

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 Fung²,
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-S6

*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.

Table S1. Chemical Functional Group List and Definitions

Functional Group	SMARTS Pattern
Carbonyl	<chem>[C,c]=O</chem>
Carboxylate	<chem>[C,c][CX3](=[OX1])[OH0-,OH]</chem>
Isopropyl	<chem>CC(C)[C;!R]</chem>
Carboxamide	<chem>[C,c][CX3](=[OX1])[NX3H2]</chem>
NH	<chem>[C,c][N;H2,H1]</chem>
OH	<chem>[C,c][O;H1]</chem>
Thiol	<chem>[CX4;H2][S;H1,H0]</chem>
Thioether	<chem>CSC</chem>
Guanidine	<chem>N=C(N)N</chem>
Phenol	<chem>[O;H1]c1ccccc1</chem>
N+	<chem>[C,c][N+;H3]</chem>
Indole	<chem>c1c[nH]c2ccccc12</chem>
Imidazole	<chem>c1c[n;H0,H1]cn1</chem>
Phenyl	<chem>c1ccccc1</chem>
Isopropanol	<chem>[C,c][C;!R]([C;!R])[O;H1]</chem>
Ethanol	<chem>[C,c][C;H2;!R][O;H1]</chem>
Aromatic Secondary Amine	<chem>c[n;H0,H1]c</chem>
Pro Ring	<chem>[\$([NX3H,NX4H2+]),\$([NX3](C)(C)(C))1[CX4H]([CH2][CH2][CH2]1)</chem>
Val Side	<chem>[CHX4]([CH3X4])[CH3X4]</chem>
Ile Side	<chem>[CHX4]([CH3X4])[CH2X4][CH3X4]</chem>

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.

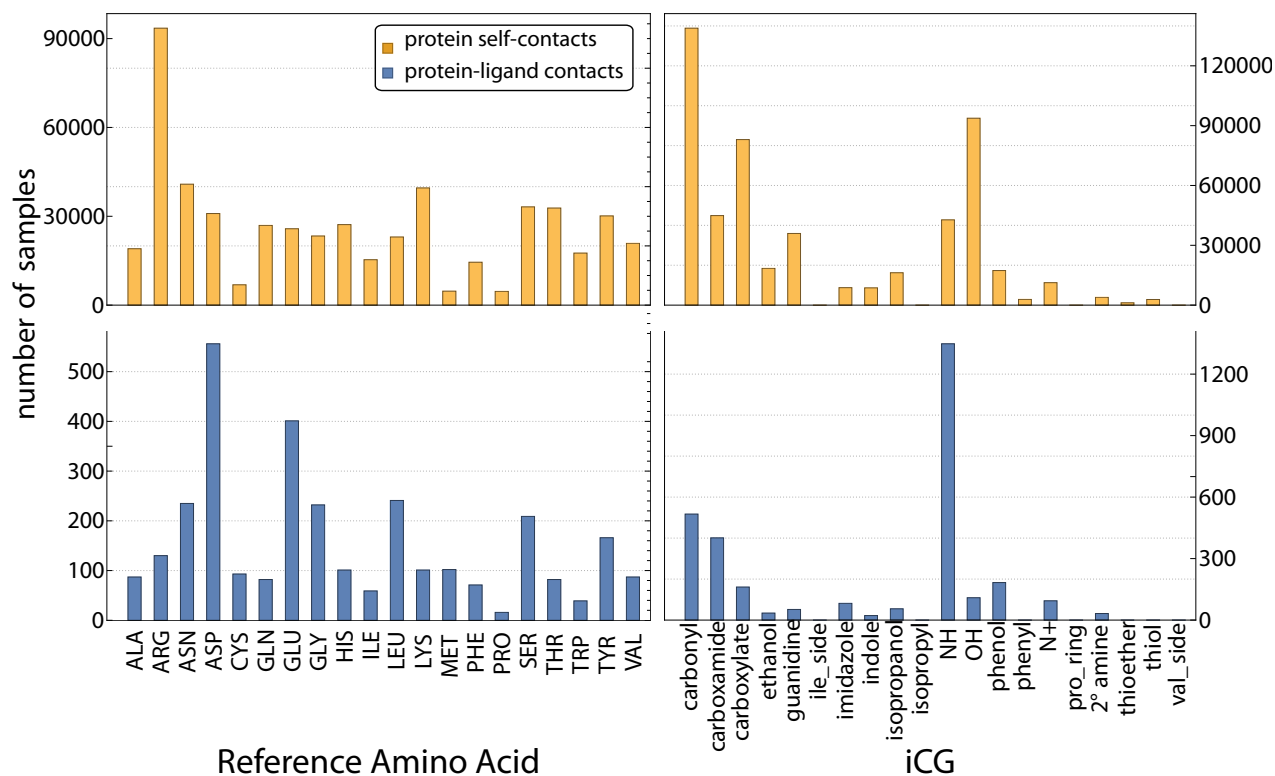


Fig. S1. Frequencies of each iCG and amino acid in both datasets

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																					
		Alanine	Arginine	Asparagine	Aspartic acid	Cysteine	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophan	Tyrosine	Valine		
iCG	Carbonyl	5794 9	29845 39	10893 92	3123 22	1347 46	6895 44	2482 5	6523 86	6683 14	5836 10	7528 20	11261 43	1239 18	4118 14		7863 51	8268 25	4078 6	6823 39	8265 12		
	NH	2496 66	1220 15	2514 81	4146 381	617 44	1779 13	3475 239	2363 119	1042 24	2381 37	3076 169	1055 14	648 78	2189 29	1104 15	2643 64	2696 37	1154 13	2857 36	3301 65		
	OH	2565 2	23204 6	7413 15	3474 16	844 0	4950 7	2628 12	3662 0	5948 17	1700 2	2856 0	8816 3	651 0	1748 5	449 0	5924 6	5827 3	3566 4	5083 10	2409 1		
	N+	437 2	168 0	591 11	2647 22	81 0	377 1	2414 31	489 12	199 0	282 0	514 0	182 0	98 0	310 0	200 0	666 1	626 1	171 0	470 12	337 1		
	Thiol	70 0	117 0	142 0	68 0	477 0	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 0	62 0		
	Ethanol	703 0	2087 0	1725 6	1339 9	270 0	1181 1	933 3	835 0	975 4	462 1	722 0	969 0	167 0	480 4	187 0	1269 2	1394 2	901 2	1159 0	708 0		
	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 24	281 0	268 0	497 26	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 1	4666 1	4747 2	3107 10	460 0	2962 2	2371 16	1987 4	1705 0	1471 3	2166 15	2479 6	547 0	1426 0	684 0	3002 5	3161 2	1597 0	2794 5	1750 2		
	Carboxylate	1904 9	26185 59	5931 11	1018 11	412 5	3441 9	795 10	3196 8	6389 26	1099 5	1908 2	11939 34	371 0	869 2	3 0	5263 35	4310 14	2490 10	4035 39	1453 4		
	Guanidine	1476 0	751 0	1883 1	7622 38	277 0	1554 0	7103 5	1782 2	734 0	818 1	1861 0	575 0	311 0	929 0	1135 0	1995 2	1816 0	640 1	1679 1	1011 1		
	Isopropanol	611 0	1917 0	1532 9	1033 3	181 0	1196 1	700 22	763 3	817 1	375 0	545 0	821 1	171 0	365 9	142 0	1184 5	1607 1	735 0	981 0	560 0		
	Imidazole	242 0	612 1	522 0	1047 1	122 0	339 0	684 1	394 0	504 0	177 0	319 2	292 0	97 0	390 0	129 0	723 2	640 0	417 0	860 1	244 0		
	Phenol	494 2	1798 18	1595 12	1265 35	512 2	1156 3	1162 40	621 1	1008 22	371 0	772 6	718 6	180 0	587 8	310 0	1166 15	1071 5	770 4	1365 2	419 2		
	Indole	296 0	434 1	674 0	738 2	379 2	587 0	782 1	375 0	464 0	197 0	457 0	187 0	104 0	285 0	248 0	595 7	598 0	374 0	674 0	206 0		

Fig. S2. Protein-Ligand sample sizes including only contacts with hydrogen bonds. For each vdM 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 vertices 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.

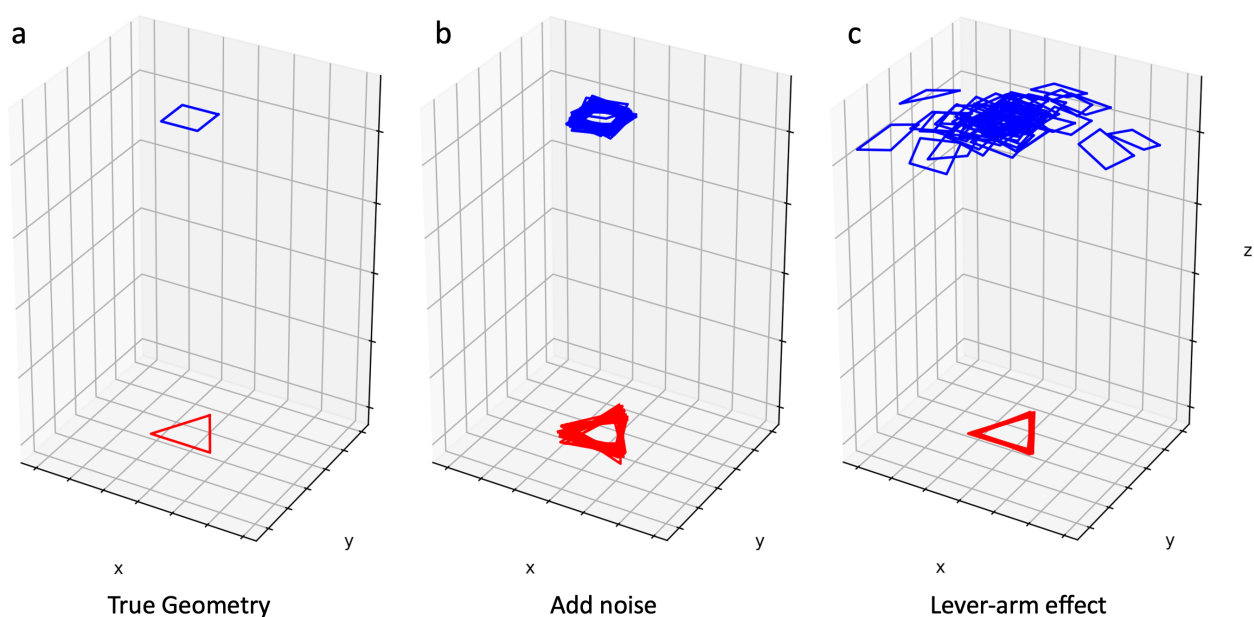


Fig. S3. Simulation of the lever-arm effect. a) The true geometry of a simulated contact between a red triangle and a blue square is shown. b) 50 versions of the contact plus noise are sampled. The noise has a component that is added independently to each vertex as well as random independent rotations 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. Atomic angle definition

The atomic angle statistic defined and reported in the main text is illustrated in Figure S4.

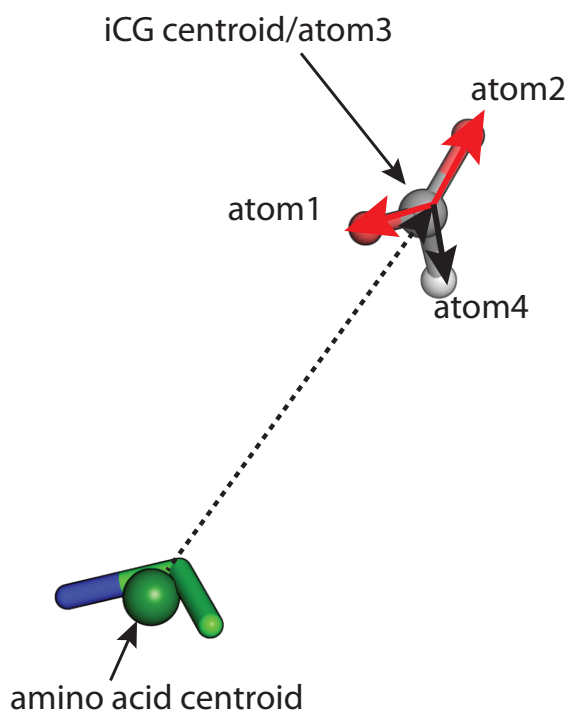


Fig. S4. The atomic angle statistic described in the main text is illustrated here. For each iCG atom, we compute the vector from the iCG centroid to the atom (shown as solid arrows). We then also compute the vector from the amino acid centroid to the iCG centroid (shown as dotted arrow). The atomic angle statistic for a particular atom is the cosine of the angle between the centroid-centroid vector and the centroid-atom vector

6. 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 S5. Hydrogen bonding atoms almost always have a strong negative skew indicating that they point inwards towards the amino acid backbone.

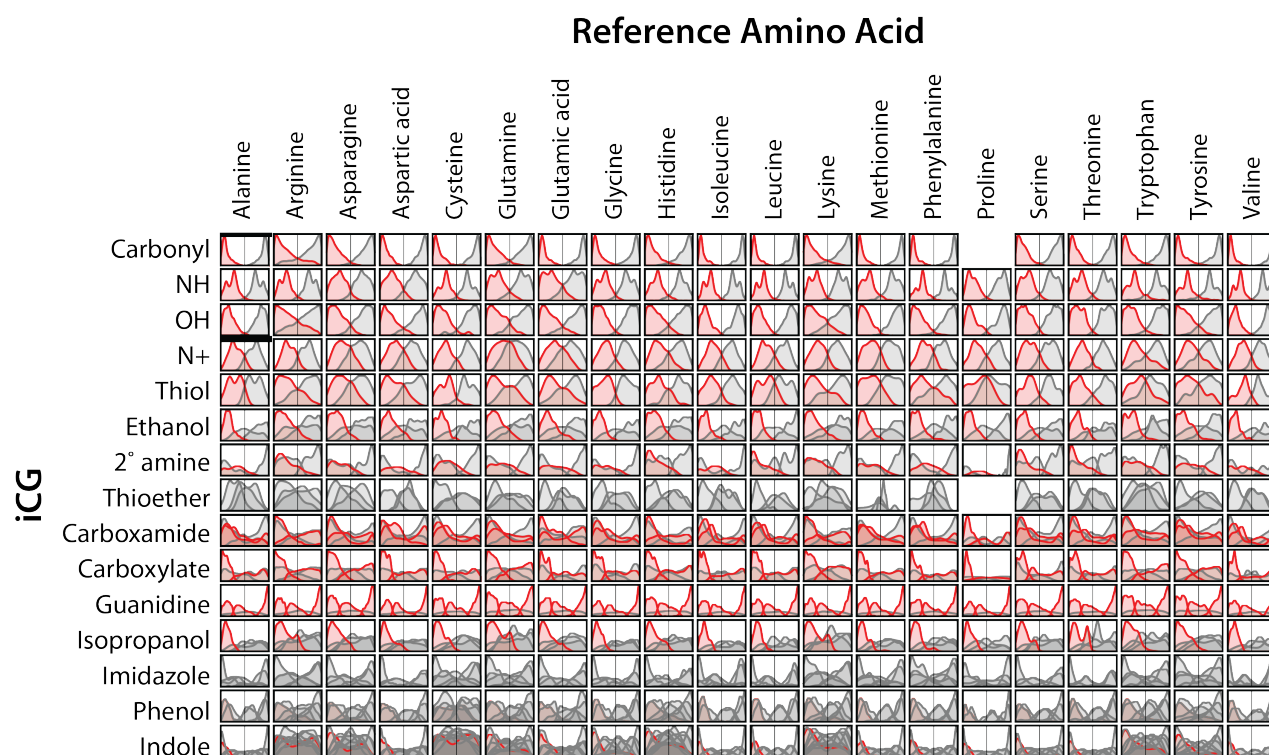


Fig. S5. Orientation of iCG for each vdM type relative to the centroid vector. H-bonding atoms of the iCG are shown in red. There is a clear preference for orientations that place H-bonding atoms towards the reference amino acid centroid

7. Radius testing p -values

For the transferability test that is described in the main text, we had sufficient samples for 32 contact types. The distribution of p -values along with the correlation to sample size in the protein-ligand dataset are shown in Figure S6.

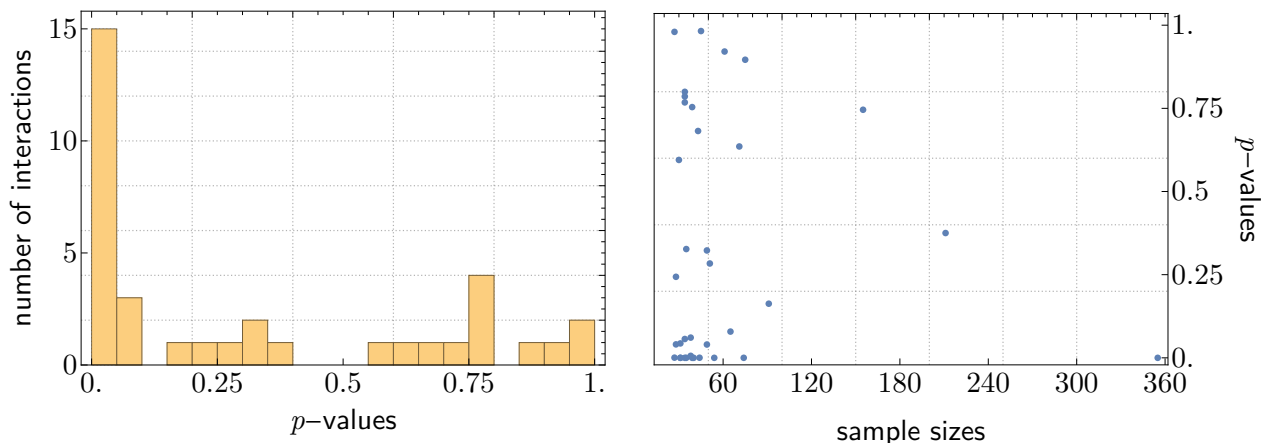


Fig. S6. Distribution of p -values for hypothesis test of transferability on the distance between iCG centroid and amino acid backbone centroid. 32 contact types had sufficient samples for the test to be valid and of these, 15 had p -values less than 1. The p -values do not have a clear correlation with 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).