

SUPPLEMENTARY INFORMATION FOR: THE TOPOLOGICAL PROPERTIES OF THE PROTEIN UNIVERSE

TABLES

Table 1 provides a description for each column in the Postgres database collecting protein metadata.

Supplementary Table 1. Database schema. The description of each column of each table in the database.

Table	Column	Type	Description
AF	acc	string	UniProt accession ID
AF	tax	integer	Taxonomy ID used by AlphaFold2
AF	n	integer	Number of residues
AF	maxrep1	integer	Maximum number of residues in a feature (dim=1)
AF	maxrep2	integer	Maximum number of residues in a feature (dim=2)
AF	maxpers1	float	Maximum persistence of a feature (dim=1)
AF	maxpers2	float	Maximum persistence of a feature (dim=2)
AF	meanplddt	float	pLDDT averaged across α -carbons
AF	nrep1	integer	Number of representatives (dim=1)
AF	nrep2	integer	Number of representatives (dim=2)
AF	nrep1_tT	integer	nrep1 with persistence $> T$ ($T \in [1, 10]$)
AF	nrep2_tT	integer	nrep2 with persistence $> T$ ($T \in [1, 10]/10$)
JSON	acc	string	UniProt accession ID
JSON	path	string	Path in database to JSON file
Tax	acc	string	UniProt accession ID
Tax	tax	integer	Taxonomy ID
TaxTree	tax	integer	Taxonomy ID for species and lower ranks
TaxTree	parent	integer	Taxonomy ID for species
TaxTree	rankp	string	“species”
TaxTree	rank	string	Rank of tax (“species” or lower)
TaxTree	domain	string	Domain of tax (“A”, “B”, “E”, or “V”)
TaxParent	tax	integer	Taxonomy ID
TaxParent	parent	integer	Taxonomy ID for direct or indirect parent of tax
TaxParent	rankp	string	Rank of parent
TaxParent	rank	string	Rank of tax
TreeNode	tax	integer	Taxonomy ID
TreeNode	domain	string	Domain (“A”, “B”, “E”, or “V”)
TreeNode	proteins	integer	Number of proteins
TreeNode	avg_V	float	AF values V averaged across child nodes
TreeNode	avg_V_pp	float	AF values V averaged across proteins
TreeNode	var_V	float	Variance of AF values V averaged across child nodes
TreeNode	var_V_pp	float	Variance of AF values V
TreeEdge	child	integer	Child node taxonomy ID
TreeEdge	parent	integer	Parent node taxonomy ID
TreeEdge	V_f	float	$V_{\text{child}}/V_{\text{parent}}$

Table 2 provides a summary of the datasets discussed in the biological analysis.

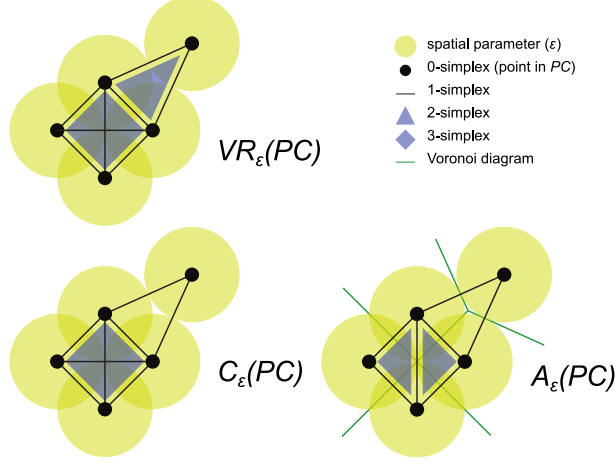
Supplementary Table 2. Dataset overview. A brief summary of the datasets discussed in the biological analysis. We analysed all the AlphaFold2 structures and in the biological comparisons focused on the high confidence structures. More detailed discussions of each data set are given in the text below.

Dataset	AlphaFold2	High confidence
RCSB	2712	2637
M-CSA	866	862
CATH	73749	62861
Thermo-/mesophiles	1815	1656
Mutations (disease/neutral)	471/855	419/681

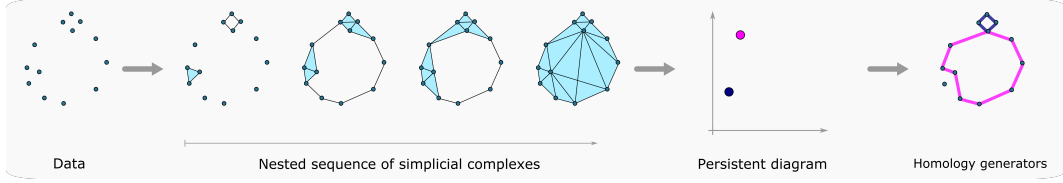
Supplementary Table 3. Significance tests for differences between meso- and thermophile distribution shapes. Two-sided two-sample Kolmogorov–Smirnov tests were performed comparing distributions of TIFs between meso- and thermophiles for each amino acid. AA: Amino acids sorted from lowest to highest volume. n_{meso} : number of observations from mesophiles. n_{thermo} : number of observations from thermophiles. D : test statistic (largest difference in cumulative distributions). Adjusted p -value: adjusted by false discovery correction for multiple testing, individually for the two sets of tests.

AA	n_{meso}	n_{thermo}	D	p -value	Adjusted p -value
G	46160	8910	0.039	2.035×10^{-10}	4.523×10^{-10}
A	49753	8993	0.056	1.690×10^{-21}	6.759×10^{-21}
S	32915	6164	0.073	1.047×10^{-24}	6.979×10^{-24}
C	5508	642	0.073	4.454×10^{-03}	4.688×10^{-03}
D	35764	6601	0.032	1.970×10^{-05}	2.815×10^{-05}
P	23845	5375	0.075	4.713×10^{-22}	2.357×10^{-21}
N	28171	5121	0.061	1.328×10^{-14}	4.428×10^{-14}
T	30573	5764	0.057	4.493×10^{-14}	1.284×10^{-13}
E	33253	7627	0.074	4.523×10^{-30}	9.047×10^{-29}
V	34965	7986	0.046	1.462×10^{-12}	3.655×10^{-12}
Q	23701	3404	0.033	2.920×10^{-03}	3.244×10^{-03}
H	11417	2550	0.072	7.381×10^{-10}	1.476×10^{-09}
M	11812	2607	0.067	9.618×10^{-09}	1.749×10^{-08}
L	45969	9325	0.022	1.476×10^{-03}	1.736×10^{-03}
I	32161	6679	0.028	2.634×10^{-04}	3.512×10^{-04}
K	30712	6478	0.023	7.538×10^{-03}	7.538×10^{-03}
R	23862	5714	0.040	6.933×10^{-07}	1.067×10^{-06}
F	22215	4976	0.088	1.576×10^{-27}	1.576×10^{-26}
Y	24260	4953	0.031	8.389×10^{-04}	1.049×10^{-03}
W	11281	2267	0.063	6.658×10^{-07}	1.067×10^{-06}

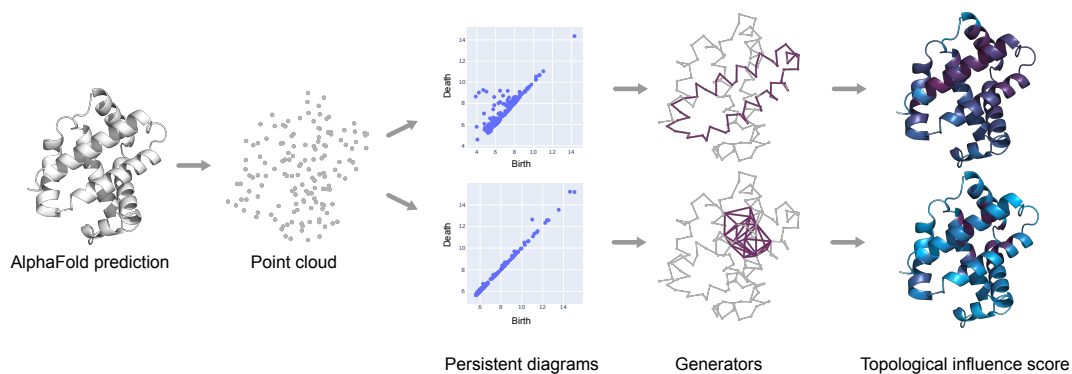
SUPPLEMENTARY FIGURES



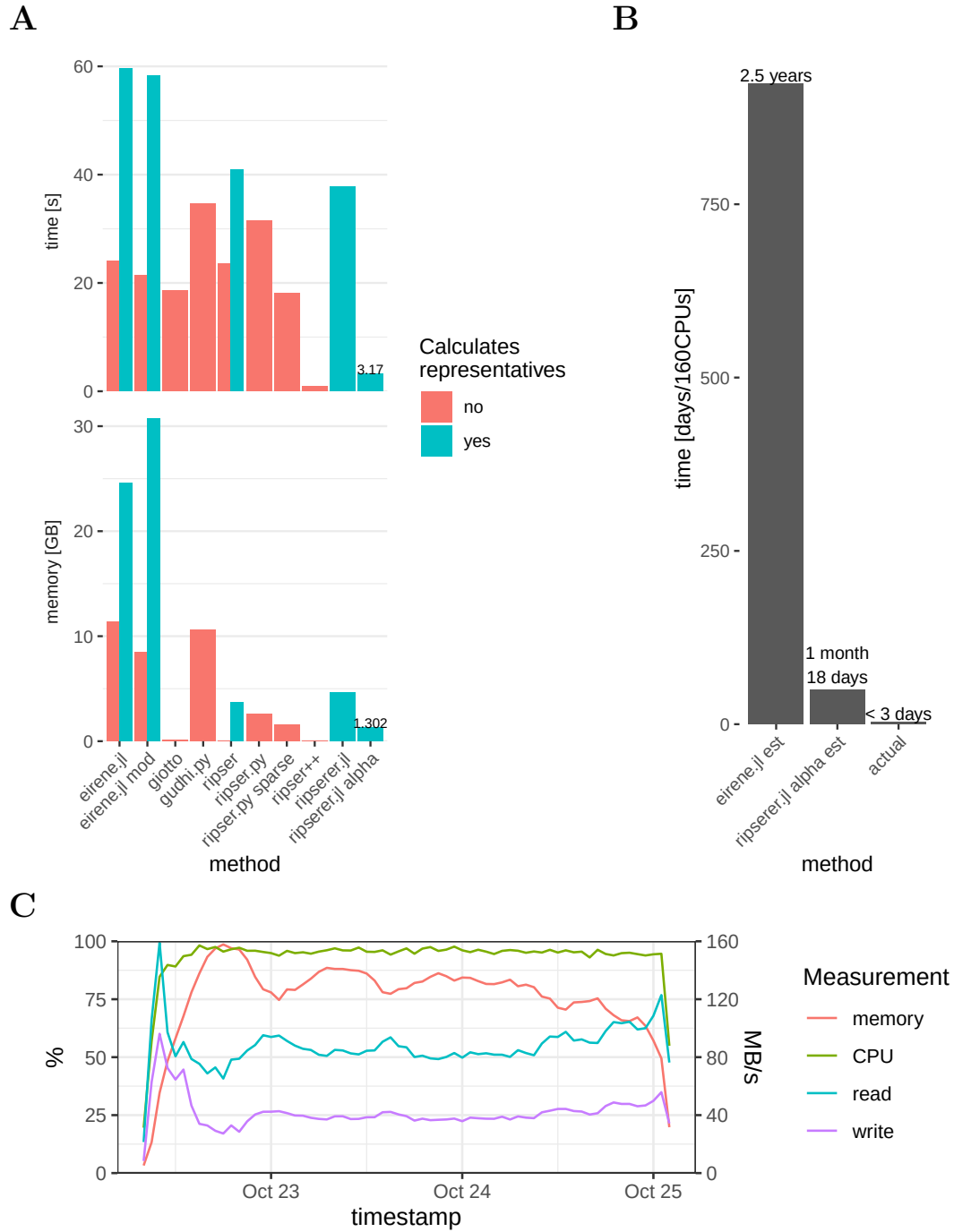
Supplementary Figure 1. An example of Vietoris-Rips complex, Čech complex and Alpha complex constructed on the same point cloud, for the same parameter ε .



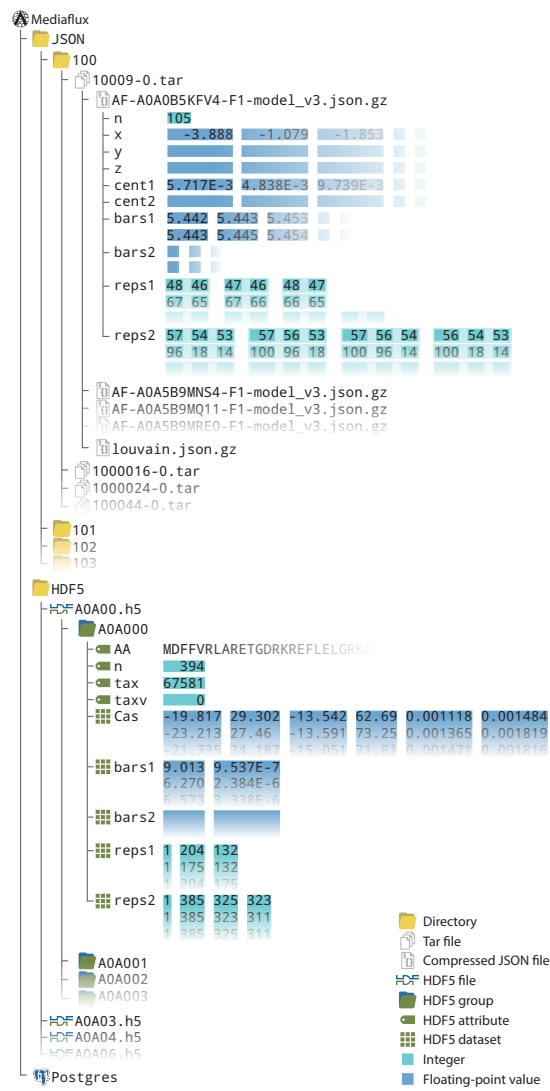
Supplementary Figure 2. In the standard persistent homology pipeline, the input is a point cloud $PC = \{p_1, \dots, p_n\} \subset \mathbb{R}^n$. The first step in the algorithm is to construct a nested sequence of simplicial complexes $PC_{\varepsilon_0} \hookrightarrow PC_{\varepsilon_1} \hookrightarrow \dots \hookrightarrow PC_{\varepsilon_N}$ that approximate the data at different spatial scales. This depends on a choice of *filtration*; common choices are the Vietoris-Rips, Čech and Alpha complexes. The figure shows a schematic of the Vietoris-Rips filtration built on the initial point cloud. The simplicial complexes are then analysed by computing their simplicial homology and corresponding persistent diagrams. The figure shows the output for 1-dimensional features; a total of two loops appear in the filtration. The fact that one is more persistent than the other reveals the *circular* structure of the data. For each homology class in a persistent diagram, it is possible to compute representative cycles. The figure shows a possible choice of generator for each 1-dimensional feature in the persistent diagram.



Supplementary Figure 3. From a predicted protein structure, we extract the point cloud $PC = \{p_1, \dots, p_n\}$ given by xyz -coordinates of its α -carbon atoms. We then compute the corresponding persistent diagrams in dimension 1 and 2, and a generator for each topological feature. For each dimension considered, the information consisting of persistent diagram, generators, and their persistence is used to compute the point-wise topological influence score (TIF), which ranks residues based on how often they contribute to topological features, and how important these are. Source data are provided as a Source Data file in `Source_data/source_data.FigureS3.json`.



Supplementary Figure 4. Performance benchmarks. Performance comparison for a number of tested persistent homology tools (A). Bar fill indicates if representative cycles were included in the output. All benchmarks were on a single large structure with 600 residues. Actual computation on all 214M structures compared to estimates based on the benchmark of the initial method and the chosen method (B). Percentage use of the 1 TB memory and 160 CPUs, as well as disk read/write in MB/s during the 3 day computation (C). Source data are provided as a Source Data file in `Source_data/source_data.FigureS4AB.tsv` and `Source_data/source_data.FigureS4BC.tsv`.



Supplementary Figure 5. Data organisation on Mediaflux. One example protein is shown for JSON format as well as HDF5.

AF

214,175,790 × 30

acc	taxon	n	maxrep1	maxrep2	maxpers1	maxpers2	meanpiddt	nrep1	nrep2	nrep1_t1	nrep1_t2
A0A4Y2VRE5	182803	143	74	36	4.869422	2.108323	80.50973	310	189	27	4
A0A4Y2VRE6	182803	200	24	46	3.9872656	1.7559824	77.5748	390	283	19	8
A0A4Y2VRE9	182803	104	5	6	0.67140913	0.15324402	88.001144	198	167	0	0

JSON

214,175,790 × 2

acc	path
A0A3S0EAG4	PH/100/100-0/AF-A0A3S0EAG4-F1-model_v3.json.gz
A0A3S0EAU0	PH/100/100-0/AF-A0A3S0EAU0-F1-model_v3.json.gz
A0A3S0EB17	PH/100/100-0/AF-A0A3S0EB17-F1-model_v3.json.gz

Tax

888,163,507 × 2

acc	tax
A0A009IHW8	1310613
A0A011QK89	522306
A0A017SE81	1388766

TaxTree

4,367,074 × 5

tax	parent	rankp	rank	domain
3040113	3040113	species	species	V
3040114	3040114	species	species	V
3040115	3040115	species	species	V

TaxParent

47,595,713 × 4

tax	parent	rankp	rank
1211725	2626993	no rank	species
1211726	264000	genus	species
1211727	2646371	no rank	species

TreeNode

879,207 × 40

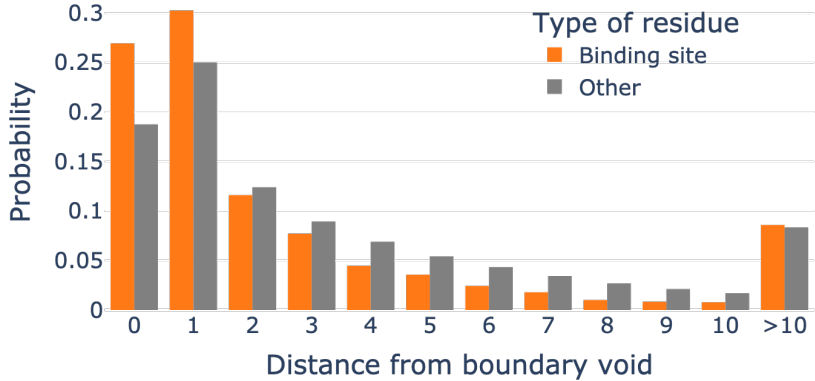
tax	domain	proteins	avg_n	avg_maxrep1	avg_maxrep2	avg_maxpers1	avg_maxpers2	avg_nrep1	avg_nrep2	var_n
1955042	B	2758	294.741	109.33828861	38.094996374	6.612734582391	1.85362138966	711.630529	463.369108	39083
1956187	E	14966	365.271	118.60216784	38.791365116	10.6211077741	1.84818517341	822.681202	531.995778	61255
1961800	E	11	322.151	98.7333333333	52.7555555555	6.632029215494	2.00836224026	699.222222	383.955555	837.65

TreeEdge

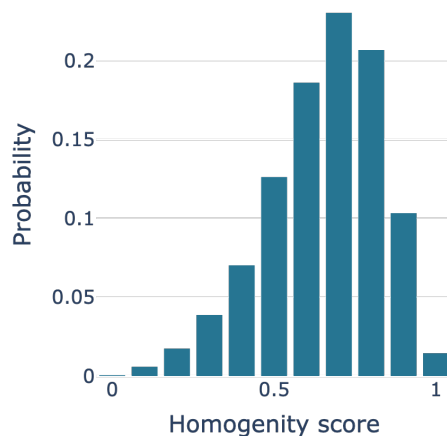
871,760 × 17

child	parent	proteins_f	avg_n_f	avg_maxpers1_f	avg_maxpers2_f	avg_nrep1_f	avg_nrep2_f	var_n_f
2039065	30080	0.000431220	0.48383411	0.205448916060810	0.070549373925334	0.29337834246	0.35278577095	0.0000000
2039066	30080	0.001293661	1.00488621	1.052807995476650	0.968394986139928	0.96395741096	0.93714374028	2.0785836
2039094	30080	0.000431220	1.46390831	1.3356359567923	2.093408533313980	1.64743223078	1.487127711414	0.0000000

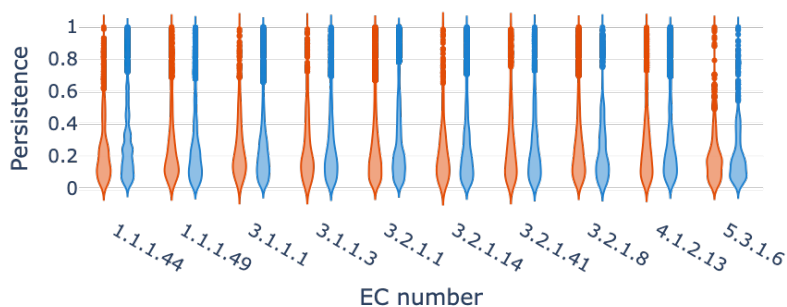
Supplementary Figure 6. Example entries from database tables. The first three rows to exemplify the data contained in the protein metadata. Table sizes indicated with rows × columns notation.



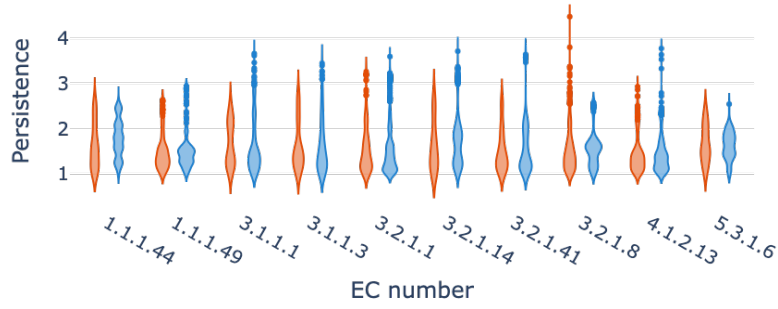
Supplementary Figure 7. The bar plot shows the distribution of distances (in number of residues) of residues between binding sites from boundaries of 2-dimensional topological clusters in experimentally solved structures. These boundary points are enriched for binding sites from the Mechanism and Catalytic Site Atlas (M-CSA) dataset. Source data are provided as a Source Data file in Source_data/source_data_FigureS7.tsv.



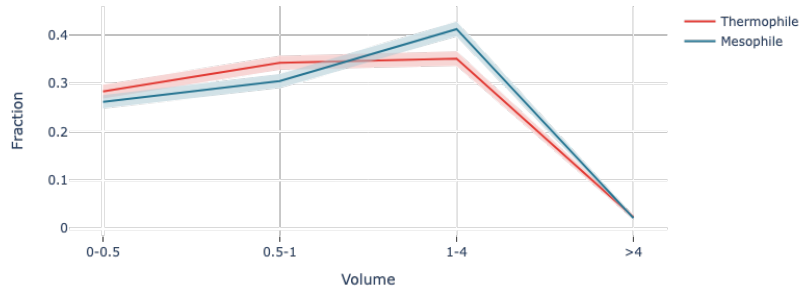
Supplementary Figure 8. The bar plot shows the distribution of homogeneity scores for all the 62,861 high-confidence AlphaFold2 predictions with identified CATH domains. Source data are provided as a Source Data file in `Source_data/source_data.Figure3C-S8.tsv`.



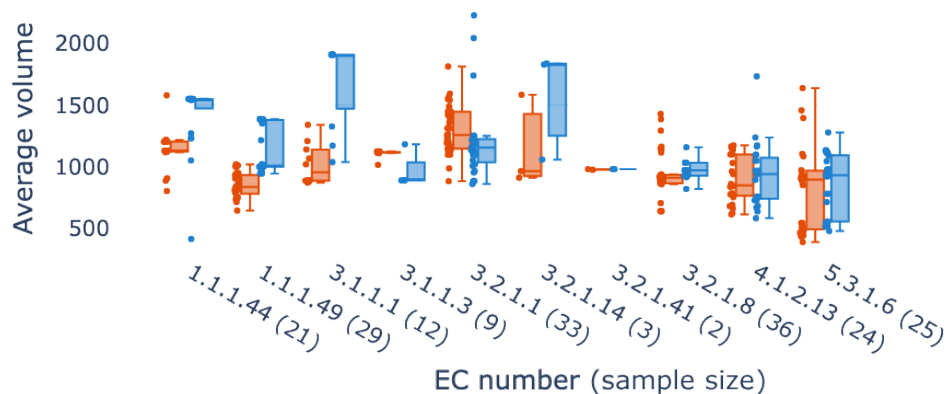
Supplementary Figure 9. The distribution of persistence values of 2-dimensional topological features across the different EC numbers, for low-persistence (< 1) features. Thermophiles are in red and mesophiles in blue. Source data are provided as a Source Data file in `Source_data/source_data.FigureS9-S10-S12.tsv`.



Supplementary Figure 10. The distribution of persistence values of 2-dimensional topological features across the different EC numbers, for high-persistence (≥ 1) features. Thermophiles are in red and mesophiles in blue. Source data are provided as a Source Data file in `Source_data/source_data.FigureS9-S10-S12.tsv`.

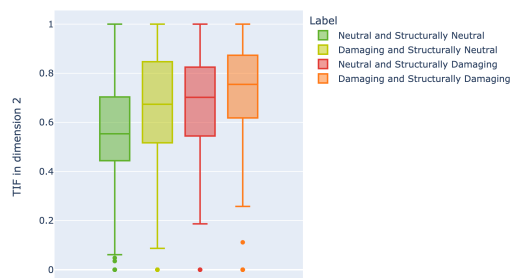


Supplementary Figure 11. The distribution of volumes of 2-dimensional topological features, for high-persistence (≥ 1) features. Error bands are given by standard deviation. Thermophiles are in red and mesophiles in blue. Source data are provided as a Source Data file in `Source_data/source_data.FigureS11.tsv`.

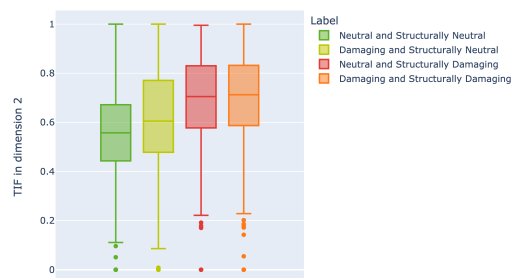


Supplementary Figure 12. The distribution of average volumes of 2-dimensional topological features across the different EC numbers, for high-persistence (≥ 1) features. The pattern is maintained across all but two groups. Thermophiles are in red and mesophiles in blue. Sample sizes in parentheses are numbers of proteins for thermophiles plus mesophiles. Source data are provided as a Source Data file in `Source_data/source_data_FigureS9-S10-S12.tsv`.

A Main dataset



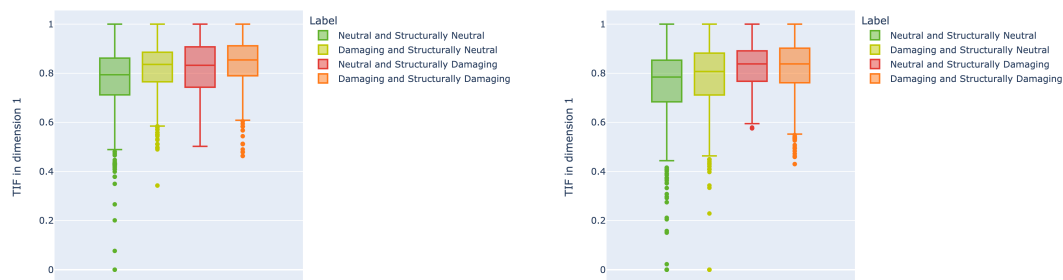
B Control group



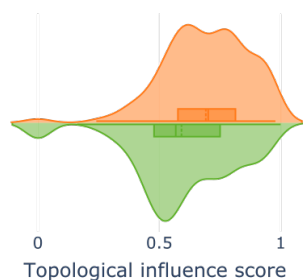
Supplementary Figure 13. The distribution of 2-dimensional TIFs in residues accommodating missense variation considered in the analysis for the main dataset (A) and for the control group (B). Source data are provided as a Source Data file in `Source_data/source_data_Figure5B-S13-S14.tsv`.

A Main dataset

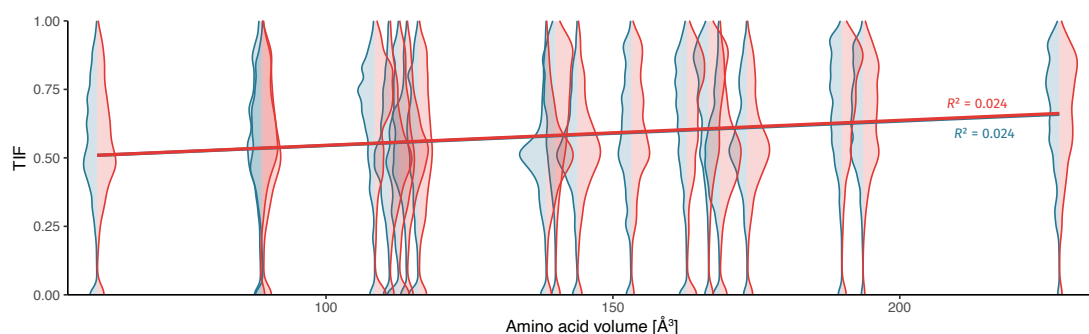
B Control group



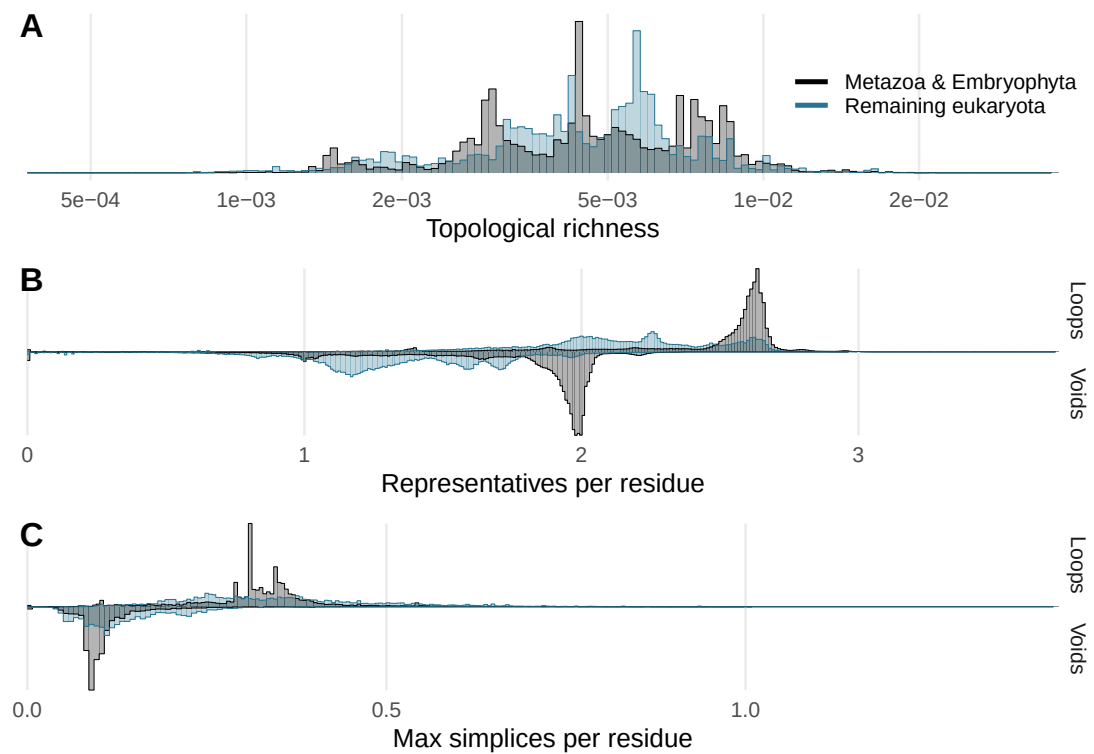
Supplementary Figure 14. The distribution of 2-dimensional TIFs in residues accommodating missense variation considered in the analysis for the main dataset (A) and for the control group (B). Source data are provided as a Source Data file in `Source_data/source_data.Figure5B-S13-S14.tsv`.



Supplementary Figure 15. The distribution of 2-dimensional TIFs for residues in human CFTR that are predicted to accommodate structural damaging mutations. Source data are provided as a Source Data file in `Source_data/source_data.FigureS15.tsv`.



Supplementary Figure 16. Amino acid volumes in the topological voids for meso- and thermophiles. Here, TIF is the topological influence score from two-dimensional homology as indicator of the void occupancy for a given amino acid. This is contrasted with the amino acid volumes [1]. Violin plots are shown in place of a regular scatter plot for clarity. Violin plots correspond to the violins in Fig. 4 of the main text, except each violin is scaled locally as individual densities in the main text, and here as two global densities. A simple linear trendline shows no detectable difference in linear relationships for mesophiles and thermophiles. Source data are provided as a Source Data file in `Source_data/source_data.Figure4C-S16.tsv`.



Supplementary Figure 17. Histogram computed on proteins as in Fig 2 of the main text, here shown for a grouping with multicellular species from Metazoa and Embryophyta, and a second grouping of remaining eukaryotic proteins. (A) shows topological richness on a log-scale for numbers of excluded zero-valued entries, (B) shows the number of topological representatives – either “loops” or “voids” – normalized by protein length, and (C) shows the size of the single largest loop and void in terms of the number of simplices normalized by protein length. Source data are provided as a Source Data file in `Source_data/source_data.FigureS17.tsv`.

SUPPLEMENTARY REFERENCES

- [1] Zamyatnin, Alexander A. *Protein volume in solution*. Progress in biophysics and molecular biology 24 (1972): 107-123.