Bioinformatics, 2025, **41(11)**, btaf588
https://doi.org/10.1093/bioinformatics/btaf588
Advance Access Publication Date: 27 October 2025
Applications Note

OXFORD

Structural bioinformatics

AFragmenter: schema-free, tuneable protein domain segmentation for AlphaFold protein structures

Stefaan Verwimp[1,2], Rob Lavigne[2], Cédric Lood[3,*], Vera van Noort[1,4,*]

[1]Department of Microbial and Molecular Systems, KU Leuven, Leuven 3001, Belgium
[2]Department of Biosystems, KU Leuven, Leuven 3001, Belgium
[3]Department of Biology, University of Oxford, Oxford OX1 3SZ, United Kingdom
[4]Institute of Biology Leiden, Leiden University, Leiden 2333 BE, The Netherlands

*Corresponding authors. Vera van Noort, KU Leuven, Department of Microbial and Molecular Systems, Kasteelpark Arenberg 20, Leuven 3001, Belgium. E-mail: vera.vannoort@kuleuven.be; Cédric Lood, University of Oxford, Department of Biology, Life and Mind Building, South Parks Road, Oxford OX1 3EL, United Kingdom. E-mail: cedric.lood@biology.ox.ac.uk.

Associate Editor: Jianlin Cheng

Abstract

Summary: Protein domain segmentation is a crucial aspect of understanding protein functions and interactions, and it is vital for protein modelling exercises and evolutionary studies. Current segmentation methods often rely on predefined classification schemes, leading to inconsistencies and biases. AFragmenter provides a schema-free and tuneable approach to protein domain segmentation based on network analysis of AlphaFold-predicted structures. Utilizing Predicted Aligned Error values, AFragmenter constructs a fully connected network of protein residues and identifies distinct structural domains by using Leiden clustering. This method empowers users to adjust parameters including contrast threshold and resolution, providing control over the segmentation process.

Availability and implementation: AFragmenter is implemented in Python3 and freely available under an MIT license. It can be found as a Python library and command line tool at https://github.com/sverwimp/AFragmenter, pip, and Conda.

1 Introduction

Protein domains are generally defined as self-stabilizing regions that can support independent biological functions. However, this definition only captures one of several perspectives, as partitioning protein structures into domains can be based on several criteria, such as evolutionary conservated motifs, dynamically coupled segments, or spatial separation (Dawson *et al.* 2017, Postic *et al.* 2017). Regardless of the framing, delineating these domains is important to studying their functions and supporting applications such as enzyme engineering, drug design, and comparative genomics (Poleszak *et al.* 2012, Tong *et al.* 2022).

This multi-criteria nature of domain definition has resulted in discrepancies in domain assignment across different studies, as researchers may interpret structural features and functional roles differently (Taylor 1999, Veretnik *et al.* 2004). These inconsistencies have propagated into the various domain databases, where identical queries return different results. For example, MET8 (UniProt: P15807, PDB: 1KYQ) is classified as a three-domain protein in both CATH (Knudsen and Wiuf 2010) and ECOD (Cheng *et al.* 2014). In contrast, SCOPe (Chandonia *et al.* 2022) categorizes it as a two-domain protein, while InterPro (Paysan-Lafosse *et al.* 2023) contains two domains but also includes an additional unintegrated Pfam domain (Mistry *et al.* 2021). Meanwhile, SCOP (Lo Conte *et al.* 2000) considers MET8 to be a single-domain protein. Such discrepancies across classification systems highlight the inherent complexity of defining protein domains. Real-world protein structures often exhibit diverse and irregular domain architectures, including discontinuous domains that are not contiguous in the amino acid sequence (Dawson *et al.* 2017, Wang *et al.* 2021). These complexities make domain segmentation, the process of identifying and separating protein domains, a particularly challenging task.

Efforts have been made to create and improve domain segmentation tools to alleviate manual work from researchers. At least fourteen new or improved methods have been published between 2019 and 2024 (Hong *et al.* 2019, Jiang *et al.* 2019, Shi *et al.* 2019, Eguchi and Huang 2020, Zheng *et al.* 2020, Mulnaes *et al.* 2021, Cretin *et al.* 2022, Mahmud *et al.* 2022, Wang *et al.* 2022, Lau *et al.* 2023, Yu *et al.* 2023, Zhang *et al.* 2023, Zhu *et al.* 2023, Wells *et al.* 2024). Among the newly published methods, ten out of fourteen are machine-learning based. Merizo (Lau *et al.* 2023) and Chainsaw (Wells *et al.* 2024) are two such methods and are among the best performing domain segmentation tools available, at the time of writing this paper, based on benchmarking results against data from CATH. The absence of definite standards on protein domains can lead to machine-learning based approaches to incorporate a bias towards the domain classification scheme used by their training data. Similarly, heuristic methods frequently suffer from this issue if their methods were fine-tuned using this same data. A second issue with the aforementioned referenced domain segmentation

Received: 19 March 2025; **Revised:** 27 September 2025; **Accepted:** 21 October 2025

methods is their lack of flexibility (see Table 1, available as supplementary data at *Bioinformatics* online). These tools do not allow the user to significantly influence the segmentation process of the protein structure. Although this inflexibility may not always pose a major problem, it can render these tools impractical when their results do not align with researchers' interpretations or appear illogical.

To address the bias towards certain domain segmentation schemes and the lack of flexibility, we developed a schema-free, flexible protein segmentation method, based on network clustering of AlphaFold structures. Our approach leverages AlphaFold's Predicted Aligned Error values, which captures predicted confidence in inter-residue alignment and thus provide a basis for identifying independent structural units within a protein. This method does not rely on any predefined classification scheme and allows users to fine-tune a limited set of parameters, enabling them to explore multiple partitionings of a protein structure. Our method aims to strike a balance between performance, user control, and ease of use.

2 Implementation

AFragmenter is an AlphaFold structure segmentation tool based on a network clustering approach. It constructs a fully connected network where the nodes represent the residues of a protein, and the weights are derived from the Predicted Aligned Error (PAE) values from AlphaFold. PAE values measure the estimated error of inter-residue distances, reflecting the confidence of the model in relative residue positions. High PAE between segments indicates lower confidence in their relative arrangement, suggesting they may behave as independent entities or flexible linkers. Conversely, low PAE signifies high confidence in a self-contained structural unit (Jumper *et al.* 2021, McCafferty *et al.* 2023, Roca-Martínez *et al.* 2024, Varadi *et al.* 2024). Recent studies also indicate AlphaFold models, including PAE, implicitly capture protein dynamics (Guo *et al.* 2022), a property frequently associated with distinct domain movements. Consequently, intra-domain residue pairs are expected to have lower PAE values compared to inter-domain pairs. This difference distinguishes well-structured regions within a protein structure. By using these values as edge weights for a network and applying Leiden clustering, AFragmenter clusters protein residue pairs of well-structured regions together, thus achieving protein structure segmentation.

2.1 Creation of the protein structure network

The python interface to the igraph library (Csárdi and Nepusz 2006) is used to construct the protein structure network and to perform Leiden clustering. Each residue of the protein structure is represented as a single node in the network, and the edges between these nodes are weighted based on PAE values. The weight between nodes i and j is calculated as follows: $w_{ij} = \frac{1}{1+e^{(PAE_{ij}-T)}}$, where T represents a 'contrast' threshold that serves as a soft cut-off to increase the contrast between low and high PAE values. This increased contrast leads to more distinct, better-defined clusters. The threshold T can be changed by the user as it is sometimes necessary when working with AlphaFold predictions that may be overly confident due to the inclusion of identical or very similar protein in the training data of the AlphaFold model. The aforementioned MET8 protein illustrates this issue, as shown in Fig. 1A and B, where the pLDDT and PAE values around all but one disordered segment is much better than we would typically expect. By adjusting the contrast threshold, the difference in inter- versus intra-domain PAE values between various structural regions can be more accurately captured, leading to a clearer distinction.

2.2 Clustering to segment the protein

The Leiden clustering algorithm is used to partition the weighted protein structure graph into distinct clusters, representing various segments of the protein. This clustering algorithm features a tuneable parameter known as the resolution parameter, which modulates the granularity of the clusters by balancing intra-cluster density against inter-cluster separation. Adjusting this parameter allows for control over the number of clusters: higher resolution values yield a greater number of smaller clusters, while lower values result in fewer, larger clusters (Traag *et al.* 2019). The resulting clusters are subsequently mapped back onto the protein structure, delineating the different structural partitions of the protein (see Fig. 1C).

Recommended parameter settings, based on benchmarking with 1000 randomly selected samples from the CATH and ECOD databases (see Supplementary Information, available as supplementary data at *Bioinformatics* online), are a resolution of 0.7 and a threshold of 2. These values serve as effective starting points for domain segmentation. Benchmarking involved calculating the Intersection over Union (IoU) across the AFragmenter parameter space, with the 10 best parameter combinations yielding mean IoU values of 0.71–0.75 for CATH and 0.71–0.72 for ECOD (see Fig. 1, available as supplementary data at *Bioinformatics* online). For initial exploration, resolution values in the range of 0.4–0.8 and threshold values between 0 and 4 are suggested. Higher values for either parameter tend to yield a larger number of smaller domains, while lower values result in fewer, broader domains. However, depending on the specific application, values outside these ranges may also be informative.

3 Discussion

Protein domain segmentation is a complex problem, often with multiple possible solutions. This can be explained by the absence of a standardized definition of what constitutes a protein domain. Our AFragmenter tool tackles two key limitations of existing methods: First, all but one of the tools surveyed return single solutions, and their uses may miss alternative interpretations (see Fig. 1D). Second, these tools are influenced by the specific domain segmentation scheme used by the data source used during training, optimization and evaluation. We suspect that there might be a bias towards protein domains from well-studied organisms, making the performance on proteins of lesser studied organisms less reliable due to the lack of ground truths for validation.

One existing tool that addresses the lack of flexibility is SWORD2 (Cretin *et al.* 2022) which returns multiple solutions through generating protein units and testing different merges. This is an elegant way to circumvent the schema issue, but it does not provide any level of control over the segmentation process. AFragmenter provides this control through the contrast threshold and resolution parameters. This distinction sets AFragmenter apart in terms of the control provided to the user.

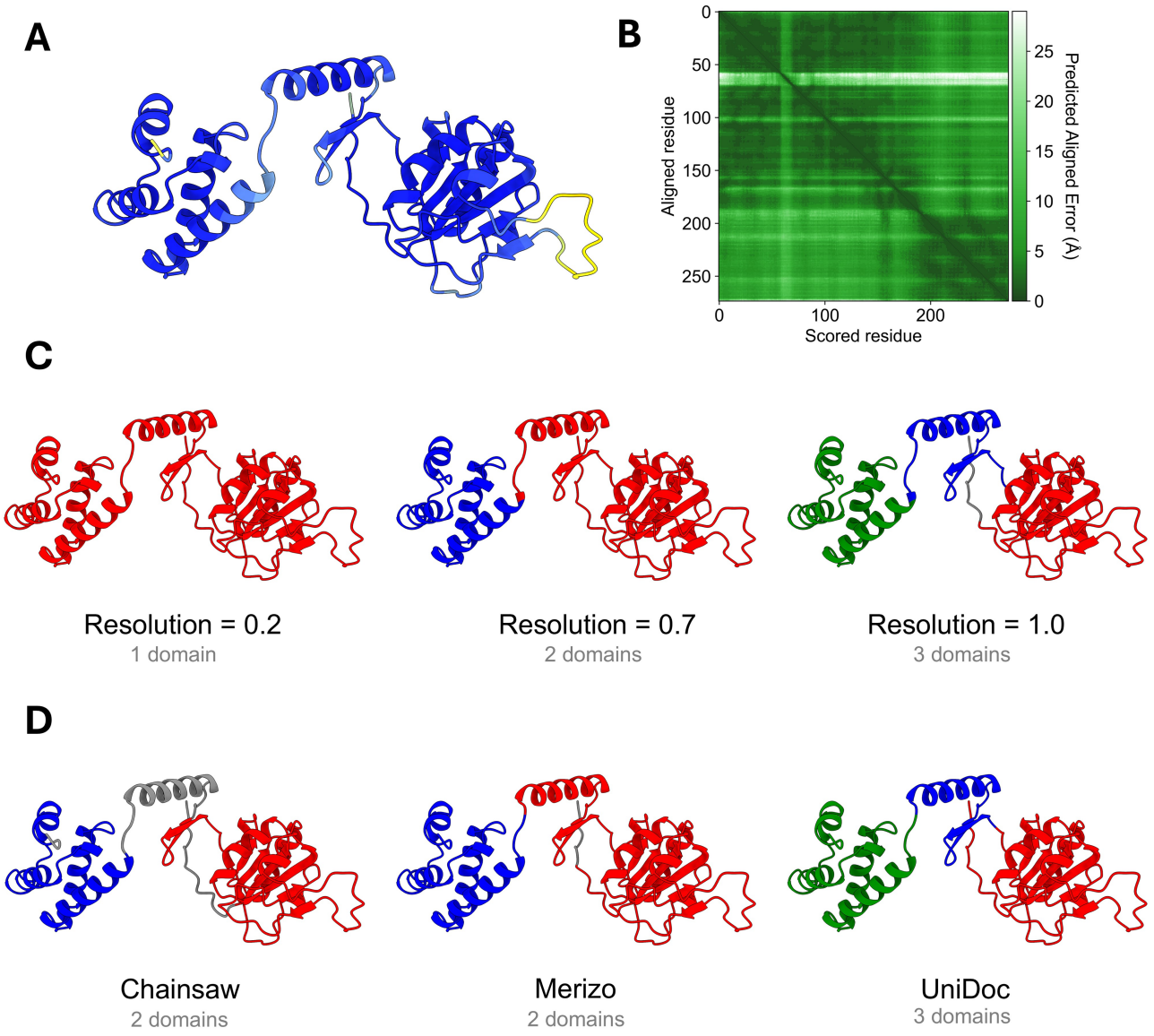

Figure 1. (A) The predicted AlphaFold protein structure of MET8 (P15807). Colours correspond to the pLDDT score values; dark blue indicates high confidence regions, while lighter colours represent lower confidence regions. (B) The PAE plot of MET8. One noticeable band of high PAE scores corresponds to the disordered segment with low pLDDT scores. (C) Segmentation results achieved by tuning the resolution parameter of AFragmenter (other parameters default). (D) A comparison of different domain segmentation methods for the AlphaFold structure of MET8. Domains are coloured red, blue, and green. Grey indicates residues that are not assigned to any domain.

There are currently two tools that have used AlphaFold structures in their method design, namely Merizo and DPAM (Zhang *et al.* 2023). Merizo used AlphaFold structures to fine-tune its machine-learning model. DPAM uses PAE values to calculate probabilities of residue pairs belonging to the same domain. However, these PAE derived probabilities only account for one tenth of the final scoring, and much more of the final score is dependent on sequence- and structure-based hits to known domains in ECOD.

The clustering approach implemented within AFragmenter is not entirely unique, as a similar method is used by the 'AlphaFold predicted Alignment Error' tool within the ChimeraX protein viewer software for colouring protein structure domains (Meng *et al.* 2023). However, AFragmenter differs by using a soft cut-off for enhanced contrast between low and high PAE values, and uses the more robust Leiden algorithm over the Clauset-Newman-Moor greedy algorithm (Blondel *et al.* 2008, Traag *et al.* 2019). Importantly, AFragmenter was designed for ease of use and does not necessitate a full GUI application like ChimeraX.

In individual cases, we observed that recent segmentation methods such as Chainsaw and Merizo perform well without parameter tuning, making them particularly well suited for processing large datasets without supervision. In contrast, AFragmenter is intended to provide a tuneable method that allows finer control over the segmentation. This user control can be valuable for investigating proteins where conventional domain definitions are ambiguous or where specific research questions necessitate alternative structural interpretations, contexts where a single, pre-defined domain solution may not suffice. Such tunability enables users to generate interpretations aligned with specific biological hypotheses or to investigate domain organizations in less-characterized protein systems.

In summary, AFragmenter offers a flexible and schema-free approach to protein structure segmentation by leveraging AlphaFold's PAE scores and graph theory. By avoiding reliance on predefined segmentation schemas and incorporating user-adjustable parameters, our method provides researchers with an easy-to-use, exploratory, and tuneable strategy for semi-automatic segmentation of protein structures. This capability is particularly beneficial in research areas where standard domain classification may not fully capture functional or evolutionary nuances, offering a complementary and exploratory perspective to database-driven domain assignment.

Author contributions

Stefaan Verwimp (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Rob Lavigne (Funding acquisition [equal], Writing—review & editing [equal]), Cédric Lood (Supervision [equal], Writing—review & editing [equal]), and Vera van Noort (Conceptualization [equal], Funding acquisition [equal], Supervision [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by KU Leuven project 'PHAGEFORCE' [ID-N(20/24)].

Data availability

The data underlying this article and the benchmarking are available on Github at https://github.com/sverwimp/AFragmenter and on Zenodo at https://doi.org/10.5281/zenodo.17215723.

References

Blondel VD, Guillaume J-L, Lambiotte R *et al.* Fast unfolding of communities in large networks. *J Stat Mech* 2008;**2008**:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Chandonia J-M, Guan L, Lin S *et al.* SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Res* 2022;**50**:D553–9. https://doi.org/10.1093/nar/gkab1054

Cheng H, Schaeffer RD, Liao Y *et al.* ECOD: an evolutionary classification of protein domains. *PLoS Comput Biol* 2014;**10**:e1003926. https://doi.org/10.1371/journal.pcbi.1003926

Cretin G, Galochkina T, Vander Meersche Y *et al.* SWORD2: hierarchical analysis of protein 3D structures. *Nucleic Acids Res* 2022;**50**:W732–8. https://doi.org/10.1093/nar/gkac370

Csárdi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Syst* 2006;**1695**.

Dawson N, Sillitoe I, Marsden RL *et al.* The classification of protein domains. In: Keith JM (ed.), *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*. New York, NY: Springer, 2017, 137–64. https://doi.org/10.1007/978-1-4939-6622-6_7

Eguchi RR, Huang P-S. Multi-scale structural analysis of proteins by deep semantic segmentation. *Bioinformatics* 2020;**36**:1740–9. https://doi.org/10.1093/bioinformatics/btz650

Guo H-B, Perminov A, Bekele S *et al.* AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* 2022;**12**:10696. https://doi.org/10.1038/s41598-022-14382-9

Hong SH, Joo K, Lee J. ConDo: protein domain boundary prediction using coevolutionary information. *Bioinformatics* 2019;**35**:2411–7. https://doi.org/10.1093/bioinformatics/bty973

Jiang Y, Wang D, Xu D. DeepDom: Predicting protein domain boundary from sequence alone using stacked bidirectional LSTM. *Pac Symp Biocomput* 2019;**24**:66–75.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2

Knudsen M, Wiuf C. The CATH database. *Hum Genomics* 2010;**4**:207–12. https://doi.org/10.1186/1479-7364-4-3-207

Lau AM, Kandathil SM, Jones DT. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. *Nat Commun* 2023;**14**:8445. https://doi.org/10.1038/s41467-023-43934-4

Lo Conte L, Ailey B, Hubbard TJP *et al.* SCOP: a structural classification of proteins database. *Nucleic Acids Res* 2000;**28**:257–9.

Mahmud S, Guo Z, Quadir F *et al.* Multi-head attention-based U-Nets for predicting protein domain boundaries using 1D sequence features and 2D distance maps. *BMC Bioinformatics* 2022;**23**:283. https://doi.org/10.1186/s12859-022-04829-1

McCafferty CL, Pennington EL, Papoulas O *et al.* Does AlphaFold2 model proteins' intracellular conformations? An experimental test using cross-linking mass spectrometry of endogenous ciliary proteins. *Commun Biol* 2023;**6**:421. https://doi.org/10.1038/s42003-023-04773-7

Meng EC, Goddard TD, Pettersen EF *et al.* UCSF ChimeraX: tools for structure building and analysis. *Protein Sci* 2023;**32**:e4792. https://doi.org/10.1002/pro.4792

Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9. https://doi.org/10.1093/nar/gkaa913

Mulnaes D, Golchin P, Koenig F *et al.* TopDomain: exhaustive protein domain boundary metaprediction combining multisource information and deep learning. *J Chem Theory Comput* 2021;**17**:4599–613. https://doi.org/10.1021/acs.jctc.1c00129

Paysan-Lafosse T, Blum M, Chuguransky S *et al.* InterPro in 2022. *Nucleic Acids Res* 2023;**51**:D418–27. https://doi.org/10.1093/nar/gkac993

Poleszak K, Kaminska KH, Dunin-Horkawicz S *et al.* Delineation of structural domains and identification of functionally important residues in DNA repair enzyme exonuclease VII. *Nucleic Acids Res* 2012;**40**:8163–74. https://doi.org/10.1093/nar/gks547

Postic G, Ghouzam Y, Chebrek R *et al.* An ambiguity principle for assigning protein structural domains. *Sci Adv* 2017;**3**:e1600552. https://doi.org/10.1126/sciadv.1600552

Roca-Martínez J, Kang H-S, Sattler M *et al.* Analysis of the inter-domain orientation of tandem RRM domains with diverse linkers: connecting experimental with AlphaFold2 predicted models. *NAR Genom Bioinform* 2024;**6**:lqae002. https://doi.org/10.1093/nargab/lqae002

Shi Q, Chen W, Huang S *et al.* DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network. *Bioinformatics* 2019;**35**:5128–36. https://doi.org/10.1093/bioinformatics/btz464

Taylor WR. Protein structural domain identification. *Protein Eng Des Select* 1999;**12**:203–16. https://doi.org/10.1093/protein/12.3.203

Tong CL, Kanwar N, Morrone DJ *et al.* Nature-inspired engineering of an artificial ligase enzyme by domain fusion. *Nucleic Acids Res* 2022;**50**:11175–85. https://doi.org/10.1093/nar/gkac858

Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:5233. https://doi.org/10.1038/s41598-019-41695-z

Varadi M, Bertoni D, Magana P *et al.* AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;**52**:D368–75. https://doi.org/10.1093/nar/gkad1011

Veretnik S, Bourne PE, Alexandrov NN *et al*. Toward consistent assignment of structural domains in proteins. *J Mol Biol* 2004;**339**:647–78. https://doi.org/10.1016/j.jmb.2004.03.053

Wang L, Zhong H, Xue Z *et al*. Res-Dom: predicting protein domain boundary from sequence using deep residual network and Bi-LSTM. *Bioinform Adv* 2022;**2**:vbac060. https://doi.org/10.1093/bioadv/vbac060

Wang Y, Zhang H, Zhong H *et al*. Protein domain identification methods and online resources. *Comput Struct Biotechnol J* 2021;**19**:1145–53. https://doi.org/10.1016/j.csbj.2021.01.041

Wells J, Hawkins-Hooker A, Bordin N *et al*. Chainsaw: protein domain segmentation with fully convolutional neural networks. *Bioinformatics* 2024;**40**:btae296. https://doi.org/10.1093/bioinformatics/btae296

Yu Z-Z, Peng C-X, Liu J *et al*. DomBpred: protein domain boundary prediction based on domain-residue clustering using inter-residue distance. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**:912–22. https://doi.org/10.1109/TCBB.2022.3175905

Zhang J, Schaeffer RD, Durham J *et al*. DPAM: a domain parser for AlphaFold models. *Protein Sci* 2023;**32**:e4548. https://doi.org/10.1002/pro.4548

Zheng W, Zhou X, Wuyun Q *et al*. FUpred: detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics* 2020;**36**:3749–57. https://doi.org/10.1093/bioinformatics/btaa217

Zhu K, Su H, Peng Z *et al*. A unified approach to protein domain parsing with inter-residue distance matrix. *Bioinformatics* 2023;**39**:btad070. https://doi.org/10.1093/bioinformatics/btad070

Bioinformatics, 2025, 41, 1–5
https://doi.org/10.1093/bioinformatics/btaf588
Applications Note