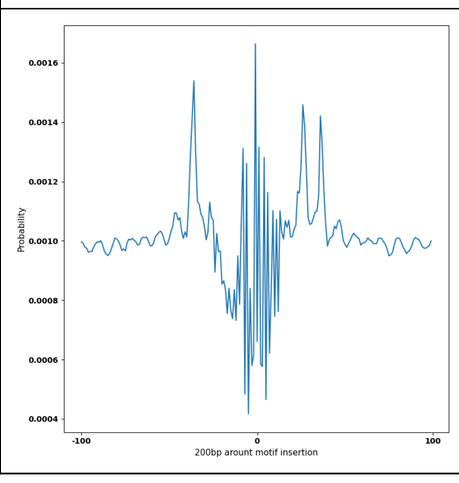
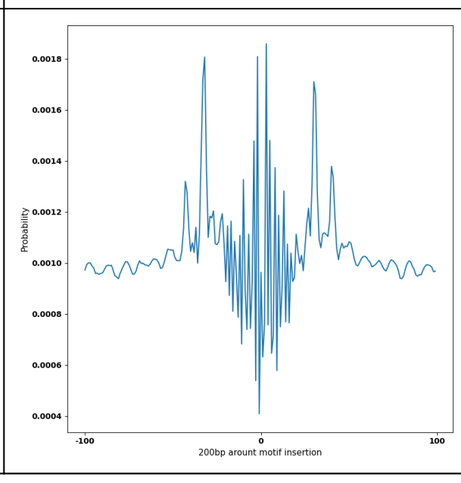
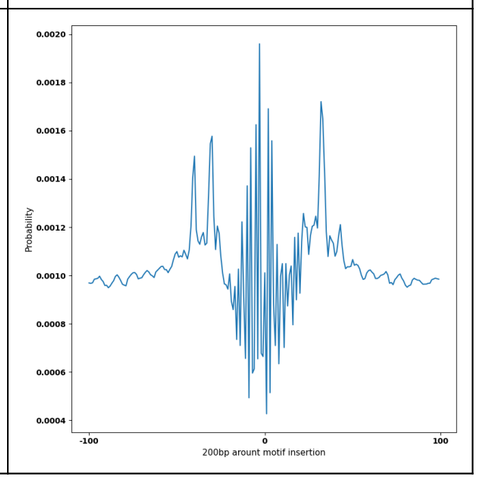
	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.459501	0.295409

ChromBPNet marginal footprints on tn5 motifs

The marginal footprints are the response of the ChromBPNet no bias model to the hetergenous bias motifs. If the bias correction is complete the max of the profiles below should be below 0.003 on all the bias motifs.

For your convenience we calculate here the average of the max of the profiles: 0.002 And the model according to this is **corrected**

What to do if your model looks uncorrected (i.e max of profiles is greater than 0.003)?
Look at the motifs below captured by TFModisco and you should be able to see motifs that closely look like the bias motifs showing incomplete bias correction. This indicates that your bias model was not completely capturing the response of the bias. We recommend that you use a different pre-trained bias model. For more intuition on choosing the correct pre-trained model or retraining your bias model refer to FAQ section in wiki.

tn5 motif 1	tn5 motif 2	tn5 motif 3	tn5 motif 4	tn5 motif 5
				

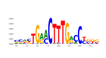
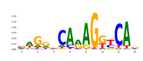
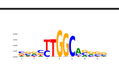
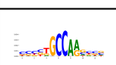
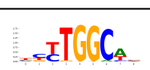
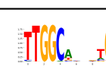
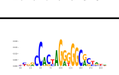
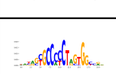
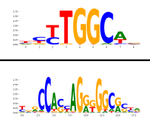
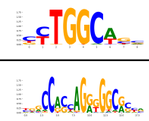
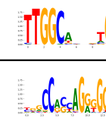
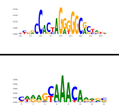
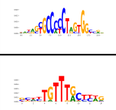
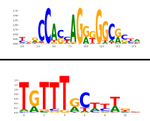
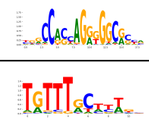
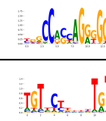
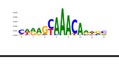
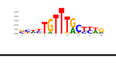
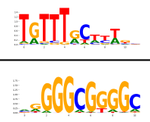
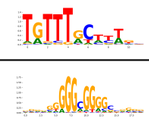
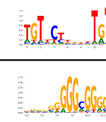
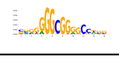
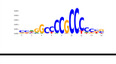
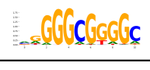
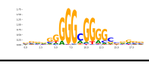
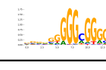
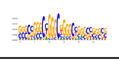
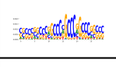
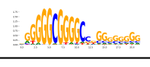
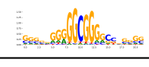
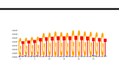
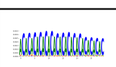
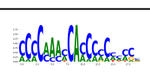
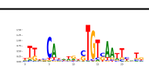
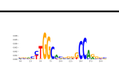
TFModisco motifs learnt from ChromBPNet after bias correction (chrombpnet_nobias.h5) model

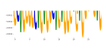
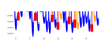
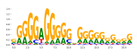
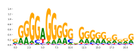
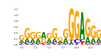
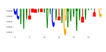
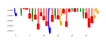
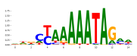
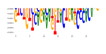
TFModisco motifs generated from profile contribution scores of the ChromBPNet after bias correction model. cwm_fwd, cwm_rev are the forward and reverse complemented consolidated motifs from contribution scores in subset of random peaks. These CWM motifs should be free from any bias motifs and should contain only Transcription Factor (TF) motifs. 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 low (< 0.0001) for most of the closest TF motif hits (i.e indicating that the closest match is the correct match) - this is also generally verifiable by eye as the closest match will look closely like the CWMs (atleast part of it in case of heterodimers). All the motifs in the list should look nothing like the enzyme bias motif.

What to do if you find an obvious bias motif in the list?
This indicates that your bias model was not completely capturing the response of the bias. We recommend that you use a different pre-trained bias

model. For more intuition on choosing the correct pre-trained model or retraining your bias model refer to FAQ section in wiki.

What to do if you find an obvious bias motif in the list?

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_
pos__0	2905			HNF4G_MA0484.1	1.459050e-06		NR1H3_MOUSE.H11MO.0.A	9.166910e-06		HNF4A_nuclearreceptor_4	1.654290e-05	
pos__1	2251			NFIC_MOUSE.H11MO.0.A	1.201570e-03		NFIB_MOUSE.H11MO.0.C	1.430660e-02		NFIA_NFI_1	1.430660e-02	
pos__2	1988			CTCF_MA0139.1	1.867750e-15		CTCF_HUMAN.H11MO.0.A	1.291760e-11		CTCF_MOUSE.H11MO.0.A	3.302560e-10	
pos__3	1805			FOXM1_HUMAN.H11MO.0.A	6.480970e-04		FOXM1_MOUSE.H11MO.0.B	6.480970e-04		ANDR_HUMAN.H11MO.0.A	9.641230e-03	
pos__4	1570			KLF12_HUMAN.H11MO.0.C	1.462570e-04		SP3_HUMAN.H11MO.0.B	1.462570e-04		SP3_MOUSE.H11MO.0.B	1.462570e-04	
pos__5	1373			SP1_HUMAN.H11MO.0.A	3.077050e-05		SP2_HUMAN.H11MO.0.A	3.077050e-05		SP2_MOUSE.H11MO.0.B	3.077050e-05	
pos__6	1328			RREB1_MA0073.1	1.000000e+00		BRAC_MOUSE.H11MO.0.B	1.000000e+00		None	NaN	
pos__7	1226			NFIC_HUMAN.H11MO.0.A	4.721690e-08		NFIA_HUMAN.H11MO.0.C	1.982610e-06		NFIA_MOUSE.H11MO.0.C	1.982610e-06	
pos__8	1191			CEBPA_MA0102.3	1.697470e-04		CEBPA_MOUSE.H11MO.0.A	3.898770e-04		CEBPB_MOUSE.H11MO.0.A	3.898770e-04	
pos__9	732			PRDM6_HUMAN.H11MO.0.C	4.719480e-02		ZNF384_MA1125.1	4.719480e-02		ANDR_HUMAN.H11MO.0.A	1.508780e-01	
pos__10	681			HNF1A_HUMAN.H11MO.0.C	5.603710e-07		HNF1B_MA0153.2	6.773500e-06		HNF1B_homeodomain_1	6.773500e-06	
pos__11	613			NR2C1_HUMAN.H11MO.0.C	1.302750e-04		NR2C1_MOUSE.H11MO.0.C	1.302750e-04		NR1H3_MOUSE.H11MO.0.A	8.955510e-04	
pos__12	612			SP1_HUMAN.H11MO.0.A	9.628350e-07		SP2_HUMAN.H11MO.0.A	9.628350e-07		SP2_MOUSE.H11MO.0.B	9.628350e-07	
pos__13	593			VEZF1_HUMAN.H11MO.0.C	5.328610e-02		MAZ_HUMAN.H11MO.0.A	5.328610e-02		MAZ_MOUSE.H11MO.0.A	5.328610e-02	
pos__14	549			HNF6_HUMAN.H11MO.0.B	5.179930e-03		HNF6_MOUSE.H11MO.0.A	5.179930e-03		CUX2_MOUSE.H11MO.0.C	7.564820e-02	
pos__15	544			IRF3_HUMAN.H11MO.0.B	5.919960e-02		IRF3_MOUSE.H11MO.0.A	5.919960e-02		MAZ_HUMAN.H11MO.0.A	8.682610e-02	
pos__16	434			SP5_MOUSE.H11MO.0.C	1.307870e-01		MAZ_HUMAN.H11MO.0.A	1.307870e-01		MAZ_MOUSE.H11MO.0.A	1.307870e-01	
pos__17	356			RORA_MA0072.1	1.000000e+00		RORC_MA1151.1	1.000000e+00		OLIG2_bHLH_1	1.000000e+00	
pos__18	298			RREB1_MA0073.1	1.000000e+00		KLF9_MA1107.1	1.000000e+00		TBX20_TBX_1	1.000000e+00	
pos__19	250			TN5_7	2.647030e-02		TN5_2	5.094400e-02		ZN341_HUMAN.H11MO.0.C	9.340580e-02	
pos__20	229			SP2_HUMAN.H11MO.0.A	3.179430e-07		SP2_MOUSE.H11MO.0.B	3.179430e-07		SP1_HUMAN.H11MO.0.A	2.641090e-06	
pos__21	204			ZNF384_MA1125.1	4.394330e-02		PRDM6_HUMAN.H11MO.0.C	4.394330e-02		ANDR_HUMAN.H11MO.0.A	1.587060e-01	
pos__22	193			RREB1_MA0073.1	1.180590e-01		KLF9_MA1107.1	1.000000e+00		RUNX2_HUMAN.H11MO.0.A	1.000000e+00	
pos__23	187			FOXB1_forkhead_2	2.330730e-03		FOXC1_forkhead_2	2.330730e-03		FOXC2_forkhead_1	2.330730e-03	
pos__24	165			CUX1_HUMAN.H11MO.0.C	6.867330e-01		CUX1_MOUSE.H11MO.0.C	6.867330e-01		CUX1_CUT_3	6.867330e-01	
pos__25	162			NFAC4_HUMAN.H11MO.0.C	8.008100e-01		NFAC4_MOUSE.H11MO.0.C	8.008100e-01		Pou5f1+Sox2_MA0142.1	1.000000e+00	
pos__26	155			SP1_HUMAN.H11MO.0.A	4.416510e-13		SP3_HUMAN.H11MO.0.B	1.749880e-09		SP3_MOUSE.H11MO.0.B	1.749880e-09	
pos__27	150			E2F4_E2F_1	5.853870e-01		E2F3_E2F_1	1.000000e+00		E2F1_E2F_1	1.000000e+00	
pos__28	149			NFAC4_HUMAN.H11MO.0.C	5.420360e-01		NFAC4_MOUSE.H11MO.0.C	5.420360e-01		NFATC1_NFAT_1	5.420360e-01	
pos__29	128			NR1D2_MOUSE.H11MO.0.A	2.007540e-01		IRF3_HUMAN.H11MO.0.B	2.007540e-01		IRF3_MOUSE.H11MO.0.A	2.007540e-01	
pos__30	125			NFIC_HUMAN.H11MO.0.A	3.769880e-03		EWSR1-FLI1_MA0149.1	2.143350e-02		NFIB_MOUSE.H11MO.0.C	3.258620e-01	
pos__31	122			ZNF76_HUMAN.H11MO.0.C	8.769280e-19		ZN143_HUMAN.H11MO.0.A	1.318390e-17		THA11_HUMAN.H11MO.0.B	1.845570e-13	
pos__32	117			RREB1_MA0073.1	1.000000e+00		VEZF1_HUMAN.H11MO.0.C	1.000000e+00		ZN281_MOUSE.H11MO.0.A	1.000000e+00	
pos__33	116			NRF1_HUMAN.H11MO.0.A	4.246560e-09		NRF1_MOUSE.H11MO.0.A	4.246560e-09		NRF1_MA0506.1	6.137980e-05	
pos__34	109			MAFK_bZIP_3	7.352050e-01		MAFK_bZIP_1	7.352050e-01		Arid5a_MA0602.1	7.352050e-01	
pos__35	107			FOXM1_HUMAN.H11MO.0.A	4.336300e-01		FOXM1_MOUSE.H11MO.0.B	4.336300e-01		NR2E1_nuclearreceptor_2	4.336300e-01	
pos__36	102			ZNF263_MA0528.1	6.050220e-08		SP5_MOUSE.H11MO.0.C	1.121610e-06		ZN263_HUMAN.H11MO.0.A	5.995220e-06	
pos__37	95			GRHL1_CP2_1	6.230610e-01		Foxc1.mouse_forkhead_2	6.230610e-01		ERG_ETS_4	6.230610e-01	
pos__38	91			ZN121_HUMAN.H11MO.0.C	1.000000e+00		SREBF1_MA0595.1	1.000000e+00		PAX5_HUMAN.H11MO.0.A	1.000000e+00	
pos__39	82			CTCF_MOUSE.H11MO.0.A	9.375630e-04		CTCF_MA0139.1	2.022440e-03		CTCF_HUMAN.H11MO.0.A	2.022440e-03	
pos__40	78			RUNX1_MOUSE.H11MO.0.A	7.603560e-01		RUNX1_MA0002.2	7.603560e-01		NFYC_MOUSE.H11MO.0.B	7.603560e-01	
pos__41	73			NFIA_NFI_1	1.498450e-01		Creb3l2.mouse_bZIP_2	1.498450e-01		CREB3L1_bZIP_3	1.498450e-01	
pos__42	70			HIC2_C2H2_1	4.455650e-01		HIC2_MA0738.1	4.455650e-01		USF1_HUMAN.H11MO.0.A	4.455650e-01	
pos__43	70			ATF3_HUMAN.H11MO.0.A	4.969720e-03		ATF1_MOUSE.H11MO.0.B	4.969720e-03		CREM_HUMAN.H11MO.0.C	4.969720e-03	
pos__44	65			VDR_nuclearreceptor_1	4.471510e-02		VDR_HUMAN.H11MO.0.A	4.471510e-02		VDR_MOUSE.H11MO.0.A	4.471510e-02	
pos__45	57			NFATC1_NFAT_1	4.527940e-01		RXRA_MOUSE.H11MO.0.A	4.527940e-01		RARA_MOUSE.H11MO.0.A	4.951260e-01	
pos__46	53			RARG_nuclearreceptor_1	8.122110e-01		SOX15_HMG_1	8.122110e-01		Nr2e1.mouse_nuclearreceptor_2	8.122110e-01	
pos__47	50			FOXJ2_HUMAN.H11MO.0.C	1.531350e-04		FOXJ2_MOUSE.H11MO.0.C	1.531350e-04		FOXC2_forkhead_2	1.470340e-03	
pos__48	43			ZN331_HUMAN.H11MO.0.C	1.000000e+00		ZNF524_C2H2_1	1.000000e+00		ZNF524_C2H2_2	1.000000e+00	
pos__49	38			FOXO1_HUMAN.H11MO.0.A	1.000000e+00		None	NaN		None	NaN	
pos__50	27			NFIB_MOUSE.H11MO.0.C	1.000000e+00		DNASE_5	1.000000e+00		CLOCK_HUMAN.H11MO.0.C	1.000000e+00	
pos__51	27			Gfi1_MA0038.1	1.000000e+00		Nr5a2_MA0505.1	1.000000e+00		ZN418_HUMAN.H11MO.0.C	1.000000e+00	
pos__52	26			ZNF8_HUMAN.H11MO.0.C	1.686260e-01		FOXD3_forkhead_2	1.686260e-01		FOXI1_MA0042.2	1.686260e-01	
pos__53	24			ZN257_HUMAN.H11MO.0.C	1.000000e+00		FOXO3_forkhead_3	1.000000e+00		ZNF75A_C2H2_1	1.000000e+00	

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_
neg__0	57			MAZ_HUMAN.H11MO.0.A	7.887650e-06		MAZ_MOUSE.H11MO.0.A	7.887650e-06		VEZF1_HUMAN.H11MO.0.C	7.887650e-06	
neg__1	48			MEF2C_MA0497.1	1.000000e+00		Arid3b_MA0601.1	1.000000e+00		MEF2C_MOUSE.H11MO.0.A	1.000000e+00	
neg__2	32			FOXA3_HUMAN.H11MO.0.B	6.674050e-01		FOXA3_MOUSE.H11MO.0.A	6.674050e-01		FOXM1_HUMAN.H11MO.0.A	6.674050e-01