Structural Dynamics ARTICLE pubs.aip.org/aip/sdy

Identifying protein conformational states in the Protein Data Bank: Toward unlocking the potential of integrative dynamics studies

Cite as: Struct. Dyn. **11**, 034701 (2024); doi: 10.1063/4.0000251
Submitted: 7 March 2024 · Accepted: 8 May 2024 ·
Published Online: 17 May 2024

View Online Export Citation CrossMark

Joseph I. J. Ellaway,[1] Stephen Anyango,[1] Sreenath Nair,[1] Hossam A. Zaki,[2] Nurul Nadzirin,[1] Harold R. Powell,[3] Aleksandras Gutmanas,[4] Mihaly Varadi,[1] and Sameer Velankar[1,a)]

AFFILIATIONS

[1]Protein Data Bank in Europe, European Bioinformatics Institute, Hinxton, United Kingdom
[2]The Warren Alpert Medical School of Brown University, Providence, Rhode Island 02903, USA
[3]Imperial College London, Department of Life Sciences, London, United Kingdom
[4]WaveBreak Therapeutics Ltd., Clarendon House, Clarendon Road, Cambridge, United Kingdom

Note: Paper published as part of the special topic Tribute to Olga Kennard (1924-2023).
[a)]**Author to whom correspondence should be addressed:** sameer@ebi.ac.uk

ABSTRACT

Studying protein dynamics and conformational heterogeneity is crucial for understanding biomolecular systems and treating disease. Despite the deposition of over 215 000 macromolecular structures in the Protein Data Bank and the advent of AI-based structure prediction tools such as AlphaFold2, RoseTTAFold, and ESMFold, static representations are typically produced, which fail to fully capture macromolecular motion. Here, we discuss the importance of integrating experimental structures with computational clustering to explore the conformational landscapes that manifest protein function. We describe the method developed by the Protein Data Bank in Europe – Knowledge Base to identify distinct conformational states, demonstrate the resource's primary use cases, through examples, and discuss the need for further efforts to annotate protein conformations with functional information. Such initiatives will be crucial in unlocking the potential of protein dynamics data, expediting drug discovery research, and deepening our understanding of macromolecular mechanisms.

INTRODUCTION

As of February 2024, the Protein Data Bank (PDB),[1] the global repository of experimentally determined structures, hosts over 215 000 macromolecular structures. Recent advances in protein structure prediction—made by the new generation of AI-based tools such as AlphaFold2,[2] RoseTTAFold,[3] and ESMFold[4]—have predicted almost 1×10^9 further structures, archived in the AlphaFold Protein Structure Database (AFDB),[5] the ESM Metagenomic Atlas,[4] and the Model Archive.[6] Although significant work is ongoing to generate ensemble models, these tools generally predict a single structure per sequence.[7]

To realize the relationship between protein sequence, structure, and function, we must consider their dynamics—relative movements between residues. The structure of a protein navigates a high-dimensional conformational landscape, where stable conformations occupy free energy minima.[8] The transitions between these minima represent conformation changes, often crucial for protein function, both under physiological conditions or in disease progression.[9,10] Changes to the landscape's topology may be induced via ligand association, solvent packing, oligomerization, pH changes, or post-translational modification[11–16] [Fig. 1(a), right]. On the far end of the conformational flexibility spectrum are the intrinsically disordered proteins (IDPs), whose free energy landscapes lack deep energy minima [Fig. 1(b)], instead being littered with shallow dips that could become more favorable upon environment changes.[17–19] Investigating these landscapes requires diverse experimental techniques, each contributing unique insights into conformational states or motion of proteins[20–22] [Fig. 1(c)].

X-ray crystallography has been instrumental in providing atomic-resolution models of proteins. Despite its tendency to capture proteins in static states due to crystal packing, advancements such as temperature-jump and time-resolved serial femtosecond crystallography (SFX) can observe local dynamic processes within crystallized proteins.[23–27] In contrast, small-angle x-ray scattering (SAXS) is a low-resolution method

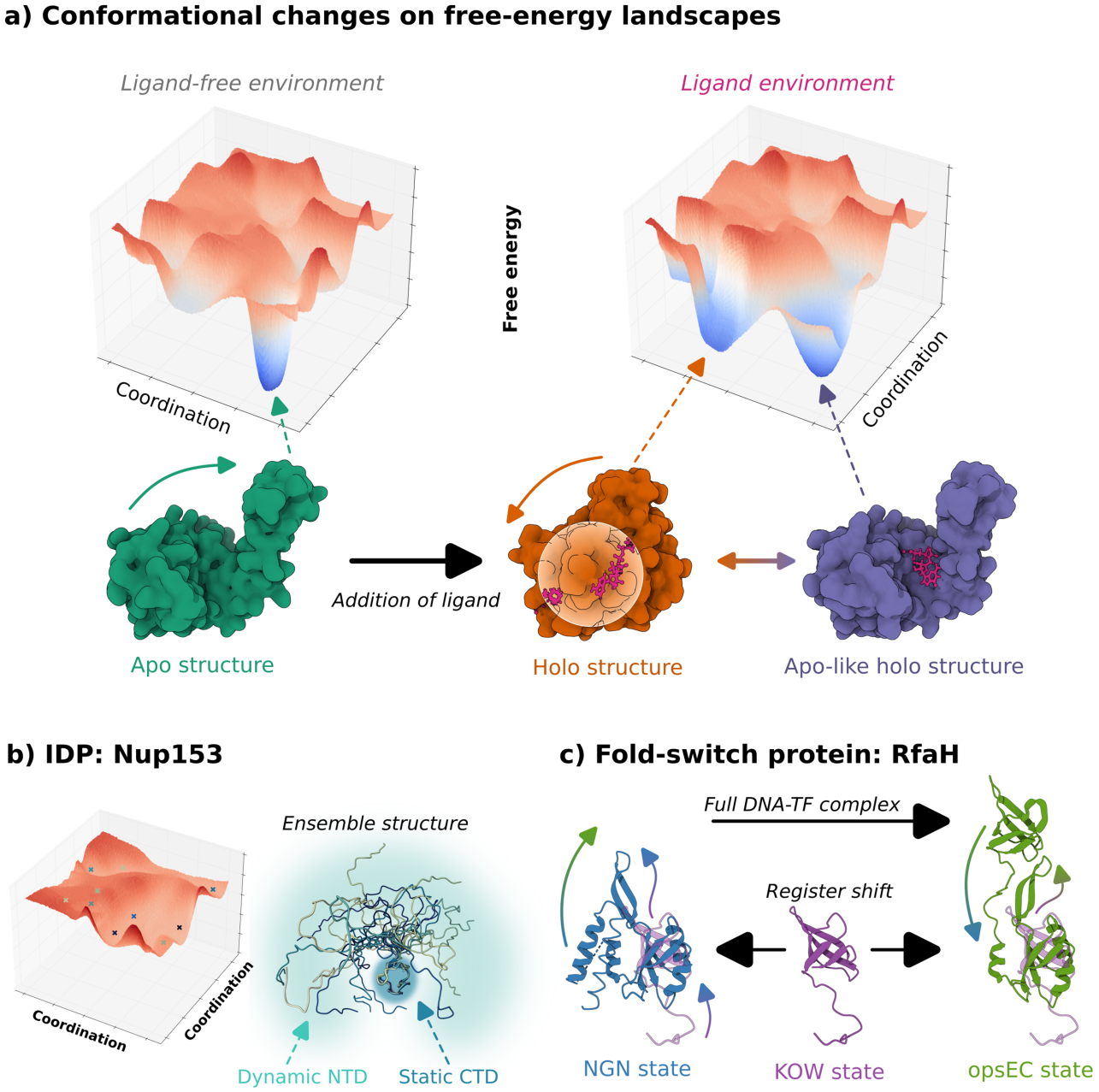

FIG. 1. Illustration of functional protein conformation changes. (a) Hypothetical free-energy landscape (top) of adenylate kinase's coordination state before (left) and after (right) ligand binding. A dominant minimum is plotted in the ligand-free environment, facilitating one apo conformation (PDB: 6S36, green). Addition of ADP changes the landscape to accommodate a second minimum for the adoption of a closed conformation (PDB: 8CRG, orange), while permitting the existence of the original conformation (PDB: 6F7U, indigo). An energy barrier between these states must be overcome to transition between the conformations. (b) Hypothetical free energy landscape of the human nuclear pore complex protein Nup153—an IDP (PDB: 2EBV). (c) Conformational states of *E. coli* transcription factor RfaH binding the NusG N-terminal domain (NGN, left), the NusG C-terminal domain (KOW, middle), and when bound to an operon polarity suppressor (*ops*) DNA sequence in the transcription elongation complex (opsEC, right).[33] KOW-bound structure is truncated—solved for the N-terminal domain only. PDB: 2OUG, blue; PDB: 2LCL, purple; PDB: 6C6S, green.

for studying larger, global conformation changes in solution.[28–31] Combined with experimentally derived or predicted atomic models, integrative SAXS models can offer impressively comprehensive views of macromolecular states and dynamics.[24,32]

In recent years, cryogenic electron microscopy (cryoEM) has dramatically improved to solve thousands of macromolecules—particularly those that are difficult to crystallize.[34] Like x-ray crystallography, cryoEM has traditionally produced single, static models from

three-dimensional (3D) projections of many images.[35,36] However, advances in direct-electron detectors and image classification software facilitate the reconstruction of conformational ensembles,[37–41] offering views of proteins in states closer to their physiological conditions. Furthermore, the study of individual molecules from cryo-electron tomography (cryoET) reveals the conformation of macromolecules *in situ*.[42–47] Such physiological insight was once the preserve of nuclear magnetic resonance (NMR) spectroscopy, which excels at detailing structure and dynamics over a range of timescales in near-physiological conditions.[48–51] NMR can detect both transient states[52] and intrinsically disordered regions,[53,54] providing insight into protein movements and interactions crucial for biological function. However, its application is typically limited to smaller proteins and complexes—complementing the data collected by cryoEM, which struggles to resolve smaller macromolecules.[38,48]

The success of AI-based tools at predicting protein structures from amino acid sequences has marked a significant milestone in structural biology.[2–4] However, modeling the conformational states of proteins remains a frontier,[55–59] as demonstrated by the general tendency of AlphaFold2 to predict structures in similar conformations.[56,60–65] Innovations have emerged where modifications to the multiple sequence alignments (MSAs), a key input for many structure prediction tools, enable the exploration of more diverse protein conformations.[7,61,64,66,67] For example, the AF-Cluster technique has demonstrated through experimental validation that AlphaFold can predict multiple states of the fold-switching protein KaiB.[65,68]

While structure prediction tools can help investigate conformational heterogeneity, molecular dynamics (MD) simulations remain indispensable for probing the theoretical dynamic behavior of macromolecules, complementing the generally static models provided by AI-based predictions and experimental data.[69–71] Despite their computational cost and the challenges associated with force field accuracy, MD simulations are invaluable tools for exploring the conformation space and potential biological activities of proteins, helping to identify novel ligand-binding sites crucial for drug discovery.[63,72–74]

Here, we describe the method the Protein Data Bank in Europe – Knowledge Base[75] (PDBe-KB) uses to aggregate and cluster protein conformational states, primarily from x-ray, cryoEM, and NMR structures deposited in the PDB.

METHODS

The first step of the clustering process is to collate polypeptide chains from the PDB with 100% sequence identity into groups called *segments* [Fig. 2(a)]. A single segment will contain only structures mapping to a contiguous section of their corresponding UniProt sequence, potentially resulting in multiple segments per UniProt sequence (such as truncated N- or C-terminal domains). Each polypeptide in the PDB archive is mapped to a corresponding UniProt sequence using the SIFTS annotation tool.[76,77] Only chains within segments are subsequently considered for clustering.

Next, we calculate the Euclidean distances between Cα atoms per residue pair, leading to a transformation-independent Cα distance matrix. Polypeptides are compared pairwise by calculating the absolute difference between their Cα distance matrices, capturing the chain–chain differences in Cα position, independent of the chains' original Cartesian coordinates. The distance matrix is filtered by reducing elements to zero if below 3 Å, removing small discrepancies in Cα placement between structures. To condense this filtered difference matrix, the upper diagonal elements are summated and normalized by multiplication with the fraction of modeled residues, penalizing any gaps in the structures [Fig. 2(b)]. This measure captures the GLObal CONformation (*GLOCON*) difference as a dissimilarity score between chains.

Next, we use UPGMA agglomerative clustering to group chains based on their GLOCON scores, splitting the segment into *clusters*—approximating potential conformational states [Figs. 2(c) and 2(d)]. Based on the GLOCON dissimilarity score, small structural differences (such as changes in loop position) are noticeable by this clustering method, such as in the manganese ABC transporter's Leu127-Lys135 region (UniProt accession: P0A4G2). However, small differences could be obscured where small and large differences occur (such as domain movements or fold switches). Reasonable separation into clusters is generally achievable at 70% of the maximum GLOCON score, although this threshold could be further optimized per segment. All chains are superposed (independently of the clustering step) using GESAMT, which identifies structurally conserved regions between possibly heterogeneous structures.[78] Where NMR structures are clustered, the first model of the ensemble is selected as a reference. PDBe runs this pipeline weekly,[75] predicting conformations for the entire PDB archive.

Alongside the experimentally derived structures, our process allows users to superpose the corresponding AlphaFold2 model, supplementing the cluster results. The root-mean-squared deviation of the AlphaFold2 model from each cluster's representative chain is calculated and displayed, allowing identification of the conformational state predicted by AlphaFold2. This comparison allows users to quickly identify the conformational state predicted by the full-length AlphaFold2 protein, potentially expediting functional characterization.

To test the clustering pipeline, we manually curated a benchmark dataset of polypeptide chains in the PDB archive that adopt open or closed conformations,[89] similar to previous datasets characterizing distinct secondary structure changes during fold switching.[79] An initial search identified 630 unique entries with descriptions of open or closed in their PDB entry title before filtering the results for spurious substrings (e.g., cyclo*pent*adienyl). Publications for the remaining 315 entries were read to designate labels of conformational states. The dataset comprises a range of structural variations at different scales, such as a ∼5 Å loop movement in α-fucosidase (UniProt accession: J9UN47), a set of intra-domain rearrangement of residues in NMR structures (e.g., PDB code: 6qeb) of human carbonic anhydrase (UniProt accession: P00918), and a ∼20 Å C-terminal domain movement at 5′-deoxynucleotidase's Glu332 hinge (UniProt accession: P21589). We make the dataset available through the PDBe-KB's FTP server and Kaggle.

All the data from the clustering process are openly accessible from the PDBe-KB FTP area, through API end points in the PDBe Aggregated API and via the PDBe-KB aggregated views of proteins. The code is open source and available on GitHub under the Apache 2.0 license.

RESULTS: NOTABLE EXAMPLES FROM THE ARCHIVE

The PDB provides a rich sampling of protein conformation space, where independently solved structures have identical sequences. Although a significant portion of the biologically meaningful conformation space has been captured, it is non-trivial to identify distinct conformations across all PDB entries.[80,81] For example, hexokinase from *Sulfurisphaera tokodaii* (UniProt accession: Q96Y14) is the first glycolytic enzyme that initializes respiration and is essential during anaerobic conditions [Fig. 3(a)]. The kinase is moderately promiscuous

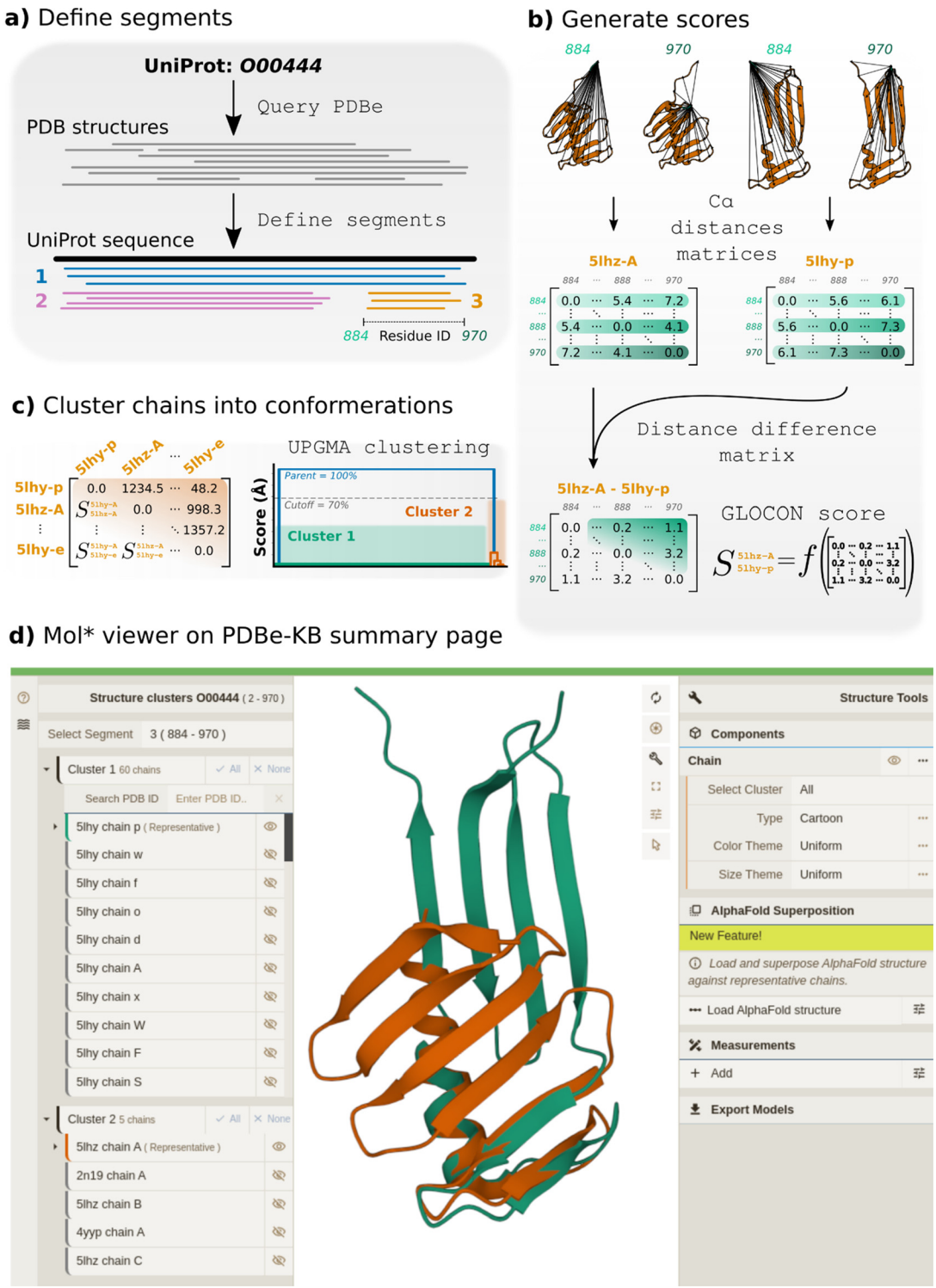

FIG. 2. Automated identification of protein conformational states across the PDB archive. (a) All chains of a given UniProt accession (100% sequence identity) are assigned to segments based on their overlap with the reference UniProt sequence. Non-overlapping sequences are grouped into separate segments. (b) Chains are superposed to all other chains within their assigned segment. (c) Chain–chain GLOCON scores are calculated for all polypeptides within a segment (refer to Ref. 89 for formal definition) before (d) agglomerative clustering is performed. The results are displayed in 3D on PDBe-KB aggregated views of proteins pages.

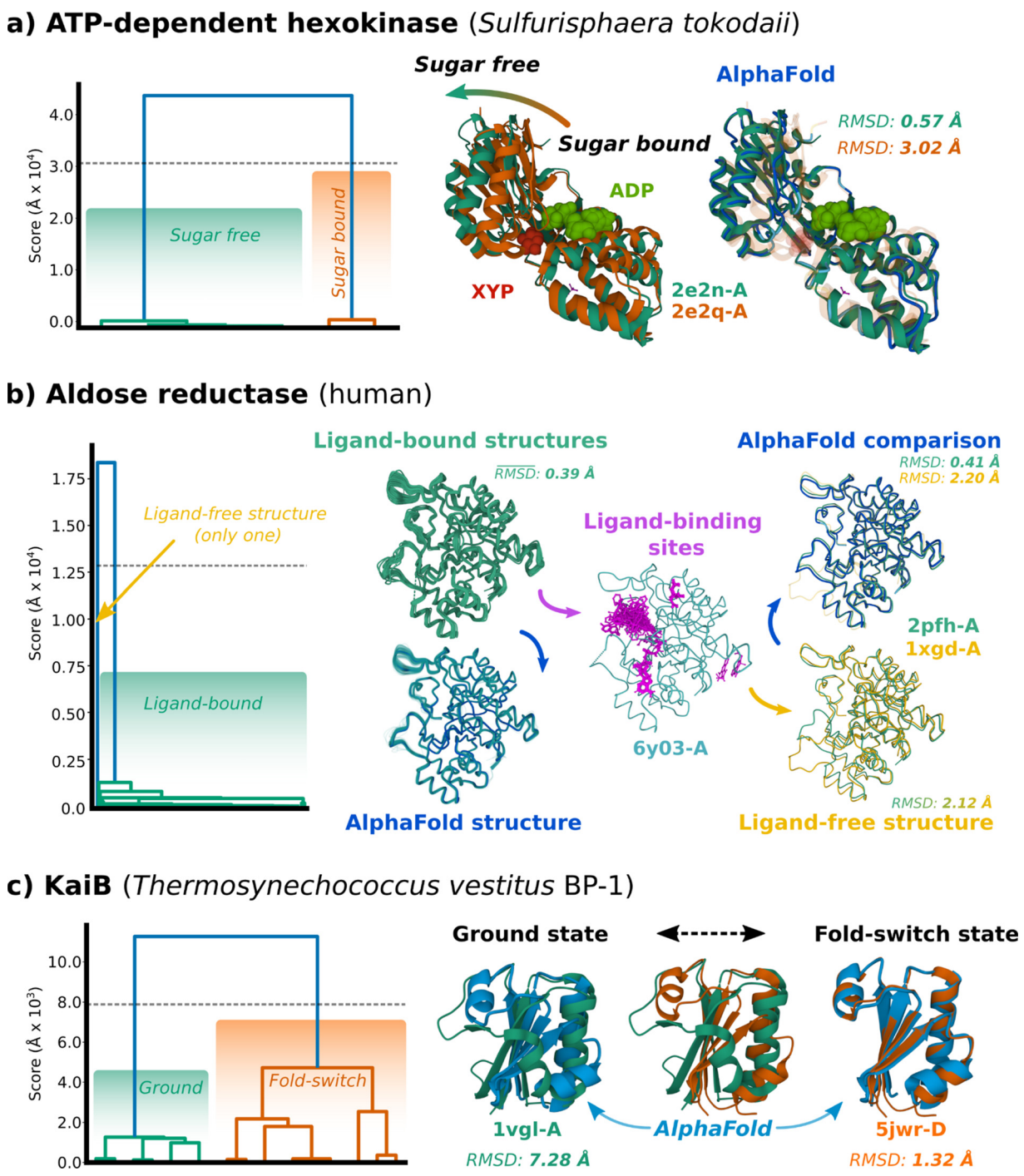

FIG. 3. Notable examples of predicted conformational states by the PDBe-KB. (a) Clustering results in dendrogram (left) and structures (right) of the open–closed conformation change made by UniProt: Q96Y14. XYP (red) denotes β-D-xylopyranose and ADP (light green) denotes adenosine triphosphate, both bound to 2E2Q chain A. 2E2N chain A is an apo-form of the polypeptide. RMSD calculated between AlphaFold2 model and experimental structures. (b) Substrate promiscuity illustrated by consistent binding of diverse ligands (magenta), despite the polypeptide (UniProt: P15121) adopting a consistent conformation. Mean RMSD displayed for the collection of ligand-bound structures (top left), the AlphaFold2 structure to the two representative chains (top right), and between representative chains (bottom right). Structural variation between ligand-bound structures is relatively low, with a standard deviation in RMSD of 0.16 Å. Ligand-free structure (yellow) has a displaced loop in the Pro211-Asp230 region. (c) Fold-switch protein (UniProt: Q79V61) transitioning to control day–night cycle. Clustering dendrogram (left) with AlphaFold2 structure superposed alongside experimentally determined models. RMSD calculated between AlphaFold2 model and experimental structures.

to sugar substrates,[82] allowing it to associate with glucose, mannose, glucosamine, xylose, and N-acetylglucosamine. Hexokinase adopts an open or a closed conformation, dependent on sugar binding, although ADP binding has a marginal effect on the protein's shape. Our automated pipeline can discern between the open and closed states, even identifying the open and closed chains solved within the asymmetric unit of 2E2Q [Fig. 3(a)].

Additionally, human aldose reductase (UniProt accession: P15121) accepts a diverse range of carbonyl-based substrates, reducing them to alcohol products using NADH as an electron source [Fig. 3(b)]. Many structures of this protein have been independently solved with a variety of ligands, providing information on the conformational heterogeneity within the holo state.[83] Individual PDB entries fail to capture the structural heterogeneity in the β-sheet region spanning Val121-Arg156, but our pipeline can separate the only non-liganded structure in the PDB (1XGD) from all other ligand-bound chains. Superposition of all chains highlights a structural deviation in the unliganded structure within the Pro211-Asp230 loop.

Finally, the circadian rhythm protein KaiB (UniProt: Q79V61) helps regulate the day–night cycle in cyanobacteria and has been previously characterized as a fold-switch protein[79] [Fig. 3(c)]. Associating with KaiA and KaiC, KaiB from *Thermosynechococcus vestitus* partakes in a concerted cycle of complex formation, autophosphorylation, and autodephosphorylation of KaiC, completing each oscillation every ~24 h.[68] KaiB adopts a homotetrameric ground state during the day and a thioredoxin-like "fold-switch" state at night. The fold-switch state is ordinarily stabilized upon oligomerization with KaiC and KaiB subunits, forming a multimeric complex.[68] The clustering method described here identifies the structures solved in these two states and highlights that the protein's AlphaFold2 model from the AFDB is closer in conformation to the night-dominant fold-switch state.

DISCUSSION

Exploring protein dynamics and conformational heterogeneity is essential for understanding molecular mechanisms and disease progression. However, capturing the full range of biologically relevant conformations—even for a single polypeptide—poses significant challenges beyond solving or predicting a static structure.[22,81] Numerous experimental and computational methods characterize macromolecular dynamics, but a lack of standardization hinders comprehensive data integration. The next generation of integrative methodologies promises to combine diverse experimental data and computational techniques to achieve accurate and meaningful representations of conformational heterogeneity.[22] Here, we have presented the method of clustering the static structures archived in the PDB. These clusters may depict some of the most stable, highly populated protein conformations (at 100% sequence identity) but cannot represent the complete free-energy topology nor the pathways traversed during conformational state transitions.

Nevertheless, even a high-accuracy representation of structural dynamics will be of limited value in answering biological questions unless contextualized with functional information. Attributing biological significance to conformational differences becomes much more challenging without annotations, such as ligand binding, oligomeric state, post-translational modifications, and point mutations, to name a few.[11–16] When comparing more distantly related proteins, ontological annotations and domain mappings from resources such as CATH[84] and SCOP[85] can help systematically explore sequence–structure–function relationships. Automated annotation methods, utilizing structural motifs, domain composition, and comparative modeling, will be useful for predicting functions of uncharacterized proteins and their distinct conformations. Tools such as DALI,[86] SSAP,[87] and Foldseek[88] are currently available for the identification of evolutionary relationships and functional similarities via structural comparison. As more conformations are determined experimentally—improving ensemble-model prediction algorithms—high-quality functional annotations necessitate integration to enable systematic analysis of structural diversity across different structure data archives. The PDBe-KB superposition and clustering pipeline presented here is a step toward this goal, but the collation of annotations is now needed before biological relevance can be systematically mapped to distinct protein conformations.

CONCLUSION

Here, we present a deterministic data pipeline that clusters all proteins in the PDB archive based on model coordinates, independently of superposition. We demonstrated that the process can automatically identify distinct conformations, which, due to lack of standardized labeling in the archive, would otherwise be non-trivial to find in the PDB.

However, the lack of systematic, high-quality conformational state annotations impedes our understanding of the biological implications of protein dynamics. As such, functional annotations become available, and high-throughput mapping to conformations could be driven by initiatives such as the PDBe-KB consortium that has laid the groundwork for creating unified data access mechanisms and standard data exchange formats for a broad range of functional annotations.

Unlocking the potential of protein dynamics involves a multifaceted approach to understand their roles in biological mechanisms. It demands application of the innovative multimodal approaches seen by integrative modeling, combined with continued infrastructure improvement for high-throughput annotation and data access. As the field advances, these efforts will help with the development of novel therapeutic strategies and help us realize the relationship between protein sequence, structure, and function.

ACKNOWLEDGMENTS

The authors would like to thank the UKRI-Biotechnology and Biological Sciences Research Council for providing funding under the FunCLAN (No. BB/V016113/1) project and the European Molecular Biology Laboratory-European Bioinformatics Institute for supporting development of the service.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Joseph I. J. Ellaway: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Stephen Anyango:** Resources (equal); Software (equal). **Sreenath Nair:** Resources (equal); Software (equal). **Hossam A. Zaki:** Conceptualization (equal); Methodology (supporting); Software (supporting); Writing – review & editing (supporting).

Nurul Nadzirin: Software (equal). **Harold R. Powell:** Methodology (equal); Software (equal); Writing – review & editing (supporting). **Aleksandras Gutmanas:** Conceptualization (equal); Software (equal); Writing – review & editing (equal). **Mihaly Varadi:** Conceptualization (equal); Funding acquisition (equal); Investigation (supporting); Methodology (supporting); Project administration (lead); Supervision (lead); Writing – original draft (equal); Writing – review & editing (lead). **Sameer Velankar:** Conceptualization (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1]wwPDB consortium, "Protein Data Bank: The single global archive for 3D macromolecular structure data," Nucl. Acids Res. **47**, D520–D528 (2019).
[2]J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," Nature **596**, 583–589 (2021).
[3]M. Baek *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," Science **373**, 871–876 (2021).
[4]Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science **379**, 1123–1130 (2023).
[5]M. Varadi *et al.*, "AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," Nucl. Acids Res. **50**, D439–D444 (2022).
[6]T. Schwede *et al.*, "Outcome of a workshop on applications of protein models in biomedical research," Structure **17**, 151–159 (2009).
[7]B. Jing, B. Berger, and T. Jaakkola, "AlphaFold meets flow matching for generating protein ensembles," preprint arXiv:2402.04845 (2024).
[8]K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins," Nature **450**, 964–972 (2007).
[9]L. Xue *et al.*, "Visualizing translation dynamics at atomic detail inside a bacterial cell," Nature **610**, 205–211 (2022).
[10]C. Weng, A. J. Faure, A. Escobedo, and B. Lehner, "The energetic and allosteric landscape for KRAS inhibition," Nature **626**, 643–652 (2024).
[11]G. Pozzati *et al.*, "Limits and potential of combined folding and docking," Bioinformatics **38**, 954–961 (2022).
[12]M. H. Høie, M. Cagiada, A. H. B. Frederiksen, A. Stein, and K. Lindorff-Larsen, "Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation," Cell Rep. **38**, 110207 (2022).
[13]T.-M. Fu *et al.*, "Cryo-EM structure of caspase-8 tandem DED filament reveals assembly and regulation mechanisms of the death-inducing signaling complex," Mol. Cell **64**, 236–250 (2016).
[14]D. Rimmerman *et al.*, "Revealing fast structural dynamics in pH-responsive peptides with time-resolved x-ray scattering," J. Phys. Chem. B **123**, 2016–2021 (2019).
[15]N. Zimmermann, A. Noga, J. M. Obbineni, and T. Ishikawa, "ATP-induced conformational change of axonemal outer dynein arms revealed by cryo-electron tomography," EMBO J. **42**, e112466 (2023).
[16]D. Adamoski *et al.*, "Molecular mechanism of glutaminase activation through filamentation and the role of filaments in mitophagy protection," Nat. Struct. Mol. Biol. **30**, 1902–1912 (2023).
[17]A. Abyzov, M. Blackledge, and M. Zweckstetter, "Conformational dynamics of intrinsically disordered proteins regulate biomolecular condensate chemistry," Chem. Rev. **122**, 6719–6748 (2022).
[18]F. E. Thomasen and K. Lindorff-Larsen, "Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins," Biochem. Soc. Trans. **50**, 541–554 (2022).
[19]S. Qin and H.-X. Zhou, "Effects of macromolecular crowding on the conformational ensembles of disordered proteins," J. Phys. Chem. Lett. **4**, 3429–3434 (2013).
[20]G. F. Schröder, "Hybrid methods for macromolecular structure determination: Experiment with expectations," Curr. Opin. Struct. Biol. **31**, 20–27 (2015).
[21]H. van den Bedem and J. S. Fraser, "Integrative, dynamic structural biology at atomic resolution—It's about time," Nat. Methods **12**, 307–318 (2015).
[22]R. Grandori, "Protein structure and dynamics in the era of integrative structural biology," Front. Biophys. **1**, 1219843 (2023).
[23]A. M. Wolff *et al.*, "Mapping protein dynamics at high spatial resolution with temperature-jump X-ray crystallography," Nat. Chem. **15**, 1549–1558 (2023).
[24]S. Du *et al.*, "Refinement of multiconformer ensemble models from multi-temperature X-ray diffraction data," Methods Enzymol. **688**, 223–254 (2023).
[25]P. Nogly *et al.*, "Retinal isomerization in bacteriorhodopsin captured by a femtosecond X-ray laser," Science **361**, eaat0094 (2018).
[26]N. Coquelle *et al.*, "Chromophore twisting in the excited state of a photoswitchable fluorescent protein captured by time-resolved serial femtosecond crystallography," Nat. Chem. **10**, 31–37 (2018).
[27]K. Oda *et al.*, "Time-resolved serial femtosecond crystallography reveals early structural changes in channelrhodopsin," eLife **10**, e62389 (2021).
[28]R. P. Rambo and J. A. Tainer, "Accurate assessment of mass, models and resolution by small-angle scattering," Nature **496**, 477–481 (2013).
[29]H. S. Cho *et al.*, "Dynamics of quaternary structure transitions in R-state carbonmonoxyhemoglobin unveiled in time-resolved X-ray scattering patterns following a temperature jump," J. Phys. Chem. B **122**, 11488–11496 (2018).
[30]I. Josts *et al.*, "Photocage-initiated time-resolved solution X-ray scattering investigation of protein dimerization," IUCrJ **5**, 667–672 (2018).
[31]F. Caporaletti *et al.*, "Small-angle x-ray and neutron scattering of MexR and its complex with DNA supports a conformational selection binding model," Biophys. J. **122**, 408–418 (2023).
[32]T. Narayanan *et al.*, "A multipurpose instrument for time-resolved ultra-small-angle and coherent X-ray scattering," J. Appl. Crystallogr. **51**, 1511–1524 (2018).
[33]J. Y. Kang *et al.*, "Structural basis for transcript elongation control by NusG family universal regulators," Cell **173**, 1650–1662.e14 (2018).
[34]E. Nwanochie and V. N. Uversky, "Structure determination by single-particle cryo-electron microscopy: Only the sky (and intrinsic disorder) is the limit," Int. J. Mol. Sci. **20**(17), 4186 (2019).
[35]A. Punjani, H. Zhang, and D. J. Fleet, "Non-uniform refinement: Adaptive regularization improves single-particle cryo-EM reconstruction," Nat. Methods **17**, 1214–1221 (2020).
[36]H. Gupta, M. T. McCann, L. Donati, and M. Unser, "CryoGAN: A new reconstruction paradigm for single-particle cryo-EM via deep adversarial learning," IEEE Trans. Comput. Imaging **7**, 759–774 (2021).
[37]E. D. Zhong, A. Lerer, J. H. Davis, and B. Berger, "CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), pp. 4046–4055.
[38]W. S. Tang, E. D. Zhong, S. M. Hanson, E. H. Thiede, and P. Cossio, "Conformational heterogeneity and probability distributions from single-particle cryo-electron microscopy," Curr. Opin. Struct. Biol. **81**, 102626 (2023).
[39]L. F. Kinman, B. M. Powell, E. D. Zhong, B. Berger, and J. H. Davis, "Uncovering structural ensembles from single-particle cryo-EM data using cryoDRGN," Nat. Protoc. **18**, 319–339 (2023).
[40]M. Chen and S. J. Ludtke, "Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM," Nat. Methods **18**, 930–936 (2021).
[41]M. Chen, B. Toader, and R. Lederman, "Integrating molecular models into cryoEM heterogeneity analysis using scalable high-resolution deep gaussian mixture models," J. Mol. Biol. **435**, 168014 (2023).
[42]R. Rangan *et al.*, "Deep reconstructing generative networks for visualizing dynamic biomolecules inside cells," preprint arXiv:18.553799 (2023).
[43]H. Zhang *et al.*, "A method for restoring signals and revealing individual macromolecule states in cryo-ET, REST," Nat. Commun. **14**, 2937 (2023).
[44]X. Zeng *et al.*, "High-throughput cryo-ET structural pattern mining by unsupervised deep iterative subtomogram clustering," Proc. Natl. Acad. Sci. U. S. A. **120**, e2213149120 (2023).

[45]W. J. H. Hagen, W. Wan, and J. A. G. Briggs, "Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging," J. Struct. Biol. **197**, 191–198 (2017).
[46]S. Khavnekar *et al.*, "Multishot tomography for high-resolution in situ subtomogram averaging," J. Struct. Biol. **215**, 107911 (2023).
[47]I. de Teresa-Trueba *et al.*, "Convolutional networks for supervised mining of molecular patterns within cellular context," Nat. Methods **20**, 284–294 (2023).
[48]T. A. Ramelot, R. Tejero, and G. T. Montelione, "Representing structures of the multiple conformational states of proteins," Curr. Opin. Struct. Biol. **83**, 102703 (2023).
[49]P. Wapeesittipan, A. S. J. S. Mey, M. D. Walkinshaw, and J. Michel, "Allosteric effects in cyclophilin mutants may be explained by changes in nano-microsecond time scale motions," Commun. Chem. **2**, 41 (2019).
[50]N. Karschin, S. Becker, and C. Griesinger, "Interdomain dynamics via paramagnetic NMR on the highly flexible complex calmodulin/Munc13-1," J. Am. Chem. Soc. **144**, 17041–17053 (2022).
[51]J. H. Overbeck, D. Stelzig, A.-L. Fuchs, J. P. Wurm, and R. Sprangers, "Observation of conformational changes that underlie the catalytic cycle of Xrn2," Nat. Chem. Biol. **18**, 1152–1160 (2022).
[52]J. B. Stiller *et al.*, "Structure determination of high-energy states in a dynamic protein ensemble," Nature **603**, 528–535 (2022).
[53]M. R. Jensen, M. Zweckstetter, J. Huang, and M. Blackledge, "Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy," Chem. Rev. **114**, 6632–6660 (2014).
[54]A. R. Camacho-Zarco *et al.*, "NMR provides unique insight into the functional dynamics and interactions of intrinsically disordered proteins," Chem. Rev. **122**, 9331–9356 (2022).
[55]G. Ahdritz *et al.*, "OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization," preprint arXiv:20.517210 (2022).
[56]D. Chakravarty, J. W. Schafer, E. A. Chen, J. R. Thole, and L. L. Porter, "AlphaFold2 has more to learn about protein energy landscapes," preprint arXiv:12.571380 (2023).
[57]H.-B. Guo *et al.*, "AlphaFold2 models indicate that protein sequence determines both structure and dynamics," Sci. Rep. **12**, 10696 (2022).
[58]T. J. Lane, "Protein structure prediction has reached the single-structure frontier," Nat. Methods **20**, 170–173 (2023).
[59]D. Sala, F. Engelberger, H. S. Mchaourab, and J. Meiler, "Modeling conformational states of proteins with AlphaFold," Curr. Opin. Struct. Biol. **81**, 102645 (2023).
[60]R. A. Stein and H. S. Mchaourab, "SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2," PLOS Comput. Biol. **18**, e1010483 (2022).
[61]D. del Alamo, D. Sala, H. S. Mchaourab, and J. Meiler, "Sampling alternative conformational states of transporters and receptors with AlphaFold2," eLife **11**, e75751 (2022).
[62]Y. J. Huang *et al.*, "Assessment of prediction methods for protein structures determined by NMR in CASP14: Impact of AlphaFold2," Proteins Struct. Funct. Bioinform. **89**, 1959–1976 (2021).
[63]L. Heo and M. Feig, "Multi-state modeling of G-protein coupled receptors at experimental accuracy," Proteins Struct. Funct. Bioinform. **90**, 1873–1885 (2022).
[64]T. Saldaño *et al.*, "Impact of protein conformational diversity on AlphaFold predictions," Bioinformatics **38**, 2742–2748 (2022).
[65]H. K. Wayment-Steele, S. Ovchinnikov, L. Colwell, and D. Kern, "Prediction of multiple conformational states by combining sequence clustering with AlphaFold2," preprint arXiv:17.512570 (2022).
[66]B. Wallner, "AFsample: Improving multimer prediction with AlphaFold using massive sampling," Bioinformatics **39**, btad573 (2023).
[67]T. K. Karamanos, "Chasing long-range evolutionary couplings in the AlphaFold era," Biopolymers **114**, e23530 (2023).
[68]R. Tseng *et al.*, "Structural basis of the day-night transition in a bacterial circadian clock," Science **355**, 1174–1180 (2017).
[69]A. Banerjee, S. Saha, N. C. Tvedt, L.-W. Yang, and I. Bahar, "Mutually beneficial confluence of structure-based modeling of protein dynamics and machine learning methods," Curr. Opin. Struct. Biol. **78**, 102517 (2023).
[70]A. Gupta, S. Dey, A. Hicks, and H.-X. Zhou, "Artificial intelligence guided conformational mining of intrinsically disordered proteins," Commun. Biol. **5**, 610 (2022).
[71]G. Janson, G. Valdes-Garcia, L. Heo, and M. Feig, "Direct generation of protein conformational ensembles via machine learning," Nat. Commun. **14**, 774 (2023).
[72]C. W. Park *et al.*, "Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture," Npj Comput. Mater. **7**, 73 (2021).
[73]J. Tian *et al.*, "Revealing the conformational dynamics of UDP-GlcNAc recognition by O-GlcNAc transferase via Markov state model," Int. J. Biol. Macromol. **256**, 128405 (2024).
[74]D. Wang *et al.*, "Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics," Nat. Comput. Sci. **2**, 20–29 (2021).
[75]M. Varadi *et al.*, "PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education," Protein Sci. **31**, e4439 (2022).
[76]J. M. Dana *et al.*, "SIFTS: Updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins," Nucl. Acids Res. **47**, D482–D489 (2019).
[77]S. Velankar *et al.*, "SIFTS: Structure integration with function, taxonomy and sequences resource," Nucl. Acids Res. **41**, D483–D489 (2012).
[78]E. Krissinel, "Enhanced fold recognition using efficient short fragment clustering," J. Mol. Biochem. **1**, 76–85 (2012).
[79]L. L. Porter and L. L. Looger, "Extant fold-switching proteins are widespread," Proc. Natl. Acad. Sci. **115**, 5968–5973 (2018).
[80]P. V. Burra, Y. Zhang, A. Godzik, and B. Stec, "Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure," Proc. Natl. Acad. Sci. U. S. A. **106**, 10505–10510 (2009).
[81]M. D. Miller and G. N. Phillips, "Moving beyond static snapshots: Protein dynamics and the Protein Data Bank," J. Biol. Chem. **296**, 100749 (2021).
[82]H. Nishimasu, S. Fushinobu, H. Shoun, and T. Wakagi, "Crystal structures of an ATP-dependent hexokinase with broad substrate specificity from the hyperthermophilic archaeon *Sulfolobus tokodaii*," J. Biol. Chem. **282**, 9923–9931 (2007).
[83]A. Sandner *et al.*, "Which properties allow ligands to open and bind to the transient binding pocket of human aldose reductase?," Biomolecules **11**, 1837 (2021).
[84]I. Sillitoe *et al.*, "CATH: Increased structural coverage of functional space," Nucl. Acids Res. **49**, D266–D273 (2021).
[85]A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin, "The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures," Nucl. Acids Res. **48**, D376–D382 (2020).
[86]L. Holm, A. Laiho, P. Törönen, and M. Salgado, "DALI shines a light on remote homologs: One hundred discoveries," Protein Sci. **32**, e4519 (2023).
[87]C. A. Orengo and W. R. Taylor, "SSAP: Sequential structure alignment program for protein structure comparison," in *Methods in Enzymology* (Academic Press, 1996), Vol. 266, pp. 617–635.
[88]M. van Kempen *et al.*, "Fast and accurate protein structure search with Foldseek," Nat. Biotechnol. **42**, 243–246 (2024).
[89]See the supplementary material for details. We include a copy of our manually curated benchmark dataset of 315 proteins across a range of conformational states and a supplementary methods document, formally describing the algorithm.