High-Order Neural Networks and Kernel Methods for Peptide-MHC Binding Prediction

Pavel P. Kuksa, Martin Renqiang Min*, Rishabh Dugar, and Mark Gerstein*

**To whom correspondence should be addressed*

Supplementary Information

1 Model hyperparameters

We set the hyperparameters for both DNN and HONN by optimizing their performance on the validation set (20% of the training data) for each target allele. The found sets of parameters were used for training on the full train dataset(IEDB) and making predictions for the corresponding test data of each target allele. The selected set of hyperparameters is shown in Table S1.

Table S1: DNN and HONN hyperparameters used for training and pre-training

Model	param	A0201	A0206	A2402
DNN	# hidden units h_1 (layer 1)	16	32	16
	# hidden units h_2 (layer 2)	16	16	16
	# hidden units h_3 (layer 3)	4	4	8
	fine-tuning learning rate	0.008	0.008	0.008
	fine-tuning batch size	1	1	1
	fine-tuning epochs	30	60	30
	pre-training learning rate	0.005	0.05	0.05
	pre-training batch size	100	100	100
	pre-training epochs	30	10	30
HONN	# mean hidden units h^m	4	8	8
	# high-order (covariance) units h^g	4	4	4
	# factors	8	8	8
	fine-tuning learning rate	0.0015	0.0011	0.005
	fine-tuning batch size	1	1	1
	fine-tuning epochs	30	30	30
	pre-training learning rate	0.05	0.05	0.05
	pre-training batch size	100	100	100
	pre-training epochs	20	20	20

2 Comparison of the performance for pre-trained models vs models without pre-training

Tables S2, S4, S6 show the quality of ranking (nDCG) MHC-peptide interactions on test data using DNN trained with or without pre-training. Similarly, Tables S3, S5, S7 show the quality of ranking MHC-peptide interactions using high-order neural network (HONN) trained with or without pre-training. Pre-training of the models consistently improves their performance.

Table S2: A0201-Japanese data. Relevance/ranking assessment (nDCG)

method	nDCG@10	nDCG@20	nDCG@30	nDCG@50	nDCG
DNN (no pre-training)	47.26	54.36	61.56	67.98	81.32
DNN (with pre-training)	63.89	65.59	70.12	74.57	86.33

Table S3: A0201-Japanese data. Relevance/ranking assessment (nDCG)

method	nDCG@10	nDCG@20	nDCG@30	nDCG@50	nDCG
HONN (no pre-training)	63.07	64.30	66.81	74.54	85.54
HONN (with pre-training)	63.93	65.94	70.61	75.55	86.46

Table S4: A0206-Japanese data. Relevance/ranking assessment (nDCG)

method	nDCG@10	nDCG@20	nDCG@30	nDCG
DNN (no pre-training)	75.83	79.30	79.03	89.81
DNN (with pre-training)	77.50	82.21	81.72	91.74

Table S5: A0206-Japanese data. Relevance/ranking assessment (nDCG)

method	nDCG@10	nDCG@20	nDCG@30	nDCG
HONN (no pre-training)	74.14	77.86	76.90	88.72
HONN (with pre-training)	75.39	78.06	79.92	90.80

Table S6: A2402-Japanese data. Relevance/ranking assessment (nDCG)

method	nDCG@10	nDCG@20	nDCG@30	nDCG
DNN (no pre-training)	48.49	51.22	55.11	82.21
DNN (with pre-training)	51.07	53.78	56.88	84.36

Table S7: A2402-Japanese data. Relevance/ranking assessment (nDCG)

method	nDCG@5	nDCG@10	nDCG@20	nDCG@30	nDCG
HONN (no pre-training)	52.17	50.38	56.83	60.60	84.96
HONN (with pre-training)	55.73	57.36	57.20	60.82	85.20

3 Evaluation on benchmark IEDB data

To further evaluate performance of proposed methods, we used recent benchmark MHC-I binding datasets from (Kim *et al.*, 2014). We first trained a deep neural network (DNN), a proposed high-order semi-RBM (HONN), and a high-order kernel SVM (hkSVM) on BD2009 binding datasets (Kim *et al.*, 2014) for each allele. We then tested on the independent blind dataset (BLIND) for each MHC-I allele. Table S8 shows training (BD2009) and independent test (BLIND) datasets used for evaluation.

To train HONN, we first select hyper-parameters (number of hidden units and learning rate) on the validation set (20% of the training BD2009 allele data). Selected hyper-parameters are then used to train on the full training dataset. Trained models are then tested on the independent BLIND test sets for each allele.

Table S9 shows AUC and ROC-n scores for the high-order kernel SVM (hkSVM), high-order semi-RBM (HONN), and NetMHC method (Nielsen *et al.*, 2003; Buus *et al.*, 2003; Lundegaard *et al.*, 2011). As can be seen from the results in the table, while the area under ROC curve (AUC) scores are very similar for both our method and the NetMHC method (AUC scores of 92.875 and 92.375), for the very highest ranked peptides (low false positive (FP) rates), both hkSVM and HONN+hkSVM perform better on average compared to NetMHC as measured by ROC-n scores (e.g., ROC-1 scores of hkSVM or HONN are higher in about 67% (31/46) of the tested alleles and tied in 22% (10/46) of the tested alleles).

Peptide binding prediction scores on the test (BLIND) datasets for each allele are provided in the supplementary data (`hksvm_honn_blind_predict.tar.gz`).

Table S8: MHC-I BD2009 (train), MHC-I BLIND (test) benchmark datasets (Kim *et al.*, 2014) (human MHC-I alleles)

Train	#peptides	#binders	#non-binders	Test	#peptides	#binders	#non-binders
A0201-BD2009	6759	2070	4689	A0201-BLIND	1778	935	843
A2601-BD2009	2337	247	2090	A2601-BLIND	1132	86	1046
B0801-BD2009	2068	459	1609	B0801-BLIND	707	191	516
B5101-BD2009	1294	162	1132	B5101-BLIND	673	46	627
B5701-BD2009	1649	207	1442	B5701-BLIND	647	145	502
B3901-BD2009	876	168	708	B3901-BLIND	641	25	616
B1501-BD2009	3114	828	2286	B1501-BLIND	633	215	418
A0301-BD2009	4533	944	3589	A0301-BLIND	591	330	261
B0702-BD2009	2914	539	2375	B0702-BLIND	589	365	224
B1517-BD2009	841	267	574	B1517-BLIND	582	93	489
A3101-BD2009	3266	676	2590	A3101-BLIND	521	263	258
A1101-BD2009	3798	1119	2679	A1101-BLIND	514	193	321
B0802-BD2009	486	19	467	B0802-BLIND	509	18	491
B1801-BD2009	1626	181	1445	B1801-BLIND	503	28	475
A0206-BD2009	3122	1211	1911	A0206-BLIND	482	251	231
A0101-BD2009	3102	383	2719	A0101-BLIND	479	77	402
A3001-BD2009	1933	567	1366	A3001-BLIND	470	151	319
A6901-BD2009	2073	220	1853	A6901-BLIND	470	26	444
B1509-BD2009	345	16	329	B1509-BLIND	466	29	437
A0203-BD2009	3822	1205	2617	A0203-BLIND	460	255	205
A6802-BD2009	3085	607	2478	A6802-BLIND	457	188	269
B2703-BD2009	433	0	433	B2703-BLIND	441	0	441
A2501-BD2009	519	66	453	A2501-BLIND	416	5	411
A2602-BD2009	202	67	135	A2602-BLIND	413	99	314
A8001-BD2009	774	111	663	A8001-BLIND	379	7	372
B4601-BD2009	1355	77	1278	B4601-BLIND	378	0	378
B3801-BD2009	136	3	133	B3801-BLIND	351	141	210
A2402-BD2009	1904	316	1588	A2402-BLIND	346	136	210
B3501-BD2009	1945	478	1467	B3501-BLIND	341	162	179
A2603-BD2009	205	25	180	A2603-BLIND	312	37	275
A6801-BD2009	1665	629	1036	A6801-BLIND	312	186	126
B5301-BD2009	576	190	386	B5301-BLIND	296	120	176
A3301-BD2009	1579	223	1356	A3301-BLIND	288	161	127
B5801-BD2009	2371	367	2004	B5801-BLIND	275	158	117
A3002-BD2009	867	230	637	A3002-BLIND	267	103	164
A3201-BD2009	569	275	294	A3201-BLIND	256	120	136
B4402-BD2009	1262	101	1161	B4402-BLIND	250	129	121
B2705-BD2009	2294	335	1959	B2705-BLIND	249	26	223
B4001-BD2009	2288	274	2014	B4001-BLIND	243	100	143
A2301-BD2009	1429	269	1160	A2301-BLIND	242	86	156
B0803-BD2009	217	9	208	B0803-BLIND	234	9	225
B4403-BD2009	474	96	378	B4403-BLIND	205	103	102
B1503-BD2009	404	331	73	B1503-BLIND	165	33	132
A0202-BD2009	2212	1003	1209	A0202-BLIND	126	44	82
A2902-BD2009	1736	427	1309	A2902-BLIND	118	27	91
B5401-BD2009	577	127	450	B5401-BLIND	79	5	74
B4002-BD2009	450	154	296	B4002-BLIND	74	18	56
B4501-BD2009	459	101	358	B4501-BLIND	65	5	60

Table S9: Comparison of AUC scores on independent (BLIND) dataset for human MHC-I alleles. Proposed methods display state-of-the-art prediction performance when tested across 46 human MHC-I alleles. Observed improvements in ROC-n scores (low FP rates) are significant (paired signed rank test P-values=7e-3 and 1.38e-2 for hkSVM and HONN+hkSVM, respectively).

	NetMHC				hkSVM				HONN				HONN+hkSVM			
Allele	AUC	ROC-1	ROC-5	ROC-10	AUC	ROC-1	ROC-5	ROC-10	AUC	ROC-1	ROC-5	ROC-10	AUC	ROC-1	ROC-5	ROC-10
A0101-BLIND	**92.84**	9.09	26.23	**36.49**	88.04	11.69	17.14	23.77	88.60	7.79	**27.79**	35.84	88.53	**16.88**	18.96	25.45
A0201-BLIND	**93.83**	2.57	**9.09**	**16.85**	**93.31**	**3.97**	**9.05**	14.13	91.80	2.57	8.70	14.96	**93.31**	**3.97**	**9.05**	14.01
A0202-BLIND	**89.05**	50.00	60.45	67.95	**88.83**	**61.36**	**66.82**	**72.27**	86.34	43.18	61.36	67.73	88.69	56.82	65.45	71.59
A0203-BLIND	**97.12**	67.06	79.22	84.08	**96.90**	**69.02**	**81.57**	**85.14**	96.00	38.82	70.98	78.27	**96.89**	67.84	81.49	85.10
A0206-BLIND	**89.00**	13.15	44.22	**57.21**	**88.91**	20.72	**48.84**	**57.17**	87.20	**26.69**	35.22	46.22	**88.94**	20.72	48.53	**57.17**
A0301-BLIND	**92.43**	4.85	11.39	22.76	91.05	**9.39**	17.52	24.24	**92.12**	8.48	**23.15**	**36.36**	91.34	8.79	17.45	24.61
A1101-BLIND	**95.19**	9.33	**41.66**	**52.38**	**94.77**	13.99	38.55	48.50	94.13	**25.91**	35.85	49.07	**94.79**	13.47	39.27	48.81
A2301-BLIND	**91.31**	8.14	**26.98**	**36.05**	88.93	16.28	24.88	32.56	85.52	9.30	17.44	26.28	88.93	**18.60**	24.42	32.91
A2402-BLIND	**84.11**	6.62	15.59	**22.50**	82.13	**14.71**	18.82	21.84	81.19	5.88	16.03	21.18	82.14	13.97	**19.56**	**22.28**
A2501-BLIND	**99.17**	40.00	**72.00**	**76.00**	96.01	40.00	52.00	56.00	**98.59**	**60.00**	60.00	60.00	96.45	40.00	52.00	60.00
A2601-BLIND	**94.51**	16.28	24.42	**30.93**	92.61	**20.93**	**25.81**	30.47	91.90	15.12	19.30	27.21	92.74	18.60	24.88	29.53
A2602-BLIND	93.16	20.20	41.21	50.00	94.24	**33.33**	49.70	59.29	90.99	24.24	29.29	36.16	**94.27**	31.31	**50.91**	**60.00**
A2603-BLIND	**89.01**	21.62	35.68	45.68	**88.98**	**37.84**	**43.78**	**45.95**	70.43	2.70	8.11	16.22	**88.87**	27.03	42.16	44.59
A2902-BLIND	86.94	59.26	59.26	61.11	87.71	**59.26**	62.22	66.30	**88.81**	40.74	60.00	**68.89**	87.87	**59.26**	**62.96**	67.78
A3001-BLIND	**92.48**	5.30	12.32	21.92	89.56	**7.28**	15.23	25.50	91.49	**9.93**	**29.27**	**39.27**	89.96	7.28	14.97	25.36
A3002-BLIND	77.94	0.99	4.36	7.72	77.74	2.97	**8.51**	**12.08**	77.62	**3.96**	7.52	10.10	**77.97**	2.97	7.92	11.58
A3101-BLIND	**88.65**	2.28	14.75	21.22	87.92	8.37	**17.26**	21.33	87.57	**9.51**	13.23	18.78	**88.03**	8.37	**17.26**	**21.56**
A3201-BLIND	83.27	6.67	14.67	22.25	84.55	**10.83**	**19.83**	**27.08**	77.03	6.67	10.33	15.50	**84.71**	**10.83**	**19.83**	26.92
A3301-BLIND	**93.06**	0.00	19.13	43.60	92.20	**6.21**	**34.16**	43.79	91.48	4.35	21.49	34.22	**92.42**	**6.21**	33.91	**44.72**
A6801-BLIND	**94.00**	24.73	**48.17**	**64.03**	93.06	26.34	36.13	54.30	92.35	**30.11**	42.80	53.49	93.09	26.34	36.77	54.57
A6802-BLIND	**96.93**	**81.91**	**82.98**	**84.31**	**96.36**	54.26	69.15	77.93	95.48	49.47	62.02	67.55	**96.38**	55.32	69.47	78.09
A6901-BLIND	**96.01**	34.62	41.54	**56.15**	93.59	**38.46**	**41.54**	43.85	92.30	3.85	20.77	31.15	93.54	34.62	40.00	41.54
A8001-BLIND	**98.35**	28.57	51.43	54.29	97.70	28.57	45.71	51.43	**98.16**	**71.43**	**71.43**	**71.43**	97.70	28.57	45.71	54.29
B0702-BLIND	**92.09**	20.55	27.29	32.88	**91.77**	28.22	**32.60**	**37.97**	87.79	22.47	28.55	32.14	91.73	**29.59**	32.27	37.23
B0801-BLIND	**97.19**	23.04	27.02	**36.02**	96.21	**24.08**	**28.06**	35.86	93.60	8.90	21.26	29.16	96.22	**24.08**	27.75	35.65
B0802-BLIND	**98.99**	5.56	**57.78**	**67.78**	**98.55**	27.78	40.00	48.33	92.15	**33.33**	34.44	41.11	**98.69**	**33.33**	42.22	51.67
B0803-BLIND	**96.84**	0.00	13.33	36.67	95.56	0.00	**20.00**	**42.22**	87.51	0.00	13.33	17.78	95.56	0.00	**20.00**	41.11
B1501-BLIND	**93.44**	11.16	20.84	**33.44**	**93.17**	10.23	22.88	30.00	91.67	**13.49**	**24.65**	32.79	**93.23**	12.09	23.07	30.79
B1503-BLIND	**86.48**	**21.21**	**26.67**	**34.55**	77.43	18.18	23.64	27.27	68.39	0.00	12.12	16.97	77.18	18.18	23.64	27.27
B1509-BLIND	**90.10**	17.24	**24.14**	**29.31**	76.35	**17.24**	20.00	20.34	74.08	0.00	8.28	12.76	76.95	**17.24**	19.31	20.00
B1517-BLIND	90.27	24.73	**36.34**	**41.61**	**91.72**	**27.96**	34.84	39.57	84.90	6.45	24.30	35.70	91.71	**27.96**	33.98	38.49
B1801-BLIND	79.47	3.57	**19.29**	**23.93**	**81.87**	**3.57**	10.00	13.93	74.23	0.00	1.43	6.07	**81.87**	**3.57**	9.29	13.57
B2705-BLIND	**95.34**	3.85	12.31	25.77	92.51	3.85	22.31	43.85	90.91	**7.69**	**26.92**	40.38	92.57	3.85	24.62	**45.38**
B3501-BLIND	84.48	**5.56**	**19.38**	**27.10**	86.21	4.32	16.54	26.23	84.70	2.47	18.02	25.31	**86.31**	4.32	16.67	26.36
B3801-BLIND	92.57	0.00	0.43	9.08	**94.96**	**1.42**	**29.79**	**47.45**	88.42	0.00	2.41	9.15	94.56	0.00	23.83	41.63
B3901-BLIND	96.51	40.00	55.20	**65.60**	**97.12**	44.00	51.20	57.60	96.30	**64.00**	**64.00**	64.00	**97.31**	48.00	54.40	59.20
B4001-BLIND	**94.88**	19.00	**48.40**	**57.60**	93.06	29.00	41.40	48.30	93.12	14.00	43.20	51.30	93.53	**31.00**	42.00	49.00
B4002-BLIND	**92.06**	**77.78**	**81.11**	**85.00**	91.37	44.44	57.78	70.00	84.52	16.67	47.78	58.89	**91.87**	44.44	61.11	72.78
B4402-BLIND	**93.45**	26.36	**45.12**	**54.81**	**93.10**	**26.36**	37.83	45.74	90.87	1.55	16.43	31.55	**93.14**	22.48	36.90	46.82
B4403-BLIND	**89.35**	**26.21**	**35.73**	**46.89**	85.94	17.48	23.69	31.17	84.61	22.33	29.71	36.41	86.37	17.48	24.08	32.04
B4501-BLIND	100.00	100.00	100.00	100.00	**100.00**	**100.00**	**100.00**	**100.00**	99.67	80.00	96.00	98.00	**100.00**	**100.00**	**100.00**	**100.00**
B5101-BLIND	**92.91**	**21.74**	**26.52**	**30.65**	92.09	15.22	20.87	24.35	91.65	19.57	23.04	26.30	92.33	17.39	22.61	25.87
B5301-BLIND	**86.98**	7.50	11.67	18.58	84.80	5.00	12.50	19.83	81.14	**7.50**	**14.33**	19.00	84.82	5.00	11.83	**20.08**
B5401-BLIND	81.62	80.00	80.00	80.00	**82.16**	**80.00**	**80.00**	**80.00**	81.08	**80.00**	**80.00**	**80.00**	**82.16**	**80.00**	**80.00**	**80.00**
B5701-BLIND	**96.93**	**37.93**	**43.45**	50.48	**96.77**	24.83	40.00	**51.24**	94.93	15.86	24.00	32.48	**96.76**	24.83	39.59	50.34
B5801-BLIND	**95.85**	27.22	44.81	67.09	94.67	44.94	54.56	63.86	**95.61**	**53.80**	**61.52**	**71.39**	94.79	44.30	55.57	64.68
Median	**92.875**	19.6	31.485	42.605	92.145	**20.825**	**33.38**	**43.82**	90.89	11.71	24.475	34.96	**92.375**	19.66	**33.09**	41.585

Table S10: Peptide dissimilarity between train (IEDB) and test (Japanese) datasets. Sequence identity column shows average sequence identity between test peptide sequence and most similar peptide sequence in the training dataset.

Train dataset	Test dataset	Sequence identity
A0201-IEDB	A0201-Japanese	55.2%
A0206-IEDB	A0206-Japanese	46.8%
A2402-IEDB	A24020-Japanese	51.1%

4 Naturally-processed peptide prediction

To train our models, we used the data from Machine Learning in Immunology competition (MLI-II) `http://bio.dfci.harvard.edu/DFRMLI/HTML/natural.php`. For each of the MHC-I alleles, a set of 8- to 11-mer peptides is given for training and testing. We directly train our models to recognize naturally processed and presented (NP) peptides, using "eluted" peptides as a positive set, and all other peptides (non-binders + non-eluted binders) as a negative set. We then test our models on the data composed of non-eluted binding peptides, non-binding peptides, and naturally processed ("eluted") peptides. We used the same training and test split as specified in the competition. In Table S11, we compare our approach hkSVM with the popular NetMHC method, which was used as a benchmark in the competition, as well as the recently introduced MHC-NP (Gigure *et al.*, 2013) method that yielded state-of-the-art accuracy for naturally processed (NP) peptide prediction.

Table S11: Prediction of naturally processed and presented (NP) peptides (MLI-II competition). Comparison of AUC scores.

Allele	MHC-NP (E vs B+N)† AUC	MHC-NP† AUC	NetMHC† AUC	Our method AUC
B0702	91.57	92.36	90.42	**93.45**
B3501	91.87	**93.67**	90.9	**93.54**
B4403	81.78	79.47	81.04	**90.35**
B5301	84.01	85.15	66.51	**91.14**
B5701	80.17	**82.58**	81.81	**82.93**

† quoted from (Gigure *et al.*, 2013)

References

Buus, S., Lauemøller, S., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A., and Brunak, S. (2003). Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, **62**(5), 378–384.

Gigure, S., Drouin, A., Lacoste, A., Marchand, M., Corbeil, J., and Laviolette, F. (2013). MHC-NP: Predicting peptides naturally processed by the MHC. *Journal of Immunological Methods*, **400401**(0), 30 – 36.

Kim, Y., Sidney, J., Buus, S., Sette, A., Nielsen, M., and Peters, B. (2014). Dataset size and composition impact the reliability of performance benchmarks for peptide-mhc binding predictions. *BMC Bioinformatics*, **15**(1), 241.

Lundegaard, C., Lund, O., and Nielsen, M. (2011). Prediction of epitopes using neural network based methods. *Journal of Immunological Methods*, **374**(1–2), 26 – 34. High-throughput methods for immunology: Machine learning and automation.

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science*, **12**(5), 1007–1017.