STRUCTURAL
BIOLOGY

Volume 79 (2023)

Supporting information for article:

A scoring function for the prediction of protein complex interfaces based on the neighborhood preferences of amino acids

Mulpuri Nagaraju and Haiguang Liu

Table S1 Performance of scoring functions, CAPRI, Nepre and iNepre on CAPRI decoy sets.

Decoys Sets (15 complexes)	Total number of accepted CAPRI models in X × 15 predictions	Number of selected models	Percentage of acceptable models		
			CAPRI	Nepre	iNepre
T (29,30,32,35, 36,37,38,39,40, 41,46,47,50,53, 54)	863	1	0.0	0.1	0
		5	0.3	0.2	0.4
		10	0.5	0.7	0.8
		20	1.1	1.2	2.4
		50	3.6	3.7	5.9

Table S2 Performance of scoring functions on individual CAPRI decoy sets.

Decoy Sets (Predictions X)	Number of accepted CAPRI models in X predictions	Number of selected models	Number of acceptable models		
			CAPRI	Nepre	iNepre
T29 (2083)	87	1	-	-	-
		5	-	-	-
		10	-	-	1
		20	-	-	1
		50	-	3	1
T30 (1343)	2	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-
		50	-	-	-
T32 (599)	12	1	-	-	-
		5	-	-	-
		10	-	1	-
		20	2	1	1
		50	3	4	2
T35 (499)	3	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-
		50	-	-	-
T36 (309)	NO	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-

		50	-	-	-
T37 (1500)	42	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-
		50	-	-	-
T38 (899)	NO	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-
		50	-	-	-
T39 (1400)	1	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-
		50	-	-	-
T40 (2180)	189	1	-	-	-
		5	-	-	1
		10	-	-	1
		20	1	-	4
		50	4	-	16
T41 (1200)	249	1	-	-	-
		5	2	-	1
		10	2	-	2
		20	4	-	5
		50	12	-	15
T46 (1699)	24	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-

		50	-	-	-
T47 (1051)	26	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-
		50	2	-	-
T50 (1451)	97	1	-	-	-
		5	1	-	1
		10	2	1	1
		20	3	2	3
		50	6	10	3
T53 (1400)	113	1	-	1	-
		5	-	2	1
		10	-	4	2
		20	-	7	7
		50	3	15	14
T54 (1400)	18	1	-	-	-
		5	-	-	-
		10	-	-	-
		20	-	-	-
		50	1	-	-

Table S3 Performance of HADDOCK and ZDOCK scoring energies vs RMSD with respect to native structure. In HADDOCK, interfacial RMSD (i-RMSD) and ligand RMSD (l-RMSD) are taken, whereas in ZDOCK top scored 1000 complexes are considered out of 2000 complexes generated in ZDOCK procedure and RMSD for these complexes are calculated based on native structure. RMSD cutoff 2.5, 3.5 and 5Å are used to identify top rank models in ZDOCK decoy datasets.

Number of selected models	HADDOCK Decoy dataset (25)		ZDOCK Set-II Decoy dataset (135)			ZDOCK Set-I Decoy dataset (43)		
	i-RMSD	l-RMSD	2.5Å	3.5Å	5Å	2.5Å	3.5Å	5Å
1	3	0	4	6	11	0	0	0
5	4	0	6	10	18	0	0	0
10	5	2	7	12	21	0	0	1
20	6	3	7	13	24	0	1	1
50	7	3	7	14	28	0	1	1

Table S4 A benchmark study on interfacial residue cutoff distance (at 4Å, 5Å and 6Å) in iNepre scoring function, in all cases amino acid residues neighborhood distance cutoff is 6.0 Å.

Number of selected models	HADDOCK Decoy dataset (25)			GRAMM-X Decoy dataset (43)			ZDOCK Set-II Decoy dataset (130)			ZDOCK Set-I Decoy dataset (36)		
	4Å	5Å	**6Å**	4Å	5Å	**6Å**	4Å	5Å	**6Å**	4Å	5Å	**6Å**
1	1	9	**13**	2	5	**13**	6	26	**38**	2	4	**12**
5	4	9	**13**	7	8	**15**	11	31	**51**	3	6	**13**
10	6	10	**14**	8	10	**17**	17	39	**59**	5	6	**14**
20	7	12	**15**	11	16	**18**	28	49	**69**	7	7	**16**
50	7	12	**17**	25	22	**25**	65	71	**87**	16	14	**21**

Table S5 Testing the convergence of interfacial amino acid residue pair data points collected at 6.0 Å interfacial distance cutoff. Amino acid – Amino acid data points are considered at 50%, 75% and 100% to generate energy matrix file, in all cases amino acid residues neighborhood distance cutoff is 6.0 Å.

Number of selected models	HADDOCK Decoy dataset (25) % Data			GRAMM-X Decoy dataset (43) % Data			ZDOCK Set-II Decoy dataset (130) % Data			ZDOCK Set-I Decoy dataset (36) % Data		
	50%	75%	**100%**	50%	75%	**100%**	50%	75%	**100%**	50%	75%	**100%**
1	9	12	**13**	11	12	**13**	38	37	**38**	8	10	**12**
5	12	13	**13**	14	16	**15**	46	47	**51**	10	13	**13**
10	12	14	**14**	17	18	**17**	53	54	**59**	11	13	**14**
20	12	15	**15**	20	21	**18**	67	67	**69**	15	15	**16**
50	15	18	**17**	26	26	**25**	84	86	**87**	20	23	**21**

Table S6 Protein complexes used for the ZDOCK Set-II data set.

Complex	Category [a]	Complex	Category [a]	Complex	Category [a]
1AHW_AB:C	AA	1TMQ_A:B	EI	1GPW_A:B	OX
1DQJ_AB:C	AA	1UDI_E:I	EI	1H9D_A:B	OX
1JPS_HL:T	AA	1US7_A:B	ER	1HCF_AB:X	OR
1MLC_AB:E	AA	1WDW_BD:A	ER	1HE1_C:A	OG
1VFB_AB:C	AA	1YVB_A:I	EI	1I4D_D:AB	OG
1WEJ_HL:F	AA	1Z5Y_D:E	ES	1J2J_A:B	OG
2FD6_HL:U	AA	2A9K_A:B	ES	1K74_AB:DE	OR
2I25_N:L	AS	2ABZ_B:E	EI	1KAC_A:B	OR
2VIS_AB:C	AA	2AYO_A:B	ER	1KLU_AB:D	OX
2VXT_HL:I	AA	2B42_B:A	EI	1KTZ_A:B	OR
2W9E_HL:A	AA	2GAF_D:A	ER	1KXP_A:D	OX
3EOA_LH:I	AA	2J0T_A:D	EI	1M27_AB:C	OX
3MXW_LH:A	AA	2MTA_HL:A	ES	1ML0_AB:D	OR
3RVW_CD:A	AA	2O8V_A:B	ES	1OFU_XY:A	OX
4DN4_LH:M	AA	2OOB_A:B	ES	1PVH_A:B	OR
4G6M_HL:A	AA	2OOR_AB:C	ER	1QA9_A:B	OX
1AVX_A:B	EI	2PCC_A:B	ES	1RLB_ABCD:E	OX
1AY7_A:B	EI	2SIC_E:I	EI	1RV6_VW:X	OR
1BUH_A:B	EI	2SNI_E:I	EI	1S1Q_A:B	OX
1BVN_P:T	EI	2UUY_A:B	EI	1SBB_A:B	OR
1CLV_A:I	EI	2YVJ_A:B	ER	1T6B_X:Y	OR
1D6R_A:I	EI	3A4S_A:D	EI	1XD3_A:B	OX
1DFJ_E:I	EI	3K75_D:B	ER	1XU1_ABD:T	OR
1E6E_A:B	ES	3PC8_A:C	ER	1Z0K_A:B	OG
1EAW_A:B	EI	3SGQ_E:I	EI	1ZHH_A:B	OR
1EWY_A:C	ES	3VLB_A:B	EI	1ZHI_A:B	OX
1EZU_C:AB *	EI	4CPA_A:I	EI	2A5T_A:B	OX
1F34_A:B	EI	4H03_A:B	ES	2AJF_A:E	OR
1F51_AB:E	ER	4HX3_BD:A	EI	2B4J_AB:C	OX
1FLE_E:I	EI	7CEI_A:B	EI	2BTF_A:P	OX
1GL1_A:I	EI	1A2K_C:AB	OG	2FJU_B:A	OG

1GLA_G:F	ER	1AK4_A:D	OX	2G77_A:B	OG
1GXD_A:C	EI	1AKJ_AB:DE	OX	2GTP_A:D	OG
1HIA_AB:I	EI	1AZS_AB:C	OG	**2HLE_A:B**	OR
1JTD_B:A	EI	1E96_A:B	OG	2HQS_A:H	OX
1JTG_B:A	EI	1EFN_B:A	OX	2X9A_D:C	OR
1MAH_A:F	EI	1FCC_AB:C	OX	3BIW_A:E	OX
1OC0_A:B	ER	1FFW_A:B	OX	3BP8_AB:C	OX
1OPH_A:B	EI	1FQJ_A:B	OG	3D5S_A:C	OX
1PPE_E:I	EI	1GHQ_A:B	OR	3H2V_A:E	OX

Table S7 Protein complexes used for the ZDOCK Set-I data set and GRAMM-X data sets.

Complex	Category [a]	Complex	Category [a]	Complex	Category [a]
3EO1_AB:CF	AA	2Z0E_A:B	ER	1WQ1_R:G *	OG
3G6D_LH:A	AA	4FZA_A:B	ER	1XQS_A:C	OX
3HI6_XY:B	AA	4IZ7_A:B	EI	2CFH_A:C	OX
3L5W_LH:I	AA	4LW4_AB:C	ES	2H7V_A:C	OG
3V6Z_AB:F	AA	1B6C_A:B	OX	2HRK_A:B	OX
1CGI_E:I	EI	1GP2_A:BG	OG	2OZA_B:A	OX
1IJK_A:BC	ER	1GRN_A:B *	OG	3AAA_AB:C	OX
1JIW_P:I	EI	1HE8_B:A	OG	3AAD_A:D	OX
1KKL_ABC:H	ES	1I2M_A:B	OG	3BX7_A:C	OX
1M10_A:B	ER	1IB1_AB:E	OX	3CPH_G:A	OG
1NW9_B:A	ER	1K5D_AB:C	OG	3DAW_A:B	OX
1R6Q_A:C	ER	1LFD_B:A	OG	3R9A_AC:B	OR
1ZM4_A:B	ES	1MQ8_A:B	OX	3S9D_B:A	OR
2NZ8_A:B	ER	1SYX_A:B	OX	3SZK_DE:F	OX
				4JCV_ADBC:E	OX

Note: The category of protein complexes in Table S6 and Table S7 are labeled as below.

Complex category labels:

EI = Enzyme-Inhibitor
ES = Enzyme-Substrate
ER = Enzyme complex with a regulatory or accessory chain
AA = Antibody-Antigen
AS = Antigen – Single domain Antibody
OG = Others, G-protein containing
OR = Others, Receptor containing
OX = Others, miscellaneous

Table S8 Protein complexes used for the HADDOCK data set. The data set and decoys are downloaded from the following website: https://data.sbgrid.org/dataset/131/

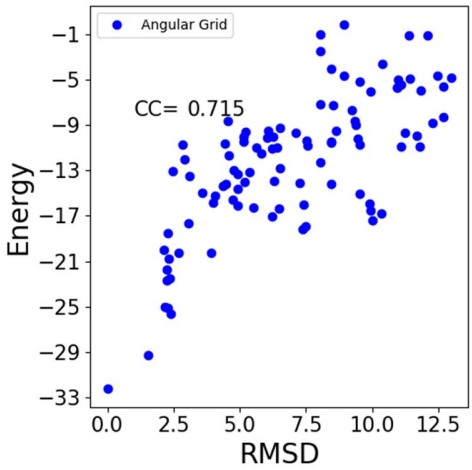
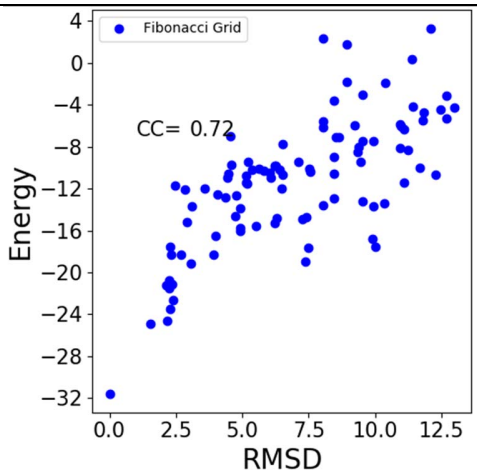
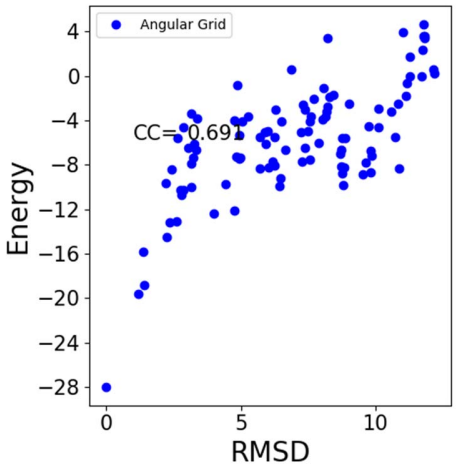
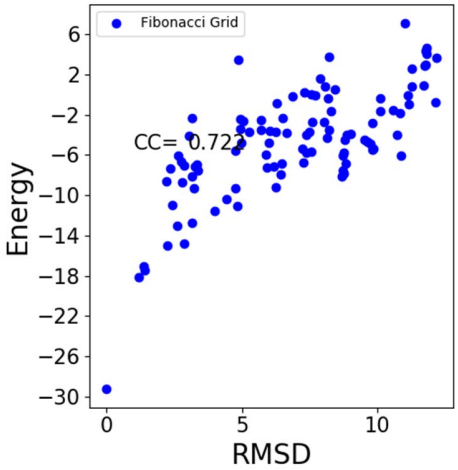
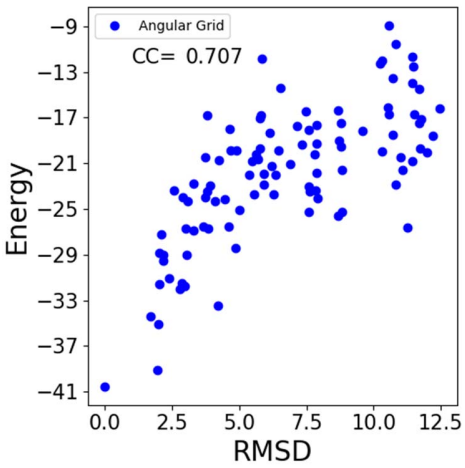
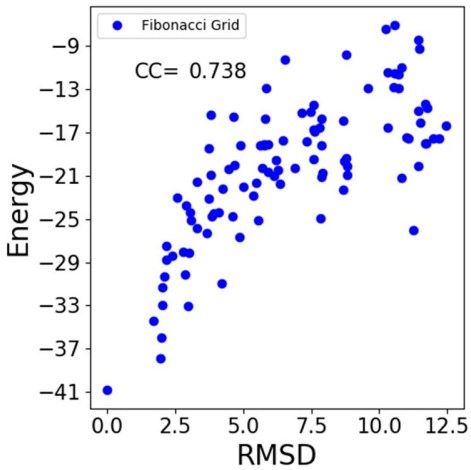
Complex	Category [a]	Complex	Category [a]	Complex	Category [a]
1JTD		3AAA		3MXW	AA
1RKE		3AAD		3RVW	AA
2A1A		3BX7		4DN4	AA
2GAF		3DAW		4FZA	
2GTP		3F1P		4G6J	AA
2VXT	AA	3FN1		4G6M	AA
2W9E	AA	3G6D	AA	4H03	
2YVJ		3H11		4IZ7	
				4M76	

Note: The category of protein complexes in Table S8 labeled as below.

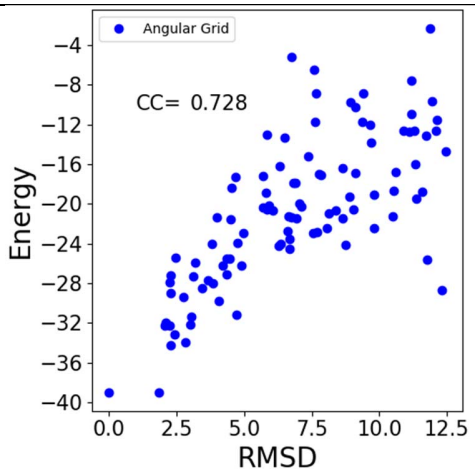
Complex category labels:

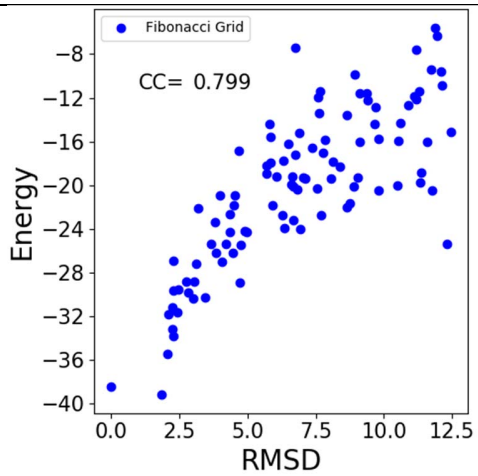
AA = Antibody-Antigen

Others are non-antibody antigen

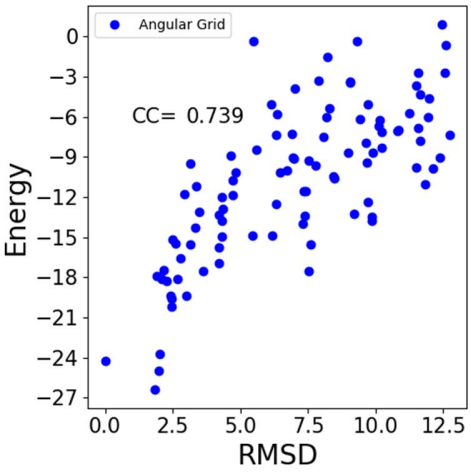
PDB ID	Angular grid	Fibonacci grid
RO1a68		
RO1bq9		
RO1dhn		

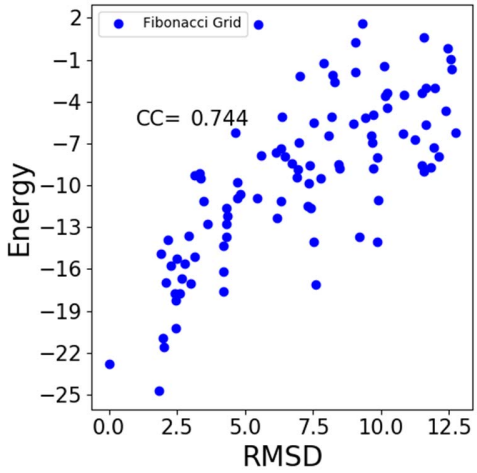
RO1e6i



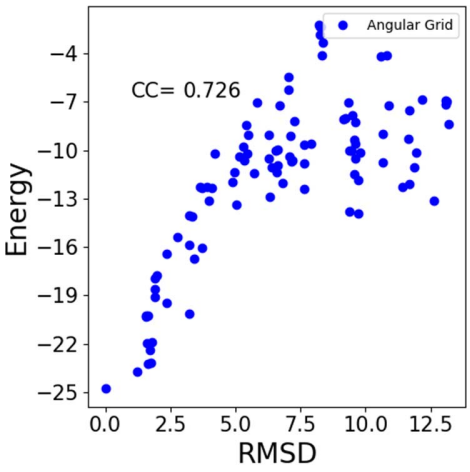


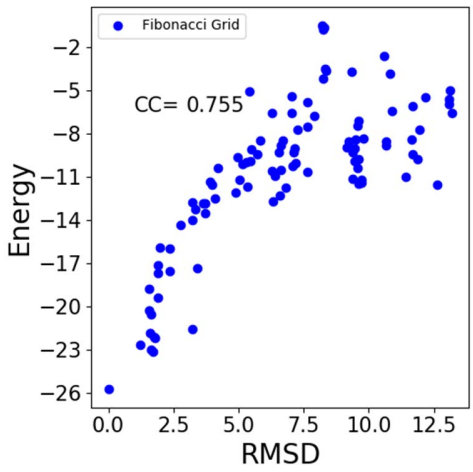
RO1gvp

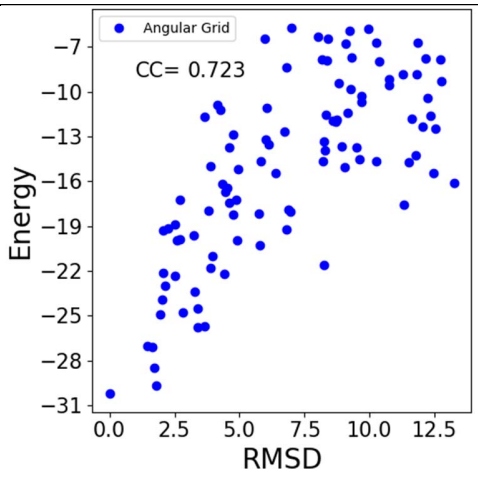
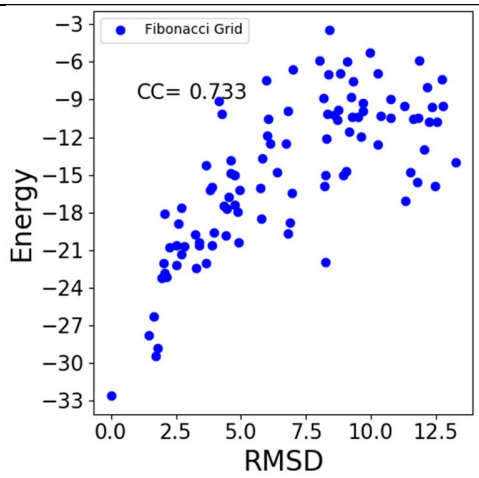
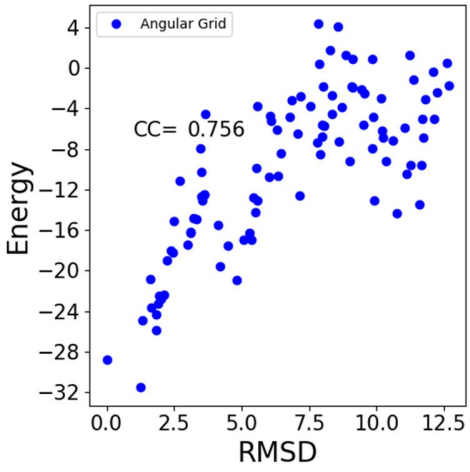
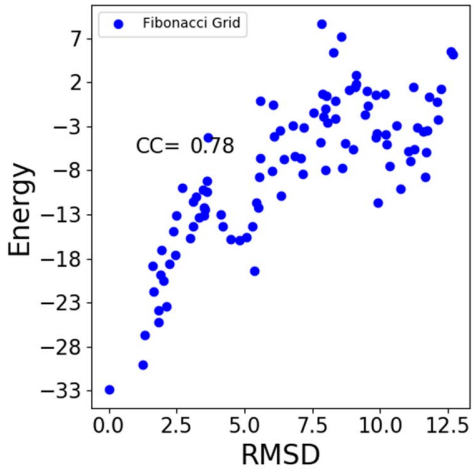
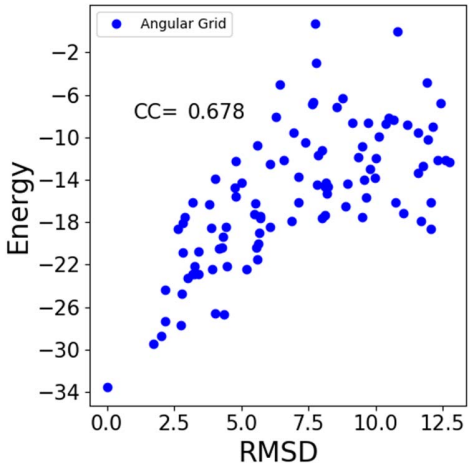
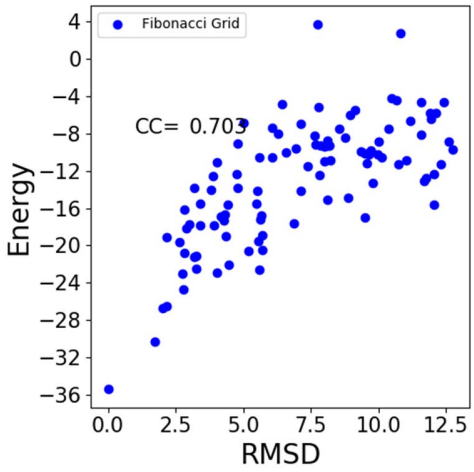




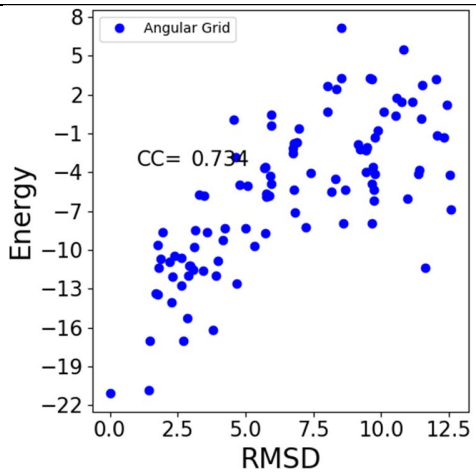
RO1hz6

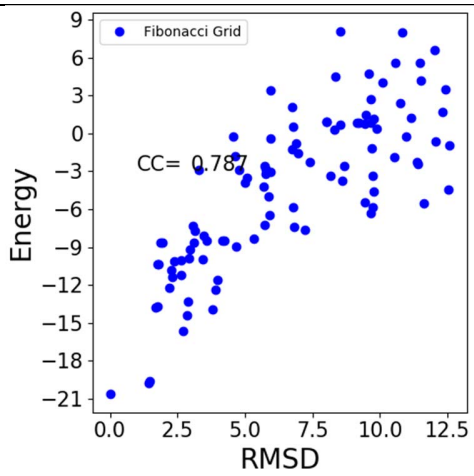




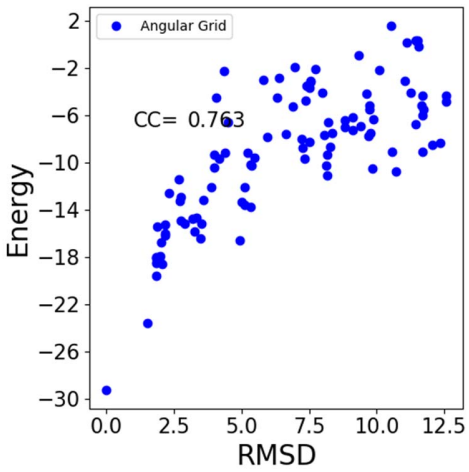
RO1lou		
RO1nps		
RO1rnb		

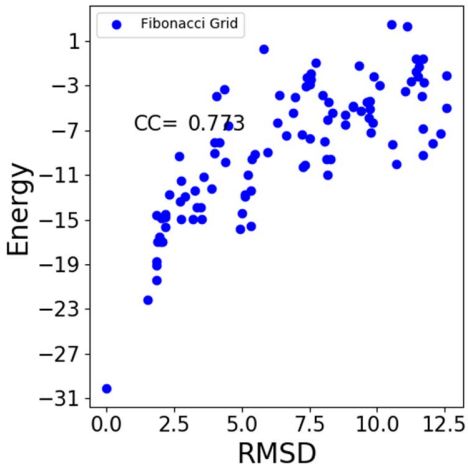
RO1ten



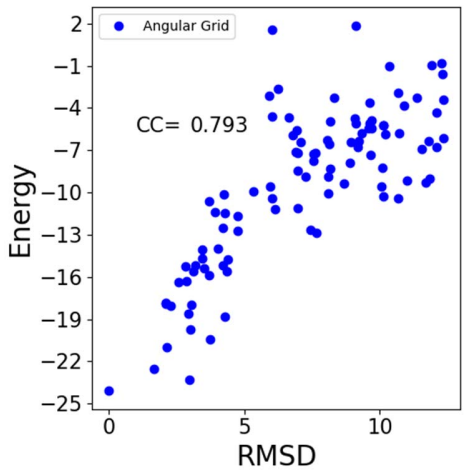


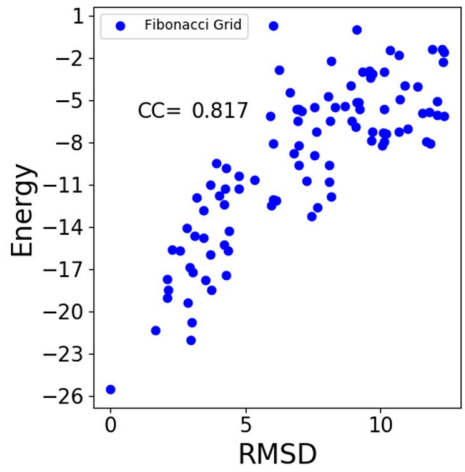
RO1ubi





RO1vcc





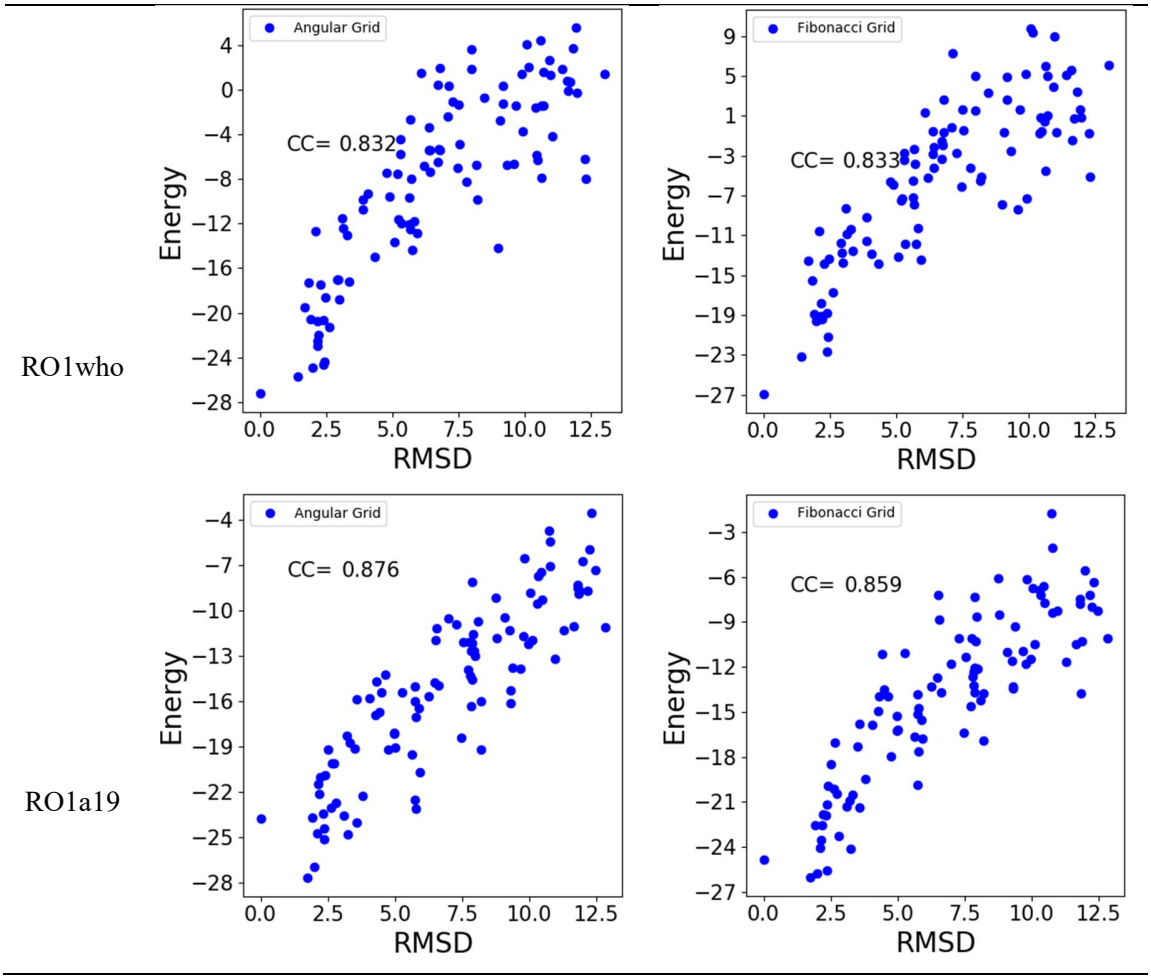
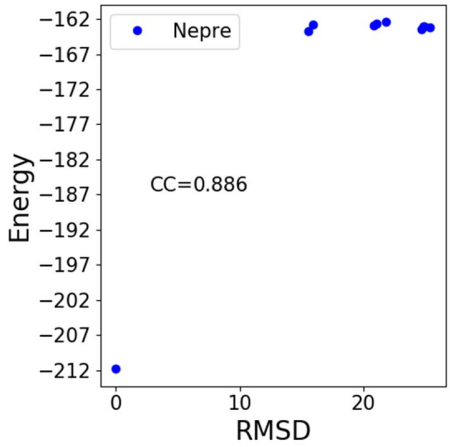
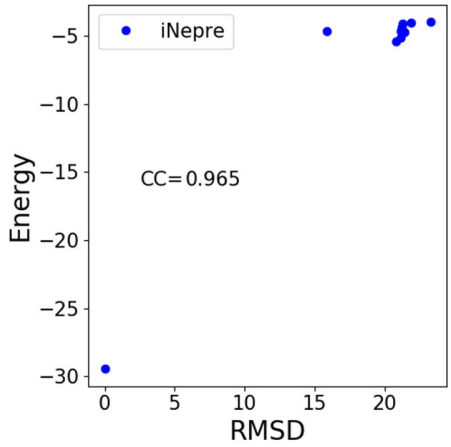
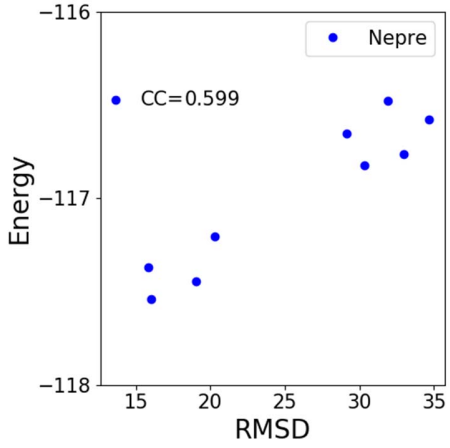
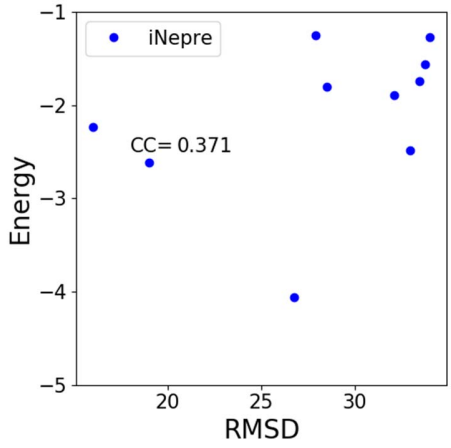
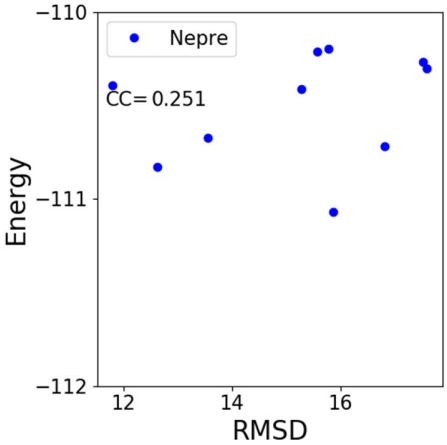
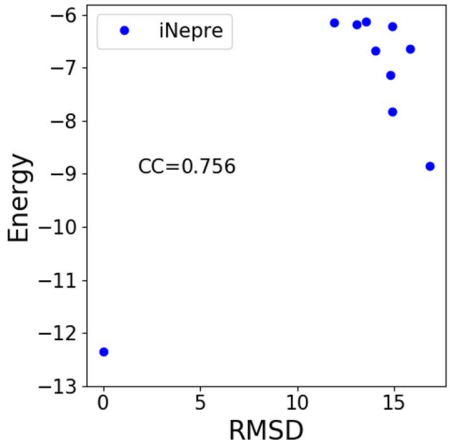
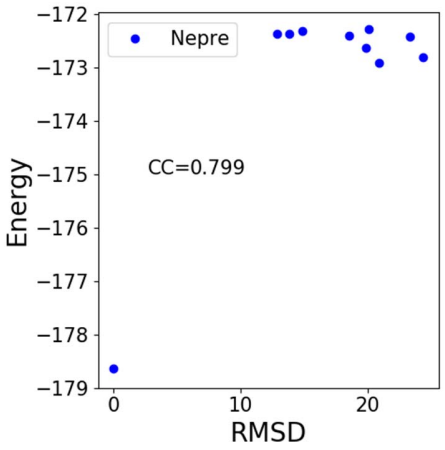
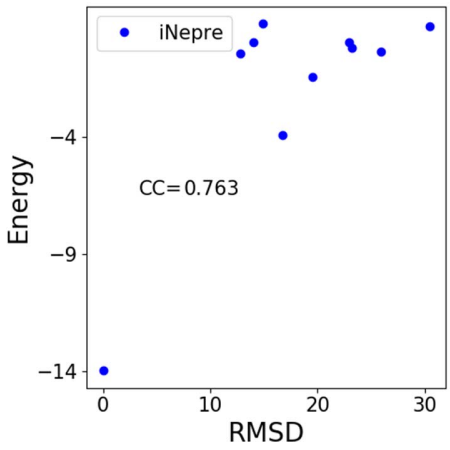


Figure S1 Pearson Coefficients calculated on some Rosetta decoy data sets using Nepre scoring function with Angular grid (left panel) and Fibonacci grid (right panel).

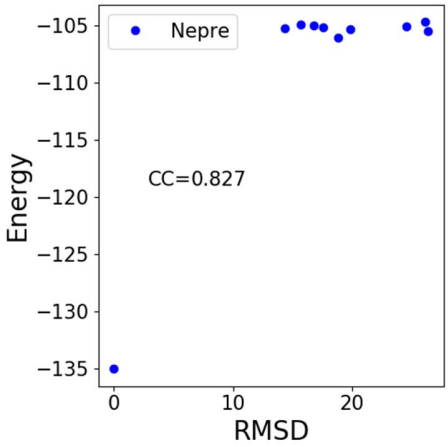
PDB ID	Nepre	iNepre
1KKL		
1MQ8		
2CFH		

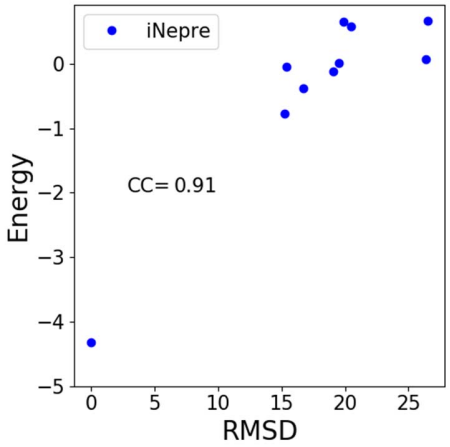
3AAA



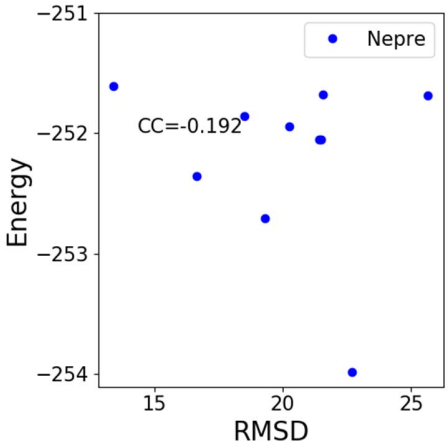


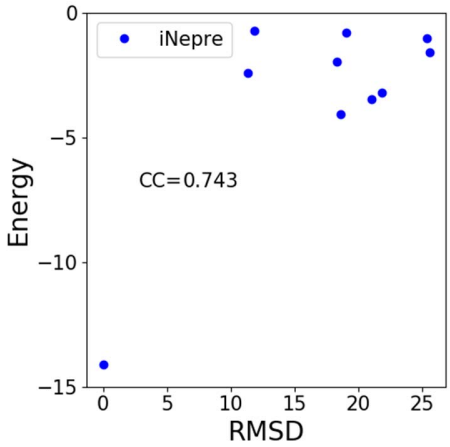
3DAW





3EO1





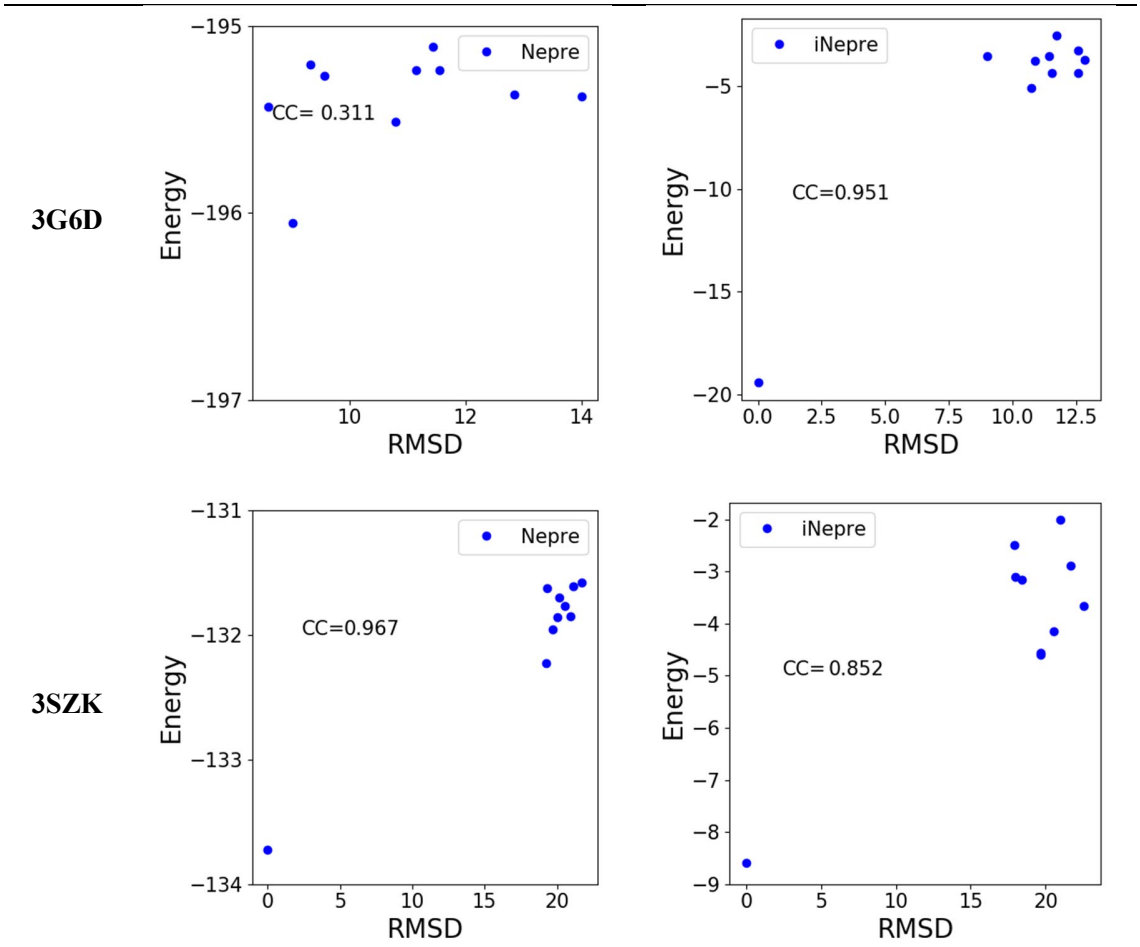


Figure S2 Pearson Coefficients calculated on TOP 10 decoys of some of ZDOCK Set-I decoys' data set using Nepre (left side panel) and iNepre (right side panel) scoring functions by applying Fibonacci grid.

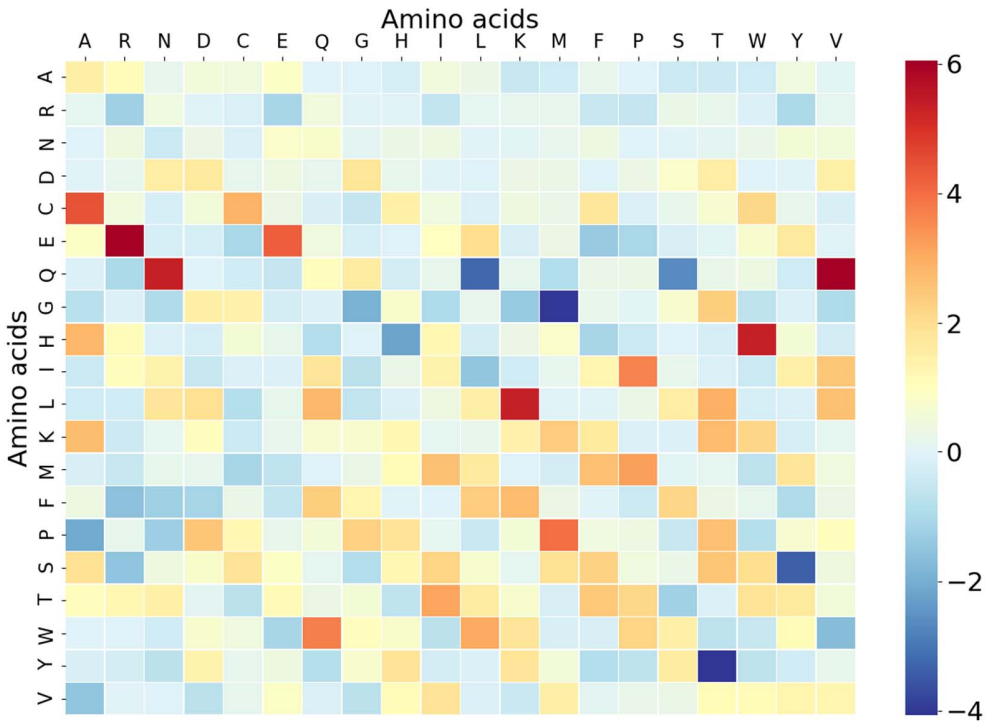


Figure S3 The comparison of neighborhood preferences (Nepre) derived from single chain protein structures and that from protein complex interfaces. Each row summaries the neighborhoods centered at a particular amino acid type; and each entry in a row corresponds to a neighboring amino acid distributed around the centered amino acid. The columns for alanine, cysteine, glycine and valine reveal large differences in the neighborhood preferences; the divergences are small in other cases, indicating similar preferences.

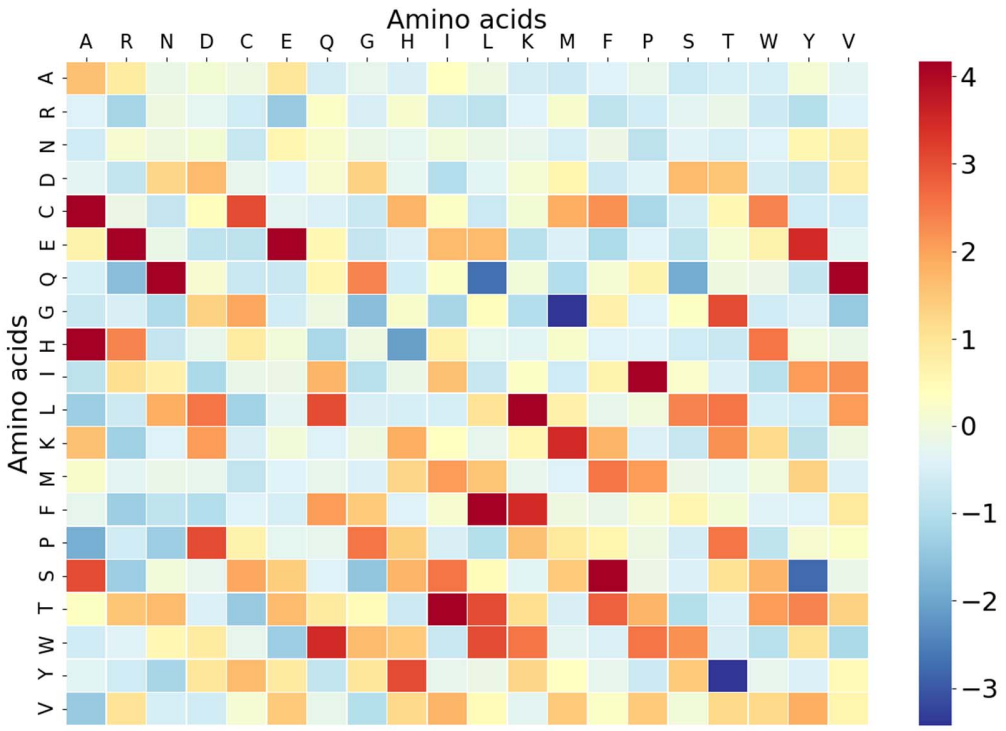


Figure S4 The comparison of neighborhood preferences (iNepre) derived from 2-4 chain protein structures interface residues and that from protein complex interfaces. Each row summaries the neighborhoods centered at a particular amino acid type; and each entry in a row corresponds to a neighboring amino acid distributed around the centered amino acid. The columns for leucine, methionine, threonine, and tyrosine reveal large differences in the neighborhood preferences; the divergences are small in other cases, indicating similar preferences.

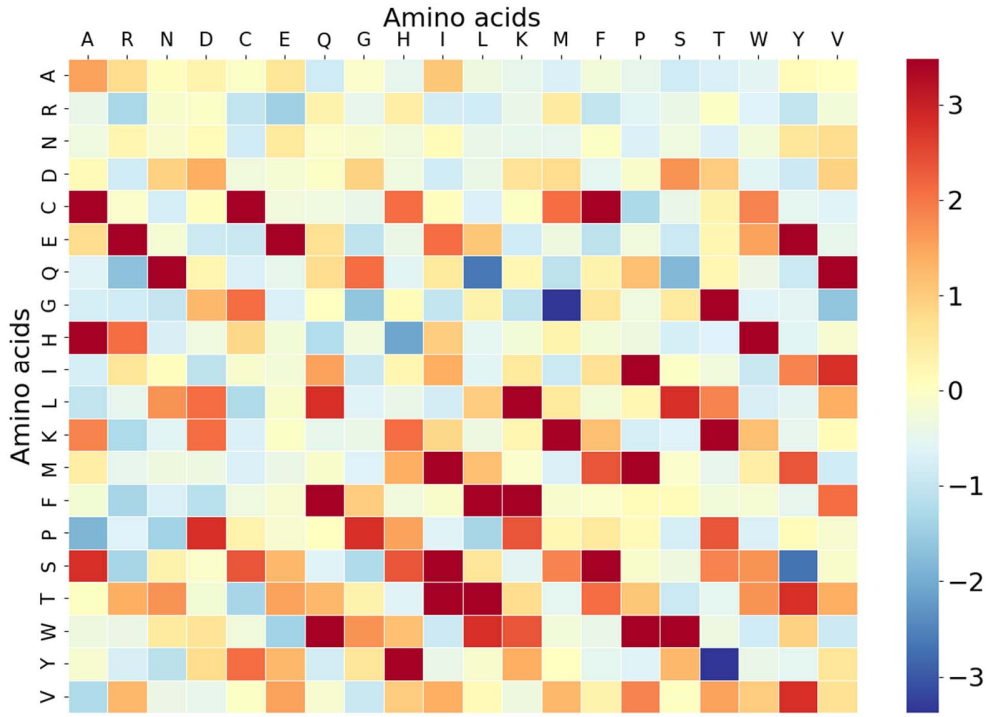


Figure S5 50 % of amino acid – amino acid pair interfacial residues data used in the comparison of neighborhood preferences (iNepre) derived from 2-4 chain protein structures interface residues and that from protein complex interfaces. Each row summaries the neighborhoods centered at a particular amino acid type; and each entry in a row corresponds to a neighboring amino acid distributed around the centered amino acid. The columns for leucine, methionine, threonine, and tyrosine reveal large differences in the neighborhood preferences; the divergences are small in other cases, indicating similar preferences.

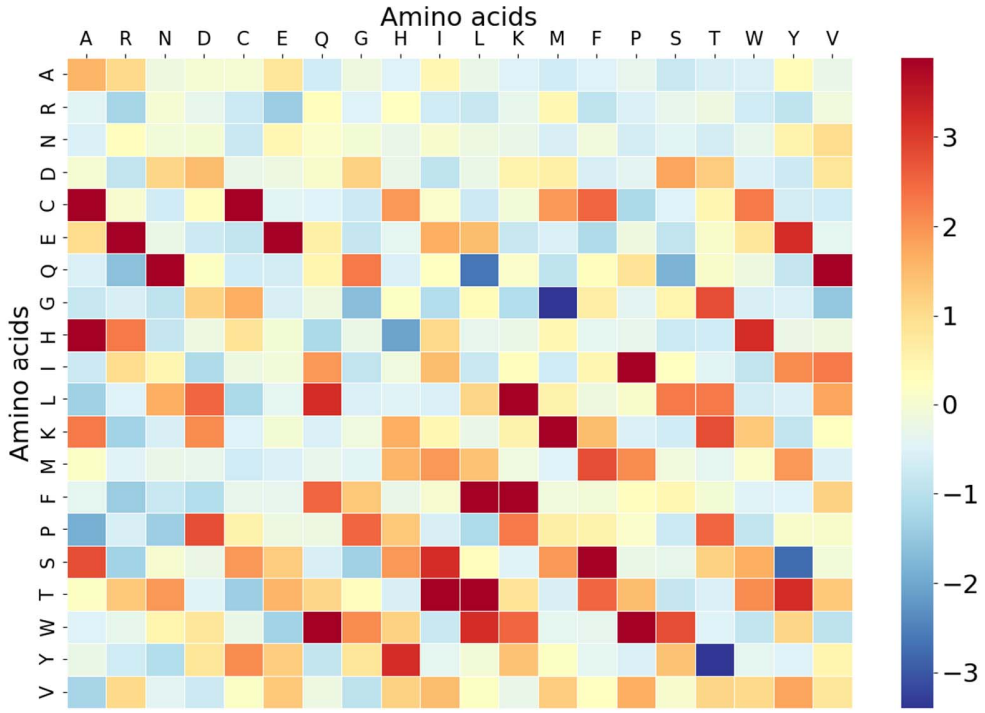


Figure S6 75 % of amino acid – amino acid pair interfacial residues data used in the comparison of neighborhood preferences (iNepre) derived from 2-4 chain protein structures interface residues and that from protein complex interfaces. Each row summaries the neighborhoods centered at a particular amino acid type; and each entry in a row corresponds to a neighboring amino acid distributed around the centered amino acid. The columns for leucine, methionine, threonine, and tyrosine reveal large differences in the neighborhood preferences; the divergences are small in other cases, indicating similar preferences.

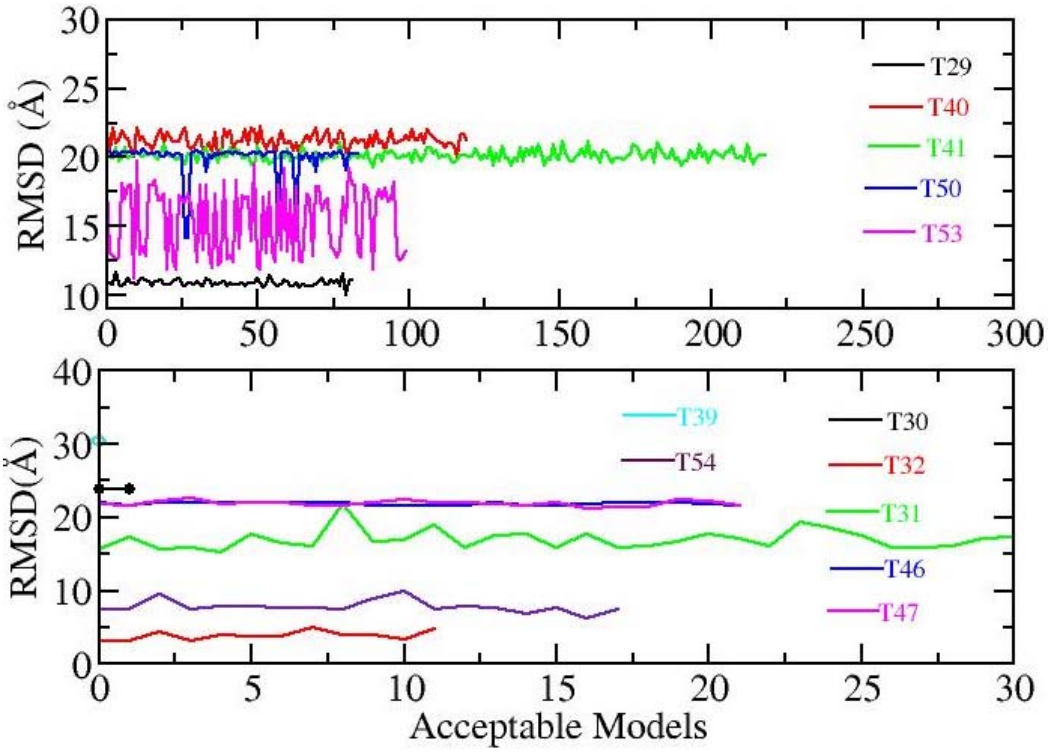


Figure S7 RMSD of CAPRI acceptable models (T29, T30, T32, T37, T39, T40, T41, T46, T47, T50, T53, T54) with respect to their corresponding native structures, divided into two groups according to acceptable number of models for better visualization.