Supplemental Material

Transferability of Geometric Patterns from Protein Self-Interactions to Protein-Ligand Interactions

Antoine Koehl[1,*], Milind Jagota[2,*], Dan D. Erdmann-Pham[3,*], Alexander Fung[2], and Yun S. Song[1,2,4,†]

[1]*Department of Statistics,* [2]*Department of Electrical Engineering and Computer Sciences,*
[3]*Department of Mathematics*
University of California, Berkeley, CA 94720, USA
[4]*Chan Zuckerberg Biohub, San Francisco, CA 94158*
[†]*E-mail: yss@berkeley.edu*

This file contains:

- Supplementary Text
- Supplementary Table S1
- Supplementary Figures S1-S7

[*]These authors contributed equally to this work.

1. Definitions of Interacting Chemical Groups

We examine patterns of interaction geometry for 20 different chemical groups that are found in the 20 amino acids. These are the same chemical groups studied by Polizzi and Degrado [1]. We list them here along with the SMARTS string definition for each in Table S1.

2. Frequencies of different interaction types

In addition to comparing geometric patterns within each interaction type, one can also ask how common the different chemical groups and amino acids are in protein self-interactions as compared to protein-ligand interactions. We plot both distributions in Figure S1.

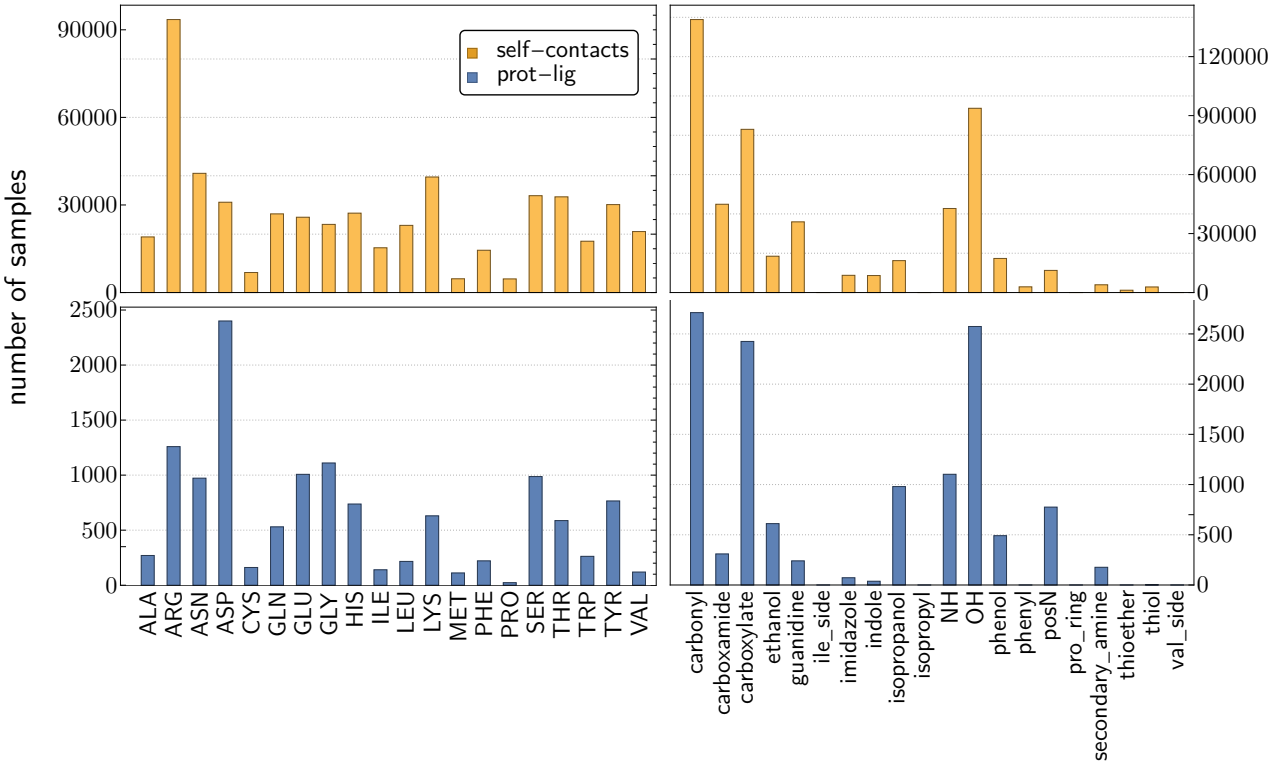

Fig. S1. Frequencies of each iCG and amino acid in both datasets

Table S1. Chemical Functional Group List and Definitions

Functional Group	**SMARTS Pattern**	**Amino Acid** Atoms
Carbonyl	[C,c]=O	**GLY** C,O
Carboxylate	[C,c][CX3](=[OX1])[OH0-,OH]	**ASP** CB, CG, OD1, OD2
Isopropyl	CC(C)[C;!R	**LEU** CB, CG, CD1, CD2
Carboxamide	[C,c][CX3](=[OX1])[NX3H2]	**ASN** CB, CG, OD1, ND2
NH	[C,c][N;H2,H1]	**GLY** CA, N
OH	[C,c][O;H1]	**SER** CB, OG
Thiol	[CX4;H2][S;H1,H0]	**CYS** CB, SG
Thioether	CSC	**MET** CG, SD, CE
Guanidine	N=C(N)N	**ARG** NH2, CZ, NE, NH1
Phenol	[O;H1]c1ccccc1	**TYR** OH, CZ, CE2, CD2, CG, CD1, CE1
N+	[C,c][N+;H3]	**LYS** CE, NZ
Indole	c1c[nH]c2ccccc12	**TRP** CG, CD1, NE1, CE2, CZ2, CH2, CZ3, CE3, CD2
Imidazole	c1c[n;H0,H1]cn1	**HIS** CG, CD2, NE2, CE1, ND1
Phenyl	c1ccccc1	**PHE** CG, CD1, CE1, CZ, CE2, CD2
Isopropanol	[C,c][C;!R]([C;!R])[O;H1]	**THR** CG2, CB, CA, OG1
Ethanol	[C,c][C;H2;!R][O;H1]	**SER** CA, CB, OG
Secondary Amine	c[n;H0,H1]c	**HIS** CE1, NE2, CD2
Pro Ring	[$([NX3H,NX4H2+]),$([NX3](C)(C)(C))]1[CX4H]([CH2][CH2][CH2]1)	**PRO** N, CD, CG, CB, CA
Val Side	[CHX4]([CH3X4])[CH3X4]	**VAL** CG1, CB, CG2
Ile Side	[CHX4]([CH3X4])[CH2X4][CH3X4]	**ILE** CG2, CB, CG1, CD1

3. Sample sizes

Sample sizes for the protein-ligand dataset are very small, as shown in Figure S2. We also show sample sizes for the protein self-interaction dataset for comparison.

Reference Amino Acid

iCG	Alanine	Arginine	Asparagine	Aspartic acid	Cysteine	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophan	Tyrosine	Valine
Carbonyl	5794 100	29845 225	10893 306	3123 249	1347 99	6895 207	2482 29	6523 359	6683 115	5836 79	7528 84	11261 153	1239 44	4118 48		7863 160	8268 142	4078 36	6823 229	8265 48
NH	2496 47	1220 28	2514 71	4146 113	617 16	1779 28	3475 82	2363 350	1042 50	2381 8	3076 39	1055 14	648 18	2189 31	1104 16	2643 82	2696 54	1154 6	2857 28	3301 21
OH	2565 17	23204 244	7413 237	3474 639	844 3	4950 120	2628 350	3662 103	5948 243	1700 5	2856 11	8816 125	651 28	1748 44	449 4	5924 89	5827 68	3566 135	5083 93	2409 16
N+	437 20	168 0	591 36	2647 218	81 0	377 21	2414 177	489 83	199 0	282 0	514 14	182 0	98 0	310 13	200 0	666 59	626 69	171 0	470 53	337 12
Thiol	70 0	117 0	142 0	68 0	477 1	78 0	42 0	81 0	169 0	55 0	53 0	49 0	55 0	455 0	12 0	150 0	141 0	213 0	341 2	62 0
Ethanol	703 4	2087 30	1725 51	1339 192	270 0	1181 23	933 99	835 12	975 57	462 5	722 2	969 14	167 0	480 29	187 0	1269 30	1394 12	901 27	1159 23	708 1
2˚ amine	101 0	245 0	242 0	262 0	67 0	155 0	193 0	172 0	199 0	73 0	172 0	114 0	59 0	290 0	61 0	378 115	281 0	268 0	497 61	112 0
Thioether	23 0	94 0	179 0	25 0	92 0	127 0	18 0	45 0	70 0	24 0	33 0	38 0	5 0	20 0		79 0	103 0	72 0	87 0	35 0
Carboxamide	1814 4	4666 8	4747 11	3107 89	460 2	2962 7	2371 22	1987 62	1705 7	1471 4	2166 20	2479 11	547 13	1426 0	684 1	3002 25	3161 4	1597 3	2794 12	1750 4
Carboxylate	1904 53	26185 646	5931 161	1018 51	412 30	3441 90	795 46	3196 80	6389 152	1099 30	1908 22	11939 246	371 6	869 20	3 0	5263 309	4310 208	2490 33	4035 233	1453 9
Guanidine	1476 11	751 1	1883 3	7622 110	277 0	1554 8	7103 27	1782 26	734 2	818 4	1861 7	575 1	311 2	929 3	1135 1	1995 23	1816 4	640 5	1679 1	1011 1
Isopropanol	611 1	1917 36	1532 61	1033 598	181 0	1196 16	700 87	763 16	817 58	375 2	545 2	821 32	171 0	365 16	142 0	1184 24	1607 11	735 9	981 11	560 0
Imidazole	242 0	612 2	522 2	1047 17	122 0	339 1	684 10	394 9	504 2	177 0	319 3	292 0	97 0	390 1	129 0	723 17	640 4	417 1	860 3	244 0
Phenol	494 13	1798 37	1595 31	1265 115	512 8	1156 9	1162 76	621 8	1008 51	371 3	772 8	718 34	180 0	587 16	310 1	1166 42	1071 11	770 7	1365 14	419 7
Indole	296 0	434 2	674 2	738 9	379 2	587 0	782 2	375 2	464 0	197 0	457 4	187 0	104 0	285 0	248 0	595 12	598 0	374 0	674 2	206 0

Fig. S2. Protein-Ligand sample sizes including only contacts with hydrogen bonds. For each contact type, note the number of observations in the protein-protein(top number-yellow) and protein-ligand(bottom number- blue) datasets

4. Lever-arm effect

In their work, Polizzi and DeGrado observed the "lever-arm effect", whereby aligning observed contacts by amino acid backbone amplifies noise in the iCG [1]. We illustrate this effect with a simulation in Figure S3. In Figure S3a, we start out with a true contact geometry between a red triangle and a blue square. In Figure S3b, we sample 50 instances of the true geometry plus noise. The noise has both a component that is applied to all verticies independently as well as a random rotation around the centroid that is applied independently to the triangle and square. In Figure S3c, we then align these instances on the triangle. Although the true geometry is still recognizable after adding the noise, aligning on the triangle amplifies the noise on the square so that the true geometry is hard to see.

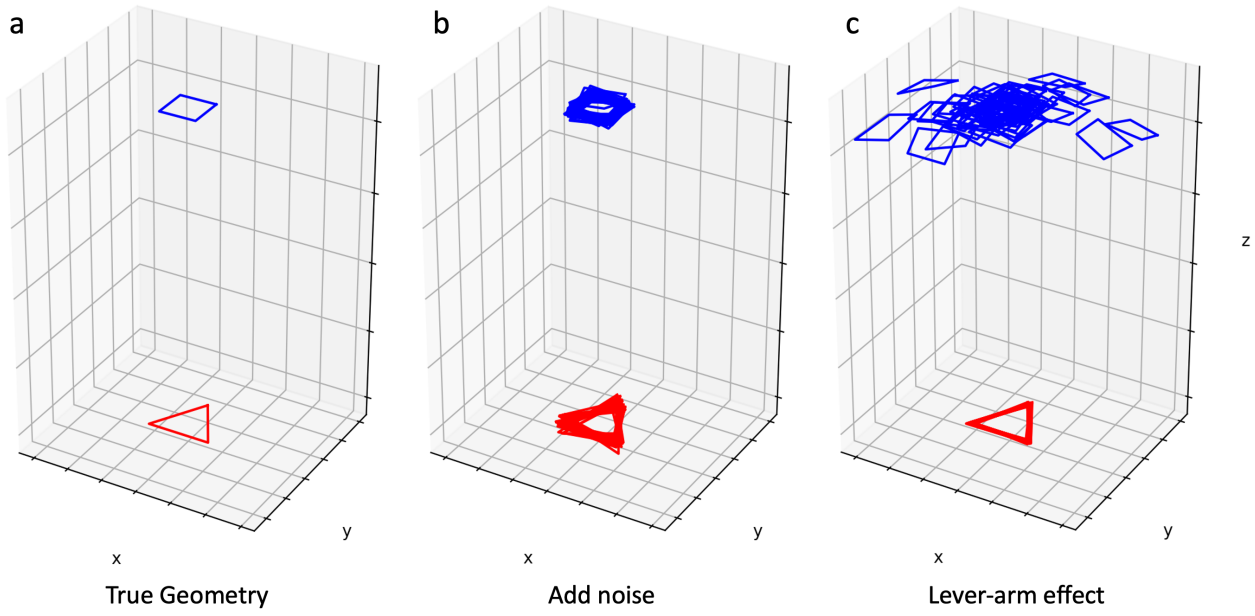

Fig. S3. Simulation of the lever-arm effect. a) The true geometry of a simulated contact between a red triangle and a blue square is shown. b) 50 versions of the contact plus noise are sampled. The noise has a component that is added independently to each vertex as well as random independent rotations of both the triangle and square around their centroids. c) After aligning the noisy contacts on the triangle, the noise of the square becomes amplified so that the true geometry is hard to see.

5. Plots of atomic angles

We show the distribution of cosine angles for every iCG atom in every contact type that can have hydrogen bonds in Figure S4. Hydrogen bonding atoms almost always have a strong negative skew indicating that they point inwards towards the amino acid backbone.

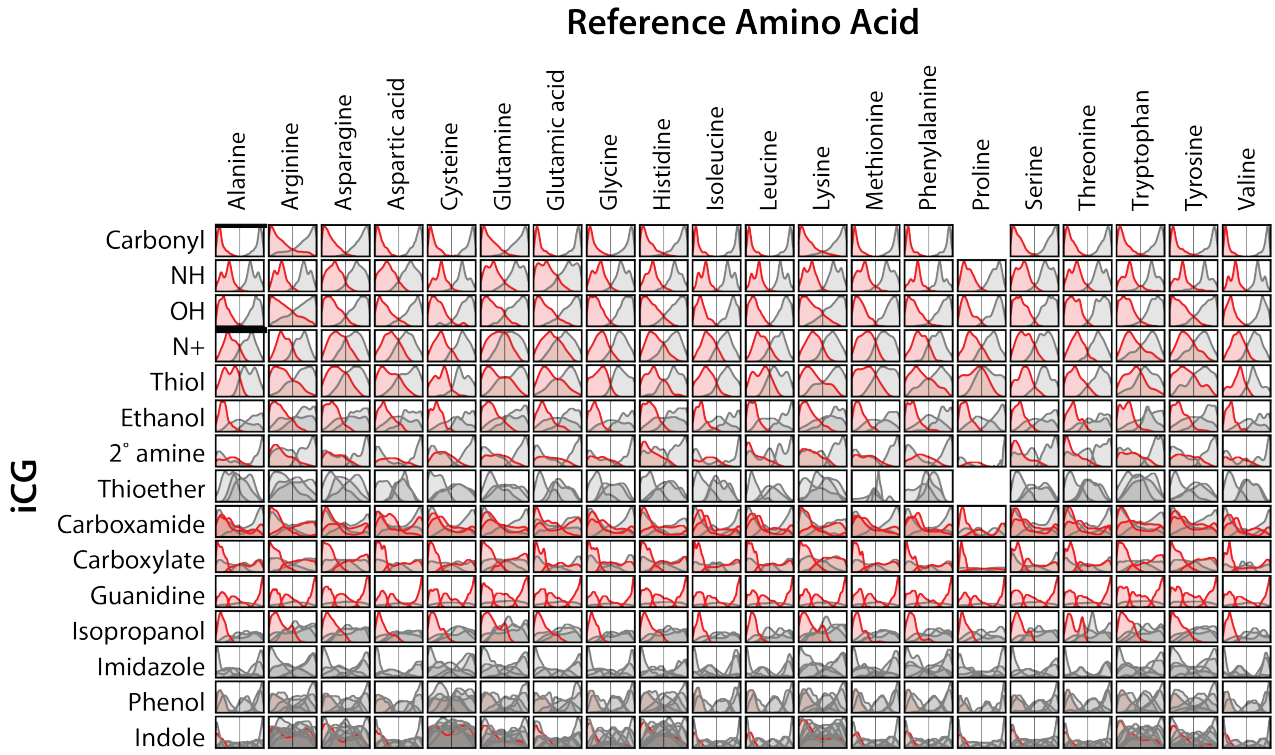

Fig. S4. Orientation of iCG for each contact type relative to the centroid vector. H-bonding atoms of the iCG are shown in red. There is a clear preference for orientations that place H-bonding atoms towards the reference amino acid centroid

6. Radius testing p-values

For the transferability test that is described in the main text, we had sufficient samples for 87 contact types. The distribution of p-values along with the correlation to sample size in the protein-ligand dataset are shown in Figure S5.

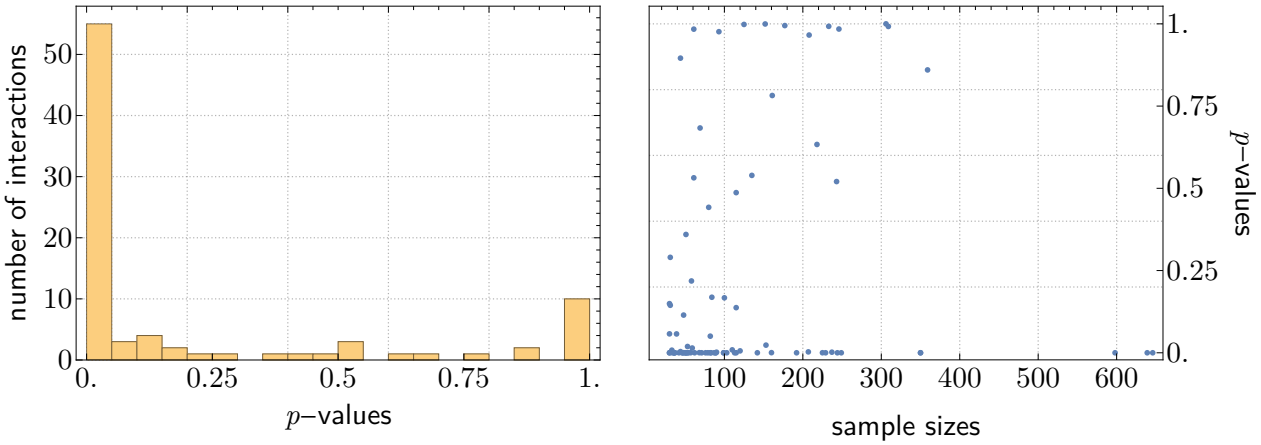

Fig. S5. Distribution of p-values for hypothesis test of transferability on the distance between iCG centroid and amino acid backbone centroid. 87 contact types had sufficient samples for the test to be valid and of these, 57 had p-values less than 0.05. The p-values do not have a clear correlation with sample size in the protein-ligand data indicating that failures to reject are not just due to lack of samples.

7. Radius testing distributions

We provide amino acid- iCG centroid distance distributions from protein-protein and protein-ligand contacts. In the majority of cases, both distributions show similar interaction modes - however, imbalances in the density of contacts in each mode will lead to rejection by our test. Figure S6 provides a qualitative view of these distributions to aid in understanding similarities and differences between protein-ligand and protein self-contact contacts.

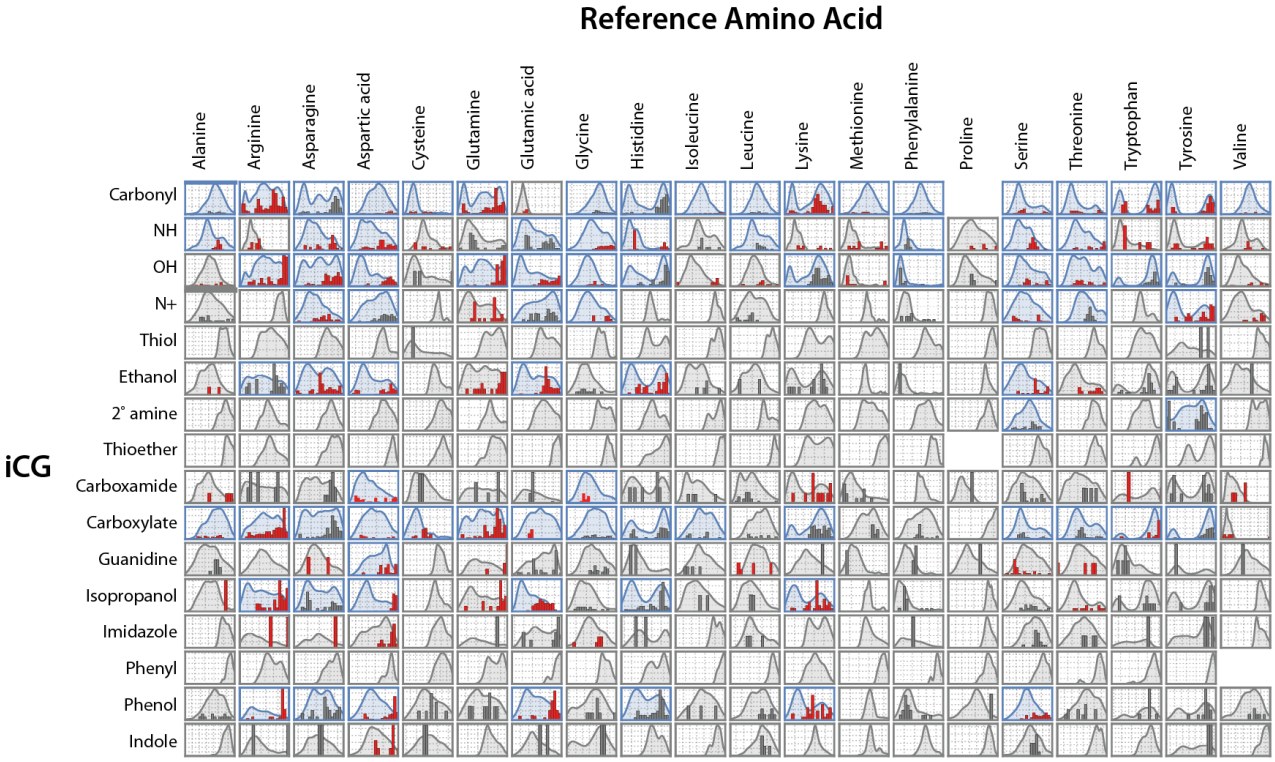

Fig. S6. Distribution of iCG-Backbone centroid distances of all contacts from protein self-interactions (smooth curves) and protein-ligand interactions (histrograms). Contact types that meet the asymptotic assumptions of our test are colored blue, all others are colored grey. Contact types with matching protein-ligand and protein self contact distributions have gray histogram bars, while those rejected by our test are colored red.

8. Orientation testing *p*-values

For the orientational transferability test that is described in the main text, we had sufficient samples for 87 contact types. The distribution of p-values along with the correlation to sample size in the protein-ligand dataset are shown in Figure S7

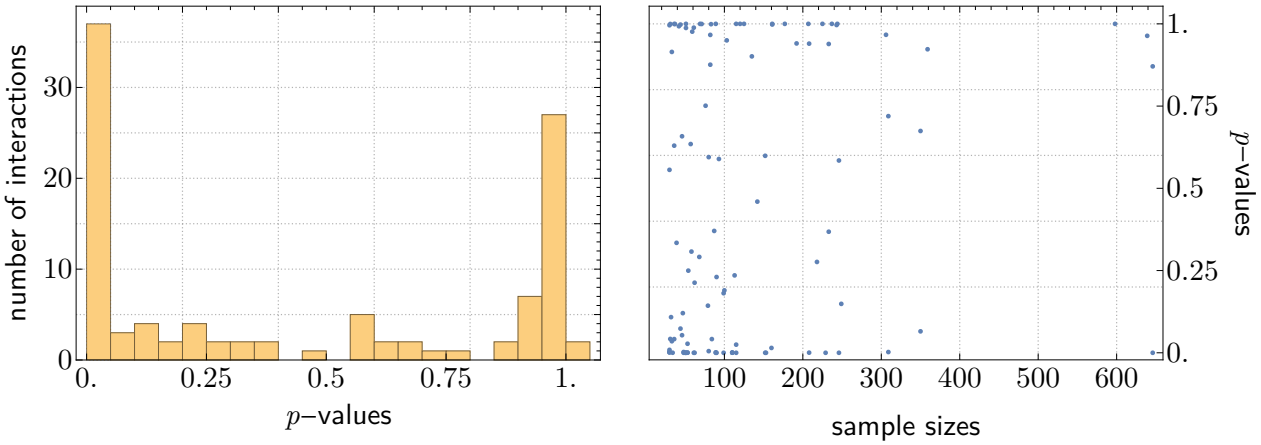

Fig. S7. Distribution of p-values for hypothesis test of transferability on the distance between iCG centroid and amino acid backbone centroid. 87 contact types had sufficient samples for the test to be valid and of these, 37 had p-values less than 0.05. The p-values do not have a clear correlation with sample size in the protein-ligand data indicating that failures to reject are not just due to lack of samples.

References

[1] N. F. Polizzi and W. F. DeGrado, A defined structural unit enables de novo design of small-molecule-binding proteins, *Science* **369**, 1227 (2020).