Supplementary Material:
Computational design of the Tiam1 PDZ domain and its ligand binding

David Mignon[1], Nicolas Panel[1], Xingyu Chen[1], Ernesto J. Fuentes[2] and Thomas Simonson[1]

[1]Laboratoire de Biochimie (UMR CNRS 7654), Ecole Polytechnique, Palaiseau, France
[2]Department of Biochemistry, Roy J. and Lucille A. Carver College of Medicine, and Holden Comprehensive Cancer Center, University of Iowa, Iowa City, Iowa 52242-1109, United States

Details on the Proteus model variant optimized using $\epsilon_P = 4$

We list in Table 1 below the amino acid composition obtained with the Proteus model optimized using a protein dielectric of $\epsilon_P = 4$. The corresponding reference energies are listed in Table 2 of the main text. We also show (Fig. 1) a histogram of Blosum similarity scores for the Proteus sequences designed with the $\epsilon_P = 4$ model, which are very good, and comparable to the results with the $\epsilon_P = 8$ model (main text).

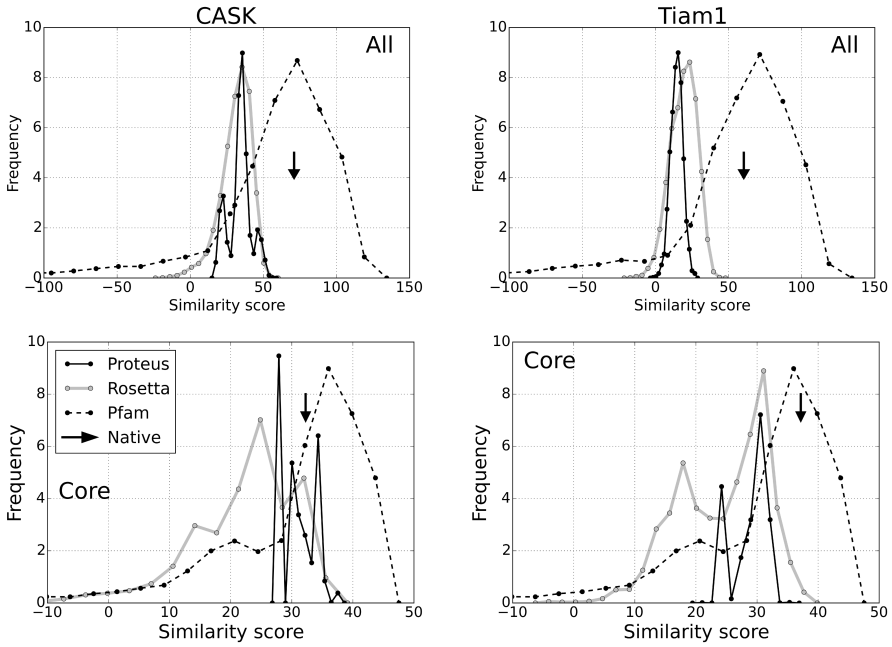

Figure 1: Histogram plots showing similarity scores for designed PDZ sequences, using the model with $\epsilon_P = 4$. Similarity scores for Tiam1 (left) and Cask (right), relative to the Pfam-RP55 alignment. The scores were computed for all positions (top) and 14 core positions (bottom). Values are shown for Proteus, Rosetta, and Pfam sequences (all compared to RP55). The similarity score of the wildtype sequence is indicated in each panel by a vertical arrow.

Table 1: Amino acid composition (%) of natural and designed PDZ proteins ($\epsilon_P = 4$)

type	Natural sequences				Designed sequences			
	Buried		Exposed		Buried		Exposed	
	type	class	type	class	type	class	type	class
A	5.9		4.6		8.1		5.6	
C	1.5	11.2	1.2	13.4	4.4	12.5	3.1	13.8
T	3.8		7.6		0.0	[1.3]	5.1	[0.4]
S	4.7	4.7	10.2	10.2	4.4	4.4 [-0.3]	10.7	10.7 [0.5]
D	3.5	9.6	6.2	16.7	3.3	11.0	0.9	16.9
E	6.1		10.5		7.7	[1.4]	16.0	[0.2]
N	1.9	2.7	7.4	16.1	1.0	2.5	3.5	15.7
Q	0.8		8.7		1.5	[-0.2]	12.2	[-0.4]
H^+	0.7		4.7		0.5		3.9	
H_ϵ	0.0	0.7	0.0	4.7	0.2	0.8	0.7	4.7
H_δ	0.0		0.0		0.1	[0.1]	0.1	[0.0]
I	15.7		4.1		15.4		2.7	
V	13.5	49.6	5.5	14.4	18.0	48.8	2.1	14.4
L	20.4		4.8		15.4	[-0.8]	9.6	[0.0]
M	5.0	5.0	1.4	1.4	3.9	3.9 [-1.1]	1.4	1.4 [0.0]
K	6.5	6.5	10.1	10.1	6.9	6.9 [0.4]	11.6	11.6 [1.5]
R	1.8	1.8	9.5	9.5	1.5	1.5 [-0.3]	9.8	9.8 [0.3]
F	5.0	5.0	0.4	0.4	3.1	5.1	0.1	0.4
W	0.0		0.0		2.0	[0.1]	0.3	[0.0]
Y	2.9	2.9	0.9	0.9	2.6	2.6 [-0.3]	0.8	0.8 [-0.1]
G	0.0	0.3	1.7	2.1	0.0	0.0	0.0	0.0
P	0.3		0.4		0.0	[-0.3]	0.0	[-0.4]
	type	class	type	class	type	class	type	class

Compositions are given for buried/exposed positions, for individual amino acid types (left) and for classes (right); values in brackets (right) are the deviations between design and experiment per class. The experimental target set included the Tiam1 and Cask homologs.

Cross-validation tests of the Proteus model

Cross-validation tests of the Proteus model are described below. Two different parameterizations (reference energy sets) were derived using natural homologs of two sets of PDZ domains, leading to two model variants, called T+C and n=6, described in the Table below. Both parameterizations used a protein dielectric constant of ϵ_P=8. The first, T+C model used Tiam1 and Cask homologs for parameterization, then was tested on DLG2 and syntenin. The second, n=6 model used DLG2, syntenin, PSD95, GRIP, INAD, and NHERF for parameterization, then was tested on Tiam1 and Cask. Superfamily scores were given in the main text. Below, we show histograms (Figs. 2, 3) of Blosum similarity scores for sequences designed with Proteus using the two models. Performance of the cross-validated models is slightly degraded, as expected, compared to models parameterized then applied to the same proteins.

Table 2: Proteus cross-validation tests

model name	domains whose homologs were used for parameterization	domains whose structures were used for cross validation tests
T+C	Tiam1, Cask	DLG2, syntenin
n=6	DLG2, syntenin, PSD95, GRIP, INAD, NHERF	Tiam1, Cask

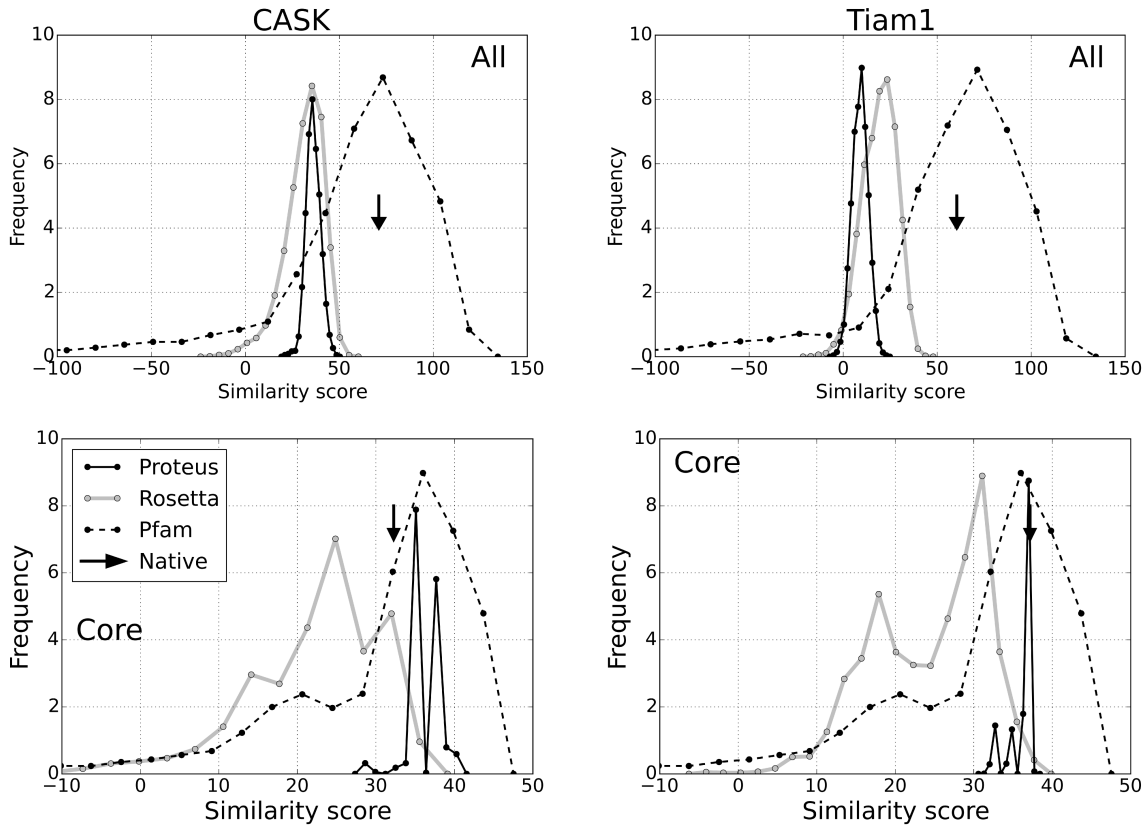

Figure 2: Histogram plots showing similarity scores for PDZ sequences designed with the n=6 model using the Tiam1 and Cask backbone structures. Similarity scores for Cask (left) and Tiam1 (right), relative to the Pfam-RP55 alignment. The scores were computed for all positions (top) and 14 core positions (bottom). Values are shown for Proteus, Rosetta, and Pfam sequences (all compared to RP55). The similarity score of the wildtype sequence is indicated in each panel by a vertical arrow. The Proteus scores can be compared to those obtained with the T+C model variant, shown in the main text (Fig. 5, main text).

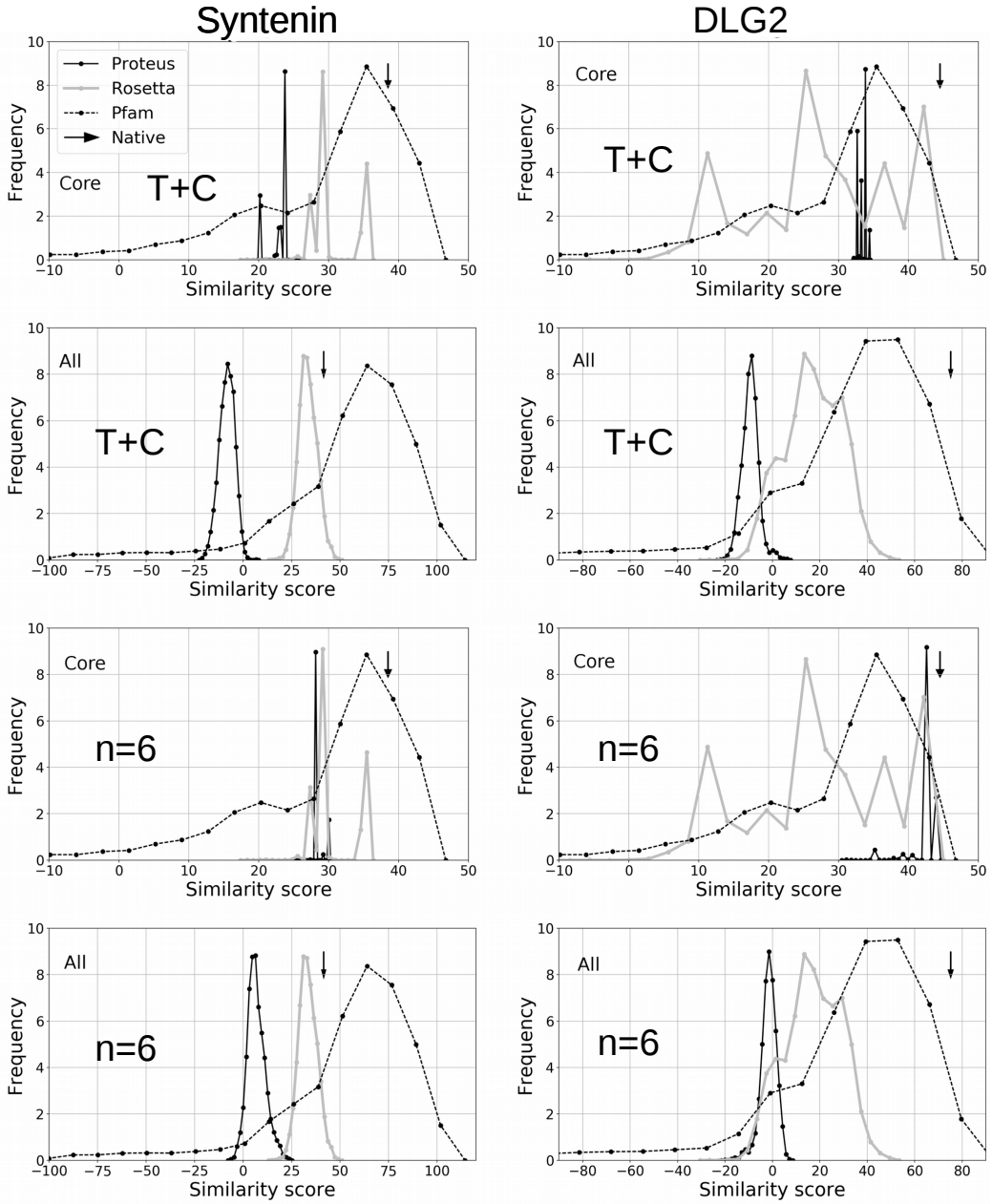

Figure 3: Histogram plots showing similarity scores for PDZ sequences designed with the T+C and n=6 models using the syntenin and DLG2 backbone structures. Similarity scores relative to the Pfam-RP55 alignment. The scores were computed for all positions and 14 core positions. Values are shown for Proteus, Rosetta, and Pfam sequences (all compared to RP55).