Supporting Information

Conservation of Potentially Druggable Cavities in Intrinsically Disordered Proteins

Bin Chong,[†] Maodong Li,[‡] Tong Li,[§] Miao Yu,[†] Yugang Zhang,[∥] and Zhirong Liu[*,†,‡,⊥]

[†]College of Chemistry and Molecular Engineering, [‡]Center for Quantitative Biology, and [⊥]Beijing National Laboratory for Molecular Sciences (BNLMS), Peking University, Beijing 100871, China

[§]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[∥]Department of Chemistry and Chemical Biology, Cornell University, New York 14850, USA

[*] Corresponding author. E-mail address: LiuZhiRong@pku.edu.cn

Table S1. Properties of the discarded IDPs

pE-DB ID	Name	Method	Length	Conf. number	Cavity number (total)	Druggable cavity number
1AAB	Heat shock protein beta-6 (HSPB6) fragment (57–160) V67G mutant	SAXS	208	5	36	0
3AAD	Mengovirus Leader Protein	NMR	71	10	30	0
4AAD	Mengovirus Leader Protein	NMR	71	10	34	0
5AAD	Mengovirus Leader Protein	NMR	71	10	24	0
6AAD	Mengovirus Leader Protein	NMR	71	10	35	2
9AAA	Sic1	SAXS & NMR	92	44	162	0

Table S2. Result of classification for examined single-chain IDPs

pE-DB ID	Average cavity number	Druggable cavity number	δ Range	Number of group	δ Range of I	δ Range of II	δ Range of III
1AAA	3.5	8	41~77	1	41~77	-	-
1AAD	5.20	431	13~118	3	13~47	48~83	84~118
2AAA	4.03	6	44~51	1	44~51	-	-
2AAD	6.41	511	21~116	3	21~52	53~84	85~116
4AAB	2.63	1300	12~39	2			
5AAA	2.52	20	12~41	2	12~29	31~41	-
6AAA	4.90	1967	9~98	1	9~38	39~68	69~98
6AAC	3.22	9	30~83	1	30~83	-	-
7AAC	3.27	46	33~97	3	33~55	56~78	79~97
8AAC	3.21	64	13~46	3	13~23	24~33	34~46
9AAC	5.78	400	11~126	3	11~49	50~88	89~126
n.a.	1.12	47	17~25	1	17~25	-	-

Table S3. Result of classification for examined multi-chain IDPs

pE-DB ID	Average cavity number	Druggable cavity number	δ Range	Number of group	δ Range of I	δ Range of II	δ Range of III
2AAB	9.38	7	43~105	1	43~105	-	-
3AAA	24.18	11	277~542	2	277~366	429~542	-
3AAB	18.5	15	88~115	2	88~98	110~115	-
4AAA	24.12	13	248~630	2	248~378	404~630	-
5AAC	19.91	71	43~499	3	43~183	231~407	411~499
7AAA	10.17	18	21~116	2	21~82	83~116	-
8AAA	17.33	8	107~117	1	107~117	-	-

Table S4. Information of the ordered proteins dataset

PDB ID	ResidueNum of potentially druggable cavity	Surface area ($Å^2$)	Volume ($Å^3$)
1A4J	76.33	1624.083	3483.625
1BRQ	46	406.5	408.875
1BYA	75	1046.75	1692.375
1CHG	34	444.5	634.875
1HSI	39	1215.5	2103.625
1IFB	52	570.75	579.375
1PHC	74	891	894.375
1PSN	65	986.5	1660.25
1PTS	34	347.5	396.75

1QIF	60	766	980.75
2CTB	32.5	460.75	571
2RTA	37	347.75	397.375
3APP	85	1247.75	2687.625
3LCK	72.5	1040.125	2280.688
3P2P	31	407.75	389
5CPA	24	455	555.375

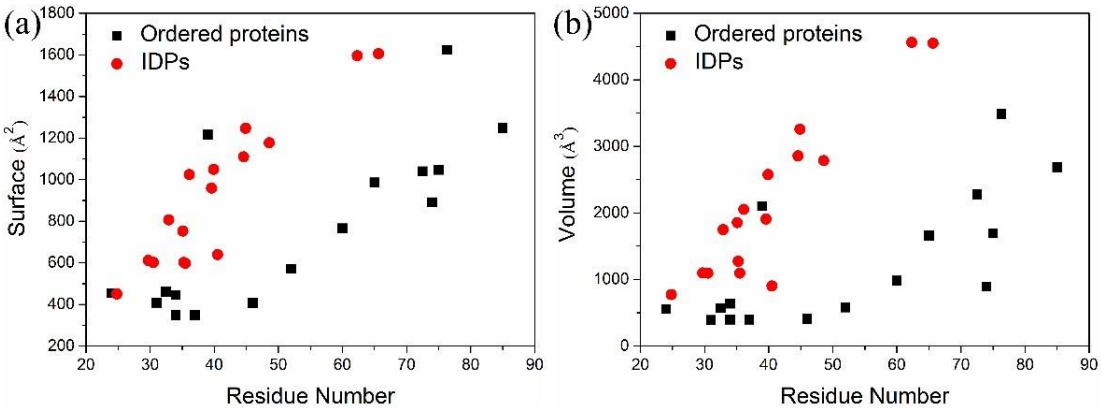

Figure S1. The average surface and volume of druggable cavities in comparison between ordered proteins and IDPs (single-chain IDPs from pE-DB and Disprot-pdb), as a function of the average druggable-cavity residue number of each protein.

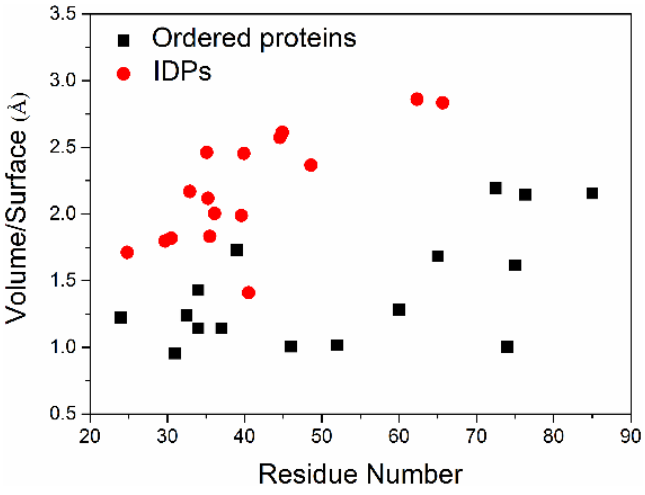

Figure S2. The ratio between the average volume and surface of druggable cavities as a function of the average druggable-cavity residue number in each protein.

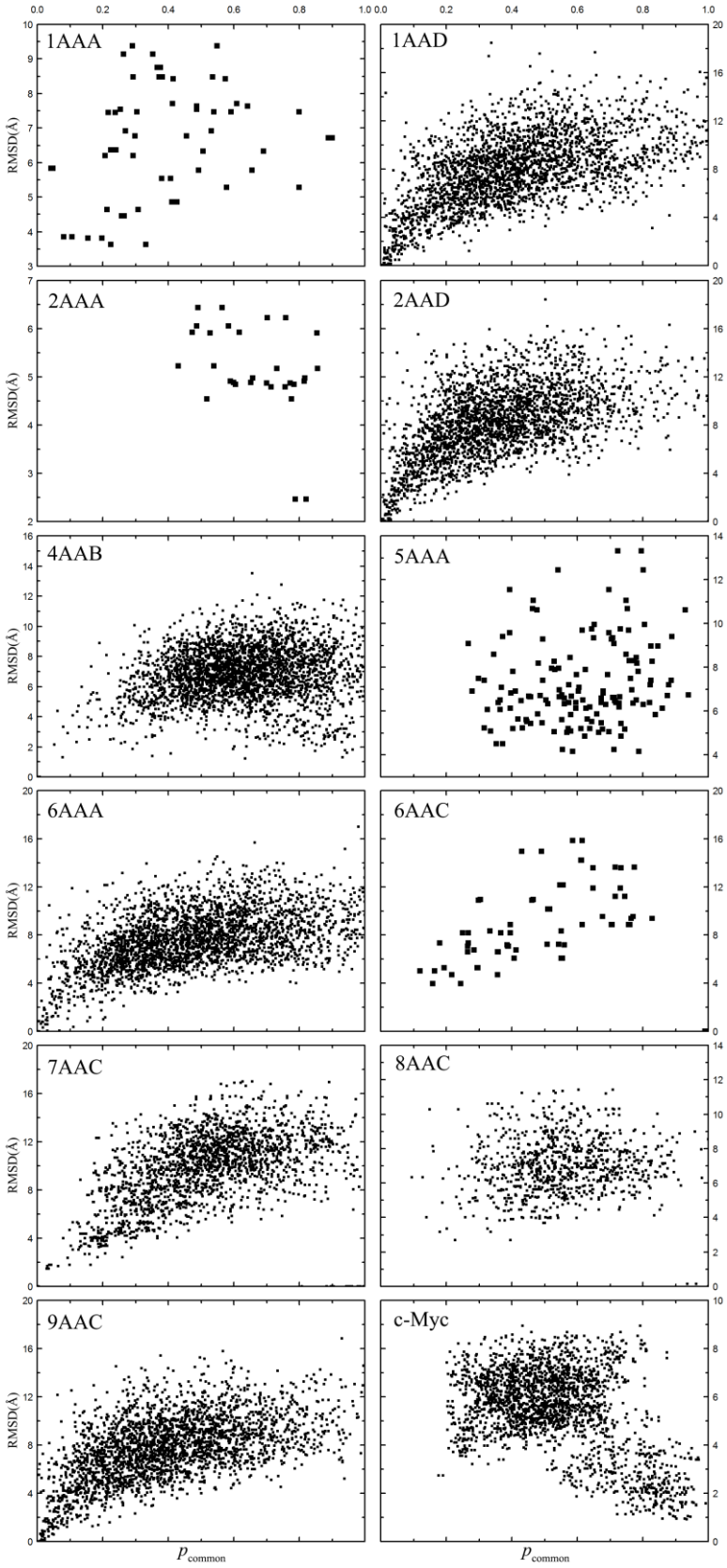

Figure S3. The graph is plotted by percentage on the horizontal axis and corresponding RMSD on the vertical. Some values were omitted to avoid crowding. Better conservation is observed when data points are clustered in the bottom right corner.

Surface area/Volume and pK_d (Multi-chain IDPs)

We first extracted the cavity information of surface area, volume and pK_d values from the cavity file. The statistical result is shown in Fig. S4.

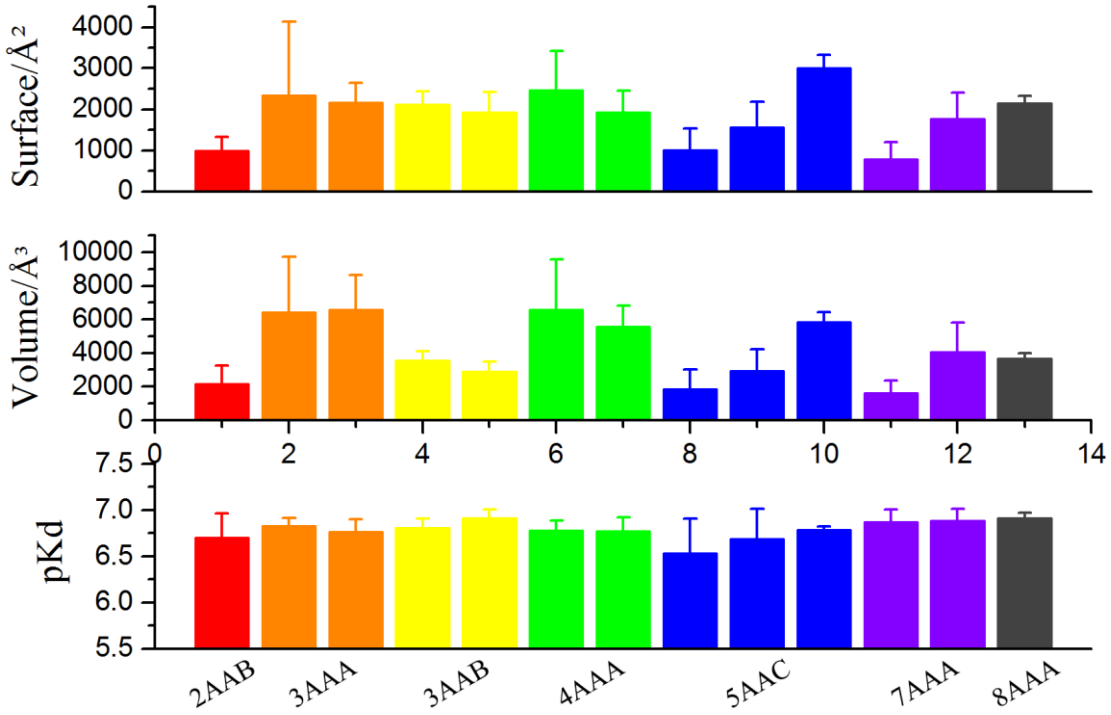

Figure S4. Histograms illustrate the properties (surface area, volume and pK_d) of potentially druggable cavities from different ensembles. Cavities in each ensemble were divided into 1–3 groups as described in Table SIII.

It can be seen from Fig. S4 that there are significant differences in the conservativeness of different systems according to the error bars. Interestingly, for 5AAC, the conservativeness of the third group looks better than the first two groups, which however needs further analysis.

Figure Factor (shape parameter of a cavity)

The figure factor and maximum depth of the cavities for oligomeric IDPs were also calculated and analyzed. The results are shown in Fig. S5. The results of Fig. S5 show that the standard error deviation of 3AAB and 8AAA is relatively small and their conservativeness are good. As for the 5AAC system, we can see that the standard deviation of the third group is smaller than the first two groups, which also proved

that the last part of the cavities for 5AAC is conserved better.

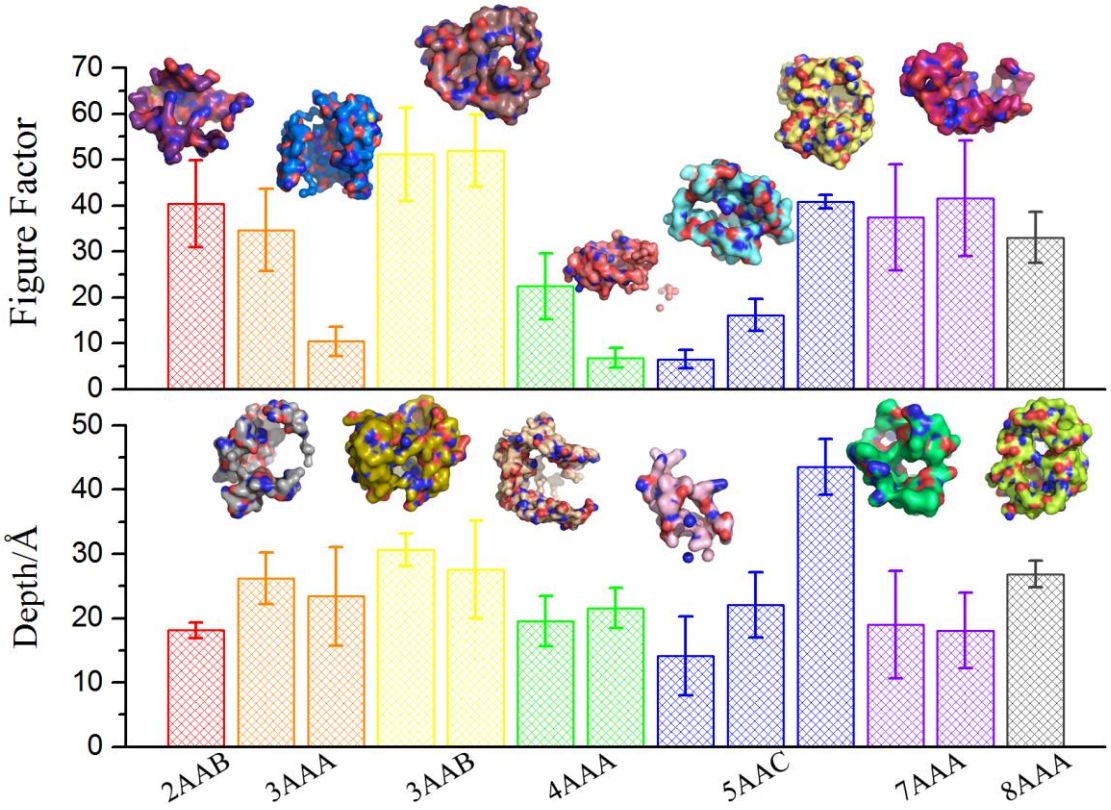

Figure S5. Histograms show the average figure factor and maximum depth of potentially druggable cavities in different ensembles. Error bars represent the standard deviations. Representative cavities with figure factor and depth values similar to the average values in each ensemble are chosen to display in graphics.

Common atom percentage (composition parameter of a cavity)

Since oligomeric proteins are composed of multiple chains, we need to ensure that both chain and atom simultaneously correspond to each other when looking for the same atoms between the two cavities. So the proportion of the same atoms may be reduced and the results of the calculation are shown in Fig. S6.

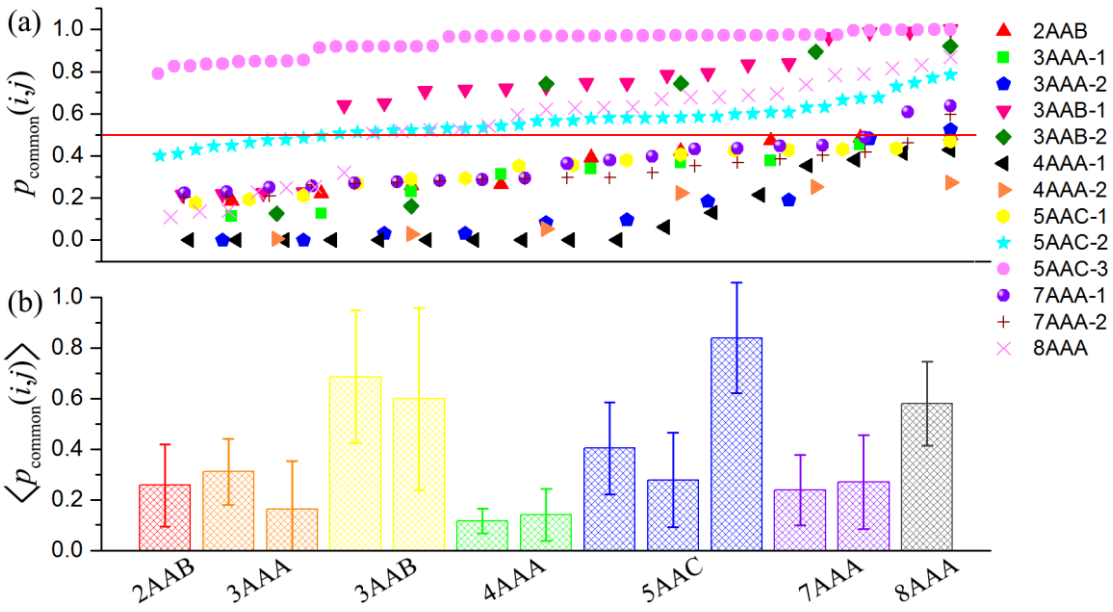

Figure S6. The common-atom percentage of potentially druggable cavities in different ensembles. (a) $p_{common}(i,j)$ plotted in ascending order for each ensemble. The data points are evenly spaced within each ensemble to span the full horizontal range. Some values were omitted to avoid crowding. Red lines indicate the 50% level. (b) The average of $p_{common}(i,j)$ in different ensembles. Error bars represent the standard deviations.

As shown in Fig. S6, oligomeric proteins 3AAA and 4AAA are not as conserved, whereas 3AAB, 8AAA and the latter half of 5AAC are more conserved. In general, oligomeric proteins are more conserved compared with that of single chains. That is because the multi-chain protein should be more conformationally stable than single chain protein.

RMSD (parameter of conformation change of a cavity)

Since the number of potentially druggable cavities for oligomeric proteins is too small, a statistical histogram analysis is not suitable because of insufficient data points. Fig. S7 is composed of the common atom percentage as the abscissa and the corresponding RMSD as the ordinate.

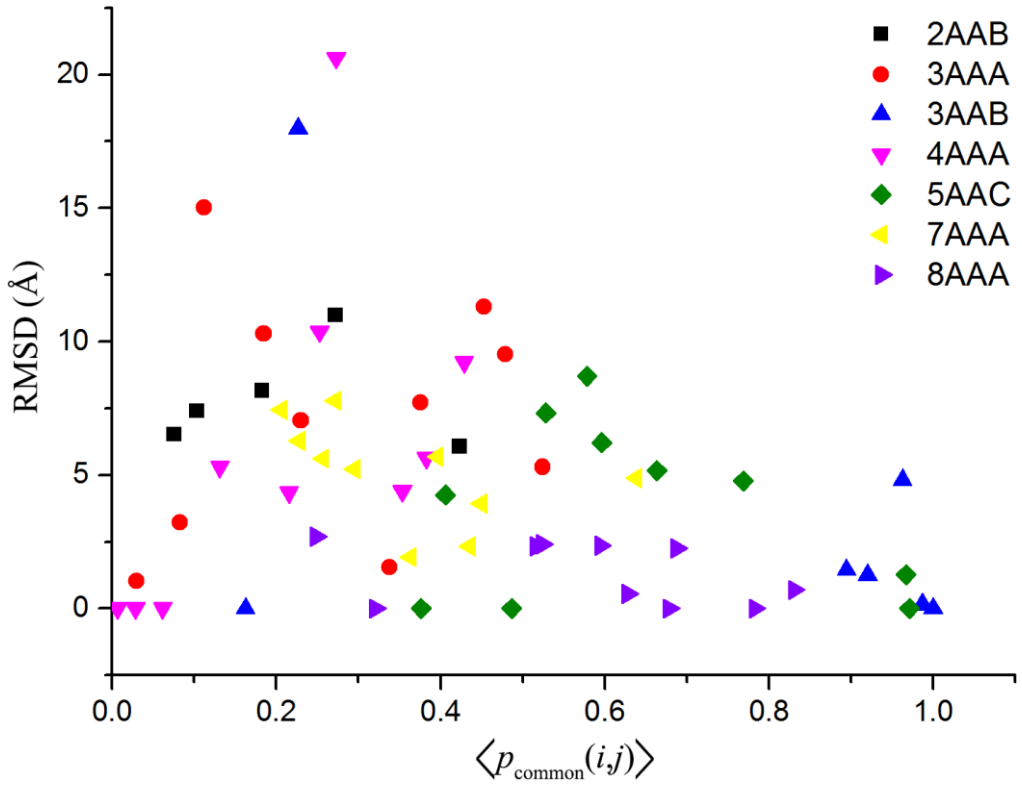

Figure S7. The graph is plotted by percentage on the horizontal axis and corresponding RMSD on the vertical. Conservation is higher when data points are located in the bottom right corner.

The results of Fig. S7 show that conservation of the different ensembles is quite different. The data points representing 2AAB, 3AAA, 4AAA concentrate to the left of the graph, whereas data for 3AAB, 5AAC and 8AAA are located to the bottom right and data for 7AAA are centrally located. Data that is located on the right side of the graph shows higher conservation of the ensemble, whereas data that is located on the left side of the graph indicates poorer conservation of the ensemble.

Since the number of potentially druggable cavities in the 5AAC is large, the statistics of the common atom percentage and RMSD of 5AAC are shown in Fig. S8.

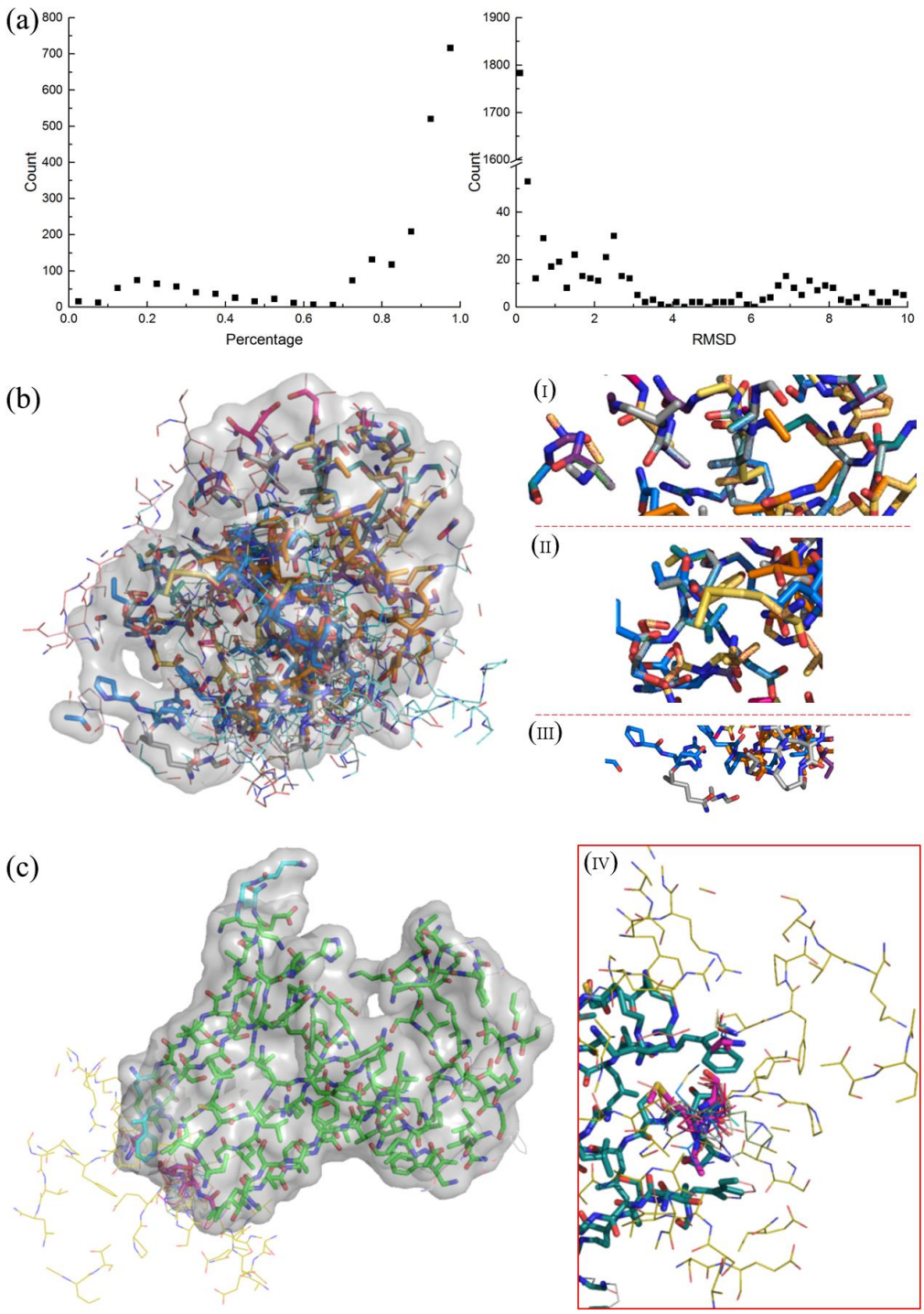

Figure S8. 5AAC as an example where (a) the common atom percentage and RMSD are statistically analyzed, (b-c) the potentially druggable cavities of different parts of 5AAC are aligned (same method). Insets are magnified sections.

The statistical results for 5AAC shown in Fig. S8(a) deviate considerably from a Gaussian distribution. The graphs show that most of the data points concentrate on the

area of the common atom percentage > 90%, RMSD < 1. The results show that the conservation of 5AAC is best, which may be related to the conservatism of the 5AAC ensemble and the method of obtaining the conformation. The second and third parts of the 5AAC ensemble are also aligned and presented using PyMOL (Fig. S8(b-c))

Fig. S8(b) is the aligned image of the cavities in the second part of the 5AAC, and the insets (I, II and III) show details. Insets (I) and (II) show that many of the molecular structures overlap with each other, indicating good conservation. In contrast, inset (III) illustrates that some structures are very different. These observations show that the cavities for the second part of 5AAC appear better conserved in some locations and poorer in other locations. Thus, for a protein as a whole, the conservatism of different parts will be different, which also verifies that it is necessary to divide the protein into 1–3 parts.

Fig. S8(c) is the aligned image of the cavities in the third part of 5AAC. Common atom parts are represented by green sticks and parts that are not exactly the same are represented by other colors. Completely different atom parts are represented by lines. The inset (IV) is a magnified image of the completely different parts. As can be seen from Fig. S8(c), because the third part of the 5AAC ensemble is more conserved, the third part of the potentially druggable cavity is almost identical and indicates that this part has high potential for drug design.

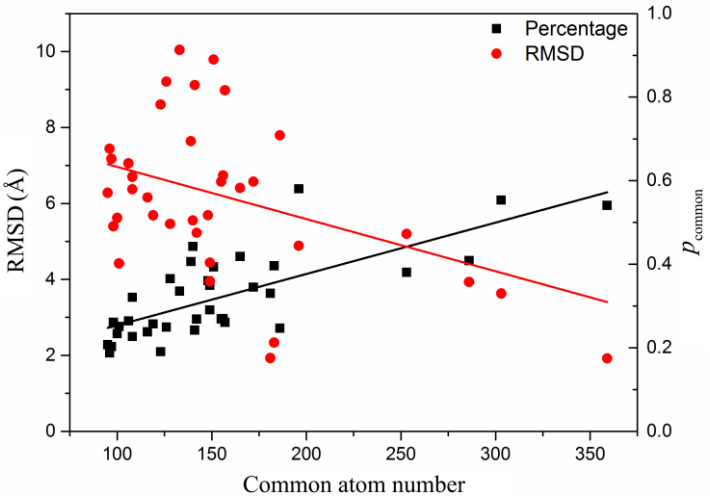

Figure S9. The graph is plotted by the number of common atoms on the horizontal axis and corresponding p_{common} and RMSD on the vertical axis for 7AAA. RMSD is observed to drop

unexpectedly with increasing common atom number and p_{common}.

The statistical analysis about sample size

In statistics, the influence of sample size is well understood (e.g., refer to https://en.wikipedia.org/wiki/Sample_size_determination). In our case, the database pE-DB (with 3 to 13718 conformations for each protein entry) can be regarded to contain samples from the true conformation ensembles (with a vast number of conformations). Statistically, the reliability of a sample with a selection of replicates, i.e., to what extent it can represent the total population, is less affected by the size of the total population. This is actually in conflict with the common sense since one may expect the required sample size to be proportional to the population size. Take the mean (average value) of an inspected property as an example, when estimating the population mean using an independent sample of size n, where each data value has variance σ^2, the standard error of the sample mean is:

$$\frac{\sigma}{\sqrt{n}} \tag{S1}$$

which is independent with the population size N when N is very large. This expression describes quantitatively how precise the estimate is under the adopted sample size. Using the central limit theorem to justify approximating the sample mean with a normal distribution yields an approximate 95% confidence interval of the form

$$\left(\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right) \tag{S2}$$

The required n is much smaller than what was usually thought.

REFERENCE

1. Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An Energy-based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics* 2005, 21, 1908–1916.