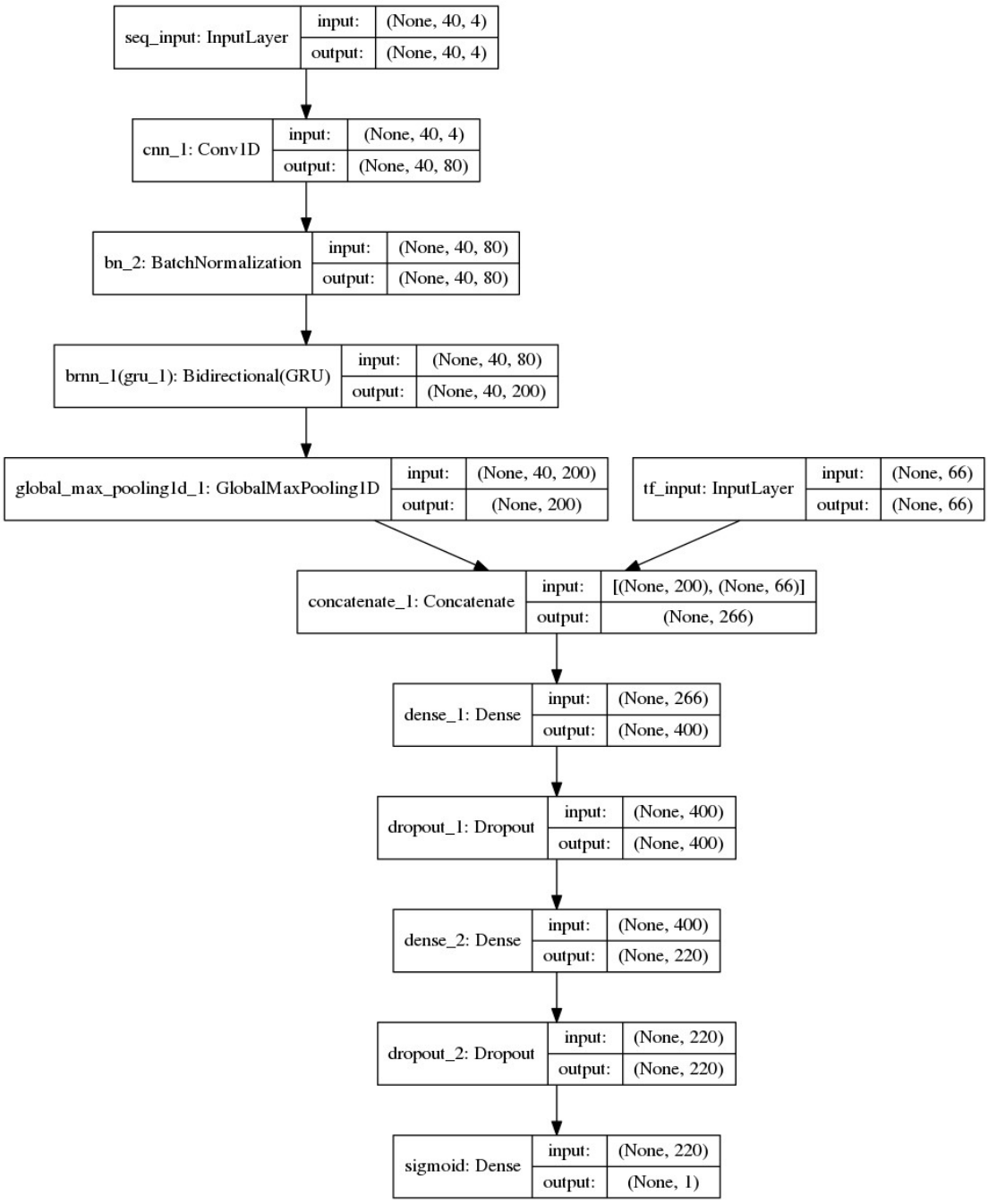
SemanticBI: quantifying intensities of transcription factor-DNA binding by learning from an ensemble of experiments

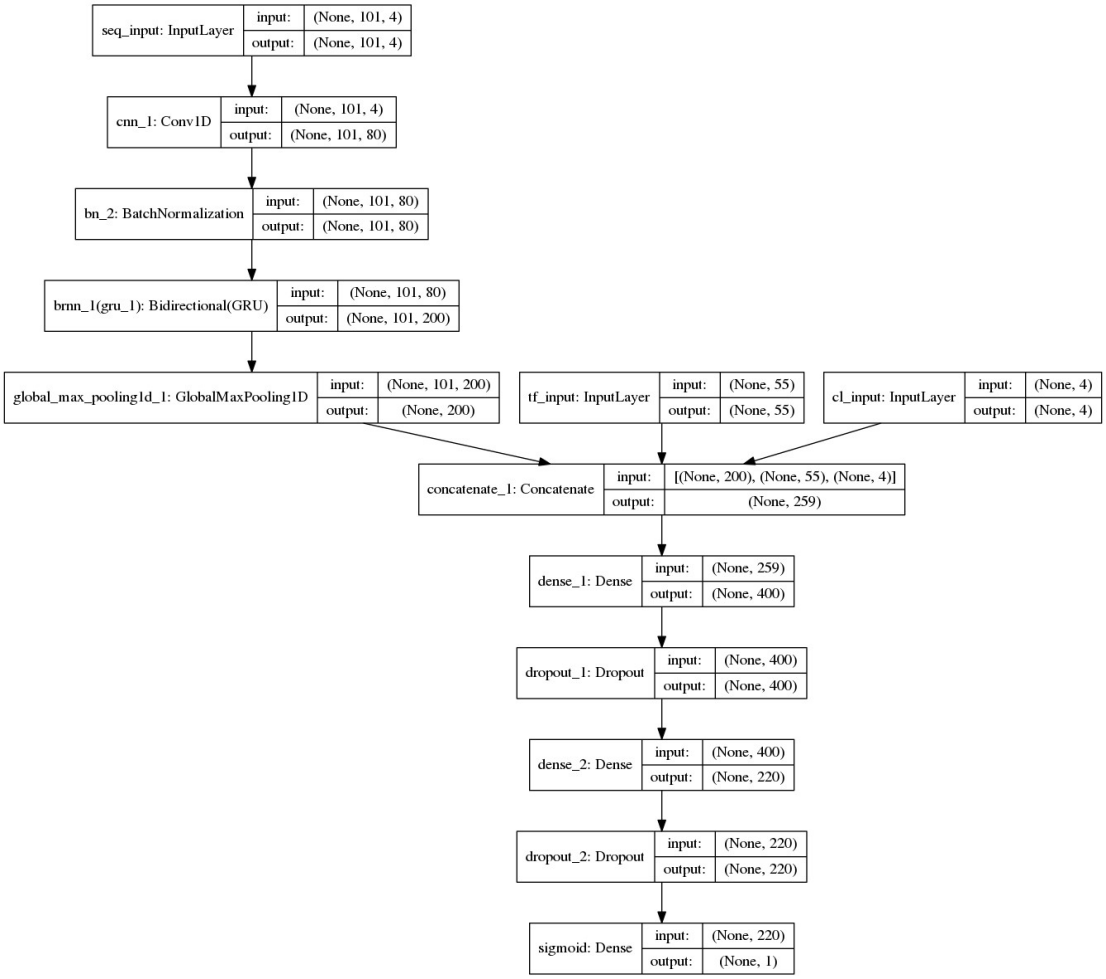
Supplementary Information 1. The hyperparameter configuration and model structure of SemanticBI. (A) Table S1 is the hyperparameter configuration of SemanticBI used on both PBM_66 data and Chip-seq_83 data. (B) Figure S1 is the model structure of SemanticBI for PBM_66 data. (C) Figure S2 is the model structure of SemanticBI for ChIP-seq_83 data.

Supplementary Information 1 - Table S1

Type of hyperparameter	Value of hyperparameter
CNN, LSTM, dense kernel initializer	Glorot_uniform
CNN, first dense, second dense activation	ReLU
CNN filters	80
CNN kernel size	9
CNN padding	Same padding
BatchNormalization momentum	0.99
LSTM units	200
LSTM recurrent activation	Hard sigmoid
LSTM recurrent initializer	Orthogonal
BLSTM merge mode	Sum
First dense units	400
Second dense units	220
Dropout rate	0.25
Third dense units	1
Third dense activation	Sigmoid
Loss function	Mean Squared Error / Binary Crossentropy
Optimizer	Adaptive Moment Estimation(Adam)
Adam learning rate	0.001
Adam beta_1	0.9
Adam beta_2	0.999
Adam decay	0
Batch size	1000
Epochs	100 epochs
Rate of validation	0.1
Checkpoint monitor	Validation loss
ReduceLR monitor	Train loss
ReduceLR factor	0.1
ReduceLR patience	6 epochs
ReduceLR minimum learning rate	0.00001
EarlyStop patience	10 epochs
EarlyStop monitor	Validation loss
Power_rate in binding affinity of train set	0.3

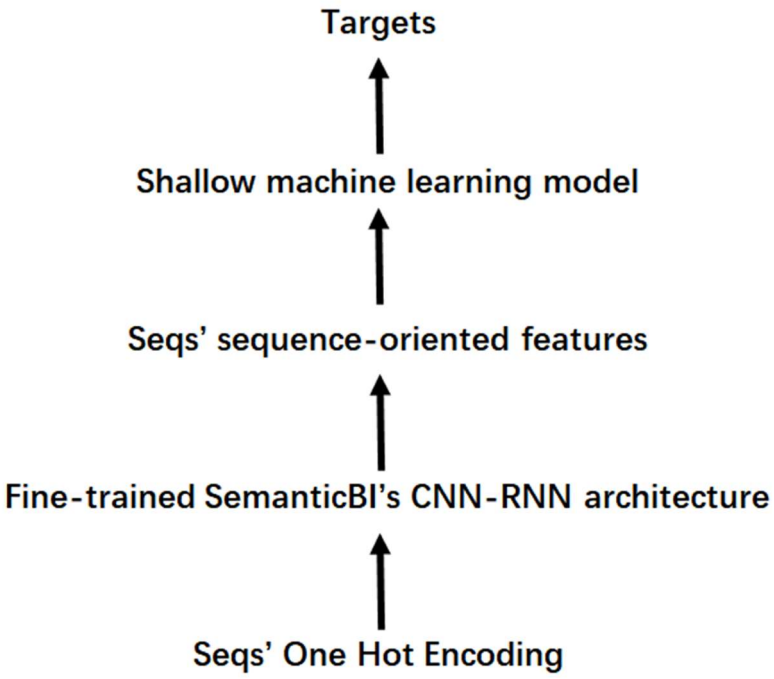


Supplementary Information 1 - Figure S1



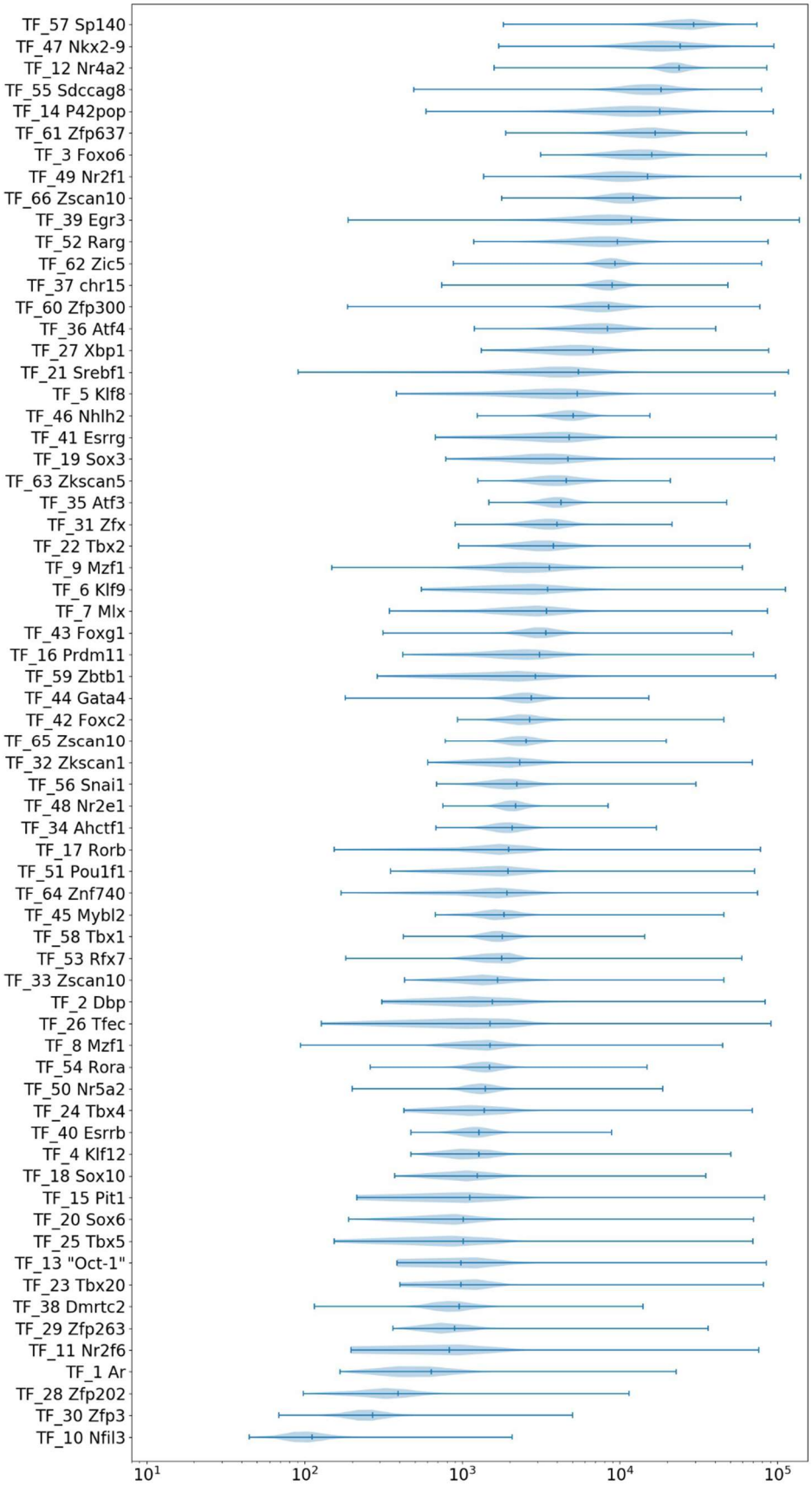
Supplementary Information 1 - Figure S2

Supplementary Information 2. The details of SemanticFeature experiments. (A) SemanticFeature is an abstract model, whose protocol is shown in Figure S3. First, feed one hot encoding sequence to the fine-trained SemanticBI model, and obtain the sequence-oriented features by SemanticBI's CNN-RNN architecture. Second, fit targets/labels using a shallow machine learning model with sequence-oriented features as the representation of sequences. (B) Details of the experiment from CADD framework: The representation of sequence is from PBM_66 SemanticBI or from both PBM_66 SemanticBI and ChIP-seq_83 SemanticBI. The input of paired sequence is the combination of positive sequence's sequence-oriented feature and the element-wise subtraction of positive sequence's sequence-oriented feature and negative sequence' sequence-oriented feature. The shallow machine learning here was a two-layer fully connected layer, first layer had 400 units and second layer had 200 units.

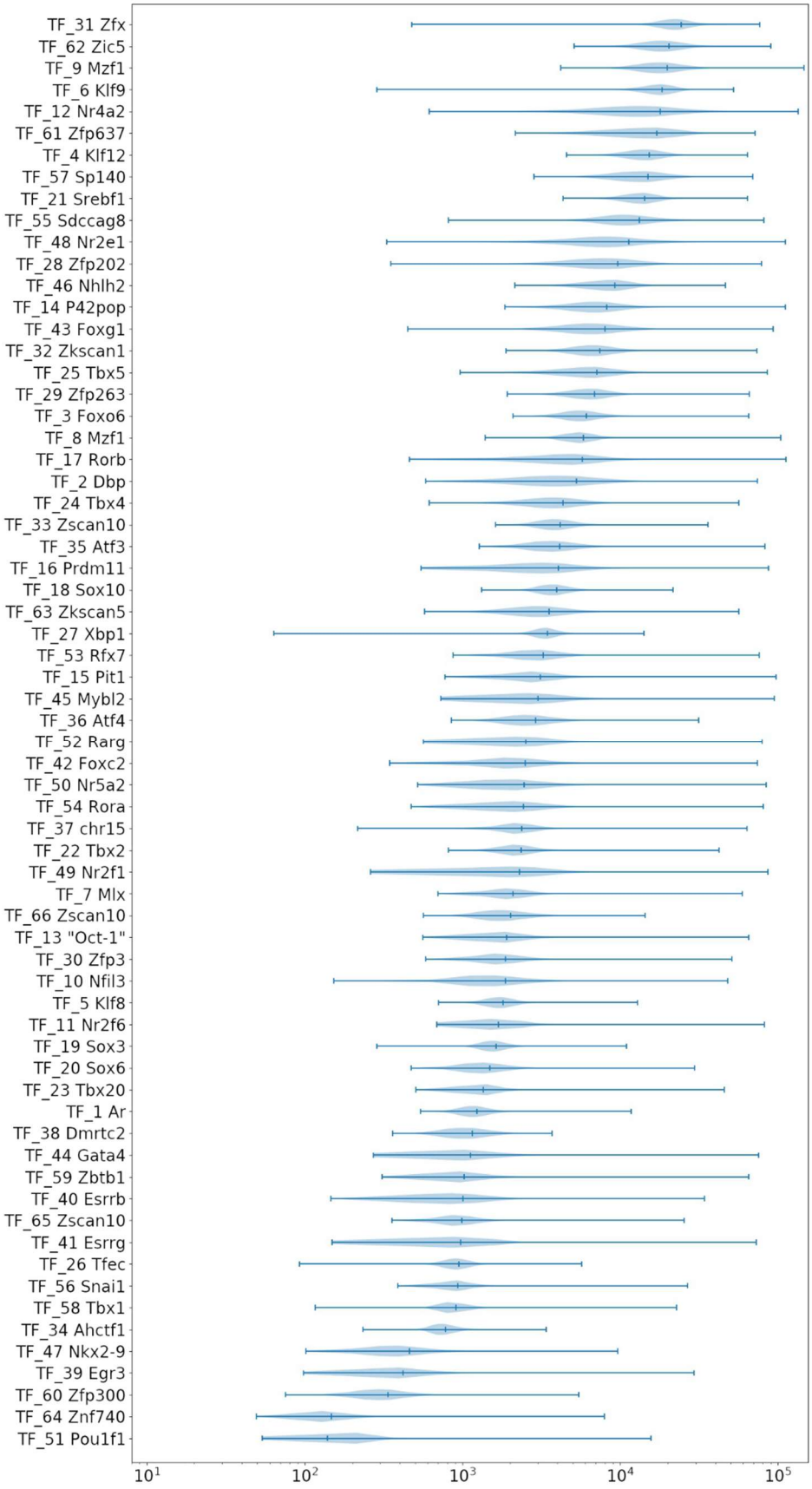


Supplementary Information 2 - Figure S3

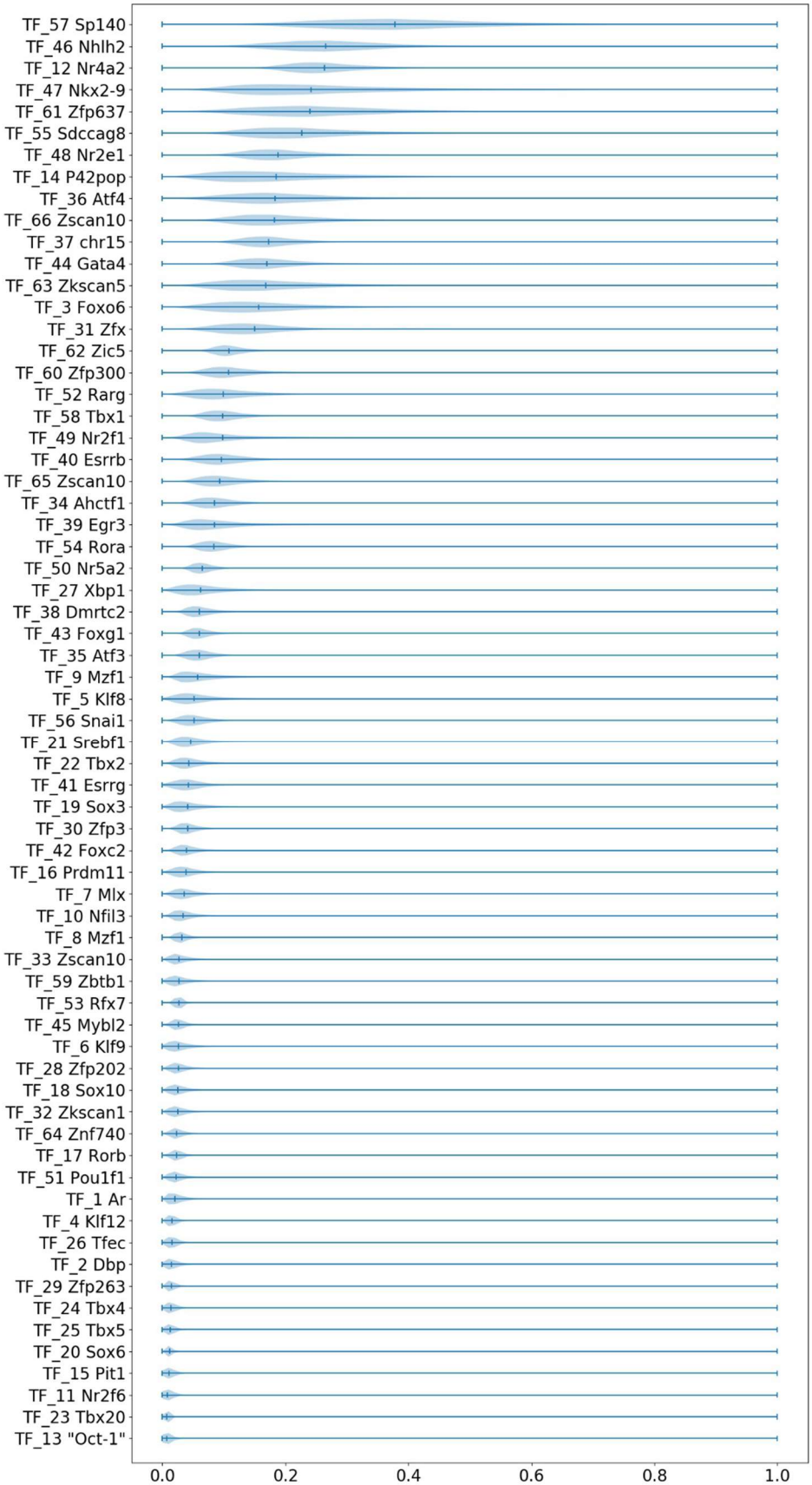
Supplementary Information 3. The distribution of binding intensities for each PBM datasets. (A) Figure S4 is the sorted distribution of binding intensities for each training set of PBM datasets. (B) Figure S5 is the sorted distribution of binding intensities for each testing set of PBM datasets. (C) Figure S6 is the sorted distribution of binding intensities for each normalized training set of PBM datasets. Most of intensities for some TFs locate to a small region in the left side of axis. (D) Figure S7 is the sorted distribution of binding intensities for each normalized and powered training set of PBM datasets.



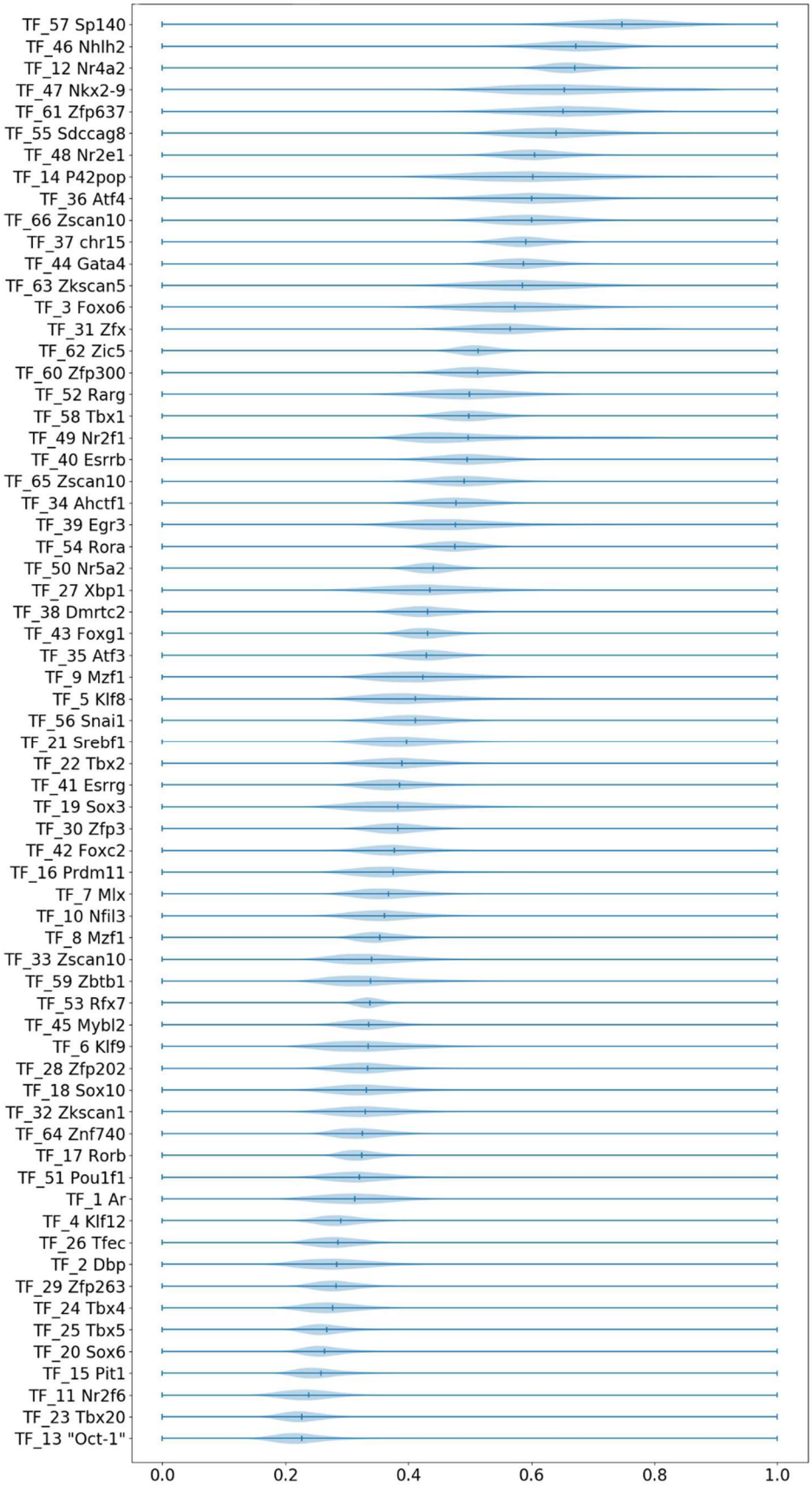
Supplementary Information 3 - Figure S4



Supplementary Information 3 - Figure S5

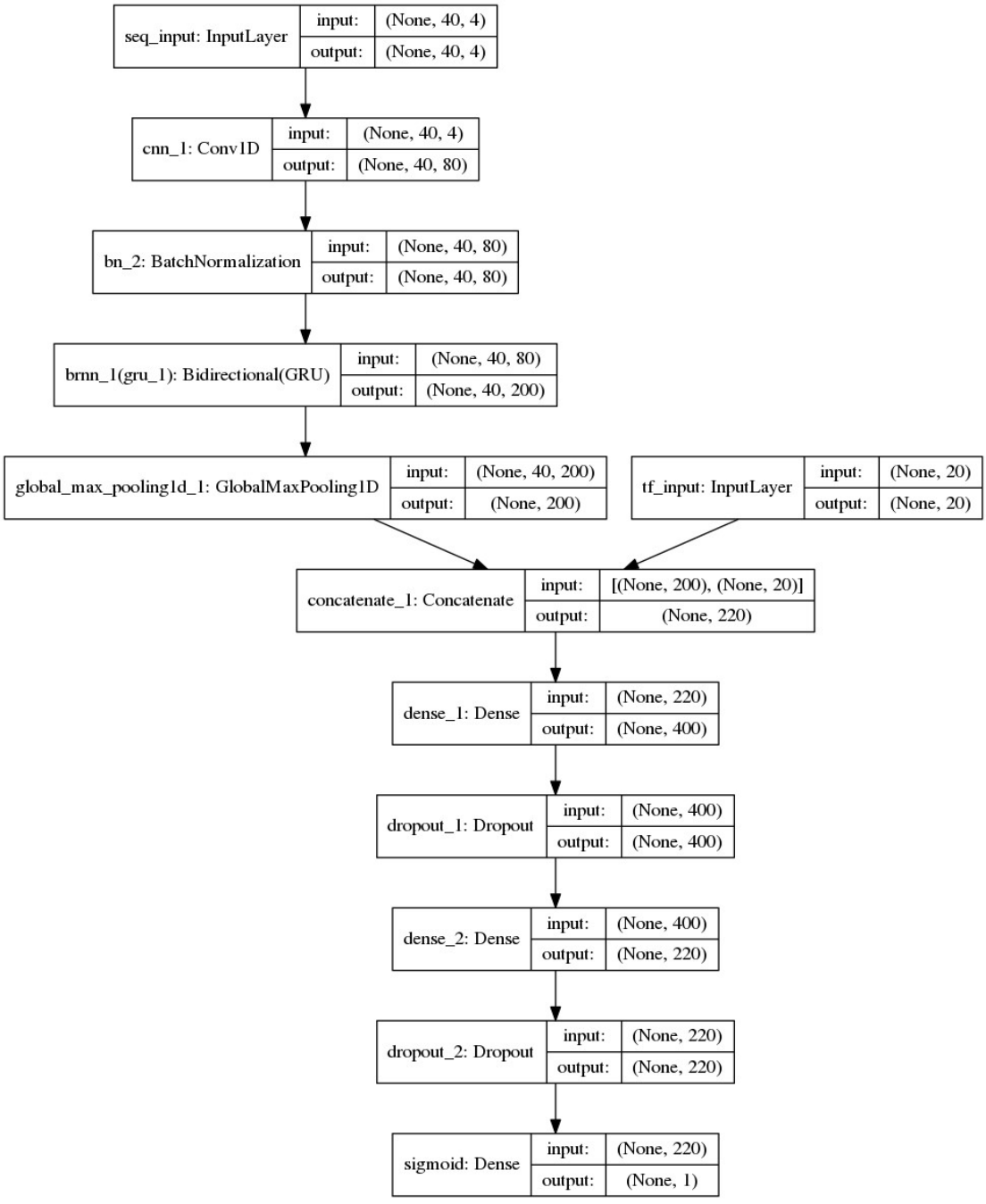


Supplementary Information 3 - Figure S6



Supplementary Information 3 - Figure S7

Supplementary Information 4. The details of PBM_20 experiment. (A) Figure S8 is the model structure of PBM_20 SemanticBI. The hyperparameter configuration of PBM_20 SemanticBI is shown in Table S1. (B) Table S2 contains the 20 TFs that were selected to train SemanticBI. More details on how we reduced PBM_66 to PBM_20 are shown in Supplementary Information PBM_20.



Supplementary Information 5 - Figure S8

Supplementary Information 5 - Table S2: TF list of PBM_20

TF index	TF name	TF family
TF_3	Foxo6	Forkhead
TF_4	Klf12	C2H2 ZF
TF_6	Klf9	C2H2 ZF
TF_8	Mzf1	C2H2 ZF
TF_12	Nr4a2	Nuclear receptor
TF_16	Prdm11	Myb/SANT
TF_20	Sox6	Sox
TF_22	Tbx2	T-box
TF_26	Tfec	bHLH
TF_28	Zfp202	C2H2 ZF
TF_29	Zfp263	C2H2 ZF
TF_31	Zfx	C2H2 ZF
TF_36	Atf4	bZIP
TF_38	Dmrtc2	DM
TF_39	Egr3	C2H2 ZF
TF_44	Gata4	GATA
TF_47	Nkx2-9	Homeodomain
TF_51	Pou1f1	Pou+Homeodomain
TF_53	Rfx7	RFX
TF_56	Snai1	C2H2 ZF