Supporting Information for:

Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design

TABLE OF CONTENTS

TEXT

FIGURES

TABLES

REFERENCE

TEXT

Text S1: Overlapping function

We clustered the FUZZLE database using a density-based clustering method. To check if our network topology is a consequence of this clustering we defined an overlapping function. The function is defined as follows: Let D1 and D2 be two domains define a hit A. $D1_A$ and $D2_{A,}$ are the two fragments that define the sequence and structural alignment. Thus, $D1_A$ and $D2_A$ are a unique sub-domain sized fragment present in domains from different folds, that besides being evolutionary related superimpose spatially. Now, let another alignment B between D1 and D3 match subsections $D1_B$ and $D3_B$. If the residues in fragment $D1_B$ are virtually the same ones as those in $D1_A$, then $D1_B$, $D3_{B,}$ $D1_A$ and $D2_A$ are alternative names for the same fragment. There are 208944 that surpass the cutoffs for the construction of the network (see main text). Instead of clustering the domains in these hits by a density method we iteratively computed the overlap among fragments of the same domain with the following formula:

$$\frac{\max(e_A, e_B) - \min(s_A, s_B)}{\min(l_A, l_B)} < 1.11$$

Where e_A, e_B, s_A, and s_B constitute the alignment's ends and starts of domain D1 in the alignment A and B, respectively, and l_A, l_B define the alignment lengths. If $D1_B$ and $D1_A$ overlap at least an x % in position and size, a single node can define this domain, otherwise two nodes will be defined. We constructed networks at several overlap cutoffs (ranging from 1 to 1.5). The number of nodes and connected components is represented in **Fig. S4.**

FIGURES

Figure S1: Protein similarity networks using different cutoffs: TMscore **(a)**, RMSD **(b)**, and sequence/structural length ratio (S_{Aln}/S_{Str}) **(c)** cutoffs. Probability and structural alignment length were kept as in the main manuscript (probability over 70 % and length between 10 and 200 amino acids). For each parameter, the plot indicates the number of fragments (blue axis) and nodes in the major component (red axis) as a function of the parameter. Two networks are shown for every parameter (denoted as 1 and 2), one with a looser stringent cutoff, and another with more stringent cutoff than the one in the main manuscript, depicted with a green line.

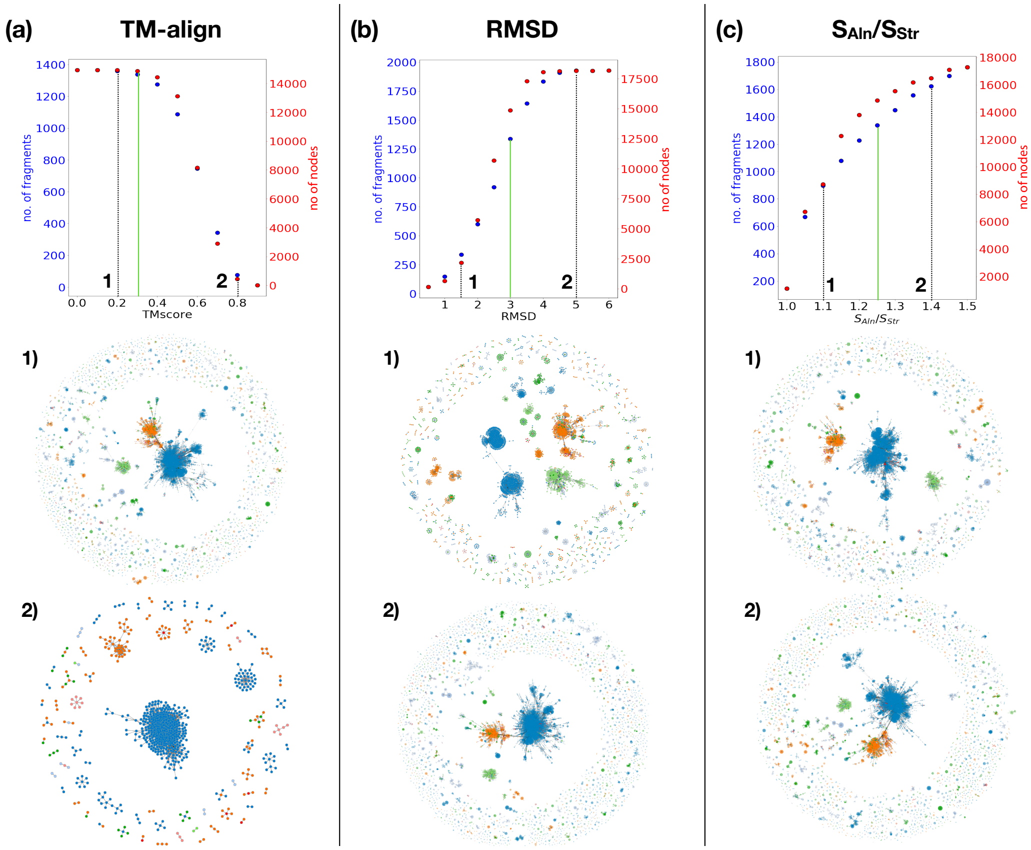


Figure S2: Log-log distribution of domains vs degree of connectivity. The y-axis represents the number of domains/nodes that present a certain degree of connectivity (x-axis).

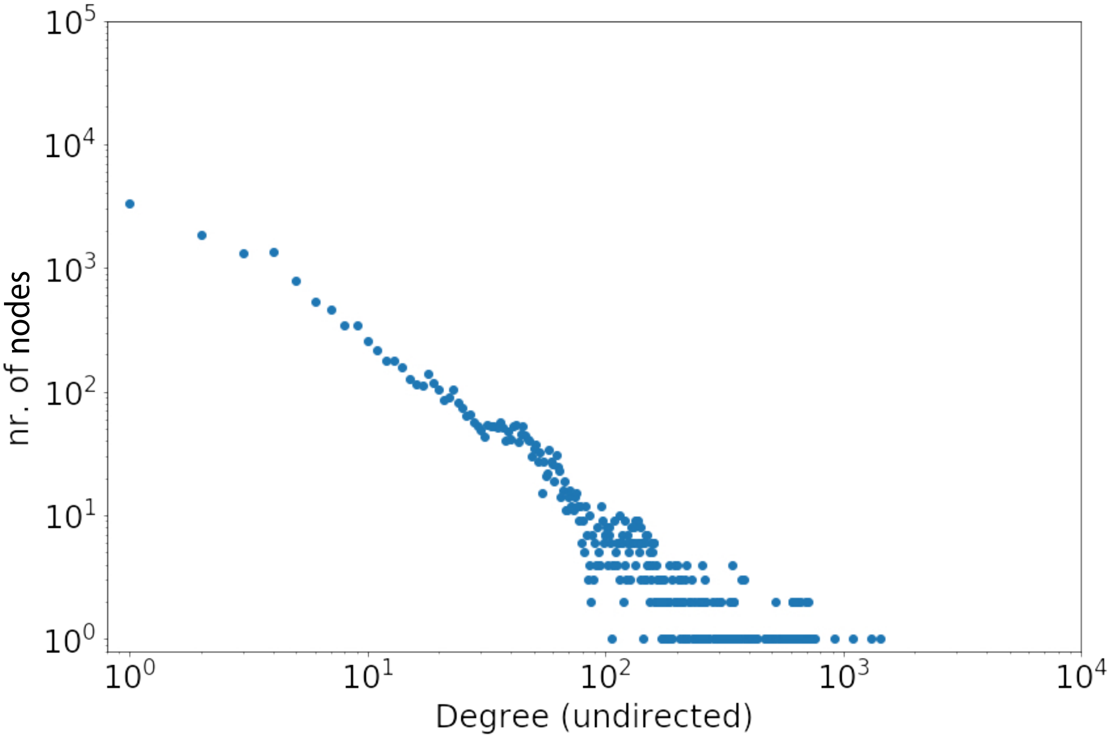


Figure S3: Network built from 1,000 randomly chosen domains from each of the main four classes. We took 4,000 random Fuzzle hits where query and subject belonged to the main four SCOPe classes as follows: 700 hits where query and subject belong to the same class, and 100 hits where query and subject belong to different classes. For example, for the *a* class, 700 random hits were taken such as query and subject belonged to a class *a* (*a*-*a*), and another 100 where query and subject belonged to different classes (*a*-*b*, *a*-*c*, and *a*-*d*).

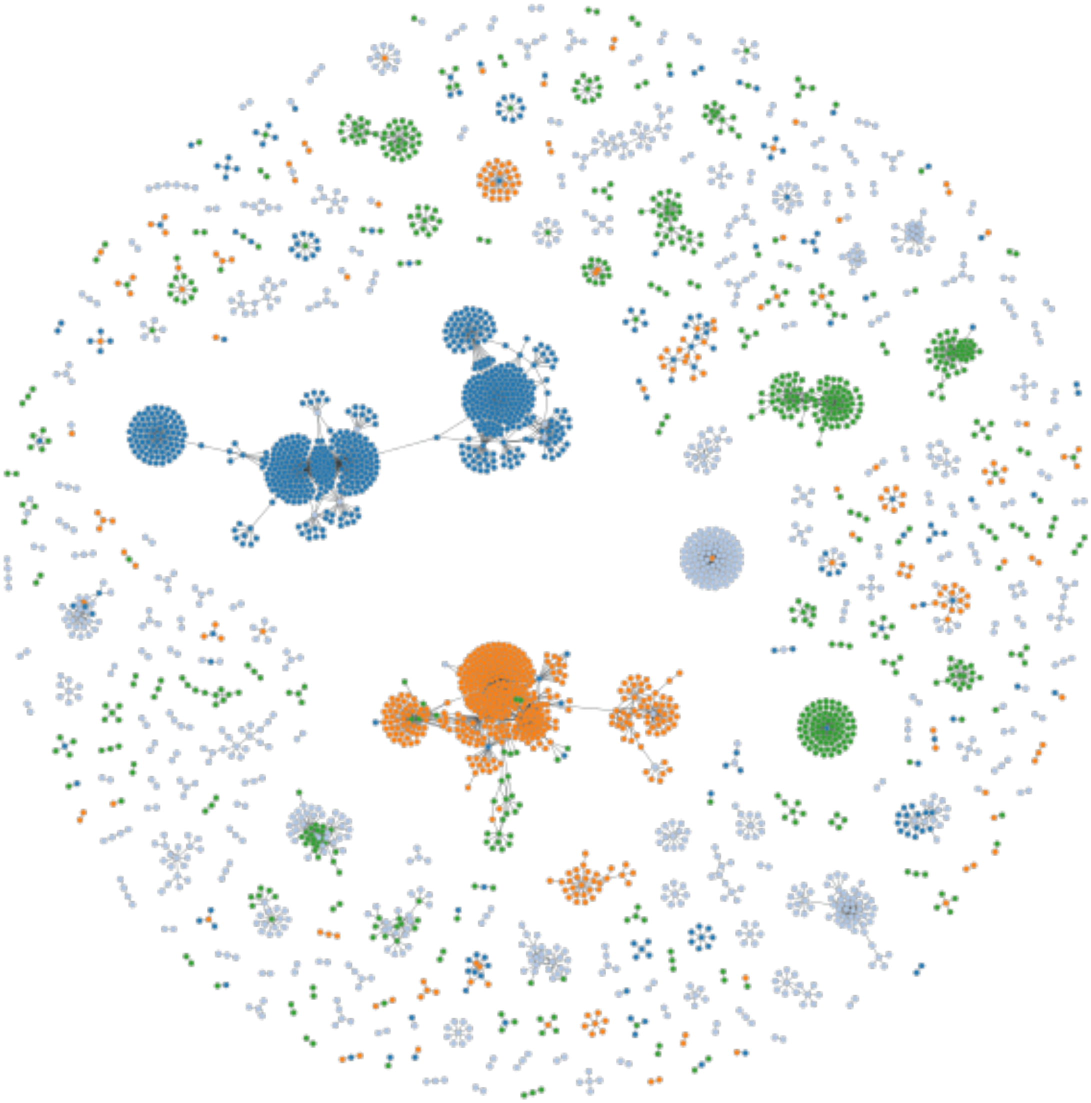


Figure S4: Number of fragments and nodes in the major component using an overlap function (**Text S1**). The rest of the database parameters were kept as in the main manuscript: probability > 70%, structural alignment between 10 and 200 amino acids, RMSD < 3 Å, TM-score > 0.3, and sequence/structural length ratio (S_{Aln}/S_{Str}) < 1.25 **(a)** Number of fragments and nodes in the major component as a function of the overlap. The overlap that resembles most the results in the manuscript is an overlap of 1.1, which leads to 1,245 fragments and 14,893 nodes in the major component **(c)**. Protein similarity networks for an overlap of 90% **(c)** and 60% **(d)**.

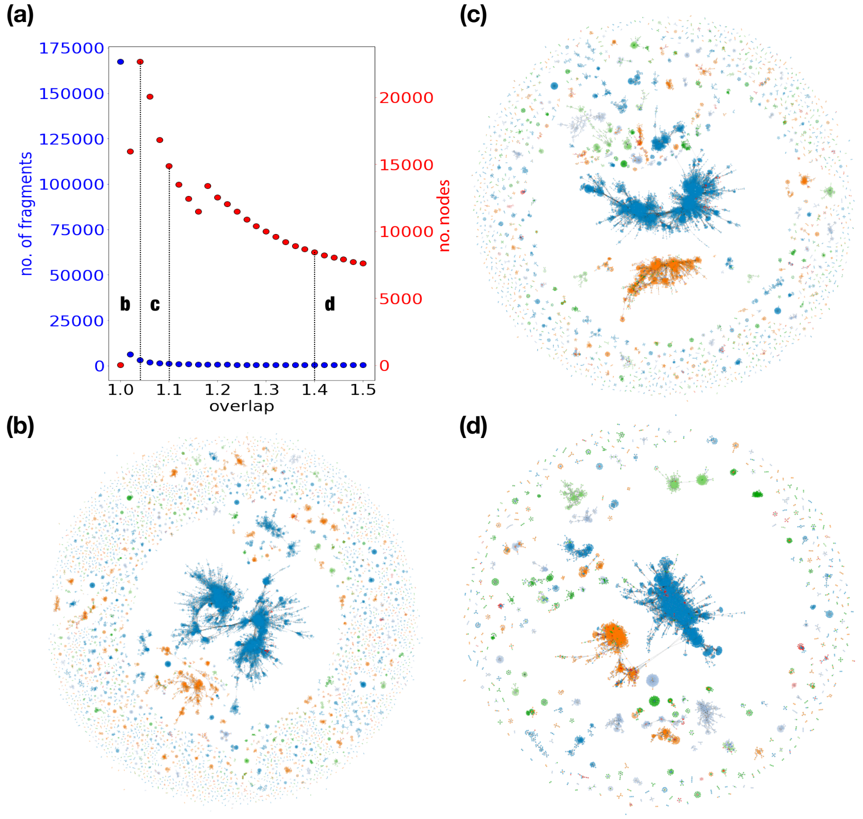


Figure S5: Fragments between the all-α and α/β classes in the major component (fragment 0). Superpositions of three fragments that are shared between folds of the all-α (green) and α/β (blue) are shown as cartoons.

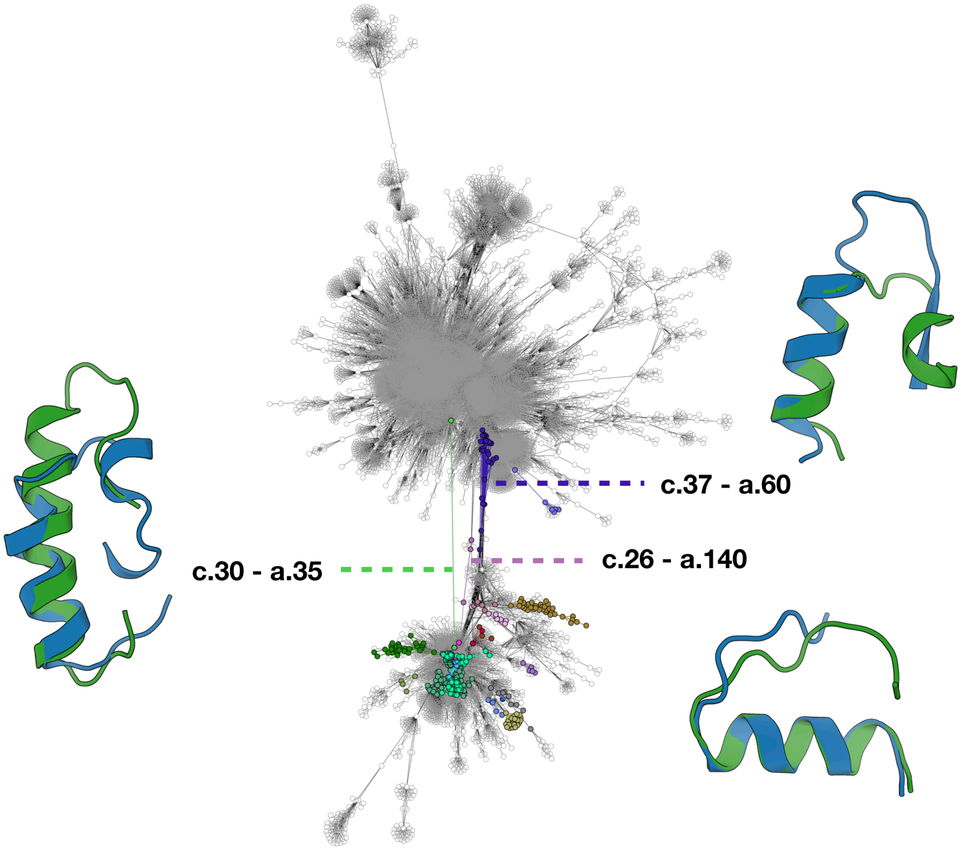
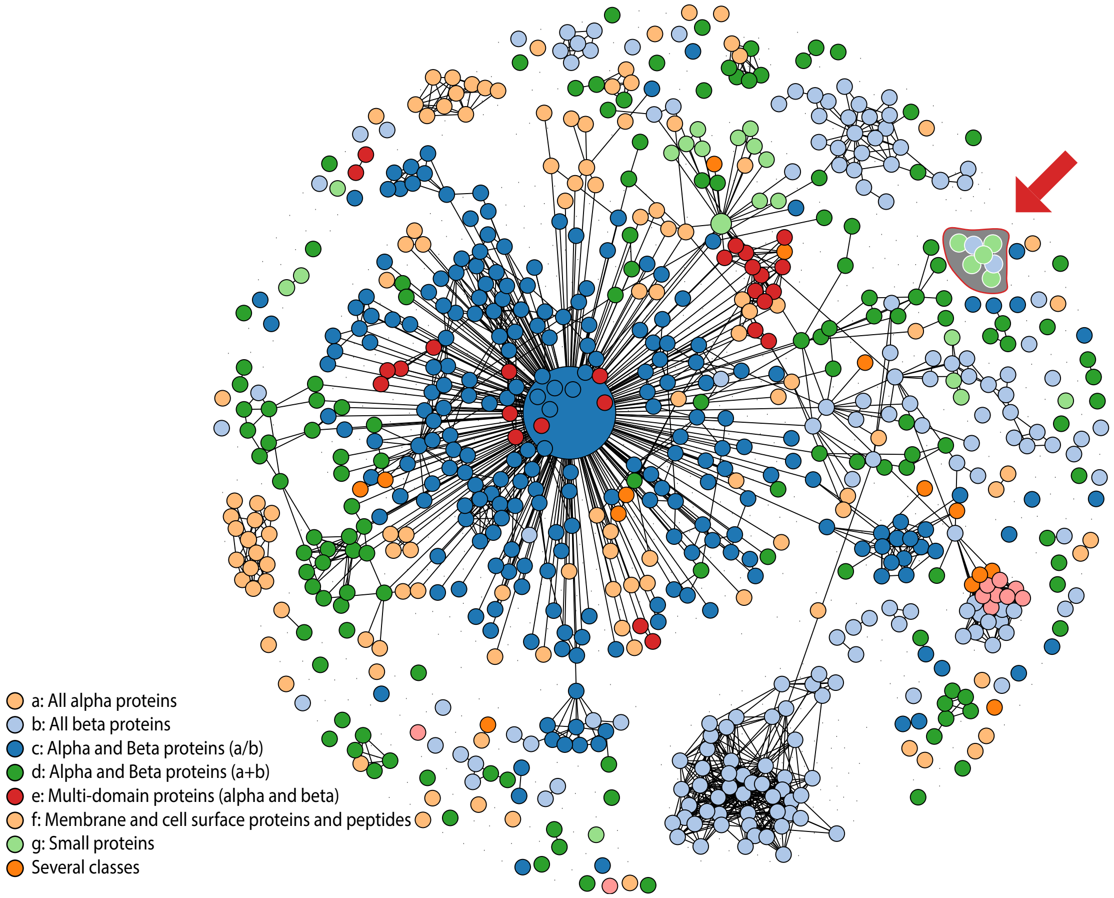


Figure S6: Fragment universe as shown on the Fuzzle website. In contrast to Fig. 4 in the manuscript, here nodes initially represent components instead of single domains to enable interactive browsing. Here each node contains several domains that share a common fragment. The individual domains are shown upon clicking on one of the black-circled nodes. In this figure the user clicked on one node indicated by the red arrow, which expands to a grey area. All the white-circled nodes within this area now show the domains contained in this cluster.


a: All alpha proteins
b: All beta proteins
c: Alpha and Beta proteins (a/b)
d: Alpha and Beta proteins (a+b)
e: Multi-domain proteins (alpha and beta)
f: Membrane and cell surface proteins and peptides
g: Small proteins
Several classes

TABLES

Table S1: 10 most connected hubs in the network ordered by decreasing degree.

Compontent/ Fragment No.	Domain	Degree	SCOP fold	Different folds among neighbors
182	d1j6ua1	1,423	c.5	25
184	d1p3da1	1,317	c.5	26
183	d4hv4a1	1,096	c.5	20
196	d2x5oa1	921	c.5	20
626	d1ebda2	756	c.3	12
639	d1v59a2	738	c.3	8
558	d1jw9b_	731	c.111	8
657	d1c0pa1	712	c.4	8
678	d2bi7a1	712	c.4	9

Table S2: 10 most promiscuous component in the network ordered by decreasing number of fold neighbors.

Component/ Fragment No.	Domain	Degree	SCOP fold	Different folds among neighbors
184	d1p3da1	1,317	c.5	26
182	d1j6ua1	1,423	c.5	25
183	d4hv4a1	1,096	c.5	20
196	d2x5oa1	921	c.5	20
1096	d2x5oa1	367	c.5	20
623	d1a9xa4	552	c.30	18
1042	d1b0nb_	106	a.34	18
722	d1seza1	667	c.3	17
2750	d1lssa_	214	c.2	17

Table S3: Connections between domains from the all-α and α/β classes in the major component shown by folds. Connections that are shown in **Fig. S5** are highlighted in gray.

Fold pair	**Number of links**
c.23 - a.4	155
c.47 - a.4	69
c.66 - a.156	68
c.43 - a.43	39
c.37 - a.60	35
c.23 - a.35	23
c.45 - a.5	23
c.55 - a.60	8
c.37 - a.5	8
c.113 - a.60	6
c.47 - a.140	5
c.93 - a.35	4
c.123 - a.60	4
c.1 - a.34	3
c.25 - a.43	3
c.26 - a.140	2
c.15 - a.5	2
c.30 - a.35	1
c.25 - a.5	1
c.37 - a.34	1
c.23 - a.43	1
c.26 - a.34	1

Table S4: Top 20 most popular fold pairs in the major component.

Fold pair	Number of links
c.3 - c.2	39,822
c.66 - c.2	21,306
a.4 - a.35	15,059
c.30 - c.2	7,811
c.91 - c.37	5,804
c.78 - c.2	5,596
c.4 - c.2	4,610
c.5 - c.2	3,412
c.2 - c.111	2,787
c.23 - c.1	2,512
a.6 - a.4	2,397
c.65 - c.2	2,177
c.37 - c.2	2,094
c.93 - c.23	1,635
c.72 - c.2	1,074
c.4 - c.3	1,014
c.79 - c.2	903
c.72 - c.37	894
a.74 - a.4	864
c.30 - c.3	802

REFERENCES

1. Alva V, Söding J, Lupas AN: **A vocabulary of ancient peptides at the origin of folded proteins**. *Elife* 2015, **4**.