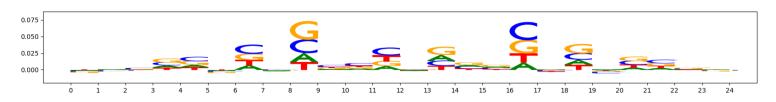
Bias model training and quality check report

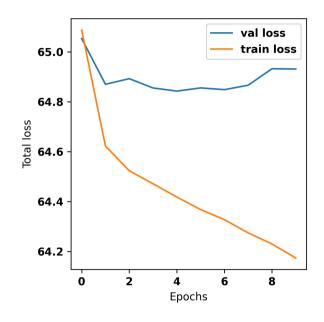
Preprocessing report

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 <u>chrombpnet training</u> 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	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 <u>FAQ</u>)

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos0	9856	ese.c.sec]hedybec.		TN5_2	7.010100e-05	for the ford a few.	TN5_1	1.068740e-04		TN5_3	0.000200	
pos1	8564		Esc. esc. Calzaladase	TN5_2	3.801010e-08		TN5_1	4.532620e-08		TN5_7	0.000358	Sec. Sec. Sec. Sec. Sec. Sec. Sec. Sec.
pos2	2257			TN5_3	1.767860e-02		TN5_7	4.006500e-02		TN5_1	0.040065	······································
pos3	1728			TN5_3	7.637230e-07		TN5_1	2.144860e-03	и. 	TN5_4	0.002145	
pos4	1241			TN5_1	7.391490e-04		TN5_3	3.168080e-02		TN5_4	0.067799	
pos5	680			TN5_3	7.048660e-10		CTCF_C2H2_1	3.527470e-02	_z z <mark>i)</mark> J2a,TJ7-9)))2a,	TN5_1	0.035275	
pos6	606	- cher hr. the can	Encrea borth all taken	FOSL1+JUN_MA1128.1	1.000000e+00	TGActcA	JUNB_HUMAN.H11MO.0.A	1.000000e+00	TGAGTCA	FOS_HUMAN.H11MO.0.A	1.000000	T GAGTCA
pos7	504	httl. to began		TN5_6	4.848480e-04	astraligned in the strate	PAX5_HUMAN.H11MO.0.A	1.668430e-01		ZN322_MOUSE.H11MO.0.B	0.174026	6.66 I6.14 6.199.70
pos8	423			TBX21_TBX_6	6.009990e-02	TcACACct_agaCCTCTCA	TBX21_TBX_3	6.009990e-02		TN5_3	0.085444	
pos9	422		=	TN5_3	7.890980e-03		TN5_1	1.323850e-01		NKX22_MOUSE.H11MO.0.A	0.132385	T-SAGTO
pos10	416			ZNF384_MA1125.1	8.281610e-02		PRDM6_HUMAN.H11MO.0.C	1.142550e-01	,,, GAAAA	STAT1_MOUSE.H11MO.0.A	0.215604	
pos11	405			TN5_3	2.406420e-04		TBX21_TBX_3	6.125340e-02	Tealling agitited	TN5_1	0.061253	
pos12	247			ZNF384_MA1125.1	3.337830e-02		DNASE_2	9.231380e-02	10- 10- 10- 10- 10- 10- 10- 10- 10- 10-	FOXC1_forkhead_1	0.306734	A**IAAAIAA**
pos13	222			FOXC1_forkhead_1	7.527850e-02	essigadese	ONECUT3_CUT_1	9.203230e-02	, AAAATCAATA,	ONECUT3_MA0757.1	0.092032	AAAAATCAATAA
pos14	145		AND	FOXB1_forkhead_2	6.522890e-01	TATELANTITACATA	FOXB1_MA0845.1	1.000000e+00	TATGTAATA	FOXB1_forkhead_3	1.000000	TATGIANTA
pos15	141	Enterment all the for the second	entropy of the second of the s	Tcf12_MA0521.1	1.000000e+00		TFE2_HUMAN.H11MO.0.A	1.000000e+00		TFE2_MOUSE.H11MO.0.A	1.000000	
pos16	141	TATLATIATIATIATIATIATIATIATIATIATIATIATIATIA	THE REPORT OF	DNASE_2	9.148280e-01	10 10 10 10 10 10 10 10 10 10 10 10 10 1	CPEB1_RRM_1	9.148280e-01	AAAA	Arid3b_MA0601.1	1.000000	
pos17	90		dre,khe,khe,khe,khe,khe,l	VEZF1_HUMAN.H11MO.0.C	3.152690e-01	SESSES SWARSSELSES	ZBT17_HUMAN.H11MO.0.A	3.152690e-01		MAZ_HUMAN.H11MO.0.A	0.315269	COMANGE GEESESSES
pos18	62		Juney Cally Jaroll & Con	SOX14_HMG_1	3.361650e-01	ACAATA_CATTG	TBX21_TBX_3	3.361650e-01	Teologia activited	TBX21_TBX_6	0.336165	TsACACci_ogAGGIGIGA
pos19	39	Lessiell less for]	TBX1_TBX_5	5.187510e-02	TCACACCI _{seasc} utura.	ZN329_HUMAN.H11MO.0.C	1.000000e+00	CIGUAICEAGCE_I.CTOA	T_MA0009.2	1.000000	T ₌₀ CAC,IA_GTUI5A
pos20	23	. 6. ALACTOTAGE GT TEACH	Effletten al attraction of	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
pos21	23	Essler and Wisser he	Excly reself the frace of	TBX1_TBX_5	1.000000e+00	TICKIACCI - ACTIVICA.	NaN	NaN		NaN	NaN	
pos22	21		_bergh_let]theybee	TBP_MOUSE.H11MO.0.A	1.000000e+00	TATAAA As	TBP_HUMAN.H11MO.0.A	1.000000e+00	STATAAAASS	HXB13_HUMAN.H11MO.0.A	1.000000	Ţ ĪĪTAI _{zSe}

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
pos0	4578]].	TN5_7	0.021638	sel fe see office	TN5_1	0.060915		ZN331_HUMAN.H11MO.0.C	0.074815	GSCTEGIC CALCTEGS2SS
pos1	4308	seculal filterese		TN5_2	0.000396		TN5_1	0.241832		ZIC3_C2H2_1	0.414188	<u>_^(((((,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,</u>
pos2	3073			ZFX_MOUSE.H11MO.0.B	0.000507		SP1_MOUSE.H11MO.0.A	0.000843	Ge. GE. GECHENCLE, GGG	SP3_HUMAN.H11MO.0.B	0.000843	
pos3	2780	Esenci Millanssonne].	SP2_HUMAN.H11MO.0.A	0.004209	Esser Scould See Fores	SP2_MOUSE.H11MO.0.B	0.004209		SP3_HUMAN.H11MO.0.B	0.009132	
pos4	1996		-	SP1_MOUSE.H11MO.0.A	0.005426	<u>çe ce, ce çççîlçîlçîle, ççç</u>	SP2_HUMAN.H11MO.0.A	0.005426		SP2_MOUSE.H11MO.0.B	0.005426	eseo coomic co
pos5	1822	Essects Hellipselses	makahlalidaa	TN5_8	0.000105		ZN331_HUMAN.H11MO.0.C	0.073823	Geeteller Contractions	ZFX_MOUSE.H11MO.0.B	0.164062	
pos6	749	seas Wars half issue		CTCFL_MOUSE.H11MO.0.A	0.122171		SP1_HUMAN.H11MO.0.A	0.229363		CTCFL_HUMAN.H11MO.0.A	0.280913	cC_ <mark>66066665</mark> _
pos7	393	Excepted to a list second].	CTCF_MOUSE.H11MO.0.A	0.000019		CTCF_MA0139.1	0.000019	z-sulectilise forses	CTCFL_HUMAN.H11MO.0.A	0.000019	
pos8	392	and the lace of the and	= pace (bl bal align	TN5_6	0.174220	ait.athdddilithianita.ata	TN5_8	0.363248		STA5A_MOUSE.H11MO.0.A	0.583670	TTC580GAA_
pos9	285	where a della marke	for the second second	PITX1_homeodomain_3	0.132031	_TAAtcc_	Pitx1_MA0682.1	0.132031	TAATCC_	PITX1_homeodomain_2	0.258662	
pos10	285	, approxille chill	Cide Chilesee of som	Rarb.mouse_nuclearreceptor_2	0.672683	ACTCA	CTCFL_MA1102.1	0.672683		RARG_nuclearreceptor_3	0.672683	"A <mark>GGTCARAGGTCA</mark>
pos11	277	for the for th	${}_{\mu}{}_{\alpha,\alpha}{}_{\alpha}{}$	TN5_6	0.006895	asistji.dd/d/ddiaipsta	PAX5_HUMAN.H11MO.0.A	0.090715		ZN121_HUMAN.H11MO.0.C	0.162496	CTUGCCAACA_AUG_AUACC
pos12	123	Lice of Calib. Gase	south the state	Tcf12_MA0521.1	1.000000		Myog_MA0500.1	1.000000	Recage TG	TFE2_HUMAN.H11MO.0.A	1.000000	
pos13	100	endre faller i drifte	ecd, hi ad the lake	BHLHA15_bHLH_1	0.396092	SCATATG	MAX+MYC_MA0059.1	0.396092	_{≠A=} CACGTGG _∓	OLIG2_MA0678.1	0.396092	A-CATATG-I
pos14	61	seed lacks blace	<u>ercheledecthers</u>	COT2_HUMAN.H11MO.0.A	0.592170	LAGAGGICA	COT2_MOUSE.H11MO.0.A	0.592170	AGGTCA	RARA_MA0730.1	0.592170	a <mark>GGTCA, , , , aaGGTCA</mark>
pos15	50	ended etchood	<u> <u>kaldelekter</u>ete</u>	KLF5_MOUSE.H11MO.0.A	1.000000		GATA2_HUMAN.H11MO.0.A	1.000000		GATA1_HUMAN.H11MO.0.A	1.000000	
pos16	40	merfel eff. al fil sprey	program he have been an	TN5_1	0.397694		TN5_7	0.397694		KLF8_HUMAN.H11MO.0.C	0.750509	Ç<u>A</u>ÇĢÛĢÛ<u>T</u>G
pos17	38	free, a likelike e. a. o	<u>elessillatellit</u> son	TN5_7	0.336944		Nfe2l2_MA0150.2	1.000000	ŢŢŢ	KLF5_MOUSE.H11MO.0.A	1.000000	<u><u></u> ₽₽₽₽₽<mark>₽</mark>₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽</u>
pos18	21	englieler flen	accut contraction	NF2L2_HUMAN.H11MO.0.A	0.282132	TÇCTGAGTCA ı	NFE2_MOUSE.H11MO.0.A	0.282132	a <mark>tçactcaçca_</mark> ~*	ZN667_HUMAN.H11MO.0.C	0.341423	FIRE LICONNELL