W44–W50 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 24 May 2022*
https://doi.org/10.1093/nar/gkac388

AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods

Wanda Niemyska[1,2], Pawel Rubach[1,3], Bartosz A. Gren[1], Mai Lan Nguyen[1,4], Wojciech Garstka[1,5], Fernando Bruno da Silva[1], Eric J. Rawdon[6] and Joanna I. Sulkowska[1,*]

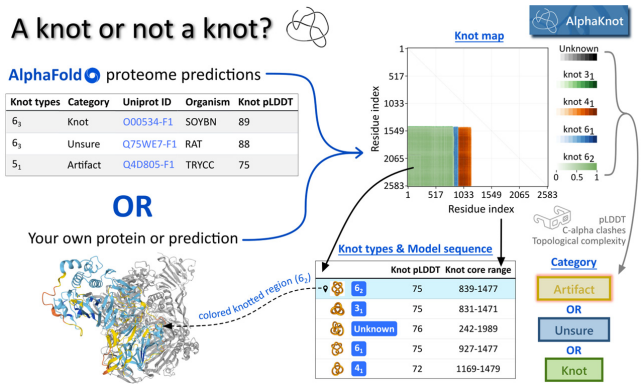
[1]Centre of New Technologies, University of Warsaw, Banacha 2c, Warsaw, Poland, [2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, [3]Warsaw School of Economics, Al. Niepodleglosci 162, 02-554, Warsaw, Poland, [4]Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland, [5]Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland and [6]University of St. Thomas, 2115 Summit Ave, Saint Paul, MN 55105, USA

Received March 29, 2022; Revised April 20, 2022; Editorial Decision April 29, 2022; Accepted May 06, 2022

ABSTRACT

AlphaKnot is a server that measures entanglement in AlphaFold-solved protein models while considering pLDDT confidence values. AlphaKnot has two main functions: (i) providing researchers with a web-server for analyzing knotting in their own AlphaFold predictions and (ii) providing a database of knotting in AlphaFold predictions from the 21 proteomes for which models have been published prior to 2022. The knotting is defined in a probabilistic fashion. The knotting complexity of proteins is presented in the form of a matrix diagram which shows users the knot type for the entire polypeptide chain and for each of its subchains. The dominant knot types as well as the computed locations of the knot cores (i.e. minimal portions of protein backbones that form a given knot type) are shown for each protein structure. Based mainly on the pLDDT confidence values, entanglements are classified as Knots, Unsure, and Artifacts. The database portion of the server can be used, for example, to examine protein geometry and entanglement-function correlations, as a reference set for protein modeling, and for facilitating evolutional studies. The AlphaKnot server can be found at https://alphaknot.cent.uw.edu.pl/.

GRAPHICAL ABSTRACT



INTRODUCTION

The presence of entanglements in proteins is an important phenomenon, inspiring multidisciplinary research involving biology, biophysics, chemistry and mathematics. Proteins with knots and slipknots, intensively studied over last years, provide an important example of this phenomenon. However, the 3D structures of knotted proteins are not easily discovered experimentally.

Currently around 2% (1) of solved protein structures from the PDB are considered to contain non-trivial topologies: knots, slipknots, or links (2). However, the PDB contains only experimentally-solved structures and, therefore, is not a representative set of structures. The set of solved structures is not sufficient to answer the most important questions about the biological role of knots, their evolution, and the identification of sequence motifs responsible for coding non-trivial 3D topologies (3).

However, in the last year the amount of data crucial for structural biology increased significantly as a result of ground-breaking developments based on methods from ar-

*To whom correspondence should be addressed. Tel: +48 22 55 43 675; Fax: +48 22 822 02 11 (Ext 320); Email: j.sulkowska@cent.uw.edu.pl

tificial intelligence: the AlphaFold 2 (AF 2) program by DeepMind (4) predicted structures of >990 000 proteins from 48 complete proteomes, which exceeds the amount of experimental data gathered to date. The quality of some of the data is extraordinary. Furthermore, analysis of AF 2 results for the human genome shows that AF 2 predicts knot types not observed in experimentally-solved structures, as well as new knotted families (5). On the other hand, since there is no sufficient training set, the AlphaFold database includes families whose members (homological proteins from different organisms) possess high pLDDT scores but different topologies, implying that some of the predictions form artificial knots (3) and, thus, are likely not accurate models.

Herein, we present the AlphaKnot server—the first server to identify entanglements in AlphaFold-solved protein models while taking into account the pLDDT confidence values. AlphaKnot acts as a submission-server and as a database for some AlphaFold published predictions. First, the submission-server provides a simple interface that allows researchers to upload AlphaFold predicted models for full topological analysis. Users can upload single chains or multiple models (including multiple models that overlap), in which case the server displays a visual comparison of the knotting in different models. Second, AlphaKnot is a database for the 21 proteomes for which AlphaFold models have been published prior to December 2021 with a browsing feature that allows for filtering by different criteria (such as certain proteomes, knot types, and core lengths).

Thus, in addition to finding entanglements in uploaded or AlphaFold-predicted proteins, AlphaKnot can be used as a tool for improving structure prediction. For example, topological differences between related proteins in combination with pLDDT values could be used to reveal potential areas of improvement (or strengths and weaknesses) for AlphaFold predictions.

MATERIALS AND METHODS

Detection of knots

The algorithm computes knotting in the protein backbone using the coordinates of Cα atoms based on a CIF or PDB file (for details see on-line help). The general idea of knot identification follows earlier works (1).

The full structure (processed either by the server or in the database) is first checked using the HOMFLY-PT polynomial (in the case of probabilistic closing, with a high number 1000 of random closures). If a non-trivial knot is found, then the structure is further processed and the knot core or whole knot map is calculated. By 'finding a knot using probabilistic closure' we mean that the dominant knot type is non-trivial with a probability of at least 48%. When computing a knot map, the faster Alexander polynomial is used. Once we have computed the knot map, we can determine the knot fingerprint and knot cores of all knot types formed by the structure. The whole algorithm is implemented in C/C++.

Graphical content and structure smoothing

The structures are visualized using PDBe Mol* v1.2 which was modified via JavaScript to highlight subchains using different coloring schemes and to hide side chains after highlighting. The PDBe version of Mol* was used to take advantage of implemented confidence coloring and pLDDT information for AlphaFold structures. We provide an additional menu bar to allow users to select the coloring mode, the representation mode (cartoon or backbone) and to highlight interesting subchains in the structure.

The so-called knot map image is generated using the Matplotlib library and manipulated by D3 and JQuery libraries to make an interactive information box of residue index, to highlight a knot in the structure, and to display information corresponding to a given knot cutoff. Users can also display arrows to show interpretations of different knots in the structure, i.e. the knot core, tails, their locations, and their lengths. The web interface is built with the newest versions of popular web libraries, e.g. Bootstrap 5, including the implementation of some simple interface animations and transitions.

Server and database implementation

The frontend of the server is implemented in Python 3 using the Flask framework. The data is stored in a MySQL database and accessed using SQLAlchemy. An asynchronous cluster tasks management solution—the kafka−slurm−agent (https://github.com/prubach/kafka-slurm-agent) was developed to enable seamless large-scale computations necessary to obtain the presented results for all organisms. Thanks to this solution, the knot detection and identification is computed on three independent Linux clusters managed by slurm (https://slurm.schedmd.com/). The kafka-slurm-agent is also used to manage the computing of tasks submitted by online users to our server. The knot detection and identification is computed using the Topoly package (6). Information about the proteins is downloaded from PDBe, RCSB and PFAM SIFTS and RESTful services. The whole service is installed on multicore Linux nodes.

SERVER DESCRIPTION

AlphaKnot consists of two parts: the submission server and the database.

Submission server

The server part of AlphaKnot performs a comprehensive topological analysis of single or multiple models uploaded by users, in the CIF, CIF.gz or PDB formats. The files can be models predicted by AlphaFold, in which case the topology confidence analysis will be most complete. But users can also upload PDB files resulting from other sources, e.g. from RoseTTa (7) predictions or from the RCSB database.

Users can select from several choices which affect the speed and accuracy of the computations. Figure 1 shows the submission page on the left with the various choices on the right. For the lists on the right, the options are shown with the speed of the computation from fastest to slowest. The results page, an example of which is shown in Figure 2, shows the types of knots, pLDDT confidence information for the knotted chain, an image matrix showing the positions of

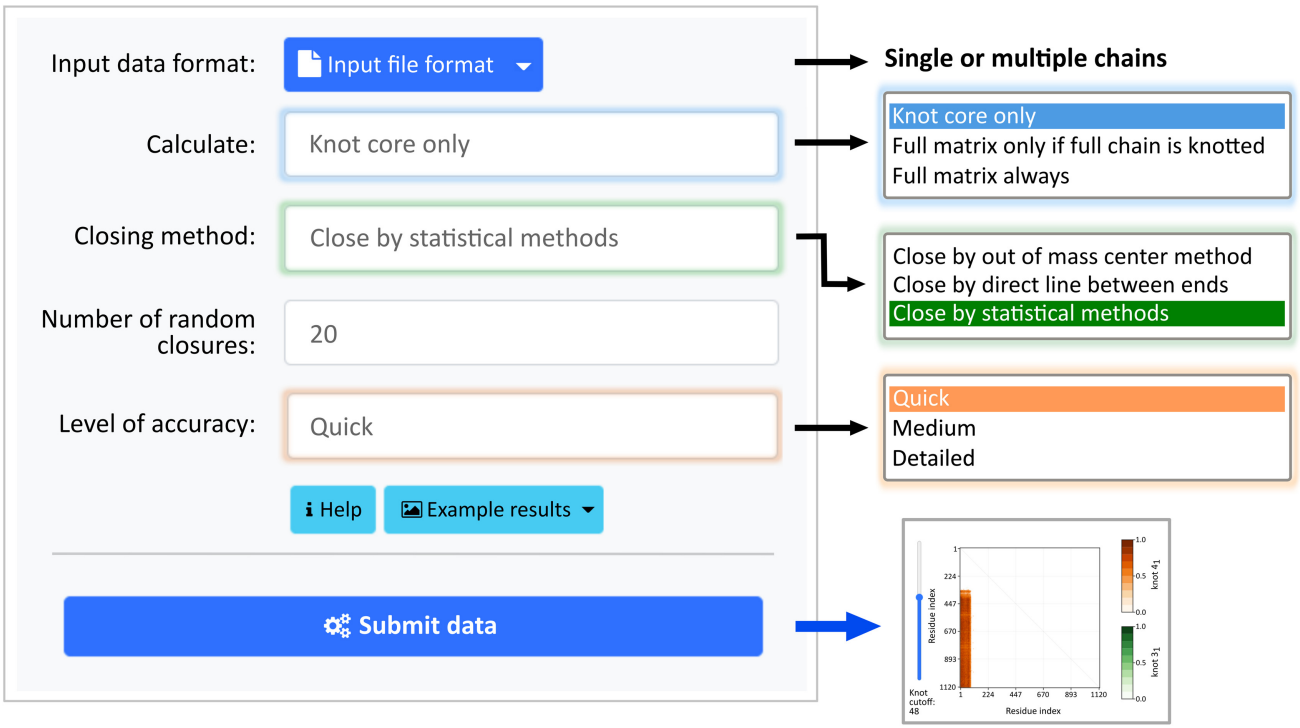


Figure 1. Submission options. The user can choose from several options which determine what methods will be used to describe the entanglement in the structure, and also determine the accuracy of the calculations. The calculation time on the server is directly related to the selected parameters (and to the length and complexity of the structure). For a single structure, computation time varies from a few seconds to several hours.

knotting along the chain, a 3D manipulatable model of the protein with colored knot locations, information about the position of the knot along the amino acid sequence, and other information about the protein.

Database of AlphaFold predictions from 21 proteomes

The database part of AlphaKnot contains knotting information for structures predicted prior to 2022 based on 21 proteomes (with pLDDT >50) published in (4). A browse feature allows users to see and compare all of the knotted proteins. The proteins are categorized manually (by visual inspection) into the categories of Knots (high confidence), Unsure (significant, but not insurmountable, issues exist in the model), and Artifact (very likely the entanglement is due to issues with the model).

This classification does not always agree with pLDDT confidence, e.g. in the Artifacts, users can find proteins with high pLDDT values but very unlikely topologies. Some unusual structures were also checked with the RoseTTa tool (7). For example, proteins in which AlphaFold predicted the new 5_1 knot type, were classified in AlphaKnot as Unsure (e.g. P53336) and Artifact (e.g. Q75JZ0) due to low pLDDT values and the fact that all known models of methyltransferases analyzed to date form 3_1 knots. (8). RoseTTa's models of these structures confirm AlphaKnot's classification, P53336 forms the 5_1 knot, and Q75JZ0 is unknotted. One way to explain the unknotted topology in the case of Q75JZ0 is that it could be due to the wrong template being used for prediction: the protein with PDB code 5ZYO (a protein which is artificially unknotted). We do not see the origin of the 5_1 knot. It would be interesting to conduct an experiment to investigate the existence of the 5_1 knot (potentially the first non-twisted knot type in a protein (9)) in these structures.

A search results in a list which links to an individual page for each protein. The individual pages contain the same information as is shown for submitted jobs, as well as other information about the protein and a list of proteins with similar sequences.

Features unique to AlphaKnot

The functions available on AlphaKnot are not available on any other server/website/software. AlphaFold's pLDDT measures the confidence of a predicted model, computed per amino acid. This is critical information which is needed in order to interpret the reliability of the knotting within a model. While there are very few resources for researchers trying to analyze entanglement in chains in general, AlphaKnot is the only platform for computing knotting of protein chains which includes pLDDT values. For user-submitted files, or within the included database, AlphaKnot identifies the knot cores, i.e. shortest chains realizing a given knot, within the chain and uses the pLDDT values to characterize to what extent the results should be trusted. Furthermore, in the 3D model, the full chain (or a given subchain) can be colored relative to the pLDDT values to allow the user to visualize the confidence one might have in the model. The other coloring modes, default (which is monochrome) and

Figure 2. Example page with the output from a submission to AlphaKnot. Here the multiple chains option was chosen and four proteins with at least 40% sequence similarity were analyzed: I1L257, Q2JIZ5, P93527, B4YB07 (UniProtKB ID). (**A**) Job status table. The table shows details of the job, information about the methods used, the chosen parameters for the job, and the main results. (**B**) Structure choice. For multiple chain uploads, the user can choose the structure to view. Here I1L257 is selected, which forms the knot 4_1. (**C**) Information and Artifact check. The tables present basic information on the topology of the structure and results of some automatic tests on the quality of the structure and its knotted region. These results can help the user to recognize if the knotted topology may be an artifact—any information in red suggests that the user should be wary of the results. (**D**) The knot map (if full matrix was computed). The user can choose the cutoff—knot frequency required to show it on the matrix. (**E**) View of the structure. The user can choose the coloring method and select the fragment of the chain which is colored. In the figure the knot core region is colored by the pLDDT values, which is helpful in assessing whether the knot is an artifact. (**F**) Knot types and model sequence. The table contains detailed information about all knots formed by fragments of the chain (if full matrix was computed). By selecting a knot, the corresponding sequence is highlighted. (**G**) Comparison of the location and types of knots in each (knotted) structure. The color of the bar corresponds to the type of knot. The user can see on the chart that three of the four structures (besides Q2JIZ5) form the knot 4_1.

rainbow (where the color changes as one passes through the protein) provide modes that are more useful for visualizing the 3D nature of the protein.

Sometimes AlphaFold models are created by predicting smaller segments of the protein which overlap. As is shown in Figure 2, users can upload portions of AlphaFold predictions that overlap. In such a case, the models can disagree on the location or type of knot that is formed over a subchain. When multiple portions of a common protein are uploaded for analysis, AlphaKnot provides a tool for visualizing the similarity and differences in the knotting between the models, as is seen in Figure 2G. In that figure, we see that three of the four predictions predict a 4_1 knot, but with slightly different minimal cores. From these results, one can assess how stable the knotting is within the set of predictions. Since the knot does not appear in one of the models, the predictions are not consistent and this suggests areas within the structure that garner further attention.

Improvements over our KnotProt server

AlphaKnot also has a number of improvements over our server KnotProt (1,10). Many of these were inspired by the unique opportunities (such as the pLDDT values mentioned above) and challenges due to the size and complexity of knots observed in the AlphaFold models. First, the Topoly package (freely available, see (6)), is fully integrated into AlphaKnot and is able to recognize more complex knot types than KnotProt. More precisely on the KnotProt, knot types through eight crossings are recognized, while on the AlphaKnot, knot types through 12 crossings are recognized. Second, AlphaKnot allows users to specify a cut-off percentage in the interactive knot map. At the cut-off value of 0.3, for example, one sees all subchains whose predominate non-trivial knot type occurs with probability at least 0.3. As in KnotProt, the opacity of the color in each cell corresponds to this probability value. Increasing the cut-off value results in fewer, but more robustly knotted chains. This feature allows users to visualize how and where the knotting is created within the protein. Third, AlphaKnot computes knot cores for uploaded models. This functionality does exist in some other software for proteins, like KymoKnot (11), however without the critical pLDDT analysis specific for AlphaFold predicted structures. Fourth, AlphaKnot implements a number of speed improvements to make the calculations more efficient. The chain lengths of the protein models from AlphaFold are much larger than what is seen in the PDB. Furthermore, the complexity of knotting observed is much higher. These improvements were needed so that the submission server would be feasible.

User experience of the AlphaKnot

The AlphaKnot server and database has a number of features focused on providing an intuitive and robust user experience. First, the advanced search allows users to search by knot type, core length, tail length, or fingerprint filtered by proteome (or proteomes) and within the classes of Knots, Unsure or Artifacts. The filtering choices are shown in Figure 3. Second, by hovering over an amino acid in the 3D model or hovering over an amino acid in the sequence, one

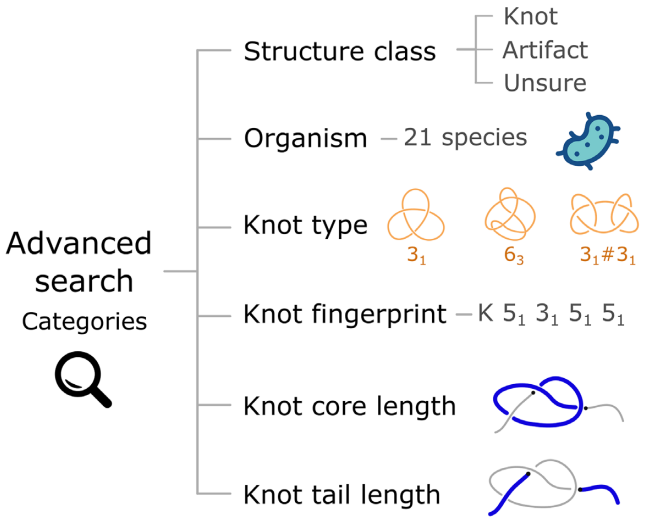


Figure 3. Features that can be searched for knotted proteins in Advanced Search in the AlphaKnot database.

sees the protein information, and the amino acid's type, number and pLDDT score for the amino acid. Third, the knot core search algorithm is improved and provides a more accurate calculation of the core and N- and C-termini tails. Fourth, each database protein (or uploaded model) provides output from a number of automated checks on the integrity of the model. For example, if the distance (in space) between amino acids is too small the user receives a warning. Also, low pLDDT scores for the full chain, and within knotted subchains, are highlighted for the user. Lastly, for every structure in the database, there are links which direct the user to the corresponding entries in UniProt and AlphaFold. When a model overlaps with an experimentally-solved structure from the PDB, the corresponding PDB and KnotProt links are also given.

ONLINE DOCUMENTATION

AlphaKnot provides a rich set of online documentation with sections (i) about—providing an overview of the website, (ii) knot detection—detailing the methods used to compute the knotting, (iii) how to interpret the knotting data—explaining how one can interpret the entanglement based on the knot map, (iv) how to use the server—showing how users can upload proteins, the choices available, the speed-effect on the different choices, and how to interpret the results from single or multiple chain output, (v) how to search and browse the database—guiding the user on how to search within the database of AlphaFold predictions, (vi) database statistics—presenting information about the protein models in the database as well as downloadable text files which list information about these models, (vii) API—documenting an API for downloading sets of data according to different search criteria.

APPLICATIONS

Artificial intelligence (AI) methods are used commonly to determine protein structures from less investigated genomes

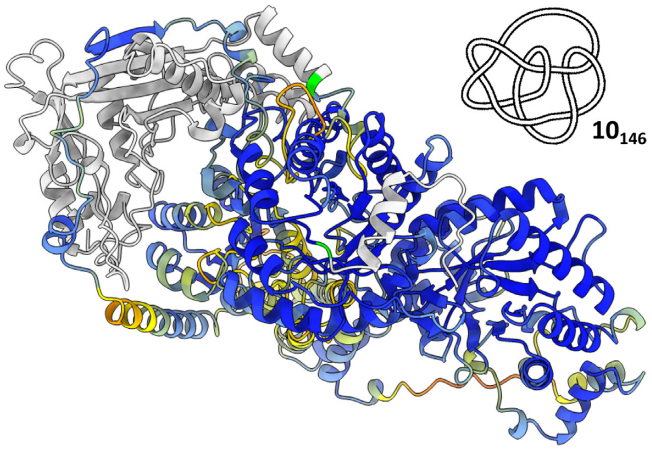


Figure 4. Protein (UniProtKB ID: Q54P92) with very high pLDDT value, but whose topology (knot with 10 crossings $K10_{146}$, shown with schematic) suggests that the prediction is not reliable.

or with unknown homological models. AI methods and the constantly growing AlphaFold database facilitate a wide range of research based on 3D protein structures. However, a careful analysis of knotting within these models can provide unique insights into the the predicted structures which can be used for many different applications. Below we provide some examples.

The analysis of all proteins from the PDB has shown that the topology is conserved between proteins with the same function but low sequence similarity (3) with one exception (12). Computations by the AlphaKnot server on proteins in AlphaFold database with the same biological function but from different organisms show that in some cases the topology indeed is strictly conserved, implying that a detected topology has a high probability of being correctly predicted. As a result, new knotted protein families have been discovered, e.g. tRNA-uridine aminocarboxypropyltransferase 1 (PF03942) with a 3_1 knot located in the active site. On the other hand, in some cases the topologies are seemingly random (5), strongly suggesting that the modeled structures are not correct, e.g. as is shown in Figure 4. There are also cases when the topology is conserved only between some homological proteins, e.g. integrin as noted in (5). Such families open the door for an evolutional analysis (determining the origin of a non-trivial topology) or even the detection of an amino acid motif responsible for creating a non-trivial topology. Also, some potential new types of knots have been found, e.g. in Putative methyltransferase (UniProtKB ID: YGR283C). This knot type, denoted 5_1 in mathematical notation, would necessitate a more complex folding path than any other knotted protein characterized to date. This knot is called potential since not all homological structures deposited in the AlpahFold possess it.

Other potential applications include in silico drug design or structural biology studies. For example proteins from the SPOUT family, such as the TrmD protein are on the list of high importance proteins due to 12 drug-resistant bacteria (based on World Health Organization) that should be treated as high-priority pathogens. This makes TrmD, a vital bacterial protein present in all of these strains, an excellent antimicrobial target for drug design. On the other hand, in all experimentally-solved structures from the SPOUT clan, the active site is embedded in a tightly packed knot even though they share low sequence similarity. Therefore, a correct prediction of topology is necessary for a successful search for a potent drug. Moreover, the majority of knotted proteins are enzymes which have an active site embedded in the knot. Accurate predictions of knots are crucial for further structural biology investigations.

We present one further example which highlights a possible issue with relying on pLDDT values and how AlphaKnot can provide critical insights in such cases. Figure 4 shows a protein structure (UniProtKB ID: Q54P92) with a high pLDDT reliability which is certainly not a realistic model from the topological perspective. The knot formed is 10_{146}, which is much more complicated than any protein deposited in the PDB. The high pLDDT values suggest a good quality fold prediction for the structure. The pLDDT values along the knot core are also high. However, the AlphaKnot topological analysis reveals that the topological arrangement is highly unlikely. We used the AlphaKnot server to analyze 27 homological structures deposited in the AlphaFold database and found that the majority have the trivial topology (although two form the 3_1 knot and some form the 10_{146} knot). Furthermore, we used RoseTTa to predict the structure for Q54P92 and found that the model is unknotted. Together, these results suggest that all members are probably unknotted. Without the use of AlphaKnot, this prediction could be interpreted as being reliable based on the pLDDT values. However, AlphaKnot has exposed that researchers should be wary of this structure.

We would like to point out that some of the potentially wrongly predicted topologies may be due to the fact that some structures in the PDB database contain gaps. Modeling long gaps as straight segments can easily change the types of knots observed. Since the structures predicted by AlphaFold do not have gaps, one may not be aware that AlphaFold used structures with gaps in the learning process. One example is the family including the sodium-driven chloride bicarbonate exchanger. The original predicted structure, based on the human genome UniProtKB ID: Q6U841, has good quality (high pLDDT score over 70) and forms a 3_1 knot. Currently 86 homologues can be found in AlphaFold. We found that 85 out of its 86 homologs from other organisms are unknotted. That is, although AF 2 achieves incredible effectiveness in protein structure prediction, there is clearly still room for improvement in specific cases.

The AlphaKnot server is based on protein analysis, therefore it has many biological applications. Nevertheless, its versatility makes it useful in other areas of research as well.

SUMMARY

AlphaKnot provides a user-friendly submission server combined with a rich database of knotting analysis for AlphaKnot predicted structures. The use of AlphaFold's pLDDT values is critical for interpreting the reliability of the knotting within models and is only found on this server. The 3D models colored by pLDDT and multi-chain analysis for overlapping protein subchains allows users to efficiently vi-

sualize the quality of the models. In addition, AlphaKnot provides unique insights into the reliability of predictions which are not available by using the pLDDT alone. Furthermore, the AlphaKnot server is a tool to conduct fast topological analysis of sets of related structures (e.g. from different organisms) to assess the quality of the predictions and suggest possible improvements for the AlphaFold algorithm.

DATA AVAILABILITY

The AlphaKnot web server, with the online documentation, is publicly available at https://alphaknot.cent.uw.edu.pl.

ACKNOWLEDGEMENTS

We thank Center for Machine Learning from University of Warsaw (Center4ML) for help with modeling protein structures using AlphaFold v2.0 pipeline.

FUNDING

National Science Centre [UMO-2018/31/B/NZ1/04016 to J.I.S.]; European Biology Organization [EMBO Installation Grant 2057 to J.I.S.]; National Science Foundation [1720342 to E.J.R.]. Funding for open access charge: National Science Centre [UMO-2018/31/B/NZ1/04016 to J.I.S.]
Conflict of interest statement. None declared.

REFERENCES

1. Jamroz,M., Niemyska,W., Rawdon,E.J., Stasiak,A., Millett,K.C., Sułkowski,P. and Sulkowska,J.I. (2015) KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res.*, **43**, D306–D314.
2. Sulkowska,J.I. (2020) On folding of entangled proteins: knots, lassos, links and θ-curves. *Curr. Opin. Struct. Biol.*, **60**, 131–141.
3. Sułkowska,J.I., Rawdon,E.J., Millett,K.C., Onuchic,J.N. and Stasiak,A. (2012) Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Nat. Acad. Sci. U.S.A.*, **109**, E1715–E1723.
4. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žìdek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
5. Perlinska,A.P., Niemyska,W.H., Gren,B.A., Rubach,P. and Sulkowska,J.I. (2022) New 63 knot and other knots in human proteome from AlphaFold predictions. bioRxiv doi: https://doi.org/10.1101/2021.12.30.474018, 01 January 2022, preprint: not peer reviewed.
6. Dabrowski-Tumanski,P., Rubach,P., Niemyska,W., Gren,B.A. and Sulkowska,J.I. (2020) Topoly: Python package to analyze topology of polymers. *Brief. Bioinformatics*, **22**, bbaa196.
7. 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.
8. Jarmolinska,A.I., Perlinska,A.P., Runkel,R., Trefz,B., Ginn,H.M., Virnau,P. and Sulkowska,J.I. (2019) Proteins' knotty problems. *J. Mol. Biol.*, **431**, 244–257.
9. Bölinger,D., Sułkowska,J.I., Hsu,H.-P., Mirny,L.A., Kardar,M., Onuchic,J.N. and Virnau,P. (2010) A Stevedore's protein knot. *PLoS Comput. Biol.*, **6**, e1000731.
10. Dabrowski-Tumanski,P., Rubach,P., Goundaroulis,D., Dorier,J., Sułkowski,P., Millett,K.C., Rawdon,E.J., Stasiak,A. and Sułkowska,J.I. (2018) KnotProt 2.0: a database of proteins with knots and other entangled structures. *Nucleic Acids Res.*, **47**, D367–D375.
11. Tubiana,L., Polles,G., Orlandini,E. and Micheletti,C. (2018) KymoKnot: a web server and software package to identify and locate knots in trajectories of linear or circular polymers. *Eur. Phys. J. E*, **41**, 72.
12. Virnau,P., Mirny,L.A. and Kardar,M. (2006) Intricate knots in proteins: function and evolution. *PLoS Comput. Biol.*, **2**, e122.