De novo identification of protein domains in cryo-electron tomography maps from AlphaFold2 models

Yi Lu[1], Guanglin Chen[1], Fei Sun[3,4,5,6,*], Yun Zhu[3,*], Zhiyong Zhang[1,2,*]

[1]Department of Physics and Anhui Center for Fundamental Sciences in Theoretical Physics, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China

[2]MOE Key Laboratory for Cellular Dynamics, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China

[3]National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China.

[4]University of Chinese Academy of Sciences, Beijing, China.

[5]Center for Biological Imaging, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China.

[6]Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou, Guangdong, China

[*]Correspondence: Fei Sun, feisun@ibp.ac.cn; Yun Zhu, zhuyun@ibp.ac.cn; Zhiyong Zhang, zzyzhang@ustc.edu.cn

Abstract

In recent years, in situ cryo-electron tomography (cryo-ET) has emerged as a powerful method for visualizing structures of protein complexes within their native cellular environment. However, the resolution of in situ cryo-ET remains lower than that of in vitro cryo-electron microscopy single particle analysis (cryo-EM/SPA). Identifying proteins in a cryo-ET density map is particularly challenging owing to the limited visibility of side chains. To address this issue, we developed DomainSeeker, a tool designed for de novo identification of protein domains in cryo-ET maps of protein complexes. DomainSeeker integrates AlphaFold2-predicted models to partition proteins into domains using predicted aligned errors and a clique-based graph cluttering algorithm, and subsequently fits these domains into segmented density maps. Finally, it uses a two-dimensional scoring metric to identify the most probable domain of each density map. On a test set containing 26 model-based densities segmented from a sperm microtubule doublet map, DomainSeeker efficiently identified nearly all of these domains at both 6.5 Å and 10 Å. On another test set of 20 auto-segmented densities, 17 out of them were also determined. Furthermore, we successfully identified a new protein within an unmodeled region of this map.

Introduction

In recent decades, structural biology has experienced significant advancements[1]. Traditionally, structural biologists decompose cellular proteomes, purify individual components, and study them separately[2]. With the improvement in cryo-electron microscopy (cryo-EM) and the growing use of artificial intelligence (AI), it has become feasible to directly analyze biomolecular structures within cellular environments and explore subcellular structures as a whole[3]. Notably, in situ cryo-electron tomography (cryo-ET) stands out for resolving biomolecular structures in their native cellular environment[4,5,6] and visualizing the positions and interactions of the intracellular proteome, which is often referred to as visual proteomics[7,8]. This technique bridges the gap between protein structures and cellular architecture, providing deeper insights into protein functions and organization within cells[9].

However, while structures of in vitro protein complexes can often be determined through cryo-EM single particle analysis (cryo-EM/SPA) at atomic resolution[10,11], the overall resolution of in situ cryo-ET remains comparatively lower[12]. Consequently, for regions with invisible sidechains of in situ cryo-ET maps, identifying proteins and constructing atomic models directly from the density maps alone is not feasible[13]. Hence, the developing visual proteomics requires efficient methods to facilitate the model building of these densities.

Under these circumstances, a general strategy is employed to solve the problem. First, segment the unsolved target density from the full density map. Then, achieve atomic structures of a set of candidate proteins and separate them into domains. Next, fit all candidate domains into the target density. Finally, assess the match between all fitted domains with the target density to identify the most probable domain. This process can be combined with subsequent experiments and computations to finalize the modeling of the unknown density[14,15,16].

In this process, the candidate protein list can be obtained from mass spectrometry (MS)[17] experiments, or alternatively, the entire proteome of the species can be utilized. Over the decades, numerous tools for predicting protein structures, such as homology modeling[18], AlphaFold2[19], AlphaFold3[20], RoseTTAFold[21] and ColabFold[22], have been developed. Among them, AlphaFold2 is the first to achieve predictions approaching experimental accuracy in a majority of cases[19]. In addition, the AlphaFold Protein Structure Database (AFDB) contains more than 200 million entries covering the proteomes of human and 47 other key organisms[23]. Thus, the predicted atomic structures of the candidate proteins can be easily obtained from AFDB, making it feasible to automate this process.

There have been several studies that fully or partially followed this process. By fitting 21,615 AlphaFold2-predicted protein models of the mouse proteome into a mouse sperm axonemal microtubule structure, several microtubule inner proteins (MIPs), such as Tektin 5, CCDC105, and SPACA9, were identified as novel microtubule-associated proteins[14]. In another cryo-ET map of mouse sperm flagella at approximately 6.5 Å resolution, parts of the model were constructed with the aid of MS and AlphaFold2[15].

However, all the four steps in this procedure require exploration for better overall performance. First, the cryo-ET density segmentation is usually done manually[14] and often results in more or less density compared to the corresponding domain, complicating subsequent domain fitting and ranking. Second, while AlphaFold2 models demonstrate high accuracy for compact domains, the relative positions and orientations between these domains may not always be correct[24]. And this should be taken into consideration in domain parsing. Third, some candidate domains may be wrongly fitted into the target density, potentially leading to the correct domain being missed. Fourth, most protocols rank the fitted domains on the basis of the overall cross-correlation (CC)[14,16], which means that a highly ranked model tends to exhibit a similar overall shape to the target density but not necessarily correct detailed structures. To the best of our knowledge, DomainFit[16] is the only software to date that has partially explored these issues.

In this work, we developed DomainSeeker, a program designed to accurately identify the protein domain corresponding to a given density map. We mainly explored two steps of this process: domain parsing and the scoring/ranking of the fitted models. For domain parsing, we combined the AlphaFold2 output, that is, the predicted aligned error (PAE)[25,26,27], and a graph cluttering algorithm based on cliques[28,29] to parse the input predicted models. For the scoring and ranking, we proposed a two-dimensional scoring metric, overlap volume and overlap correlation, based on Laplacian filtering[30] instead of the traditional CC. DomainSeeker then calculates local z-scores assessing the vertical deviation of data points from the average curve on the 2D plot, and ranks fitted domains by local z-scores to identify the most probable domains for each density map. As for the other two steps, we utilized ChimeraX[31] to segment the density map and fit domains into target densities.

To validate DomainSeeker, we selected 26 domains from 20 different MIPs of the solved atomic structures corresponding to the sperm microtubule doublet[15]. We then segmented the densities around these 26 domains from the cryo-ET map of the sperm microtubule doublet as target densities. DomainSeeker identified the correct domains with high precision. For 23 densities, the correct domains ranked first based on the local z-score. For the remaining 3 densities, the correct domain ranked within the top seven. Additionally, to test whether DomainSeeker can identify domains for densities segmented without reference models, we automatically segmented the entire density map using ChimeraX and selected 20 densities that fully or partially covered our selected domains for testing. DomainSeeker successfully identified 17 of them as well. Furthermore, we identified a new protein for an unmodeled density in the sperm microtubule doublet.

Results

Workflow of DomainSeeker

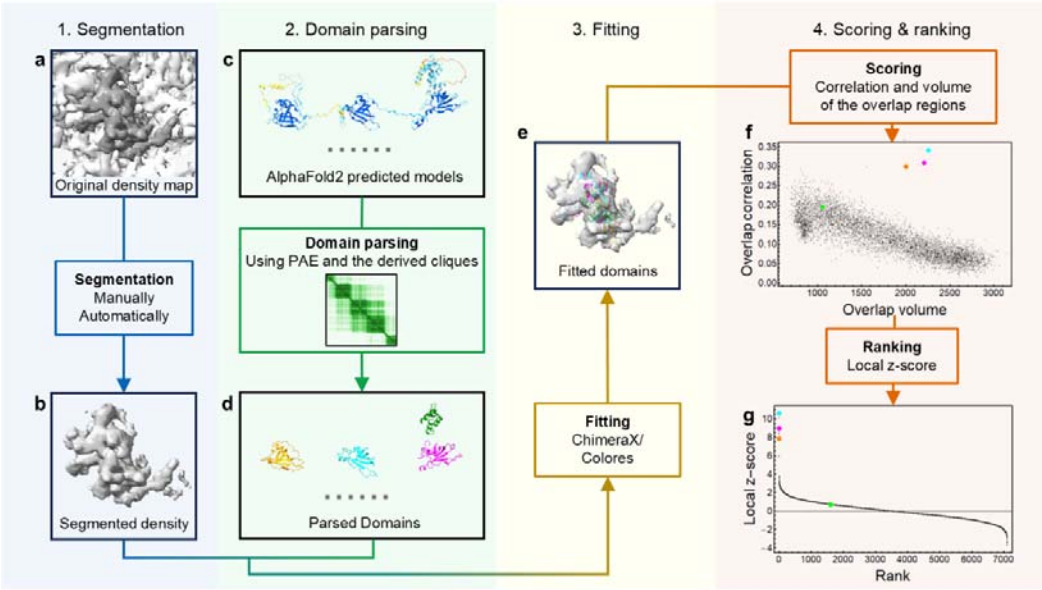

Figure 1. Workflow of DomainSeeker

(a) The full cryo-ET map. (b) Segment the target density out of the full map. (c) Download predicted models from AFDB. (d) Parse AlphaFold2-predicted models into domains using PAE and a clique-based graph cluttering algorithm. (e) Fit all domains into the target density. (f) Calculate the correlation and volume of the overlap region for each fitted domain and plot them on a 2D figure. (g) Rank all fitted domains by a local z-score. The UniProt ID of the sample protein is Q9D9T8.

Figure 1 illustrates the workflow of DomainSeeker to identify the most probable domain for a target density. First, segment the target density out of the full map manually or automatically. Second, download the predicted models and PAE files of candidate proteins from AFDB. The list of candidate proteins can be obtained from MS, or alternatively, one can take the whole proteome. Third, divide the protein structures into domains based on cliques derived from the PAE matrix. Fourth, fit each domain into the target density. After that, DomainSeeker calculates the correlation and volume of the overlap region between each fitted domain and the target density, and plot these on a two-dimensional graph. Usually, the most probable domain stands out in the upper-right corner. Finally, DomainSeeker calculates the local z-score for each data point based on the two-dimensional graph to assess its vertical deviation from the average value curve of the data, and all fitted domains are ranked by local z-scores to identify the most probable domain.

Target densities and candidate proteins for testing

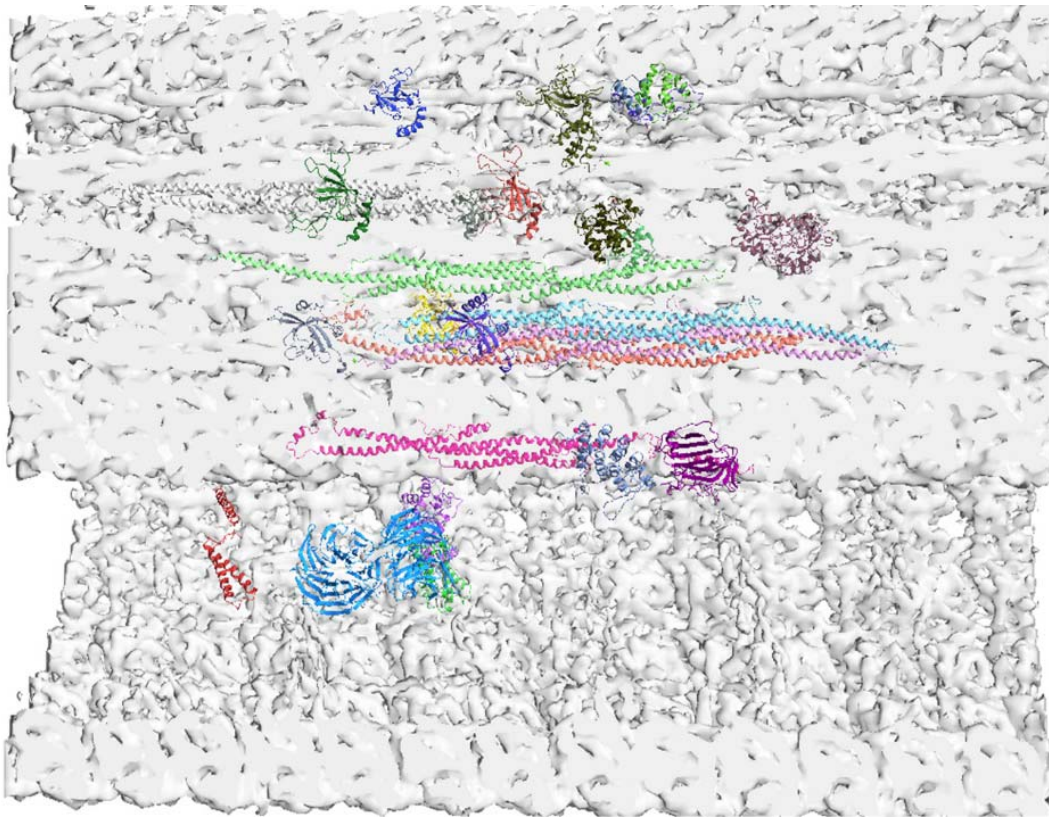

Figure 2. 26 domains for testing DomainSeeker

These 26 domains were selected from the atomic models (PDB entry 8I7R) corresponding to a cryo-ET map (EMD-35230) of the sperm microtubule doublet. Each domain is represented in a different color.

To test the performance of DomainSeeker, we used a cryo-ET map (EMD-35230) of the sperm microtubule doublet at about 6.5 Å resolution[15] to generate the test set. The atomic structures solved from this map was recorded in PDB 8I7R. Among the 36 MIPs in PDB 8I7R, 16 either lack structured domains or consist solely of a long single helix. These proteins require alternative identification methods. Consequently, we selected 26 domains from the remaining 20 different MIPs, as shown in Figure 2 and Table S1.

We simultaneously loaded the map and the corresponding domains into ChimeraX. Using the Color Zone tool, we painted densities within 5 Å of the 26 target domains. Subsequently, we separated the colored densities from the original map via the Split Map tool. Additionally, we lowered the resolution of these densities to 10 Å to further evaluate DomainSeeker.

There are 4,303 proteins listed in the MS data, and 4,227 AlphaFold2 predicted models were downloaded as candidates. Several proteins are absent from AFDB either because of their lengthy sequences or because they are missing from the FASTA file in UniProt, adhering to the "one protein sequence per gene" rule[32]. Then these predicted models are separated into 7129 domains by DomainSeeker.

Domain parsing based on cliques derived from PAE

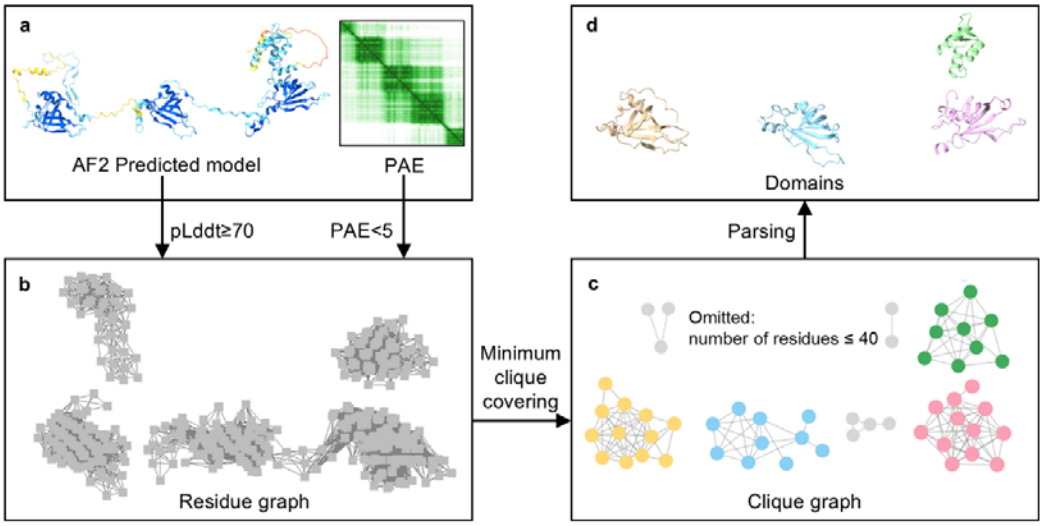

Figure 3. Domain parsing

(a) The input model and the PAEs. (b) Exclude residues with pLDDT below 70 and mutual PAEs above 5 to form a residue graph. (c) Apply minimum cliques to cover this graph, treating the resulting cliques as nodes in a new clique graph. Two cliques are linked if their shared residue ratio is over 50% or if their PAE edge ratio exceeds 60%. (d) All residues within these interconnected cliques form a domain. The UniProt ID of the sample protein is Q9D9T8.

To partition the predicted protein structures into suitable domains, we used the AlphaFold2 PAE, which is indicative of secondary structural elements and interdomain flexibilities, alongside the per-residue local distance difference test (pLDDT), which quantifies local confidence levels[23]. The procedure is shown in Figure 3. First, residues with a pLDDT score below 70, denoted as 'low confidence' by AlphaFold2[23], are excluded. Subsequently, each remaining residue is treated as a node in a residue graph. An edge is established between each pair of nodes where the average PAE value is below 5. Notably, we observe the emergence of large-scale clique structures within the PAE graph. Each clique inherently constitutes an accurate and rigid entity, as evidenced by the presence of an edge between every pair of nodes within a clique. Consequently, we seek to minimize the coverage of the PAE graph with cliques, thereby forming resulting cliques as nodes in a new clique graph. Cliques with fewer than 5 nodes are omitted, as this is typical for free loops. The connection between two cliques is established if the common residue ratio exceeds 50% or the PAE edge ratio surpasses 60% between them. Consequently, all residues encompassed within a group of interconnected cliques constitute a domain. Notably, clique groups with fewer than 40 residues (Fig. 3c, in gray) are omitted for they can usually fit well in most densities or represent short flexible regions.

Figure 4a illustrates the AlphaFold2-predicted structure of Q9D485, colored based on the pLDDT. Correspondingly, Figure 4b displays the PAE plot, where darker

green shades indicate smaller PAE values. A PAE diagram typically provides clear illustrations of distinct domains. As shown in Figure 4b, there exist four main domains in the predicted model. DomainSeeker successfully separates these four domains (Fig. 4c).

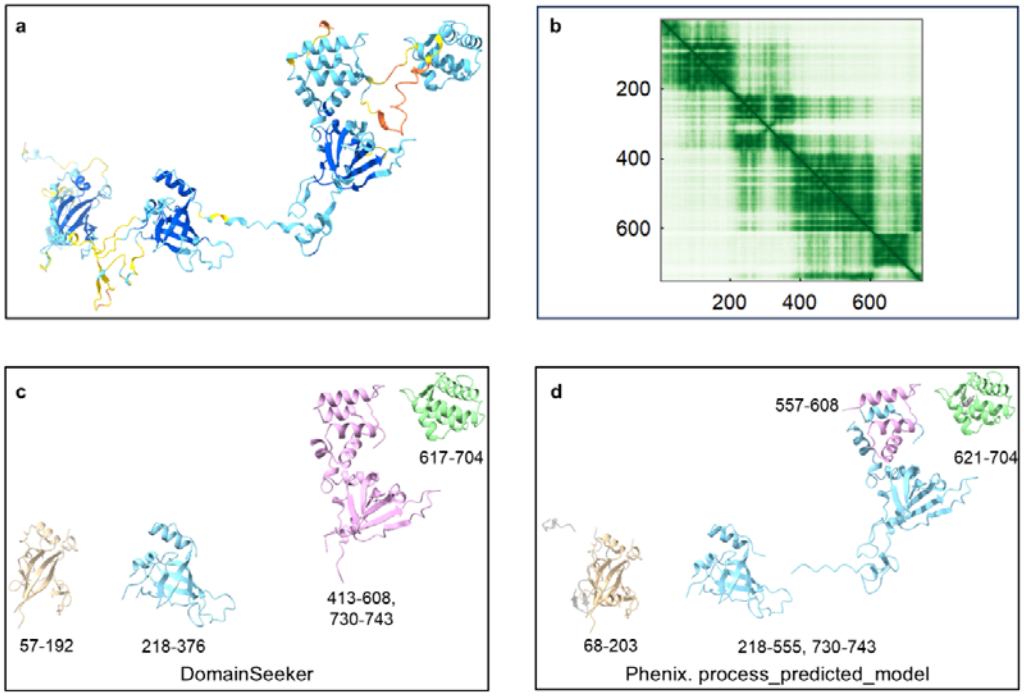

Figure 4. Domains parsed by DomainSeeker and Phenix

(a) AlphaFold2-predicted model of Q9D485. Residues with pLDDT above 90 are displayed in blue, 70 to 90 in cyan, 50 to 70 in yellow, and below 50 in orange. (b) PAE for the predicted model of Q9D485. The darker the green color is, the smaller the PAE. (c) Domains parsed by DomainSeeker. (d) Domains parsed by *phenix.process_predicted_model.*

For comparison, we also present results of the *phenix.process_predicted_model*[33] program in Figure 4d, while Phenix is a widely used package for determining macromolecular structures[34]. This program refines the input model by cropping low-confidence residues and organizing it into compact domains on the basis solely of its 3D structure. As observed, it mistakenly merges the middle two domains despite their insufficiently reliable relative positions. Additionally, the separation of residues 557-608 disrupts the integrity of the EF-hand structure, potentially resulting in an inaccurate fit.

Other results of domain parsing are drawn in Figure S1. DomainSeeker can reasonably parse predicted models with varying domain shapes and different features of PAE. Notably, DomainSeeker may not necessarily divide predicted models into traditional domains. For example, in Figures S1a-f, Phenix always maintains the whole Tektin-like structures as domains, while DomainSeeker may cut them because

the PAE plots suggest that the predicted models could be partially incorrect.

Overall, DomainSeeker uses PAE to evaluate whether the relative positions of the domains in a model are sufficiently reliable for merging and employs a clique-based approach to avoid unjustified segmentation. We examined thousands of samples for DomainSeeker beyond these instances and did not encounter any instance of obvious unreasonable partitioning.

Scoring and ranking

After fitting domains into the target density, the overall CC between the simulated density derived from the fitted domain and the target density is usually calculated to assess the match between each fitted domain and the target density. It is an overall estimation. Hence, a high cross-correlation indicates that the input domain closely resembles the target density in overall shape.

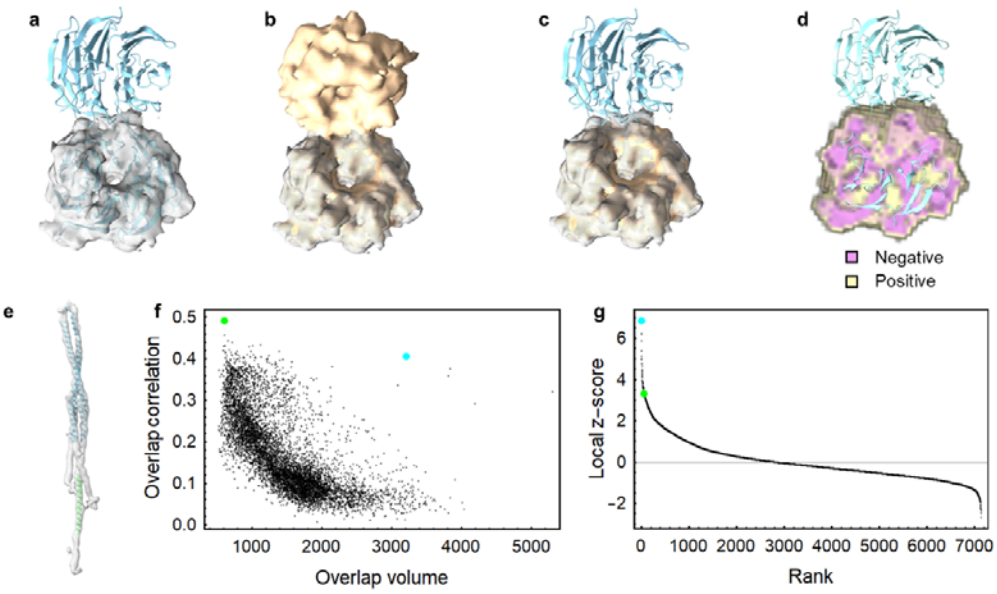

Figure 5. Scoring and ranking of the fitted domains

(a) A domain fitted to its corresponding target density. (b) The target density (white) and the simulated density (orange). (c) The target density (white) and the simulated density (orange) of the overlapping region. (d) Density after Laplacian filtering. negative regions in pink, positive regions in yellow. (e) A domain of Tektin1 (cyan) and a short helix (green) fitted to the Tektin1 density. (f) Correlation versus volume of the overlapping region. Tektin1 domain in cyan, short helix in green. (g) Local z-scores versus their rankings. The domain of Tektin1 in cyan ranks first.

However, the correct domain and the target density may differ in shape and size. For example, the target density may only cover a portion of the correct domain as shown in Figure 5a. In Figures 5b and 5c, the simulated density is shown in orange. The overall correlation of the simulated density and the target density is 0.70 in Figure 5b while the correlation of overlapping region in Figure 5c is 0.96. Therefore, conventional metrics for assessing overall relevance may yield lower scores for the

correct domain in such an instance.

In Figure 5d, we apply a Laplacian filter[30] to the target density. After filtering, the negative pink regions are closer to the atoms in the input model, while the positive yellow regions correspond to voxels with no nearby atoms. Clearly, Laplacian filtering enhances the contrast within a domain. However, positive voxels may extend beyond the boundary of the overlapping region, potentially reducing the correlation. To address this, we set a threshold for the positive regions to confine them near the overlapping boundary. Consequently, we calculate the correlation between the overlapping regions of the simulated and target densities after applying Laplacian filtering.

However, the overlap correlation alone does not determine which fitted domain best matches the target density. As shown in Figure 5e, both the larger cyan domain and the smaller green helix fit well in the density. Figure 5f indicates that the overlap correlation for the green helix is higher than that for the cyan domain. Given that the small helix fits well in most densities, this result is not surprising. Nevertheless, Figure 5e clearly shows that the cyan domain matches the density better than the green helix. Therefore, we also calculated the volume of the overlapping regions, considering that a fitted domain which matches the target density well should have a prominent overlap correlation among domains with similar overlap volumes. Therefore, we plotted all fitted domains on a two-dimensional graph depicting the volume and correlation of the overlapping regions. As shown in Figure 5f, the correct domain in cyan stands out in the upper-right corner of the two-dimensional graph.

Therefore, the volume-correlation graph itself is capable of determining the most probable domain. Furthermore, to rank the domains quantitatively, we defined a local z-score as the deviation of each data point from the average value curve vertically, measured in standard deviations. This local z-score metric can identify points that are notably higher than the local average scores. In Figure 5g, the cyan domain ranks first according to the local z-score.

Performance of DomainSeeker on model-based densities

To validate DomainSeeker, we employed it to identify protein domains of target densities. As shown in Figure 6a, we identified all the 26 densities at 6.5 Å from the top seven candidates, with 23 of them ranking first. The 2D plots are shown in Figure S2. The structures of these densities with fitted correct domains are shown in Figure S3. Specifically, densities 10 to 15 represent similar domains from two proteins, Efhc1 and Efhc2. Their structures are shown in Figure 6b. DomainSeeker can effectively distinguish these highly similar domains. Maps of density 3 (Tektin3) and density 4 (Tektin4) are shown in Figure 6c. They both show a notable structural difference at the right end of the density. In addition, Tektin1-5 and Tektl1 exhibit similar structures. Thus, the correct domains of density 3 and density 4 rank slightly lower. Nevertheless, their rankings remain higher than those of other proteins, allowing them to be identified as the most probable domains.

When examining the 10 Å results (Figure 6a, square), DomainSeeker performed nearly as effectively as it did at a 6.5 Å resolution (Figure 6a, circle), with the

exception of density 18. The density 18 represents part of the protein EFHB. We noticed a remarkable decline in its ranking for the correct domain. Interestingly, the correct domain stands out prominently in the 2D scoring plot (Figure 6d). The local standard deviation around the correct point is larger than that around the other points, primarily due to a turn in the data points toward the right end. Consequently, its ranking declines. Hence, users should always put an eye on the 2D scoring plot to avoid some evitable ranking errors. Despite the excellent performance of the local z-score on other densities, this particular case highlights the need for a more effective algorithm to accurately estimate deviations from the main data.

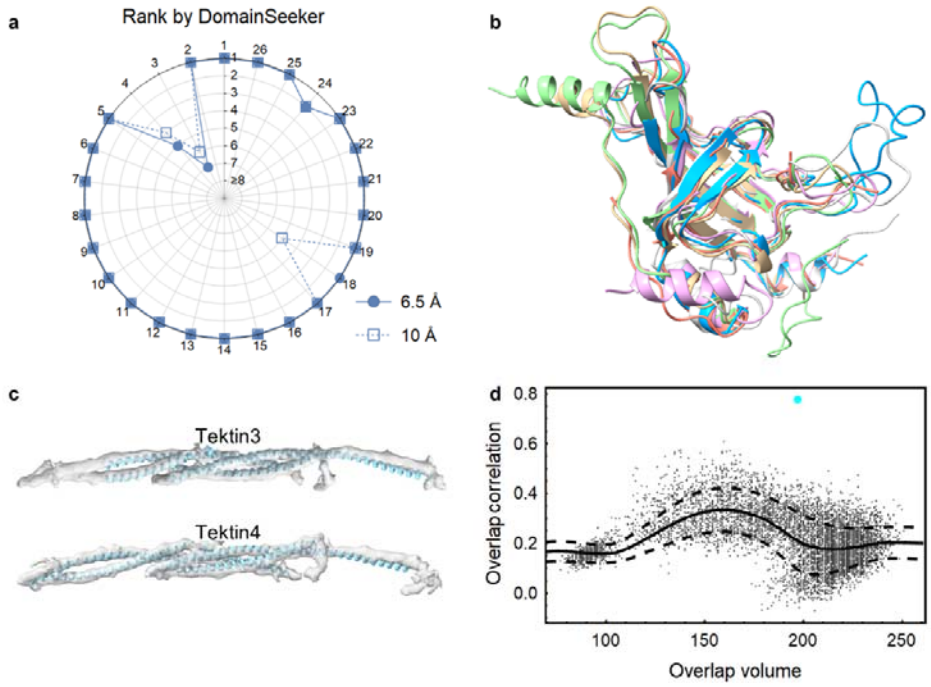

Figure 6. Performance of DomainSeeker on model-based densities

(a) Rankings of correct domains in their respective densities at resolutions of 6.5 Å (circle) and 10 Å (square). (b) Aligned structures corresponding to densities 10 through 15. (c) Tektin 3 and Tektin 4 aligned within their respective densities. (d) Plot of correlation versus volume of the overlapping region for the density 18. The cyan data point represents the correct domain. The solid line is the average curve. Points between the dashed lines deviate by no more than one standard deviation.

Comparison with DomainFit

DomainFit parses domains based on *phenix.process_predicted_model*. Then it fits domains into segmented densities using ChimeraX. Finally, it ranks all domains based on the p-value of the correlations for different fitting positions. Here, the p-value reflects the distinctiveness of the best fit among all fitting positions. We also tested

DomainFit on our test set using 200 initial searching placements for comparison. As shown in Figures 7a and 7d, DomainSeeker performs better than DomainFit on our test set at both 6.5 Å and 10 Å resolutions.

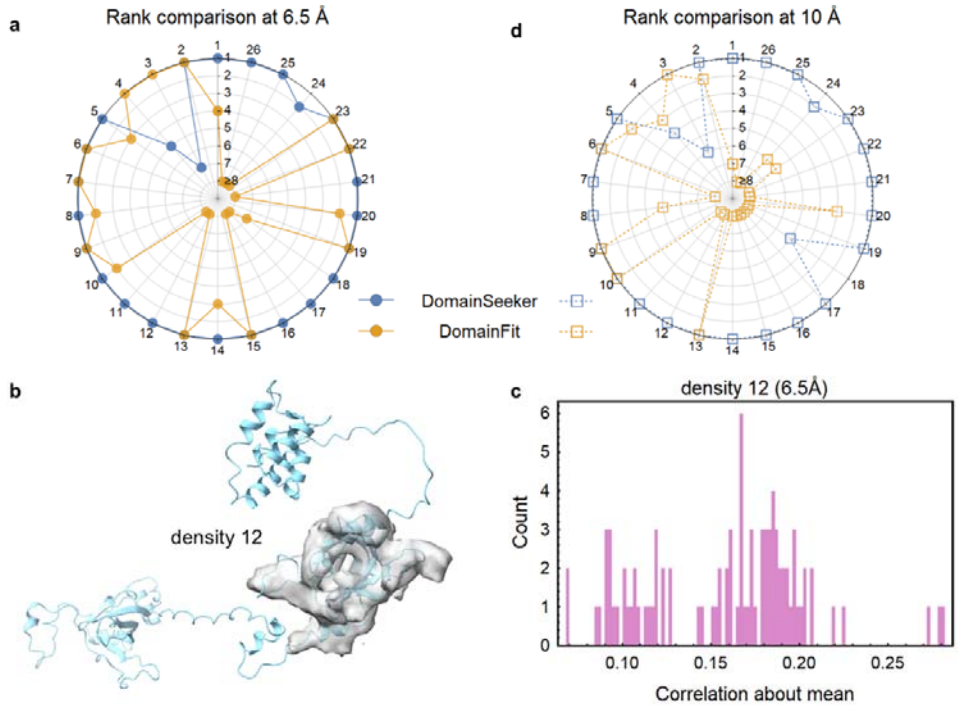

Figure 7. Comparison of performance between DomainSeeker and DomainFit

(a) Rankings of correct domains in their respective densities at the resolutions of 6.5 Å given by DomainSeeker (blue) and DomainFit (orange). (b) The cyan structure, which contains a domain corresponding to density 12, was parsed and correctly fitted into the density by DomainFit. (c) The histogram for the correlation about mean of diverse fitting positions between the cyan structure and density 12 in (b). (d) Rankings of correct domains in their respective densities at 10 Å given by DomainSeeker (blue) and DomainFit (orange).

At 6.5 Å (Figure 7a), DomainFit ranked the correct domains within the top seven for 18 densities, while for the other 8 densities, the correct domains were ranked outside the top twenty. One of the main reasons is the over-sized domains. The over-sized domain may lead to a broader range of correlations at different fitting positions. For instance, in Figure 7b, there are 3 sub-domains, if the sub-domain corresponding to the density is initially placed within the density, the fitting correlation will be relatively larger (0.15 to 0.20 in Figure 7c). Conversely, placing the other two sub-domains within the density initially results in a lower fitting correlation (0.10 to 0.15 in Figure 7c). This leads to the worse p-value.

In Figure 7d, at 10 Å resolution, DomainFit shows a considerable decline in performance. In contrast, DomainSeeker maintains performance at 10 Å that is comparable to its performance at 6.5 Å. Thus, DomainSeeker can sustain its accuracy

in identification up to 10 Å.

Performance of DomainSeeker on auto-segmented densities

The above analysis is based on densities segmented based on the solved models. Actually, reference models for segmentation are not available when working with an unmodeled cryo-ET map, which means that the segmented densities might not precisely match their corresponding domains in size. To test whether DomainSeeker can identify domains for densities segmented without reference models, we also used ChimeraX to automatically segment the entire density map at 6.5 Å and selected 20 segmented densities that either fully or partially covered our 20 chosen domains for testing purposes. Here, 6 of the 26 selected domains are omitted because they are Tektin-like proteins, which are elongated and difficult to be split automatically by ChimeraX.

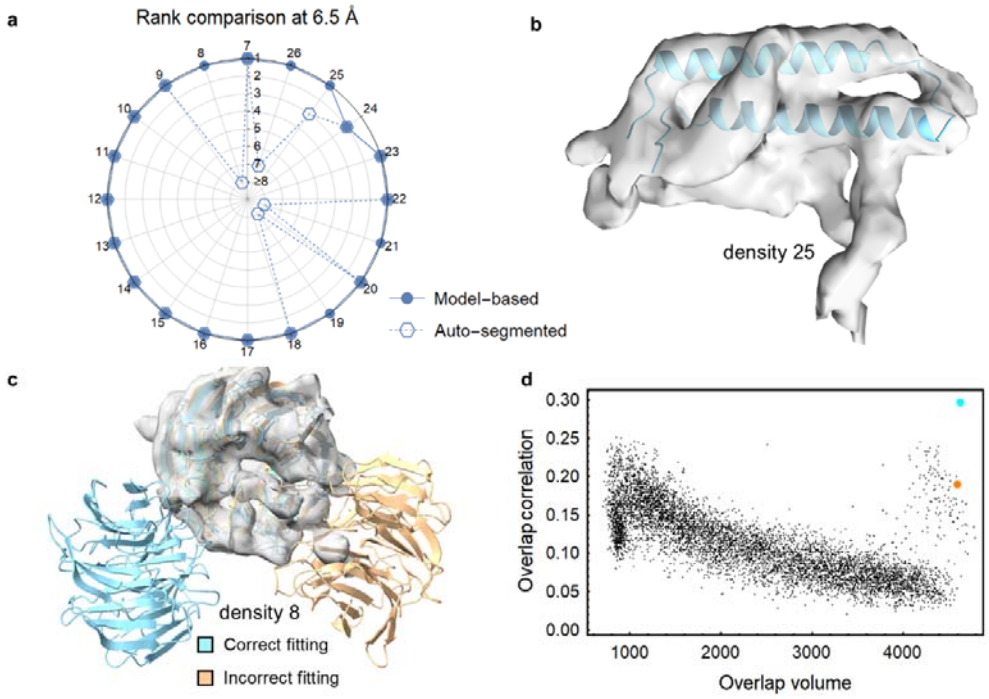

Figure 8. Performance of DomainSeeker on auto-segmented densities

(a) Ranks of the correct domains in densities segmented based on models (circle) and automatically (hexagon). (b) The correct domain fitted into the auto-segmented density 25. (c) The correct domain fitted into the auto-segmented density 8 is shown in cyan for the correct fitting position. While ChimeraX fitted it incorrectly in the orange position. (d) Correlation versus volume of the overlapping region for density 8. The correct fit position is shown by a cyan point, and the incorrect fit position by an orange point.

As shown in Figure 8a, DomainSeeker identified 17 densities successfully and 15 of them exhibit the same accuracy with the model-based densities. For the other two densities 25 and 26, they are susceptible to the redundant noise in the auto-segmented

densities because there are only about 60 residues in these two domains, as shown in Figure 8b. As a result, their rankings experience a slight decline.

For the three unrecognized densities 8, 19 and 20, ChimeraX assigned incorrect fitting positions for the corresponding domains. In Figure 8c, the cyan structure represents the correct domain for density 8 and is accurately fitted into the density. In contrast, the orange structure shows the incorrect fitting position assigned to the same domain by ChimeraX. Figure 8d further illustrates that although the orange data point for the incorrectly fitted domain is obscured by other data points, the corresponding cyan data point would prominently appear in the upper-right corner of the two-dimensional graph if the domain were correctly fitted. All two-dimensional scoring plots are shown in Figure S4. And all domains fitted into the auto-segmented densities are shown in Figure S5.

Therefore, the scoring metric of DomainSeeker works well. To further enhance the performance of DomainSeeker, we need to explore for a better fitting algorithm.

A new protein identified by DomainSeeker

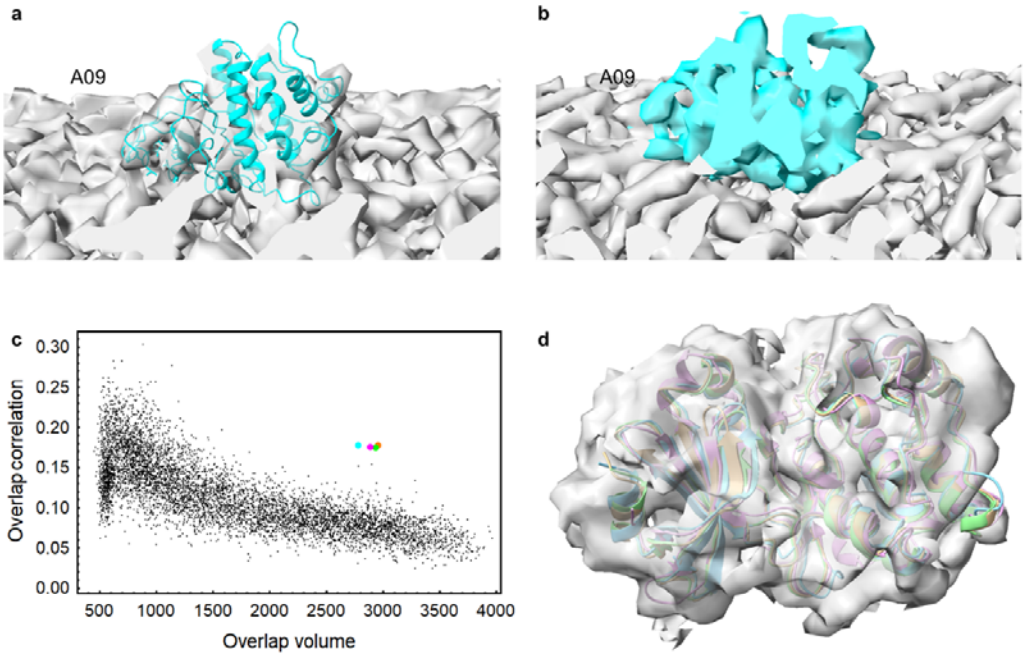

Figure 9. A new protein identified by DomainSeeker

(a) Tssk6 in the cryo-EM/SPA map outside A09. (b) An unmodeled region of the in-situ map outside A09. (c) 2D scoring figure for the unmodeled density. The cyan point represents Tssk6. Colored points are ranked in the top four by the local z-score. (d) Predicted structures of the top 4 domains. Tssk6 is colored in cyan.

Comparing the in situ cryo-ET map EMD-35230 with a cryo-EM/SPA map EMD-35823 at 3.5 Å[35], we noticed an unmodelled density adjacent to the outer side of protofilament A09. Figure 9a depicts part of the cryo-EM/SPA density, while Figure 9b shows the corresponding region in the in-situ density. This region in the

cryo-EM/SPA map was identified as Tssk6. However, due to a lack of sufficient evidence, it was not modeled in the in-situ map.

We employed DomainSeeker to analyze this density. In Figure 9c, the 2D scoring plot highlights Tssk6 in the upper right corner as a cyan point. Tssk6 ranks fourth in the local z-score ranking, with scores of the top four domains, as painted in vibrant colors in Figure 9c, being nearly identical. As depicted in Figure 9d, their structures closely resemble each other. Consequently, we have successfully identified the correct structure corresponding to this unmodeled density.

Discussion

In this work, we developed DomainSeeker to identify protein domains in segmented cryo-ET densities at an intermediate resolution from the AlphaFold2-predicted structures. We proposed two critical algorithms. First, using the PAE information, we employed a graph clustering method based on cliques to partition the predicted models into domains. We demonstrated that PAE is indispensable and effective for accurately parsing domains in AlphaFold2 predicted models. Second, we employed a local z-score metric based on the correlation and volume of the overlap region between the target and simulated densities to rank the candidate domains, which has been shown to be more effective than traditional cross-correlation metrics.

For the 26 model-based densities from MIPs in the sperm microtubule doublet, DomainSeeker identified every structure of these densities with high precision. We also showed that DomainSeeker can work efficiently up to a resolution of 10 Å, which could significantly aid in the reconstruction of in-situ cryo-ET densities. Besides, to test whether DomainSeeker can identified densities segmented without reference models, we used ChimeraX to automatically segment the entire density map and selected 20 densities that fully or partially covered our domains for testing. DomainSeeker successfully identified 17 of these densities.

Furthermore, we applied DomainSeeker to an unmodeled density of the in-situ cryo-ET map of the sperm microtubule doublet and successfully identified Tssk6 as one of the most probable candidates for that density.

Overall, DomainSeeker holds promise for effectively identifying proteins from in-situ cryo-ET maps. Although the 2D scoring plot performs exceptionally well, the local z-score may not be the optimal metric for distinguishing the most probable domains from the data points. We plan to refine our algorithms further, particularly by optimizing the ranking metric based on the 2D volume-correlation plot, to increase its performance. Additionally, the density segmentation and domain fitting steps in the identification process have a significant impact and require further exploration.

Methods

Segmentation

Densities of interest can be segmented from the full map either manually or automatically using software like ChimeraX.

In ChimeraX, users can place markers on regions that may be potential domains.

Then they can use the Color Zone tool to color the regions around these markers and the *volume splitbyzone* command to segment the densities. Alternatively, the Segment Map tool[36] can be used for automatic map segmentation, allowing users to select and refine densities of interest.

Domain parsing

The AlphaFold2 predicted models and PAE files can be downloaded from AFDB using the *fetch_pdb_pae.py* script from DomainSeeker with an input list of UniProt IDs for the candidate proteins.

Then, DomainSeeker provides a python script *parse_with_pae.py* to divide the candidate proteins into domains. It groups the highly connected cliques, derived from the PAE matrices, into domains. We recommend setting the minimum number of residues in a domain to 40.

Fitting

ChimeraX[31] and COLORES[37] can both be used for fitting. The *fit_with_chimerax.py* script in DomainSeeker uses the *fitmap* command in ChimeraX to fit the input domains into the target densities. ChimeraX fits each domain within the density based on a fixed number of randomly generated initial search positions. For the test set of this work, an initial search number of 200 in ChimeraX is sufficient to find the correct fitting position and orientation. Users need to adjust this number corroding to the map size. Additionally, users should type in an option to specify the level of the density map above which the main volume voxels are located.

Scoring and ranking

The *score_rank.py* script calculate the volume and correlation of the overlapping regions between the input fitted domains and the target density. It generates a two-dimensional plot of volume versus correlation of the overlapping regions and calculates the local z-scores for each domain. Domains that are most likely to be accurate typically stand out in the 2D plot, and the ranking based on local z-scores should be used as a reference.

The expression of the local z-score is

$$z_i = \frac{c_i - \mu_i}{s_i}$$

where c_i is the overlap correlation of the candidate domain i, μ_i is the average score of the domains with similar overlapping areas, and s_i is the corresponding standard deviation.

Code availability

The code for DomainFit is available at https://github.com/zyzhangGroup/DomainSeeker.

Acknowledgments

This work is supported by the National Key Research and Development Program of

China (2021YFA1301504), the Chinese Academy of Sciences Strategic Priority Research Program (XDB37040202), and the National Natural Science Foundation of China (91953101). We are grateful to Mr. Yundong Zhang for technical support.

References

1. Moving structural biology forward together. *Cell* **187**, 511-512 (2024).
2. Bauerlein FJB, Baumeister W. Towards Visual Proteomics at High Resolution. *J. Mol. Biol.* **433**, 167187 (2021).
3. Beck M, Covino R, Hanelt I, Muller-McNicoll M. Understanding the cell: Future views of structural biology. *Cell* **187**, 545-562 (2024).
4. Doerr A. Cryo-electron tomography. *Nat. Methods* **14**, 34-34 (2017).
5. Toro-Nahuelpan M, Zagoriy I, Senger F, Blanchoin L, Thery M, Mahamid J. Tailoring cryo-electron microscopy grids by photo-micropatterning for in-cell structural studies. *Nat. Methods* **17**, 50-54 (2020).
6. Turk M, Baumeister W. The promise and the challenges of cryo-electron tomography. *FEBS Lett.* **594**, 3243-3261 (2020).
7. Strack R. Visual proteomics. *Nat. Methods* **20**, 1868-1868 (2023).
8. Klykov O, Kopylov M, Carragher B, Heck AJR, Noble AJ, Scheltema RA. Label-free visual proteomics: Coupling MS- and EM-based approaches in structural biology. *Mol. Cell* **82**, 285-303 (2022).
9. Nickell S, Kofler C, Leis AP, Baumeister W. A visual approach to proteomics. *Nat. Rev. Mol. Cell Biol.* **7**, 225-230 (2006).
10. Renaud JP, *et al.* Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471-492 (2018).
11. Cheng Y. Single-particle cryo-EM—How did it get here and where will it go. *Science* **361**, 876-880 (2018).
12. Hong Y, Song Y, Zhang Z, Li S. Cryo-Electron Tomography: The Resolution Revolution and a Surge of In Situ Virological Discoveries. *Annu. Rev. Biophys. Biomol. Struct.* **52**, 339-360 (2023).
13. Beckers M, Mann D, Sachse C. Structural interpretation of cryo-EM image reconstructions. *Prog. Biophys. Mol. Biol.* **160**, 26-36 (2021).
14. Chen Z, *et al.* De novo protein identification in mammalian sperm using in situ cryoelectron tomography and AlphaFold2 docking. *Cell* **186**, 5041-5053 e5019 (2023).
15. Tai L, Yin G, Huang X, Sun F, Zhu Y. In-cell structural insight into the stability of sperm microtubule doublet. *Cell Discov.* **9**, 116 (2023).
16. Gao J, Tong M, Lee C, Gaertig J, Legal T, Bui KH. DomainFit: Identification of protein domains in cryo-EM maps at intermediate resolution using AlphaFold2-predicted models. *Structure* **32**, 1248-1259.e1245 (2024).
17. Kaklamanos G, Aprea E, Theodoridis G. Mass spectrometry: principles and instrumentation. In: *Chemical Analysis of Food*) (2020).
18. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinform.* **54**, 5.6.1-5.6.37 (2016).

19. Jumper J, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
20. Abramson J, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
21. Baek M, *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
22. Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679-682 (2022).
23. Varadi M, *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368-D375 (2024).
24. Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct Target Ther.* **8**, 115 (2023).
25. Agarwal V, McShan AC. The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. *Nat. Chem. Biol.* **20**, 950-959 (2024).
26. Guo HB, *et al.* AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci. Rep.* **12**, 10696 (2022).
27. Akdel M, *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056-1067 (2022).
28. Eblen JD, Phillips CA, Rogers GL, Langston MA. The maximum clique enumeration problem: algorithms, applications, and implementations. *BMC Bioinformatics* **13**, S5 (2012).
29. Valiente G. Clique, Independent Set, and Vertex Cover. In: *Algorithms on Trees and Graphs: With Python Code* (ed Valiente G). Springer International Publishing (2021).
30. Chacón P, Wriggers W. Multi-resolution contour-based fitting of macromolecular structures11Edited by W. Baumeister. *J. Mol. Biol.* **317**, 375-384 (2002).
31. Meng EC, *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
32. Tunyasuvunakool K, *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596 (2021).
33. Oeffner RD, *et al.* Putting AlphaFold models to work with phenix.process_predicted_model and ISOLDE. *Acta Crystallogr. D: Struct. Biol.* **78**, 1303-1314 (2022).
34. Liebschner D, *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D: Struct. Biol.* **75**, 861-877 (2019).
35. Zhou L, *et al.* Structures of sperm flagellar doublet microtubules expand the genetic spectrum of male infertility. *Cell* **186**, 2897-2910 e2819 (2023).
36. Pintilie GD, Zhang J, Goddard TD, Chiu W, Gossard DC. Quantitative

analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* **170**, 427-438 (2010).

37. Wriggers W, Milligan RA, McCammon JA. Situs: A Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy. *J. Struct. Biol.* **125**, 185-195 (1999).