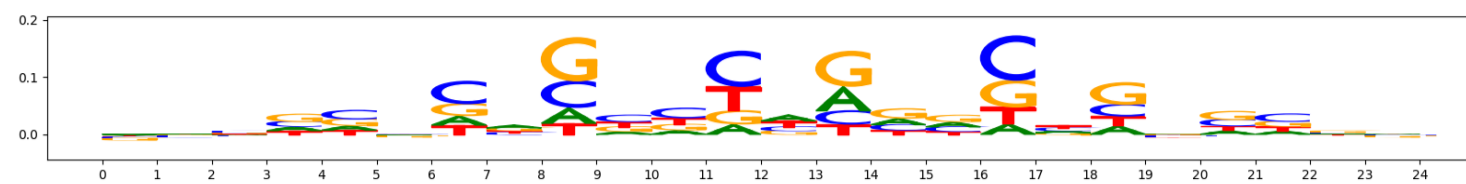


Bias factorized ChromBPNet training and quality check report

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.

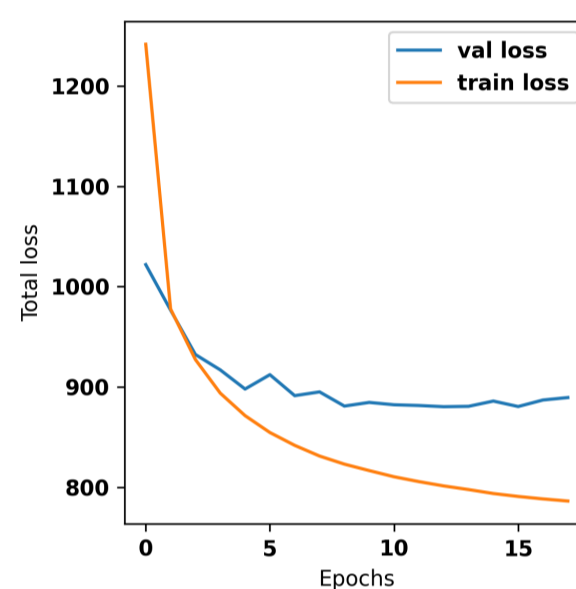
Profile Metrics: Median JSD (Jensen Shannon Divergence between observed and predicted) lower the better. Median norm JSD is median of the min-max normalized JSD where min JSD is the worst case JSD i.e JSD of observed with uniform profile and max JSD is the best case JSD i.e 0. Median norm JSD is higher 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 pipeline will automatically throw an error.

	peaks.pearsonr	peaks.mse
counts_metrics	0.188456	15.097933
	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.456584	0.303574

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	0.778526	0.646662
	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.371181	0.429503

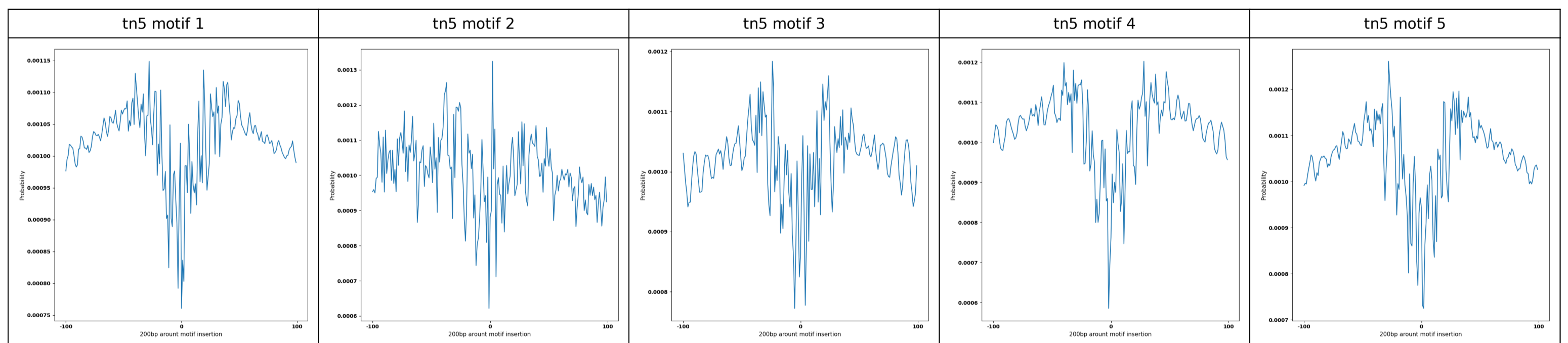
ChromBPNet marginal footprints on tn5 motifs

The marginal footprints are the response of the ChromBPNet no bias model to the heterogenous 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: 0.001 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.



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

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
neg_14	36			NFIC_HUMAN.H11MO.0.A	5.420420e-02		MXI1_HUMAN.H11MO.0.A	1.624110e-01		MXI1_MOUSE.H11MO.0.A	1.624110e-01	
neg_15	31			EWSR1-FLI1_MA0149.1	8.923250e-03		FLI1_HUMAN.H11MO.0.A	3.422790e-01		ETS2_HUMAN.H11MO.0.B	4.636750e-01	