Supplementary Information for

Docking Model Evaluation by 3D Deep Convolutional Neural Networks

Xiao Wang[1], Genki Terashi[2], Charles W. Christoffer[1], Mengmeng Zhu[2], and Daisuke Kihara[2, 1, *]

[1]Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA, [2]Department of Biological Sciences, Purdue University, West Lafayette, IN, 47907, USA.

* Contact: dkihara@purdue.edu

Supplementary Table S1. Grouping of ZDock benchmark dataset by structural similarity.

Group 1: 1AKJ
Group 2: 1AY7
Group 3: 1AZS
Group 4: 1B6C
Group 5: 1BUH, 1FQ1, 1JWH, 2OZA
Group 6: 1DFJ
Group 7: 1FC2, 1FCC
Group 8: 1A2K, 1I2M, 1BKD, 1HE1, 1HE8, 1WQ1, 2FJU, 2OT3, 3CPH, 1I4D, 2NZ8
Group 9: 1EER
Group 10: 1EFN, 1GCQ
Group 11: 1F34
Group 12: 1F51
Group 13: 2O8V
Group 14: 1FFW
Group 15: 3D5S
Group 16: 1GLA
Group 17: 1GPW
Group 18: 1GXD, 2J0T
Group 19: 1H9D
Group 20: 1HCF
Group 21: 1IJK
Group 22: 1JIW
Group 23: 1JK9
Group 24: 1JTG
Group 25: 1JZD, 1Z5Y
Group 26: 1K74
Group 27: 1KAC
Group 28: 1SBB
Group 29: 1KTZ
Group 30: 1MAH
Group 31: 1ML0
Group 32: 1MQ8

Group 33: 1NW9
Group 34: 1OC0
Group 35: 1ACB, 1AVX, 1CGI, 1D6R, 1EAW, 1EZU, 1FAK, 1FLE, 1GL1, 1HIA, 2SNI, 2UUY, 1AHW, 1BJ1, 1BVK, 1DQJ, 1E6J, 1FSK, 1I9R, 1IQD, 1JPS, 1KXQ, 1VFB, 1WEJ, 2FD6, 2JEL, 1AK4, 1RV6, 2I25, 1BVN, 1CLV, 1TMQ, 1R0R, 3SGQ, 1K4C, 1QFW, 2SIC, 2I9B
Group 36: 1PVH
Group 37: 1PXV
Group 38: 1QA9
Group 39: 1R6Q
Group 40: 1RLB
Group 41: 1S1Q, 1XD3, 2AYO, 2OOB
Group 42: 1SYX
Group 43: 1T6B
Group 44: 1UDI
Group 45: 1US7
Group 46: 1XQS
Group 47: 1XU1
Group 48: 1ZHI
Group 49: 2ABZ
Group 50: 2A5T
Group 51: 2AJF
Group 52: 2B42
Group 53: 2B4J
Group 54: 2CFH
Group 55: 2HLE
Group 56: 2HQS
Group 57: 2HRK
Group 58: 2IDO
Group 59: 2MTA
Group 60: 2O3B
Group 61: 2OOR
Group 62: 2VDB
Group 63: 2Z0E

120 target protein complexes from ZDock benchmark (Ver. 4.0) was classified into groups considering their structural similarity. If both of the two protein structures of complexes have over a TM-score of 0.5 and the sequence identity over 30%, then the complexes were clustered in the same group.

Supplementary Table S2. Splits of the 63 target groups for training and testing the network.

Split	PDB ID
1	1AKJ(1), 1B6C(4), 1FFW(14), 1GXD(18), 2J0T(18), 1IJK(21), 1JK9(23), 1KAC(27), 1ML0(31), 1OC0(34), 1PXV(37), 1S1Q(41), 1XD3(41), 2AYO(41), 2OOB(41), 1US7(45), 2B4J(53), 2HQS(56), 2MTA(59), 2Z0E(63)
2	1BUH(5), 1FQ1(5), 1JWH(5), 2OZA(5), 1A2K(8), 1I2M(8), 1BKD(8), 1HE1(8), 1HE8(8), 1WQ1(8), 2FJU(8), 2OT3(8), 3CPH(8), 1I4D(8), 2NZ8(8), 1F34(11), 3D5S(15), 1H9D(19), 1JTG(24), 1SBB(28), 1MQ8(32), 1QA9(38), 1SYX(42), 2A5T(50), 2HRK(57), 2O3B(60)
3	1AY7(2), 1DFJ(6), 1EER(9), 1F51(12), 1GLA(16), 1HCF(20), 1JZD(25), 1Z5Y(25), 1KTZ(29), 1ACB(35), 1AVX(35), 1CGI(35), 1D6R(35), 1EAW(35), 1EZU(35), 1FAK(35), 1FLE(35), 1GL1(35), 1HIA(35), 2SNI(35), 2UUY(35), 1AHW(35), 1BJ1(35), 1BVK(35), 1DQJ(35), 1E6J(35), 1FSK(35), 1I9R(35), 1IQD(35), 1JPS(35), 1KXQ(35), 1VFB(35), 1WEJ(35), 2FD6(35), 2JEL(35), 1AK4(35), 1RV6(35), 2I25(35), 1BVN(35), 1CLV(35), 1TMQ(35), 1R0R(35), 3SGQ(35), 1K4C(35), 1QFW(35), 2SIC(35), 2I9B(35), 1R6Q(39), 1T6B(43), 1XQS(46), 1ZHI(48), 2AJF(51), 2CFH(54), 2IDO(58), 2OOR(61)
4	1AZS(3), 1FC2(7), 1FCC(7), 1EFN(10), 1GCQ(10), 2O8V(13), 1GPW(17), 1JIW(22), 1K74(26), 1MAH(30), 1NW9(33), 1PVH(36), 1RLB(40), 1UDI(44), 1XU1(47), 2ABZ(49), 2B42(52), 2HLE(55), 2VDB(62)

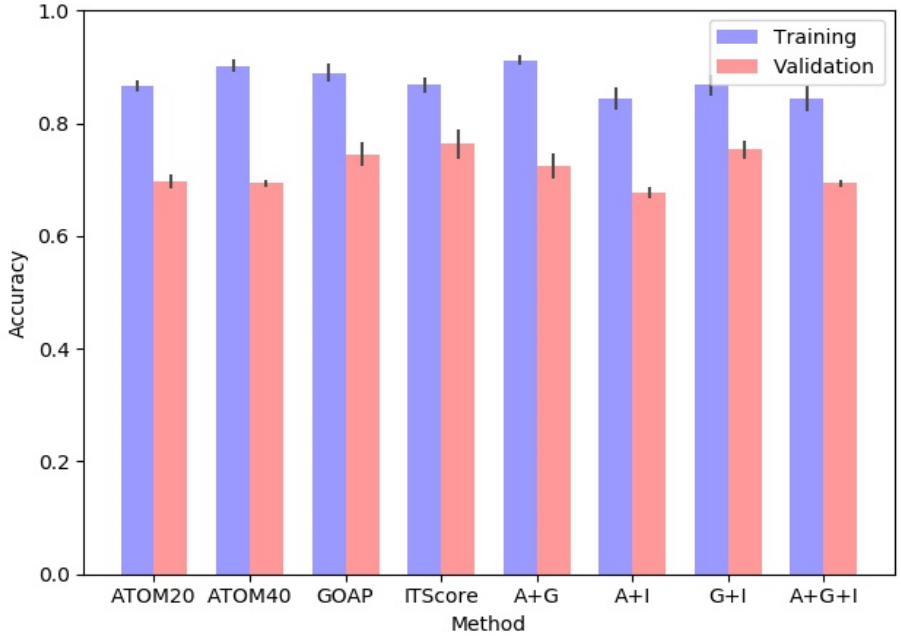
These are the four subsets of the 63 groups that were used for the four-fold cross validation. In the parentheses, the group IDs from the Supplementary Table S1 are shown.

Supplementary Table S3. Hyper-parameters determined during the training and validation.

Features	Learning Rate	L2 Regularization
ATOM20	0.0005	0.0002
	0.0005	0.0002
	0.0005	0.0002
	0.0008	0.0008
ATOM40	0.001	0.0001
	0.0005	0.0001
	0.0005	0.001
	0.0005	0.0003
GOAP	0.0005	0.0001
	0.0005	0.0001
	0.0005	0.0001
	0.0005	0.0001
ITScore	0.0005	0.0002
	0.0006	0.0002
	0.0005	0.0002
	0.0005	0.0002
ATOM40-GOAP	0.0005	0.0002
	0.0005	0.0002
	0.0005	0.0003
	0.0005	0.0005
ATOM40-ITScore	0.0005	0.0002
	0.0005	0.0002
	0.0006	0.0004
	0.0005	0.0001
GOAP-ITScore	0.0005	0.0002
	0.0005	0.0002
	0.0005	0.0002
	0.0005	0.0002
ATOM40-GOAP-ITScore	0.0005	0.0001
	0.0005	0.0002
	0.0005	0.0002
	0.0005	0.0002

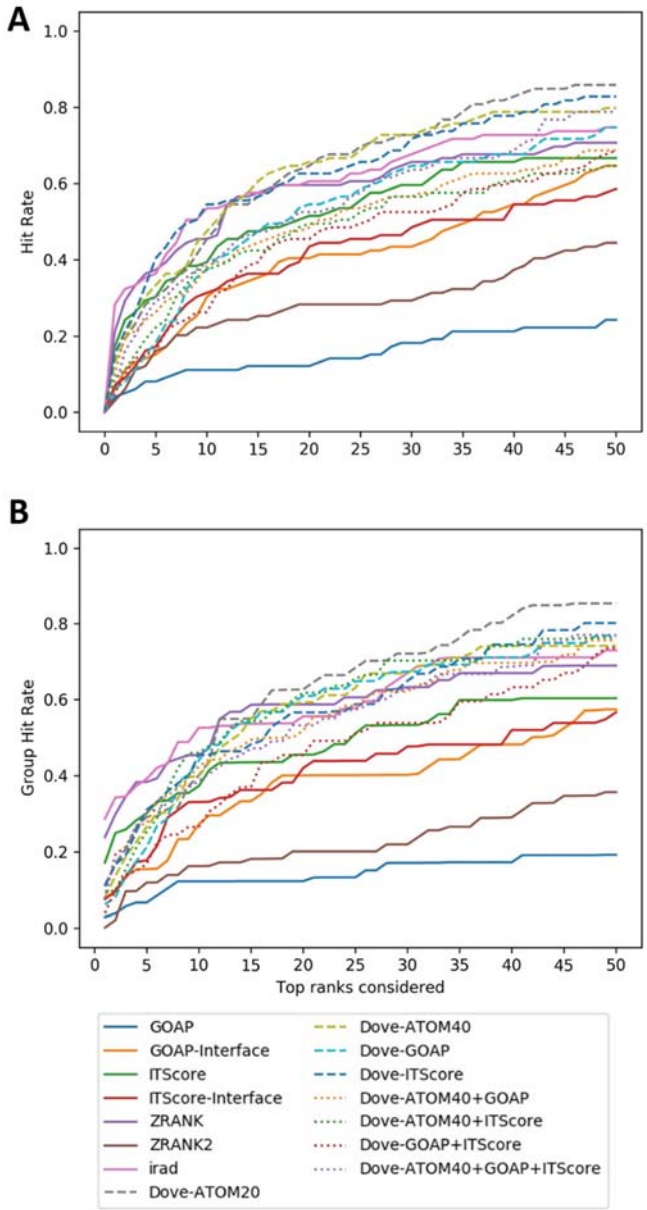
For each feature combination, four hyper-parameter value combinations are shown that come from four-fold cross validation. The range of the values tested were 1e-9, 1e-8,1e-7,1e-6,..., 0.1. After the best combinations of the learning rate and the L2 normalization were determined from these combinations, then values within the ranges were randomly tried to seek for better results.

Supplementary Figure S1. Prediction accuracy in Training and validation.



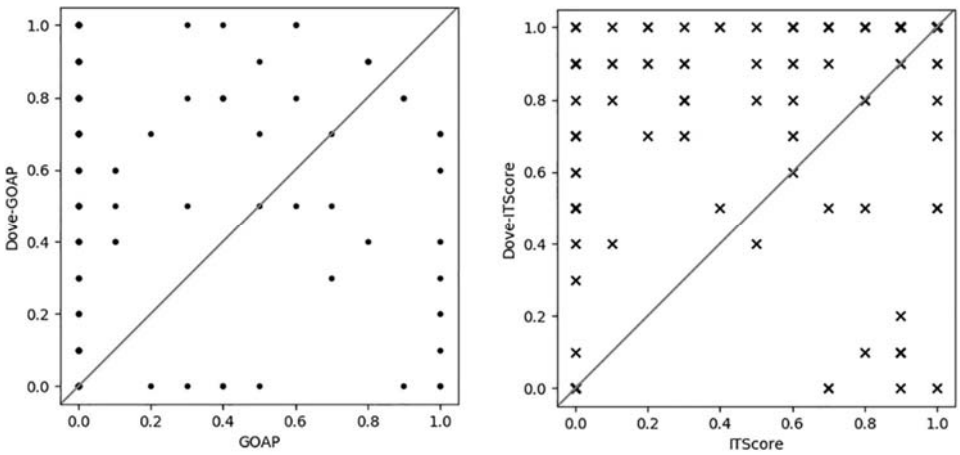
The accuracies (i.e. the number of correctly recognized correct and incorrect decoys over the total number of decoys) during the training and validation phase using eight different feature combinations are shown. Results shown are the average and the standard deviation observed in the four training and four validation sets in the four-fold cross validation.
The feature combinations are indicated with abbreviations: A+G, Atom40+GOAP; A+I, Atom+ITScore; G+I, GOAP+ITscore; A+G+I, Atom40+GOAP+ITScore.

Supplementary Figure S2. Comparison on the ZDOCK Benchmark dataset considering only medium quality models. 99 complexes among the 120 complexes in the benchmark set that hare at least one medium quality models were used for this evaluation. On average there were 51.0 medium quality models in a decoy set of a complex.
A, The fraction of complexes among the 99 complexes for which each method selected at least one medium quality model (within top x scored models) was plotted. Results shown are from test sets. In addition to DOVE with eight different feature combinations, performance of GOAP, GOAP-Interface, ITScore, ITScore-Interface, Zrank, Zrank2, and irad are shown. **B**, Considering the similar complexes that were grouped into 54 groups (Supplementary Table S1), the hit rates for complexes in each group were averaged and re-averaged over the 54 groups for each x.



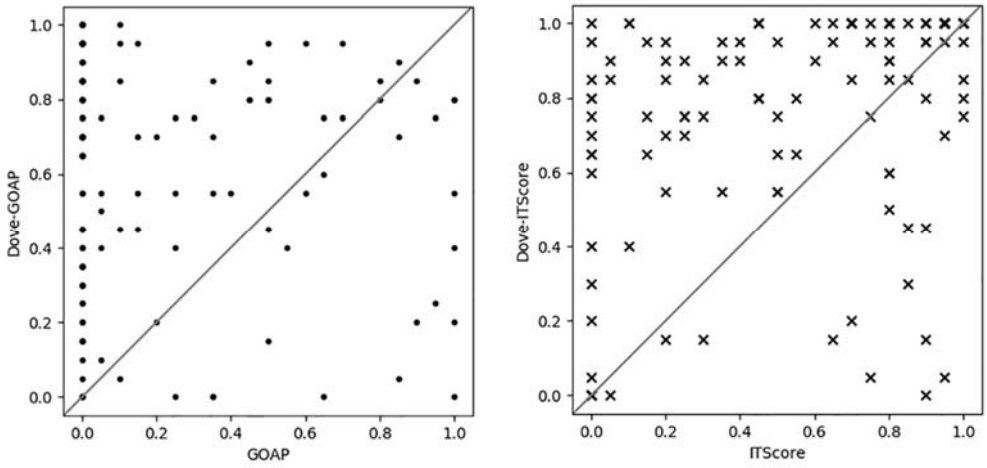
Supplementary Figure S3. Comparison of the fraction of correct decoys among top 10 and top 50 by GOAP/ITScore and DOVE-GOAP/ITScore. Results on 120 targets were compared. Left panel, GOAP; right panel, ITScore. Comparison on the top 20 ranked decoys are shown as Fig. 4 in the main manuscript.

Top 10



Dove-GOAP showed a higher hit fraction than GOAP for 85 cases, tied for 14 cases, and worse for 21 cases. Dove-ITScore showed a higher hit fraction than ITScore for 73 cases, tied for 29 cases, and worse for 18 cases.

Top 50

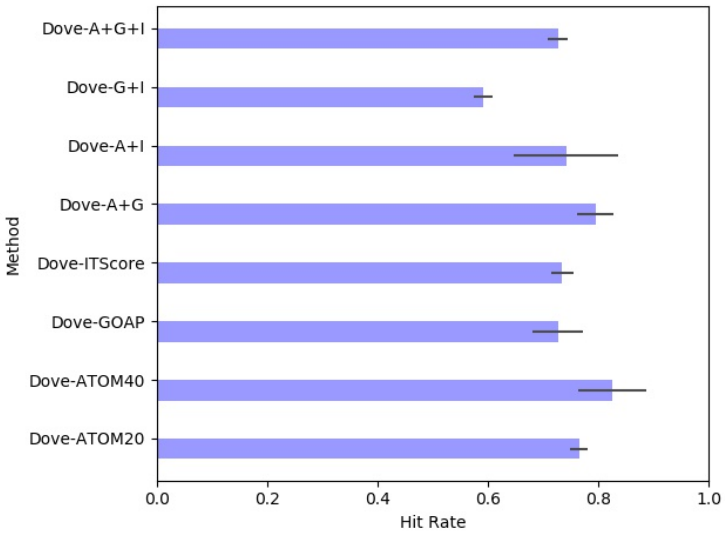


Dove-GOAP showed a higher hit fraction than GOAP for 100 cases, tied for 4 cases, and worse for 16 cases. Dove-ITScore showed a higher hit fraction than ITScore for 91 cases, tied for 5 cases, and worse for 24 cases.

Supplementary Figure S3. The average and the standard deviation of the top 10 hit rates of Dove on the Dockground dataset.

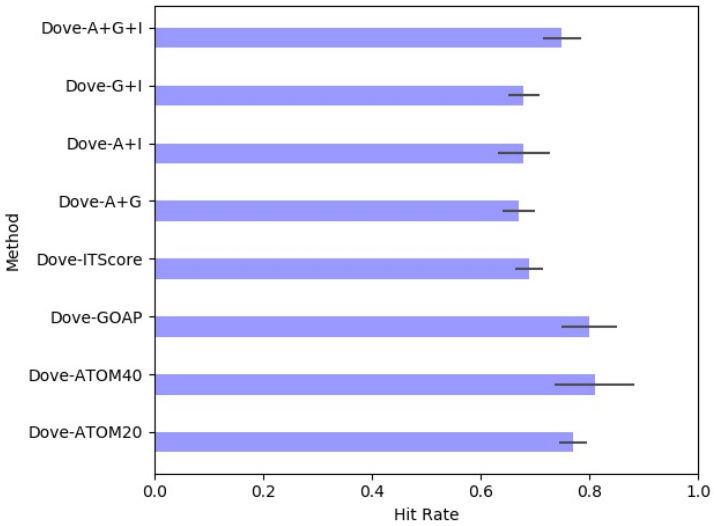
Results using four models from the four-fold cross validation on the ZDOCK dataset were used to compute the average and standard deviation.

A. 33 independent targets.



The 33 target complexes are not similar (do not satisfy TM-score > 0.5 for both proteins in a complex) to any of the complexes in the ZDOCK benchmark dataset. A+G, Atom40+GOAP; A+I, Atom40 + ITScore; G+I, GOAP + ITScore; A+G+I, Atom40+GOAP+ITscore. The standard deviation values were: 0.019, 0.017, 0.094, 0.033, 0.020, 0.046, 0.061, and 0.017, respectively, from the top to the bottom bar.

B. 25 targets that are similar to complexes in ZDOCK dataset.



The standard deviations were 0.036, 0.028, 0.047, 0.030, 0.026, 0.051, 0.073, and 0.026, respectively, from the top to the bottom bar.