CASTpFold: Computed Atlas of Surface Topography of the universe of protein Folds

Bowei Ye[1] boweiye2@uic.edu
Wei Tian[2] jksrtw@gmail.com
Boshen Wang[3] Boshen.Wang@utsouthwestern.edu
Jie Liang[1,*] jliang@uic.edu

Abstract

Geometric and topological properties of protein structures, including surface pockets, interior cavities, and cross channels, are of fundamental importance for proteins to carry out their functions. Computed Atlas of Surface Topography of proteins (CASTp) is a widely used web server for locating, delineating, and measuring these geometric and topological properties of protein structures. Recent developments in AI-based protein structure prediction such as AlphaFold2 (AF2) have significantly expanded our knowledge on protein structures. Here we present CASTpFold, a continuation of CASTp that provides accurate and comprehensive identifications and quantifications of protein topography. It now provides (i) results on an expanded database of proteins, including the Protein Data Bank (PDB) and non-singleton representative structures of AlphaFold2 structures, covering 183 million AF2 structures; (ii) functional pockets prediction with corresponding Gene Ontology (GO) terms or Enzyme Commission (EC) numbers for AF2-predicted structures; and (iii) pocket similarity search function for surface and protein-protein interface pockets. The CASTpFold web server is freely accessible at https://cfold.bme.uic.edu/castpfold/.

1 Introduction

Protein structures are complex, containing a multitude of surface pockets, internal cavities, and interconnected channels. These distinctive topographical and topological features provide the micro-environments essential for the biochemical functions of the proteins, such as binding with ligands, interacting with DNA, and catalyzing enzymatic reactions. Identification and measurement of these topographical features are of fundamental importance in deciphering the relationship between protein structures and functions (1), in studying protein fitness (2; 3), and in developing therapeutic interventions (4).

The CASTp server provides detailed quantitative analysis of protein topographical and topological features (5–7) and is widely used in various applications, including exploring therapeutics for neuropsychiatric disorders (8), drugging "undruggable" pockets (9), elucidating G-protein coupling specificity (10), unraveling lipid translocation mechanisms in autophagy (11), understanding bacteriophage host ranges (12), and investigating plant metabolism processes (13).

Recently, the advent of advanced AI tools such as AlphaFold2 (AF2) (14) has significantly expanded the repository of available protein structures. The AlphaFold Protein Structure Database (AFDB) provides more than 214 million predicted protein structures. As AF2-predictions provide valuable resources of protein structures, here we present the CASTpFold

[1]Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA.
[2]Salk Institute for Biological Studies, La Jolla, CA 92037, USA.
[3]UT Southwestern Medical Center, Dallas, TX 75390, USA.
[*]To whom correspondence should be addressed. Tel: +1 312 355 1789; Fax: +1 312 413 218.

server, which retains all critical functionalities of its predecessors while providing topographical and topological quantification to the AF2-predicted structures, so that the research community could have a comprehensive tool for analyzing and understanding the spatial arrangement, function, and similarity of protein topography in both predicted and PDB protein structures, facilitating more informed investigations.

2 MATERIALS AND METHODS

2.1 The CASTpFold server

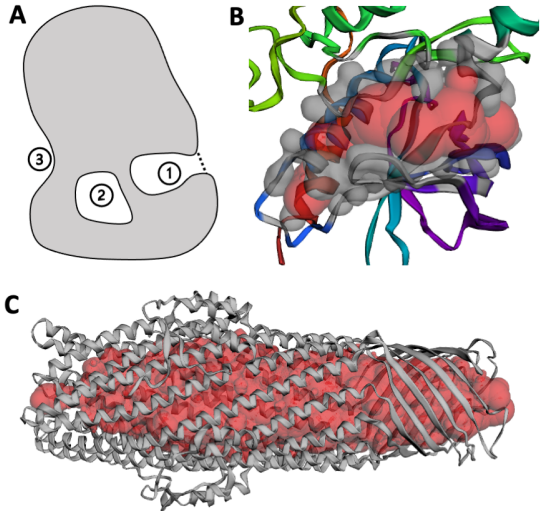

Figure 1: The concave regions identified on protein structures. (A) Protein surfaces contain three types of concave regions: ① an accessible surface pocket with a narrowed entrance, indicated by a dashed line; ② a completely enclosed cavity (void) lacking an entrance; and ③ a shallow depression. (B) A surface pocket supported by atoms in grey as identified by CASTpFold, which exhibits kinase activity (GO:0016301, Pocket ID: 2, style: "surface") in the AF2-predicted structure (AF2 ID: A0A2N0QT79). Its imprint is in red. (C) A cross channel in the outer membrane protein ST50 with its imprint highlighted in red balls (PDB ID: 5BUN, Pocket ID: 1).

The CASTpFold server is based on the alpha shape method (15) from topological data analysis to identify topographical and topological features and to compute their areas, volumes, and imprints (16–20). Topographical concave regions of proteins include surface pockets and depressions, CASTpFold considers pockets, which are on the protein surface and are accessible through a narrow entrance from the outside openings large enough for a probe ball, e.g., a water molecule, to access (Fig 1A-①). Pockets differ from depressions, which are concave surface areas without entrance constriction (Fig 1A-③). As an example, Fig 1B depicts a pocket in an AF2-predicted structure, which is predicted to be a functional region involved in kinase activity. Cavities are topological features and are internal voids buried inside a protein that the probe ball cannot access (Fig 1A-②). Channels are a class of pockets with two mouth openings at opposite sides. An example is shown in Fig 1C (PDB ID: 5BUN, Pocket ID: 1). CASTpFold identifies and measures surface pockets and interior cavities on a protein structure, as well as protein-protein interface (PPI) pockets and cavities where multiple proteins or subunits interact to form protein complexes.

The CASTpFold server also includes other information about the protein. The secondary structures are assigned using the DSSP method (21). The protein residue annotations are obtained from the UniProt database (22) and aligned with the PDB structure via the SIFTS

database (23). Functional pockets for AF2 structures identified by CASTpFold are annotated with GO terms or EC numbers by integrating results obtained from the deep learning algorithm DeepFRI (24; 25). The pocket similarity is measured using an approach adapted from the Foldseek method (26).

2.2 New features of CASTpFold

2.2.1 Enlarged database.

CASTpFold expands its scope beyond the PDB database (27) and now includes all non-singleton AF2 representative structures (28), providing an exhaustive analysis of protein topography of most of the protein-universe. Specifically, the 214 million AF2-predicted structures (as of Nov 2022) can be grouped into 18,661,407 structural clusters, by the criterion that their representative structures are recognized by Foldseek (26). For protein sequences in the UniProt database, after removing those labeled as "fragments", the remaining 2,302,908 non-singleton sequences clusters (as of Dec 2023) can be mapped to the 183,581,108 AF2-predicted structures (28), whose representative structures are all included in the CASTpFold server. When querying any of the 183,581,108 AF2-predicted structures, the search is automatically directed to its appropriate representative AF2 structure. For example, a query for the structure (AF2 ID: R4G0B4) redirects to its representative structure (AF2 ID: A0A6B1EM21), ensuring users can efficiently find the most relevant structural information.

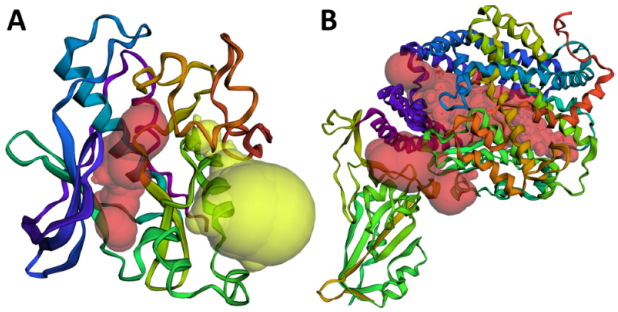

Figure 2: Surface pockets and protein-protein interface pockets: (A) Two surface pockets identified on the AF2-predicted structure (AF2 ID: A0A0C3Q0M4, Pocket IDs: 1 in lime and 2 in red). They are collectively predicted to carry out kinase activity; (B) A member of a PPI pocket cluster (cluster size: 294), located at the interface of the SARS-CoV-2 Spike/ACE2 Complex (PDB ID: 6M0J, Pocket ID: 1)

2.2.2 Predicted functional pocket for AF2 structures.

The UniProt database (22), with its repository of over 250 million sequences, has only about 0.3% of its entries manually reviewed in UniProtKB/Swiss-Prot. Given that protein functionalities are enabled by specific local surface regions, pinpointing relevant topographical features such as pockets and cavities can help decipher the mechanisms of protein functions. To identify potential functional pockets and cavities in unannotated AF2-predicted structures, we combine protein topography analysis with functional residue identification obtained through DeepFRI (24; 25), so our computed functional pockets along with DeepFRI derived GO terms or EC numbers are provided. An example is shown In Fig 2A, where two surface pockets (Pocket IDs: 1 in lime and 2 in red) in an AF2-predicted structure (AF2 ID: A0A0C3Q0M4) are predicted to facilitate kinase activity. Here the computed negative imprint of the pocket volume are colored lime and red, respectively. Fig 2B illustrates a member of a PPI pocket cluster (cluster size: 294), highlighted in red, located at the functional interface between the SARS-CoV-2 Spike pro-

tein and the human angiotensin-converting enzyme-2 (ACE2) protein (PDB ID:6M0J, Pocket ID: 1).

2.2.3 Pocket similarity search.

In our dataset of PDB structures and 2.3 million AF2-predicted representative structures, we have collected 4,108,408 surface pockets and 433,078 protein-protein interface (PPI) pockets, each satisfying the requirement of containing a minimum of 14 residues. The surface and PPI pocket databases have been clustered using the Foldseek algorithm (26) for rapid identification of similar pockets. Users can explore the relationship among the surface and PPI pockets by clicking the "Pocket Similarity" panel, where they can access, download, and visualize lists of similar pockets.

3 INPUT AND OUTPUT

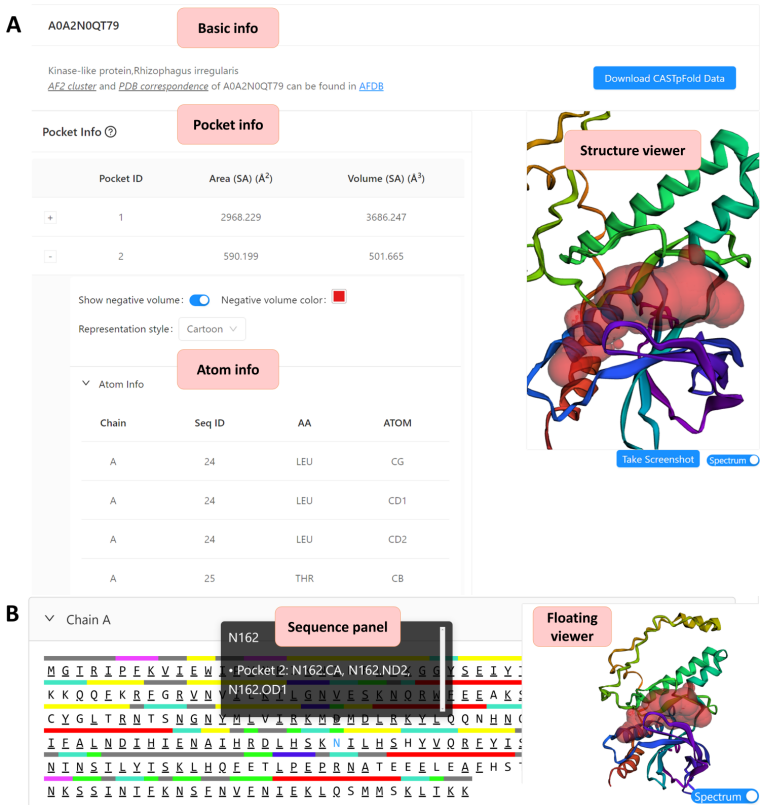

Figure 3: The main user interface of the CASTpFold server supports both precomputed and user-uploaded structures. (A) The server displays data on pocket surface area, volume, and detailed information on atoms contributing to the pocket. Here the second pocket is highlighted in "cartoon" style. (B) The sequence panel presents information on atomic contributions and annotations of residues along the sequence. Here the residue "N162" and its atoms contributing to the pocket formation are listed(Pocket ID:2). In addition, a structure viewer and a floating viewer facilitate visualization of the results.

3.1 Input

The CASTpFold server processes protein structures in PDB/mmcif format, incorporating a user-specified probe radius for detailed topographic analysis of the solvent-accessible surface. Users

can either explore pre-computed results by searching with standard PDB/AF2 IDs through the server's interface, or submit their own protein structures for computation. Pre-computed results were derived using a default probe radius of 1.4 Å for water. For customized computation requests, users have the option to adjust the probe radius within the range of 0 to 10 Å, enabling customized analysis.

3.2 Output

The CASTpFold server identifies all surface pockets, interior cavities, and cross channels in a protein structure, offering precise mapping of every atom involved in these topographic features. Additionally, it quantifies their exact volumes and surface areas, including the dimensions of any mouth openings. These calculations are performed analytically through the solvent-accessible surface model (SA) (29) and the molecular surface model (MS) (30). The "Pocket Info" panel will display only SA-related data, while MS-related information is available in the downloadable content, including the absolute values and ratios of atom-level surface exposed areas and volumes for both SA and MS models. User submitted structures will be subject to detailed topographical analyses, the results of which can be accessed and downloaded via CASTpFold. Predictions of functional pockets and measurements of pocket similarities are also accessible, although they are currently restricted to the precomputed structures.

3.3 Case study

The predicted functional pocket shown in Fig 1B is further illustrated in Fig 3, which presents a pocket related to potential kinase activity identified in the AF2-predicted structure (AF2 ID: A0A2N0QT79, Pocket ID: 2).

In Fig 3A, the "Basic Info" panel for protein (AF2 ID: A0A2N0QT79) displays information from the UniProt database. The "Pocket Info" panel depicts the selected pocket (Pocket ID: 2), which is rendered in cartoon style in the "Structure viewer" panel. Additionally, an expandable "Atom Info" panel reveals the atomic details of the selected pocket. Fig 3B shows the "Sequence Info" panel for the structure, including pocket residue with atom information and annotations. The viewer panel is designed to float, maintaining visual access while navigating the site, and users can save the current view by using the "Take Screenshot" button.

While navigating the CASTpFold server, users exploring AF2-predicted structures can find a "Predicted Function" panel, which lists the predicted function of the pocket (Fig 4A), including the associated GO terms or EC numbers, functionally relevant positions, and links to details of potential functions. For PDB structures, an "Annotation" panel is available (Fig 4B), providing annotations extracted from the UniProt database (22). AF2 predicted functionally relevant residues or PDB annotated residues can be visualized together with their respective pockets in a combined view (Fig 4A and 4B). Additionally, a "Pocket Similarity" panel is accessible for both AF2 and PDB structures, allowing users to discover pockets similar to their query protein across other proteins in the database. For example, Fig 4C shows a pocket from an AF2-predicted structure (AF2 ID: A0A2N0QT79, Pocket ID: 2) that has a total of 93 pockets across both PDB and AF2 structures exhibiting similarity to the queried pocket. One of these is from the PDB structure (PDB ID: 7NQQ, Pocket ID: 1), which is visualized in the "Similar pocket viewer". The "PDB counterpart in AFDB" panel is designed for PDB structures to display their AF2 counterparts, enabling users to compare PDB structures with their AF2 equivalents by clicking the corresponding links (Fig 4D).

4 DISCUSSION

CASTpFold represents a major update of the CASTp server, with the following significant enhancements: (i) an expanded database that includes PDB entries and 2.3 million AF2 struc-

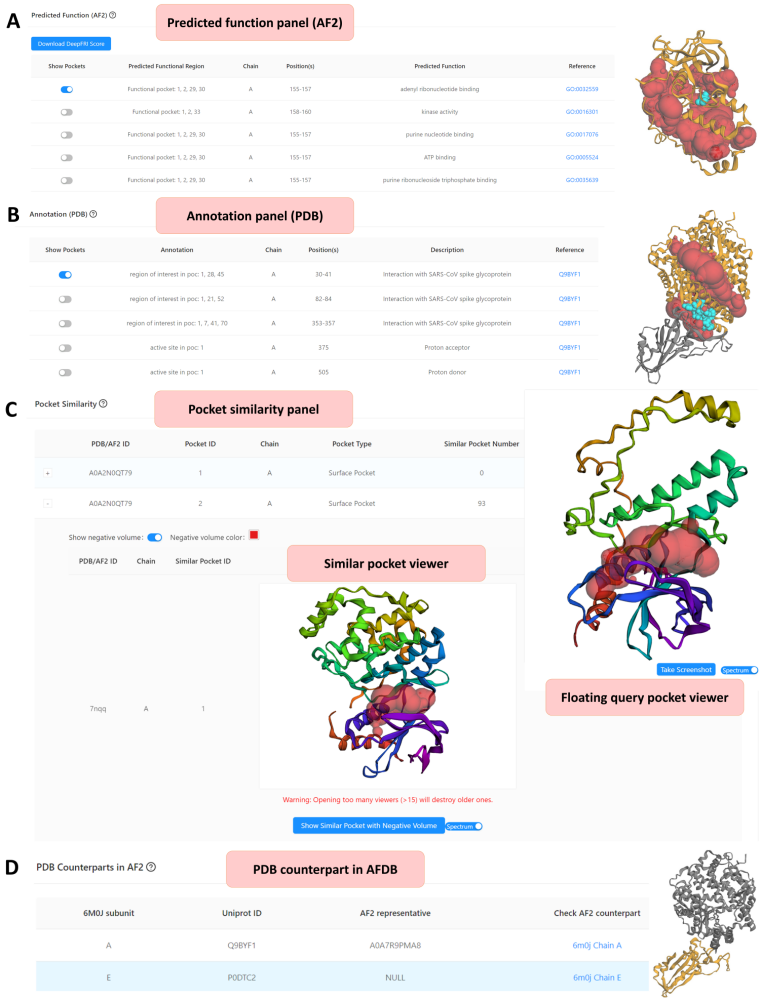

Figure 4: Additional features of the user interface of the CASTpFold server for precomputed structures. (A) The predicted function panel displays functional pockets in AF2 structures with links to relevant GO terms or EC numbers. The predicted functionally relevant positions are shown as cyan balls, their corresponding chains in orange, and the imprints of the pockets in red for combined viewing. (B) The annotation panel displays UniProt-sourced annotations for PDB structures, highlighting annotated residues, their respective chains, and their relevant pocket imprints in cyan, orange, and red, respectively. (C) The pocket similarity panel offers a list of pockets from other proteins similar to the query pocket, for both AF2 and PDB structures. (D) The "PDB counterpart in AFDB" panel shows the AF2 counterparts of all subunits of the query PDB structure, where the corresponding Uniport ID is mapped to its AF2 representative entry. Users can perform a detailed comparison by exploring the topographies of the AF2 counterpart by opening the counterpart link.

tures representing 183 million AF2 structures, covering over 85% of the AFDB; (ii) predicted functional pockets of AF2 structures, complete with associated GO terms or EC numbers; (iii) a pocket similarity search function of surface and PPI pockets for both PDB and AF2 structures; (iv) the user interface has been redesigned to be more informative, ensuring improved user engagement. These advancements are expected to further facilitate the exploration and analysis of protein structures and their functions. CASTpFold is free and open to all users and there is no login requirement.

5 FUNDING

The grant support of NIH R35GM127084 is gratefully acknowledged.

5.0.1 Conflict of interest statement.

None declared.

References

1. Toh, S., Holbrook-Smith, D., Stogios, P. J., Onopriyenko, O., Lumba, S., Tsuchiya, Y., Savchenko, A., and McCourt, P. (2015) Structure-function analysis identifies highly sensitive strigolactone receptors in *Striga*. *Science,* **350**(6257), 203–207.

2. Wang, B., Lei, X., Tian, W., Perez-Rathke, A., Tseng, Y.-Y., and Liang, J. (2023) Structure-based pathogenicity relationship identifier for predicting effects of single missense variants and discovery of higher-order cancer susceptibility clusters of mutations. *Briefings in Bioinformatics,* **24**(4), bbad206.

3. Ye, B., Wang, B., and Liang, J. (2023) Predicting Pathology of Missense Mutations through Protein-Specific Evolutionary Pattern. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* pp. 1–4.

4. Khan, I., Li, S., Tao, L., Wang, C., Ye, B., Li, H., Liu, X., Ahmad, I., Su, W., Zhong, G., Wen, Z., Wang, J., Hua, R.-H., Ma, A., Liang, J., Wan, X.-P., Bu, Z.-G., and Zheng, Y.-H. (2024) Tubeimosides are pan-coronavirus and filovirus inhibitors that can block their fusion protein binding to Niemann-Pick C1. *Nature Communications,* **15**(1), 162.

5. Binkowski, T. A., Naghibzadeh, S., and Liang, J. (2003) CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Research,* **31**(13), 3352–3355.

6. Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., and Liang, J. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research,* **34**(suppl$_2$), W116–W118.

7. Tian, W., Chen, C., Lei, X., Zhao, J., and Liang, J. (2018) CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Research,* **46**(W1), W363–W367.

8. Kim, K., Che, T., Panova, O., DiBerto, J. F., Lyu, J., Krumm, B. E., Wacker, D., Robertson, M. J., Seven, A. B., Nichols, D. E., Shoichet, B. K., Skiniotis, G., and Roth, B. L. (2020) Structure of a Hallucinogen-Activated Gq-Coupled 5-HT$_{2A}$ Serotonin Receptor. *Cell,* **182**(6), 1574–1588.e19.

9. Kessler, D., Gmachl, M., Mantoulidis, A., Martin, L. J., Zoephel, A., Mayer, M., Gollner, A., Covini, D., Fischer, S., Gerstberger, T., Gmaschitz, T., Goodwin, C., Greb, P., Häring, D., Hela, W., Hoffmann, J., Karolyi-Oezguer, J., Knesl, P., Kornigg, S., Koegl, M., Kousek, R., Lamarre, L., Moser, F., Munico-Martinez, S., Peinsipp, C., Phan, J., Rinnenthal, J., Sai, J., Salamon, C., Scherbantin, Y., Schipany, K., Schnitzer, R., Schrenk, A., Sharps, B., Siszler, G., Sun, Q., Waterson, A., Wolkerstorfer, B., Zeeb, M., Pearson, M., Fesik, S. W., and McConnell, D. B. (2019) Drugging an undruggable pocket on KRAS. *Proceedings of the National Academy of Sciences,* **116**(32), 15823–15829.

10. Maeda, S., Qu, Q., Robertson, M. J., Skiniotis, G., and Kobilka, B. K. (2019) Structures of the M1 and M2 muscarinic acetylcholine receptor/G-protein complexes. *Science,* **364**(6440), 552–557.

11. Matoba, K., Kotani, T., Tsutsumi, A., Tsuji, T., Mori, T., Noshiro, D., Sugita, Y., Nomura, N., Iwata, S., Ohsumi, Y., Fujimoto, T., Nakatogawa, H., Kikkawa, M., and Noda, N. N. (2020) Atg9 is a lipid scramblase that mediates autophagosomal membrane expansion. *Nature Structural & Molecular Biology,* **27**(12), 1185–1193.

12. Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M. J., and Kilcher, S. (2019) Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. *Cell Reports,* **29**(5), 1336–1350.e4.

13. Akbudak, M. A., Yildiz, S., and Filiz, E. (2020) Pathogenesis related protein-1 (PR-1) genes in tomato (Solanum lycopersicum L.): Bioinformatics analyses and expression profiles in response to drought stress. *Genomics,* **112**(6), 4089–4099.

14. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature,* **596**(7873), 583–589.

15. Edelsbrunner, H. and Mücke, E. P. (1994) Three-dimensional Alpha Shapes. *ACM Transactions on Graphics,* **13**(1), 43–72.

16. Liang, J., Woodward, C., and Edelsbrunner, H. (1998) Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science,* **7**(9), 1884–1897.

17. Edelsbrunner, H., Facello, M., and Liang, J. (1998) On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics,* **88**(1), 83–102 Computational Molecular Biology DAM - CMB Series.

18. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V., and Subramaniam, S. (1998) Analytical shape computation of macromolecules: I. molecular area and volume through alpha shape. *Proteins: Structure, Function, and Bioinformatics,* **33**(1), 1–17.

19. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V., and Subramaniam, S. (1998) Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins: Structure, Function, and Bioinformatics,* **33**(1), 18–29.

20. Ebalunode, J. O., Ouyang, Z., Liang, J., and Zheng, W. (2008) Novel Approach to Structure-Based Pharmacophore Search Using Computational Geometry and Shape Matching Techniques. *Journal of Chemical Information and Modeling,* **48**(4), 889–901.

21. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**(12), 2577–2637.

22. Consortium, T. U. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research,* **47**(D1), D506–D515.

23. Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2012) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research,* **41**(D1), D483–D489.

24. Ye, B. and Liang, J. (2024) Predicting Functional Surface Topographies Combining Topological Data Analysis and Deep Learning Across the Human Protein Universe. Accepted for publication.

25. Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., Xavier, R. J., Knight, R., Cho, K., and Bonneau, R. (2021) Structure-based protein function prediction using graph convolutional networks. *Nature Communications,* **12**(1), 3168.

26. van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. (2024) Fast and accurate protein structure search with Foldseek. *Nature Biotechnology,* **42**(2), 243–246.

27. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Research,* **28**(1), 235–242.

28. Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C. L. M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., and Steinegger, M. (2023) Clustering predicted structures at the scale of the known protein universe. *Nature,* **622**(7983), 637–645.

29. Lee, B. and Richards, F. (1971) The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology,* **55**(3), 379–IN4.

30. Connolly, M. L. (1983) Solvent-Accessible Surfaces of Proteins and Nucleic Acids. *Science,* **221**(4612), 709–713.