OXFORD (GIGA)n SCIENCE

GigaScience, 2022, **11**, 1–8
DOI: 10.1093/gigascience/giac118
Technical Note

3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources

Mihaly Varadi [1,*,†], Sreenath Nair [1,†], Ian Sillitoe [2,†], Gerardo Tauriello [3,13,†], Stephen Anyango [1], Stefan Bienert [3,13], Clemente Borges [4,5], Mandar Deshpande [1], Tim Green [6], Demis Hassabis [6], Andras Hatos [7,8,14,15,16], Tamas Hegedus [9], Maarten L. Hekkelman [10], Robbie Joosten [10], John Jumper [6], Agata Laydon [6], Dmitry Molodenskiy [4,5], Damiano Piovesan [7], Edoardo Salladini [7], Steven L. Salzberg [11], Markus J. Sommer [11], Martin Steinegger [12], Erzsebet Suhajda [9], Dmitri Svergun [4,5], Luiggi Tenorio-Ku [7], Silvio Tosatto [7], Kathryn Tunyasuvunakool [6], Andrew Mark Waterhouse[3,13], Augustin Žídek [6], Torsten Schwede [3,13,*], Christine Orengo [2,*] and Sameer Velankar [1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton CB10 1SA, UK
[2]Department of Structural and Molecular Biology, UCL, London WC1E 6BT, UK
[3]Biozentrum, University of Basel, Basel 4056, Switzerland
[4]Computational Structural Biology, SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland
[5]European Molecular Biology Laboratory, EMBL Hamburg, Hamburg 69117, Germany
[6]DeepMind, London EC4A 3TW, UK
[7]Department of Biomedical Sciences, University of Padova, Padova 35129, Italy
[8]Department of Oncology, Lausanne University Hospital, Lausanne 1015, Switzerland
[9]Department of Biophysics and Radiation Biology, Semmelweis University, Budapest 1094, Hungary
[10]Netherlands Cancer Institute, Amsterdam 1066 CX, The Netherlands
[11]Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA
[12]School of Biology, Seoul National University, Seoul 82-2-880-6971, 6977, South Korea
[13]Computational Structural Biology, SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland
[14]Department of Computational Biology, University of Lausanne, Lausanne 1015, Switzerland
[15]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland
[16]Swiss Cancer Center Leman, Lausanne 1005, Switzerland
*Correspondence address. Mihaly Varadi, PDBe team, Wellcome Trust Genome Campus, Saffron Walden CB10 1SA, UK. E-mail: mvaradi@ebi.ac.uk
[†]These authors contributed equally.

Abstract

While scientists can often infer the biological function of proteins from their 3-dimensional quaternary structures, the gap between the number of known protein sequences and their experimentally determined structures keeps increasing. A potential solution to this problem is presented by ever more sophisticated computational protein modeling approaches. While often powerful on their own, most methods have strengths and weaknesses. Therefore, it benefits researchers to examine models from various model providers and perform comparative analysis to identify what models can best address their specific use cases. To make data from a large array of model providers more easily accessible to the broader scientific community, we established 3D-Beacons, a collaborative initiative to create a federated network with unified data access mechanisms. The 3D-Beacons Network allows researchers to collate coordinate files and metadata for experimentally determined and theoretical protein models from state-of-the-art and specialist model providers and also from the Protein Data Bank.

Keywords: structural biology, experimentally determined structures computationally predicted structures, federated data network, bioinformatics

Introduction

Proteins are essential building blocks of almost every biological process; therefore, understanding their functions is critical to many applications, from drug discovery [1, 2] to tackling environmental challenges such as plastic pollution [3]. Accurate information on the structure of a protein, especially in the context of its biological assembly, can help scientists understand and modulate its function [4, 5].

Unfortunately, gaining such insights regarding the function of proteins through their structures is severely hampered by the lack of high-quality, experimentally determined structures. As of 2022, the Universal Protein Resource (UniProt) contains around 204 million nonredundant amino acid sequences, while the Protein Data Bank (PDB) [6, 7] contains around 190,000 PDB entries mapped to approximately 52,000 UniProt accessions. In other words, less than 0.03% of all the known protein sequences have experimentally determined atomic resolution structures. As sequencing becomes more accessible, the gap between protein sequences and structures increases (Fig. 1).

Received: August 11, 2022. **Revised:** September 20, 2022. **Accepted:** November 11, 2022

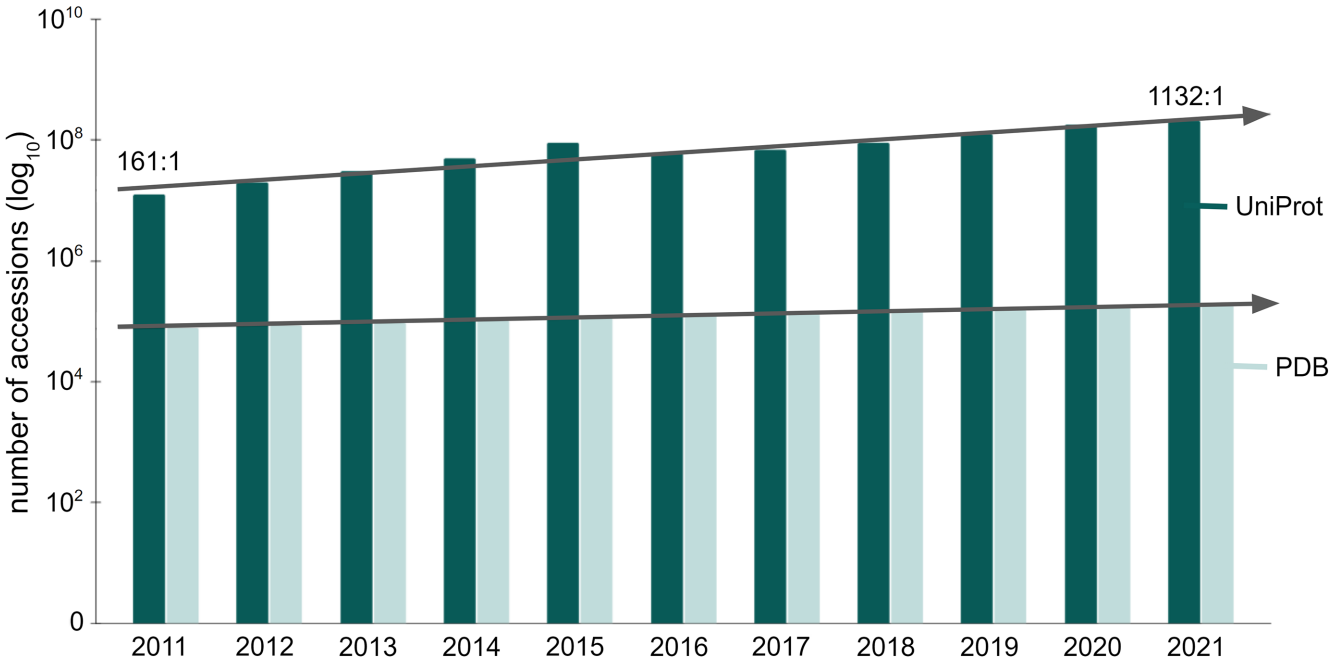

Figure 1: Growth of the UniProt and the PDB databases. This figure shows the number of accessions (on a logarithmic scale) throughout the past decade. In 2011, the UniProt had 161× as many protein sequences as the number of PDB entries. This ratio grew by an order of magnitude and was 1,132 to 1 in 2021, showing that the gap between known protein sequences and their structures keeps increasing.

A practical approach to addressing this challenge relies on high-accuracy computational models to complement the experimentally determined structures when the latter are unavailable for a certain protein of interest [8]. The thermodynamic hypothesis postulates that within certain limitations, the native structure is determined only by the protein's amino acid sequence [9, 10]. Indeed, the past 50 years saw the development of many algorithms and scientific software to predict protein structures [11, 12]. An approach developed early in this field was to use homologous protein structures as templates. Several modeling tools and data resources have long provided access to such models, for example, the SWISS-MODEL and the ModBase web services and databases [13–15]. In 2021, the field saw tremendous advances with tools such as AlphaFold and RoseTTAFold achieving much higher accuracy for *de novo* predictions without homologous templates than ever before [16, 17]. This new generation of prediction tools makes it possible to try and predict the structure of virtually any known protein based on its sequence.

While these new techniques are increasingly accurate, it is important that they are supplemented with reliable estimates of model confidence both for the whole model and locally for each residue. Researchers should not expect all predictions to be equally accurate neither globally nor in every region, and confidence estimates should hence be used to determine if a predicted structure can be used for downstream analysis [18]. Commonly used model confidence methods aim to predict the global and local similarity of the model compared to the correct coordinates if those coordinates were provided by an experimentally determined structure. In recent years, several model prediction methods such as SWISS-MODEL [14], RoseTTAFold [17], and AlphaFold [16] have chosen the superposition-free local distance difference test (lDDT) score [19] as a similarity metric to provide model confidence for their own models. The lDDT score measures differences in interatomic distances within a short radius between model and reference structure. It has been shown that superposition-free measures are robust with respect to domain movements and have advantages for the analysis of local structural details [20]. Similarly, superposition-free measures have been used for a long time in the creation of experimental structure models [21].

Another important consideration when relying on any structure prediction tool is to consider its limitations. While structures in the PDB have the advantage of experimental data backing the coordinates, enabling experimental as well as geometric validation, it is a relatively small dataset, as discussed above. Template-based models have the distinct advantage of enabling the mapping of a model to homologues with known structures, thus mapping to experimentally derived structures that can be in distinct conformational states or in complex with other molecules. Some tools excel at general-purpose protein structure modeling; others specialize in placing relevant ligands in the context of a model or representing conformational flexibility with ensembles of potential conformations [14, 16, 22–24] (Fig. 2). For example, AlphaFold 2.0 cannot perform docking of small molecules, even if they are obligate ligands of the proteins, such as Zinc-finger proteins. However, data resources such as AlphaFill can tackle this problem by building on existing models and adding known ligands to these structures [23] (Fig. 2A). On the other hand, the central repository of AlphaFold models, the AlphaFold Structure Database, only contains predictions for single polypeptide chains and not necessarily the functional forms of proteins [25]. In the case of multimeric complexes, the functional form can include several polypeptide chains. Since the number of known protein complexes is immense, having a comprehensive database for complex structures soon is rather unlikely. Therefore, integrating 3-dimensional data from experts in specialized fields of proteins is important, as demonstrated by physiologically and pathologically relevant transmembrane ABC half transporters [26] and by a set of computed structures of core eukaryotic protein complexes deposited in the ModelArchive [27]. Databases such as the Small-Angle Scattering Biological Data Bank (SASBDB) [28] and the Pro-

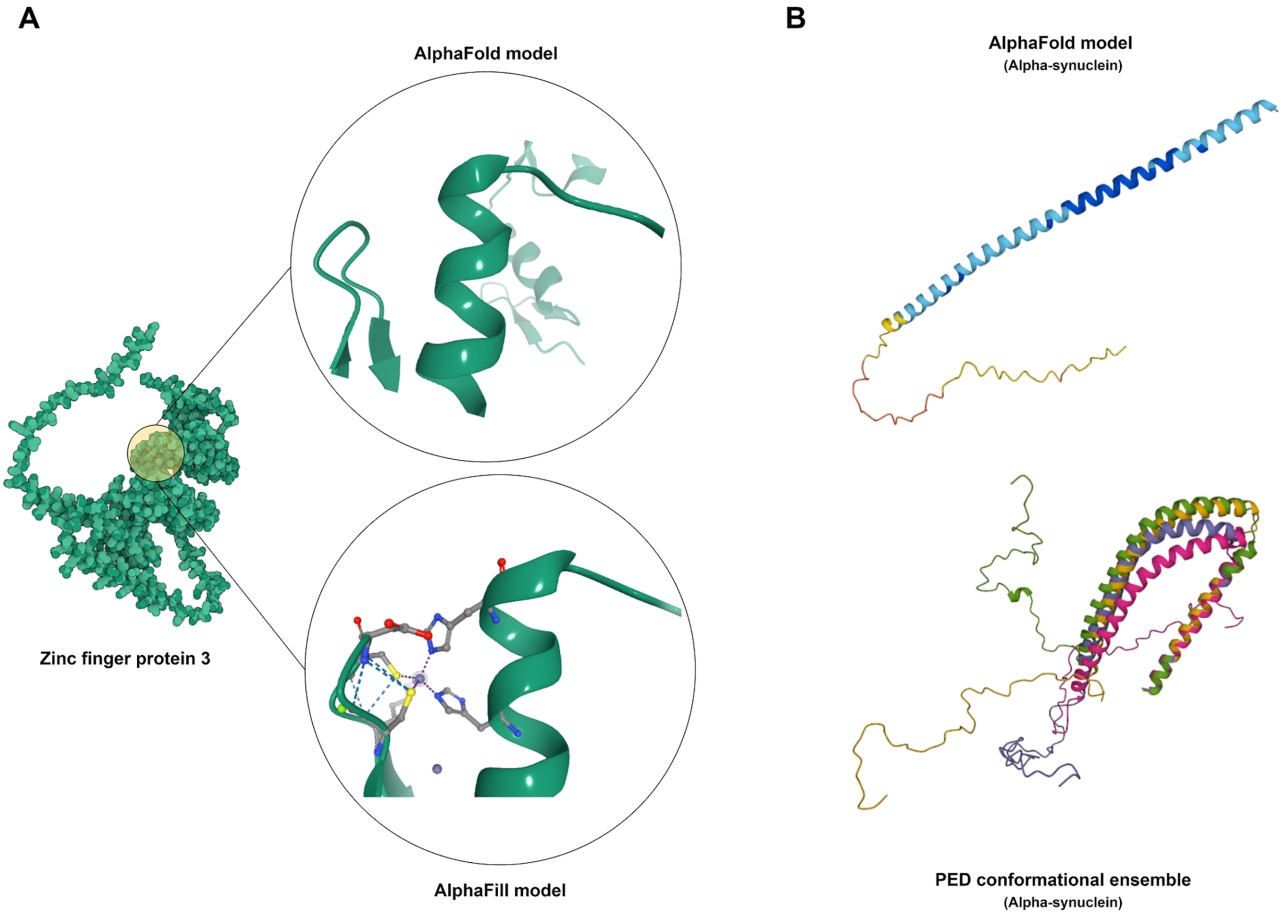

Figure 2: Highlighting the strengths and weaknesses of modeling techniques. Each modeling approach has limitations and specific strengths. For example, AlphaFill complements AlphaFold models by placing obligate ligands in their contexts (A). Other data providers, such as the Protein Ensemble Database, provide conformational ensembles for intrinsically disordered proteins (IDPs), for example, for the human Alpha-synuclein (B).

tein Ensemble Database (PED) [22] highlight the dynamic nature of intrinsically disordered proteins (Fig. 2B). Small-angle scattering provides low-resolution information on the shape and size of biological macromolecules in solution, but it also offers powerful means for the quantitative analysis of flexible systems, including intrinsically disordered proteins (IDPs) [29]. Theses data, together with *ab initio* modeling approaches, can be used to generate an experimentally validated pool of IDP models. PED provides access to such conformational ensembles but also those based on other experimental approaches. Considering the limitations of certain tools highlights the importance of using models and methods from various synergistic software and data providers to mitigate the weaknesses of individual modeling techniques.

While many prediction software and several publicly accessible data resources host and archive protein structures, these resources are fragmented and often rely on their own data standards to describe the necessary meta-information essential for providing context for a specific model. They also offer distinct data access mechanisms, requiring the users to learn multiple sets of technical details when interacting with various resources. The lack of standardization can severely impede the comparative analysis of these models, making it difficult to gain valuable insights.

Here, we present the 3D-Beacons Network, an open, collaborative platform for providing programmatic access to 3-dimensional coordinates and their standardized meta-information from both experimentally determined and computationally modeled protein structures.

Results

The 3D-Beacons Network is an open collaboration between providers of experimentally determined and computationally predicted protein structures. To date, 10 data providers make their protein structures available through this platform (Table 1). The consortium is guided by a collaboration agreement that prospective data providers agree to comply with. We encourage and invite macromolecular structure providers from research teams focusing on small, curated datasets to large data resources to join the 3D-Beacons Network and take advantage of its infrastructure to make their models more accessible to the scientific community. Importantly, all the data provided through the network must be freely available for academic and commercial use under Creative Commons Attribution 4.0 license terms.

The 3D-Beacons Network is based on an infrastructure that helps providers of protein structures to standardize their meta-information and easily link their model files to a centralized search engine, called the 3D-Beacons Hub API (application programming interface) (Fig. 3). Each data provider has its 3D-Beacon connected to the central hub. The hub is the public access point through which the users (or other data services) can retrieve mod-

Table 1: Members of the 3D-Beacons Network

Data provider	Model category	Number of structures*
AlphaFill	Template based	995,411
AlphaFold DB	*Ab initio*	214,684,311
Genome3D	Template based	*In progress*
HegeLab	*Ab initio*	15
isoform.io	*Ab initio*	48,551
ModelArchive	*Ab initio*/template based	1,106
PDBe	Experimentally determined	190,639
PED	Conformation ensembles	275
SASBDB	Experimentally determined	3,912
SWISS-MODEL Repository	Template based	2,216,915

*Numbers are accurate as of 29 July 2022.

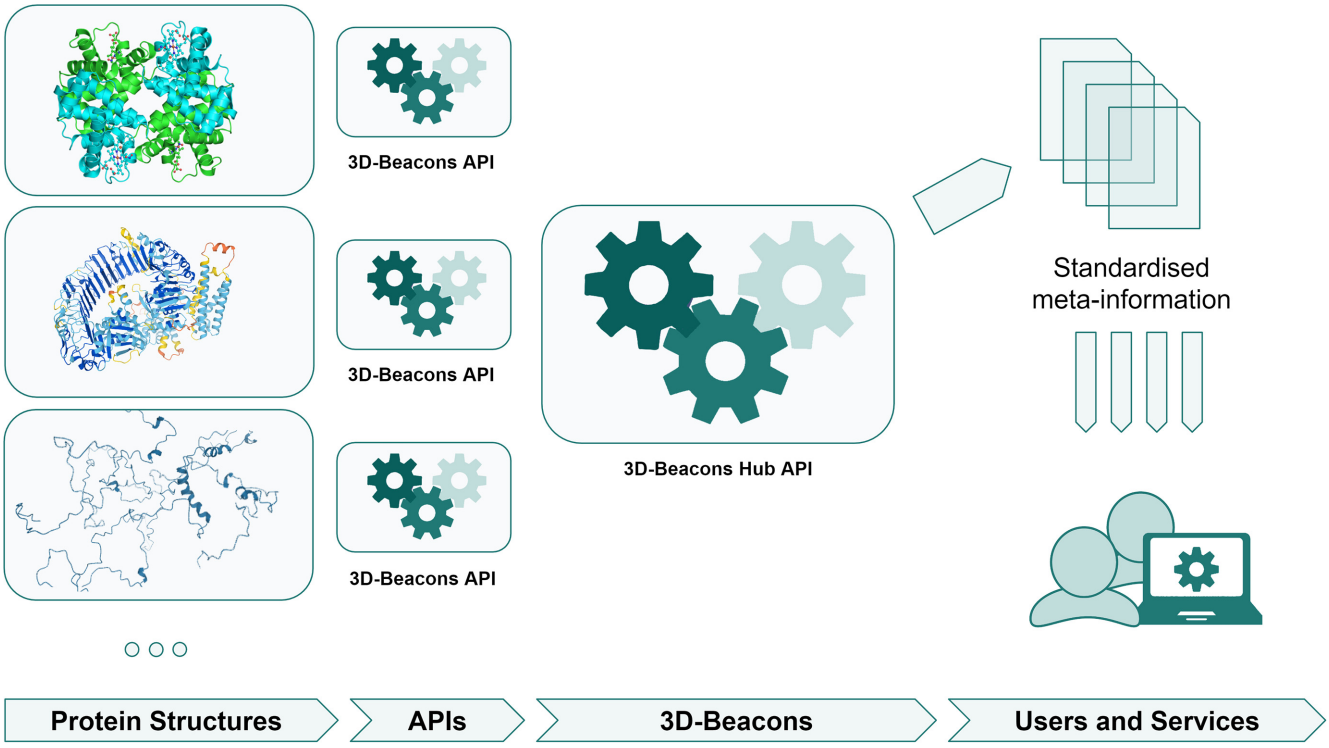

Figure 3: Schematic overview of the 3D-Beacons Network. Data providers standardize their meta-information and make their models available through 3D-Beacons API instances. The 3D-Beacons Registry links these instances to the central 3D-Beacons Hub API, which can be openly accessed by the scientific community and other data services.

els from any members. This allows users to get all structures for a given UniProt accession instead of manually retrieving them from all the different structure providers.

Thanks to the standardized data formats, the infrastructure ensures complete transparency in data provenance and allows users to easily compare protein structures and their relevant meta-information. This initiative has evolved in parallel with efforts to improve the standardization of the coordinate files for theoretical models. In particular, members of the 3D-Beacons Network contributed to the ModelCIF extension of the PDBx/mmCIF format, which supports more exhaustive meta-information and includes mappings to the corresponding UniProt accessions next to the atomic coordinates.

While the primary purpose of 3D-Beacons is to provide efficient and scalable programmatic access to protein structures, we also offer a graphical user interface that allows researchers to get an overview of the available protein structures. For example, users can view all the available data from any member data provider for the human cellular tumor antigen p53 protein by searching based on its UniProt accession (Fig. 4).

We divided the protein structures into 4 categories: (i) experimentally determined, (ii) template based, (iii) *ab initio*, and (iv) conformational ensembles. We defined the categories as follows:

Experimentally determined structures are based on data from techniques such as X-ray crystallography, cryo-electron microscopy, nuclear magnetic resonance spectroscopy, or small-angle scattering. This category is exemplified by structures in the PDB and the SASBDB databases.

Template-based models use alignments to similar sequences with known structure (i.e., templates) as their main input. SWISS-MODEL is an example of data providers with such models.

Ab initio models can use templates as an auxiliary input but do not depend on them. AlphaFold models are considered *ab initio* in this framework.

Finally, **conformational ensembles** are created using a combination of experimental data and computational modeling, yield-

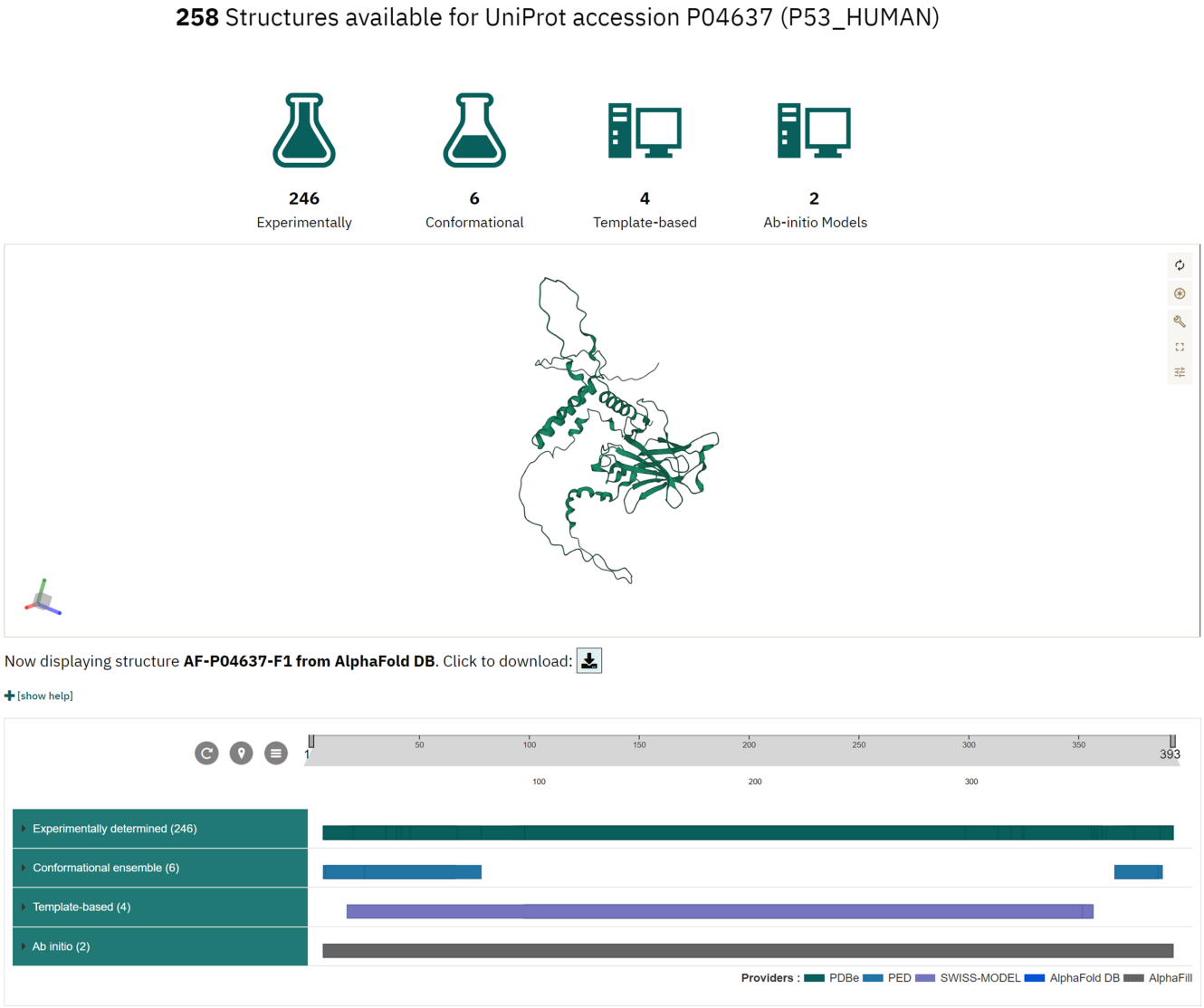

Figure 4: Graphical user interface of 3D-Beacons. While the main focus of the 3D-Beacons Network is to provide programmatic access to experimentally determined and computationally predicted protein structures, we also provide a graphical user interface where researchers can query for specific proteins using UniProt accessions. This interface displays which section of the protein sequence the models cover and provides an interactive 3-dimensional view.

ing a large number of possible conformations. Ensembles in the PED database are an example of this category.

Researchers can view the number of models under each category and inspect which parts of the amino acid sequences are covered by which models in a 2D viewer, PDB ProtVista [30]. Users can also display the structures using an embedded 3-dimensional molecular graphics viewer, Mol* [31], and download the models in PDB or mmCIF formats.

Discussion

The purpose of the 3D-Beacons Network is to standardize the representation of protein structure models and associated metadata and to provide efficient, high-throughput programmatic access to experimentally determined and theoretical models and their standardized metadata. The current version (as of 29 July 2022) of 3D-Beacons supports querying any number of UniProt accessions, while future updates are planned to collate models based on other identifiers such as taxonomy IDs or domain IDs. This platform enables both the scientific community and developers of data visualization and data-providing services to access and seamlessly integrate 3-dimensional models from various protein structure data providers.

While designing the data access points and data formats, we had extensive discussions with scientists and developers who provided specific use cases that are relevant to their work. We used these data to drive the development of 3D-Beacons, starting with the most frequently requested data (i.e., information keyed on UniProt accessions) that can answer the question, "What experimental or theoretical structures are available for my protein of interest?" Going forward, we will address more of the collated use cases, such as searching by sequence or by gene identifiers and selecting structures based on protein families. Already, the API endpoints of 3D-Beacons provide easy access to models from sparse and fragmented data resources, supporting researchers and software developers alike.

For example, the 3D-Beacons infrastructure allows users of Jalview, a workbench for creating multiple sequence alignments

(MSAs) and analyzing them, to discover 3-dimensional models for MSAs of proteins from the UniProt and place them in the context of genetic variation from Ensembl [32]. It can also visualize local model quality scores such as pLDDT (Predicted Local Distance Difference Test).

The Protein Data Bank in Europe–Knowledge Base (PDBe-KB) [33] displays all the experimentally determined and computationally predicted structures for proteins of interest on their aggregated views of proteins. To retrieve metadata and the location of model files, it uses the 3D-Beacons Hub API. This integration also allows PDBe-KB to display functional and biophysical annotations both for theoretical models in addition to experimentally determined structures.

The SWISS-MODEL Repository (SMR) [13] fetches models from AlphaFold DB and the ModelArchive using the 3D-Beacons Hub API. SMR displays these models next to homology models from SWISS-MODEL [14] and experimental structures from the PDB [6] to facilitate comparative analysis. SMR also takes advantage of the confidence measure information, and the models are displayed with a consistent coloring based on these confidence metrics.

By providing easy access to experimentally determined and computationally predicted protein structures, we aim to make these data an essential part of the toolbox of researchers in the broader scientific fields of life sciences. Establishing an infrastructure of federated model providers if a scalable and expandable approach can efficiently adjust to include new models and provides a more sustainable model than if a single data repository would try and archive all the data in one place. By taking advantage of the 3D-Beacons Network, protein structures can better realize their full impact on fields from structure-based drug discovery [2, 34] to structural bioinformatics [35, 36] and from scientific software development [37] to experimental structure determination [38, 39]. The amount of available protein structures has never been as large as it is now, and providing convenient access to these models is a key service that will enable further research.

Methods

The infrastructure of the 3D-Beacons Network consists of a registry, a hub, and the data access implementations. The 3D-Beacons Network is open to data providers of protein structures. Such data resources are invited to contact the 3D-Beacons consortium to discuss ways their data can be linked. Briefly, the common steps are as follows: data providers review the consortium guidelines and the latest API specification. The data providers then convert their metadata to the specified format and make these data available either through their APIs or by setting up a 3D-Beacon client. Once these steps are completed, the registry can be updated to link the new data resource with the 3D-Beacons Hub API. The following sections give more detailed information on each of these elements of the infrastructure.

3D-Beacons Registry

The 3D-Beacons Registry is a transparent, publicly accessible registry that stores information on all the data providers linked to the 3D-Beacons Network. The registry is available on GitHub. It contains information on the public URLs of data providers, a brief description of the protein structures they provide, and a list of API endpoints they support. For example, PDBe [40, 41] supports the API endpoint that is keyed on a UniProt accession and that provides high-level information about the models, while SMR [13] supports both the high-level and the detailed API endpoints, which additionally provides per-chain and per-residue information on the models.

3D-Beacons data exchange format

The API endpoints comply with the data exchange format, which the 3D-Beacons members collaboratively design and improve. We defined the data exchange format as a JavaScript Object Notation (JSON) specification, an industry-standard format for sharing textual meta-information. The specification is available on Apiary and GitHub.

3D-Beacons client

Members of the 3D-Beacons Network can either implement their own API endpoints according to the API specification described above or install a local instance of the 3D-Beacons client. This client is a Docker-containerized, lightweight Python package that can import and parse PDB or mmCIF formatted protein structure files and their corresponding meta-information (in JSON format). It also includes capabilities to add model confidence scores using QMEANDisCo [42] if models do not already include comparable scores such as pLDDT. QMEANDisCo, which is used internally by SWISS-MODEL, can be applied to models from any provider and has proven to be an accurate confidence predictor for homology modeling and some *ab initio* methods [20]. The client indexes the collated data in an embedded MongoDB database instance and exposes the information through an embedded API implementation that complies with the 3D-Beacons API specifications. The client is freely available on GitHub.

3D-Beacons Hub API

At the core of the 3D-Beacons infrastructure lies the Hub API, a programmatic aggregator of the meta-information from all the member data providers. We implemented the Hub API using the FastAPI framework. This API relies on the previously described registry to retrieve information on which data provider supports which specific API endpoints. It aggregates data and provides its own API endpoints that researchers, services, and software can directly access to retrieve the location of available model files and their corresponding meta-information, such as the overall model quality or residue-level confidence measures. It is important to note that in the current implementation, the model confidence measures are provided by the original data sources, and different providers might have different approaches to estimating confidence. This can hamper effective comparison of the models based on these scores, and it is an active focus area both within the 3D-Beacons Network and the broader modeling community to design a broadly applicable confidence measure.

3D-Beacons front-end

Finally, we provide a graphical user interface that contains documentation and showcases the information one can retrieve using the 3D-Beacons Hub API. We implemented this interface using the Angular framework, and it relies on the sequence feature viewer, PDB ProtVista [30], and the 3D molecular graphics viewer, Mol* [31]. The source code of this front-end application is available from GitHub.

Availability of Supporting Source Code and Requirements

The source codes of the 3D-Beacons Registry, Client, Hub API, and front-end application are all publicly available:

Project name: 3D-Beacons
Project homepage: https://3d-beacons.org
Operating system(s): Platform independent
Programming language: Python, TypeScript
Other requirements: Python 3.7 or higher, Angular 11.1.3 or higher
License: Apache License 2.0
biotools: 3d-beacons

Data Availability

All the data provided through the network are freely available for academic and commercial use under Creative Commons Attribution 4.0 license terms. Documentation of the 3D-Beacons Hub API is available at https://www.ebi.ac.uk/pdbe/pdbe-kb/3dbeacons/api/. The specification of the data exchange format is available at https://3dbeacons.docs.apiary.io/#. An archival copy of the code and other supporting data are also available via the *GigaScience* database GigaDB [43].

Abbreviations

API: Application Programming Interface; IDP: intrinsically disordered proteins; JSON: JavaScript Object Notation; lDDT: local distance difference test; MSA: multiple sequence alignment; PDB: Protein Data Bank; PDBe-KB: Protein Data Bank in Europe–Knowledge Base; PED: Protein Ensemble Database; SASBDB: Small-Angle Scattering Biological Data Bank; SMR: SWISS-MODEL Repository; UniProt: Universal Protein Resource.

Competing Interests

The authors declare no competing interests.

Authors' Contributions

M.V. created the initial draft, handled project and data management at PDBe and AlphaFold DB, and designed and developed the 3D-Beacons webpages. S.N. led the development of the 3D-Beacons registry, Hub API, and PDBe and AlphaFold API implementation, as well as contributed to the development of the 3D-Beacons client and the webpages. I.S. worked on the 3D-Beacons client and on the Genome3D beacon. G.T. contributed to the management of 3D-Beacons. M.V., G.T., A.M.W., and S.B. contributed to the API design. A.L. and N.A. provided management support for AlphaFold DB. A.H., S.T., L.T.-K., E.S., and D.P. contributed to the PED beacon. A.M.W., S.B., and G.T. contributed to the SWISS-MODEL and ModelArchive beacons. S.B. contributed QMEANDisCo for the client. S.B. and G.T. connected with the ModelCIF working group. T.H. and E.S. contributed to the HegeLab beacon. M.L.H. and R.P.J. contributed to the AlphaFill beacon. C.B., Dm.S, and D.M. contributed to the SASBDB beacon. M.S., M.J.S., and S.L.S. contributed to the isoform.io beacon. M.D. and S.A. worked on data visualization and infrastructure. S.V., C.O., and T.S. provided oversight as co-principal investigators. I.S., G.T., D.M., T.H., S.V., R.J., and E.S. contributed to the manuscript drafts. Every coauthor reviewed the final manuscript.

Funding

Work on creating the 3D-Beacons infrastructure was primarily funded by the BBSRC grant BB/S020071/1. We acknowledge the contribution of ELIXIR BioHackathon participants in 2020 and 2021 who helped improve various aspects of the 3D-Beacons infrastructure. G.T., A.M.W., S.B., and T.S. acknowledge funding from ELIXIR and the SIB Swiss Institute of Bioinformatics. D.I.S. and D.S.M. acknowledge funding from the German Ministry of Science and Education Grant Number: 16QK10A-SAS-BSOFT. M.S. acknowledges support from the National Research Foundation of Korea (NRF) Grant Number: 2021-R1C1-C102065.

References

1. Batool, M, Ahmad, B, Choi, SA. Structure-based drug discovery paradigm. *Int J Mol Sci* 2019;**20**(11):2783.
2. Ochoa, D, Hercules, A, Carmona, M, *et al.* Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res* 2021;**49**(D1):D1302–10.
3. Zhu, B, Wang, D, Wei, N. Enzyme discovery and engineering for sustainable plastic recycling. *Trends Biotechnol* 2022;**40**(1):22–37.
4. Lee, D, Redfern, O, Orengo, C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 2007;**8**(12):995–1005.
5. Varadi, M, Berrisford, J, Deshpande, M, *et al.* PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res* 2020;**48**(D1):D344–53.
6. wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;**47**(D1):D520–8.
7. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**(D1):D480–9.
8. Akdel, M, Pires, DEV, Pardo, EP, *et al.* A structural biology community assessment of AlphaFold 2 applications. *Nat Struct Mol Biol* 2022;**29**:1056–1067.
9. Anfinsen, CB. Principles that govern the folding of protein chains. *Science* 1973;**181**(4096):223–30.
10. Hirata, F, Sugita, M, Yoshida, M, *et al.* Perspective: structural fluctuation of protein and Anfinsen's thermodynamic hypothesis. *J Chem Phys* 2018;**148**(2):020901.
11. Masrati, G, Landau, M, Ben-Tal, N, *et al.* Integrative structural biology in the era of accurate structure prediction. *J Mol Biol* 2021;**433**(20):167127.
12. Pereira, J, Simpkin, AJ, Hartmann, MD, *et al.* High-accuracy protein structure prediction in CASP14. *Proteins Struct Funct Bioinf* 2021;**89**(12):1687–99.
13. Bienert, S, Waterhouse, A, de Beer, TAP, *et al.* The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res* 2017;**45**(D1):D313–9.
14. Waterhouse, A, Bertoni, M, Bienert, S, *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;**46**(W1):W296–303.
15. Pieper, U, Webb, BM, Dong, GQ, *et al.* ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 2014;**42**(D1):D336–46.
16. Jumper, J, Evans, R, Pritzel, A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**(7873):583–9.
17. Baek, M, Dimaio, F, Anishchenko, I, *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**(6557):871–6.
18. Schwede, T. Protein modeling: what happened to the "protein structure gap"? *Structure* 2013;**21**:1531–40.
19. Mariani, V, Biasini, M, Barbato, A, *et al.* lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 2013;**29**(21):2722–8.

20. Olechnoviä, K, Monastyrskyy, B, Kryshtafovych, A, *et al.* Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics* 2019;**35**(6):937–44.
21. Smart, OS, Womack, TO, Flensburg, C, *et al.* Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr D Biol Crystallogr* 2012;**68**(4):368–80.
22. Lazar, T, Martínez-Pérez, E, Quaglia, F, *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res* 2021;**49**(D1):D404–11.
23. Hekkelman, ML, de Vries, I, Joosten, RP, *et al.* AlphaFill: enriching the AlphaFold models with ligands and co-factors. bioRxiv 2021. https://doi.org/10.1101/2021.11.26.470110.
24. Waman, VP, Blundell, TL, Buchan, DWA, *et al.* The Genome3D Consortium for structural annotations of selected model organisms. *Methods Mol Biol* 2020;**2165**:27–67.
25. Varadi, M, Anyango, S, Deshpande, M, *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**(D1):D439–44.
26. Tordai, H, Suhajda, E, Sillitoe, I, *et al.* Comprehensive collection and prediction of ABC transmembrane protein structures in the AI era of structural biology.*Int J Mol Sci* 2022 **23**(16): 8877.
27. Humphreys, IR, Pei, J, Baek, M, *et al.* Computed structures of core eukaryotic protein complexes. *Science* 2021;**374**(6573): eabm4805.
28. Kikhney, AG, Borges, CR, Molodenskiy, DS, *et al.* SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Sci* 2020;**29**(1):66–75.
29. Kikhney, AG, Svergun, DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* 2015;**589**(19, Pt A):2570–7.
30. Deshpande, M, Varadi, M, Paysan-Lafosse, T, *et al.* PDB ProtVista: a reusable and open-source sequence feature viewer. 2022. https://doi.org/10.1101/2022.07.22.500790.
31. Sehnal, D, Bittrich, S, Deshpande, M, *et al.* Mol∗ Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 2021;**49**(W1):W431–7.
32. Procter, JB, Mungo Carstairs, G, Soares, B, *et al.* Alignment of biological sequences with Jalview. *Methods Mol Biol* 2021;**2231**:203–24.
33. Varadi, M, Anyango, S, Armstrong, D, *et al.* PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res* 2022;**50**(D1):D534–42.
34. Xue, HT, Stanley-Baker, M, Kong, AWK, *et al.* Data considerations for predictive modeling applied to the discovery of bioactive natural products. *Drug Discovery Today* 2022;**27**(8):2235–43.
35. Bludau, I, Willems, S, Zeng, W-F, *et al.* The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol* 2022;**20**(5):e3001636.
36. Tian, R, Li, Y, Wang, X, *et al.* A pharmacoinformatics analysis of artemisinin targets and de novo design of hits for treating ulcerative colitis. *Front Pharmacol* 2022;**13**:843043.
37. Bordin, N, Sillitoe, I, Nallapareddy, V, *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *bioRxiv* 2022. https://doi.org/10.1101/2022.06.02.494367.
38. Cai, SW, Zinder, JC, Svetlov, V, *et al.* Cryo-EM structure of the human CST-Pol*α*/primase complex in a recruitment state. *Nat Struct Mol Biol* 2022;**29**:8813–9.
39. Yu, Y, Li, S, Ser, Z, *et al.* Cryo-EM structure of DNA-bound Smc5/6 reveals DNA clamping enabled by multi-subunit conformational changes. *Proc Natl Acad Sci* 2022;**119**(23):e2202799119.
40. Armstrong, DR, Berrisford, JM, Conroy, MJ, *et al.* PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res* 2020;**48**(D1): D335–43.
41. Nair, S, Váradi, M, Nadzirin, N, *et al.* PDBe aggregated API: programmatic access to an integrative knowledge graph of molecular structure data. *Bioinformatics* 2021;**37**(21):3950–2.
42. Studer, G, Rempfer, C, Waterhouse, AM, *et al.* QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics* 2020;**36**(6):1765–71.
43. Varadi, M, Nair, S, Sillitoe, I, *et al.* Supporting data for "3D-Beacons: Decreasing the gap between protein sequences and structures through a federated network of protein structure data resources." *GigaScience Database*. 2022. http://dx.doi.org/10.5524/102328.