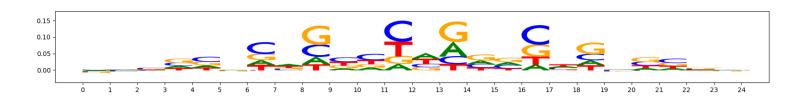
## Bias model training and quality check report

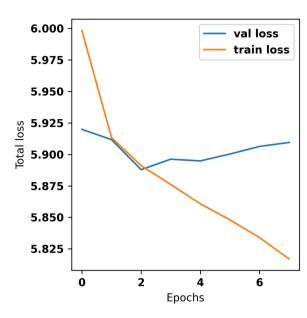
### **Preprocessing report**

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 <u>FAQ</u> section in wiki. If the value is less than -0.5 the <u>chrombpnet training</u> will automatically throw an error.

	nonpeaks.pearsonr	nonpeaks.pearsonr		peaks.pearsonr	peaks.mse		
counts_metrics	-0.19	-0.19		-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 <u>FAQ</u> 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  $\overline{\text{FAQ}}$ )

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos0	7428	-cerlettedhere.	-cessaldsaldselvesc.	TN5_2	2.288990e-05	. se talesto e.	TN5_1	0.000088	Le see ste	TN5_8	0.011999	
pos1	4242			TN5_1	2.355870e-07		TN5_3	0.000004	se villedane e s	TN5_2	0.002699	s.s. G. Accord
pos2	4135	-cascettles liles bases	Lega Note Tagage	TN5_6	2.403020e-04	aid Middle Mineralia and	TN5_8	0.034761		TN5_3	0.078611	
pos3	3709	- sochechechecheles	enceled his weeks extence,	TN5_2	7.359000e-09	ese fracció es	TN5_1	0.003010		TN5_4	0.007352	
pos4	3258	Les Challed Viter Les Const.	Lakeloge Wateless	TN5_3	1.298910e-09	ssoluteless s es	TN5_1	0.000113	L Comment	TN5_7	0.018416	se fre line office
pos5	2750	Leve for Misser form		TN5_3	8.151590e-04	ssolutions s	TN5_1	0.009594	E See Ale Se	TN5_4	0.016195	coen la
pos6	598	-, acre front be so from a	- crle bayedle liciec	TN5_1	6.257520e-04	\$ 5.58 At C	TN5_2	0.000626	. Ess fractions e.s.	TN5_3	0.001863	Scotton Caree S. e.
pos7	515	= ch. letthrochec	Madalisa	TN5_3	9.802630e-05	se telledam s ex	TN5_1	0.000610		ZN554_HUMAN.H11MO.0.C	0.178798	eC CaCeVa Tec
pos8	438		-c.knoljeblodocika,	TN5_1	9.919950e-03	\$ 558 At C	TN5_2	0.027255	. Ess fraction s.s.	TN5_3	0.048034	Scotton Carrows & sec
pos9	304			PRDM6_HUMAN.H11MO.0.C	8.247210e-02		ZNF384_MA1125.1	0.082472	AAAAA	STAT1_MOUSE.H11MO.0.A	0.236510	0250HA . CHINET A.C
pos10	92	Lxx_brd_tablobropac_	Tracket III We be	TN5_6	5.732510e-02		FOXB1_forkhead_1	1.000000	ead <mark>TGAÇAÇ</mark> eçe	ZN121_HUMAN.H11MO.0.C	1.000000	CTGGCAACA_AGc.AGACC
pos11	84	IGYABAARARA.	AQATETTETS A	HOXC12_homeodomain_1	8.251390e-02	<b>AAATAA</b>	HOXD12_homeodomain_1	0.082514	ĘŢĠĄŢĠĄĠ	HOXD12_homeodomain_4	0.082514	STAATAAS
pos12	34	====IACACA J. Laboro Lac.	Focherety habyetet faces	TN5_3	5.717810e-02		NR2F1_MA0017.2	0.057178	EACUTCA_ G	NR2F1_nuclearreceptor_4	0.057178	==eeQUTCA§

**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 <u>FAQ</u> 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	7907	instituted to Winisters.	r.en.Trial Problem Array	TN5_2	3.311020e-04	so frakty	TN5_1	3.551440e-04		TN5_4	1.852220e-02	Com Ade
pos1	5058	xzzedádo TTOMMO TTEADAZOX		DNASE_2	1.000000e+00	10. 10. 10. 10. 10. 10. 10. 10. 10. 10.	LHX3_HUMAN.H11MO.0.C	1.000000e+00	A.TAATTA	None	NaN	
pos2	3462	azo Tyráli nali olbadad, Isra	m.a.T.tacTack o helitary.ca.	TN5_4	1.534940e-02	Coette Ages	TN5_5	1.534940e-02	cord Les	TN5_2	1.427680e-01	s.s. Carlos
pos3	1592			ZNF384_MA1125.1	6.524530e-02	AAAAA	PRDM6_HUMAN.H11MO.0.C	6.524530e-02		FOXJ3_HUMAN.H11MO.0.A	2.204240e-01	*IĞİİİAİ*Tİİ*
pos4	1017	a emercijani likelijanijam	Indeption of some	TN5_2	3.199300e-04	e see to restrict see	TN5_4	7.584600e-04	. Cott	TN5_5	7.584600e-04	Court Age
pos5	998	9. Id-90.T9T/WIED days a describ	Treetitife HATOAbanib.x	TN5_8	4.318060e-02	& Mercycle, C18	TN5_2	4.318060e-02	· se factorises	TN5_4	4.662670e-02	Cott Ade
pos6	992	Topa AAA AAT I AAAA AA AA AA AA AA AA AA AA AA AA AA	يرجيه المستوالية المستوادية المستودية المستوادية المستوادية المستوادية المستوادية المستوادية المستودية	PRDM6_HUMAN.H11MO.0.C	1.637360e-01	eGAAAee	ONECUT3_CUT_1	1.637360e-01	_aaaAATCaATAa_	ONECUT3_MA0757.1	1.637360e-01	_AAAAATCAATA
pos7	951	amii thii hidin	Exist Ist Ist Individua	TN5_4	4.033330e-02	CSETE ALONG CO.	TN5_5	4.033330e-02	COTT Los you	TN5_7	2.249750e-01	s.s. La s.s. Order
pos8	792	Enzurittelittelittelitte	المعروبال والخالخالجالجاموم	ZNF384_MA1125.1	1.005810e-01	AAAAAA	SRY_HUMAN.H11MO.0.B	1.005810e-01	TTGT_T.~	SRY_MOUSE.H11MO.0.B	1.005810e-01	_xIIGIT.
pos9	628			ZNF384_MA1125.1	1.505310e-03	AAAAAA	PRDM6_HUMAN.H11MO.0.C	4.278660e-01		FOXJ3_HUMAN.H11MO.0.A	4.278660e-01	* ON TALETIL
pos10	596	Treathe late los - a	في وسيرا والتبطيرات ويورون	TN5_6	6.803010e-24	sist, silada dilikulisadisis	ZSC31_HUMAN.H11MO.0.C	9.422140e-01	*ccollollority	TN5_8	9.422140e-01	\$0. Determine \$18
pos11	381		fectal fall daries in	TN5_6	4.329560e-04	iii, siladd dddinainai	TN5_7	1.089260e-01	Le see tres offers	NKX25_MOUSE.H11MO.0.A	6.389360e-01	====\$AGTG
pos12	365	grantistellis kellesati		NKX25_MOUSE.H11MO.0.A	2.713330e-01		NKX21_MOUSE.H11MO.0.A	2.713330e-01		TN5_4	2.713330e-01	Cocto Agas
pos13	309	Envelopping the properties of	= cccn/m/m/micc	RREB1_MA0073.1	1.000000e+00	cocheles (Veces es	BRAC_MOUSE.H11MO.0.B	1.000000e+00	.IIQxp.s	None	NaN	
pos14	215	Tanasall Mirashilkan	= erzeMylecripyliceoebocv	RREB1_MA0073.1	1.509250e-01	ZŢŹſŦĬŦĔſŊĬŢĬŸŢĔŦĔĒ	TN5_4	4.970310e-01	. Colti dec.	TN5_5	4.970310e-01	Cocto Agos
pos15	174	أبعرهم أبية الإناساليال تربي	Indeptitute	ZNF384_MA1125.1	3.464990e-02		FOXJ3_HUMAN.H11MO.0.A	3.092960e-01	<b>IGIIIAI</b>	FOXJ3_MOUSE.H11MO.0.A	3.092960e-01	#IĞİİTATETİL
pos16	128	إبابيا بهايها يهابها	<u></u>	ZN502_HUMAN.H11MO.0.C	2.924840e-07	GAAT <mark>O</mark> GAAT-GAATO <mark>GAA</mark> T-	SMCA5_MOUSE.H11MO.0.C	2.191630e-06	TGGALT SAATGGAA	ZN394_HUMAN.H11MO.0.C	9.933130e-06	sdergeder evele
pos17	102	<u></u>	<u></u>	TN5_6	8.559250e-15	aid, teleber de de de de de de de de de de de de de	TN5_8	9.394500e-01		ZSC31_HUMAN.H11MO.0.C	9.394500e-01	**č°ÝČÝČČȰ***
pos18	83	eleteteteleteletereteleter	talda.adddaaaa	RREB1_MA0073.1	1.000000e+00	ECCLOHOE NECESTRATE	BRAC_MOUSE.H11MO.0.B	1.000000e+00	II. A. xy. 8 Medarix 1.	None	NaN	
pos19	70	E-re-trady Collection - construction	<u> </u>	FEZF1_HUMAN.H11MO.0.C	8.828710e-02	~ <u>IQ</u> ësël <u>I</u> I	ZN667_HUMAN.H11MO.0.C	1.000000e+00	<sub>5.752</sub> V <sub>5.</sub> I <sub>1.646</sub> MVCC. C	ZN264_HUMAN.H11MO.0.C	1.000000e+00	*Tro-ball-karihile-éair
pos20	51	cilyoft, Whateroph survey.	= weet I gradulti lytigelis	MEIS2_MEIS_1	1.000000e+00	#TGACAGETGTCAA	ZN350_HUMAN.H11MO.0.C	1.000000e+00	****** <u>1</u> 1*1*****	MEF2B_MA0660.1	1.000000e+00	<sub>*</sub> CTA <sub>X</sub> AAATAÇ <sub>*</sub>
pos21	29		Trest of the transfer of	TN5_6	5.545980e-04	sist, stilladd ddddinaethasta	PAX9_MA0781.1	3.572190e-01	e <mark>JTCACÇC</mark> AT <u>eA</u> JÇs.	PAX9_PAX_1	3.572190e-01	e <mark>(]CAC(</mark> CA]EA_]G <sub>E</sub> _
pos22	27	alakaliaka lighiriphalisis	econolistality dus des	RREB1_MA0073.1	1.000000e+00	coclehe (yelete ee	ZN281_MOUSE.H11MO.0.A	1.000000e+00	**************************************	SRBP2_HUMAN.H11MO.0.B	1.000000e+00	st@_st@a_
pos23	25	actishetathetahketahtau	red diddidddda	RREB1_MA0073.1	1.708960e-01	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	EGR2_HUMAN.H11MO.0.A	9.264730e-01	s_s_s_G, (TG())_\$0;s	BRAC_MOUSE.H11MO.0.B	1.000000e+00	.ias