Current Research in Structural Biology 9 (2025) 100163

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology

Research Article

An easy-to-use three-dimensional protein-structure-prediction online platform "DPL3D" based on deep learning algorithms

Yunlong Gao [c], He Wang [c], Jiapeng Zhou [b,**], Yan Yang [a,*]

[a] The College of Health Humanities, Jinzhou Medical University, Jinzhou, 121001, China
[b] College of Life Sciences, Hunan Normal University, Changsha, 410000, China
[c] NewInsyght Biotech (Guangdong) Co., Ltd. DongGuan 523000, China

ARTICLE INFO

Handling Editor: Dr A Wlodawer

Keywords:
Structural biology
Protein three-dimensional structure
Gene mutation

ABSTRACT

The change in the three-dimensional (3D) structure of a protein can affect its own function or interaction with other protein(s), which may lead to disease(s). Gene mutations, especially missense mutations, are the main cause of changes in protein structure. Due to the lack of protein crystal structure data, about three-quarters of human mutant proteins cannot be predicted or accurately predicted, and the pathogenicity of missense mutations can only be indirectly evaluated by evolutionary conservation. Recently, many computational methods have been developed to predict protein 3D structures with accuracy comparable to experiments. This progress enables the information of structural biology to be further utilized by clinicians. Thus, we developed a user-friendly platform named DPL3D (http://nsbio.tech:3000) which can predict and visualize the 3D structure of mutant proteins. The crystal structure and other information of proteins were downloaded together with the software including AlphaFold 2, RoseTTAFold, RoseTTAFold All-Atom, and trRosettaX-Single. We implemented a query module for 210,180 molecular structures, including 52,248 human proteins. Visualization of protein two-dimensional (2D) and 3D structure prediction can be generated via LiteMol automatically or manually and interactively. This platform will allow users to easily and quickly retrieve large-scale structural information for biological discovery.

1. Introduction

Over 400,000,000 non-redundant protein sequences are available in public databases, with new sequences continuously being discovered and collected (https://www.uniprot.org/uniparc/) (The UniProt and Consortium, 2021), (Jin et al., 2022). Meanwhile, only slightly more than 100,000 unique protein structures were identified experimentally, despite progress made in high-throughput structural genomics initiatives (Berman et al., 2003), (Li et al., 2020). A number of experimental processes have been developed to provide high-resolution images of biomolecules. For example, cryo-electron microscopy has enabled the determination of biological structures with resolutions as high as ~1.2 Å, which is demonstrated with apoferritin (Hoff et al., 2024). It takes a considerable workforce and material resources to obtain the structures experimentally. Therefore, computational methods have become valuable alternatives. Generally, protein structure prediction methods can be broadly divided into three classes: homology modeling (John, 2003), protein threading (Wu, 2004), and ab initio protein modeling (Rollins et al., 2019).

Homology modeling methods are utilized to align the target to a set of proteins with similar sequences and known structures, assuming evolutionarily related proteins have similar structures (Johnston, 2005), (Marc et al., 2000). These methods typically require more than 30% consistency with the known protein sequence(s) to predict the three-dimensional shape of a query protein (Adam et al., 2008). It has been reported that 90% of the protein pairs were homologous and had similar structures when their amino acid sequences had 30% identity or above; less than 10% were homologous when the pairs had below 25% identity (Adam et al., 2008). Protein threading, similar to homology modeling, is a series of template-based approaches to structure prediction (Guo, 2003). However, threading techniques aim to identify templates whose structures best accommodate the target protein, without relying on sequence similarity (Torda et al., 2004). Ab initio modeling constructs a protein 3D conformation solely from a sequence rather than

* Corresponding author.
** Corresponding author.
E-mail addresses: zhoujp111@126.com (J. Zhou), 360226919@qq.com (Y. Yang).

https://doi.org/10.1016/j.crstbi.2024.100163
Received 6 April 2024; Received in revised form 20 November 2024; Accepted 30 December 2024
Available online 3 January 2025

a template, seeking a conformation with the lowest free energy (Pruitt et al., 2013), (Baker and Sali, 2001). Ab initio modeling can work well for proteins with fewer than 120 amino acids (Roy et al., 2010).

Traditional de novo protein structure prediction involves exhaustively searching for an optimal conformation among all possibilities, which is achieved through molecular dynamics simulations using empirical force fields, or Monte Carlo simulations based on fragment-assembly or threading-assembly approaches (Ding and Gong, 2020). These simulations rely on experimentally determined structures as templates, guiding the target protein's fragments to compose specific conformations. However, the effectiveness of traditional methods diminishes for more complex structures and proteins with low sequence homology, such as the free-modeling (FM) targets in the critical assessment of protein structure prediction (CASP) competitions. Notable improvements in protein structure prediction have been observed in CASP12 compared to CASP11, with the application of deep learning algorithms and co-evolution information recognized as the main drivers (Schaarschmidt et al., 2018).

Several deep learning-based approaches have been developed for protein structure prediction in recent years. AlphaFold 1, a convolutional neural network-based system, achieved higher accuracy than its competitors in CASP13 (Senior et al., 2020). This model uses a neural network to predict distances between residue pairs, which aids in constructing a potential of mean force to describe protein shape. AlphaFold 2 adopts Evoformer, a modified transformer that refines evolutionary information from multiple sequence alignments (MSA) and spatial data from structural template searches before determining the final protein structure (Jumper et al., 2021). The latest version, AlphaFold 3, can predict various types of biomolecules with high accuracy. Key updates include a simplified MSA representation and a diffusion module for predicting raw atom coordinates, enabling the prediction of arbitrary chemical components (Abramson et al., 2024). Similar to AlphaFold 2, RoseTTAFold uses a three-track network as its core module in which the information of amino acid sequence, distance map, and 3D coordinates is transformed and integrated (Baek et al., 2021). This approach delivers accuracy comparable to AlphaFold 2. In 2024, the Baker Lab released RoseTTAFold All-Atom, which employs the three-track network and incorporates information on chemical element types of non-polymer atoms, chemical bonds between atoms, and chirality. The updated model can predict the structures of a wide diversity of biomolecules (Krishna et al., 2024). TrRosetta processes MSA and homologous templates to predict inter-residue geometries, followed by structure prediction through energy minimization using the Rosetta framework (Du et al., 2021). Its latest iteration, trRosettaX-Single, adopts an MSA-free algorithm to better predict orphan protein structures, and its performance is superior to AlphaFold 2 and RoseTTAFold (Wang et al., 2022).

The Human Gene Mutation Database (HGMD) has recorded over 310,537 disease-causing or likely disease-causing variants, including frameshift, nonsense, splicing, and missense mutations (Stenson et al., 2020). Missense mutations make up 46.8% of these variations. However, the pathogenicity assessment of the same missense mutation in different databases is inconsistent, such as HGMD vs ClinVar (https://www.ncbi.nlm.nih.gov/clinvar). Accurate prediction of the pathogenicity of missense mutations is crucial. Researchers may gain significant insights into how genetic changes affect protein function if the 3D structures of the associated proteins are available (Capriotti et al., 2011), (Diwan et al., 2021), (Sarkar et al., 2017).

Official online versions of these prediction tools, such as the AlphaFold Server and trRosetta server, offer fast processing of prediction tasks. However, locally installed structure viewers may be necessary to visualize the predicted structures, as the embedded molecule viewers on the websites typically have limited functionality. The online prediction tools don't provide access to experimentally determined protein structures, although the prediction processes often require these structures as templates. Researchers typically need to obtain the structure of a known protein from an external source to compare it with a predicted structure. To address these inconveniences, we developed the DPL3D platform, which includes various structure prediction software, advanced visualization tools and an extensive collection of protein structural data.

DPL3D is a comprehensive 3D structure platform that includes the full databases required for AlphaFold 2, RoseTTAFold, RoseTTAFold All-Atom, and trRosettaX-Single, as well as a query service for 210,180 molecular structure entries. It stores 3D structure annotations for model organisms (Table 1), and all data are freely accessible and downloadable for academic use (Deelen et al., 2019). DPL3D allows users to predict the structure of novel or mutant proteins, assisting in the discovery of underlying biological mechanisms. It incorporates the latest RoseTTAFold All-Atom for fast and accurate structure prediction, as well as established tools such as AlphaFold 2 and RoseTTAFold. TrRosettaX-Single can be used to predict the structures of orphan proteins, a task in which this approach specializes. The platform also provides a powerful structure viewer, which supports commonly used visualization styles, including cartoons, surfaces, and ball-and-stick models, along with various coloring options. With the integrated molecular viewer and extensive structure database, both predicted and archived protein structures can be viewed and compared online without additional software.

2. Materials and methods

2.1. Applications and system configuration

The databases required for AlphaFold 2, RoseTTAfold, and RosettaFold All-Atom were downloaded and decompressed as instructed in their Github repositories. The full databases of AlphaFold 2 need 2.6 TB of disk space after decompression, including BFD, MGnify, PDB (Senior et al., 2020), (Shahzad et al., 2023), etc. RoseTTAfold can share BFD database with AlphaFold 2, while it needs about 460 GB of additional disk space for PDB100 and Uniref30. AlphaFold 2, RoseTTAfold, RosettaFold All-Atom, and trRosettaX-Single programs were installed in a local high-performance computer—CPU: Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30 GHz *2, RAM: 128G, GPU: NVIDIA Tesla P40, Hard Drive: WD Ultrastar DC HC310 HUS726T6TALE6L4 *2. Users can choose any of the four programs to predict the 3D structure of a mutant or novel protein sequence. The image of the 3D structure will be generated by the LiteMol software.

2.2. Protein structure predicting models

AlphaFold 2 is a protein structure prediction model developed by DeepMind in 2021. The model comprises two main modules. The first module, Evoformer, receives input features derived from MSA and structural template searching and exchanges evolutionary and spatial information iteratively. The refined MSA and pair representation output by Evoformer are then processed by the structure module, the second part of the AlphaFold 2 pipeline. The backbone and side chain conformations are determined by the structure module and a final predicted protein structure is generated (Jumper et al., 2021).

RoseTTAFold pipeline starts with homologous sequence searching against UniRef30 and BFD sequence databases. The generated MSA is

Table 1
Summary of DPL3D data.

Category	Count
Species	8111
Protein 3D structures	210,180
Homo sapiens	54,332
Arabidopsis thaliana	1962
Caenorhabditis elegans	457
Drosophila melanogaster	1170
Danio rerio	570
Mus musculus	7719

used for structural template searching against the PDB100 database. RoseTTAFold network predicts inter-residue geometries with MSA and top 10 templates. RoseTTAFold provides the PyRosetta procedure and the end-to-end procedure for generating the final structure with inter-residue geometries. The PyRosetta version is selected in our platform because it has relatively higher accuracy than the end-to-end version and doesn't require large graphic memory for amino acid sequences longer than 400 residues. The PyRosetta version picks 5 models out of 15 sampled structures based on predicted lDDT of DeepAccNet after clustering (Baek et al., 2021).

RosettaFold All-Atom is an advanced variant of the RosettaFold protein structure prediction tool. To enhance its accuracy and utility in structural biology, several crucial updates have been applied to RosettaFold All-Atom such as including atoms that are independent of amino acid or nucleotide chains and taking into account chemical bonds between atoms and chirality (Krishna et al., 2024). RosettaFold All-Atom is built on the RoseTTAFold2 model, in which a more computationally efficient structure-biased attention is adopted (Baek et al., 2022). These enhancements enable RosettaFold All-Atom to predict multiple types of biomolecules with greater speed and efficiency compared to earlier versions.

TrRosettaX-Single is an MSA-free method for improved accuracy of predicting orphan protein 3D structure. It also works well on human-designed proteins. The full pipeline of trRosettaX-Single contains two steps: 2D geometry prediction and 3D structure folding. The input of 2D geometry prediction is a single amino acid sequence, fed into s-ESM-1 (supervised ESM-1). The output of s-ESM-1 and one-hot encoding are processed by a multi-scale network Res2Net_Single to obtain predicted 2D geometry, including inter-residue distance and orientations. 3D structure folding step receives 2D geometry information and generates 3D structure via energy minimization (Wang et al., 2022).

2.3. *Molecular structure viewer*

Protein structures are displayed on web pages using LiteMol (Sehnal et al., 2017), a web-based 3D viewer for molecular data. Designed to be user-friendly and efficient, LiteMol quickly renders large molecular structures. It offers commonly used visualizations, including cartoons, surfaces, and ball-and-stick models, along with various coloring options such as uniform and rainbow coloring. LiteMol can read both text and binary data. In our platform, predicted protein structures in PDB file format (text) can be loaded and displayed using the LiteMol plugin embedded in the webpage. The PDB file content and LiteMol library are also integrated into an HTML file, which can be downloaded and viewed in a browser.

2.4. *Workflow of DPL3D*

Fasta strings are submitted through a webpage, along with additional information such as the chosen prediction tool. The prediction task is then added to a queue and processed by one of the four available tools—AlphaFold 2, RoseTTAFold, RoseTTAFold All-Atom, or trRosettaX-Single—once all preceding tasks are completed.

The prediction tools and their dependencies are managed using Miniconda, with minor code adjustments made to accommodate updates to specific dependencies. In the RoseTTAFold pipeline, our platform utilizes the PyRosetta approach, which requires lower GPU memory usage and offers improved performance for structure prediction compared to the end-to-end approach. The full databases required by the tools, such as BFD, MGnify, and PDB, are downloaded and decompressed. To save disk space, some databases, like BFD, can be shared across multiple pipelines using soft links.

Both AlphaFold 2 and RoseTTAFold generate multiple PDB files. AlphaFold 2 outputs five unrelaxed models, with the best one determined based on the "ranking_debug" file. This top unrelaxed model is further processed to produce a relaxed version, both of which can be viewed using the LiteMol viewer embedded on the webpage. RoseTTAFold's PyRosetta version provides five final structures, with the first displayed in the viewer. In contrast, RoseTTAFold All-Atom and trRosettaX-Single each produce a single final model for every prediction.

Our platform also offers a query function for biomolecules in the PDB database. Users can search for a target molecule using either its four-character PDB ID or a description of the molecule, with the query results displayed in the molecular viewer on the page.

2.5. *Features of DPL3D*

Users can perform 3D structure analysis for novel and mutant proteins. A simple user interface allows users to input sequence data in FASTA format by copying and pasting it into the designated text area. They can choose one of the four deep learning prediction software options (Fig. 1A). Upon submitting a prediction, a 24-character token (Order ID) is generated for accessing the results. Users can save the URL containing the token to open it later in a browser.

Once the computation is complete, predicted structures can be visualized using the web-based LiteMol viewer. This viewer enables interactive exploration of structural data, quickly rendering 3D images as users zoom, drag, or rotate the molecule (Fig. 1B).

Users can download the prediction output folder as a zip file, which includes the final model in PDB format along with intermediate files such as MSA results. It also contains LiteMol components and an HTML file with the embedded protein structure, allowing users to view the 3D model directly in a browser without needing additional software.

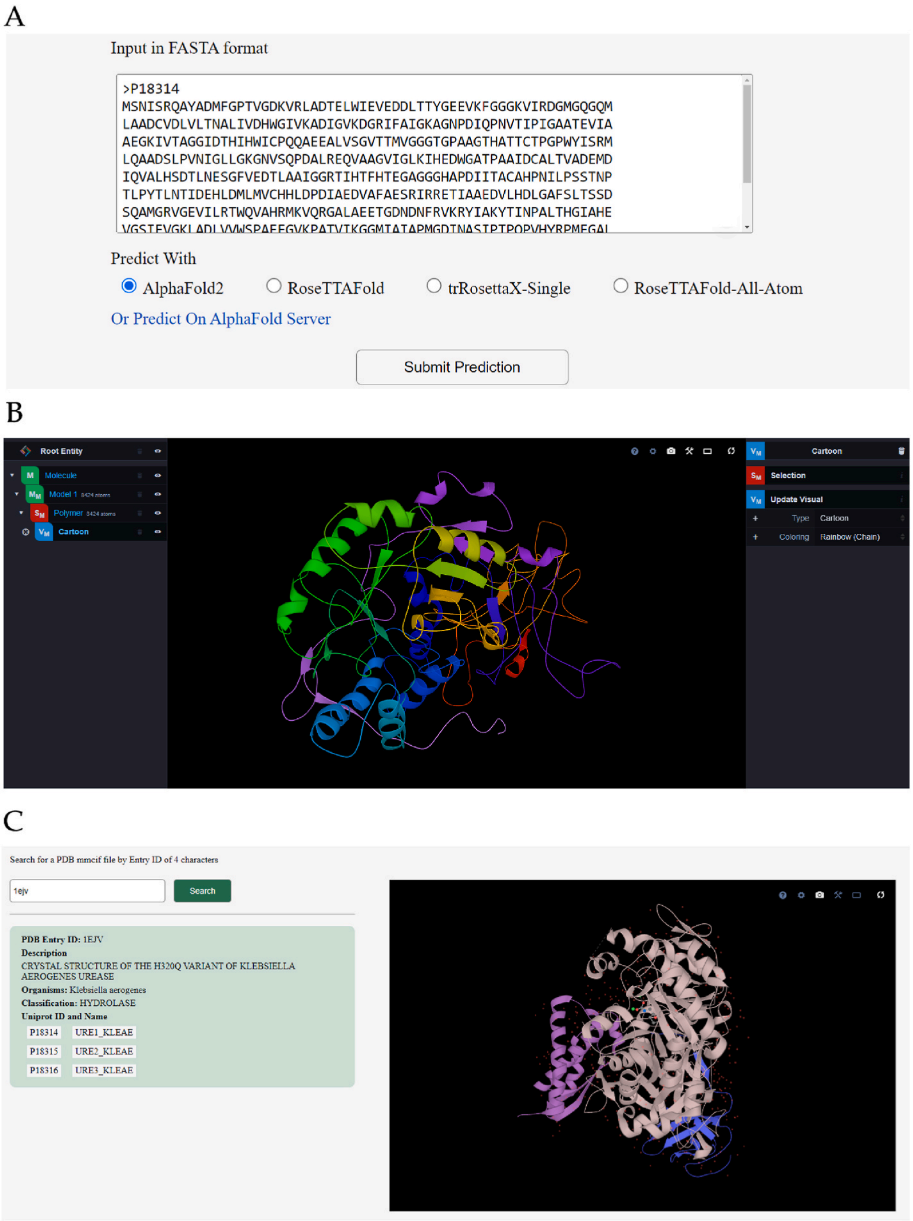
Additionally, users can access the PDB database downloaded for the prediction tools. Molecular structure data can be queried using the PDB entry ID (Fig. 1C). For example, entry ID '1EJV' can be used to find the molecule, and the structure will be displayed in LiteMol. Users can also search for molecular structure data using keywords, with related PDB entries appearing in a dropdown list for selection.

3. Results

To assess computational efficiency, we have selected proteins of varying lengths and each protein has been processed by four prediction tools in our platform. The processing time for each sequence by each prediction tool has been recorded (Fig. 2). Among the tested software, RoseTTAFold All-Atom can quickly finish the predictions for sequences among all lengths. TrRosettaX-Single is also swift at structure predictions for short sequences since it doesn't conduct MSA as other prediction tools. However, as the sequence gets longer, the time spent for prediction drastically increases and it takes more than 8 h to process a sequence of over 700 amino acids. AlphaFold and RoseTTAFold generally need more time to process the same protein sequence than RoseTTAFold All-Atom. Paired t-tests have been conducted to compare the processing time of AlphaFold, RoseTTAFold, and RoseTTAFold All-Atom. There isn't a significant difference between AlphaFold and RoseTTAFold with a p-value 0.9303. RoseTTAFold All-Atom spends significantly less time on predictions than AlphaFold (with a p-value 0.0046) and RoseTTAFold (with a p-value 0.0234).

The red frames in the results enclose the same segment of immunoglobulin transcription factor 2 (ITF2) sequence. The alpha helices (in purple) predicted by AlphaFold 2, RoseTTAFold and RoseTTAFold All-Atom resemble those observed in X-ray imaging (see Fig. 3).

According to evolutionary analysis, new genes often emerge after the differentiation of ancestral species and exhibit higher tissue specificity (Yang et al., 2023). A new gene discovered in the T2T genome assembly on chromosome 13 has been identified compared to previous genome assemblies (Sergey et al., 2021). Comparative orthology analysis has revealed that the orthologue of this new gene in humans is part of a conserved homologous syntenic block in chimpanzees (Fig. 4A). To further investigate the structural features, we have utilized our platform to predict the structures of proteins from the orthologous genes. The



Input in FASTA format

```
>P18314
MSNISRQAYADMFGPTVGDKVRLADTELWIEVEDDLTTYGEEVKFGGGKVIRDGMGQGQM
LAADCVDLVLTNALIVDHWGIVKADIGVKDGRIFAIGKAGNPDIQPNVTIPIGAATEVIA
AEGKIVTAGGIDTHIHWICPQQAEEALVSGVTTMVGGGTGPAAGTHATTCTPGPWYISRM
LQAADSLPVNIGLLGKGNVSQPDALREQVAAGVIGLKIHEDWGATPAAIDCALTVADEMD
IQVALHSDTLNESGFVEDTLAAIGGRTIHTFHTEGAGGGHAPDIITACAHPNILPSSTNP
TLPYTLNTIDEHLDMLMVCHHLDPDIAEDVAFAESRIRRETIAAEDVLHDLGAFSLTSSD
SQAMGRVGEVILRTWQVAHRMKVQRGALAEETGDNDNFRVKRYIAKYTINPALTHGIAHE
VGSIEVGKLADLVVWSPAFFGVKPATVIKGGMIATAPMGDINASIPTPQPVHYRPMFGAL
```

Predict With

● AlphaFold2 ○ RoseTTAFold ○ trRosettaX-Single ○ RoseTTAFold-All-Atom

Or Predict On AlphaFold Server

Submit Prediction

Type Cartoon
Coloring Rainbow (Chain)

Search for a PDB mmcif file by Entry ID of 4 characters

1ejv

PDB Entry ID: 1EJV
Description
CRYSTAL STRUCTURE OF THE H320Q VARIANT OF KLEBSIELLA AEROGENES UREASE
Organisms: Klebsiella aerogenes
Classification: HYDROLASE
Uniprot ID and Name

P18314	URE1_KLEAE
P18315	URE2_KLEAE
P18316	URE3_KLEAE

Fig. 1. User interfaces for sequence input, visualization of prediction results, and PDB queries. (A) The input string should be in FASTA format. (B) All structures are displayed using the LiteMol software on the webpage. (C) PDB IDs can be utilized to retrieve molecular structures.

protein expressed by the homologous gene in humans exhibits fewer helices than that in chimpanzees (Fig. 4B).

The 3D structures of many pathogenic mutants have not been reported to date in the Protein Data Bank. Therefore, this presents significant difficulties for researchers and physicians. The prediction software can help analyze the structural features of many mutated proteins in one species or across closely related species to discover underlying mechanisms. For example, we have predicted the 3D structure of the rs1559470315, NM_001904.4: c.1016_1025delinsT (p. Thr339_Arg342delinsIle, pubmed:33350591) using DPL3D (Fig. 5). The mutation can dramatically change the hydrogen bond network and even alter the ionization states of neighboring amino acids.

4. Discussion

The development of artificial intelligence (AI) technology has effectively addressed previously unresolved challenges across various domains. Its application in biological research and industry has witnessed a significant surge. Notably, many software tools have been developed, including AI models for predicting protein structures. The DPL3D platform incorporates four structure-predicting pipelines: AlphaFold 2, RoseTTAFold, RoseTTAFold All-Atom, and trRosettaX-Single. RoseTTAFold All-Atom is chosen for its superior accuracy and processing speed, as it represents the most recent and advanced model in the RoseTTAFold series for biomolecular structure prediction. Despite substantial architecture differences, AlphaFold 2 and RoseTTAFold are established MSA-based deep learning approaches known for their high accuracies, and thus both of them are integrated into our platform. TrRosettaX-Single, the latest version of trRosetta, is included in our platform due to its MSA-free design, which makes it highly effective for predicting orphan proteins. Additionally, its rapid processing of short sequences further supports its selection.

The source code for these prediction tools and their associated databases are publicly available. However, researchers may still find the installation process challenging. Proper configuration of dependencies and environment variables is essential for all prediction pipelines.

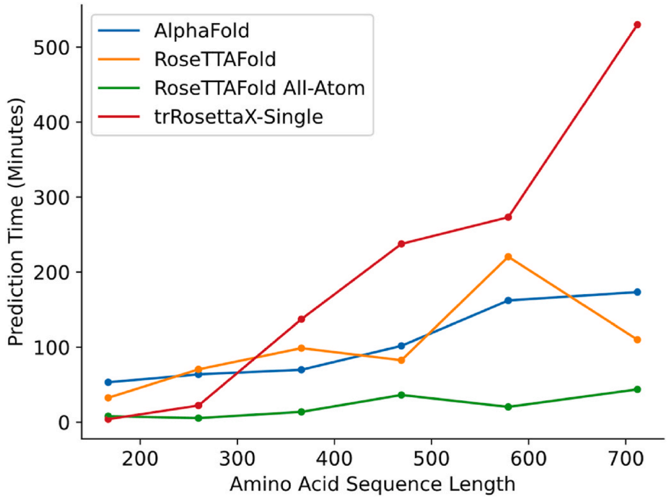


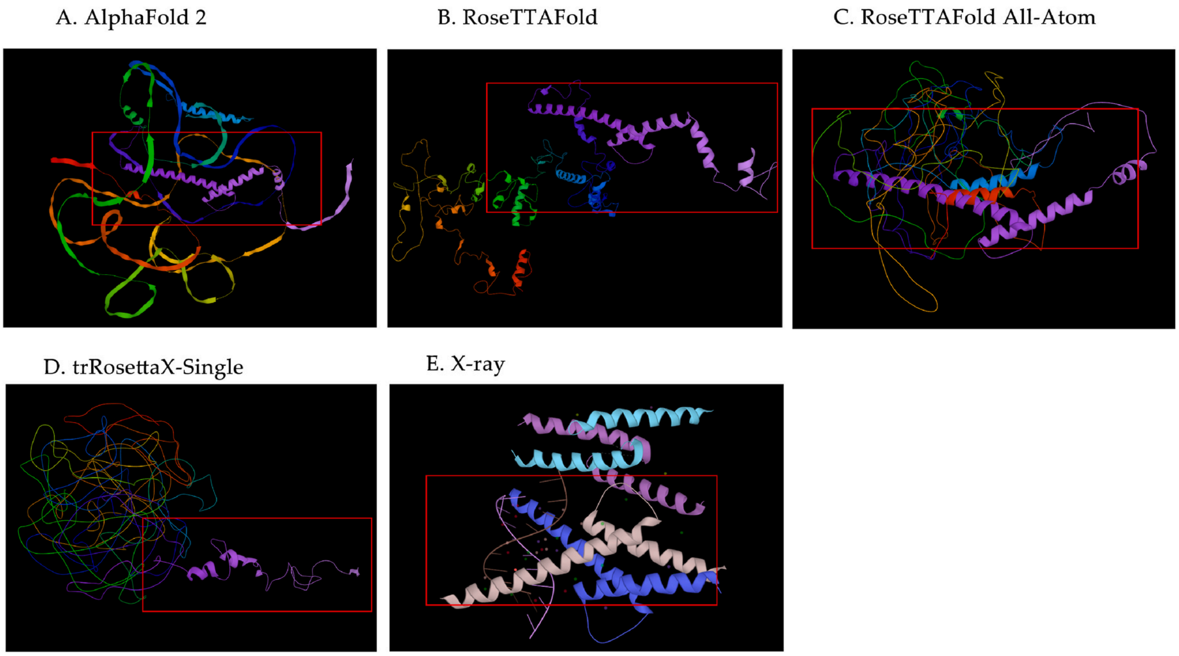
Fig. 2. The time spent by the four prediction tools on sequences of varying lengths.

Dependencies automatically installed by package managers such as Conda may not align with the versions compatible with the prediction pipelines. To solve these technical issues and streamline the prediction process, separate environments are configured for the prediction tools, and subtle adjustments are made to both the pipelines and their dependencies. A web server is set up to integrate these prediction tools. Users can upload amino acid sequences in FASTA format and choose from the four available tools to start the prediction process. The molecular structure viewer LiteMol is embedded into the browser-side UI for visualizing the predicted structure. To the best of our knowledge, the DPL3D platform stands out as the most user-friendly solution for predicting the 3D structure of proteins.

We have evaluated four protein prediction pipelines across various input sequence lengths. RoseTTAFold All-Atom proves to be the fastest for predicting the structures of amino acid sequences longer than 200 residues. Even for shorter sequences, it consistently outperforms AlphaFold 2 and RoseTTAFold in terms of speed. For the sequence with 167 amino acids, trRosettaX-Single demonstrates remarkable efficiency, completing the prediction in just 4 min, while both AlphaFold 2 and RoseTTAFold take over 30 min, primarily due to their time-consuming Multiple Sequence Alignment (MSA) steps. However, trRosettaX-Single exhibits a significant increase in computation time for sequences longer than 400 amino acids. For example, while AlphaFold 2 and RoseTTAFold predict the structure of a 712-amino acid sequence in under 2 h, trRosettaX-Single requires nearly 9 h. The first step of the trRosettaX-Single pipeline to predict distances and orientations takes less than 2 min, even for sequences slightly over 700 amino acids; the majority of the time is spent on the final structure prediction, which operates in a single-threaded mode and makes minimal use of the graphics card. Given that trRosettaX-Single focuses on structures of orphan and human-designed proteins, it is advisable to use this tool for predicting the structures of short orphan or human-designed proteins. For general-purpose predictions, RoseTTAFold All-Atom often represents the best choice due to its speed and optimizations. AlphaFold 2 and RoseTTAFold remain strong alternatives, and employing multiple pipelines enhances confidence in the results.

Each prediction tool has been tested on CASP or other datasets, demonstrating high accuracy, although the predicted structures may exhibit some inconsistency. When predicting the structure of the human ITF2 protein, AlphaFold 2, RoseTTAFold, and RoseTTAFold All-Atom identify alpha helices within the interval of 555–667 in the ITF2 amino acid sequence, consistent with observations from X-ray imaging. However, the predicted structures diverge in regions not covered by the X-ray data. This inconsistency may result from the incomplete structural data obtained experimentally; as more structural data are collected and integrated into deep learning models, this inconsistency should diminish.

The development of prediction approaches facilitates advancements across various fields of biological research and applications. One important application of protein structure prediction is assessing whether a genetic mutation alters protein structure, potentially leading to genetic disorders. The prediction tools highlight the differences between wild-type and mutated proteins associated with the mutation rs1559470315, as shown in Fig. 5. This underscores the potential of deep learning models to predict structural changes due to mutations and


A. AlphaFold 2

B. RoseTTAFold

C. RoseTTAFold All-Atom

D. trRosettaX-Single

E. X-ray

Fig. 3. The highlighted regions in the images from prediction tools and X-ray crystallography represent a shared segment of the ITF2 protein sequence. (A) AlphaFold 2. (B) RoseTTAFold. (C) RoseTTAFold All-Atom. (D) trRosettaX-Single. (E) X-ray crystallography.

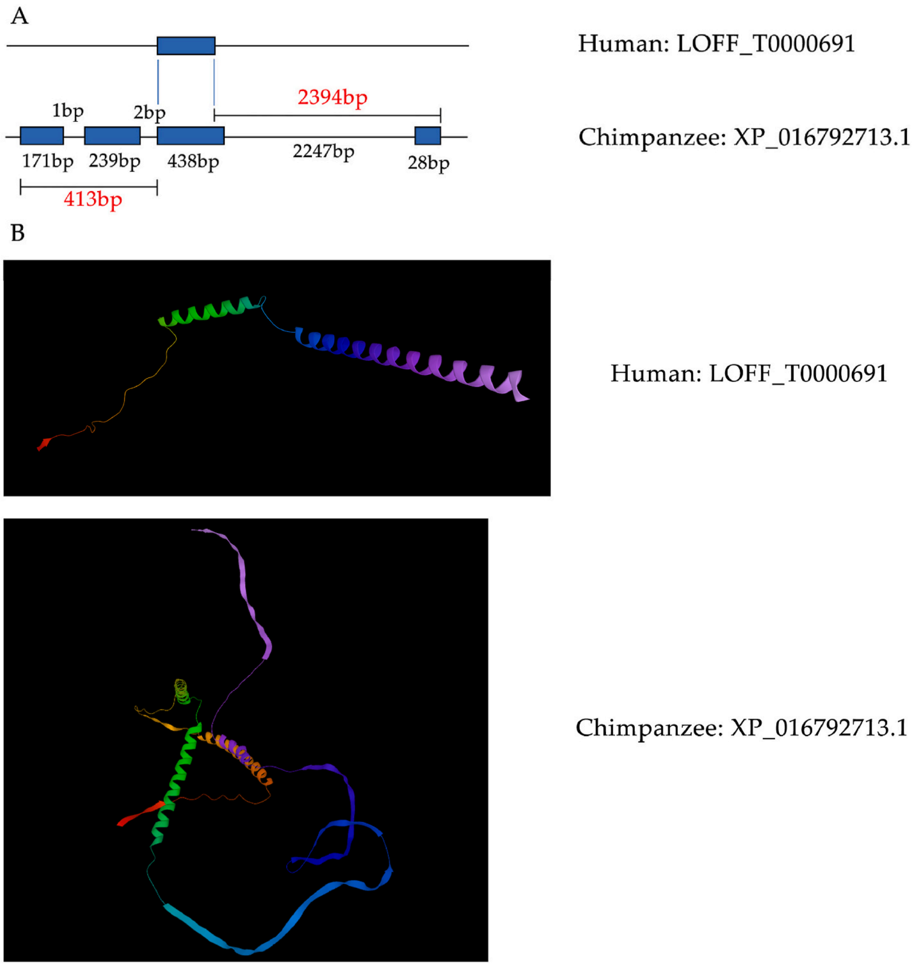


Fig. 4. Homology and 3D structure analysis of the new gene from T2T genome on chromosome 13. (A) Homology analysis of the new gene was based on the sequence alignment. (B) Structural comparisons among homologues were performed using the DPL3D platform.

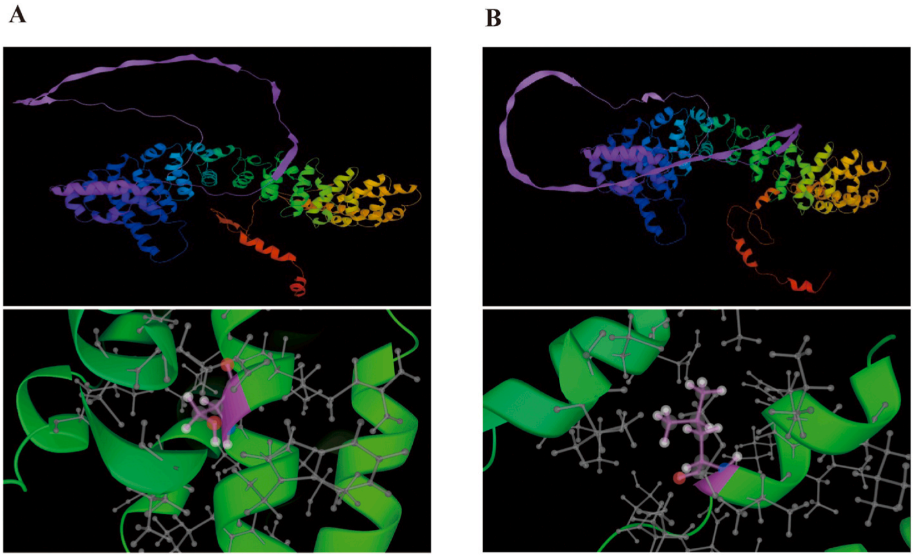


Fig. 5. The 3D structures of the mutant and wild-type proteins have been predicted to assess the effect of the mutation on the stability of the protein structure. (A) Wild-type protein. (B) Mutated protein.

evaluate the pathogenicity of variants with currently unknown clinical significance. However, researchers and clinicians should exercise caution when interpreting the clinical implications of genetic mutations using AI tools. Extensive optimization and testing are essential to ensure that these predictive approaches effectively assist in diagnosis and interventions, and results obtained from computational methods must be validated through experimental studies.

In addition to software designed for predicting structures of monomers, there are deep learning tools like AlphaFold-Multimer tailored for multimeric proteins. AlphaFold-Multimer has demonstrated precision in processing a large fraction of PDB complexes (Richard et al., 2021), (Deneke et al., 2024). Nonetheless, AlphaFold-Multimer has limitations such as the inability to predict binding of antibodies, likely caused by limited evolutionary information derived from MSA for heteromeric protein complexes. Another research suggests that AlphaFold-Multimer with modified MSA library leads to an enhanced accuracy of predicted models (Peng et al., 2023). In 2024, AlphaFold 3 (Abramson et al., 2024) and RoseTTAFold All-Atom (Krishna et al., 2024) were developed, both capable of processing various types of biomolecules and biomolecular complexes. These tools may offer promising solutions to the challenges previously mentioned, while extensive research and testing are needed to validate their effectiveness in handling complex structures (Wee and Wei, 2024), (Anusha et al., 2024).

5. Conclusions

We have developed an online platform for protein structure prediction, incorporating the capabilities of AlphaFold 2, RoseTTAFold, RoseTTAFold All-Atom, and trRosettaX-Single. Additionally, the platform offers a query service for retrieving existing protein structure data from the PDB database. The efficiency of the four prediction pipelines is also assessed when handling proteins with different lengths. This comparative analysis aims to assist researchers in selecting the most suitable pipeline.

Deep learning prediction tools can predict the structures of both wild-type and mutated proteins, showing their potential for assessing the impact of mutations on protein structure and predicting the pathogenicity of variants with uncertain clinical significance. However, further research and exploration are necessary to establish a reliable AI-assisted annotation method for clinical applications. While these tools have achieved high accuracy in structure predictions, there is still room for optimization through increased training data and improved neural network models.

With the increasing accumulation of protein structural data derived from cryo-electron microscopy and synchrotron radiation, a growing number of proteins and their regulatory roles are being unveiled. We will consistently update the platform as new data becomes available. We believe this platform will serve as a valuable resource, enabling scientists to swiftly access critical information on the 3D structures of proteins and facilitating further advancements in biological discovery.

CRediT authorship contribution statement

Yunlong Gao: Software, Validation. **He Wang:** Methodology, Supervision. **Jiapeng Zhou:** Writing – review & editing. **Yan Yang:** Conceptualization, Software, Writing – original draft.

Institutional review board statement

All data in this article are sourced from the public databases and comply with ethical review.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

The datasets supporting the conclusions of this article are included within the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Dr. Meng Li and Mr. Xianchao Guo for testing the platform and providing suggestions for modification.

Data availability

The data that has been used is confidential.

References

Abramson, J., Adler, J., Dunger, J., et al., 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500. https://doi.org/10.1038/s41586-024-07487-w.

Adam, B., Charloteaux, B., Beaufays, J., Vanhamme, L., Godfroid, E., Brasseur, R., Lins, L., 2008. Distantly related lipocalins share two conserved clusters of hydrophobic residues: use in homology modeling. BMC Struct. Biol. 8, 1.

Anusha, Zhang Z., Li, J., Zuo, H., Mao, C., 2024. AlphaFold 3 - Aided design of DNA Motifs to Assemble into Triangles. J. Am. Chem. Soc. 146 (37), 25422–25425. https://doi.org/10.1021/jacs.4c08387. Epub 2024 Sep 5. PMID: 39235269.

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 (American Association for the Advancement of Science) 373, 871–876.

Baek, M., McHugh, R., Anishchenko, I., Baker, D., DiMaio, F., 2022. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. bioRxiv. https://doi.org/10.1101/2022.09.09.507333 [Preprint].

Baker, D., Sali, A., 2001. Protein structure prediction and structural genomics. Science 294, 93–96.

Berman, H., Henrick, K., Nakamura, H., 2003. Announcing the worldwide protein Data Bank. Nat. Struct. Mol. Biol. 10, 980.

Capriotti, E., Altman, R.B., 2011. Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinf. 12 (Suppl. 4), S3.

Deelen, P., Van, D.S., Herkert, J.C., Karjalainen, J.M., Brugge, H., Abbott, K.M., Van, D.C. C., Van, D., Zwaag, P.A., Gerkes, E.H., et al., 2019. Improving the diagnostic yield of exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nat. Commun. 10.

Deneke, V.E., Blaha, A., Lu, Y., et al., 2024. A conserved fertilization complex bridges sperm and egg in vertebrates. Cell.

Ding, W., Gong, H., 2020. Predicting the Real-Valued inter-residue distances for proteins. Adv. Sci. 7, 2001314.

Diwan, G.D., Gonzalez-Sanchez, J.C., Apic, G., Russell, R.B., 2021. Next generation protein structure predictions and genetic variant interpretation. J. Mol. Biol. 433, 167180.

Du, Z., Su, H., Wang, W., et al., 2021. The trRosetta server for fast and accurate protein structure prediction. Nat. Protoc. 16, 5634–5651. https://doi.org/10.1038/s41596-021-00628-9.

Guo, J.T., 2003. Improving the performance of DomainParser for structural domain partition using neural network. Nucleic Acids Res. 31, 944–952.

Hoff, S.E., Thomasen, F.E., Lindorff-Larsen, K., Bonomi, M., 2024. Accurate model and ensemble refinement using cryo-electron microscopy maps and Bayesian inference. PLoS Comput. Biol. 20 (7), e1012180. https://doi.org/10.1371/journal.pcbi.1012180. PMID: 39008528; PMCID: PMC11271924.

Jin, X.X., Liu, J.Y., Wang, W.P., Li, J.F., Liu, G.M., Qiu, R.Q., Yang, M.Z., Liu, M., Yang, L., Du, X.F., et al., 2022. Identification of age-associated proteins and functional alterations in human retinal pigment epithelium. Dev. Reprod. Biol. 20, 633–647.

John, B., 2003. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res. 31, 3982–3992.

Johnston, C.R., 2005. A sequence sub-sampling algorithm increases the power to detect distant homologues. Nucleic Acids Res. 33, 3772–3778.

Jumper, J., Evans, R., Pritzel, A., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G.R., Morey-Burrows, F.S., Anishchenko, I., Humphreys, I.R., McHugh, R., Vafeados, D., Li, X., Sutherland, G.A., Hitchcock, A., Hunter, C.N., Kang, A., Brackenbrough, E., Bera, A.K., Baek, M., DiMaio, F., Baker, D., 2024. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science 384 (6693), eadl2528. https://doi.org/10.1126/science.adl2528. Epub 2024 Apr 19. PMID: 38452047.

Li, Y.R., Huang, Y.N., Zhao, B., Wu, M.F., Li, T.Y., Zhang, Y.L., Chen, D., Yu, M., Mo, W., 2020. RGD-hirudin-based low molecular weight peptide prevents blood coagulation via subcutaneous injection. Acta Pharmacol. Sin. 41, 753–762.

Marc, A.M., Ashley, C.S., Andras, F., Roberto, S., Francisco, M., Andrej, S., 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291–325.

Peng, Z., Wang, W., Wei, H., Li, X., Yang, J., 2023. Improved protein structure prediction with trRosettaX2, AlphaFold 2, and optimized MSAs in CASP15. Proteins 91, 1704–1711.

Pruitt, M.M., Lamm, M.H., Coffman, C.R., 2013. Molecular dynamics simulations on the Tre1 G protein-coupled receptor: exploring the role of the arginine of the NRY motif in Tre1 structure. BMC Struct. Biol. 13, 15.

Richard, E., Michael, O.N., Pritzel, A., Natasha, A., Andrew, S., Tim, G., Augustin, Z., Russ, B., Sam, B., Jason, Y., et al., 2021. Protein complex prediction with AlphaFold-Multimer. bioRxiv. https://doi.org/10.1101/2021.10.04.463034.

Rollins, N.J., Brock, K.P., Poelwijk, F.J., Stiffler, M.A., Gauthier, N.P., Sander, C., Marks, D.S., 2019. Inferring protein 3D structure from deep mutation scans. Nat. Genet. 51, 1170–1176.

Roy, A., Kucukural, A., Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5, 725–738.

Sarkar, B., Kulharia, M., Mantha, A.K., 2017. Understanding human thiol dioxygenase enzymes: structure to function, and biology to pathology. Int. J. Exp. Pathol. 98, 52–66.

Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., Bonvin, A., 2018. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins 86 (Suppl. 1), 51–66.

Sehnal, D., Deshpande, M., Vařeková, R., Sapib, M., Karel, B., Adam, M., Lukas, P., Sameer, V., Jaroslav, K., 2017. LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. Nat. Methods 14, 1121–1122.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, A., Nelson, A., Bridgland, A., et al., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710.

Sergey, N., Sergey, K., Arang, R., Mikko, R., Andrey, V.B., Alla, M., Mitchell, R.V., Nicolas, A., Lv, U., Ariel, G., et al., 2021. The complete sequence of a human genome. bioRxiv 5.

Shahzad, K., Rukhsana, A., Afifa, N., Muhammad, I.U., Alaa, A.B., Mohammed, M.A., Saiqa, I., 2023. Metabolic profiling and investigation of the modulatory effect of fagonia cretica L. aerial parts on hepatic CYP3A4 and UGT2B7 enzymes in streptozotocin—induced diabetic model. MDPI Antioxidants 12, 119.

Stenson, P.D., Mort, M., Ball, E.V., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D.S., Phillips, A.D., et al., 2020. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Hum. Genet. 139, 1197–1207.

The UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489.

Torda, A.E., Procter, J.B., Huber, T., 2004. Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. Nucleic Acids Res. 32, W532–W535.

Wang, W., Peng, Z., Yang, J., 2022. Single-sequence protein structure prediction using supervised transformer protein language models. Nat. Comput. Sci. 2, 804–814. https://doi.org/10.1038/s43588-022-00373-3.

Wee, J., Wei, G.W., 2024. Evaluation of AlphaFold 3's protein-protein complexes for predicting binding free energy changes upon mutation. J. Chem. Inf. Model. 64 (16), 6676–6683. https://doi.org/10.1021/acs.jcim.4c00976. Epub 2024 Aug 8. PMID: 39116039; PMCID: PMC11351016.

Wu, K.P., 2004. HYPROSP: a hybrid protein secondary structure prediction algorithm–a knowledge-based approach. Nucleic Acids Res. 32, 5059–5065.

Yang, Y., Wen, X.P., Wu, Z.G., Wang, K., Zhu, Y.X., 2023. Large-scale long terminal repeat insertions produced a significant set of novel transcripts in cotton. Sci. China Life Sci. https://doi.org/10.1007/s11427-022-2341-8.