*Nucleic Acids Research*, 2024, **52**, W287–W293
https://doi.org/10.1093/nar/gkae328
Advance access publication date: 15 May 2024
**Web Server issue**

# Deep learning for the PSIPRED Protein Analysis Workbench

**Daniel W.A. Buchan *, Lewis Moffat, Andy Lau, Shaun M. Kandathil and David T. Jones**

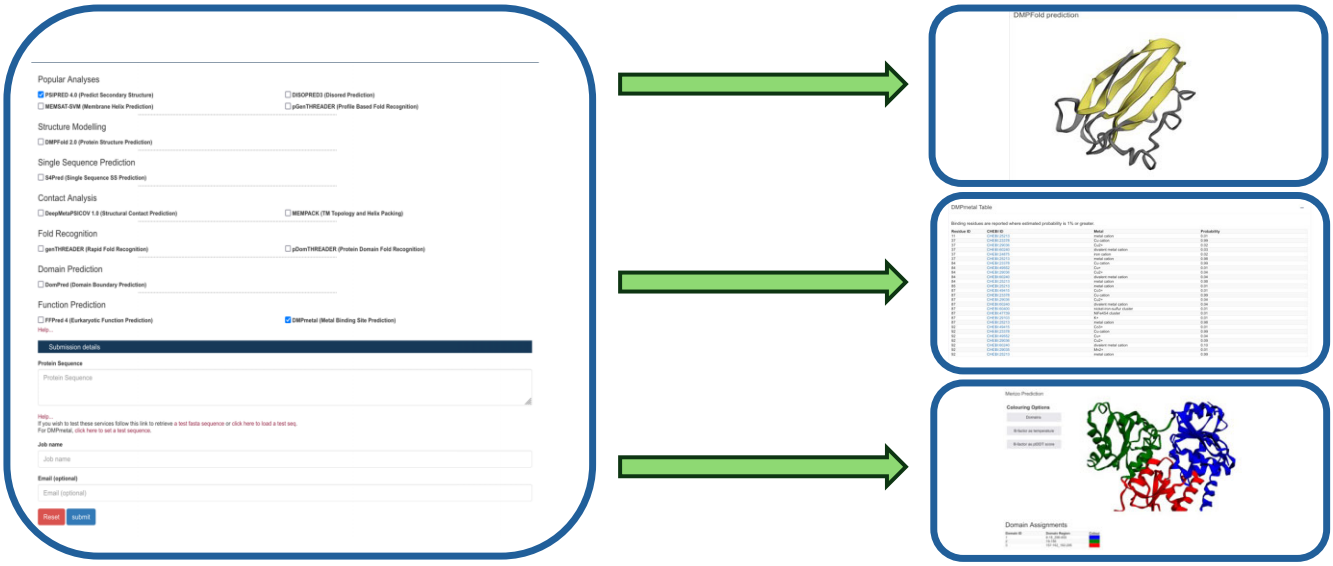
UCL Bioinformatics Group, Department of Computer Science, University College London, London, WC1E 6BT, UK

*To whom correspondence should be addressed. Tel: +44 20 7679 2000; Email: daniel.buchan@ucl.ac.uk

## Abstract

The PSIRED Workbench is a long established and popular bioinformatics web service offering a wide range of machine learning based analyses for characterizing protein structure and function. In this paper we provide an update of the recent additions and developments to the webserver, with a focus on new Deep Learning based methods. We briefly discuss some trends in server usage since the publication of AlphaFold2 and we give an overview of some upcoming developments for the service. The PSIPRED Workbench is available at http://bioinf.cs.ucl.ac.uk/psipred.

## Graphical abstract




## Introduction

The PSIPRED Workbench is part of a worldwide ecosystem of Bioscience data repositories and web services. These cover primary data repositories such as the NCBI, EBI and RCSB PDB (1–3), derived data resources such as STRING, CATH, KEGG, InterPro and UniProt (4–8), and webservices such as EBI Webservices, NCBI Webservices, among a great many others. A large number of tools and services available as code and webservices can be discovered via the Elixir BioTools web site (https://bio.tools/) (9).

We have been developing the PSIPRED Workbench for nearly 25 years. Our webservices offer a variety of machine learning-based tools focussed on characterising structural and functional features of proteins. In recent years, we have made significant headway in integrating new deep-learning based tools and techniques. In 2018, we replaced every line of code in our webserver and significantly improved both tool run times and presentation. Since then, we have seen peak annual usage rise to the order of 350 000 analyses per year.

Our services critically rely on underlying datasets from UniRef (10) and the PDB. Like all bioscience data resources, we have witnessed exponential growth in the size of these published datasets, which creates many computational challenges for bioinformatics tools and web servers. Many of our methods function by analysing evolutionary information in protein families, and protein database searching forms a critical first step in most of our tools. As the size of these resources grows, the runtimes for such analysis lengthens. To tackle this, we are increasingly looking to deep learning. Through careful model training, it is possible to embed protein sequence information such as evolutionary relationships between residues within the weights of a neural network (11,12). Consequently, we can use these and similar embeddings alongside novel deep learning-based methods and forgo the need for computationally

**Received:** February 26, 2024. **Revised:** April 8, 2024. **Editorial Decision:** April 11, 2024. **Accepted:** April 24, 2024

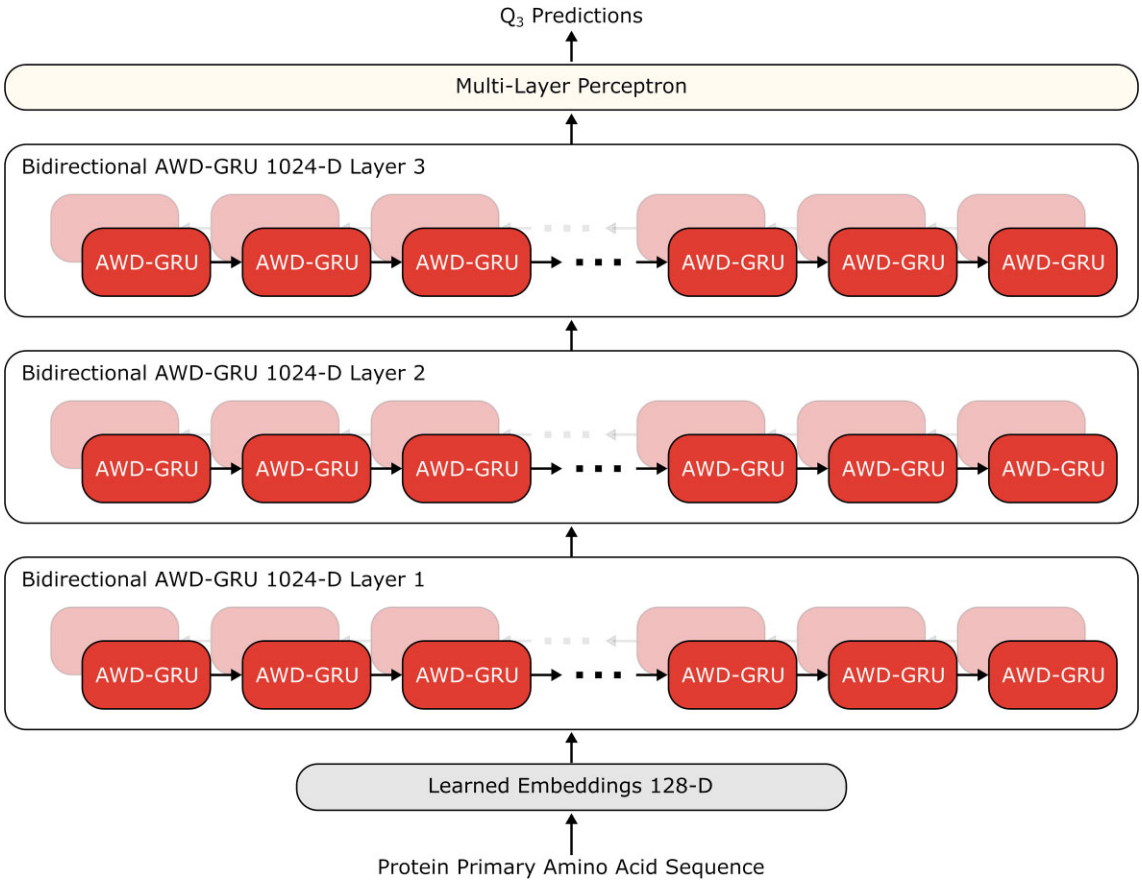



**Figure 1.** The neural network architecture of the S4PRED model. Single protein sequences are given as input to the model. Each amino acid in a given sequence is first dynamically replaced with a 128-dimensional vector embedding that is learned during training. The sequence of embeddings is then fed to a 1024-dimensional bidirectional recurrent neural network (RNN), termed the first layer. The specific RNN architecture used is the Averaged stochastic gradient descent Weight-Dropped Gated Recurrent Unit (AWD-GRU) (38,39). The RNN output is fed through two more layers and then a final multilayer perceptron which transforms the output to 3-dimensional probability predictions for each Q3 class.

expensive protein database searches while still producing accurate predictions of protein features that rely on evolutionary information.

## New methods

Since 2019, we have published a number of new methods in the UCL Bioinformatics Group and have made some of these available online via the PSIPRED Workbench. Below we give a summary of the methods we have added to the webserver.

## S4PRED

S4PRED (13) is a state-of-the-art single-sequence protein secondary structure prediction method. It is used to provide accurate secondary structure modelling for a challenging but important class of proteins, namely single orphan proteins, which have no detectable sequence relatives in current databases. Accordingly, the model takes only a protein's amino acid sequence as input, with no additional homology information, and subsequently returns 3-state secondary structure predictions for the sequence. Similarly to PSIPRED, S4PRED prediction results comprise a confidence score, a cartoon representation, 3-state prediction assignment, and the original amino acid sequence.

The model's architecture is an ensemble of five 3-layered recurrent deep neural networks (see Figure 1). It is trained using a semi-supervised learning approach to massively supplement the available number of protein sequences that can be trained on. This results in a training set in excess of a million examples. This set combines real-labelled examples, where a sequence and its secondary structure are known, and artificially labelled examples, where only the primary amino acid sequence is known. S4PRED has a Q3 secondary structure prediction accuracy of 75.3%. This is a significant improvement over our cutting edge PSIPRED method, which achieves a Q3 accuracy of 70.6% when tested on single sequences without any provided homology information. For secondary prediction tasks typical run times are of the order of seconds on contemporary CPUs.

## Merizo

Merizo is a deep learning-based method for protein domain segmentation (14). The method operates directly on structures and can produce accurate domain assignments even for discontinuous domains, as well as for predicted models from AlphaFold2 (15) which may feature long stretches of unstructured, non-domain residues.

The network of Merizo is based on an encoder-decoder architecture that utilises the invariant point attention module (introduced in AlphaFold2) to encode a structure and its sequence into an embedding. This embedding is then decoded using a Masked Transformer Decoder (16) to assign individ-

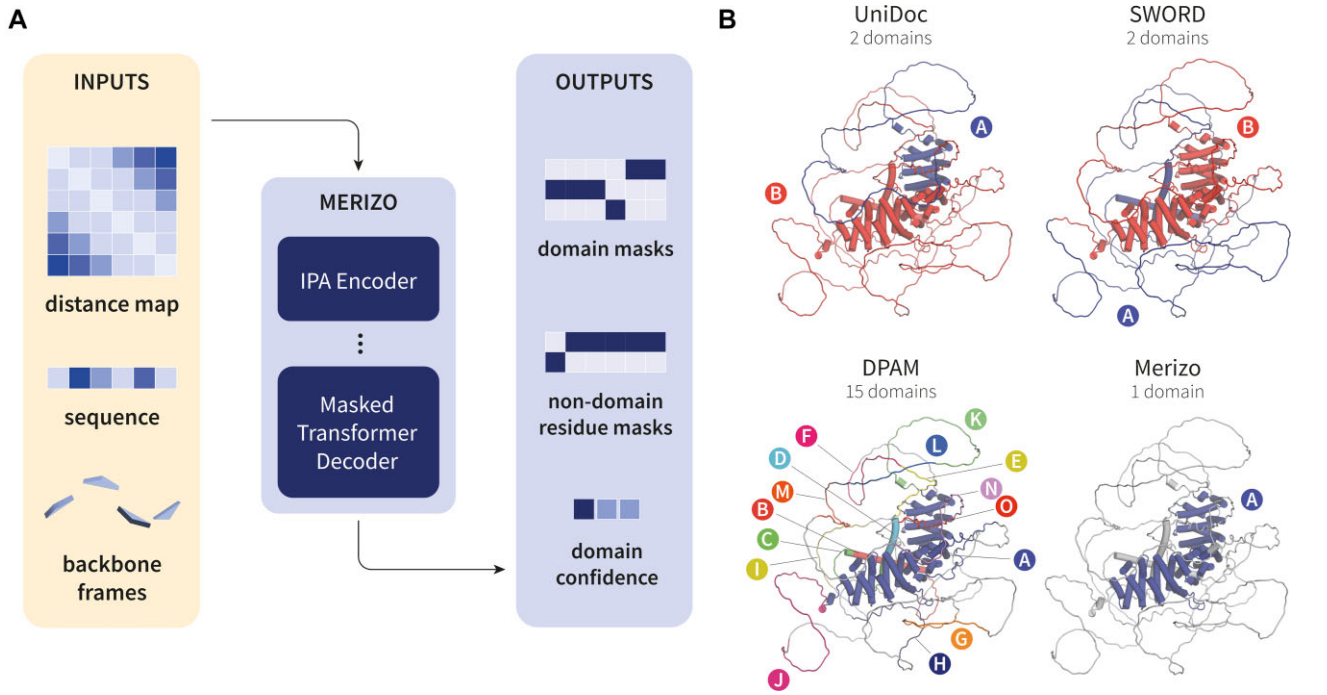



**Figure 2.** Overview of Merizo. (**A**) Summary of inputs and outputs. Merizo takes the C-alpha distance map along with the amino acid sequence and backbone frames (calculated as per Jumper *et al.*, 2021) as input. Inputs are fed into the IPA encoder which generates an embedding of the structure. The embedding is decoded by a masked transformer decoder to generate domain and non-domain residue masks, along with a confidence estimate for each predicted domain. (**B**) Example of domain assignments by several methods including Merizo on AlphaFold2 model AF-Q9UQB3-F1-model_v4. Methods include UniDoc (40), SWORD (41) and DPAM (42). Assigned domains are individually coloured and labelled from A-O. (Figure adapted from Lau *et al*, 14).

ual residues into domains in a bottom-up manner. Merizo is trained using an affinity learning strategy (17), wherein the network learns to cluster together embeddings of residues that belong to the same domain.

In a benchmark study on PDB structures, Merizo outperforms several state-of-the-art domain assignment methods, including both deep learning and non-deep learning-based methods, producing accurate assignments that are well-aligned with those documented in the CATH database. As a proof of concept, Merizo has also been applied to the human proteome, identifying over 40 000 domains that can be matched to known folds in CATH, while requiring only a fraction of the time needed by other methods.

Typical prediction times are around 1 second on modern CPUs and nearly an order of magnitude faster when using a GPU. Predictions are also shown to be highly accurate; Merizo achieves a median MCC score of approximately 1.0 when benchmarking predictions against known CATH or ECOD domain boundaries in multidomain proteins, and has a mean absolute error of ∼0.3 when predicting the number of domains within a chain correctly, outperforming other leading methods at this task. A brief overview of Merizo's inputs, outputs and illustrative performance to similar methods is given in Figure 2.

## DMPFold2

DMPfold2 (18) predicts the tertiary structure of single protein chains starting from amino acid sequence. It improves upon its predecessor DMPfold (19) in terms of both accuracy and speed of execution. The high speed of execution is enabled by a novel neural network architecture that takes as input a multiple sequence alignment (MSA) of the target protein sequence, and outputs the coordinates of C-alpha atoms of the main chain as direct outputs of the neural network. Alongside the coordinates, the network also predicts a per-residue confidence score. To predict the structure, the amino acids in the input MSA is first encoded as integers and then processed by a sequence of bidirectional Gated Recurrent Unit (biGRU) networks, the first operating on columns of the MSA to produce per-column representations. The second biGRU takes these representations as input and processes them in the horizontal direction to produce a final representation. This representation is combined along with a fast approximation of the residue precision matrix and is fed to a stack of residual convolutional layers. The output from this stack is then treated as a distance matrix and subjected to a differentiable multidimensional scaling procedure to recover the coordinates of the C-alpha atoms. The remaining main-chain atoms are added using the catomain procedure (20) and sidechain atoms can subsequently be added using tools such as SCWRL (21). Once a set of C-alpha coordinates have been generated, they can be converted into a pairwise distance map and used as an additional input to the network, and thus predictions can be recycled for iterative refinement. An overview of the method is presented in Figure 3.

Although not as accurate as AlphaFold2 and RoseTTAFold (22), DMPfold2 is orders of magnitude faster than these methods, and has considerably lower resource requirements. The former two methods require the use of GPU AI accelerators to achieve reasonable runtimes, however DMPfold2 is fast enough to be run on CPUs with runtimes ranging from sec-

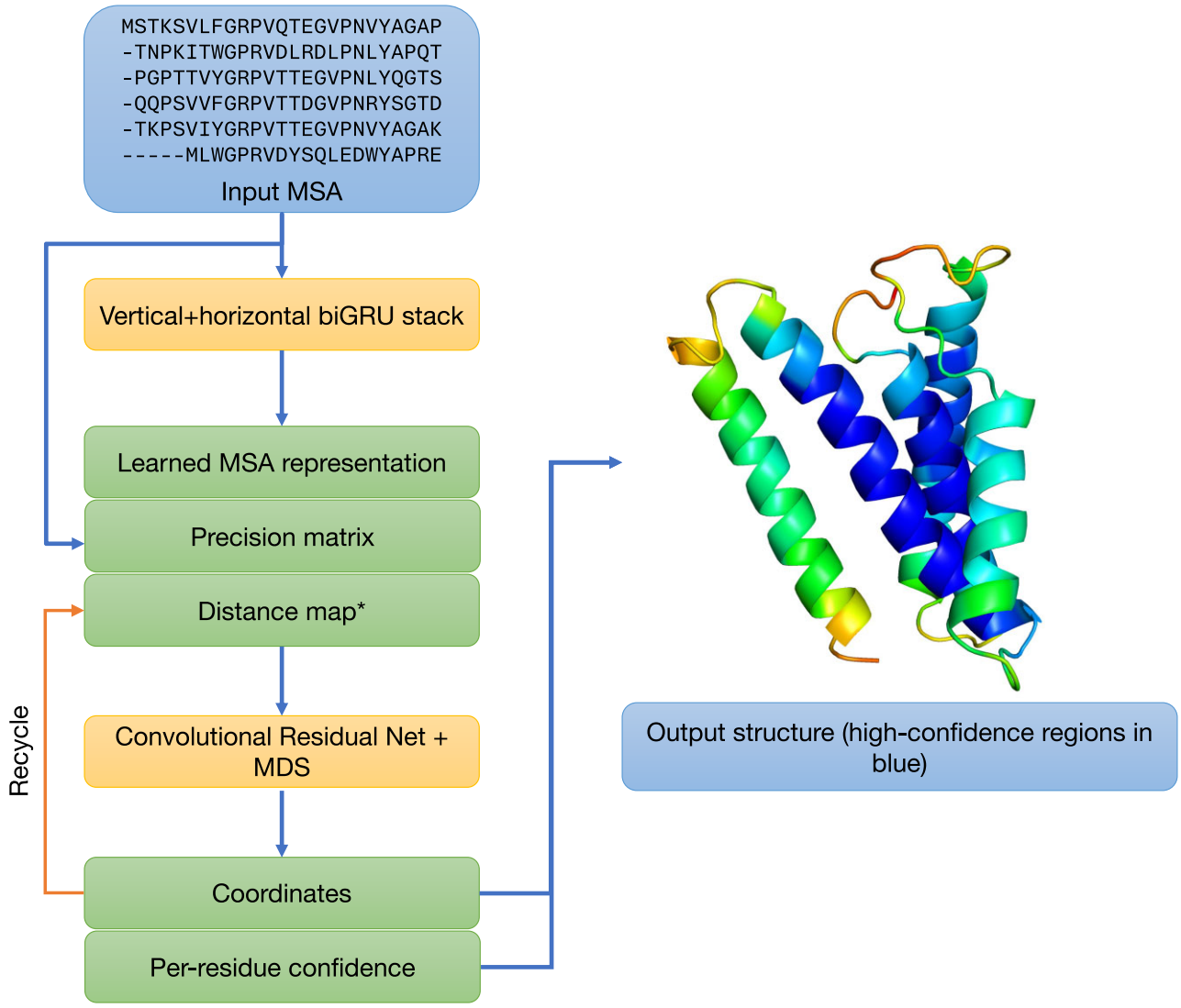



**Figure 3.** Overview of protein tertiary structure prediction using DMPfold2. *Coordinates produced at the end of the network can optionally be converted into a distance map and used to refine predictions in an iterative fashion. This distance map is zeroed out in the first iteration.

onds to a few minutes once the input MSA has been generated. On GPUs, DMPfold2 was shown to be roughly 2 orders of magnitude faster than AlphaFold2 in a head-to-head comparison.

### DMPmetal

DMPmetal is a deep learning-based method for predicting metal binding sites from amino acid sequences. It follows the approach of using a large (1.2 billion parameter) pre-trained transformer encoder protein language model (pLM) to embed the target sequences and to provide the features for simple feed-forward classifier. One difference from many other pLMs is that the DMPmetal pLM was jointly pre-trained on both sequence and structures through training on the UniRef50 subset of the AlphaFold Database (23). From a user perspective, the input to the model is a protein sequence, and the output probabilities relate to each of the 29 CHEBI metal codes. This model was ranked 1st in the UniProt Metal Binding Site Machine Learning Challenge held in 2022, and was trained on the organizers' provided NEG_TRAIN and POS_TRAIN_FULL datasets, based on curated UniProt annotations (http://insideuniprot.blogspot.com/2022/02/the-uniprot-metal-binding-site-machine.html).

### Available methods

The PSIPRED Workbench offers a number of analysis methods. We summarize these and their principal publication in Table 1.

## Retired methods

As science progresses, some of our older methods become obsolete. We now take the approach that prediction tools on our webserver which consistently see fewer than 1000 requests per year become candidates to be retired. We then assess these methods to establish if they have become obsolete; that is, they have either been replaced by a method within our group or have been made obsolete by other advances or tools

**Table 1.** Methods available via the PSIPRED workbench

| Method | Summary | Citation |
|---|---|---|
| PSIPRED 4.0 | Secondary structure prediction | Protein secondary structure prediction based on position-specific scoring matrices (24) |
| DISOPRED3 | Disordered residue prediction | DISOPRED3: precise disordered region predictions with annotated protein-binding activity (25) |
| MEMSAT-SVM | Membrane helix prediction | Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm (26) |
| GenTHREADER, pGenTHREADER & pDomTHREADER | Fold recognition | pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination (27) |
| DeepMetaPSICOV 1.0 | Structural contact prediction | Prediction of interresidue contacts with DeepMetaPSICOV in CASP13 (28) |
| DomPred | Protein domain boundary prediction | Computer-assisted protein domain boundary prediction using the DomPred server (29) |
| DMPFold 2.0 | Fast and Accurate Deep Learning Based protein structure prediciton | Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins (18) |
| FFPred3 | GO Term functional prediction | FFPred 3: feature-based function prediction for all Gene Ontology domains (30) |
| Metsite | Metal binding site prediction | Predicting metal-binding site residues in low-resolution structural mode (31) |
| HSPred | Protein-protein interaction hotspot prediction | Predictions of hot spot residues at protein-protein interfaces using support vector machines (32) |
| MEMEMBED | Membrane protein orientation prediction | Membrane protein orientation and refinement using a knowledge-based statistical potential (33) |
| Merizo | Deep Learning base structural domain segmentation | Merizo: a rapid and accurate protein domain segmentation method using invariant point attention (14) |
| S4PRED | Single Sequence Protein secondary Structure Prediction | Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework (13) |
| DMPmetal | Metal binding site prediction for protein sequences | Manuscript in preparation |

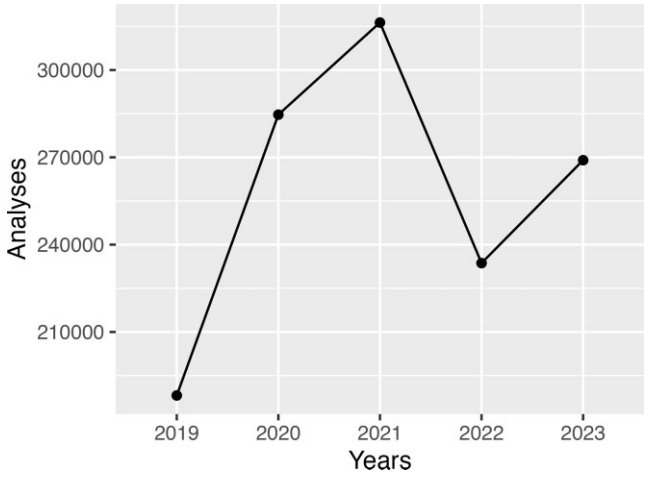



**Figure 4.** Total number of predictive analysis tasks run by the PSIPRED Workbench in the years 2019 to 2023. Y-axis is truncated.

that have emerged within protein bioinformatics. This year, we have chosen to retire our methods BioSerf and DomSerf, which were fully automated homology modelling packages which generated protein structure predictions. We see that these are amply superseded by the performance of methods such as DMPFold2 and AlphaFold2. Code for these methods will remain available in our public code repository.

## Trends In usage since 2018

Figure 4 shows the trends in usage of the PSIPRED web server since 2018, when our faster and more user-friendly web site was first launched. In the two years immediately following this, we saw substantial growth in the number of jobs submitted, aided no doubt by the increasing interest and use of novel bioinformatics tools in general during this time. However, in 2022, we saw a sharp decline in job counts, which we attribute to the availability of AlphaFold2 (15) and the associated AlphaFold structure database (23), which, at least in theory, would obviate the need for secondary structure and other predictions. Nevertheless, in 2023, submission counts for secondary structure prediction rebounded to their pre-2022 levels (see Figure 5). We suspect that over time, researchers became more familiar with the limitations of pre-calculated structure models (as observed by others, 34,35), and the additional difficulty in handling 3D structural data when only a protein sequence annotation is required. There clearly remains a demand for methods that can either corroborate the predictions made by structure modelling methods, or that can provide data that can be interpreted rapidly and more directly, for example in evaluating point mutants of some protein sequences.

## Site reliability and server developments

The principal focus of our web site development work since 2018 has been a new JavaScript front end code base. Our previous web site was implemented using the Ractive JavaScript framework (https://ractive.js.org/). This was an excellent choice in 2016 for rapidly prototyping our new web site, but as time passed and the site grew in complexity, the code became quite labyrinthine and hard to maintain. Since then, we have ported the entire website to React (https://react.dev/). We believe this gives a number of benefits; Re-


**Figure 5.** Bar chart of the proportion of predictive analyses jobs requested by user, coloured by predictive method. Data is broken down by year.

act is highly opinionated about application structure, and this should reduce any tendency towards writing unmaintainable, spaghetti code. React has also emerged as something of an industry standard for designing dynamic web applications so we anticipate that the framework will be well-supported for years to come.

Alongside this work on the web code we have recently upgraded the web server hardware. The old web server hardware had reached end-of-life and has been replaced with three new modern server machines. Alongside this, we will be adding a further data processing machine with 48 cores. This should provide sufficient additional capacity for the web site's developments in the years to come.

In our prior 2018/2019 web server publication (36), we replaced the entire code base for the website and installed new data analysis pipeline middleware. Since then, we are pleased to report that the webserver has experienced no downtime due to software failures. All server downtime has been due to scheduled hardware maintenance or unplanned hardware failures, such as power outages. Both our webserver and middleware code display excellent reliability with little need for ongoing maintenance. However, we do see a number of analyses fail. Sometimes our predictive methods have bugs or perhaps cannot handle certain edge cases; and on occasion users are able to submit erroneous input data. Nevertheless, the number of such failed jobs is small, at only around 3000 failures per year, typically <1% of all analyses each year.

## Discussion

We've reviewed in this paper some recent updates to our web services. Looking to the future, with the advent of AlphaFold2 and accurate structural modelling, we anticipate that a structural approach to protein bioinformatics will become increasingly common. With this in mind, our future developments for the service will focus on providing a novel 'structure-first' view to help integrate both structural predictions and sequence annotations in manner that makes it easy for researchers make sense of the protein sequences they are working with.

The PSIPRED workbench remains a popular and well-used bioinformatics resource for researchers across the globe. In acknowledgement of the impact and importance of our web server, the site was accepted as an Elixir Web Resource as part of the Elixir UK node in 2019 (37). This enables us to take part in the coordination of services and life science research across the UK and Europe. This will help us to continue to develop and fund the service in the years to come.

## Data availability

The web server is available at http://bioinf.cs.ucl.ac.uk/psipred/. Our principal code repository is available at https://github.com/psipred.

## Acknowledgements

We thank the following colleagues and collaborators for contributing code and methods to the server: Michael Sadowski, Sean Ward, Domenico Cozzetto, Tim Nugent, Russel Marsden, Kevin Bryson, Liam McGuffin, Jaz Sodhi, Cen Wan, Joe Greener, Federico Minneci, Stephano Lise.

## Funding

Funding for the PSIPRED Workbench is provided by the UK's Biotechnology and Biological Sciences Research Council under BBSRC [BB/T019379/1]. Funding for open access charge: UCL.

## Conflict of interest statement

None declared.

## References

1. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chao,H., Chen,L., Craig,P.A., Crichlow,G.V., Dalenberg,K., Duarte,J.M., *et al.* (2023) RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.*, **51**, D488–D508.
2. Madeira,F., Pearce,M., Tivey,A.R.N., Basutkar,P., Lee,J., Edbali,O., Madhusoodanan,N., Kolesnikov,A. and Lopez,R. (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.*, **50**, W276–W279.
3. Sayers,E.W., Beck,J., Bolton,E.E., Brister,J.R., Chan,J., Comeau,D.C., Connor,R., DiCuccio,M., Farrell,C.M., Feldgarden,M., *et al.* (2024) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **52**, D33–D43.
4. Szklarczyk,D., Kirsch,R., Koutrouli,M., Nastou,K., Mehryary,F., Hachilif,R., Gable,A.L., Fang,T., Doncheva,N.T., Pyysalo,S., *et al.* (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.*, **51**, D638–D646.
5. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N., *et al.*

(2021) CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.
6. Kanehisa,M., Sato,Y. and Kawashima,M. (2022) KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci.*, **31**, 47–53.
7. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L., *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
8. Zaru,R., Orchard,S. and UniProt,C. (2023) UniProt tools: BLAST, align, peptide search, and ID mapping. *Curr. Protoc*, **3**, e697.
9. Ison,J., Rapacki,K., Menager,H., Kalas,M., Rydza,E., Chmura,P., Anthon,C., Beard,N., Berka,K., Bolser,D., *et al.* (2016) Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res.*, **44**, D38–D47.
10. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H. and UniProt,C. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
11. Elnaggar,A., Heinzinger,M., Dallago,C., Rehawi,G., Wang,Y., Jones,L., Gibbs,T., Feher,T., Angerer,C., Steinegger,M., *et al.* (2022) ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**, 7112–7127.
12. Rives,A., Meier,J., Sercu,T., Goyal,S., Lin,Z., Liu,J., Guo,D., Ott,M., Zitnick,C.L., Ma,J., *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2016239118.
13. Moffat,L. and Jones,D.T. (2021) Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. *Bioinformatics*, **37**, 3744–3751.
14. Lau,A.M., Kandathil,S.M. and Jones,D.T. (2023) Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. *Nat. Commun.*, **14**, 8445.
15. Jumper,J. and Hassabis,D. (2022) Protein structure predictions to atomic accuracy with AlphaFold. *Nat. Methods*, **19**, 11–12.
16. Strudel,R., Garcia,R., Laptev,I. and Schmid,C. (2021) In: *Segmenter: Transformer for Semantic Segmentation*.
17. Huang,W., Deng,S., Chen,C., Fu,X. and Xiong,Z. (2022) In: *Learning to Model Pixel-embedded Affinity for Homogeneous Instance Segmentation*.
18. Kandathil,S.M., Greener,J.G., Lau,A.M. and Jones,D.T. (2022) Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2113348119.
19. Greener,J.G., Kandathil,S.M. and Jones,D.T. (2019) Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.*, **10**, 3977.
20. Aszodi,A. and Taylor,W.R. (1994) Secondary structure formation in model polypeptide chains. *Protein Eng.*, **7**, 633–644.
21. Wang,Q., Canutescu,A.A. and Dunbrack,R.L. Jr (2008) SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.*, **3**, 1832–1847.
22. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
23. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
24. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
25. Jones,D.T. and Cozzetto,D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, **31**, 857–863.
26. Nugent,T. and Jones,D.T. (2010) Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Comput. Biol.*, **6**, e1000714.
27. Lobley,A., Sadowski,M.I. and Jones,D.T. (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, **25**, 1761–1767.
28. Kandathil,S.M., Greener,J.G. and Jones,D.T. (2019) Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins*, **87**, 1092–1099.
29. Bryson,K., Cozzetto,D. and Jones,D.T. (2007) Computer-assisted protein domain boundary prediction using the DomPred server. *Curr. Protein Pept. Sci.*, **8**, 181–188.
30. Cozzetto,D., Minneci,F., Currant,H. and Jones,D.T. (2016) FFPred 3: feature-based function prediction for all gene ontology domains. *Sci. Rep.*, **6**, 31865.
31. Sodhi,J.S., Bryson,K., McGuffin,L.J., Ward,J.J., Wernisch,L. and Jones,D.T. (2004) Predicting metal-binding site residues in low-resolution structural models. *J. Mol. Biol.*, **342**, 307–320.
32. Lise,S., Buchan,D., Pontil,M. and Jones,D.T. (2011) Predictions of hot spot residues at protein-protein interfaces using support vector machines. *PLoS One*, **6**, e16774.
33. Nugent,T. and Jones,D.T. (2013) Membrane protein orientation and refinement using a knowledge-based statistical potential. *BMC Bioinf.*, **14**, 276.
34. Thornton,J.M., Laskowski,R.A. and Borkakoti,N. (2021) AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.*, **27**, 1666–1669.
35. Jones,D.T. and Thornton,J.M. (2022) The impact of AlphaFold2 one year on. *Nat. Methods*, **19**, 15–20.
36. Buchan,D.W.A. and Jones,D.T. (2019) The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.*, **47**, W402–W407.
37. Larcombe,L., Hendricusdottir,R., Attwood,T.K., Bacall,F., Beard,N., Bellis,L.J., Dunn,W.B., Hancock,J.M., Nenadic,A., Orengo,C., *et al.* (2017) ELIXIR-UK role in bioinformatics training at the national level and across ELIXIR. *F1000Res*, **6**, 952.
38. Merity,S., Keskar,N.S. and Socher,R. (2017) Regularizing and optimizing LSTM language models. arXiv doi: https://arxiv.org/abs/1708.02182, 07 August 2017, preprint: not peer reviewed.
39. Cho,K., Van Merriënboer,B., Bahdanau,D. and Bengio,Y. (2014) On the properties of neural machine translation: encoder-decoder approaches. arXiv doi: https://arxiv.org/abs/1409.1259, 07 October 2014, preprint: not peer reviewed.
40. Zhu,K., Su,H., Peng,Z. and Yang,J. (2023) A unified approach to protein domain parsing with inter-residue distance matrix. *Bioinformatics*, **39**, btad070.
41. Postic,G., Ghouzam,Y., Chebrek,R. and Gelly,J.C. (2017) An ambiguity principle for assigning protein structural domains. *Sci. Adv.*, **3**, e1600552.
42. Zhang,J., Schaeffer,R.D., Durham,J., Cong,Q. and Grishin,N.V. (2023) DPAM: a domain parser for AlphaFold models. *Protein Sci.*, **32**, e4548.

**Received:** February 26, 2024. **Revised:** April 8, 2024. **Editorial Decision:** April 11, 2024. **Accepted:** April 24, 2024