Published online 17 May 2021 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* ***W535–W540***
https://doi.org/10.1093/nar/gkab354

PredictProtein - Predicting Protein Structure and Function for 29 Years

Michael Bernhofer[1,2,†], Christian Dallago[1,2,*,†], Tim Karl[1,†], Venkata Satagopam[3,4,†], Michael Heinzinger[1,2], Maria Littmann[1,2], Tobias Olenyi[1], Jiajun Qiu[1,5], Konstantin Schütze[1], Guy Yachdav[1], Haim Ashkenazy[6,7], Nir Ben-Tal[8], Yana Bromberg[9], Tatyana Goldberg[1], Laszlo Kajan[10], Sean O'Donoghue[11], Chris Sander[12,13,14], Andrea Schafferhans[1,15], Avner Schlessinger[16], Gerrit Vriend[17], Milot Mirdita[18], Piotr Gawron[3], Wei Gu[3,4], Yohan Jarosz[3,4], Christophe Trefois[3,4], Martin Steinegger[19,20], Reinhard Schneider[3,4] and Burkhard Rost[1,21,22,*]

[1]TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr 3, 85748 Garching/Munich, Germany, [2]TUM Graduate School CeDoSIA, Boltzmannstr 11, 85748 Garching, Germany, [3]Luxembourg Centre For Systems Biomedicine (LCSB), University of Luxembourg, Campus Belval, House of Biomedicine II, 6 avenue du Swing, L-4367 Belvaux, Luxembourg, [4]ELIXIR Luxembourg (ELIXIR-LU) Node, University of Luxembourg, Campus Belval, House of Biomedicine II, 6 avenue du Swing, L-4367 Belvaux, Luxembourg, [5]Department of Otolaryngology Head & Neck Surgery, The Ninth People's Hospital & Ear Institute, School of Medicine & Shanghai Key Laboratory of Translational Medicine on Ear and Nose Diseases, Shanghai Jiao Tong University, Shanghai, China, [6]Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany, [7]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, 69978 Tel Aviv, Israel, [8]Department of Biochemistry & Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, 69978 Tel Aviv, Israel, [9]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA, [10]Roche Polska Sp. z o.o., Domaniewska 39B, 02–672 Warsaw, Poland, [11]Garvan Institute of Medical Research, Sydney, Australia, [12]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA, [13]Department of Cell Biology, Harvard Medical School, Boston, MA 02215, USA, [14]Broad Institute of MIT and Harvard, Boston, MA 02142, USA, [15]HSWT (Hochschule Weihenstephan Triesdorf | University of Applied Sciences), Department of Bioengineering Sciences, Am Hofgarten 10, 85354 Freising, Germany, [16]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA, [17]BIPS, Poblacion Baco, Mindoro, Philippines, [18]Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, [19]School of Biological Sciences, Seoul National University, Seoul, South Korea, [20]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea, [21]Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany and [22]TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany

Received February 23, 2021; Revised April 06, 2021; Editorial Decision April 21, 2021; Accepted May 10, 2021

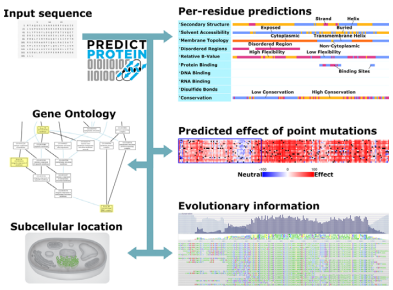
ABSTRACT

Since 1992 *PredictProtein* (https://predictprotein.org) is a one-stop online resource for protein sequence analysis with its main site hosted at the Luxembourg Centre for Systems Biomedicine (LCSB) and queried monthly by over 3,000 users in 2020. *PredictProtein* was the first Internet server for protein predictions. It pioneered combining evolutionary information and machine learning. Given a protein sequence as input, the server outputs multiple sequence alignments, predictions of protein structure in 1D and 2D (secondary structure, solvent accessibility, transmembrane segments, disordered regions, protein flexibility, and disulfide bridges) and predictions of protein function (functional effects of sequence variation or point mutations, Gene Ontology (GO) terms, subcellular localization, and

*To whom correspondence should be addressed. Tel: +49 289 17 811; Email: christian.dallago@tum.de
Correspondence may also be addressed to Burkhard Rost. Email: assistant@rostlab.org
†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

protein-, RNA-, and DNA binding). *PredictProtein's* infrastructure has moved to the LCSB increasing throughput; the use of MMseqs2 sequence search reduced runtime five-fold (apparently without lowering performance of prediction methods); user interface elements improved usability, and new prediction methods were added. *PredictProtein* recently included predictions from deep learning embeddings (GO and secondary structure) and a method for the prediction of proteins and residues binding DNA, RNA, or other proteins. PredictProtein.org aspires to provide reliable predictions to computational and experimental biologists alike. All scripts and methods are freely available for offline execution in high-throughput settings.

GRAPHICAL ABSTRACT



INTRODUCTION

The sequence is known for far more proteins (1) than experimental annotations of function or structure (2,3). This sequence-annotation gap existed when *PredictProtein* (4,5) started in 1992, and has kept expanding ever since (6). Unannotated sequences contribute crucial evolutionary information to neural networks predicting secondary structure (7,8) that seeded *PredictProtein (PP)* at the European Molecular Biology Laboratory (EMBL) in 1992 (9), the first fully automated, query-driven Internet server providing evolutionary information and structure prediction for any protein. Many other methods predicting aspects of protein function and structure have since joined under the PP roof (4,5,10) now hosted by the Luxembourg Centre of Systems Biomedicine (LCSB).

PP offers an array of structure and function predictions most of which combine machine learning with evolutionary information; now enhanced by a faster alignment algorithm (11,12). A few prediction methods now also use embeddings (13,14) from protein Language Models (LMs) (13–18). Embeddings are much faster to obtain than evolutionary information, yet for many tasks, perform almost as well, or even better (19,20). All PP methods join at PredictProtein.org with interactive visualizations; for some methods, more advanced visualizations are linked (21–23). The reliability of *PredictProtein*, its speed, the continuous integration of well-validated, top methods, and its intuitive interface have attracted thousands of researchers over 29 years of steady operation.

MATERIALS AND METHODS

PredictProtein (PP) provides

multiple sequence alignments (MSAs) and position-specific scoring matrices (PSSMs) computed by a combination of pairwise BLAST (24), PSI-BLAST (25), and MMseqs2 (11,12) on query vs. PDB (26) and query versus UniProt (1). For each residue in the query, the following per-residue predictions are assembled: secondary structure (RePROF/PROFsec (5,27) and ProtBertSec (14)); solvent accessibility (RePROF/PROFacc); transmembrane helices and strands (TMSEG (28) and PROFtmb (29)); protein disorder (Meta-Disorder (30)); backbone flexibility (relative B-values; PROFbval (31)); disulfide bridges (DISULFIND (32)); sequence conservation (ConSurf/ConSeq (33–36)); protein-protein, protein-DNA, and protein-RNA binding residues (ProNA2020 (3)); PROSITE motifs (37); effects of sequence variation (single amino acid variants, SAVs; SNAP2 (38)). For each query per-protein predictions include: transmembrane topology (TMSEG (28)); binary protein-(DNA|RNA|protein) binding (protein binds X or not; ProNA2020 (3)); Gene Ontology (GO) term predictions (goPredSim (19)); subcellular localization (LocTree3 (39)); Pfam (40) domain scans, and some biophysical features. Under the hood, PP computes more results (SOM: PredictProtein Methods; Supplementary Table S1), either as input for frontend methods, or for legacy support.

New: goPredSim embedding-based transfer of Gene Ontology (GO)

goPredSim (19) predicts GO terms by transferring annotations from the most embedding-similar protein. Embeddings are obtained from SeqVec (13); similarity is established through the Euclidean distance between the embedding of a query and a protein with experimental GO annotations. Replicating the conditions of CAFA3 (41) in 2017, goPredSim achieved F_{max} values of 37 ± 2%, 52 ± 2% and 58 ± 2% for BPO (biological process), MFO (molecular function), and CCO (cellular component), respectively (41,42). Using Gene Ontology Annotation (GOA) (43,44) to test 296 proteins annotated after February 2020, goPredSim appeared to reach even slightly higher values that were confirmed by CAFA4 (45).

New: ProtBertSec secondary structure prediction

ProtBertSec predicts secondary structure in three states (helix, strand, other) using ProtBert (14) embeddings derived from training on BFD with almost 3×10^9 proteins (6,46). On a hold-out set from CASP12, ProtBertSec reached a three-state per-residue accuracy of Q3 = 76 ± 1.5% (47). Although below the state-of-the-art (NetSurfP-2.0 (48) at 82%), this method performed on-par with other MSA-based methods, despite itself not using MSAs.

New: ProNA2020 protein–protein, protein–RNA and protein–DNA

ProNA2020 (3) predicts whether or not a protein interacts with other proteins, RNA or DNA (binary), and if so, where

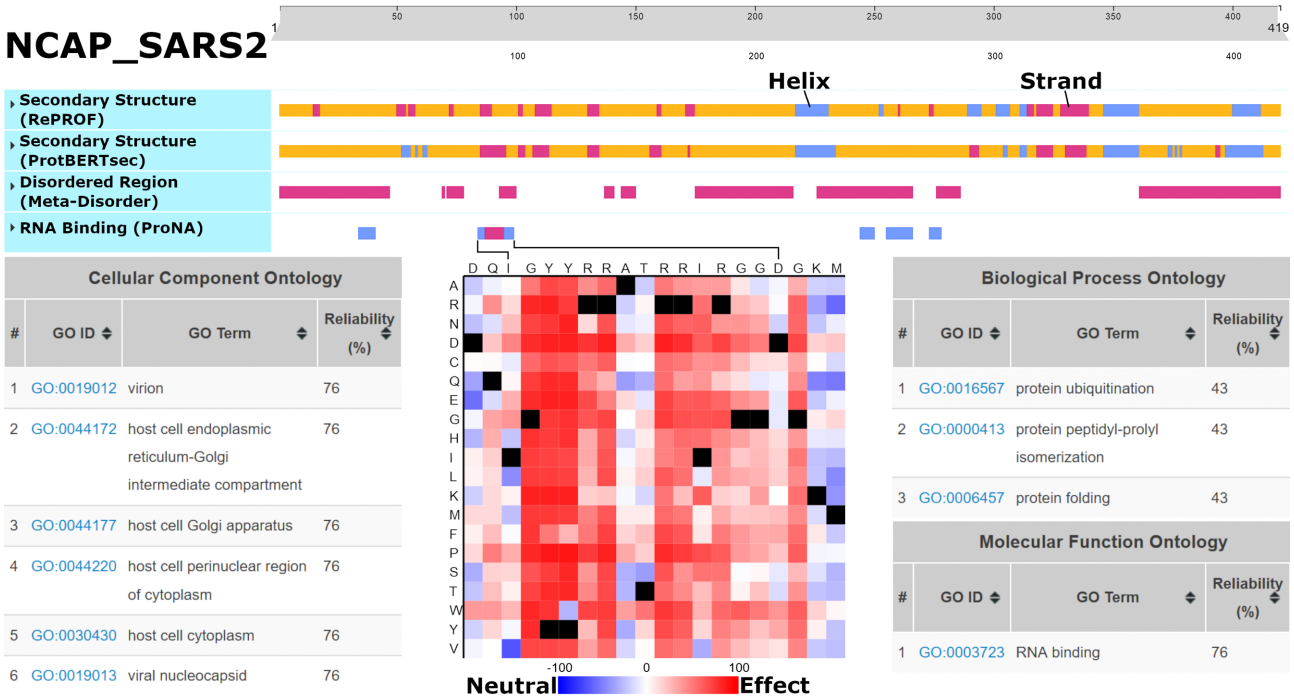


Figure 1. Predictions for SARS-CoV-2 Nucleoprotein (NCAP_SARS2). Underneath the interactive slider at the top: RePROF and ProtBertSec secondary structure (blue: helix; purple: strand; orange: other); Meta-Disorder intrinsically disordered regions (purple); ProNA2020 RNA-binding residues (low confidence: blue; medium confidence: purple). goPredSim transfers of GeneOntology (GO) terms based on embedding similarity (lower left: CCO; lower right: BPO & MFO). SNAP2 predicts the effect of point-mutations on function for the RNA-binding region from I84 to D98 (bottom-center; black: native residue). Link: predictprotein.org/visual_results?req_id=1nAmulUQY$FRPFaP8NTqLW9DzdlTG3B/.

it binds (which residues). The binary per-protein predictions rely on homology and machine learning models employing profile-kernel SVMs (49) and on embeddings from an *in-house* implementation of ProtVec (50). Per-residue predictions (where) use simple neural networks due to data shortage (51–53). ProNA2020 correctly predicted 77 ± 1% of the proteins binding DNA, RNA or protein. In proteins known to bind other proteins, RNA or DNA, ProNA2020 correctly predicted 69 ± 1%, 81 ± 1% and 80 ± 1% of binding residues, respectively.

New: MMseqs2 speedy evolutionary information

Most time-consuming for PP was the search for related proteins in ever growing databases. MMseqs2 (11) finds related sequences blazingly fast and seeds a PSI-BLAST search (25). The query sequence is sent to a dedicated MMseqs2 server that searches for hits against cluster representatives within the UniClust30 (54) and PDB (26) reduced to 70% pairwise percentage sequence identity (PIDE). All hits and their respective cluster members are turned into a MSA and filtered to the 3000 most diverse sequences.

WEB SERVER

Frontend and user interface (UI)

Users query PredictProtein.org by submitting a protein sequence. Results are available in seconds for sequences that had been submitted recently (cf. *PPcache* next section), or within up to 30 min if predictions are recomputed. Per-residue predictions are displayed online via ProtVista (55), which allows to zoom into any sequential protein region (Supplementary Figure S1), and are grouped by category (e.g. secondary structure), which can be expanded to display more detail (e.g. helix, strand, other). On the results page, links to visualize MSAs through *AlignmentViewer* (56) are available. More predictions can be accessed through a menu on the left, e.g. *Gene Ontology Terms*, *Effect of Point Mutations* and *Subcellular Localization*. Prediction views include references and details of outputs, as well as rich visualizations, e.g. GO trees for GO predictions and cell images with highlighted predicted locations for subcellular localization predictions (57).

PPcache, backend and programmatic access

The PP backend lives at LCSB, allowing for up to 48 parallel queries. Results are stored on disc in the *PPcache* (5). Users submitting sequences for which results were over the last two years obtain results immediately. Using the bio-embeddings pipeline (58), the PPcache is enriched by embeddings and embedding-based predictions on the fly. For all methods displayed on the frontend, JSON files compliant with *ProtVista* (55) are available via REST APIs (SOM: Programmatic access), and are in use by external services such as the protein 3D structure visualization suite *Aquaria* (21,23) and by *MolArt* (22).

PredictProtein is available for local use

All results displayed on and downloadable from PP are available through Docker (59) and as source code for local and cloud execution (available at github.com/rostlab).

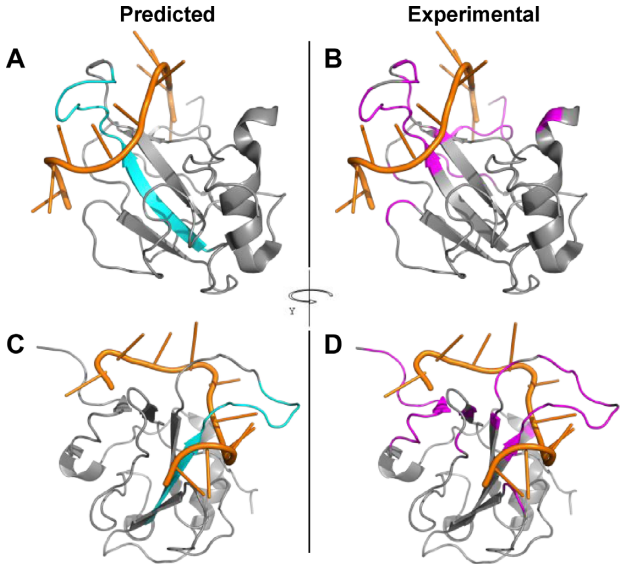


Figure 2. Experimental and predicted RNA-binding residues for NCAP2_SARS2. Predicted (via ProNA2020, in cyan, panels **A** and **C**) and observed (within 5Å, in magenta, panels **B** and **D**) RNA-binding residues for the SARS-CoV-2 nucleoprotein (gray) complexed with a 10-mer ssRNA (orange), PDB structure 7ACT (61). Two-third of the predictions are correct (precision = 0.73, recall = 0.20), which is around the expected average performance reported by ProNA2020. The important sequence consecutive central strand and loop are predicted well, while several short sequence segments that are far away in sequence space but close in structure space are missed, which is expected as ProNA2020 has no notion of 3D structure, i.e., cannot identify 'binding sites'. Panels A and B show a different orientation than panels C and D.

USE CASE

We demonstrate PredictProtein.org tools through predictions of the novel coronavirus SARS-CoV-2 (NCBI:txid2697049) nucleoprotein (UniProt identifier P0DTC9/NCAP_SARS2; Figure 1; SOM: Use Case; Supplementary Figure S2). NCAP_SARS2 has 419 residues and interacts with itself (experimentally verified). Sequence similarity and automatic assignment via UniRule (60) suggest NCAP is RNA-binding (binding with the viral genome), binding with the membrane protein M (UniProt identifier P0DTC5/VME1_SARS2), and is fundamental for virion assembly. goPredSim (19) transferred GO terms from other proteins for MFO (*RNA-binding*; GO:0003723; ECO:0000213) and CCO (compartments in the host cell and viral nucleocapsid; GO:0019013; GO:0044172; GO:0044177; GO:0044220; GO:0030430; ECO:0000255) matching annotations found in UniProt (1). While it missed the experimentally verified MFO term *identical protein binding* (GO:0042802), goPredSim predicted *protein folding* (GO:0006457) and *protein ubiquitination* (GO:0016567) suggesting the nucleoprotein to be involved in biological processes requiring protein binding. ProNA2020 (3) predicts RNA-binding regions, the one with highest confidence between I84 (Isoleucine at position 84) and D98 (Aspartic Acid at 98) (protein sequence in SOM: Use Case). While high resolution experimental data on binding is not available, an NMR structure of the SARS-CoV-2 nucleocapsid phosphoprotein N-terminal domain in complex with 10mer ssRNA (PDB identifier 7ACT (61)) supports the predicted RNA-binding site (Figure 2). Additionally, SNAP2 (38) predicts single amino acid variants (SAVs) in that region to likely affect function, reinforcing the hypothesis that this is a functionally relevant site. Although using different input information (evolutionary vs. embeddings), RePROF (5) and ProtBertSec (14) both predict an unusual content exceeding 70% non-regular (neither helix nor strand) secondary structure, suggesting that most of the nucleoprotein might not form regular structure. This is supported by Meta-Disorder (30) predicting 53% overall disorder. Secondary structure predictions match well high-resolution experimental structures of the nucleoprotein not in complex (e.g., PDB 6VYO (62); 6WJI (63)). Both secondary structure prediction methods managed to zoom into the ordered regions of the protein and predicted e.g., the five beta-strands that are formed within the sequence range I84 (Isoleucine) to A134 (Alanine), and the two helices formed within the sequence range spanned from F346 (Phenylalanine) to T362 (Tyrosine).

CONCLUSION

For almost three decades (preceding Google) *PredictProtein* (PP) has been offering predictions for proteins. PP is the oldest prediction Internet server, online for 6-times as long as most other servers (64–66). It pioneered combining machine learning with evolutionary information and making predictions freely accessible online. While the sequence-annotation gap continues to grow, the sequence-structure gap might be bridged in the near future (67). For the time being, servers such as PP, providing a one-stop solution to predictions from many sustained, novel tools are needed. Now, PP is the first server to offer fast embedding-based predictions of structure (ProtBertSec) and function (goPredSim). By slashing runtime for PSSMs from 72 to 4 min through MMseqs2 and better LCSB hardware, PP also delivers evolutionary information-based predictions fast! Instantaneously if the query is in the precomputed *PPcache*. For heavy use, the offline Docker containers provide predictors out-of-the-box. At no cost to users, *PredictProtein* offers to quickly shine light on proteins for which little is known using well validated prediction methods.

DATA AVAILABILITY

Freely accessible webserver PredictProtein.org; Source and docker images: github.com/rostlab.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Maintaining *PredictProtein* over three decades has been tough; many colleagues have helped with hands and brains, developers, and users alike. Thanks to all of you! Please find most contributors in Supplementary Table S2 or at predictprotein.org/credits. In particular, thanks to Noua Toukourou and Maharshi Vyas (both LCSB) for invaluable

help with hardware and software; to David Hoksza (Charles U, Prague) for his work on MolArt; to Marco Punta (IRCCS Milano) for his long-term support; to Inga Weise (TUM) for support with many aspects; to Roy Omond (Blue Bubble, Cambridge), Antoine de Daruvar (Univ. Bordeaux), Yanay Ofran (Bar-Ilan Univ.), Jinfeng Liu (Genentech), Tobias Hamp, Maximilian Hecht, Edda Kloppmann (all previously TUM) for contributing methods and code in the past; Johannes Söding for providing resources to develop and maintain MMseqs2.

FUNDING

Michael Bernhofer was supported by the Competence Network for Scientific High Performance Computing in Bavaria [KONWIHR-III BG.DAF]; Christian Dallago is supported by the Deutsche Forschungsgemeinschaft (DFG) [RO 1320/4-1]; Bundesministerium für Bildung und Forschung (BMBF) [031L0168]; Software Campus 2.0 (TU München), BMBF [01IS17049]; Milot Mirdita acknowledges support from the ERC's Horizon 2020 Framework Programme [`Virus-X', project no. 685778]; BMBF CompLifeSci project horizontal4meta. Martin Steinegger acknowledges support from the National Research Foundation of Korea grant funded by the Korean government (MEST) [2019R1A6A1A10073437, NRF-2020M3A9G7103933]; Creative-Pioneering Researchers Program through Seoul National University; Nir Ben-Tal acknowledges the support of Israeli Science Foundation (ISF) [450/16]; Abraham E. Kazan Chair in Structural Biology, Tel Aviv University; Haim Ashkenazy was supported by Humboldt Research Fellowship for Postdoctoral Researchers of the Alexander von Humboldt Foundation; The PredictProtein web server is hosted by ELIXIR-LU, the Luxembourgish node of the European life-science infrastructure. Funding for open access charge: Library of the Technical University of Munich.
Conflict of interest statement. None declared.

REFERENCES

1. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic. Acids Res.*, **49**, D480–D489.
2. Rost,B., Liu,J., Nair,R., Wrzeszczynski,K.O. and Ofran,Y. (2003) Automatic prediction of protein function. *Cell. Mol. Life Sci.*, **60**, 2637–2650.
3. Qiu,J., Bernhofer,M., Heinzinger,M., Kemper,S., Norambuena,T., Melo,F. and Rost,B. (2020) ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J. Mol. Biol.*, **432**, 2428–2443.
4. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.*, **266**, 525–539.
5. Yachdav,G., Kloppmann,E., Kajan,L., Hecht,M., Goldberg,T., Hamp,T., Hönigschmid,P., Schafferhans,A., Roos,M., Bernhofer,M. *et al.* (2014) PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.*, **42**, W337–W343.
6. Steinegger,M. and Söding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**, 2542.
7. Rost,B. and Sander,C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 7558–7562.
8. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
9. Rost,B. and Sander,C. (1992) Jury returns on structure prediction. *Nature*, **360**, 540.
10. Kajan,L., Yachdav,G., Vicedo,E., Steinegger,M., Mirdita,M., Angermuller,C., Bohm,A., Domke,S., Ertl,J., Mertes,C. *et al.* (2013) Cloud prediction of protein structure and function with PredictProtein for Debian. *Biomed. Res. Int.*, **2013**, 398968.
11. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
12. Mirdita,M., Steinegger,M. and Söding,J. (2019) MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, **35**, 2856–2858.
13. Heinzinger,M., Elnaggar,A., Wang,Y., Dallago,C., Nechaev,D., Matthes,F. and Rost,B. (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, **20**, 723.
14. Elnaggar,A., Heinzinger,M., Dallago,C., Rihawi,G., Wang,Y., Jones,L., Gibbs,T., Feher,T., Angerer,C., Bhowmik,D. *et al.* (2020) ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv doi: https://arxiv.org/abs/2007.06225, 04 May 2021, preprint: not peer reviewed.
15. Alley,E.C., Khimulya,G., Biswas,S., AlQuraishi,M. and Church,G.M. (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods*, **16**, 1315–1322.
16. AlQuraishi,M. (2019) End-to-end differentiable learning of protein structure. *Cell Syst.*, **8**, 292–301.
17. Rao,R., Bhattacharya,N., Thomas,N., Duan,Y., Chen,P., Canny,J., Abbeel,P. and Song,Y. (2019) Evaluating Protein Transfer Learning with TAPE. In: Wallach,H., Larochelle,H., Beygelzimer,A., d'Alché-Buc,F., Fox,E. and Garnett,R. (eds). *Advances in Neural Information Processing Systems*. Vol. **32**. Curran Associates, Inc., pp. 9689–9701.
18. Rives,A., Meier,J., Sercu,T., Goyal,S., Guo,D., Lin,Z., Liu,J., Guo,D., Ott,M., Zitnick,C.L., Ma,J. and Fergus,R. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA*, **118**, e2016239118.
19. Littmann,M., Heinzinger,M., Dallago,C., Olenyi,T. and Rost,B. (2021) Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Rep.*, **11**, 1160.
20. Rao,R., Ovchinnikov,S., Meier,J., Rives,A. and Sercu,T. (2020) Transformer protein language models are unsupervised structure learners. bioRxiv doi: https://doi.org/10.1101/2020.12.15.422761, 15 December 2020, preprint: not peer reviewed.
21. O'Donoghue,S.I., Sabir,K.S., Kalemanov,M., Stolte,C., Wellmann,B., Ho,V., Roos,M., Perdigão,N., Buske,F.A., Heinrich,J. *et al.* (2015) Aquaria: simplifying discovery and insight from protein structures. *Nat. Methods*, **12**, 98–99.
22. Hoksza,D., Gawron,P., Ostaszewski,M. and Schneider,R. (2018) MolArt: a molecular structure annotation and visualization tool. *Bioinformatics*, **34**, 4127–4128.
23. O'Donoghue,S.I., Schafferhans,A., Sikta,N., Stolte,C., Kaur,S., Ho,B.K., Anderson,S., Procter,J., Dallago,C., Bordin,N. *et al.* (2020) SARS-CoV-2 structural coverage map reveals state changes that disrupt host immunity bioinformatics. bioRxiv doi: https://doi.org/10.1101/2020.07.16.207308, 28 September 2020, preprint: not peer reviewed.
24. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
25. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
27. Rost,B. (2001) Protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
28. Bernhofer,M., Kloppmann,E., Reeb,J. and Rost,B. (2016) TMSEG: novel prediction of transmembrane helices. *Proteins*, **84**, 1706–1716.
29. Bigelow,H. and Rost,B. (2006) PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids. Res.*, **34**, W186–W188.

30. Schlessinger,A., Punta,M., Yachdav,G., Kajan,L. and Rost,B. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**, e4433.
31. Schlessinger,A., Yachdav,G. and Rost,B. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinforma. Oxf. Engl.*, **22**, 891–893.
32. Ceroni,A., Passerini,A., Vullo,A. and Frasconi,P. (2006) DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res.*, **34**, W177–W181.
33. Berezin,C., Glaser,F., Rosenberg,J., Paz,I., Pupko,T., Fariselli,P., Casadio,R. and Ben-Tal,N. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinforma. Oxf. Engl.*, **20**, 1322–1324.
34. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. and Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
35. Celniker,G., Nimrod,G., Ashkenazy,H., Glaser,F., Martz,E., Mayrose,I., Pupko,T. and Ben-Tal,N. (2013) ConSurf: using evolutionary data to raise testable hypotheses about protein function. *Isr. J. Chem.*, **53**, 199–206.
36. Ashkenazy,H., Abadi,S., Martz,E., Chay,O., Mayrose,I., Pupko,T. and Ben-Tal,N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.*, **44**, W344–W350.
37. Sigrist,C.J.A., de Castro,E., Cerutti,L., Cuche,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–347.
38. Hecht,M., Bromberg,Y. and Rost,B. (2015) Better prediction of functional effects for sequence variants. *BMC Genomics*, **16 (Suppl 8)**, S1.
39. Goldberg,T., Hecht,M., Hamp,T., Karl,T., Yachdav,G., Ahmed,N., Altermann,U., Angerer,P., Ansorge,S., Balasz,K. *et al.* (2014) LocTree3 prediction of localization. *Nucleic Acids Res.*, **42**, W350–W355.
40. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
41. Zhou,N., Jiang,Y., Bergquist,T.R., Lee,A.J., Kacsoh,B.Z., Crocker,A.W., Lewis,K.A., Georghiou,G., Nguyen,H.N., Hamid,M.N. *et al.* (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.*, **20**, 244.
42. Jiang,Y., Oron,T.R., Clark,W.T., Bankapur,A.R., D'Andrea,D., Lepore,R., Funk,C.S., Kahanda,I., Verspoor,K.M., Ben-Hur,A. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.
43. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
44. Huntley,R.P., Sawford,T., Mutowo-Meullenet,P., Shypitsyna,A., Bonilla,C., Martin,M.J. and O'Donovan,C. (2015) The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res.*, **43**, D1057–D1063.
45. El-Mabrouk,N. and Slonim,D.K. (2020) ISMB 2020 proceedings. *Bioinformatics*, **36**, i1–i2.
46. Steinegger,M., Mirdita,M. and Söding,J. (2019) Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, **16**, 603–606.
47. Abriata,L.A., Tamò,G.E., Monastyrskyy,B., Kryshtafovych,A. and Peraro,M.D. (2018) Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins Struct. Funct. Bioinforma.*, **86**, 97–112.
48. Klausen,M.S., Jespersen,M.C., Nielsen,H., Jensen,K.K., Jurtz,V.I., Sønderby,C.K., Sommer,M.O.A., Winther,O., Nielsen,M., Petersen,B. *et al.* (2019) NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinforma.*, **87**, 520–527.
49. Hamp,T., Goldberg,T. and Rost,B. (2013) Accelerating the original profile kernel. *PLoS One*, **8**, e68459.
50. Asgari,E., McHardy,A.C. and Mofrad,M.R.K. (2019) Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Sci. Rep.*, **9**, 3577.
51. Norambuena,T. and Melo,F. (2010) The protein-DNA interface database. *BMC Bioinformatics*, **11**, 262.
52. Lewis,B.A., Walia,R.R., Terribilini,M., Ferguson,J., Zheng,C., Honavar,V. and Dobbs,D. (2011) PRIDB: a protein-RNA interface database. *Nucleic. Acids. Res.*, **39**, D277–D282.
53. Hamp,T. and Rost,B. (2015) Evolutionary profiles improve protein-protein interaction prediction from sequence. *Bioinforma. Oxf. Engl.*, **31**, 1945–1950.
54. Mirdita,M., von den Driesch,L., Galiez,C., Martin,M.J., Söding,J. and Steinegger,M. (2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic. Acids. Res.*, **45**, D170–D176.
55. Watkins,X., Garcia,L.J., Pundir,S., Martin,M.J. and Consortium,U. (2017) ProtVista: visualization of protein sequence annotations. *Bioinformatics*, **33**, 2040–2041.
56. Reguant,R., Antipin,Y., Sheridan,R., Dallago,C., Diamantoukos,D., Luna,A., Sander,C. and Gauthier,N.P. (2020) AlignmentViewer: sequence analysis of large protein families. *F1000Research*, **9**, 213.
57. Dallago,C., Goldberg,T., Andrade-Navarro,M.A., Alanis-Lobato,G. and Rost,B. (2020) Visualizing human protein-protein interactions and subcellular localizations on cell images through CellMap. *Curr. Protoc. Bioinforma.*, **69**, e97.
58. Dallago,C., Schütze,K., Heinzinger,M., Olenyi,T., Littmann,M., Lu,A.X., Yang,K.K., Min,S., Yoon,S., Morton,J.T. *et al.* (2021) Learned embeddings from deep learning to visualize and predict protein sets. *Curr. Protoc. Bioinforma.*, **1**, e113.
59. Merkel,D. (2014) Docker: lightweight linux containers for consistent development and deployment. *Linux J.*, **2014**, 2.
60. MacDougall,A., Volynkin,V., Saidi,R., Poggioli,D., Zellner,H., Hatton-Ellis,E., Joshi,V., O'Donovan,C., Orchard,S., Auchincloss,A.H. *et al.* (2020) UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics*, **36**, 4643–4648.
61. Dinesh,D.C., Chalupska,D., Silhan,J., Koutna,E., Nencka,R., Veverka,V. and Boura,E. (2020) Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog.*, **16**, e1009100.
62. Chang,C., Michalska,K., Jedrzejczak,R., Maltseva,N., Endres,M., Godzik,A., Kim,Y. and Joachimiak,A. (2020) Crystal structure of RNA binding domain of nucleocapsid phosphoprotein from SARS coronavirus 2. doi:10.2210/pdb6vyo/pdb.
63. Minasov,G., Shuvalova,L., Wiersum,G. and Satchell,K. (2020) 2.05 angstrom resolution crystal structure of C-terminal dimerization domain of nucleocapsid phosphoprotein from SARS-CoV-2. doi:10.2210/pdb6wji/pdb.
64. Schultheiss,S.J., Münch,M.-C., Andreeva,G.D. and Rätsch,G. (2011) Persistence and availability of Web services in computational biology. *PLoS One*, **6**, e24914.
65. Wren,J.D., Georgescu,C., Giles,C.B. and Hennessey,J. (2017) Use it or lose it: citations predict the continued online availability of published bioinformatics resources. *Nucleic Acids. Res.*, **45**, 3627–3633.
66. Kern,F., Fehlmann,T. and Keller,A. (2020) On the lifetime of bioinformatics web services. *Nucleic Acids Res.*, **48**, 12523–12533.
67. Callaway,E. (2020) 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, **588**, 203–204.