Nucleic Acids Research, 2024, **52**, W182–W186
https://doi.org/10.1093/nar/gkae397
Advance access publication date: 15 May 2024
Web Server issue

AlphaFind: discover structure similarity across the proteome in AlphaFold DB

David Procházka [1,†], Terézia Slaninákova [1,2,†], Jaroslav Olha [1,2], Adrián Rošinec [2,3,4], Katarína Grešová [4], Miriama Jánošová[1], Jakub Čillík [2], Jana Porubská [3,4], Radka Svobodová [3,4], Vlastislav Dohnal [1] and Matej Antol [1,2,*]

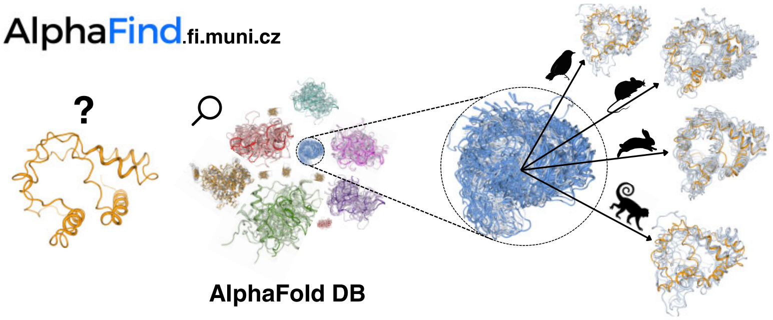
[1]Faculty of Informatics, Masaryk University, Botanická 68A, Brno 60200, Czech Republic
[2]Institute of Computer Science, Masaryk University, Šumavská 416/15, Brno 60200, Czech Republic
[3]Biological Data Management and Analysis Core Facility, CEITEC—Central European Institute of Technology, Masaryk University, Studentská, Brno 62500, Czech Republic
[4]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, Brno 62500, Czech Republic

[*]To whom correspondence should be addressed. Tel: +420 54949 6787; Email: antol@muni.cz
[†]The first two authors should be regarded as Joint First Authors.

Abstract

AlphaFind is a web-based search engine that provides fast structure-based retrieval in the entire set of AlphaFold DB structures. Unlike other protein processing tools, AlphaFind is focused entirely on tertiary structure, automatically extracting the main 3D features of each protein chain and using a machine learning model to find the most similar structures. This indexing approach and the 3D feature extraction method used by AlphaFind have both demonstrated remarkable scalability to large datasets as well as to large protein structures. The web application itself has been designed with a focus on clarity and ease of use. The searcher accepts any valid UniProt ID, Protein Data Bank ID or gene symbol as input, and returns a set of similar protein chains from AlphaFold DB, including various similarity metrics between the query and each of the retrieved results. In addition to the main search functionality, the application provides 3D visualizations of protein structure superpositions in order to allow researchers to instantly analyze the structural similarity of the retrieved results. The AlphaFind web application is available online for free and without any registration at https://alphafind.fi.muni.cz.

Graphical abstract



Introduction

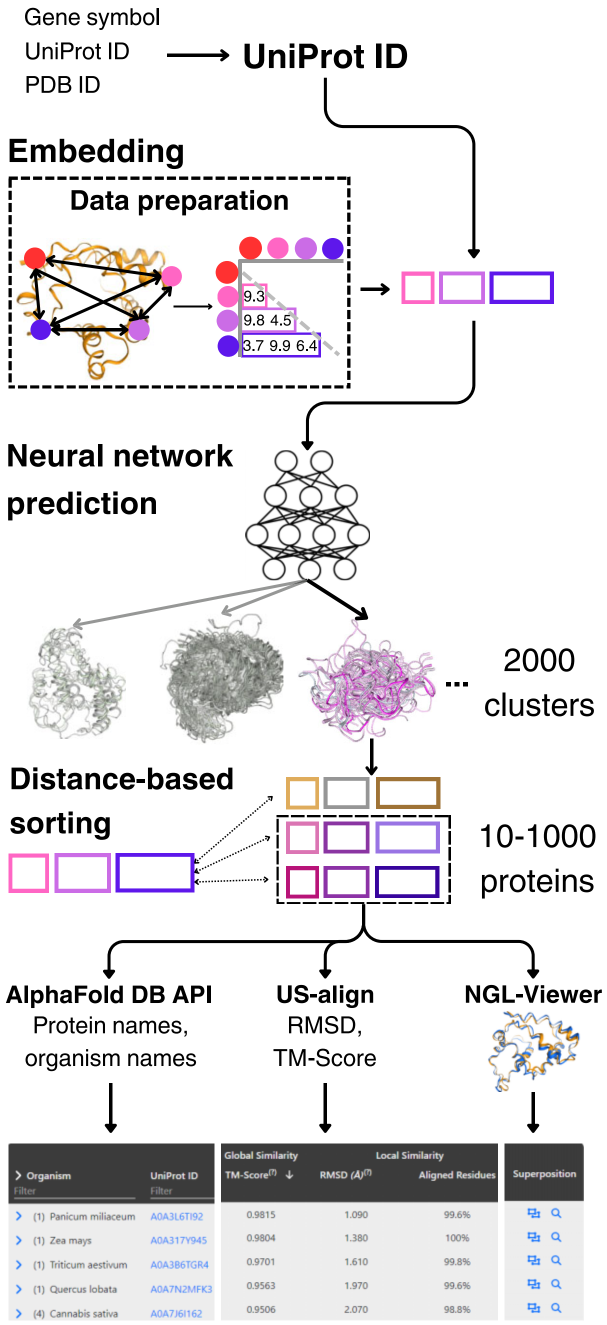
Protein structural data are highly beneficial scientific resources that serve as the basis for ever-growing and valuable research. Thanks to seven decades of intensive research, we currently have >200 000 experimental protein 3D structures deposited in the Protein Data Bank (PDB) (1). Furthermore, the AlphaFold algorithm (2), trained on these experimental data, produces highly reliable protein 3D structure predictions. As a result, the AlphaFold database has been published (3), containing >200 million 3D protein structures. In parallel, other databases of predicted protein 3D structures have been published, such as ESM Metagenomic Atlas (4), collecting 600 million protein 3D structures. Various research fields in bioinformatics benefit from such databases, with new cutting-edge technology on the horizon (5).

To take full advantage of this enormous amount of data, it is essential to organize them efficiently. Specifically, it is necessary to perform a structure-based search, because structural similarities frequently imply functional correspondence, even without high sequence similarity.

Unfortunately, conventional protein structure tools (e.g. PDBeFold) are not able to handle such huge datasets. To address this issue, novel searching tools have been developed, e.g. Foldseek (6), 3D-SURFER (7) or Dali server (8).

Received: January 30, 2024. **Revised:** April 10, 2024. **Editorial Decision:** April 28, 2024. **Accepted:** April 30, 2024

However, their functionality has limitations. The main limitation of tools such as Dali server or 3D-SURFER is that their methods do not scale well to large datasets. Foldseek search (https://search.foldseek.com/), while also not supporting search in the whole 214-million AlphaFold DB (instead using a pre-clustered 52-million subset AFDB50), handles large protein datasets very well by converting the local residue interactions into a sequence of structural patterns—this allows Foldseek to exploit various well-established and extremely efficient sequence searching techniques. However, its localized approach of focusing only on the local interactions between each residue and its closest neighbor has its own limitations when searching for broader similarity patterns within entire structures.



Organism	UniProt ID	Global Similarity	Local Similarity		Superposition
Filter	Filter	TM-Score(?) ↓	RMSD (Å)(?)	Aligned Residues	
(1) Panicum miliaceum	A0A3L6TI92	0.9815	1.090	99.6%	
(1) Zea mays	A0A317Y945	0.9804	1.380	100%	
(1) Triticum aestivum	A0A3B6TGR4	0.9701	1.610	99.8%	
(1) Quercus lobata	A0A7N2MFK3	0.9563	1.970	99.6%	
(4) Cannabis sativa	A0A7J6I162	0.9506	2.070	98.8%	

Figure 1. Visual representation of AlphaFind's workflow.

Description of the web server

The application is available as an online service free of charge. It is composed of two integral components—a front end and a back end. The front end is accessible via an internet browser, communicating with the back end that evaluates queries and returns results back to the front end. The back end starts with an offline phase, where it pre-processes the data and creates a search index for rapid search request evaluation in the online phase.

Protein structure representation

We utilize the compact data embedding method described in (9) in conjunction with data clustering and machine learning. This approach captures the semantic relationships between protein structures and quickly identifies the most relevant groups of data for a given query. The embedding of a protein is essentially an extremely compressed representation of its 3D structure, which only accounts for the broad structural features of the whole structure while neglecting primary or secondary structure information.

Data preparation/indexation

In the offline phase, we first extract semantic information from raw *cif* files into vector embeddings, thereby compressing the original AlphaFold DB size from ∼23 TiB into ∼20 GiB and obtaining representation that is more suitable for data processing algorithms. Then, we build an index on the embeddings using a learned indexing approach. Learned indexes form a new research stream defined by Kraska *et al.* for 1D numeric data (10). This research direction was later expanded to the domain of complex data where data items are compared using a distance function. The distance expresses complex similarity beyond mere comparison of two integers, as shown in (11,12). The latter two methods in conjunction with (9) establish the basis of the indexing solution presented here.

Workflow

AlphaFind was developed to provide an intuitive and useful interface for discovering structurally similar proteins within AlphaFold DB. The workflow, as depicted in Figure 1, was optimized and tested for maximum computational efficiency to provide fast and accurate results. The following steps take place from the moment that the user poses a query:

(1) *Translating the input into a UniProt ID*: AlphaFind supports three forms of input: UniProt ID, PDB ID and gene symbol. Since UniProt ID is internally used to identify a protein, other forms of input must be translated into UniProt ID using publicly available application programming interfaces (APIs). For PDB ID to UniProt ID conversion, we use https://www.ebi.ac.uk/pdbe/api/mappings/uniprot/, and for gene symbol to UniProt ID conversion, AlphaFind relies on https://rest.uniprot.org/idmapping.

(2) *Searching for a set of candidate proteins*: Using the UniProt ID, the server finds the associated protein structure embedding created during data preparation. This embedding is served to the index (a fully connected neural network), which returns the top 10 most similar clus-

ters to the embedding, narrowing down the set of candidates from 214 million to ∼10 million. Then, Euclidean distance is evaluated between the input and each candidate object embedding (13). We sort such candidates by this proximity and take just 1000 closest proteins, which reduces the search space even further.

(3) *Evaluating global and local similarity*: Based on the desired result set size ($n = 50$, initially), the nearest n proteins out of the candidate set are selected to evaluate local (root-mean-square deviation, RMSD) and global (TM-score) similarity, together with *aligned residues* and *sequence identity*, using the US-align tool (14). This is the most time-intensive step of the search process, taking up to several tens of seconds.

(4) *Downloading metadata for query and results*: After collecting the search results, AlphaFind uses the AlphaFold DB's API to collect useful information, such as the name of the protein structure and the host organism name, and groups the results by organism name for easier orientation.

(5) *Visualizing protein structures' overlap*: To represent the protein structure similarity visually, AlphaFind utilizes NGL Viewer (15) to display an overlay of every pair of the input protein and a result protein. Additionally, for a more detailed view, each result points to the Mol* Viewer (16).

(6) *Showing more results*: The user has the option to expand the search results by additional candidate proteins identified in step 2: they can choose to display 50, 100, 200 or 300 more. If this option is selected, steps 3–5 repeat, and the result table is updated.

(7) *Downloading the results*: Once searching is finished, the user can download the results for future use or archival in the CSV format using the 'Export all to CSV' button.

To reduce response times, once steps 2 and 3 are executed for a given input protein, the web server stores the computed results, causing a significant reduction in subsequent waiting times when searching for the same protein.

Implementation

The front end was implemented in TypeScript using the React framework (https://reactjs.org/) and the Bootstrap library (https://getbootstrap.com/). The back end is written in Python 3.10 with Flask (https://flask.palletsprojects.com/).

We tested AlphaFind on a diverse set of proteins varying in size, complexity and quality. AlphaFind provided biologically relevant results even for small, large and lower quality structures. When AlphaFind did not offer structures with high TM-scores, the results remained biologically relevant. The performance of AlphaFind is scalable.

Limitations

AlphaFind is constructed on AlphaFold DB in version 3, predating the v4 update. The application returns up to 1000 similar protein structures to a single query. The results returned by AlphaFind are approximate, and as such, they may not necessarily contain all the most structurally similar proteins present in AlphaFold DB.

Approximate searching trades off resource demands (such as hardware, time and money) for precision. It is customary in complex data management, as the 'objective answer', e.g. which protein structure is more similar to another, does not have an indisputable answer in the first place. Moreover, the design of the application allows for loading additional data, which expands and also refines the prior search answer. This functionality enables the user to find protein structures, which could have been missed in the previously shown search results.

AlphaFind processes the entire protein structure—its helices, sheets and its unstructured regions—and handles them with equal weight. Therefore, high occurrence of unstructured regions in the input structure can bias the search. This phenomenon is more prevalent in coiled-coil structures but can also be observed in some small structures.

Results and discussion

Searching characteristics

Due to the data embedding method used, AlphaFind is mainly concerned with global structural similarity, always comparing the entire structure and treating all parts with equal weight, regardless of primary or secondary structure. Unlike established searching methods, which tend to place more focus on occurrence of certain folds or high local similarity in particular regions, AlphaFind is more likely to find structural similarity across organisms where these regions are less conserved.

AlphaFind indexes the entire AlphaFold DB, which contains 214 million protein structures. The conversion to embeddings allows the application to search the database and return the first 50 results in an average of 7 s with negligible back-end load. The user can expand the obtained results in increments of 50, 100, 200 and 300, which takes, on average, additional 7, 9, 11 and 15 s, respectively. During periods of high load, the users' requests may be queued and processed with a slight delay, increasing with the number of concurrent users that interact with the application.

We provide three use cases to demonstrate AlphaFind performance in different conditions. The first use case (hemoglobin alpha 1) is a typical use case for testing of protein structure databases. Since this protein has been an object of intensive research for nearly five decades, a high number of very similar protein structures are present in PDB and AlphaFold DB, and can thus be found by AlphaFind. The second use case is cytochrome P450, a protein containing a high number of secondary structure elements and a complex architecture (17). This use case demonstrates that complex structural patterns do not inhibit AlphaFind's ability to reliably search for similar structures. The third use case shows how AlphaFind can help answer actual research questions. PIN proteins are currently the subject of intense research interest, but a determination of their 3D structure is difficult and has only been successful for a few PIN representatives. Thanks to AlphaFind, predicted structures of various PIN proteins can be collected and analyzed. We demonstrate this on the PIN5 protein, which is intensely studied by experimentalists at the moment.

Example I: hemoglobin alpha 1

Hemoglobin is a protein that facilitates the transport of oxygen and other gases in red blood cells. Almost all vertebrates contain hemoglobin. It consists of four protein subunits (globins), and is one of the first proteins whose 3D structure has been experimentally determined. There are many types of hemoglobin, with hemoglobin alpha 1 (encoded by the HBA1 gene) occurring in humans and being the main form of hemoglobin in adults.

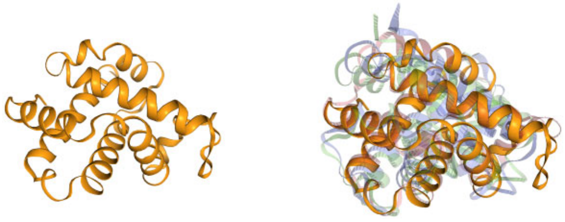


Figure 2. Hemoglobin alpha 1 from *Homo sapiens* (UniProt ID: P69905) in yellow. Examples of AlphaFind search results: hemoglobin alpha 5 from *Aquarana catesbeiana* (UniProt ID: A0A2G9QLR9) in blue, hypothetical protein from *Ranitomeya imitator* (UniProt ID: A0A821PI99) in green and globin from *Alphaproteobacteria bacterium* (UniProt ID: A0A3M1M8B2) in red.

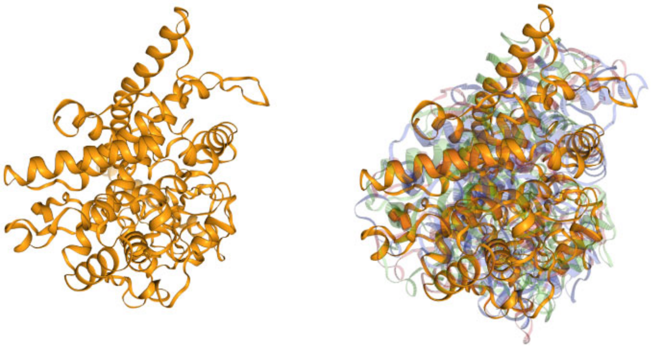


Figure 3. Cytochrome P450 family 76 subfamily C polypeptide 7 from *Z. mays* (UniProt ID: A0A1D6JW22) in yellow. Examples of AlphaFind search results: uncharacterized protein from *Oreochromis niloticus* (UniProt ID: A0A669CAC6) in blue, cytochrome P450 2C19 from *Equus caballus* (UniProt ID: A0A3Q2H4N0) in green and cytochrome P450 2H1 from *Haliaeetus albicilla* (UniProt ID: A0A091PH09) in red.

Here, AlphaFind shows us (Figure 2) that highly similar hemoglobin structures can also be found in other species.

Example II: cytochromes P450

Cytochromes P450 are enzymes that are important for the metabolism of many endogenous compounds and xenobiotics. P450 enzymes have been identified across all biological kingdoms: animals, plants, fungi, bacteria and archaea, as well as in viruses. Cytochrome P450 proteins contain one chain that is composed of >20 sheets and helices. Their sequence similarity is very low. In this use case, we can observe similarities among cytochrome P450 structures from various species (Figure 3). The search starts with a cytochrome from corn (*Zea mays*), and within the first 50 hits, we find similar structures originating from various animals (fish, eagle, mouse, cat, horse, etc.).

Example III: PIN5 protein

The PIN proteins are transmembrane proteins that regulate plant growth by influencing auxin transport from the cytosol to the extracellular space. They only occur in plants and feature a configuration of 10 main helices that collectively form a pore. Eight types of PIN proteins are known (PIN1–PIN8), and recently, the structures of three PIN proteins were uncovered and published in *Nature*. The structure of the PIN5 protein differs from other PINs (18) and has not yet been experimentally determined. This use case shows (Figure 4) that the PIN5 protein structure is strongly conserved among many different plant species.

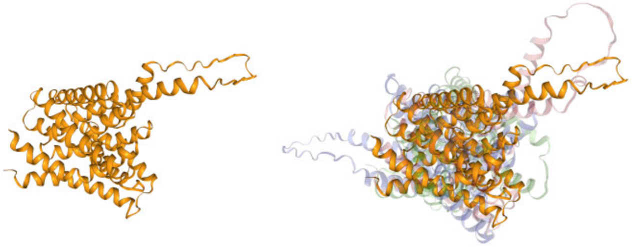


Figure 4. Auxin efflux carrier component 5 from *Arabidopsis thaliana* (UniProt ID: Q9FFD0) in yellow. Examples of AlphaFind search results: auxin efflux carrier component from *Citrus unshiu* (UniProt ID: A0A2H5NWV3) in blue, uncharacterized protein from *Eucalyptus grandis* (UniProt ID: A0A059BG64) in green and auxin efflux carrier component from *Vitis vinifera* (UniProt ID: A0A438CTF0) in red.

Conclusion

In this article, we presented AlphaFind, a novel web application for fast structure-based search of similar proteins in AlphaFold DB and PDB. AlphaFind utilizes the learned metric index (LMI) approach and a 3D feature extraction method designed for protein structures. The web application presents the search results as a table, sortable according various criteria, i.e. TM-score, RMSD and the number of aligned residues. The users can also download the results as a CSV file or visualize superpositions of input and output proteins via NGL Viewer. AlphaFind is easy to use and is platform-independent. Documentation describing its usage is referenced on the following web page: https://alphafind.fi.muni.cz.

Data availability

AlphaFind application is available at no cost and no registration at https://alphafind.fi.muni.cz. The user manual is referenced in the application and is also directly available at https://github.com/Coda-Research-Group/AlphaFind/wiki/Manual. The application's source code is accessible under the MIT license at https://github.com/Coda-Research-Group/AlphaFind and is also available on Zenodo at https://doi.org/10.5281/zenodo.11085863. The LMI is published and its performance reproducibly examined in (19). The embeddings and weights for the LMI model are available in the Czech National Repository's pilot at https://doi.org/10.48700/datst.d35zf-1ja47.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions. The authors would also like to thank Lucie Novotná for her help with the protein embeddings, Lukáš Hejtmánek for his work on the container infrastructure, and Vladimír Míč and Tomáš Raček for consultation during the early stages of our involvement with the protein structure data. Computational resources were provided by the e-INFRA CZ project (ID 90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Funding

Czech Science Foundation [GF23-07040K to D.P., T.S., J.O., V.D. and M.A.]; Ministry of Education, Youth and Sports of the Czech Republic [LM2023055 to A.R., J.P. and R.S.]. Funding for open access charge: Masaryk University as per its Read and Publish agreement with the Oxford University Press. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ [ID:90254].

Conflict of interest statement

None declared.

References

1. Burley,S.K., Bhikadiya,C., Bi,C., Sebastian,B., Chao,H., Chen,Li., Craig,P.A., Crichlow,G.V., Dalenberg,K., Duarte,J.M., *et al.* (2023) RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.*, **51**, D488–D508.
2. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Applying and improving AlphaFold at CASP14. *Proteins*, **89**, 1711–1721.
3. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2021) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
4. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
5. Varadi,M. and Velankar,S. (2023) The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*, **23**, 2200128.
6. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (20234) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.
7. La,D., Esquivel-Rodríguez,J., Venkatraman,V., Li,B., Sael,L., Ueng,S., Ahrendt,S. and Kihara,D. (2009) 3D-SURFER: software for high-throughput protein surface comparison and analysis. *Bioinformatics*, **25**, 2843–2844.
8. Holm,L. (2022) Dali server: structural unification of protein families. *Nucleic Acids Res.*, **50**, W210–W215.
9. Olha,J., Slanináková,T., Gendiar,M., Antol,M. and Dohnal,V. (2022) Learned indexing in proteins: substituting complex distance calculations with embedding and clustering techniques. In: Skopal,T., Falchi,F., Lokoč,J., Sapino,M.L., Bartolini,I. and Patella,M. (eds.) *Similarity Search and Applications*. Springer International Publishing, Cham, Switzerland, pp. 274–282.
10. Kraska,T., Beutel,A., Chi,E.H., Dean,J. and Polyzotis,N. (2018) The case for learned index structures. In: *Proceedings of the 2018 International Conference on Management of Data*. Association for Computing Machinery, NY, pp. 489–504.
11. Antol,M., Ol'ha,J., Slanináková,T. and Dohnal,V. (2021) Learned metric index—proposition of learned indexing for unstructured data. *Inform. Syst.*, **100**, 101774.
12. Slanináková,T., Antol,M., Olha,J., Kaňa,V. and Dohnal,V. (2021) Data-driven learned metric index: an unsupervised approach. In: Reyes, N., Connor,R., Kriege,N., Kazempour,D., Bartolini,I., Schubert,E. and Chen,J. (eds.) *Similarity Search and Applications*. Springer International Publishing, Cham, Switzerland, pp. 81–94.
13. Johnson,J., Douze,M. and Jégou,H. (2019) Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, **7**, 535–547.
14. Zhang,C., Shine,M., Pyle,A.M. and Zhang,Y. (2022) US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods*, **19**, 1109–1115.
15. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
16. Sehnal,D., Bittrich,S., Deshpande,M., Svobodová,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koča,J. and Rose,A.S. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structure. *Nucleic Acids Res.*, **49**, W431–W437.
17. Midlik,A., Navrátilová,V., Moturu,T.R., Koča,J., Svobodová,R. and Berka,K. (2021) Uncovering of cytochrome P450 anatomy by SecStrAnnotator. *Sci. Rep.*, **11**, 12345.
18. Ung,K.L., Winkler,M., Schulz,L., Kolb,M., Janacek,D.P., Dedic,E., Stokes,D.L., Hammes,U.Z. and Pedersen,B.P. (2022) Structures and mechanism of the plant PIN-FORMED auxin transporter. *Nature*, **609**, 605–610.
19. Slanináková,T., Antol,M., Ol'ha,J., Dohnal,V., Ladra,S. and Martínez-Prieto,M.A. (2023) Reproducible experiments with learned metric index framework. *Inform. Syst.*, **118**, 102255.

Received: January 30, 2024. **Revised:** April 10, 2024. **Editorial Decision:** April 28, 2024. **Accepted:** April 30, 2024