Supplementary Material

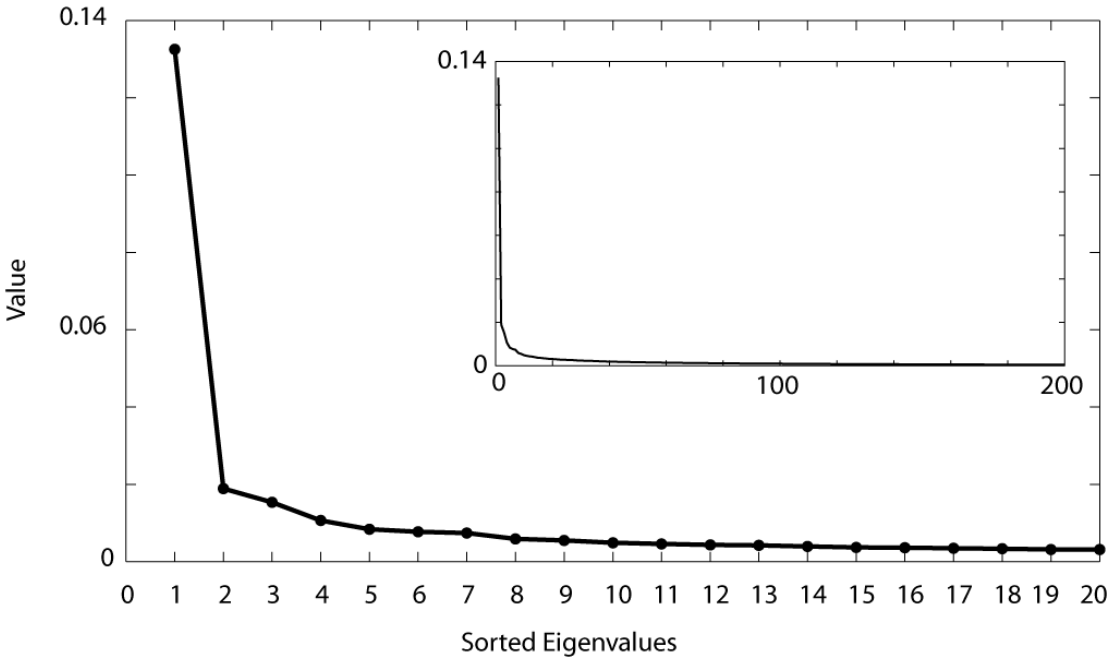

Figure 1S: Scree plot of the 400 dimensional data. The Figure shows the 20 largest eigenvalues of the (normalized) correlation matrix sorted in decreasing order; the insert shows the largest 200 eigenvalues of this matrix. The sharp drop up to the third eigenvalues suggests that three dimensions can adequately represent the essential features of protein structure space.

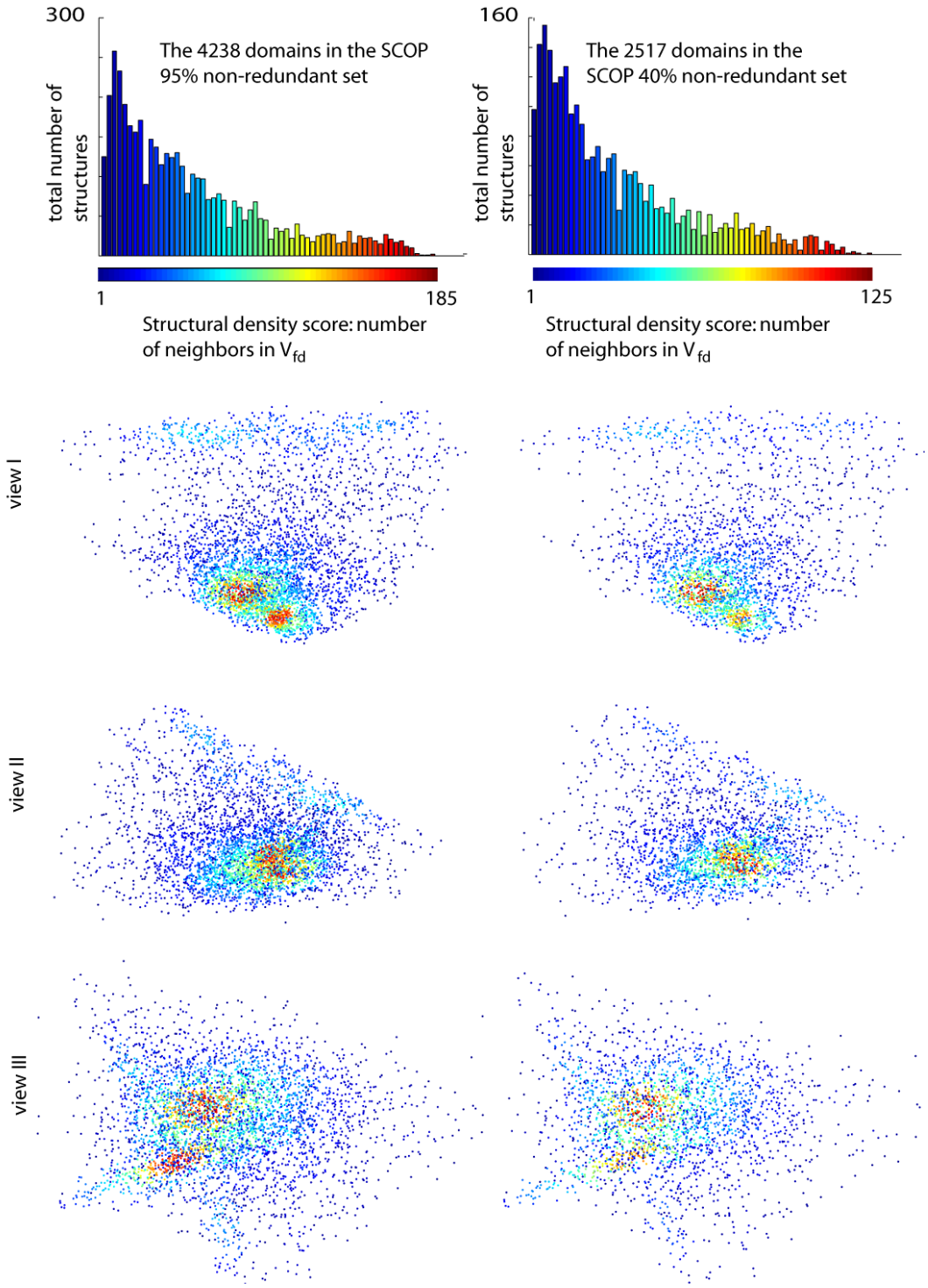

Figure 2S: Density maps of protein structure space for sequence non-redundant subsets. The points on the map are colored according to the number of domains that lie within 0.005 distance from them in the dataset considered. In the left column, the map is of a subset of size 4238, in which the sequence identity between any two proteins is at most 95%; in the right column, that map is of a subset of size 2517, in which the sequence identity is at most 40%. The correlation coefficients between the density the full sets and the 95% and 40% non redundant subsets are r=0.960 and r=0.945 respectively.

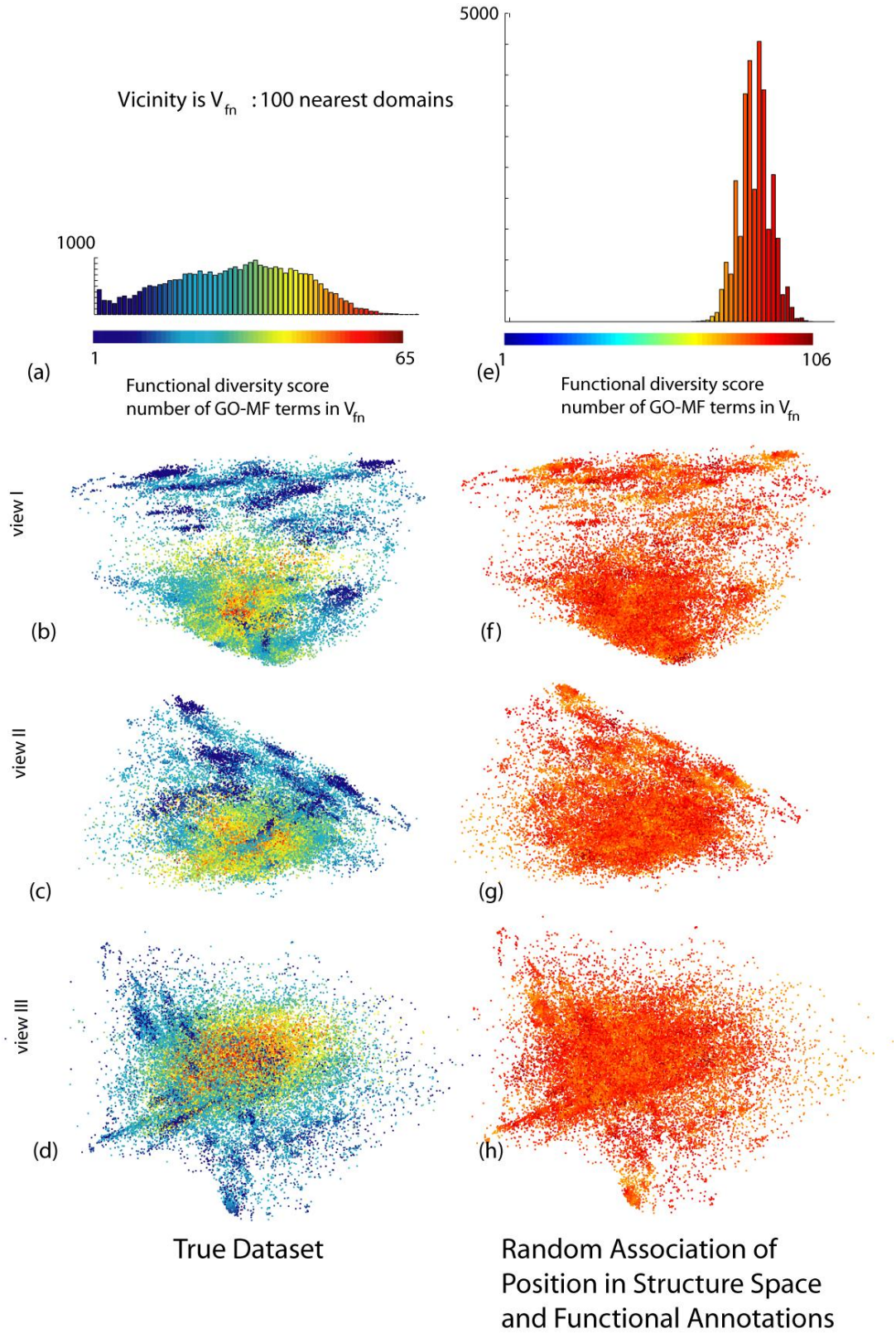

Figure 3S: Functional-diversity maps of protein structure space with vicinity defined as V_{fn}: The points on the map are colored according to their functional diversity measured by the number of distinct GO-MF terms annotating the domains in the V_{fn} vicinity of 100 nearest neighbors. In panels (a-d) we show the map for the true dataset; in panels (e-f) we randomly associate functions with domains and re-calculate the map. The map for the true dataset has a distinct high-diversity core. When functions are associated at random with the domains, the V_{fn} vicinity of all domains is (uniform and) highly diverse.

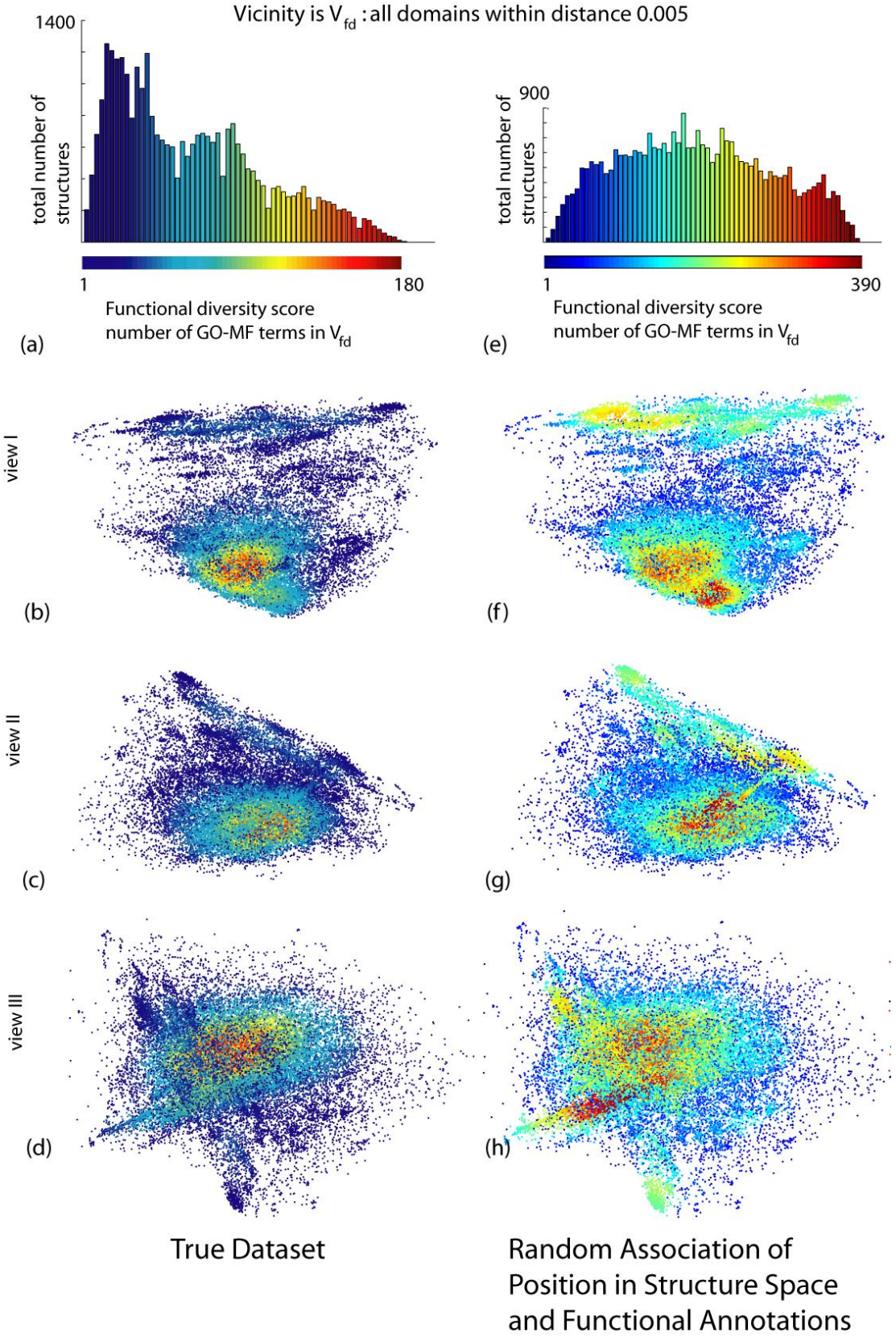

Figure 4S: Functional-diversity maps of protein structure space with vicinity defined as V_{fd}: The points on the map are colored according to their functional diversity measured by the number of distinct GO-MF terms annotating the domains in the V_{fdi} vicinity of distance 0.005. In panels (a-d) we show the map for the true dataset; in panels (e-f) we randomly associate functions with domains and re-calculate the map. The map for the true dataset has a distinct high-diversity core. When functions are associated at random the map is very similar to the structural density map (see Figure 1(e-f)).

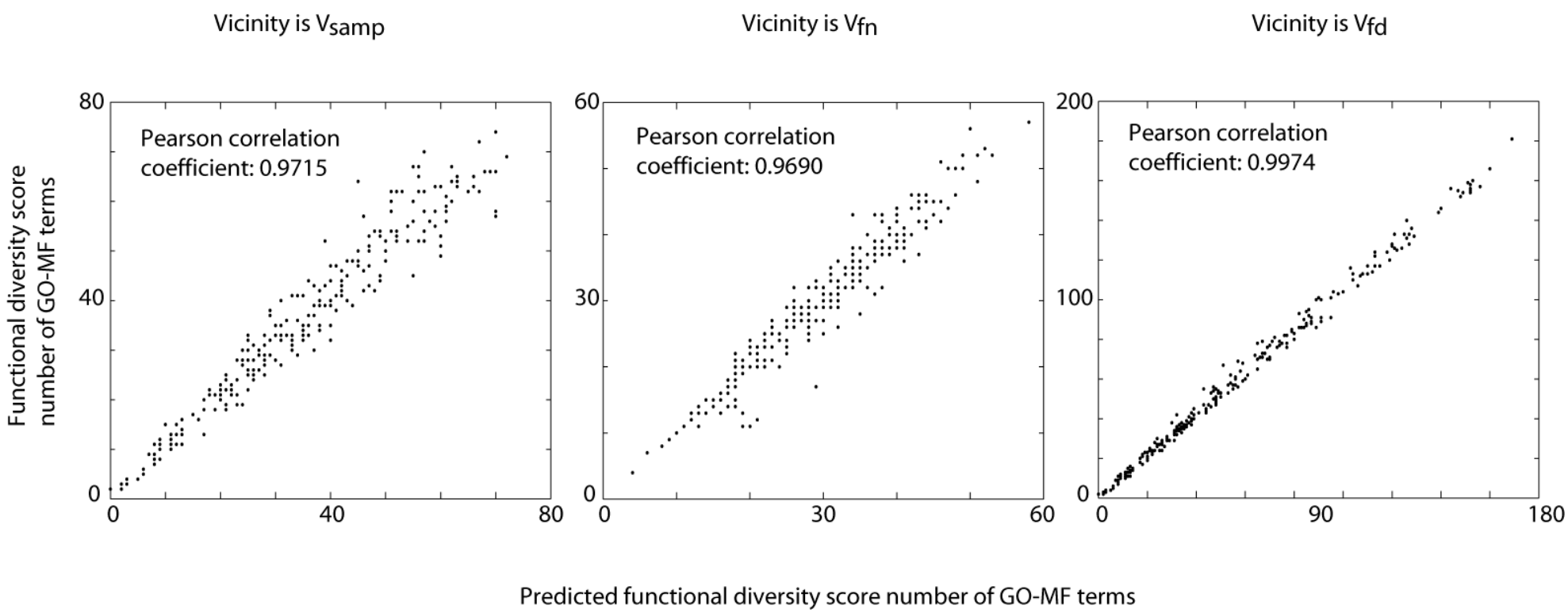

Figure 5S: The predicted functional diversity score vs. the functional diversity score calculated using the full dataset, for a test set of 250 randomly chosen structures: We consider the three definitions of local vicinities: V_{samp}, V_{fn}, and V_{fd}. We calculate the projection to three-dimensions based on set that does not include the 250 test set proteins and their sequence homologues. The predicted functional diversity of a test set protein is the number of unique GO-MF terms in the vicinity of the location calculated for the structure using that projection to a lower dimension. In all three cases, the agreement of the predicted functional diversity scores and the functional diversity score calculated using the full dataset is very good

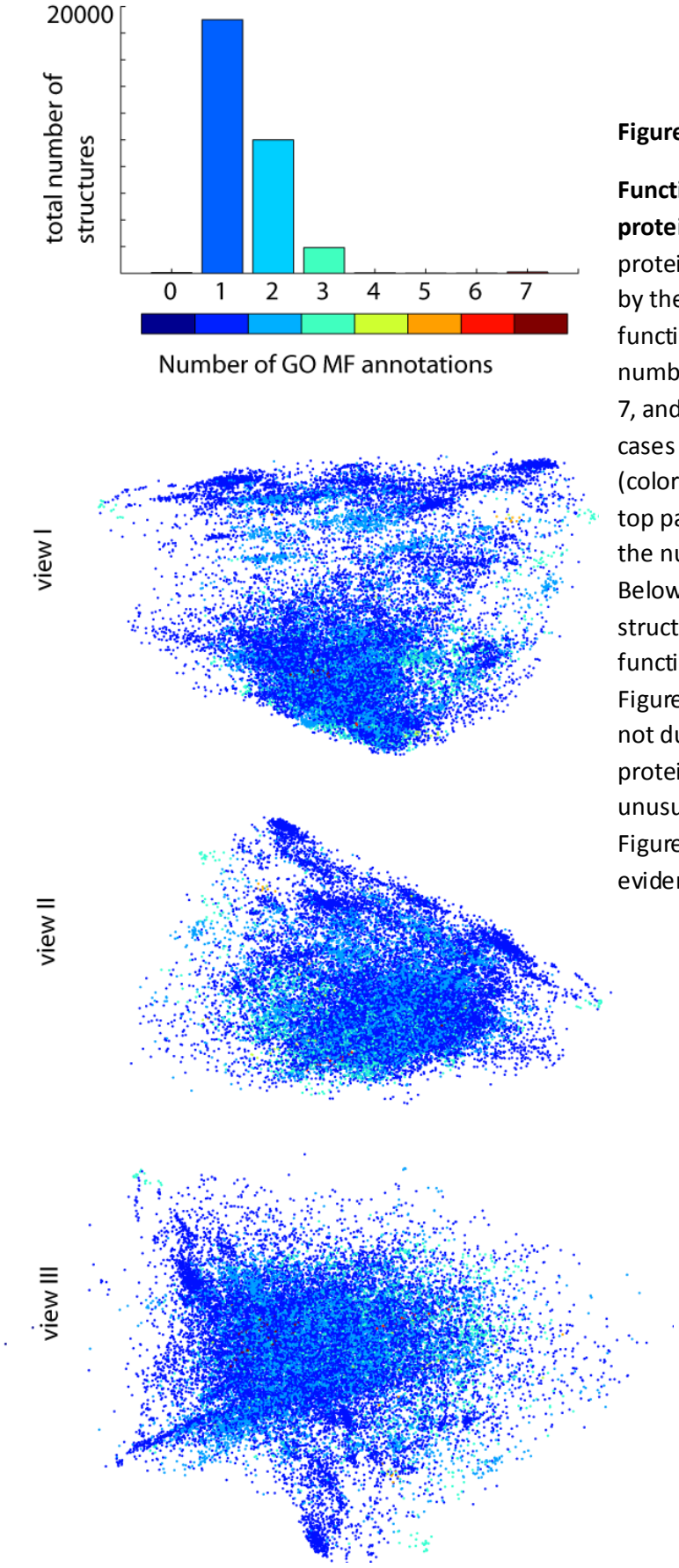

Figure 6S:

Functional multiplicity map of protein structure space. Each protein structure is color-coded by the number of its GO-MF functional annotations. The number of annotations is at most 7, and in the vast majority of the cases (99.4%) it is less than three (colored by shades of blue); the top panel shows a bar diagram of the number of annotations. Below, there are three views of structure space. The high functional diversity core seen in Figure 2 in the main document is not due to the small set of proteins that are annotated by unusually many functions (see Figure 6S below for additional evidence).

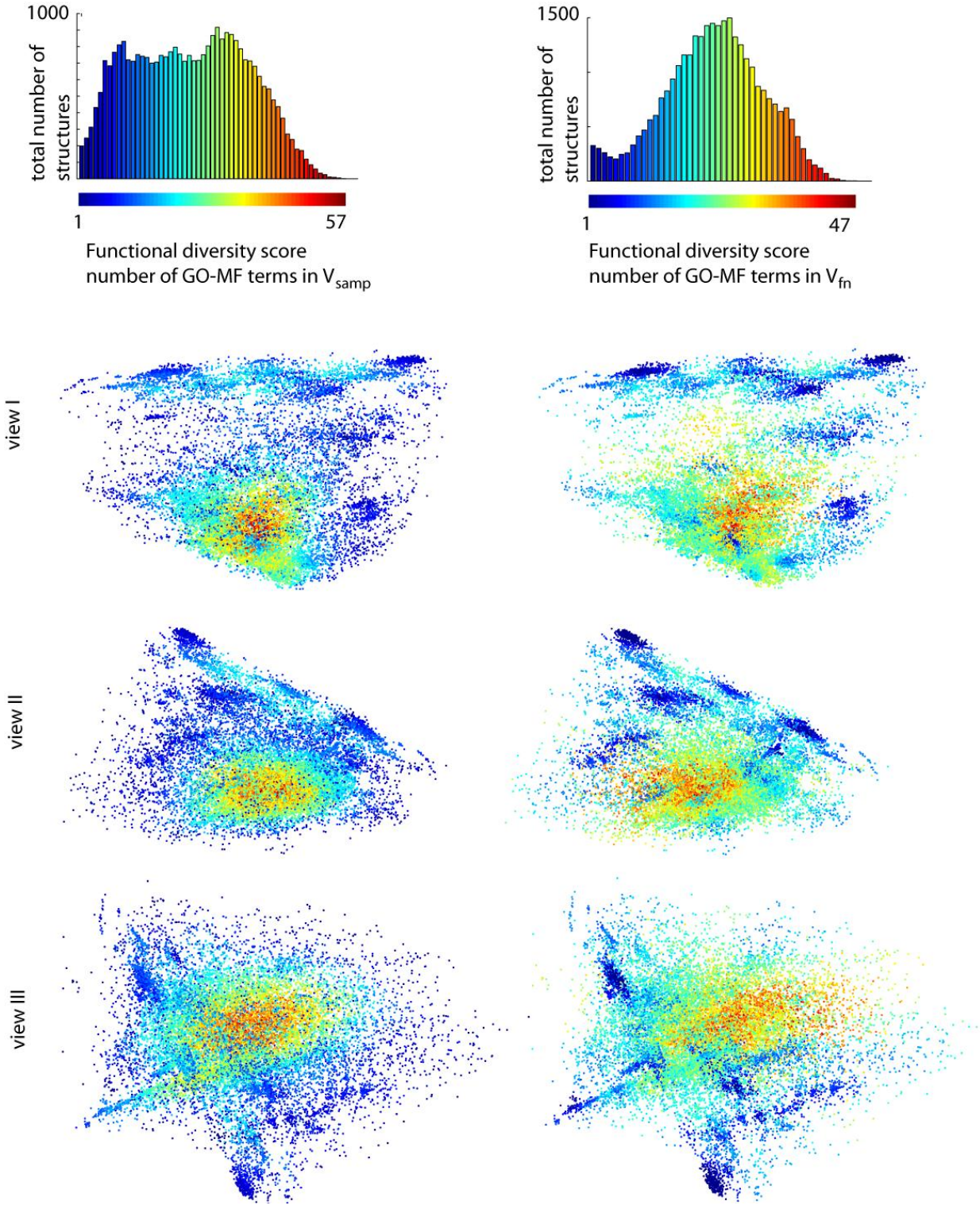

Figure 7S: Functional diversity map of protein structure space using only proteins annotated by one function: We restrict our attention to domains annotated by at one GO term (61.04% of the full data set). The correlation coefficients between the scores calculated using only this subset, and when calculating using the full dataset are listed in Table 1S below. We see the high diversity core in this dataset as well, meaning it cannot be explained away as a consequence of multiply-annotated domains.

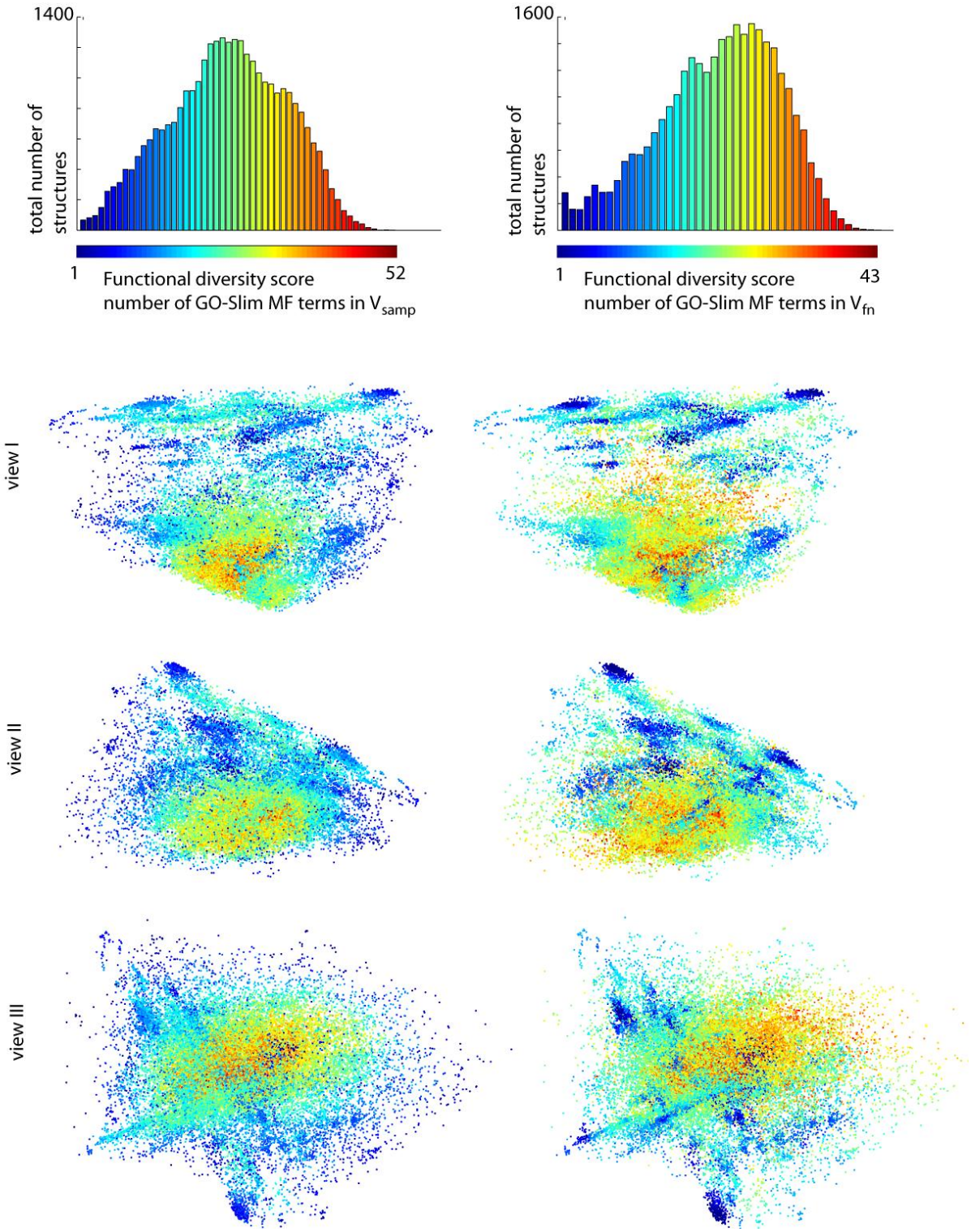

Figure 8S: Functional diversity maps of protein structure space using the GO-Slim annotation: Here we use a more restricted ontology for the annotation of function, and replace each annotation by its most specific parent in the GO-slim graph (*1*). The maps on the left column use the definition of vicinity V_{samp}, and on the right V_{fn}. The correlation coefficients between these measures and the straight-forward count of distinct GO-terms are listed in Table 1S below. We see the same characteristic highly diverse core and the same drop in diversity towards the periphery of structure space, demonstrating that our finding is not an artifact of the uneven level of detail in the GO-MF graph.

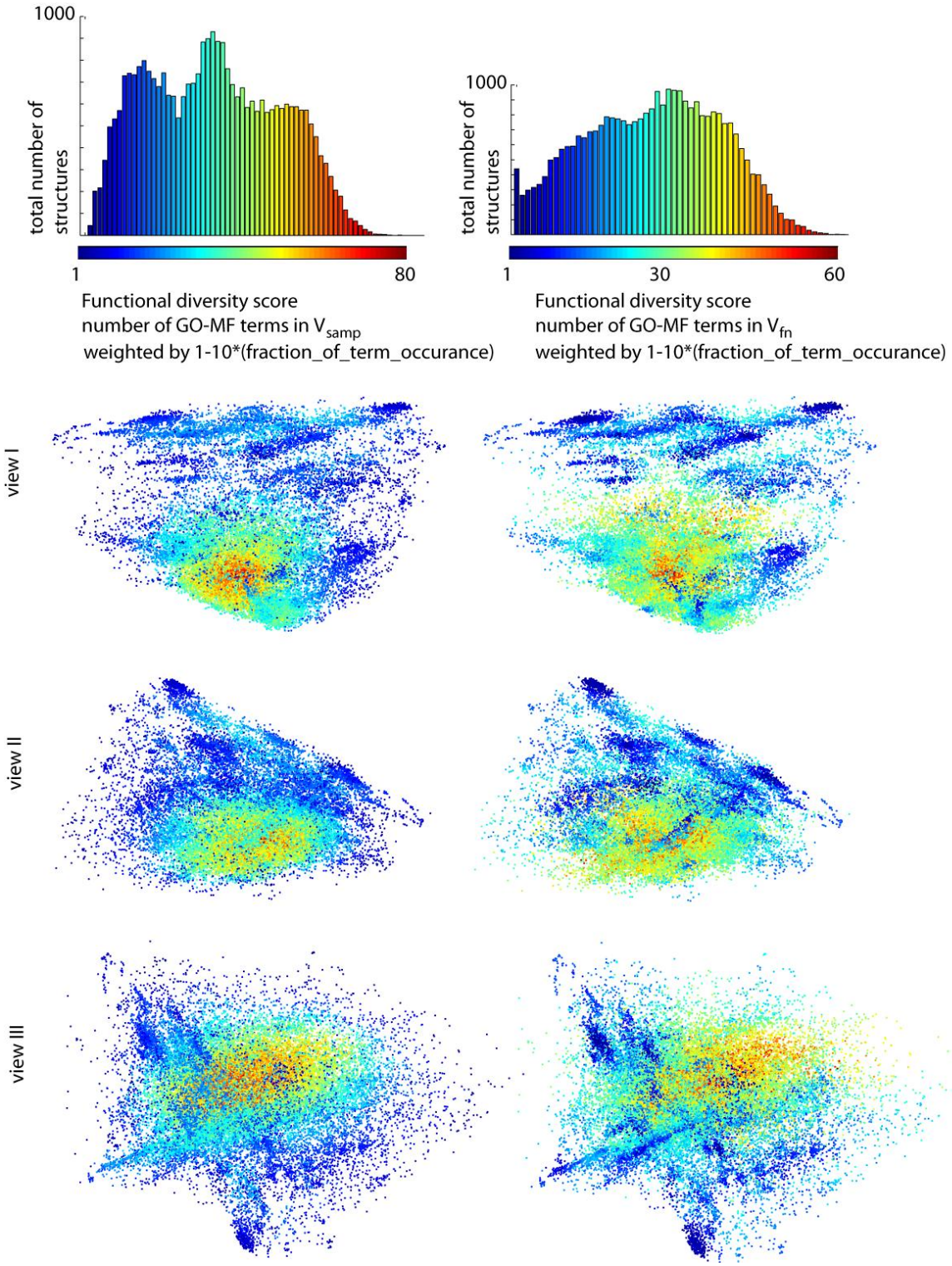

Figure 9S: Functional diversity maps using alternative scoring functions I. The functional-diversity score used to construct these maps is the weighted sum of distinct terms within a vicinity of a domain; the weight of a term depends on how common it is in the dataset, with more common terms contributing less. The maps on the left column use the definition of vicinity V_{samp}, and on the right V_{fn}. The correlation coefficients between these measures and the straight-forward count of distinct GO-terms are listed in Table 1S below. We see our main finding here too: a highly diverse core surrounded by less diverse regions.

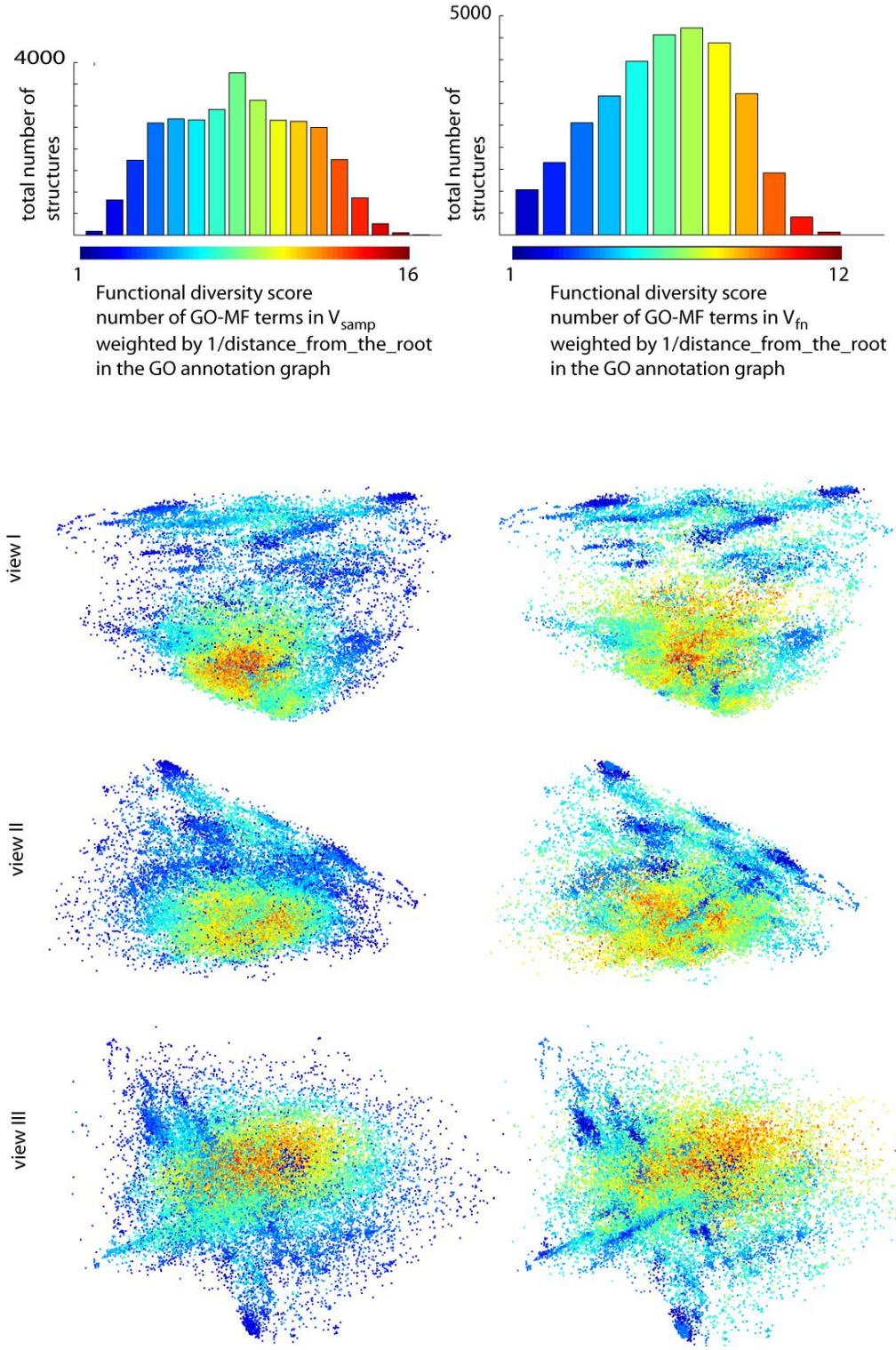

Figure 10S: Functional diversity maps using alternative scoring functions II. The functional-diversity score used to construct these maps is the weighted sum of distinct terms within a vicinity of a domain; the weight of a term is its specificity in the GO annotation graph, with more specific terms contributing less (the inverse of its distance from the root). The maps on the left column use the definition of vicinity V_{samp}, and on the right V_{fn}. The correlation coefficients between these measures and the straight-forward count of distinct GO-terms are listed in Table 1S below. Here too, we see our main finding: a highly diverse core surrounded by less diverse regions.

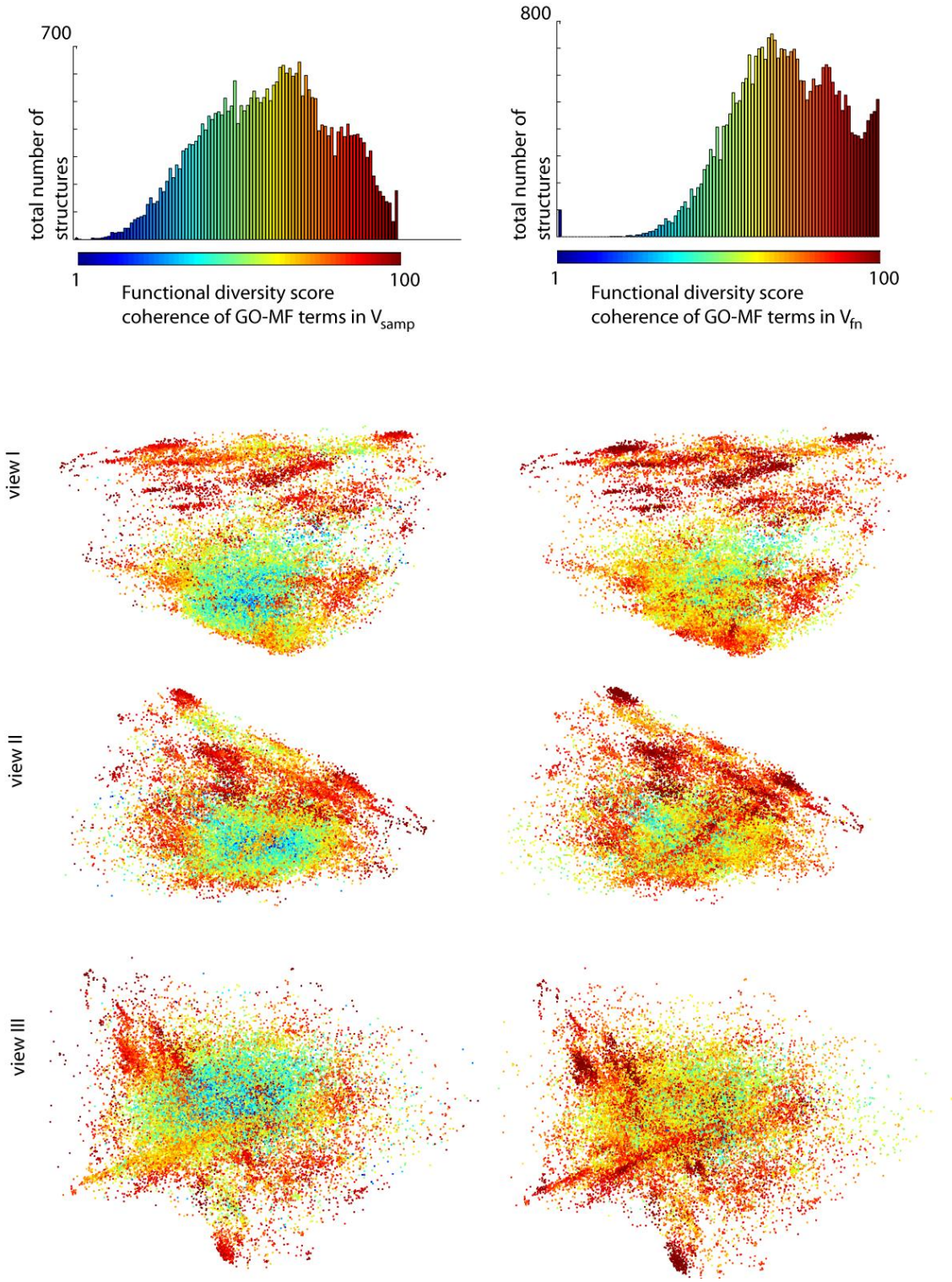

Figure 11S: Functional diversity maps alternative scoring function III. The functional-diversity score used to construct these maps is the 'coherence measure' of the functional GO-MF terms in the neighborhood of a protein, as suggested in (*4*) (*5*). Thus, this score ranges between 0-100%, and higher values imply proteins centered at regions that are less functionally diverse (since all their terms are unique to that region (see Methods for details). The maps on the left column use the definition of vicinity V_{samp}, and on the right V_{fn}. The correlation coefficients between these measures and the straight-forward count of distinct GO-terms are listed in Table 1S below. Using this score too, we see our main findings.

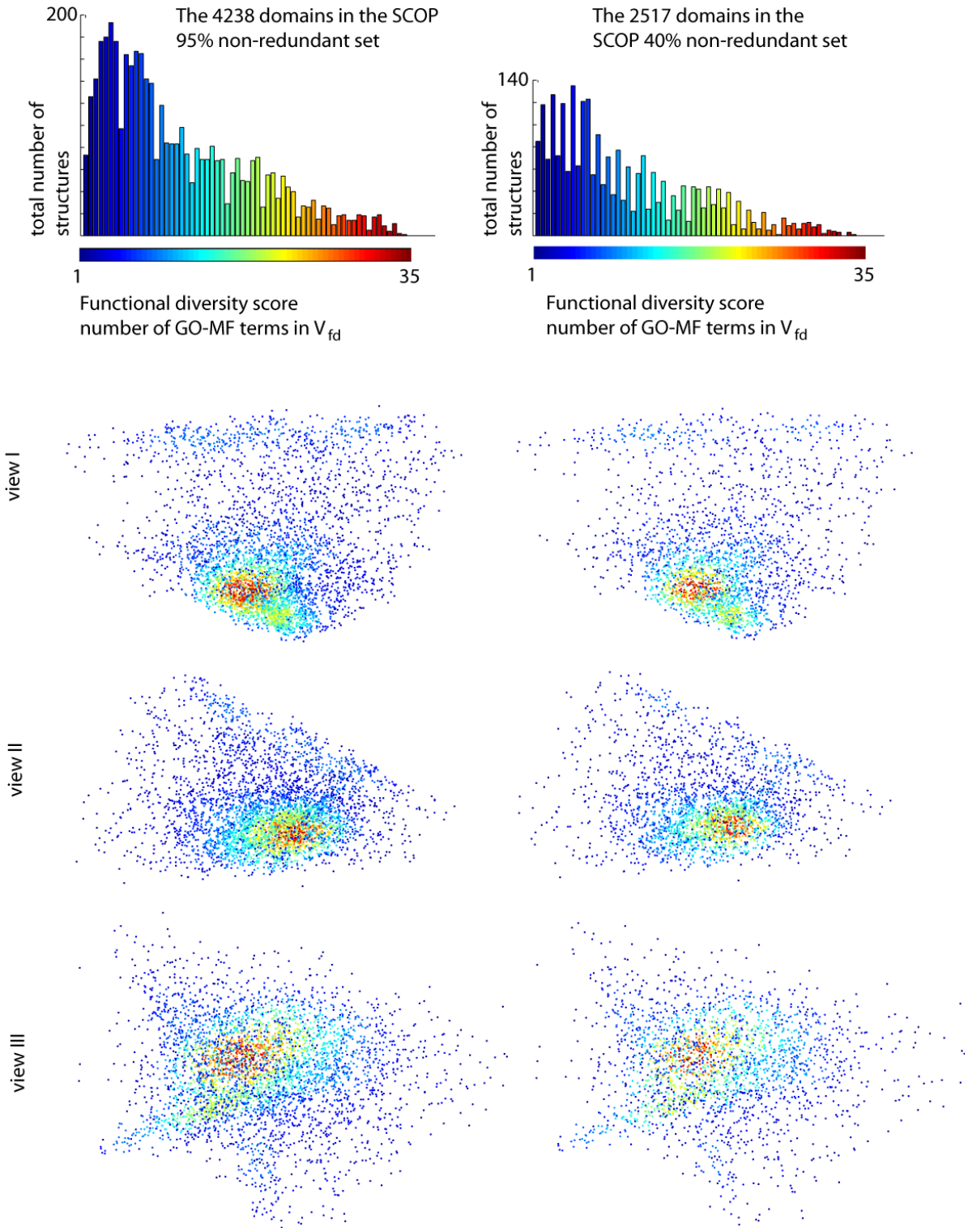

Figure 12S: Functional density scores when sampling the full dataset based on sequence. The vicinity of a protein is V_{fd}, and the diversity score is the number of distinct GO-MF terms in the annotations of the proteins in the vicinity. The correlation coefficients of this diversity score and the straightforward one using the full dataset is listed in Table 1 below. Even when using this very sparsely sampled subset, we see the same core of high functional diversity.

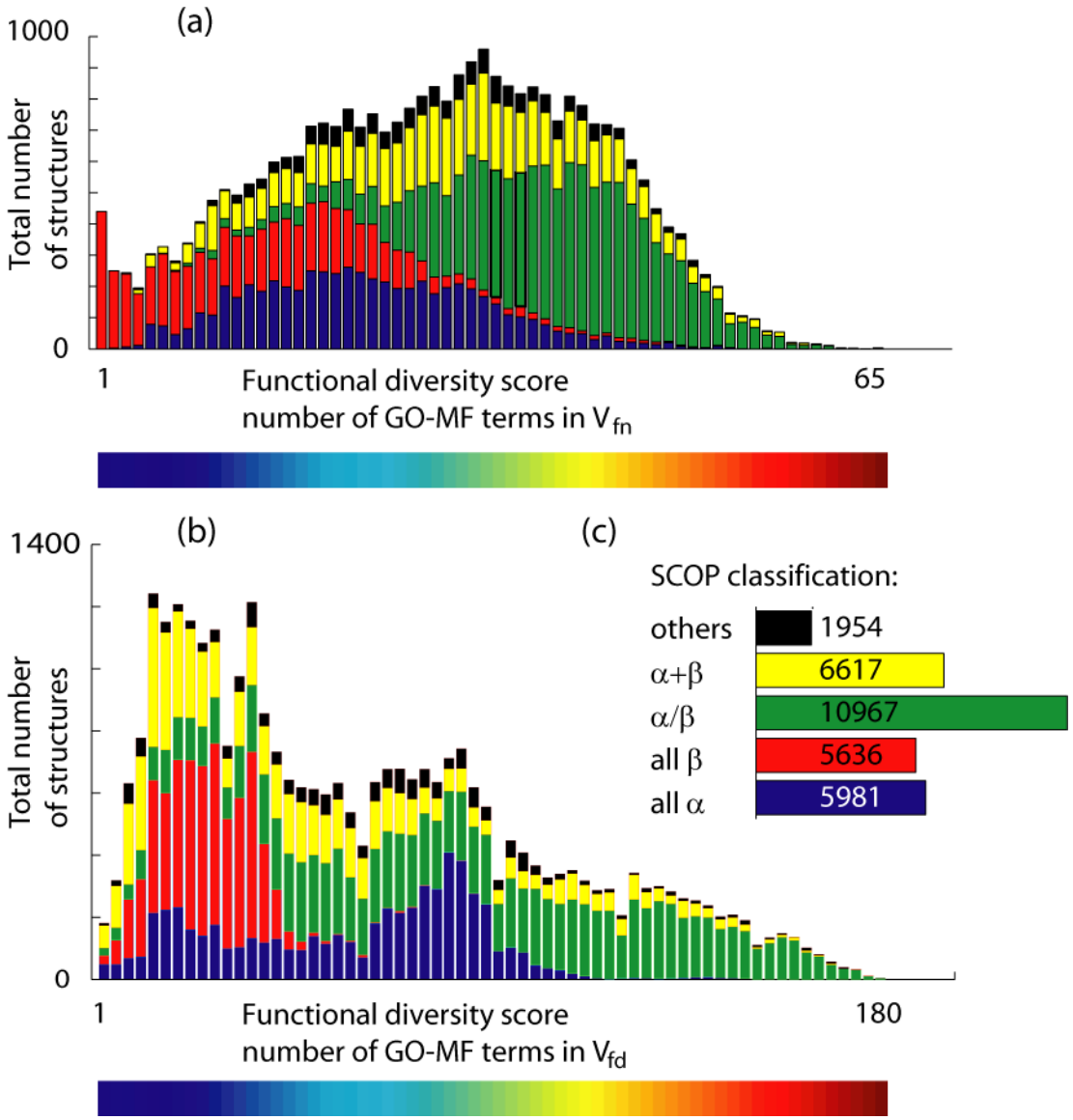

Figure 13S: Functional diversity by SCOP class with vicinity V_{fn} and V_{fd}: We calculate the separate histograms of functional diversity for each of the SCOP classes, and stack them one on top of the other. Table 2S lists the exact proportions of each of the SCOP classes, among the top 10%/20% most dense/functionally diverse domains. This supports Figure 3 in that the most functionally diverse regions are populated by the alpha/beta domains.

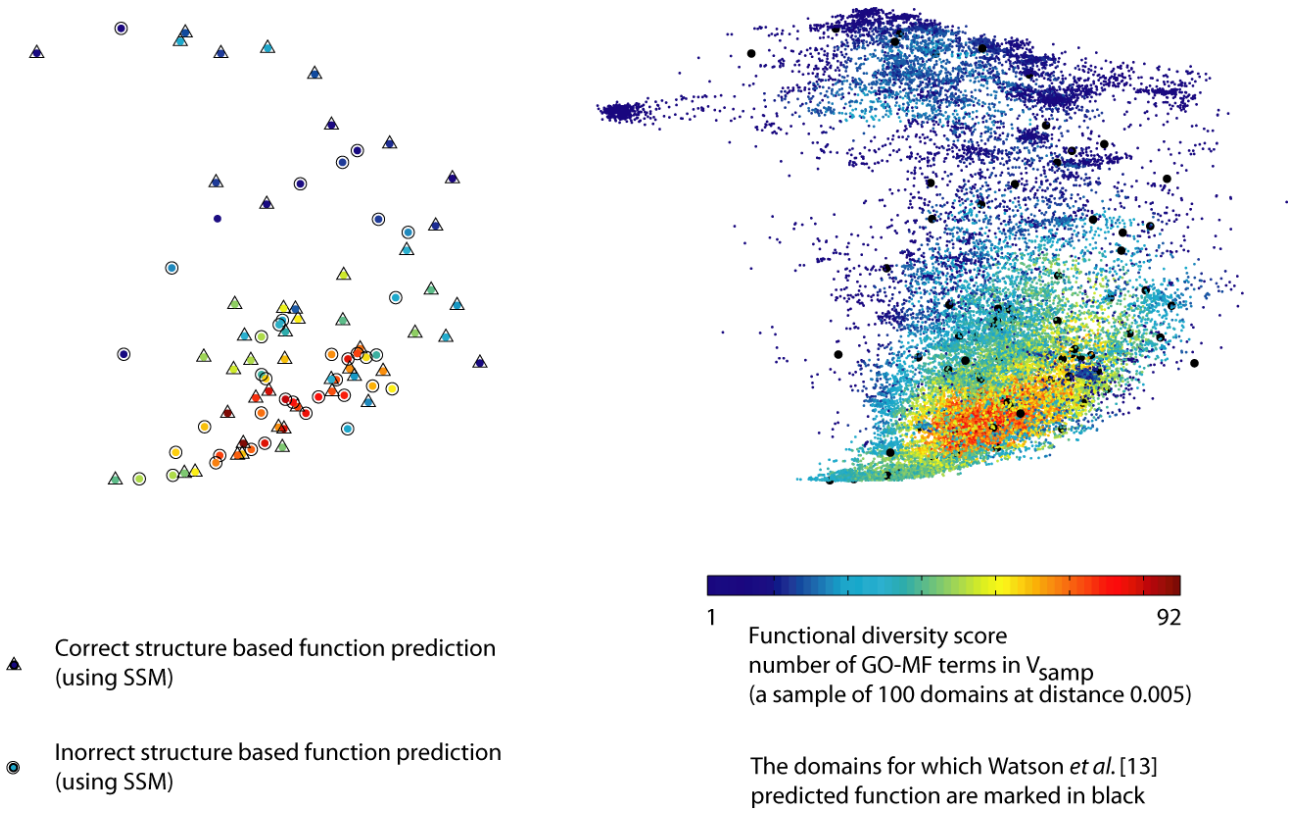

Figure 14S: A map of Watson *et al.* (*1*) "known-function" dataset. The right panel shows the functional diversity of our complete dataset, and the 90 proteins in Watson et al.'s [14] set are shown as black dots. The left panel shows the same structures and their marker depends on if their function was correctly predicted using global structural similarity (triangles), or not (circles). We see that the function of proteins in the periphery of structure space tends to be more accurate than the function of the proteins in the core. This statement is quantified in the Results section.

Table 1S: Pearson correlation coefficients between the functional diversity score[1] on the full dataset and alternative scores/datasets

Alternative set / Vicinity definition	Using only domains annotated with one function	Functional diversity calculate using GO-Slim	Weighted score by 1-10*(fraction of term occurrence)	Weighted score by 1/distance from root
	Figure 6S	Figure 7S	Figure 8S	Figure 9S
V_{samp}	0.8779	0.8952	0.9394	0.9264
V_{fn}	0.8860	0.9299	0.9977	0.9885
V_{fd}	0.9747	0.9567	0.9997	0.9983
Alternative set / Vicinity definition	Coherence as a functional (non) diversity score	Density Function	Sampling using 95% sequence non redundant	Sampling using 40% sequence non redundant
	Figure 10S	Figure 1	Figure 11S	Figure 11S
V_{samp}	-0.7934	0.4306	0.6141	0.6468
V_{fn}	-0.8117	0.0926	0.2969	0.3423
V_{fd}	-0.3030	0.7320	0.8641	0.8881

Table 2S: SCOP class composition of the most functionally diverse domains

Data Set	Vicinity definition	Within top	% all alpha	% all beta	% alpha / beta	%alpha + Beta	% others
Full	V_{samp}	10%	1.99	0.10	78.06	15.19	4.65
		20%	2.93	0.20	76.45	15.55	4.87
Full	V_{fn}	10%	2.89	1.04	72.35	19.62	4.10
		20%	3.56	1.35	70.52	19.84	4.73
Full	V_{fd}	10%	1.52	0.13	81.84	12.15	4.36
		20%	4.18	0.15	74.93	15.73	5.02
NR (95%)	V_{fd}	10%	4.16	0	79.95	10.76	5.13
		20%	19.01	0.12	63.40	11.92	5.55
NR (40%)	V_{fd}	10%	8.54	0	75.61	9.76	6.10
		20%	20.98	0.20	60.49	11.81	6.52

[1]The functional diversity score of a protein domain is the number of distinct GO-MF terms annotating the set of domains that are in the vicinity of that domain.

Table 3S: SCOP folds that lie in the functionally diverse core

V_{fd}				V_{fn}				V_{samp}			
Top 20 means		Top 20 medians		Top 20 means		Top 20 medians		Top 20 means		Top 20 medians	
SCOP Fold	Mean Score	SCOP Fold	Median Score	SCOP Fold	Mean Score	SCOP Fold	Median Score	SCOP Fold	Mean Score	SCOP Fold	Median Score
c.117	152.5	c.117	155.5	c.24	45.2	c.88	46	c.117	73.7	c.117	73
c.42	148.5	c.74	151.5	c.88	44.8	c.24	46	c.42	71.2	c. 74	72.5
c.14	133.8	c.42	151	c.6	43.9	c.74	44	c.74	69.9	c.42	72
c.74	132.8	c.14	150	c.69	43.6	c.69	44	c.67	68.2	c.67	69
c.6	128.9	c.24	135	d.165	42.6	c.6	44	c.24	68.0	c.24	69
c.24	128.1	c.6	133	c.56	42.3	a.137	43.5	c.6	67.4	d.95	68
c.67	121.8	c.93	131	c.66	41.7	c.56	43	c.14	67.1	c.36	68
c.93	118.1	d.95	122	c.74	41.6	d.165	42	c.36	66.6	c.14	68
c.36	110.6	c.67	121	c.117	41.5	c.93	42	d.174	66.1	c.6	68
d.174	107.8	d.96	115.5	c.41	41.5	c.66	42	e.26	65.8	c.69	67
c.1	107.4	c.36	114	c.42	41.3	c.41	42	c.93	65.7	e.26	66
d.95	107.3	c.1	113	c.14	41.2	c.23	42	c.69	65.3	d.174	66
c.60	105.8	d.174	112	c.23	41.1	c.14	42	c.79	64.7	c.93	66
c.79	103.8	c.69	111	c.26	40.9	d.144	41	c.60	64.7	c.56	66
c.69	103.6	c.60	111	c.45	40.7	c.117	41	c.1	64.6	c.1	66
c.56	100.1	c.56	106	a.137	40.7	c.61	41	c.7	64.4	d.144	65
c.7	99.2	c.80	102	c.53	40.4	c.53	41	c.39	64.3	d.96	65
d.96	97.5	c.79	101	c.60	40.4	c.45	41	c.56	63.7	c.79	65
d.144	97.0	c.7	101	d.144	40.3	c.42	41	d.144	63.1	c.60	65
c.39	95.8	d.144	100	c.61	40.2	c.26	41	c. 72	62.4	c.7	65