Received: 4 August 2022 | Revised: 9 November 2022 | Accepted: 21 November 2022

DOI: 10.1002/pro.4524

TOOLS FOR PROTEIN SCIENCE

LambdaPP: Fast and accessible protein-specific phenotype predictions

Tobias Olenyi[1,2] | **Céline Marquet**[1,2] | **Michael Heinzinger**[1,2] | **Benjamin Kröger**[1] | **Tiha Nikolova**[1] | **Michael Bernhofer**[2] | **Philip Sändig**[1] | **Konstantin Schütze**[1] | **Maria Littmann**[1] | **Milot Mirdita**[3] | **Martin Steinegger**[3,4,5] | **Christian Dallago**[1,6] | **Burkhard Rost**[1,7]

[1]TUM (Technical University of Munich) Department of Informatics, Bioinformatics- & Computational Biology—i12, Garching, Germany

[2]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Garching, Germany

[3]School of Biological Sciences, Seoul National University, Seoul, South Korea

[4]Korea Artificial Intelligence Institute, Seoul National University, Seoul, South Korea

[5]Korea Institute of Molecular Biology and Genetics, Seoul National University, Seoul, South Korea

[6]VantAI, New York, USA

[7]Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & TUM School of Life Sciences Weihenstephan (WZW), Freising, Germany

Correspondence

Tobias Olenyi, TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology—I12, Boltzmannstr. 3, 85748 Garching/Munich, Germany.
Email: lambda@rostlab.org

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: 031L0168; Deutsche Forschungsgemeinschaft, Grant/Award Number: RO1320/4-1; National Research Foundation of Korea, Grant/Award Numbers: 2019R1-A6A1-A10073437, 2020M3-A9G7-103933, 2021-M3A9-I4021220, 2021-R1C1-C102065; Seoul National University, Grant/Award Number: Creative-Pioneering Researchers Program; Software Campus (TUM) 2.0, Grant/Award Number: 01IS17049

Review Editor: Nir Ben-Tal

Abstract

The availability of accurate and fast artificial intelligence (AI) solutions predicting aspects of proteins are revolutionizing experimental and computational molecular biology. The webserver *LambdaPP* aspires to supersede PredictProtein, the first internet server making AI protein predictions available in 1992. Given a protein sequence as input, *LambdaPP* provides easily accessible visualizations of protein 3D structure, along with predictions at the protein level (GeneOntology, subcellular location), and the residue level (binding to metal ions, small molecules, and nucleotides; conservation; intrinsic disorder; secondary structure; alpha-helical and beta-barrel transmembrane segments; signal-peptides; variant effect) in seconds. The structure prediction provided by *LambdaPP*—leveraging *ColabFold and computed in minutes*—is based on *MMseqs2* multiple sequence alignments. All other feature prediction methods are based on the pLM *ProtT5*. Queried by a protein sequence, *LambdaPP* computes protein and residue predictions almost instantly for various phenotypes, including 3D structure and aspects of protein function. LambdaPP is freely available for everyone to use under embed.predictprotein.org, the interactive results for the case study can be found under https://embed.predictprotein.org/

Tobias Olenyi and Céline Marquet have equal first authorship.

Christian Dallago and Burkhard Rost have equal senior authorship.

o/Q9NZC2. The frontend of LambdaPP can be found on GitHub (github.com/sacdallago/embed.predictprotein.org), and can be freely used and distributed under the academic free use license (AFL-2). For high-throughput applications, all methods can be executed locally via the bio-embeddings (bioembeddings.com) python package, or docker image at ghcr.io/bioembeddings/bio_embeddings, which also includes the backend of LambdaPP.

KEYWORDS

artificial intelligence, protein annotation, protein function prediction, protein language models, protein structure prediction, web server

1 | INTRODUCTION

1.1 | PP protein prediction since dawn of internet

Launched 30 years ago, the PredictProtein (PP) web server provides a comprehensive interface for protein sequence analysis (Bernhofer et al., 2021; Rost et al., 1994; Rost & Sander, 1992; Yachdav et al., 2014). As the first internet server for predicting aspects of protein structure and function, it offers a broad overview of predicted features. Among many innovations, PP introduced the combination of evolutionary information (EI) from multiple sequence alignments (MSAs) and machine learning (Rost & Sander, 1993), a subset of artificial intelligence (AI), for protein prediction. Nature's 2021 method of the year (Marx, 2022), *AlphaFold2* (Jumper et al., 2021), peaked the innovation by essentially solving the protein structure prediction problem with models approaching experimental high-resolution, inspiring a new era of advancing methods (Ahdritz et al., 2021; Baek & Baker, 2022; Mirdita et al., 2022) and their application (Cardim-Pires et al., 2021; Kouba et al., 2021; Zhao et al., 2021). *AlphaFold2* came when more sequences than ever before (2.1 billion proteins in BFD; Steinegger & Söding, 2018) met new AI-optimized algorithms and hardware. *PredictProtein* and *AlphaFold2* work great in their domains and integrating both might help in enabling experts and novices alike to experiment, hypothesize, and generate novel insights quickly.

1.2 | EI + AI top, but not without caveats

Since the release of PredictProtein, the amount of non-annotated sequences has been rapidly increasing (Rost & Sander, 1996; Steinegger & Söding, 2018). In fact, the sequence-annotation gap continues to grow despite experimental advances, e.g., experimental residue binding annotations are currently added for only two sequence-unique proteins per month for any organism and any ligand (Littmann, Heinzinger, Dallago, Weissenow, et al., 2021). AI models mitigate this gap. Until 2020, almost all state-of-the-art prediction methods had implemented the concept introduced by PP, namely inputting MSAs into AI. Although super-fast tools relying on algorithmic and hardware advances sped up MSA generation (Buchfink et al., 2021; Mirdita et al., 2019), bio-databases continue outgrowing the pace at which computer hardware accelerates (Moore, 1965; Steinegger et al., 2019; Theis & Wong, 2017). This challenge cannot be resolved by advancing computers. On top, MSAs are not always informative, especially for small sequence families, or proteins of the Dark Proteome (Perdigao et al., 2015).

1.3 | Protein language models (pLMs) solving problems?

Developments in representation learning (Bengio et al., 2013), particularly in natural language processing (Chowdhary, 2020), let to encoding latent protein information including aspects of evolutionary information. Protein language models (pLMs) based on deep learning large sets of unannotated sequences to generate numerical representations (embeddings) (Bepler & Berger, 2019; Elnaggar et al., 2021; Heinzinger et al., 2019; Madani et al., 2020; Ofer et al., 2021; Rives et al., 2021). Embeddings from pLMs have been successfully used as input to downstream protein prediction tools (Bileschi et al., 2022; Heinzinger et al., 2022; Hie et al., 2022; Littmann, Bordin, Heinzinger, Schütze, et al., 2021; Littmann, Heinzinger, Dallago, Olenyi et al., 2021; Littmann, Heinzinger, Dallago, Weissenow, et al., 2021; Marks et al., 2021; Marquet et al., 2021; Meier et al., 2021; Singh et al., 2022; Stärk et al., 2021; Weissenow et al., 2022; Zhou et al., 2020).

Some pLM-based methods still appear inferior to top MSA-based methods (Elnaggar et al., 2021; Littmann, Heinzinger, Dallago, Olenyi, et al., 2021; Weissenow et al., 2022), others bested those (Almagro Armenteros et al., 2017; Bepler & Berger, 2021; Elnaggar et al., 2021; Høie et al., 2022; Ilzhoefer et al., 2022; Lin et al., 2022; Littmann, Heinzinger, Dallago, Weissenow, et al., 2021; Marquet et al., 2021; Stärk et al., 2021). Performance has risen so much that sequence-specific pLM-based predictions now can capture some aspects of structural and functional dynamics better than much more accurate family-averaged solutions even from *AlphaFold2* (Lin et al., 2022; Weissenow et al., 2022; Wu et al., 2022).

1.4 | pLM-based protein predictions for the web

Thirty years ago, PredictProtein offered the first access to a variety of MSA-based AI solutions. Similarly, LambdaPP now makes state-of-the-art solutions for embedding-based predictions available. The server outputs predictions for the entire query protein (per-protein) and for each of its residues (per-residue, Figure 1). All results are linked to 3D structure visualizations, currently retrieved from the AlphaFold Database (release 4, 07/2022 created using *AlphaFold Monomer v2.0 pipeline*) (Varadi et al., 2022) or if unavailable predicted using *ColabFold* (v2.1.14) simplifying *AlphaFold2* at similar performance (Mirdita et al., 2022). Currently the only non-pLM method in the LambdaPP frame, it will soon be complemented by pLM-based solutions (Lin et al., 2022; Weissenow et al., 2022; Wu et al., 2022). As novel AI tools leveraging embeddings emerge, e.g., predicting CATH (Sillitoe et al., 2021) classes (Heinzinger et al., 2022), LambdaPP will be updated to extend its breadth. All feature prediction methods integrated into the LambdaPP webserver currently use ProtT5 (Elnaggar et al., 2021) that, in our hands, outperformed ESM-1b (Rives et al., 2021) and others (Alley et al., 2019; Bepler & Berger, 2019; Elnaggar et al., 2021; Heinzinger et al., 2019) for numerous applications (Bernhofer & Rost, 2022; Heinzinger et al., 2022; Littmann, Bordin, Heinzinger, Schütze, et al., 2021; Littmann, Heinzinger, Dallago, Olenyi, et al., 2021; Littmann, Heinzinger, Dallago, Weissenow, et al., 2021; Marquet et al., 2021; Stärk et al., 2021; Weissenow et al., 2022). This consistency also increases speed as the generation of embeddings becomes a limiting step.

2 | RESULTS

2.1 | Access of server

All methods are available through embed.predictprotein.org where users submit amino acid sequences up to 2000 residues. This limit speeds-up response-time (embedding computation is non-linear in protein length). Formats currently handled are: FASTA sequence, UniProt accession number, UniProt protein name (The UniProt Consortium et al., 2021), or a string of residues ("AA format"). Results are displayed immediately if cached, if not they are computed on the fly within seconds for pLM-based methods. There is no queuing system, as it takes longer to generate the display items on the frontend than

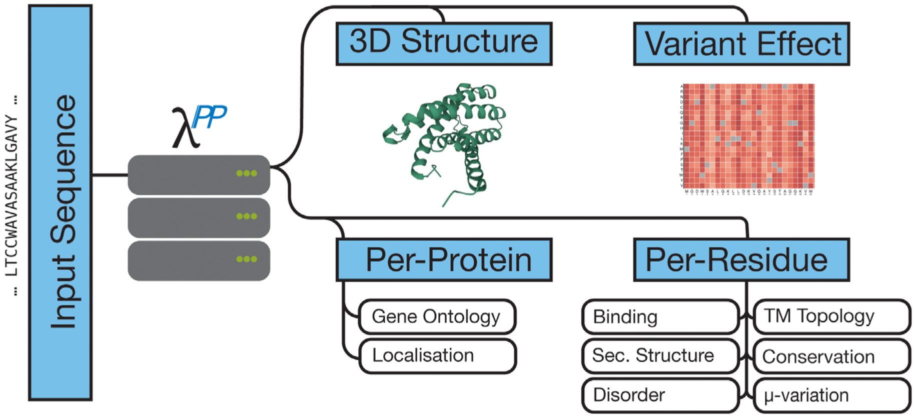


FIGURE 1 LambdaPP pipeline. Starting with an amino acid sequence, LambdaPP orchestrates the prediction of (1) protein structure using ColabFold (Mirdita et al., 2022), (2) per-protein features: gene ontology (GO) annotations using goPredSim (Littmann, Heinzinger, Dallago, Olenyi, et al., 2021), subcellular location using LA (Stärk et al., 2021); (3) per-residue features: binding residues using bindEmbed21DL (Littmann, Heinzinger, Dallago, Weissenow, et al., 2021), conservation using ProtT5cons (Marquet et al., 2021), disorder using SETH (Ilzhoefer et al., 2022), secondary structure using ProtT5-sec (Elnaggar et al., 2021), helical and barrel transmembrane (TM) regions using TMbed (Bernhofer & Rost, 2022); and (4) variant effect scores using VESPAl (Marquet et al., 2021).

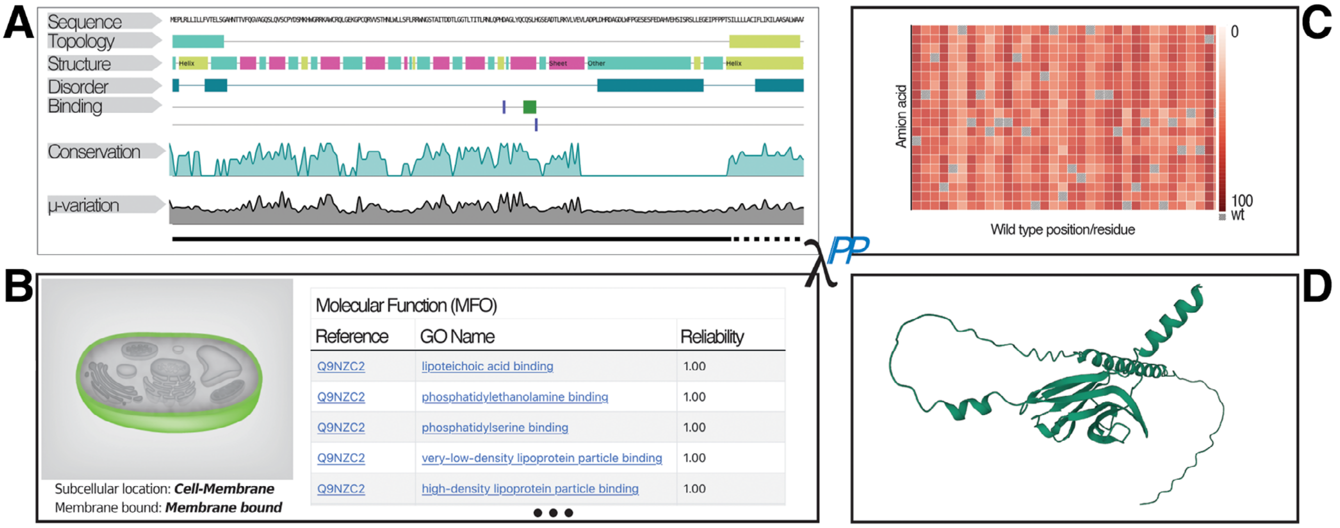
to compute the predictions in the backend. Protein 3D structures are fetched from the AlphaFold Database (Varadi et al., 2022) when inputting UniProt accessions, or predicted by ColabFold (Mirdita et al., 2022) through a *first come, first serve* queue (completing within 30 min for protein with 350 residues). Due to limited GPUs, *ColabFold* predictions are restricted to proteins shorter than 500 residues, but work is underway to transition to pLM-based 3D structure prediction providing fast and accurate predictions for longer sequences. All results are cached for 10 days before being deleted to conserve disk space and respect data privacy. Users can download results.

2.2 | Frontend and interface

The main LambdaPP interface displays the predictions in thematically ordered sections. Leveraging the *neXtProt* feature viewer (Schaeffer & Teixeira, 2017), per-residue predictions are displayed in one view-pane (Figure 2). The neXtProt plugin enables to display categorial features, e.g., binding, transmembrane-regions, and secondary structure as colored regions, and continues features, e.g., disorder, variance effect, and conservation, as line plots. An interactive connection between residue-level features and 3D structure maps predictions onto 3D visualization while displaying additional information in tooltips. The *protein-level section* visualizes subcellular location through colored images (Dallago et al., 2020; Stärk et al., 2021) and GO-term predictions as lists of predicted GO-terms along with scores reflecting reliability (RI) and links to the reference protein used for the annotation transfer (Littmann, Heinzinger, Dallago, Olenyi, et al., 2021). The *Single Amino acid Variant (SAV) effect section* features the predictions of how much point mutations (SAVs) negatively affect molecular function. By default, the effect is predicted for all 19 non-native SAVs, i.e., all point mutants irrespectively of their reachability through single nucleic variants (SNVs or SNPs). Finally, the predicted 3D structure is visualized in the *structure section*, using the Mol* plugin (Sehnal et al., 2021). To facilitate exploring predictions, we offer two alternative interfaces: the *print-page*, which displays several residue-level features in a print-ready form, and the *interactive* page, which displays the neXtProt feature viewer and 3D structure prediction in a single panel to allow easier interactive exploration. These alternative displays are reached from the main interface by clicking on the suggested alternative display buttons.

2.3 | Backend and programmatic access

Users can retrieve prediction results programmatically through the bio-embeddings REST API (details: api.bioembeddings.com/api). This interface allows, e.g., to create ProtT5 embeddings of proteins with up to 2000 residues (hardware not model restriction), and to retrieve


A
Sequence
MEPLRLLILLFVTELSGAHNTTVFQGVAGQSLQVSCPYDSMKHWGRRKAWCRQLGEKGPCQRVVSTHNLWLLSFLRRWNGSTAITDDTLGGTLTITLRNLQPHDAGLYQCQSLHGSEADTLRKVLVEVLADPLDHRDAGDLWFPGESESFEDAHVEHSISRSLLEGEIPFPPTSILLLLACIFLIKILAASALWAA
Topology
Structure
Helix
Sheet
Other
Helix
Disorder
Binding
Conservation
μ-variation
λPP
B
C
0
100
wt
Amion acid
Wild type position/residue
D

Subcellular location: ***Cell-Membrane***
Membrane bound: ***Membrane bound***

Molecular Function (MFO)

Reference	GO Name	Reliability
Q9NZC2	lipoteichoic acid binding	1.00
Q9NZC2	phosphatidylethanolamine binding	1.00
Q9NZC2	phosphatidylserine binding	1.00
Q9NZC2	very-low-density lipoprotein particle binding	1.00
Q9NZC2	high-density lipoprotein particle binding	1.00

• • •

FIGURE 2 LambdaPP output for TREM2_HUMAN. Panel a: residue level features: secondary structure, transmembrane topology, disordered residues, small molecule, nucleic or metal binding residues, residue conservation and average variation (Bernhofer & Rost, 2022; Elnaggar et al., 2021; Ilzhoefer et al., 2022; Littmann, Heinzinger, Dallago, Weissenow, et al., 2021; Marquet et al., 2021); panel b: sequence-level features: predicted subcellular localization (Stärk et al., 2021), and an excerpt of predicted GO-annotations (Littmann, Heinzinger, Dallago, Olenyi, et al., 2021); panel c: effect of SAVs (wildtype sequence on x-axis, mutations on y-axis; darker color = higher effect) (Marquet et al., 2021); and panel d: predicted 3D structure (Mirdita et al., 2022). Interactive version at https://embed.predictprotein.org/o/Q9NZC2.

the full spectrum of available annotations as generated by the backend, i.e., to download all predictions in JSON format. The backend could be hosted entirely on a standard workstation equipped with a workstation GPU (e.g., Quadro RTX 8000 46GB RAM), delivering results in seconds compared to minutes or hours for multi-node and cluster-based PP (Bernhofer et al., 2021). However, to counter machine faults and guarantee availability, the hosted LambdaPP runs on different servers at the LCSB in Luxembourg and the TUM in Munich and can process one request at a time, on average, in 4 s for proteins of 350 residues. The hosted backend can be manually scaled to respond to parallel requests during high demand.

2.4 | Availability for local deployment

To take advantage of the methods feature on LambdaPP locally, advanced users can rely on the bio-embeddings package (Dallago et al., 2021) (bioembeddings.com). Along with various use cases, it provides a docker image (ghcr.io/bioembeddings/bio_embeddings) for easy deployment on local machines. We recommend installing bio-embeddings locally to avoid length restrictions imposed on LambdaPP.

2.5 | Use case: *Triggering receptor* (Q9NZC2)

We demonstrated the LambdaPP workflow using the *triggering receptor expressed on myeloid cells 2* (UniProt accession: Q9NZC2) protein and compared results to the expert curated UniProtKB entry (The UniProt Consortium et al., 2021). We selected Q9NZC2, as it is associated with *Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalo-pathy* (PLOSL2) and has diverse annotations in different regions (all at Figure S1).

> *Per-protein*: LambdaPP trivially listed most UniProtKB GO annotations with high reliability because the protein was in goPredSim's lookup set. Subcellular location was correctly predicted as cell-membrane, and the structure from the AlphaFold Database (Varadi et al., 2022) aligned well with the structures of 5ELI (Alexander-Brett & Kober, 2015; Kober et al., 2016) (RMSD: 0.45 Å) and 5UD8 (Sudom et al., 2016, 2018) (RMSD: 0.37 Å; Figure S4).
>
> *Per-residue*: LambdaPP marked the first 17 residues as signal peptides, matching the automatic (rule-based) UniProtKB annotation (one residue shorter). The transmembrane stretch matched with the UniProtKB transmembrane stretch (four residues shorter; Figure S5).
>
> *Binding*: UniProtKB has no annotation of metal ion, small molecule, or nucleic acid binding, while bindEmbed21DL predicted two metal ions and one small-molecule binding residues with high reliability which might be interesting targets for future experiments (Littmann, Heinzinger, Dallago, Weissenow, et al., 2021).
>
> *Disorder and conservation*: UniProtKB also annotates no intrinsic disorder. Yet, the predicted high disorder content for loop regions next to the transmembrane segment and the high order content outside correlated well with *AlphaFold 2's* predicted *Local Distance Difference Test* (pLDDT), reflecting the confidence for the 3D prediction with pLDDT > 70 typically considered reliable (Figure S6; Piovesan et al., 2022; Wilson et al., 2022). For sequence conservation, we compared the predictions shown by LambdaPP to those from ConSurfDB (Ben Chorin et al., 2020) obtained for 5ELI (Alexander-Brett & Kober, 2015; Kober et al., 2016) (mean squared error [MSE] ~ 9) and 5UD8 (Sudom et al., 2016, 2018) (MSE ~ 4; Figure S7).
>
> *SAV effects*: The predictions of the effects of single SAVs upon molecular function showed a similar trend as the UniProtKB annotations: Q9NZC2 seems susceptible to mutation effects. Zooming into residues relevant for binding to PLOS2, e.g., a mutation at residue position 126 (V > G) suggests a strong mutation effect (score 71). Residues marked in LambdaPP with high scores could be an interesting target for future mutational assays (e.g., residues at position 35, 85, 105).

2.6 | Interactive selection

When users select predicted residue features on the nextProt viewer, the selection is transferred into the 3D viewer (Figure S3). This eases the identification of relevant structural regions. For instance, selecting the predicted signal peptide for Q9NZC2 (Figure S3), highlights the region on 3D structure and allows to verify the prediction visually.

2.7 | Use case: predicting a long protein

We selected a hypothetical *ice nucleation* protein from *Pseudomonas syringae* (ICEV_PSESX, O33479) due to its length, its remarkable *AlphaFold2* structure prediction (Figure 3a), and its possibly interesting ice-binding

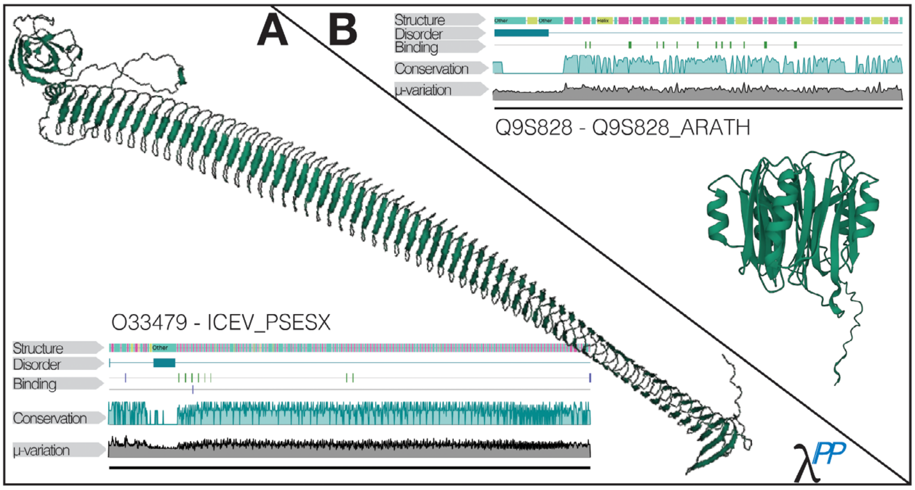


FIGURE 3 Remarkable AlphaFold2 predictions. Panel a (lower left triangle) displays the 3D structure predicted by AlphaFold2 for the ice nucleation protein ICEV_PSESX. The protein contains 1165 residues and is available through LambdaPP as part of AFDB. Panel b (upper right triangle) showcases the AlphaFold2 prediction of what might constitute a novel superfamily for the plant protein with the UniProt identifier Q9S828_ARATH.

properties (Cid et al., 2016). Most of the structure is predicted with very high confidence (pLDDT > 90; confirmed by predicted low disorder). Another region with low *AlphaFold2* pLDDT correlates with predicted low conservation and high disorder (residues 111–165, Figure 3a: loop next to top left). Three residues are predicted as metal binding (41, 208, 1196), and 13 as small molecule binding (173–175, 189–192, 222, 238–239, 253–254, 590, 606).

Disaccord: ProtT5Sec predicted secondary structure differed partially from AlphaFold2 either suggesting alternative conformations or prediction inconsistencies. Similarly, GO annotations (CC: nucleus, BP: regulation of transcription by RNA polymerase II, MF: double-stranded DNA binding) are predicted with low reliability (RI: 0.24) and differ from those inferred by homology in UniProtKB (CC: cell outer membrane, MF: ice-binding). While low reliability is a good indicator that the predicted features should be taken with a grain of salt, the result remains interesting given the lack of proteins with reliable GO annotations for GOPredSim. This could point to understudied biological functions.

2.8 | Use case: annotating family and function for an unknown protein

A plant protein with UniProt ID Q9S828 remains uncharacterized by experiment. AlphaFold2's 3D prediction suggests it folds into a new superfamily (Bordin et al., 2022). The predicted pLDDT of the 3D structure (very low 1–29, low 30–33, very low 34–45, low 46–48) matches partially with UniProtKB disorder annotations (Abriata et al., 2018; Ahdritz et al., 2021; Alexander-Brett & Kober, 2015; Alley et al., 2019; Almagro Armenteros et al., 2017; Baek & Baker, 2022; Ben Chorin et al., 2020; Bengio et al., 2013; Bepler & Berger, 2019, 2021; Berezin et al., 2004; Bernhofer et al., 2021; Bernhofer & Rost, 2022; Bileschi et al., 2022; Bordin et al., 2022; Buchfink et al., 2021; Cardim-Pires et al., 2021; Chowdhary, 2020; Cid et al., 2016; Dallago et al., 2020, 2021; Dass et al., 2020; El-Mabrouk & Slonim, 2020; Elnaggar et al., 2021; Heinzinger et al., 2019, 2022; Henikoff & Henikoff, 1992; Hie et al., 2022; Høie et al., 2022) and LambdaPP-included disorder predictions (residues 1–39; N-terminal region looping to the lower left (Figure 3b)). This region is also predicted as poorly conserved.

LambdaPP suggests that Q9S828 might be a serine–threonine kinase (GO:0004674), involved in the phosphorylation of peptidyl-serine (GO:0018105) and contribute to flower development (GO:0009908). Compared to the broader UniprotKB annotations (GO:001630, GO:0016310), functional annotations provided by LambdaPP allow hypothesizing about the protein's functional role in the plant and design targeted experiments to validate the predictions.

3 | CONCLUSIONS

LambdaPP importantly advances into the next generation of protein prediction, providing the first lightning fast, all-round protein prediction server based almost exclusively on embeddings from one protein language model—ProtT5, accompanied by high-quality 3D structure predictions from the *AlphaFold DB* or *ColabFold*. The web interface allows detailed analyses without requiring users to chip-in high-end servers, AI knowledge, or programming skills. The sub-minute turnaround time from sequence input to predictions of 15 different per-residue and per-protein features coupled with an intuitive user interface allow users to quickly generate an overview for any desired protein sequence, in turn allowing hypothesis generation and analysis. Thereby, LambdaPP might become valuable for bridging the sequence-annotation gap through predictions from high-quality methods, allowing researchers to prioritize experiments and curation efforts.

4 | MATERIALS AND METHODS

LambdaPP computes protein language model (pLM) representations (embeddings) using ProtT5 (Elnaggar et al., 2021) from single amino acid sequences. More specifically, the embeddings are derived exclusively from the encoder-part of ProtT5 in half-precision, i.e., from float32 to float16 model weights, speeding up inference and improving performance of subsequent models (Elnaggar et al., 2021). These embeddings are input to all methods provided via LambdaPP except ColabFold (Mirdita et al., 2022). Per-protein features predicted solely with ProtT5 embeddings as input currently include subcellular location (Stärk et al., 2021), and Gene Ontology terms (GO) (Littmann, Heinzinger, Dallago, Olenyi, et al., 2021). Per-residue predictions solely with ProtT5 embeddings as input include: conservation (Marquet et al., 2021); helical transmembrane regions, transmembrane beta barrels, along with signal peptides (Bernhofer & Rost, 2022); binding for various ligands (Littmann, Heinzinger, Dallago, Weissenow, et al., 2021); intrinsically disordered regions (Ilzhoefer et al., 2022); secondary structure (Elnaggar et al., 2021). LambdaPP also predicts the effect of introducing single amino acid variants (SAV) in the input sequence upon molecular function, which uses the predicted conservation with a BLOSUM62-score (Henikoff & Henikoff, 1992) of the SAV as input (Marquet et al., 2021). Additionally, the 3D structure for the query-sequence is currently predicted with ColabFold (Mirdita et al., 2022) implementing AlphaFold2 (Jumper et al., 2021), however pLM-based alternatives are in development and will be available online soon (Weissenow et al., 2022).

4.1 | Per-protein: Gene ontology (GO)

The method goPredSim (Littmann, Heinzinger, Dallago, Olenyi, et al., 2021) predicts GO terms by transferring annotations from the closest neighbor in a lookup dataset of proteins with known GO annotations (Littmann, Heinzinger, Dallago, Olenyi, et al., 2021). The closest neighbor is defined by the smallest pairwise Euclidean distance calculated between the ProtT5 embeddings of the lookup set and the target. The distance is converted to a Reliability Index (RI) ranging from 0 (weak prediction) to 1 (confident prediction). RI values above 0.35 for biological process ontology (BPO), 0.28 for molecular function ontology (MFO), and 0.29 cellular component ontology (CCO) suggest reliable results. Replicating CAFA3 (Zhou et al., 2019), goPredSim reached $F_{\max}$ (BPO): 38 ± 2%, MFO: 52 ± 3%, and CCO: 59 ± 2% using ProtT5 embeddings (Bernhofer et al., 2021; Littmann, Heinzinger, Dallago, Olenyi, et al., 2021). Tested on proteins annotated after February 2020 and confirmed by CAFA4 (El-Mabrouk & Slonim, 2020), results were slightly better (github.com/Rostlab/goPredSim—performance-assessment).

4.2 | Per-protein: subcellular location

For a given protein sequence, light attention (LA) predicts where in a cell a protein functions, i.e., its subcellular location or cellular compartment (Stärk et al., 2021). Ten subcellular localization classes are differentiated as mapped in DeepLoc (Almagro Armenteros et al., 2017). The LA network architecture using ProtT5 embeddings as input significantly outperformed MSA-based state-of-the-art (SOTA) methods by about eight percentage points (Q10). For this task, ProtT5 embeddings (Elnaggar et al., 2021) significantly outperformed all other pLM embeddings as input for the same architecture (Alley et al., 2019; Bepler & Berger, 2019; Elnaggar et al., 2021; Heinzinger et al., 2019; Rives et al., 2021).

4.3 | Per-residue: ligand-binding

bindEmbed21DL (Littmann, Heinzinger, Dallago, Weissenow, et al., 2021), a two-layer convolutional neural network (CNN) exclusively inputting ProtT5 embeddings, predicts residues binding to metal ions, nucleic acids, or small molecules, distinguishing the three classes. The pLM-based method substantially outperformed its MSA-based predecessor (F1 = 47 ± 2% vs. F1 = 34 ± 2%) (Schelling et al., 2018) on binding annotations from BioLiP (Yang et al., 2013). The seemingly low F1-score hid that the method often outperformed human annotations, in the sense that all strongly predicted (high reliability) residues annotated as non-binding investigated in detail appeared to reveal missing annotations rather than prediction mistakes.

4.4 | Per-residue: conservation

ProtT5cons (Marquet et al., 2021) is a two-layer CNN, predicting the degree to which a residue is conserved in an MSA without using an MSA as input. The conservation level is scaled from 0 (highly variable) to 8 (highly conserved) similarly to ConSurf-DB (Ben Chorin et al., 2020). ProtT5 embeddings outperformed those from the pLM ProtBERT and performed *on par* with the ESM-1b (Rives et al., 2021) pLM embeddings. While only taking embeddings as input, the performance of ProtT5-cons was similar to ConSeq (Berezin et al., 2004) using MSAs (two-state Matthews Correlation Coefficient (MCC) (embeddings) = 0.596 ± 0.006 vs. MCC (ConSeq) = 0.608 ± 0.006) when compared to conservation levels of ConSurf-DB.

4.5 | Per-residue: intrinsic disorder

SETH (Ilzhoefer et al., 2022), a two-layer CNN, predicts the degree of intrinsic disorder of a residue as defined by the chemical shift *Z*-scores (CheZOD) (Nielsen & Mulder, 2020), where values below eight signify disorder and values above eight signify order. Different pLMs were compared (ProtT5: Elnaggar et al., 2021; ProSE: Bepler & Berger, 2021; ESM-1b: Rives et al., 2021; ProtBERT: Elnaggar et al., 2021; SeqVec: Heinzinger et al., 2019) with ProtT5 numerically outperforming the others. SETH outperformed all existing SOTA approaches in terms of mean AUC (area under the receiver operating characteristic curve) and Spearman correlation (0.72 ± 0.01 for SETH vs. 0.67 ± 0.01 for next best method ODinPred (Dass et al., 2020)) as well as similar current solutions operating on ESM-1b embeddings (Redl et al., 2022).

4.6 | Per-residue: secondary structure

ProtT5-sec (Elnaggar et al., 2021), a two-layer CNN, reached a Q3 (three state per-residue accuracy) of 81 ± 1.6% for the CASP12 (Abriata et al., 2018) test set and Q3 of 84 ± 0.5% for a larger data set NEW364 (Elnaggar et al., 2021) competitive with, or even surpassing, top methods relying on MSAs.

4.7 | Per-residue: Transmembrane helices and strands

TMbed predicts for each residue one of four classes: alpha helical transmembrane (TM) region, transmembrane beta strand, signal peptide, or other (Bernhofer & Rost, 2022). For proteins with TM regions, it also predicts the inside/outside orientation within the membrane, i.e., on which side of the membrane the N-terminus begins. The model uses a four-layer CNN combined with a Gaussian filter and a Viterbi decoder. When applied to a non-redundant test set, TMbed correctly predicted 94 ± 8% of beta-barrel transmembrane proteins (TMPs) and 98 ± 1% of alpha-helical TMPs at false positive rates <1%. Furthermore, TMbed placed on average 9 out of 10 transmembrane segments within five residues of the experimental observation. TMbed performed *on par* with or better than SOTA methods. It stood out in terms of its low false positive rate and speed; both making TMbed well suited for high-thruput annotation and filtering, such as annotating millions of AlphaFold2 models.

4.8 | Per-residue: Variant effect and μ-variation

To predict the effect of SAVs, VESPAl (Marquet et al., 2021) takes the nine-state conservation prediction by ProtT5cons, and the BLOSUM62 substitution matrix (Henikoff & Henikoff, 1992) as input for a logistic regression. The simple architecture of VESPAl was close in performance to SOTA MSA-based methods (Laine et al., 2019; Riesselman et al., 2018), and the embedding-based ESM-1v (Meier et al., 2021) on 39 deep mutational scanning (DMS) experiments (with 135,665 SAV) that had not been used for development. The per-residue μ-variation describes the average effect score of the 19 possible substitutions for the respective wild type.

4.9 | 3D structure prediction

LambdaPP also includes SOTA MSA-based 3D structure predictions. If UniProt accessions are used as an input, the 3D structure is retrieved from the AlphaFold Database (AFDB) (Varadi et al., 2022). If the structure is unavailable in AFDB or the input is not a UniProt accession number, ColabFold (Mirdita et al., 2022) is employed to predict the 3D structure, easing access to AlphaFold2 (Jumper et al., 2021). Toward this end, MSAs are generated by searching with MMseqs2 (Mirdita et al., 2019) against UniRef 30 and ColabFoldDB (The UniProt Consortium et al., 2021). For single predictions, this combination is 20–30 times faster than the original AlphaFold2 at little loss of performance on the CASP14 (Kryshtafovych et al., 2021) targets (Mirdita et al., 2019). Further parameters are an early stop criterion of a prediction certainty (pLDDT) above 85 or below 40, a default recycle count of 3 and a compilation of only the best performing out of five AlphaFold2 models. As the goal of LambdaPP is to provide a single reference for pLM-based predictions, 3D structure will soon be predicted using tools recently presented in the literature (Lin et al., 2022; Weissenow et al., 2022; Wu et al., 2022), which allows structure prediction to happen in seconds rather than in minutes, at accuracy comparable with MSA-based methods.

AUTHOR CONTRIBUTIONS

Tobias Olenyi: Conceptualization (equal); formal analysis (lead); methodology (supporting); project administration (equal); software (lead); supervision (supporting); validation (equal); visualization (lead); writing – original draft (equal); writing – review and editing (equal). **Céline Marquet:** Conceptualization (equal); data curation (equal); formal analysis (supporting); methodology (lead); project administration (equal); supervision (supporting); validation (equal); visualization (supporting); writing – original draft (equal); writing – review and editing (equal). **Michael Heinzinger:** Software (supporting); writing – original draft (supporting). **Benjamin Kröger:** Software (equal). **Tiha Nikolova:** Software (equal). **Michael Bernhofer:** Software (supporting). **Philip Sändig:** Software (equal). **Konstantin Schütze:** Software (supporting). **Maria Littmann:** Software (supporting). **Milot Mirdita:** Resources (equal); software (supporting). **Martin Steinegger:** Resources (equal); software (supporting). **Christian Dallago:** Conceptualization (equal); project administration (equal); software (equal); supervision (equal); writing – original draft (supporting); writing – review and editing (supporting). **Burkhard Rost:** Supervision (equal); writing – original draft (supporting); writing – review and editing (supporting).

ACKNOWLEDGMENTS

Thanks, primarily to Tim Karl (TUM) for invaluable help with hardware and software, and to Inga Weise (TUM) for support with many other aspects of this work. Thanks to John Jumper and his team at DeepMind for the breakthrough development of AlphaFold2 and for making code and predictions freely available. Two anonymous reviews helped substantially: thanks! Last, not least, thanks to all who deposit experimental data in public databases, to those who maintain these databases, and those who make methods available that allow enriching the experimental data. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

LambdaPP is freely available for everyone to use under embed.predictprotein.org, the interactive results of the described case study can be found under https://embed.predictprotein.org/o/Q9NZC2. The code of LambdaPP can be found on GitHub (github.com/sacdallago/embed.predictprotein.org), and can be freely used and distributed under the academic free use license (AFL-2). For high-throughput applications, all methods can be executed locally via the bio-embeddings (bioembeddings.com) python package, or docker image at ghcr.io/bioembeddings/bio_embeddings.

ORCID

Tobias Olenyi https://orcid.org/0000-0002-6315-0458
Céline Marquet https://orcid.org/0000-0002-8691-5791
Michael Heinzinger https://orcid.org/0000-0002-9601-3580
Michael Bernhofer https://orcid.org/0000-0001-6103-3306
Konstantin Schütze https://orcid.org/0000-0002-3957-412X
Maria Littmann https://orcid.org/0000-0001-8533-8163
Milot Mirdita https://orcid.org/0000-0001-8637-6719
Christian Dallago https://orcid.org/0000-0003-4650-6181
Burkhard Rost https://orcid.org/0000-0003-0179-8424

REFERENCES

Abriata LA, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins. 2018;86(Suppl. 1):97–112.

Ahdritz G, Bouatta N, Kadyan S, Xia Q, Gerecke W, AlQuraishi M. OpenFold. 2021.

Alexander-Brett JM, Kober DL. Triggering receptor expressed on myeloid cells 2. 2015. https://doi.org/10.2210/pdb5ELI/pdb

Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16:1315–22.

Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33:3387–95.

Baek M, Baker D. Deep learning and protein structure modeling. Nat Methods. 2022;19:13–4.

Ben Chorin A, Masrati G, Kessel A, Narunsky A, Sprinzak J, Lahav S, et al. ConSurf-DB: an accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. Protein Sci. 2020;29:258–67.

Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE TPAMI. 2013;35:1798–828.

Bepler T, Berger B. Learning protein sequence embeddings using information from structure. (2019). arXiv.

Bepler T, Berger B. Learning the protein language: evolution, structure, and function. Cell Systems. 2021;12:654–669.e653.

Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, et al. ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics. 2004;20:1322–4.

Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, et al. PredictProtein - predicting protein structure and function for 29 years. NAR. 2021;49:W535–40.

Bernhofer M, Rost B. TMbed – Transmembrane proteins predicted through Language Model embeddings. BMC Bioinformatics. 2022;23:326.

Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, et al. Using deep learning to annotate the protein universe. Nat Biotechnol. 2022;40:932–7.

Bordin N, Sillitoe I, Nallapareddy V, Rauer C, Lam SD, Waman VP, et al. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. bioRxiv. 2022.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.

Cardim-Pires TR, Sant'Anna R, Foguel D. Peptides derived from gp43, the most antigenic protein from Paracoccidioides Brasiliensis, form amyloid fibrils in vitro: implications for vaccine development. Sci Rep. 2021;11:23440.

Chowdhary KR. Natural language processing. In: Chowdhary KR, editor. Fundamentals of artificial intelligence. New Delhi: Springer India; 2020. p. 603–49.

Cid FP, Rilling JI, Graether SP, Bravo LA, Mora MLL, Jorquera MA. Properties and biotechnological applications of ice-binding proteins in bacteria. FEMS Microbiol Lett. 2016;363.

Dallago C, Goldberg T, Andrade-Navarro MA, Alanis-Lobato G, Rost B. Visualizing human protein-protein interactions and subcellular localizations on cell images through CellMap. Curr Protocol Bioinf. 2020;69:e97.

Dallago C, Schütze K, Heinzinger M, Olenyi T, Littmann M, Lu AX, et al. Learned Embeddings from deep learning to visualize and predict protein sets. Curr Protocol. 2021;1:e113.

Dass R, Mulder FAA, Nielsen JT. ODiNPred: comprehensive prediction of protein order and disorder. Sci Rep. 2020;10:14780.

El-Mabrouk N, Slonim DK. ISMB 2020 proceedings. Bioinformatics. 2020;36:i1–2.

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. 2021;1.

Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinform. 2019;20:723.

Heinzinger M, Littmann M, Sillitoe I, Bordin N, Orengo C, Rost B. Contrastive learning on protein embeddings enlightens midnight zone. NAR Gen Bioinf. 2022;4:lqac043.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. PNAS. 1992;89:10915–9.

Hie BL, Yang KK, Kim PS. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. Cell Systems. 2022;13:274–285.e276.

Høie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, et al. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. Nucleic Acids Res. 2022;50:W510–15.

Ilzhoefer D, Heinzinger M, Rost B. SETH predicts nuances of residue disorder from protein embeddings. bioRxiv. 2022.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.

Kober DL, Alexander-Brett JM, Karch CM, Cruchaga C, Colonna M, Holtzman MJ, et al. Neurodegenerative disease mutations in TREM2 reveal a functional surface and distinct loss-of-function mechanisms. Elife. 2016;5:e20391.

Kouba T, Vogel D, Thorkelsson SR, Quemin ERJ, Williams HM, Milewski M, et al. Conformational changes in Lassa virus L protein associated with promoter binding and RNA synthesis activity. Nat Commun. 2021;12:7018.

Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—round XIV. Proteins. 2021;89:1607–17.

Laine E, Karami Y, Carbone A. GEMME: a simple and fast global Epistatic model predicting mutational effects. Mol Biol Evol. 2019;36:2604–19.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv. 2022. https://doi.org/10.1101/2022.07.20.500902

Littmann M, Bordin N, Heinzinger M, Schütze K, Dallago C, Orengo C, et al. Clustering FunFams using sequence embeddings improves EC purity. Bioinformatics. 2021;37:3449–55.

Littmann M, Heinzinger M, Dallago C, Olenyi T, Rost B. Embeddings from deep learning transfer GO annotations beyond homology. Sci Rep. 2021;11:1160.

Littmann M, Heinzinger M, Dallago C, Weissenow K, Rost B. Protein embeddings and deep learning predict binding residues for various ligand classes. Sci Rep. 2021;11:23916.

Madani A, McCann B, Naik N, Keskar NS, Anand N, Eguchi RR, et al. ProGen: language modeling for protein generation. bioRxiv. 2020.

Marks C, Hummer AM, Chin M, Deane CM. Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics. 2021;37:4041–7.

Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, et al. Embeddings from protein language models predict conservation and variant effects. Hum Genet. 2021;141:1629–47.

Marx V. Method of the year: protein structure prediction. Nat Methods. 2022;19:5–10. PMID: 35017741 {Medline}.

Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems: Curran Associates, Inc.; 2021. p. 29287–303.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19:679–82.

Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics. 2019;35:2856–8.

Moore GE. Cramming more components onto integrated circuits. Electronics. 1965;38:114.

Nielsen JT, Mulder FAA. Quantitative protein disorder assessment using NMR chemical shifts. Methods Mol Biol. 2020;2141:303–17.

Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J. 2021;19:1750–8.

Perdigao N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, et al. Unexpected features of the dark proteome. Proc Natl Acad Sci U S A PMID: 26578815 {Medline}. 2015;112:15898–903.

Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder, conditional folding and AlphaFold2. bioRxiv. 2022. https://doi.org/10.1101/2022.03.03.482768

Redl I, Fisicaro C, Dutton O, Hoffmann F, Henderson L, Owens BMJ, et al. ADOPT: intrinsic protein disorder prediction through deep bidirectional transformers. bioRxiv. 2022. https://doi.org/10.1101/2022.05.25.493416

Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 2018;15:816–22.

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS. 2021;118: e2016239118.

Rost B, Sander C. Jury returns on structure prediction. Nature. 1992;360:540.

Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 1993;232:584–99.

Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct. 1996; 25:113–36.

Rost B, Sander C, Schneider R. PHD-an automatic mail server for protein secondary structure prediction. Bioinformatics. 1994;10: 53–60.

Schaeffer M, Teixeira D, neXtProt T, Nikitin F, Amos B. calipho-sib/feature-viewer: feature-viewer - DOI (v1.0.0) (2017) Zenodo.

Schelling M, Hopf TA, Rost B. Evolutionary couplings and sequence variation effect predict protein binding sites. Proteins. 2018;86:1064–74.

Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, et al. Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 2021;49:W431–7.

Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res. 2021;49:D266–73.

Singh J, Paliwal K, Litfin T, Singh J, Zhou Y. Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. Sci Rep. 2022;12: 7607.

Stärk H, Dallago C, Heinzinger M, Rost B. Light attention predicts protein location from the language of life. Bioinform Adv. 2021; 1:vbab035.

Steinegger M, Mirdita M, Soding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods. 2019;16:603–6. PMID: 31235882 {Medline}.

Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9:2542.

Sudom A, Talreja S, Danao J, Bragg E, Kegel R, Min X, et al. Crystal structure of mutant Ig-like domain; 2016.

Sudom A, Talreja S, Danao J, Bragg E, Kegel R, Min X, et al. Molecular basis for the loss-of-function effects of the Alzheimer's disease-associated R47H variant of the immune receptor TREM2. J Biol Chem. 2018;293:12634–46.

The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.

Theis TN, Wong HSP. The end of Moore's law: a new beginning for information technology. Comput Sci Eng. 2017;19:41–50.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50: D439–44.

Weissenow K, Heinzinger M, Rost B. Protein language model embeddings for fast, accurate, alignment-free protein structure prediction. Structure. 2022;30:1169–1177.e4.

Wilson CJ, Choy W-Y, Karttunen M. AlphaFold2: a role for disordered protein/region prediction? Int J Mol Sci. 2022;23:4591. https://doi.org/10.3390/ijms23094591 {Medline}.

Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, et al. High-resolution de novo structure prediction from primary sequence. bioRxiv. 2022.

Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, et al. PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic Acids Res. 2014;42:W337–43.

Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res. 2013;41:D1096–103.

Zhao D, Liu W, Chen K, Wu Z, Yang H, Xu Y. Structure of the human RNA polymerase I elongation complex. Cell Discovery. 2021;7:97.

Zhou G, Chen M, Ju CJT, Wang Z, Jiang J-Y, Wang W. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. NAR Gen Bioinf. 2020; 2:lqaa015.

Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol. 2019;20:244.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Olenyi T, Marquet C, Heinzinger M, Kröger B, Nikolova T, Bernhofer M, et al. LambdaPP: Fast and accessible protein-specific phenotype predictions. Protein Science. 2023;32(1):e4524. https://doi.org/10.1002/pro.4524