Supplemental Material for Sequence2Vec: A novel embedding approach for modeling transcription factor binding affinity landscape

Hanjun Dai[1,*], Ramzan Umarov[2,*], Hiroyuki Kuwahara[2], Yu Li[2], Le Song[1,†], and Xin Gao[2,†]

[1] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA.

[2] King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal, 23955-6900, Saudi Arabia.

*The first two authors contributed equally to this work.

†All correspondence should be addressed to Xin Gao (xin.gao@kaust.edu.sa) and Le Song (lsong@cc.gatech.edu).

S1 Derivative computation in Algorithm 2

Let us use $f^n = g([\mu_1^n, \mu_2^n, \ldots, \mu_L^n])$ to simplify the notation, *i.e.*, the f^n is the vector representation for sequence χ^n. If we want to minimize the mean square error between the predicted and the measured binding affinity, then the corresponding partial derivative with respect to the vector representation is

$$\frac{\partial l}{\partial f^n} = (y^n - v^\top f^n)v. \tag{1}$$

To obtain $\frac{\partial l}{\partial \mu_i^n}$, *i.e.*, the operands of function g, we just need to record down the indexes which are maximum in max pooling, then propagate $\frac{\partial l}{\partial f^n}$ back to the corresponding positions.

Next using the chain rule, we can obtain the partial derivatives with respect to the messages between nodes. The derivatives of message in the final step is given by

$$\frac{\partial l}{\partial \nu_{ij}^{n(T)}} = W_4^\top \frac{\partial l}{\partial \mu_j^n} \frac{\partial \sigma}{\partial (W_3 x_j + W_4 \nu_{kj}^{(T)})}. \tag{2}$$

For the unrolling step $t = \{1, 2, \ldots, T-1\}$, the partial derivatives with respect to each pairwise message in each stage of the fixed point iteration is given by

$$\frac{\partial l}{\partial \nu_{ij}^{n(t)}} = W_2^\top \frac{\partial l}{\partial \nu_{jk'}^{n(t+1)}} \frac{\partial \sigma}{\partial (W_1 x_j + W_2 [\nu_{ij}^{n(t)}])}, \tag{3}$$

where k' is the neighbor of j besides i.

Now we are ready to get the derivatives with respect to parameters $\mathbf{W}$. As mentioned above, the embedding algorithm in Algorithm 2 is essentially a recurrent network. So in order to get the derivatives of parameters, we need to aggregate over recurrent deep and entire sequence, which is just the collection of derivatives in each unrolled step. That is

$$\frac{\partial l}{\partial v} = (y^n - v^\top f^n) f^n,$$

$$\frac{\partial l}{\partial W_1} = \sum_{t=1}^{T-1} \sum_{(i,j)\in\mathcal{E}} \frac{\partial l}{\partial \nu_{ij}^{n(t+1)}} \frac{\partial \sigma}{\partial (W_1 x_i + W_2 \nu_{ki}^{n(t)})} x_i^\top,$$

$$\frac{\partial l}{\partial W_2} = \sum_{t=1}^{T-1} \sum_{(i,j)\in\mathcal{E}} \frac{\partial l}{\partial \nu_{ij}^{n(t+1)}} \frac{\partial \sigma}{\partial (W_1 x_i + W_2 \nu_{ki}^{n(t)})} \nu_{ki}^{n(t)\top},$$

$$\frac{\partial l}{\partial W_3} = \sum_{i=1}^{L} \frac{\partial l}{\partial \mu_i^n} \frac{\partial \sigma}{\partial (W_3 x_i + W_4 \nu_{ki}^{n(T)})} x_i^\top,$$
$$\frac{\partial l}{\partial W_4} = \sum_{i=1}^{L} \frac{\partial l}{\partial \mu_i^n} \frac{\partial \sigma}{\partial (W_3 x_i + W_4 \nu_{ki}^{n(T)})} \nu_{ki}^{n(T)\top}, \tag{4}$$

where $\mathcal{E}$ denotes the set of edges between the latent variables in an HMM. Using the equations of partial derivatives above, we can perform gradient descent to update the parameters.

S2 Computation graph of Sequence2Vec

The full computational graph of our learning algorithm is shown in Figure S1, which includes the illustration of the first two rounds of message passing iterations to obtain the nonlinear feature embedding h and also the pooling feature representation g. Gradient information from the prediction residue is propagated backward based on the computation graph.

S3 Performance measures

We measured performance using the root mean square error (RMSE), Pearson product-moment correlation coefficient (PCC), and Spearman's rank correlation coefficient (SCC):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (\tilde{y}^n - y^n)^2},$$
$$\text{PCC} = \frac{\sum_{n=1}^{N} (\tilde{y}^n - \tilde{\mu}_y)(y^n - \mu_y)}{\sqrt{\sum_{n=1}^{N} (\tilde{y}^n - \tilde{\mu}_y)^2} \sqrt{\sum_{n=1}^{N} (y^n - \mu_y)^2}},$$
$$\text{SCC} = \frac{\sum_{n=1}^{N} (\tilde{z}^n - \tilde{\mu}_z)(z^n - \mu_z)}{\sqrt{\sum_{n=1}^{N} (\tilde{z}^n - \tilde{\mu}_z)^2 \sum_{n=1}^{N} (z^n - \mu_z)^2}},$$

where y^n and $\tilde{y}^n$ are the real and predicted binding affinity values, z^n and $\tilde{z}^n$ are the real and predicted rank of the affinity values, for the n-th binding sequence, respectively, and N is the number of binding sequences in the test sets. μ_y and μ_z are the average value of y and z respectively.

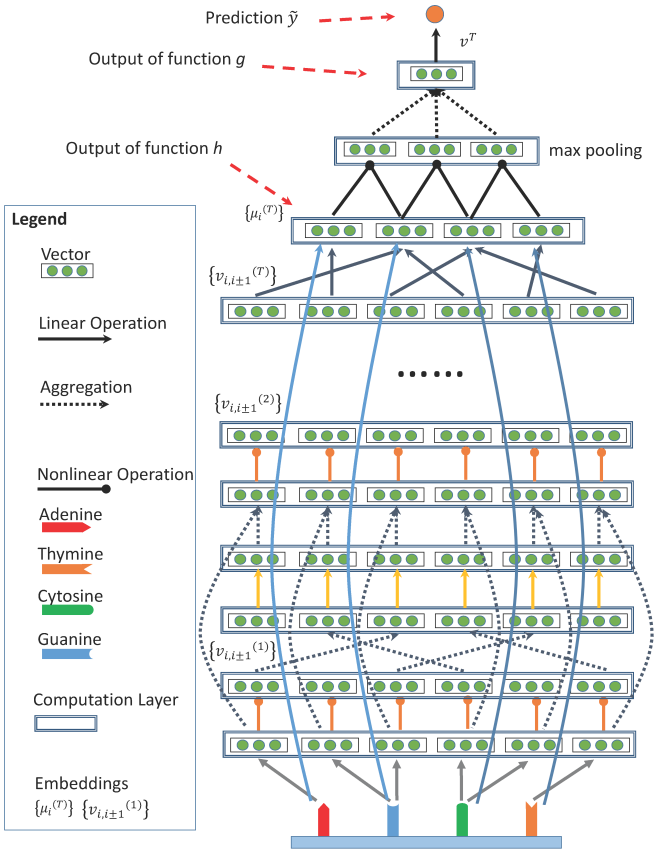


Figure S1: The computation graph of our proposed algorithm. Operation symbols with different colors correspond to different sets of parameters, while the operations with the same color represent the same parameters. The parameters are shared across nodes in the same computation layer, and across different iterations in message passing, which is similar to the recurrent neural network.

S4 Comprehensive comparison on 28 *Saccharomyces cerevisiae* MITOMI 2.0 data sets

Table S1 shows the performance of different methods over the 28 MITOMI 2.0 data sets.

S5 Comprehensive comparison to BaMM and DeepBind on DREAM5 data sets

Table S2 shows the performance of BaMM, DeepBind and Sequence2Vec for each TF measured by Pearson correlation, Spearman correlation, and AUC over the 66 PBM data sets.

S6 The sequence logos for the 28 MITOMI 2.0 data sets

Table S3 shows the sequence logos of the 8-mer motifs ranked by Sequence2Vec with those of the known motifs on the 28 *Saccharomyces cerevisiae* MITOMI 2.0 data sets. The first column shows the TF name. The second column shows the sequence logos constructed using the known motifs from the YeTFaSCo database. The third column shows the reverse complement of the sequence logos of the known motifs. The fourth column shows the sequence logos of the 8-mer motifs ranked by Sequence2Vec.

S7 Convergence of Sequence2Vec

Here we plot the convergence curves with respect to the training RMSE over all the data sets. The results of MITOMI 2.0 and HiTS-FLIP datasets are shown in Table S4 and S5. The results of 66 PBM datasets are shown from Table S6 to Table S10.

S8 Computational efficiency

We report both the training and test time in Table S11. For each data set, we report the average runtime per each sequence, as well as the total runtime, for training and testing. All the experiments were conducted on a workstation with Intel Xeon CPU E5-1620 v2 @ 3.70GHz and 32G Memory.

S9 Sensitivity analysis of hyper-parameters

We studied the performance with respect to different settings of hyper-parameters. We analyzed the effect of the range of dependencies encoded, the nonlinearity, the embedding size, and the batch size used during stochastic training. For each setting, we varied the corresponding parameter while fixing the rest. For HiTS-FLIP and MITOMI 2.0 data sets, we ran the experiments until convergence. For PBM, since there are many large TF datasets, we ran for a fixed number of iterations and report the average test performance. For HiTS-FLIP and MITOMI 2.0 data sets, we report Pearson Correlation Coefficients (PCC). While for PBM, we report AUC.

Range of dependency and nonlinearity. In Table S12, we present the results with different numbers of message passing rounds, as well as whether using the nonlinear activation function or not. It is easy to see that, using nonlinear activation function is always a good option for these datasets. Also for PBM data set which is much larger than MITOMI 2.0, the gain of performance is more significant. This implies that, when enough training data are supplied, typically the nonlinear model would have more model capacity and thus performs better than linear ones.

On the other hand, the more rounds of message passing we perform, the longer range of sequence each embedding will cover. Thus from Table S12, typically including longer range of dependency does not hurt the quality of feature. In addition, with more iterations, each embedding gets richer feature representation. Thus for large datasets like HiTS-FLIP and PBM, the performance would be even better. For small ones like MITOMI 2.0, chances are that the model suffers from overfitting.

Embedding size and batch size. Table S13 shows the performances with different embedding sizes and batch sizes. We found that, as long as the embedding is large enough (typically 64 for small data sets like MITOMI 2.0, and 128 for large ones like HiTS-FLIP), the performance does not have big differences. The performance is also robust to the batch size. These experiments show that, our model does not rely heavily on hyper-parameter tuning. It can achieve good performance under a wide range of hyper-parameter settings.

S10 Experiments on synthetic data sets

We created several synthetic experiments to check the correctness of our algorithm. Each one contains 1000 positive sequences and 1000 negative sequences. For positive ones, there is one implanted 7-mer with randomly different inserting locations and different rates of mutations (5%-40%).

The results are shown in Table S14. We can see when the noise level is acceptable, we can achieve almost perfect results. When the noise level goes to pretty high (40% for example), our method still obtains decent performance. This demonstrates the effectiveness as well as the correctness of our algorithm.

Table S1: Comparison of different methods on the MITOMI 2.0 data sets for 28 TFs in *Saccharomyces cerevisiae* (Fordyce *et al.*, 2010). No.: number of 52bp sequences in the corresponding data set; PWM: position weight matrix; BaMM: Bayesian Markov Model motif discovery (Siebert and Söding, 2016); LM: the DREAM-winning HK→ME linear model (Annala *et al.*, 2011); WD: the two round weighted degree kernel-based SVR model (Wang *et al.*, 2014); DNN: the multi-layer neural network model; CNN: the convolutional neural network model (Alipanahi *et al.*, 2015); FS: the Fisher kernel-based SVR model (Jaakkola and Haussler, 1999); and S2V: the proposed Sequence2Vec model. The best performance under each measure is in bold.

Dataset	No.	Root mean square error (RMSE)							Pearson correlation coefficient (PCC)								Spearman correlation coefficient (SCC)							
		PWM	LM	SVR	DNN	CNN	FS	S2V	PWM	BaMM	LM	SVR	DNN	CNN	FS	S2V	PWM	BaMM	LM	SVR	DNN	CNN	FS	S2V
Ace2	1456	0.047	0.076	0.041	0.042	0.036	0.041	**0.032**	0.02	0.34	0.24	0.44	0.10	0.58	0.37	**0.70**	0.03	0.22	0.11	0.25	0.09	0.25	**0.29**	0.27
Aft1	1456	0.037	0.067	0.035	0.036	0.034	0.034	**0.032**	0.06	0.19	0.22	0.29	0.04	0.30	0.26	**0.49**	0.06	0.18	0.14	0.19	0.06	0.11	0.21	**0.24**
Aft2	1456	0.049	0.081	0.042	0.044	0.038	0.042	**0.027**	0.02	0.29	0.30	0.43	0.10	0.53	0.35	**0.81**	0.05	0.30	0.14	0.28	0.13	0.25	0.31	**0.37**
Bas1	1456	0.031	0.047	0.028	0.026	0.028	**0.025**	0.027	0.11	0.09	0.20	0.22	0.19	0.15	**0.31**	0.29	0.07	0.08	0.03	0.14	0.11	0.10	**0.21**	0.11
Cad1	1456	0.055	0.105	0.053	0.053	0.059	**0.052**	0.054	0.13	0.22	0.12	0.25	0.19	0.10	0.25	**0.40**	0.16	0.24	0.14	0.27	0.22	0.13	0.27	**0.35**
Cbf1	1456	0.064	0.077	0.051	0.060	0.046	0.057	**0.040**	-0.02	0.24	0.45	0.53	0.14	0.67	0.43	**0.80**	0.05	0.26	0.15	0.36	0.18	0.31	**0.39**	0.35
Cin5	1456	0.035	0.043	0.035	0.032	0.028	0.031	**0.023**	0.02	0.16	0.31	0.36	0.07	0.54	0.28	**0.69**	0.08	0.14	0.07	0.21	0.08	0.18	0.16	**0.24**
Cup9	1456	0.034	0.067	0.034	**0.031**	0.032	**0.031**	0.035	-0.02	0.11	0.05	0.13	0.01	0.15	0.21	**0.29**	-0.01	0.11	0.03	0.12	0.00	0.14	**0.26**	0.16
Dal80	1456	0.052	0.077	0.044	0.049	0.043	0.047	**0.032**	0.10	0.23	0.31	0.50	0.19	0.53	0.41	**0.75**	0.11	0.30	0.18	0.31	0.17	0.32	0.33	**0.38**
Gat1	1456	0.051	0.071	0.039	0.046	0.036	0.044	**0.025**	0.09	0.32	0.38	0.61	0.31	0.67	0.54	**0.87**	0.05	0.40	0.20	0.37	0.20	0.34	0.46	**0.49**
Gcn4	1084	0.052	0.049	0.044	0.043	0.039	0.041	**0.028**	0.08	0.12	0.29	0.33	0.10	0.31	0.23	**0.60**	0.08	**0.15**	0.06	0.08	0.08	0.11	0.09	0.11
Mata2	1453	0.093	0.204	0.085	0.074	0.076	0.071	**0.069**	-0.02	0.12	0.12	0.19	0.07	0.24	0.29	**0.44**	0.00	0.10	0.03	0.14	0.09	0.12	0.19	**0.21**
Mcm1	1456	0.049	0.082	**0.039**	0.046	0.040	0.046	0.040	0.20	0.16	0.27	0.55	0.31	0.49	0.33	**0.57**	0.08	0.14	0.09	0.20	0.15	0.16	0.19	**0.21**
Met31	1456	0.042	0.065	0.033	0.037	**0.029**	0.039	0.033	0.12	0.18	0.33	0.56	0.41	**0.62**	0.29	0.57	0.02	0.13	0.12	0.19	0.13	0.15	**0.22**	0.13
Met32	1456	0.075	0.115	0.061	0.065	**0.057**	0.071	0.065	0.08	0.16	0.28	0.49	0.36	0.49	0.26	**0.50**	-0.01	0.12	0.10	0.14	0.07	0.07	0.15	**0.20**
Msn1	1424	0.040	0.056	0.034	0.037	0.029	0.035	**0.024**	0.04	0.30	0.33	0.50	0.08	0.62	0.41	**0.75**	0.03	0.29	0.13	0.31	0.09	0.29	0.34	**0.40**
Msn2	1456	0.094	0.143	0.067	0.077	0.063	0.081	**0.049**	0.14	0.54	0.35	0.67	0.52	0.72	0.61	**0.84**	0.22	0.50	0.26	0.51	0.37	0.51	0.51	**0.53**
Nrg2	1452	0.044	0.077	0.036	0.041	0.032	0.042	**0.029**	0.06	0.36	0.29	0.49	0.13	0.66	0.37	**0.72**	0.08	0.26	0.12	0.23	0.16	0.25	**0.31**	0.27
Pdr3	1456	0.037	0.062	0.029	0.035	0.030	0.032	**0.027**	0.15	0.38	0.26	0.60	0.28	0.60	0.51	**0.68**	0.14	0.27	0.16	0.33	0.23	0.29	0.38	**0.40**
Pho4	1456	0.043	0.066	0.037	0.038	0.031	0.036	**0.022**	-0.02	0.31	0.39	0.51	0.11	0.63	0.44	**0.81**	0.05	0.23	0.06	**0.31**	0.08	0.24	0.30	0.27
Reb1	1456	0.045	0.087	0.039	0.041	0.038	0.041	**0.037**	0.01	0.20	0.18	0.36	0.04	0.45	0.26	**0.56**	0.03	0.15	0.08	0.12	0.06	0.08	0.20	**0.21**
Rox1	1456	0.046	0.061	0.042	0.043	0.040	0.043	**0.029**	0.09	0.32	0.34	0.40	0.14	0.42	0.25	**0.76**	0.09	0.18	0.13	0.17	0.11	0.07	0.14	**0.21**
Rpn4	1456	0.054	0.108	0.053	**0.051**	0.056	**0.051**	0.053	0.04	0.07	0.08	0.08	0.05	0.04	0.14	**0.23**	0.05	0.13	0.00	0.06	0.07	-0.02	**0.18**	0.14
Sko1	1456	0.051	0.079	0.043	0.047	0.038	0.044	**0.030**	0.01	0.37	0.28	0.50	0.07	0.65	0.43	**0.78**	0.03	0.31	0.15	0.31	0.05	0.29	0.33	**0.39**
Stb5	1424	0.071	0.126	0.056	0.064	0.048	0.062	**0.045**	0.06	0.26	0.23	0.52	0.13	0.65	0.37	**0.69**	0.11	0.23	0.09	0.32	0.16	0.28	0.34	**0.36**
Yap1	1456	0.019	0.040	0.020	0.017	0.018	0.017	**0.014**	0.02	0.15	0.28	0.33	-0.01	0.31	0.26	**0.57**	0.01	0.20	0.04	0.13	-0.01	0.09	0.11	**0.21**
Yap3	1456	0.016	0.035	0.018	0.015	0.015	0.014	**0.011**	0.10	0.27	0.20	0.29	0.11	0.32	0.32	**0.65**	0.14	0.44	0.15	0.15	0.15	0.19	0.23	**0.48**
Yap7	1456	0.043	0.083	0.040	0.041	0.041	0.041	**0.038**	0.12	0.22	0.13	0.32	0.20	0.27	0.31	**0.42**	0.16	0.26	0.13	0.31	0.25	0.21	0.33	**0.37**
Average	-	0.049	0.080	0.042	0.044	0.039	0.043	**0.035**	0.06	0.24	0.26	0.41	0.16	0.45	0.34	**0.62**	0.07	0.23	0.11	0.23	0.13	0.20	0.26	**0.29**

Table S2: Comprehensive comparison of methods on the PBM data from the DREAM5 challenge. BaMM: Bayesian Markov Model motif discovery (Siebert and Söding, 2016); DeepBind: state-of-the-art binding affinity prediction method and S2V: the proposed Sequence2Vec method.

TF	Pearson			Spearman			AUC		
	BaMM	DeepBind	S2V	BaMM	DeepBind	S2V	BaMM	DeepBind	S2V
1	0.123	**0.662**	0.652	0.088	**0.736**	0.678	0.759	0.831	**0.908**
2	-0.183	0.651	**0.692**	-0.239	0.799	**0.811**	0.435	0.987	**0.993**
3	0.422	0.822	**0.863**	0.295	0.825	**0.835**	0.965	0.988	**0.992**
4	-0.214	**0.662**	0.550	-0.318	**0.689**	0.595	0.642	0.931	**0.933**
5	0.469	0.795	**0.828**	0.732	0.747	**0.809**	0.933	0.990	**0.991**
6	**0.510**	0.473	0.499	0.622	**0.645**	0.627	0.922	**0.991**	**0.991**
7	0.491	0.826	**0.844**	**0.699**	0.688	0.693	0.928	**0.999**	**0.999**
8	0.497	0.656	**0.726**	**0.581**	0.449	0.469	0.888	0.963	**0.989**
9	0.395	0.613	**0.691**	0.253	0.630	**0.706**	0.892	0.874	**0.944**
10	0.566	0.713	**0.831**	0.584	0.803	**0.822**	0.894	0.978	**0.991**
11	0.582	0.816	**0.849**	0.638	**0.710**	0.658	0.903	0.992	**0.995**
12	0.451	**0.726**	0.650	0.537	**0.719**	0.594	**0.935**	0.934	0.916
13	0.270	0.720	**0.758**	0.281	**0.802**	0.793	0.829	0.984	**0.990**
14	0.227	0.797	**0.836**	0.206	0.824	**0.848**	0.857	0.982	**0.986**
15	0.475	0.704	**0.720**	0.426	0.645	**0.747**	0.949	0.982	**0.986**
16	0.046	0.822	**0.883**	-0.089	0.823	**0.866**	0.746	0.980	**0.989**
17	0.147	0.678	**0.728**	0.062	0.580	**0.729**	0.888	0.972	**0.990**
18	0.140	**0.807**	0.792	0.024	0.821	**0.837**	0.887	0.973	**0.979**
19	0.432	**0.710**	0.674	0.348	**0.786**	0.756	**0.967**	0.931	0.933
20	0.554	0.629	**0.771**	0.573	**0.744**	0.724	0.944	0.960	**0.976**
21	0.513	**0.705**	0.699	0.338	**0.765**	0.698	0.906	0.965	**0.978**
22	0.054	0.841	**0.887**	-0.033	0.824	**0.848**	0.846	0.991	**0.993**
23	0.547	0.652	**0.683**	0.591	0.693	**0.695**	0.960	0.957	**0.989**
24	0.065	**0.634**	0.624	0.081	**0.798**	0.723	0.629	0.983	**0.984**
25	0.498	**0.637**	0.625	0.288	**0.590**	0.588	0.947	**0.991**	0.988
26	0.495	**0.707**	0.701	0.537	**0.763**	0.737	0.842	0.985	**0.989**
27	0.368	**0.729**	0.708	0.406	**0.746**	0.738	0.800	**0.987**	0.986
28	0.323	0.664	**0.722**	0.291	0.512	**0.587**	0.801	0.927	**0.945**
29	-0.137	**0.698**	0.691	-0.281	**0.684**	0.636	0.662	0.953	**0.974**
30	0.442	0.532	**0.537**	0.476	**0.724**	0.701	0.782	0.835	**0.872**
31	0.312	0.843	**0.884**	0.352	0.758	**0.828**	0.843	0.957	**0.976**
32	0.526	0.705	**0.720**	0.524	0.782	**0.820**	**0.903**	0.833	0.874
33	0.392	0.801	**0.844**	0.537	0.844	**0.858**	0.936	0.961	**0.982**
34	0.033	**0.715**	0.704	-0.079	**0.720**	0.695	0.797	0.901	**0.905**
35	0.069	0.712	**0.722**	0.072	**0.843**	0.818	0.692	0.944	**0.967**
36	0.232	**0.839**	0.833	0.251	**0.865**	0.859	0.904	0.921	**0.922**
37	0.355	0.518	**0.570**	0.452	0.645	**0.663**	0.883	0.837	**0.884**
38	0.448	**0.829**	0.798	0.456	**0.835**	0.811	0.855	**0.982**	0.981
39	0.353	0.718	**0.766**	0.419	0.799	**0.847**	0.866	0.987	**0.992**
40	0.267	**0.697**	0.679	0.039	**0.780**	0.751	0.849	**0.934**	0.929
41	-0.105	0.567	**0.664**	-0.106	0.572	**0.708**	0.437	0.986	**0.993**
42	0.523	0.800	**0.832**	0.642	0.883	**0.905**	0.833	0.976	**0.981**
43	0.176	0.756	**0.780**	0.100	0.881	**0.883**	0.691	0.889	**0.899**
44	0.368	0.572	**0.652**	0.495	0.817	**0.840**	0.843	0.922	**0.956**
45	0.532	0.726	**0.820**	0.514	0.814	**0.859**	0.952	0.969	**0.993**
46	-0.056	**0.814**	0.788	-0.251	**0.826**	0.800	0.753	**0.924**	0.905
47	0.401	0.778	**0.820**	0.664	0.859	**0.884**	0.933	0.961	**0.972**
48	0.504	0.573	**0.590**	0.438	0.616	**0.624**	0.965	0.830	**0.839**
49	-0.003	0.694	**0.746**	-0.190	0.835	**0.869**	0.741	0.967	**0.973**
50	0.039	0.792	**0.794**	-0.162	**0.891**	0.838	0.800	0.970	**0.978**
51	0.680	0.616	**0.710**	0.668	0.766	**0.810**	0.961	0.973	**0.988**
52	0.483	0.718	**0.791**	0.532	0.773	**0.810**	0.955	0.973	**0.988**
53	0.376	0.622	**0.789**	0.369	0.775	**0.820**	0.826	0.922	**0.979**
54	-0.044	0.748	**0.777**	-0.066	**0.870**	0.850	0.559	0.960	**0.980**
55	0.218	0.818	**0.835**	0.550	0.842	**0.867**	0.676	0.969	**0.970**
56	0.528	0.734	**0.807**	0.510	0.656	**0.695**	0.962	0.968	**0.974**
57	0.479	0.824	**0.832**	0.509	**0.844**	0.843	0.773	0.956	**0.968**
58	0.525	**0.732**	0.693	0.541	**0.822**	0.792	0.758	**0.945**	0.923
59	0.096	0.468	**0.635**	-0.013	0.732	**0.765**	0.788	0.935	**0.972**
60	0.429	0.807	**0.810**	0.484	0.844	**0.851**	0.807	**0.945**	0.921
61	0.311	**0.806**	0.805	0.423	**0.836**	0.834	0.974	0.823	**0.835**
62	0.222	0.714	**0.751**	0.205	0.735	**0.760**	0.848	0.930	**0.953**
63	0.319	**0.732**	0.696	0.323	**0.796**	0.778	0.763	**0.954**	0.897
64	0.228	0.737	**0.779**	0.108	**0.721**	0.720	0.941	0.998	**0.999**
65	0.390	0.612	**0.625**	0.375	**0.755**	0.750	**0.951**	0.893	0.922
66	0.530	**0.829**	0.822	0.558	**0.850**	0.840	0.871	0.959	**0.968**
Avg	0.304	0.713	**0.741**	0.291	0.758	**0.765**	0.837	0.948	**0.959**

Table S3: Comparison of the sequence logos of the 8-mer motifs ranked by Sequence2Vec with those of the known motifs on the 28 *Saccharomyces cerevisiae* MITOMI 2.0 data sets.

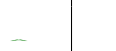
TF	Known Motifs	Reverse Complement	Sequence2Vec Motifs
Aft1			
Aft2			
Cbf1			
Pho4			
Cad1			
Cin5			
Gcn4			
Sko1			
Yap1			
Yap3			
Yap7			
Ace2			
Met31			
Met32			
Msn2			
Nrg2			
Rpn4			
Dal80			
Gat1			
Rox1			
Cup9			
Matα2			
Mcm1			
Bas1			
Reb1			
Pdr3			
Stb5			
Msn1			

Table S4: Convergence over the 28 *Saccharomyces cerevisiae* MITOMI 2.0 data sets. (part 1)

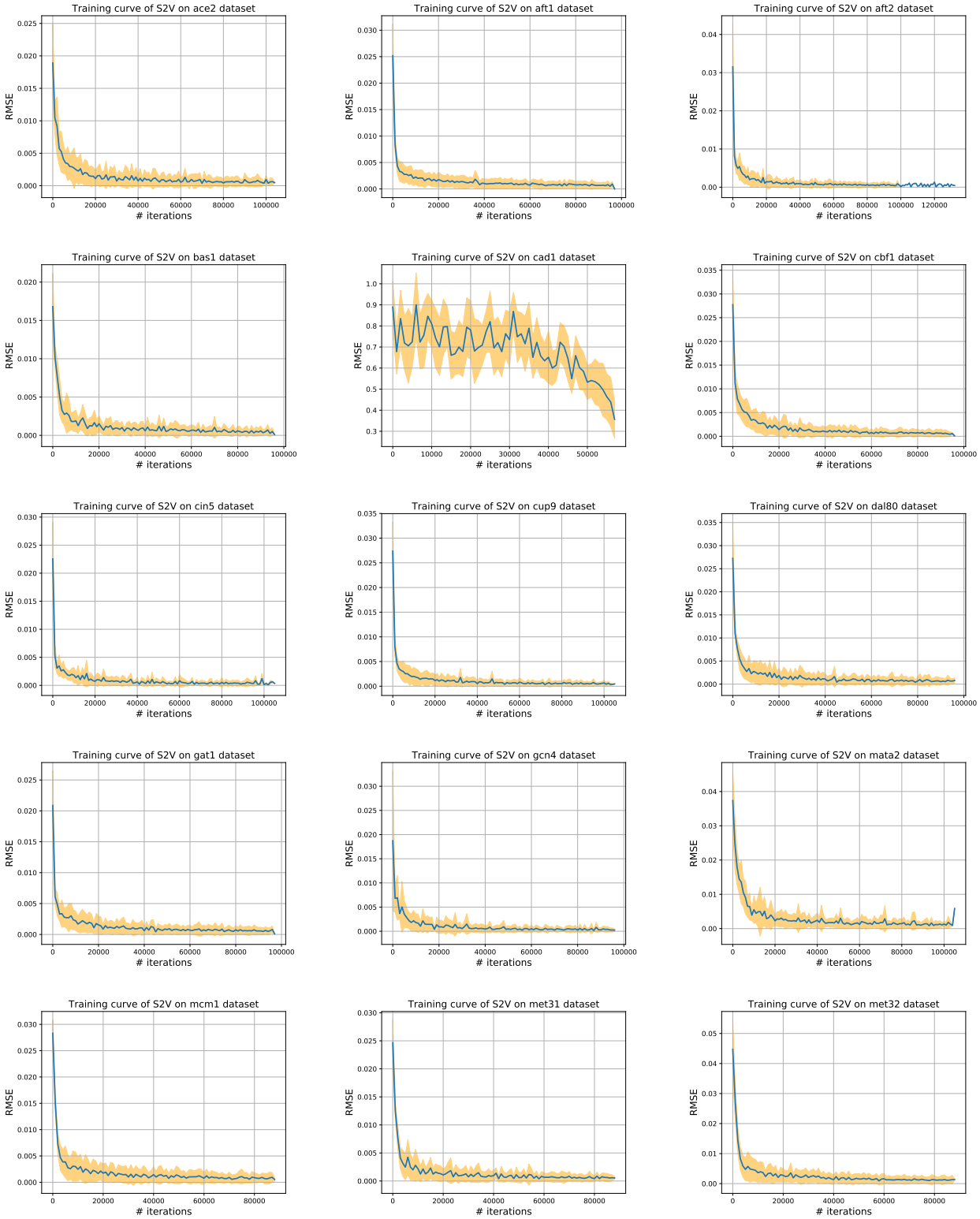


Table S5: Convergence over the 28 *Saccharomyces cerevisiae* MITOMI 2.0 data sets (part 2) and also the HiTS-FLIP data set (the last subfigure).

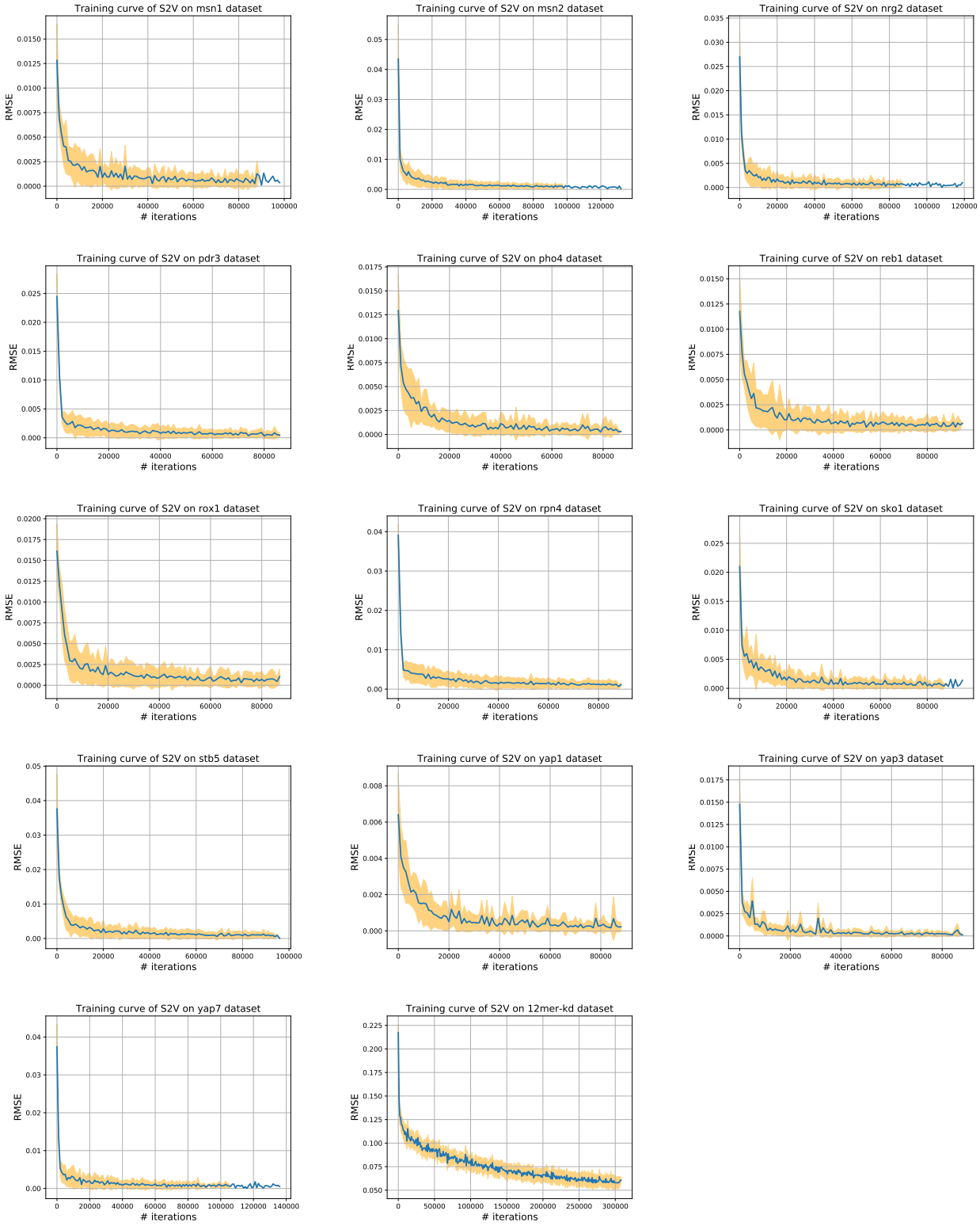
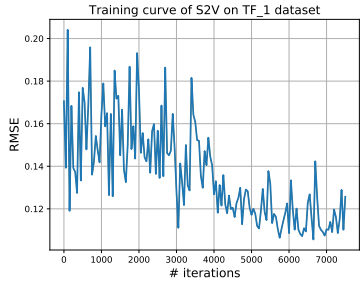
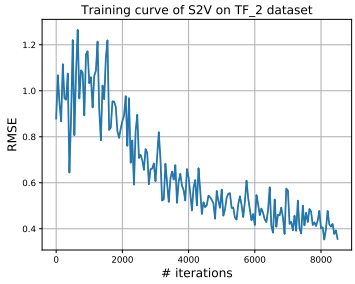
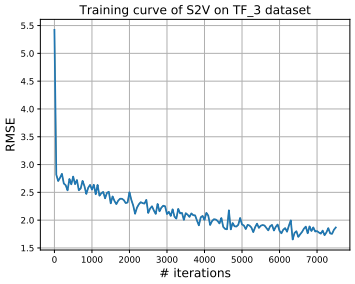
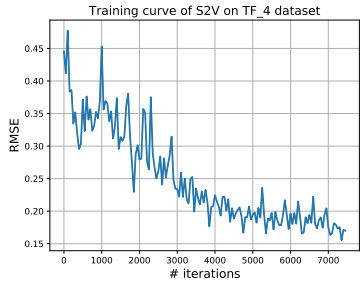


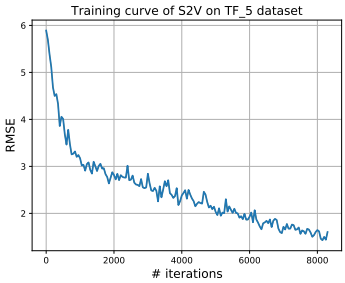
Table S6: Convergence over the 66 PBM data sets (part 1, TF_1 to TF_15)

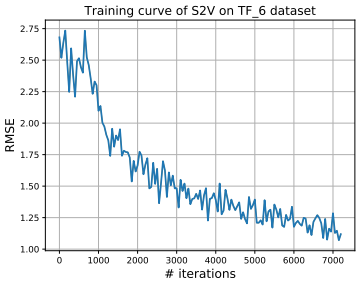


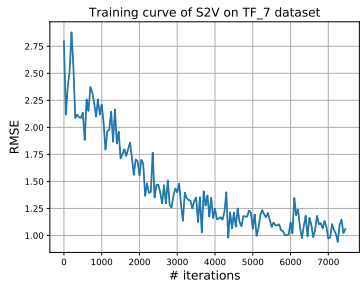


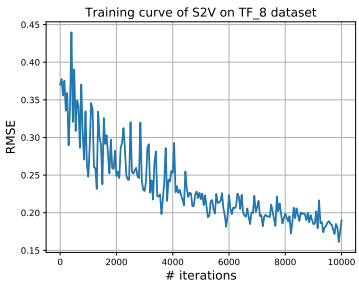


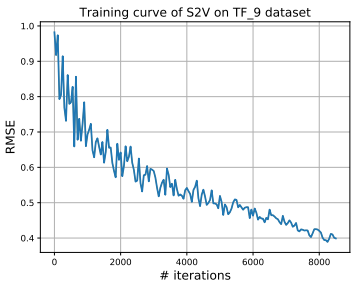


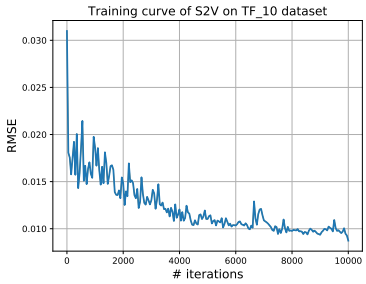


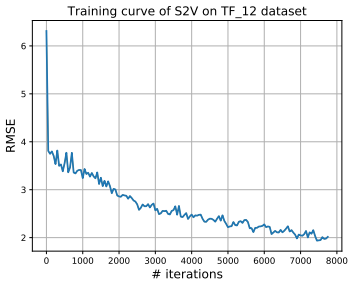


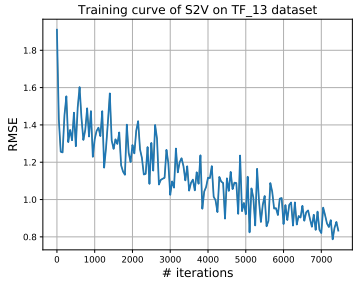












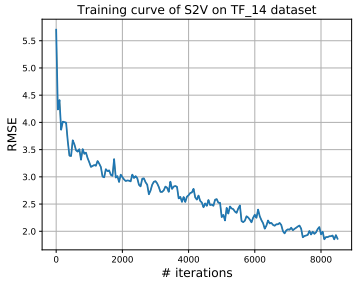
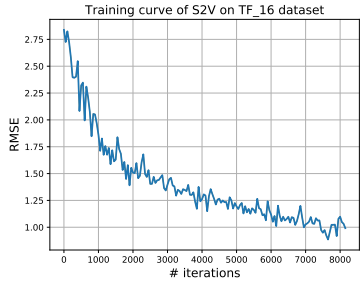
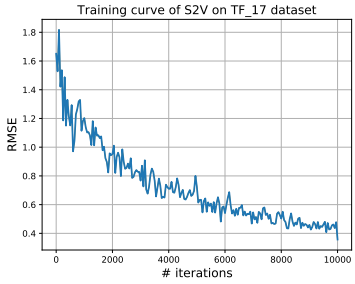
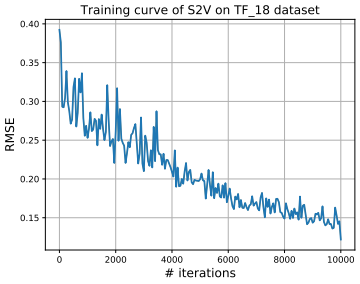
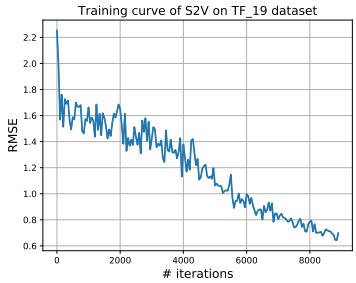


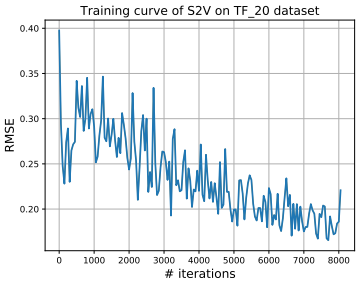
Table S7: Convergence over the 66 PBM data sets (part 2, TF_16 to TF_30)

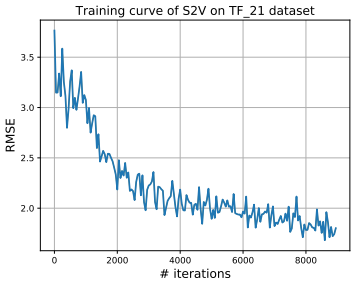


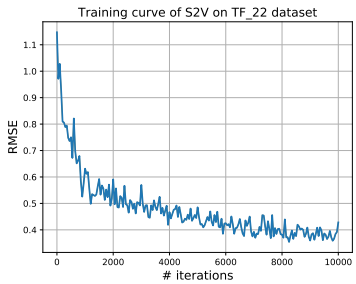


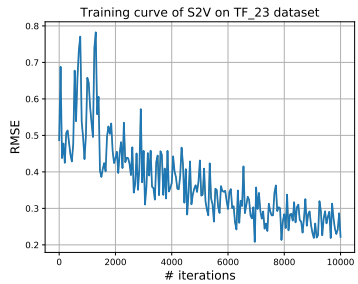


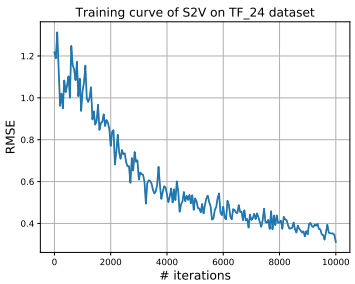


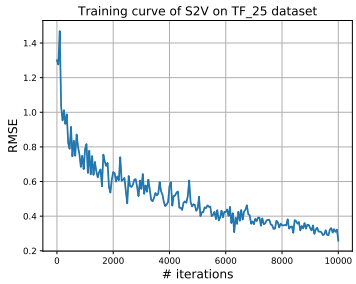


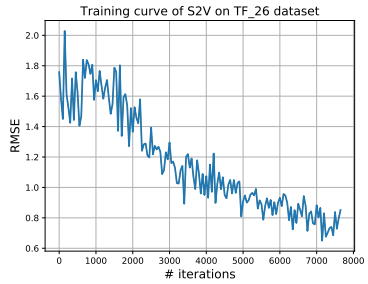


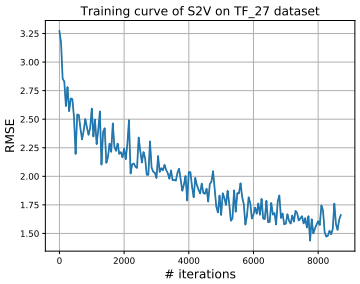


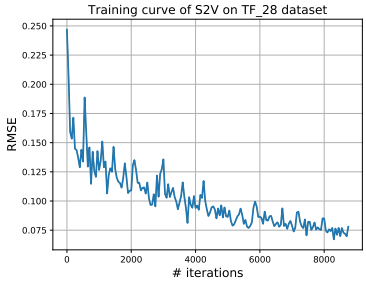


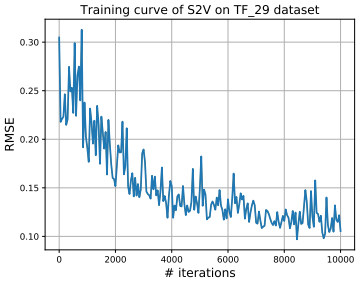












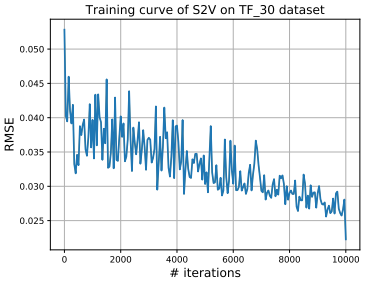
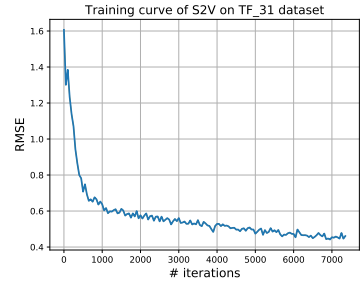
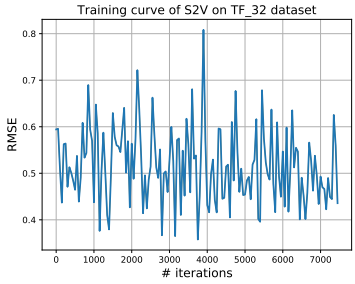
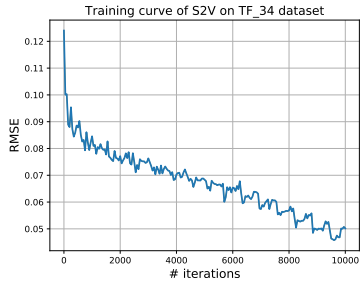
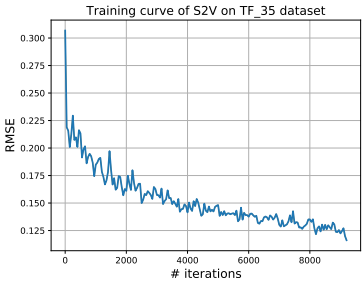


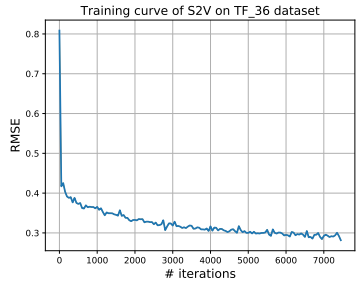
Table S8: Convergence over the 66 PBM data sets (part 3, TF_31 to TF_45)

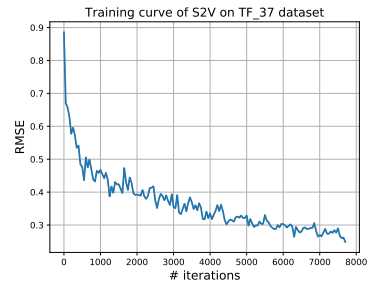


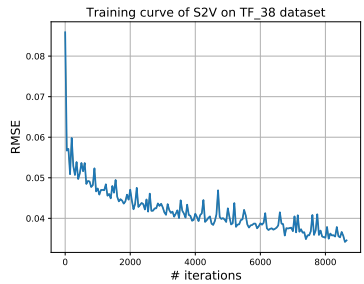


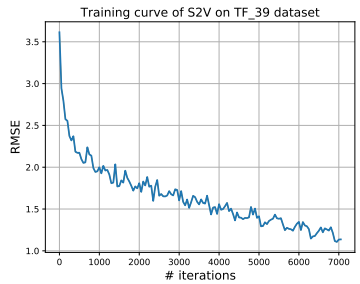


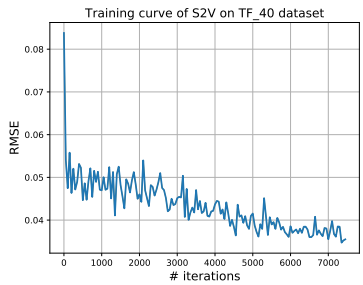


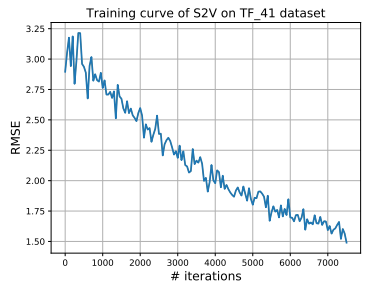


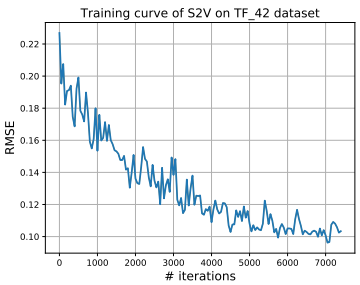


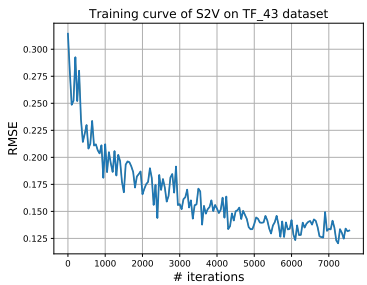


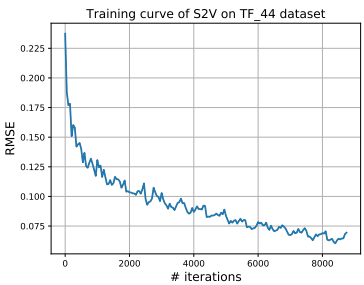












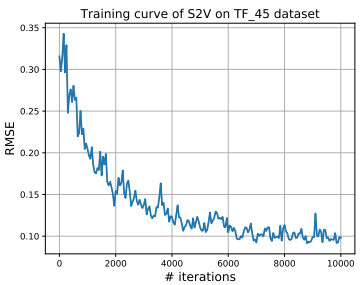
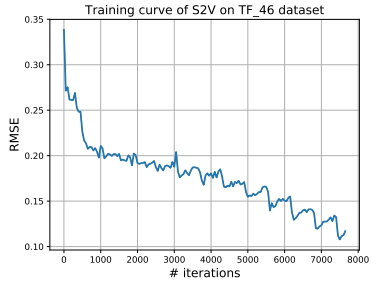
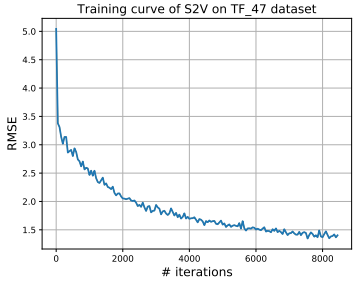
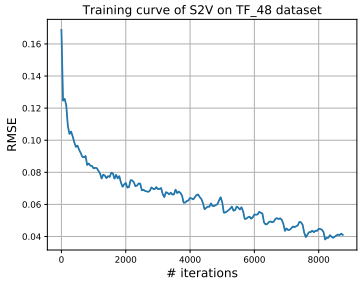
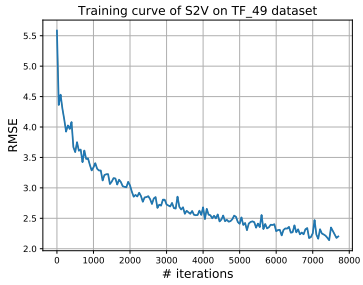


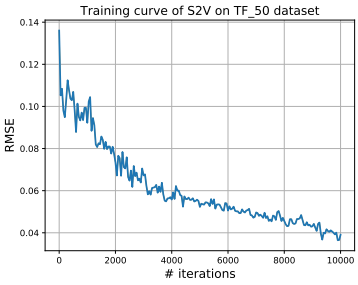
Table S9: Convergence over the 66 PBM data sets (part 4, TF_46 to TF_60)

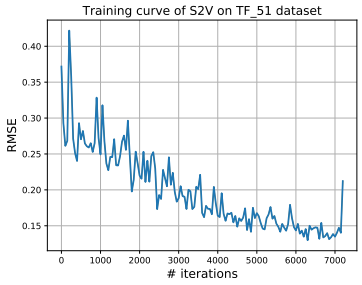


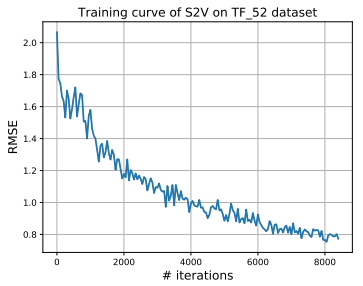


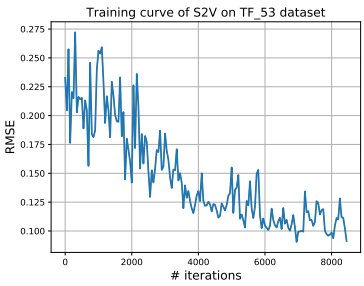


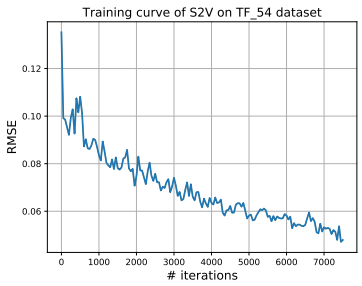


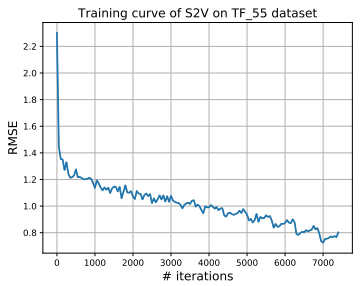


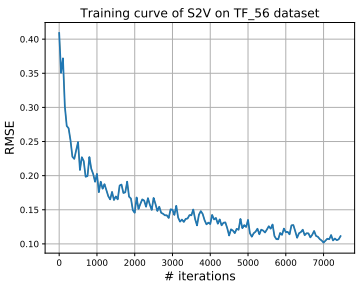


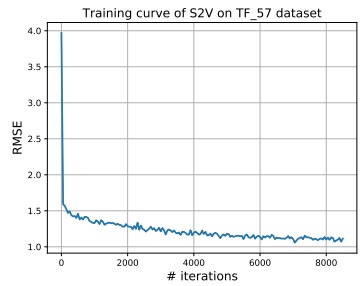


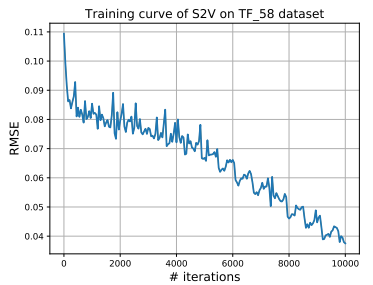


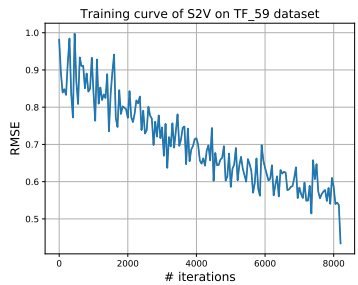












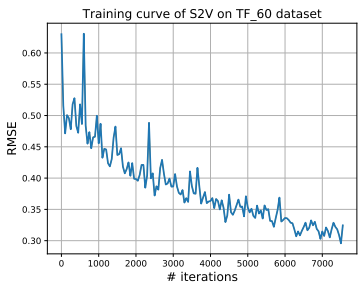


Table S10: Convergence over the 66 PBM data sets (part 5, TF_61 to TF_66)

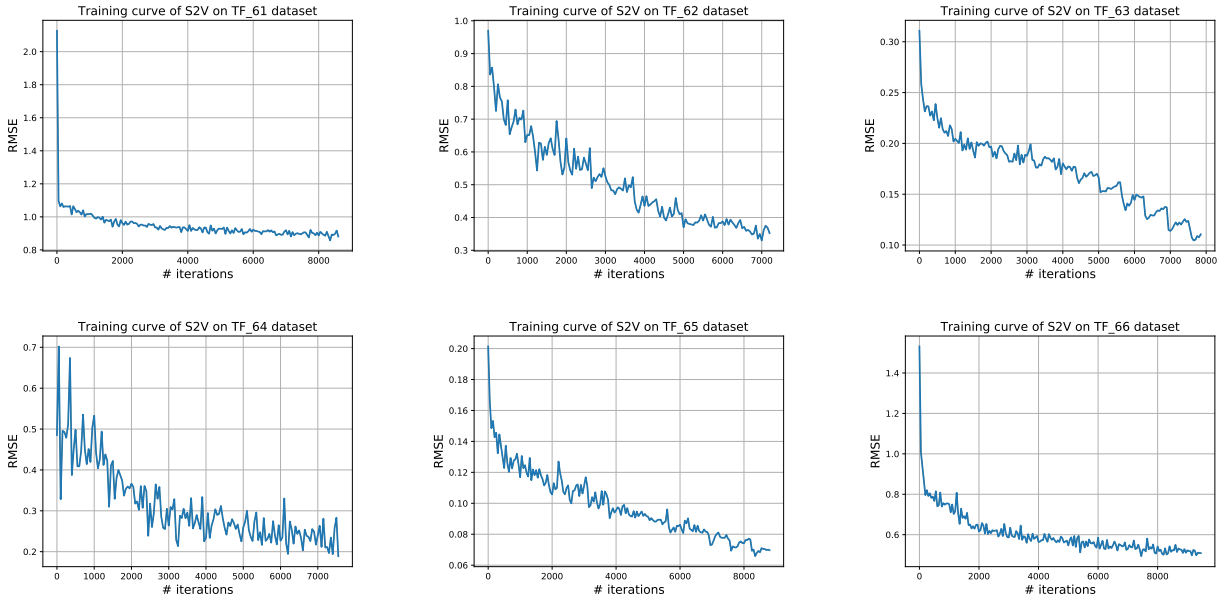


Table S11: Empirical computational cost on all three types of experimental data sets. For data sets that have multiple sets, we report the average runtime.

Dataset	Test: sec / sequence	Train: sec / sequence	Total test time (sec)	Total training time (hr)
HiTS-FLIP	$0.56 * 10^{-3}$	$1.56 * 10^{-3}$	4.66	0.68
MITOMI 2.0	$0.91 * 10^{-3}$	$2.84 * 10^{-3}$	0.13	5.02
PBM	$1.70 * 10^{-3}$	$2.98 * 10^{-3}$	68.01	1.06

Table S12: Sensitivity analysis of long range dependency and nonlinearity for Sequence2Vec. For HiTS-FLIP and MITOMI 2.0 data sets, we report Pearson correlation coefficients (PCC); and for PBM data sets, we report AUC.

	Nonlinearity		# message passing rounds			
Datasets	Linear	Nonlinear	2	3	4	5
HiTS-FLIP	0.801	**0.824**	0.824	0.825	0.826	**0.827**
MITOMI 2.0	0.611	**0.619**	0.619	0.618	**0.623**	0.619
PBM	0.883	**0.928**	0.926	0.928	0.930	**0.931**

Table S13: Sensitivity analysis of the batch size and embedding size used in Sequence2Vec. For HiTS-FLIP and MITOMI 2.0 data sets, we report Pearson correlation coefficients (PCC); and for PBM data sets, we report AUC.

	Batch Size				Embedding Size			
Datasets	16	32	64	128	32	64	128	256
HiTS-FLIP	**0.824**	0.823	**0.824**	**0.824**	0.801	0.818	0.823	**0.824**
MITOMI 2.0	**0.627**	0.618	0.619	0.618	0.595	**0.620**	0.619	0.615
PBM	0.921	0.925	**0.929**	0.928	0.912	0.922	**0.928**	**0.928**

Table S14: Synthetic experiments with Sequence2Vec. The datasets are constructed with different locations of implanted motifs and noise levels. We report AUC here.

Noise Level	0.05	0.1	0.2	0.3	0.4
AUC	0.99	0.99	0.96	0.90	0.83

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotech*, **33**(8), 831–838.

Annala, M., Laurila, K., Lähdesmäki, H., and Nykter, M. (2011). A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS One*, **6**(5), e20059.

Fordyce, P. M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J. L., and Quake, S. R. (2010). *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat Biotechnol*, **28**(9), 970–975.

Jaakkola, T. S. and Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press.

Siebert, M. and Söding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic acids research*, **44**(13), 6055–6069.

Wang, X., Kuwahara, H., and Gao, X. (2014). Modeling DNA affinity landscape through two-round support vector regression with weighted degree kernels. *BMC Syst Biol*, **8 Suppl 5**, S5.