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

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.

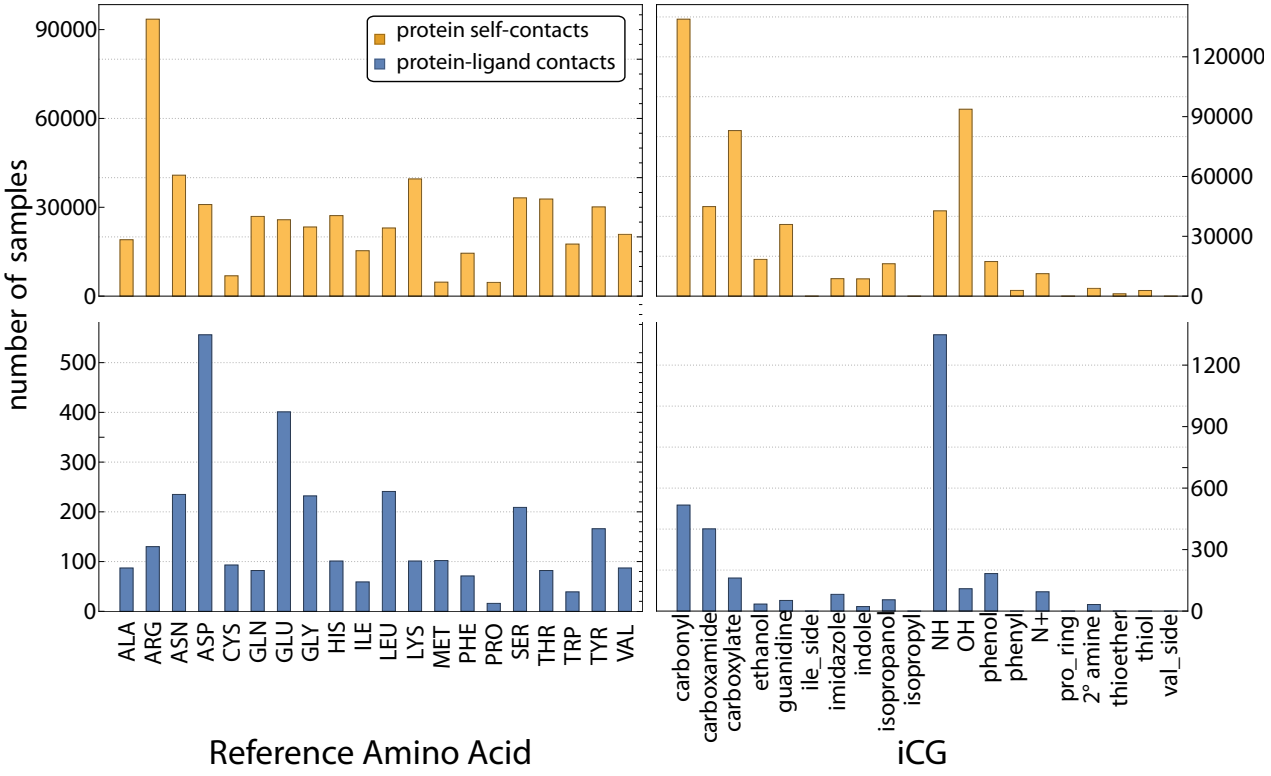

Fig. S1. Frequencies of each iCG and amino acid in both datasets

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.

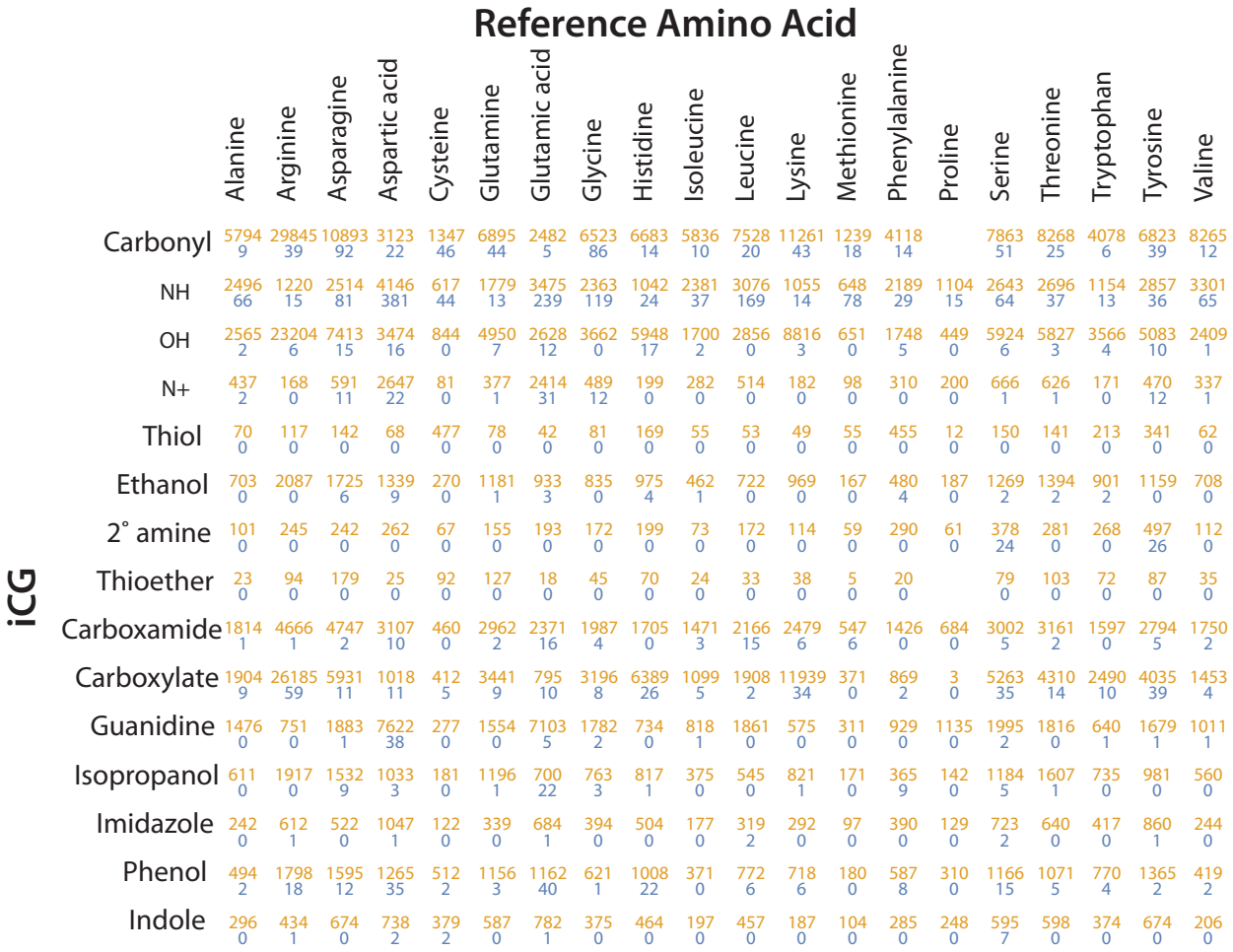

Fig. S2. Protein-Ligand sample sizes including only contacts with hydrogen bonds. For each vdM type, note the number of observations in the protein-protein(top number-yellow) and protein-ligand(bottom number- blue) datasets

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 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.

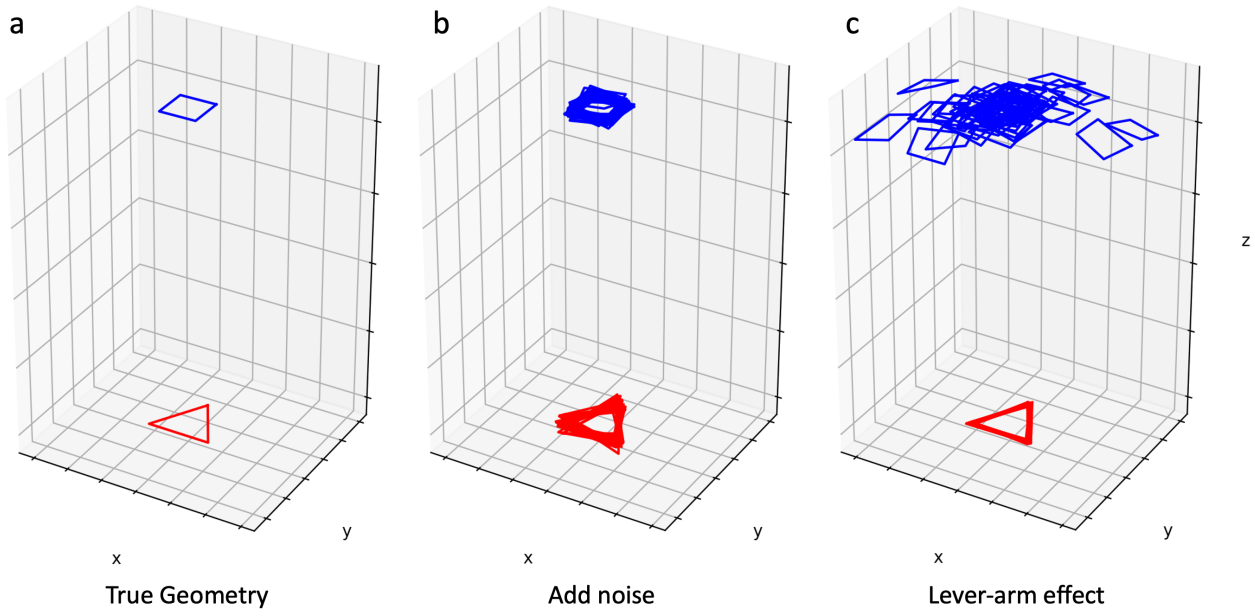

Fig. S3. Simulation of the lever-arm effect. a) The true geometry of a simulated contact between a red triangle and a blue square is shown. b) 50 versions of the contact plus noise are sampled. The noise has a component that is added independently to each vertex as well as random independent rotations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

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

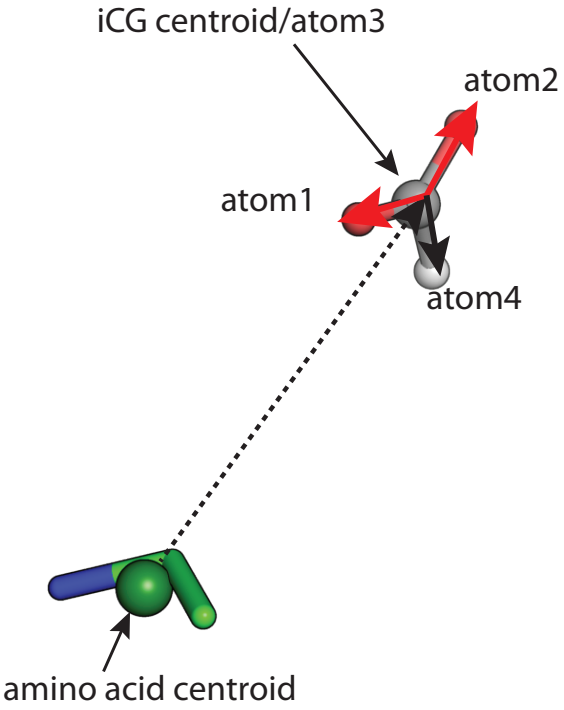

Fig. S4. The atomic angle statistic described in the main text is illustrated here. For each iCG atom, we compute the vector from the iCG centroid to the atom (shown as solid arrows). We then also compute the vector from the amino acid centroid to the iCG centroid (shown as dotted arrow). The atomic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

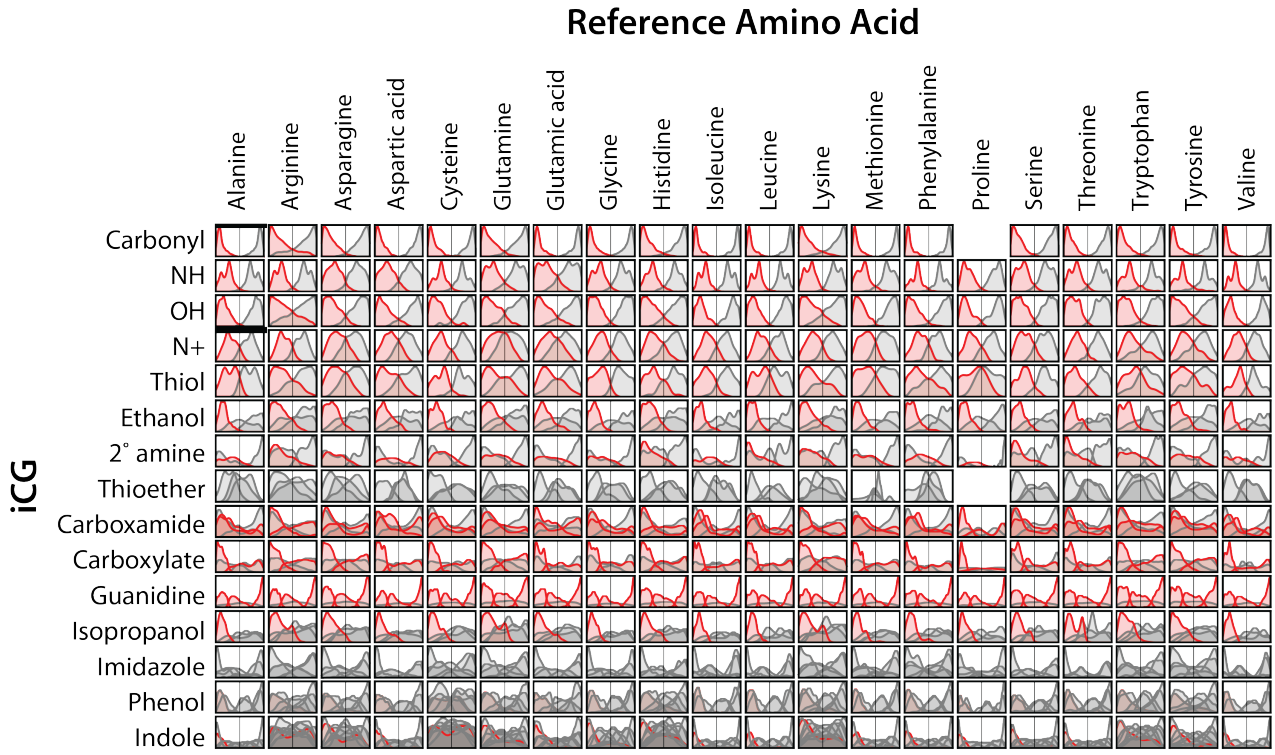

Fig. S5. Orientation of iCG for each vdM type relative to the centroid vector. H-bonding atoms of the iCG are shown in red. There is a clear preference for orientations that place H-bonding atoms towards the reference amino acid centroid

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.

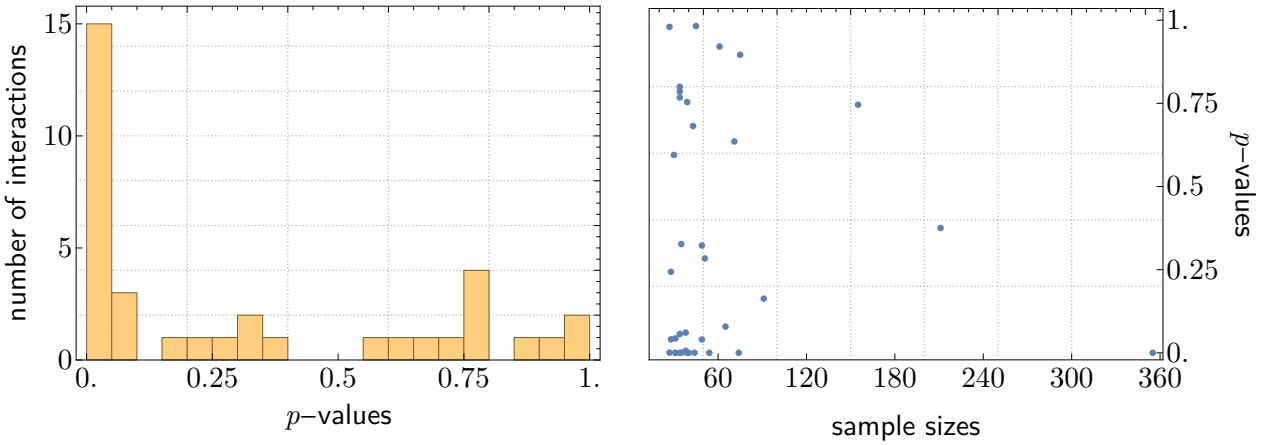

Fig. S6. Distribution of p-values for hypothesis test of transferability on the distance between iCG centroid and amino acid backbone centroid. 32 contact types had sufficient samples for the test to be valid and of these, 15 had p-values less than 1. The p-values do not have a clear correlation with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