Dual-wield NTPases: a novel protein family mined from AlphaFold DB

Koya Sakuma[1], Ryotaro Koike[1] & Motonori Ota[1,2]

1. **Graduate School of Informatics, Nagoya University**
2. **Institute for Glyco-core Research, Nagoya University**

Email: K.S. (ksakuma@i.nagoya-u.ac.jp), M.O. (mota@i.nagoya-u.ac.jp)

AlphaFold protein structure database (AlphaFold DB) archives a vast number of predicted models. We conducted systematic data mining against AlphaFold DB and discovered an uncharacterized P-loop NTPase family whose structure was surprisingly novel, showing an atypical topology for P-loop NTPases, noticeable two-fold symmetry and two pairs of independent putative active sites. Our findings show that structural data mining is a powerful approach to identifying undiscovered protein families.

Characterizing protein structures is essential for understanding their function, and structures are typically solved by experimental approaches and deposited in the Protein Data Bank[1] (PDB). More recently, public databases produced by state-of-the-art structure prediction, such as the AlphaFold protein structure database (AlphaFold DB) and ESM metagenomic Atlas (ESM Atlas), are changing this situation[2,3]. These databases are approximately three orders of magnitude larger than the PDB and contain numerous experimentally unsolved protein structures, providing opportunities for finding novel proteins based only on the structural information *in-silico*[4].

Dedicated data mining requires a clearly stated working hypothesis. Therefore, we set a simple database search question: are there monomeric proteins that contain multiple phosphate-binding loops (P-loops) on a single continuous β-sheet? The P-loop or Walker-A motif is a local functional motif that recognizes phosphate groups and is shared among P-loop NTPases, such as ATPases, GTPases and nucleotide kinases (NKs)[5–7]. In general, one P-loop resides on a single continuous β-sheet of a three-layered α/β/α sandwich architecture. Our preliminary search against the PDB supported this observation because no experimentally determined structure has multiple P-loops in a β-sheet. However, the possibility that a single β-sheet possesses multiple P-loops should not be excluded. We hypothesized that such experimentally unobserved multiple-P-loop structures exist in AlphaFold DB and can be discovered via systematic data mining.

By computationally scanning more than 0.2 billion entries in AlphaFold DB version 4, we extracted 15,977 single-chained structures possessing multiple P-loops[2,8]. We then analyzed the hydrogen-bond network and extracted 839 structures with multiple P-loops on a single continuous β-sheet. The structures were grouped into 11 clusters based on structural similarity[9]. As a result, we found an uncharacterized family of P-loop proteins, dual-wield P-loop NTPase (dwNTPase), as the largest cluster with 711 members. All structural models in this cluster were predicted with high confidence scores, i.e., the average predicted Local Distance Difference Test (pLDDT) was 94.27, indicating that the predictions were reliable[10] (Extended Data Fig. 1).

The overall architecture of dwNTPases was highly novel and showed noticeable two-fold symmetry. Fig. 1a shows the structure of a representative dwNTPase from *Bacillus*

thuringiensis (Bt. UniProt accession No.: A0A1Y0TWD8). Two P-loop domains are tightly packed and surrounded by two long bridging α-helices and two framing α-helices. Each of the domains comprises six β-strands, and twelve strands form a continuous single β-sheet, which reveals a previously unobserved single-sheeted dual-P-loop architecture (Fig. 1b). Two β-hairpins from each domain form a pier-like four-stranded β-sheet to connect two domains. Two canonical P-loops independently form two putative ligand binding sites that penetrate through the molecule and resemble tunnels rather than pockets (Fig. 1c). A search against the PDB clarified that no similar structures have been reported[9,11]. Similarly, a SwissProt subset of AlphaFold DB contained no similar structures, indicating that the dwNTPase family has no reliable annotations manually verified by UniProt curators.

We found that the P-loop domain of dwNTPases was structurally atypical for a P-loop NTPase by searching against the PDB[9,11] (Fig. 2a). A crystal structure of mutual gliding-motility protein MglAa from *Myxococcus xanthus* (PDB ID: 6h35), a bacterial small and monomeric GTPase, was the only known P-loop NTPase that showed relevant structural similarity to the dwNTPase P-loop domain[12]. The P-loop domain of dwNTPase has an additional β-strand at the N-terminus (strand 0) compared to the MglAa structure (Fig. 2b). Two strands constituting the pier sheet and a successive α-helix are also appended. In contrast, the domain lacks two C-terminal β-strands (strands 6 and 7) and some other surrounding secondary structural elements (SSEs). These unique arrangements of SSEs give rise to the atypical topology that does not resemble other P-loop NTPases (Extended Data Figs. 2 and 3). Furthermore, the P-loop domain has a long loop rather than a helix conserved in other P-loop NTPases (Extended Data Fig. 4), which we named the switch loop (Fig. 2a). These atypical features of the P-loop domain make it difficult to assign dwNTPase to known classes of P-loop NTPases.

Despite these novel features of dwNTPase, an iterative structure search by Foldseek against the entire AlphaFold DB revealed that 2,219 similar structures were deposited, most of which originated from bacteria in various Firmicutes[2,9] (Extended Data Fig. 5 and Table 1). Similar searches against ESM Atlas culled by 30% sequence identity found 748 similar structures[3,9] (Extended Data Table 2). We classified dwNTPase structures into six sub-classes based on the conservation of motifs and domains (Extended Data Fig. 6). The bona fide dwNTPase structure with two P-loops intact (class 1) was the most abundant, suggesting functional constraints exist to conserve the two active P-loops. A BLAST search against the non-redundant database revealed that dwNTPase had been classified as the PRK06851 family protein in the NCBI conserved domain database[13]. Thus, we concluded that dwNTPases constitute a conserved protein family among bacteria.

The molecular functions of dwNTPases were investigated by analyzing conserved residues (Extended Data Fig. 7). The most symmetric class of dwNTPases (class 1) possesses two clusters of conserved residues shared between both halves (Fig. 2c). We found Cys66/Cys248 (residue numbers follow Bt. dwNTPase), Asp74/Asp256, Asp87/Asp269 and His92/His274 formed putative metal binding sites. Molecular dynamics simulations of the Bt. dwNTPase structure complexed with two ATPs, two Mg^{2+} ions and two Zn^{2+} ions showed that the Zn^{2+} ions were stably coordinated by two aspartates and the γ-phosphate group of ATPs (Extended Data Fig. 8a,b), which resembles the active site structure of metal-dependent nucleotidyl-transfer enzymes[14] (Extended Data Fig. 8c). The side chains of Cys66/Cys248 and His92/His274 remained unoccupied (Extended Data Fig. 8d), suggesting that they may have roles other than metal-binding. As the pair of cysteine and histidine

residues are reminiscent of the catalytic triad/dyad in cysteine proteases (Extended Data Fig. 8e), we hypothesize that dwNTPases have additional hydrolase/ligase activity[15]. Since the biological functions of dwNTPases remain elusive, structural and biochemical studies are required and should provide greater insight into the biological significance of the dwNTPase family. In summary, we demonstrated that structural data mining is a powerful approach to exploring dark proteomes[16], which will help and encourage the design of experimental studies.

References

1. Burley, S. K. *et al. Biomolecules* **12**, 1425 (2022).
2. Varadi, M. *et al. Nucleic Acids Res.* **50**, D439–D444 (2022).
3. Lin, Z. *et al. bioRxiv* (2022) doi:10.1101/2022.07.20.500902.
4. Bordin, N. *et al. Commun. Biol.* **6**, 160 (2023).
5. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. *EMBO J.* **1**, 945–951 (1982).
6. Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. *J. Mol. Biol.* **317**, 41–72 (2002).
7. Leipe, D. D., Koonin, E. V. & Aravind, L. *J. Mol. Biol.* **333**, 781–815 (2003).
8. Kim, H., Mirdita, M. & Steinegger, M. *bioRxiv* (2022) doi:10.1101/2022.12.09.519715.
9. van Kempen, M. *et al. bioRxiv* (2022) doi:10.1101/2022.02.07.479398.
10. Jumper, J. *et al. Nature* **596**, 583–589 (2021).
11. Minami, S., Sawada, K., Ota, M. & Chikenji, G. *Bioinformatics* **34**, 3324–3331 (2018).
12. Galicia, C. *et al. Nat. Commun.* **10**, 5300 (2019).
13. Wang, J. *et al. Nucleic Acids Res.* **51**, D384–D388 (2023).
14. Yang, W. *Nat. Struct. Mol. Biol.* **15**, 1228–1231 (2008).
15. Dodson, G. & Wlodawer, A. *Trends Biochem. Sci.* **23**, 347–352 (1998).
16. Perdigão, N. *et al. Proc. Natl. Acad. Sci.* **112**, 15898–15903 (2015).

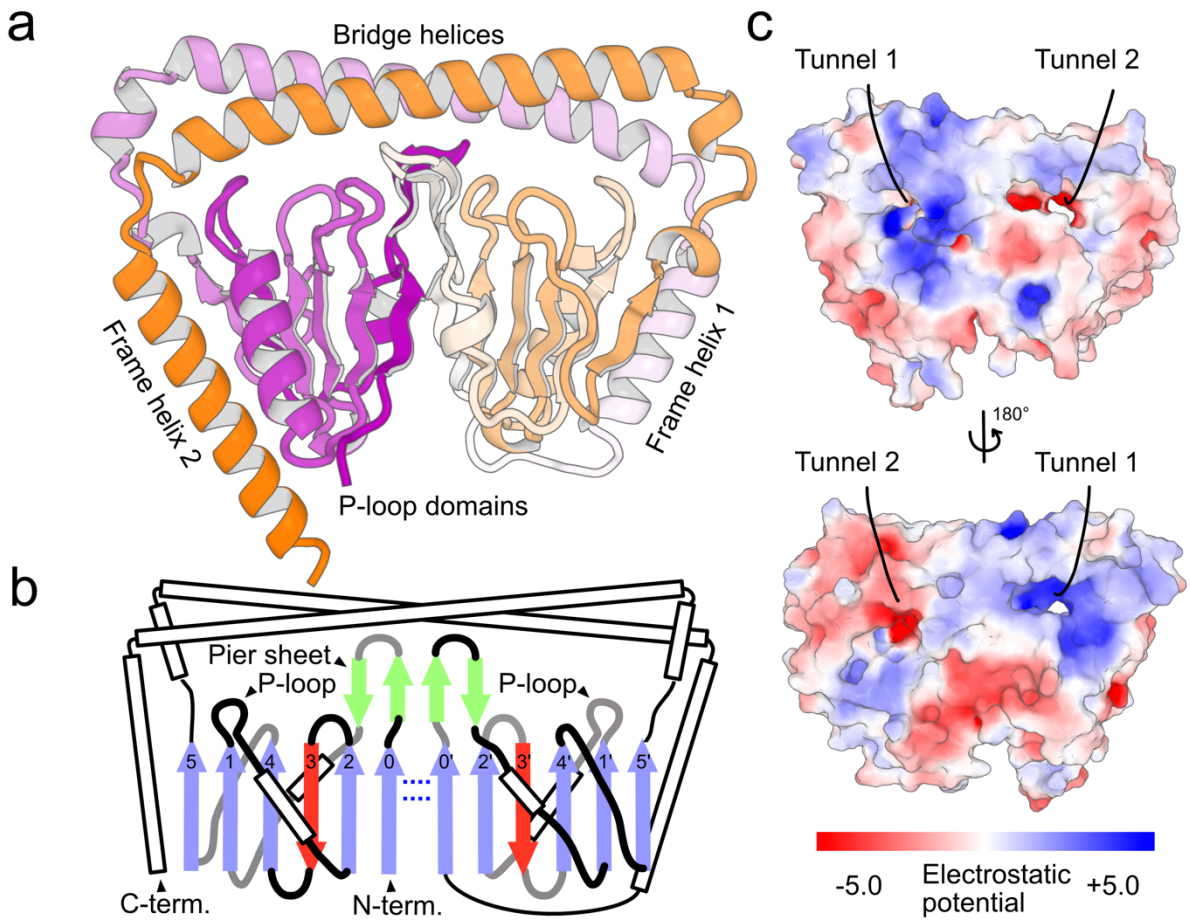


Fig. 1 | AlphaFold2 predicted model of a dual-wield NTPase structure (AF-A0A1Y0TWD8-F1-model_v4). a, Overall dwNTPase structure colored according to a purple-white-orange gradient from the N- to C-terminus. **b**, Topology diagram of dwNTPase. Blue and red arrows represent β-strands pointing up and down that form the large β-sheets in the P-loop domains. Green arrows represent the pier sheet. White rectangles are α-helices. Grey and black lines indicate junctions projecting behind and out of the β-sheets, respectively. Blue dotted lines represent hydrogen bonds connecting the two halves of the large β-sheet. **c**, The location and shape of the ligand binding tunnels. Color bar is at the bottom.

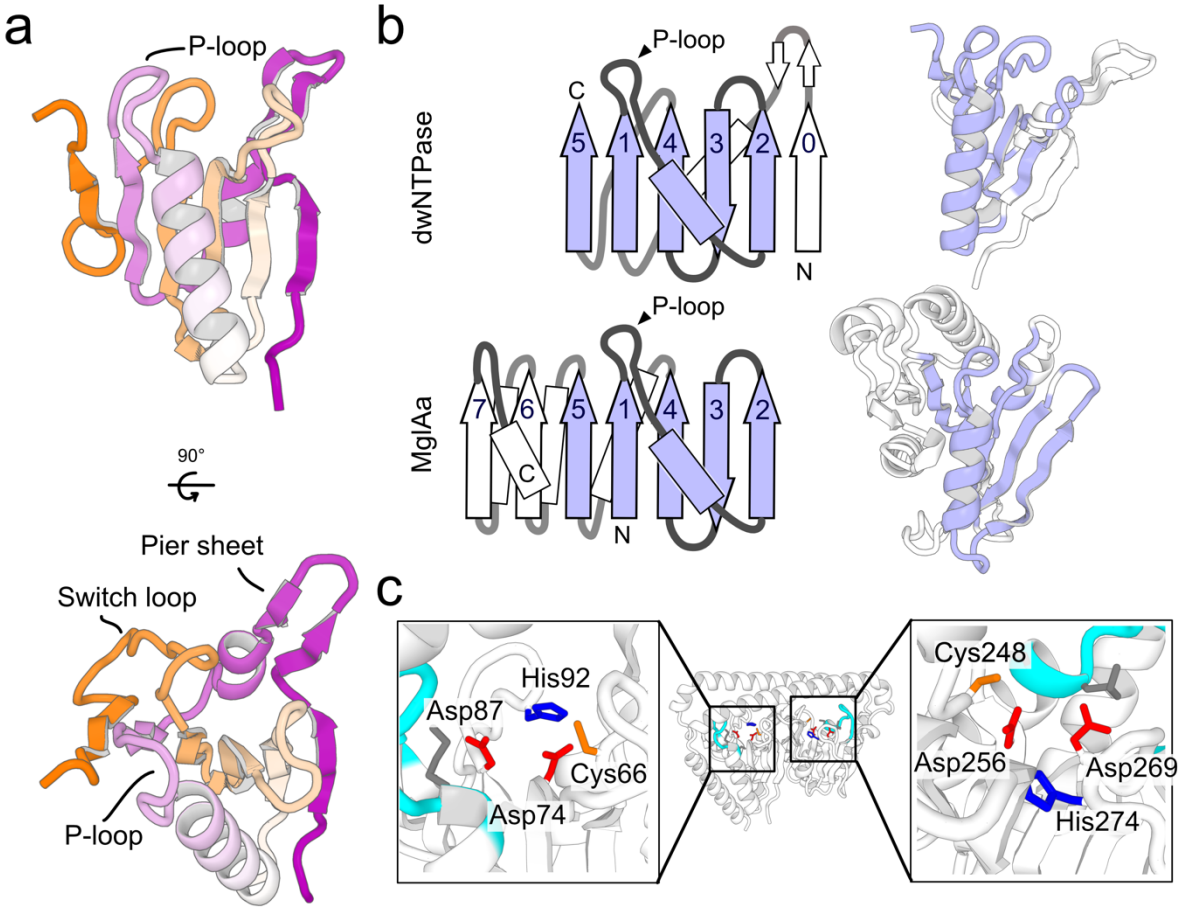


Fig. 2 | P-loop domain and its conserved residues. a, Front and top views of the dwNTPase P-loop domain colored according to a purple-white-orange gradient from the N- to C-terminus. P-loop, switch loop and pier-sheet are indicated by labels. **b**, Topology diagrams and cartoon representations of dwNTPase P-loop domains and MglAa structure. Arrows and rectangles represent β-strands and α-helices. Secondary structures that align between two structures are colored blue. **c**, Conserved residues in the putative ligand binding tunnels. Histidine, cysteine and aspartate are colored blue, orange and red, respectively. P-loops and their conserved residues are colored cyan and grey.

Methods

Identification of structures containing P-loop-like fragments

AlphaFold DB (v4 UniProt) was downloaded from the Foldcomp database[2,8]. P-loop NTPase protein structures were extracted by converting the models into the sequences of ABEGO, where A, B, E and G, respectively, denote backbone dihedral angles (phi, psi) for α, β, left-handed β and left-handed α on the Ramachandran plot[17]. O denotes other conformations unassignable on the Ramachandran plot, typically a cis-peptide conformation. Typical P-loop (Walker-A) motifs have conformations represented by EBBGAG or BBBGAG, both of which can be seen in the crystal structure of α and β subunits of bovine mitochondrial F1-ATPase[18] (chain A and chain D of PDB ID: 1BMF). Because the P-loop is a junction between a β-strand and an α-helix, we extended the ABEGO motifs to "BBBEBBGAGAAAAA" or "BBBBBBGAGAAAAA" and extracted all the structures containing any of them by sequence pattern matching. We then calculated the Cα root-mean-square deviations (RMSDs) of the matched substructures against the reference P-loop fragment (residues 166–179 of 1BMF, chain A) and filtered out substructures with RMSDs larger than 2.0 Å. We obtained 15,977 proteins containing multiple P-loop-like fragments and built a custom Foldcomp database for subsequent procedures.

Identification of dual-wield NTPases

Visual inspection revealed that most structures with multiple P-loop-like fragments within a single chain were tandem repeats of known P-loop NTPase domains connected by flexible linkers. Such proteins were excluded by analyzing structures using an in-house program that detects the β-sheet hydrogen-bond network and extracting structures possessing two P-loop-like fragments on a single continuous β-sheet[19]. We obtained 839 structures possessing two P-loop-like fragments on a single β-sheet. These structures were clustered by TM-score calculations with Foldseek into 11 clusters[9]. The largest cluster contained 711 members, which corresponded to dual-wield NTPases. For these structures, we performed all-against-all structure alignment using MICAN and defined the structure with the largest average TM-score as the representative[11] (AF-A0A1Y0TWD8-F1-model_v4).

Extraction of structures similar to dwNTPase from AlphaFold DB and ESM Atlas

We performed iterative structure searches using FoldSeek to enumerate as many structures that resemble dwNTPase as possible[9]. In the first stage, we performed a structure search against AlphaFold DB using all 711 structures initially mined from AlphaFold DB as queries[2]. After removing overlapping structures, we obtained 1377 structures. Using these structures as seeds, we again performed a Foldseek search and obtained 135 new non-overlapping structures. The third iteration of Foldseek search yielded some non-specific hits. Therefore, we stopped this iteration, manually selected similar structures and discarded the remaining structures. Consequently, we obtained 2219 dwNTPase structures from AlphaFold DB. Similarly, we performed structural searches against the highquality_clust30 subset of ESM-atlas using 711 structures found in AlphaFold DB as queries and obtained 748 structures with a TM-score larger than 0.5[3,20].

Whole structure search against the PDB and Swiss-Prot subset of AlphaFold DB

To assess the novelty of the dwNTPase structure and gain insights into the function, we performed structural searches against PDB100 and the Swiss-Prot subset of AlphaFold DB (version 4) using the Foldseek server in the TM-align mode and the representative structure as the query[9]. No relevant (TM-score ≧ 0.5) hit was found among these databases. We used MICAN to perform rigorous one-against-all searches without pre-filtering; however, no similar (TM-score ≧ 0.5) structures were found among the PDB (2023-01-09) and the Swiss-Prot subset of AlphaFold DB (version 2)[1,2,11].

Domain structure search against the PDB, Swiss-Prot subset of AlphaFold DB and SCOPe

We searched structures similar to the P-loop domain of the representative structure (residues 1–110) against the PDB100 and the Swiss-Prot subset of AlphaFold DB by using the Foldseek server[9]. No relevant hit was found. We used MICAN to perform a rigorous structure search without pre-filtering against the PDB (2023-09-01) and the Swiss-Prot subset of AlphaFold DB (version 2)[1,2,11]. We obtained 358 and 2931 relevant hits (TM-score ≧ 0.5) from the PDB and Swiss-Prot. We performed clustering by mmseq2 with sequence identity set at 35% and obtained 15 and 137 clusters[21]. The alignments were checked by visual inspection of all cluster representatives. We found that some structures showed similar topology to the P-loop domain of dwNTPase: 6h35, Q1DB04 and Q9UBK7 from the PDB and Swiss-Prot, which are annotated as GTPase or GTP-binding proteins[12]. The remaining hits showed RecA-like topology and were not topologically identical to dwNTPase because the RecA-like topology has an all-parallel β-sheet, whereas dwNTPases have anti-parallel-containing β-sheets[7]. Similarly, we performed structural comparisons against domain structures classified as G-proteins (SCOP concise classification string (SCCS): c.37.8), NKs (c.37.1) and RecA-like proteins (c.37.11) in the SCOPe version 2.08 using MICAN[11,22]. The groups of G-proteins, NKs and RecA-like proteins contained 255, 212 and 118 parsed domain structures, respectively, and we selected the structures that showed the highest TM-score in the group for visualization (Extended Data Fig. 2).

Calculation of sequence identities between two halves of dwNTPase structures

We selected 1,903 structures with more than 340 residues from the set of dwNTPases extracted from AlphaFold DB. A structure was self-aligned by MICAN in the rewiring mode, which ignores the sequential order of SSEs. The sequence identity was calculated based on the second-best alignment by MICAN[11].

Identification of putative catalytic residues and a side chain pattern search against the PDB

The potential function of dwNTPases was examined by performing a sequence search and alignment to identify conserved residues by HHblits against UniRef30_2022_02[23,24]. After three iterations, 2687 sequences were extracted from the database. To exclude fragmented sequences most likely originating from partial matches to the P-loop consensus motif, we removed aligned sequences with more than ten gaps against the representative sequence and obtained a Multiple Sequence Alignment (MSA) with 138 sequences. From this MSA, the site-wise entropy of the alignments was calculated to identify conserved residues, and the top 10 residues around the two tunnels were listed. We defined tunnel 1 as residues 61–100 and tunnel 2 as residues 243–282. From tunnel 1, residues 62, 66, 67, 69, 73, 74, 75, 87, 88 and 100 were identified. From tunnel 2, residues 244, 246, 247, 248, 252, 255, 256, 261, 263 and 274 were identified. According to the orientation of side chains toward the tunnel, we

selected Cys66, Ser67, Asp74 and Asp87 as candidates for probable functional residues in tunnel 1. Similarly, Cys248, Asp256 and His274 were selected for tunnel 2. Considering the symmetry of the dwNTPase structure, Cys66/Cys248, Asp74/Asp256, Asp87/Asp269 and His92/His274 were considered clusters of functional residues in tunnels 1 and 2. We performed a side chain pattern search against the PDB using the strucmotif-search program to determine whether protein structures possessed similar side chain configurations[25]. The set of residues Cys66, Asp74, Asp87 and His92 in the representative structure were selected as queries, and a search was performed against all structures in the PDB (2022-28-12), with the threshold for the structure similarity set to 1.0 Å. The side chain pattern search gave no hits and indicated that the putative catalytic residues have a novel configuration of conserved residues.

Docking of ATP, Mg and Zn

We transplanted ligand structures from existing PDB structures to model the complex structures. The P-loop region of an ATPase crystal structure (PDB ID: 6j18) was superposed to the P-loop of the representative structure by MICAN in PyMOL[11,26,27], and the ATP and Mg^{2+} models were extracted. Similarly, His125 from a zinc finger motif (PDB ID: 2hgh) was superposed to His92 and His274, and the coordinating Zn^{2+} ions were extracted[28]. The extracted ligand molecules were merged with the representative structure.

MD simulations

MD simulations were performed by Gromacs version 2022.04 with the charmm36 force field[29,30]. The size of simulation boxes was determined by the molecule size with margins of 13 Å. After *in-vacuo* energy minimization to remove steric clashes, the protein-ligand complex was solvated by the TIP3P water model with 0.1 M NaCl, and the system was neutralized by adding additional Na^{+} or Cl^{-} ions, depending on the total charge of the protein and ligands. The energy was minimized by the steepest descent and equilibrated by 100 ps NVT and NPT simulations with harmonic restraints on the non-hydrogen atoms. The temperature and pressure of the system were controlled to 300 K and 1 bar by the V-rescale thermostat and Parrinello-Rahman barostat. Electrostatic interactions were computed by the particle mesh Ewald method, and bonds involving hydrogen atoms were constrained by the LINCS algorithm. For each docked model, we performed 20 trajectories of 100 ns simulations with a 2 fs time step.

Figure preparation

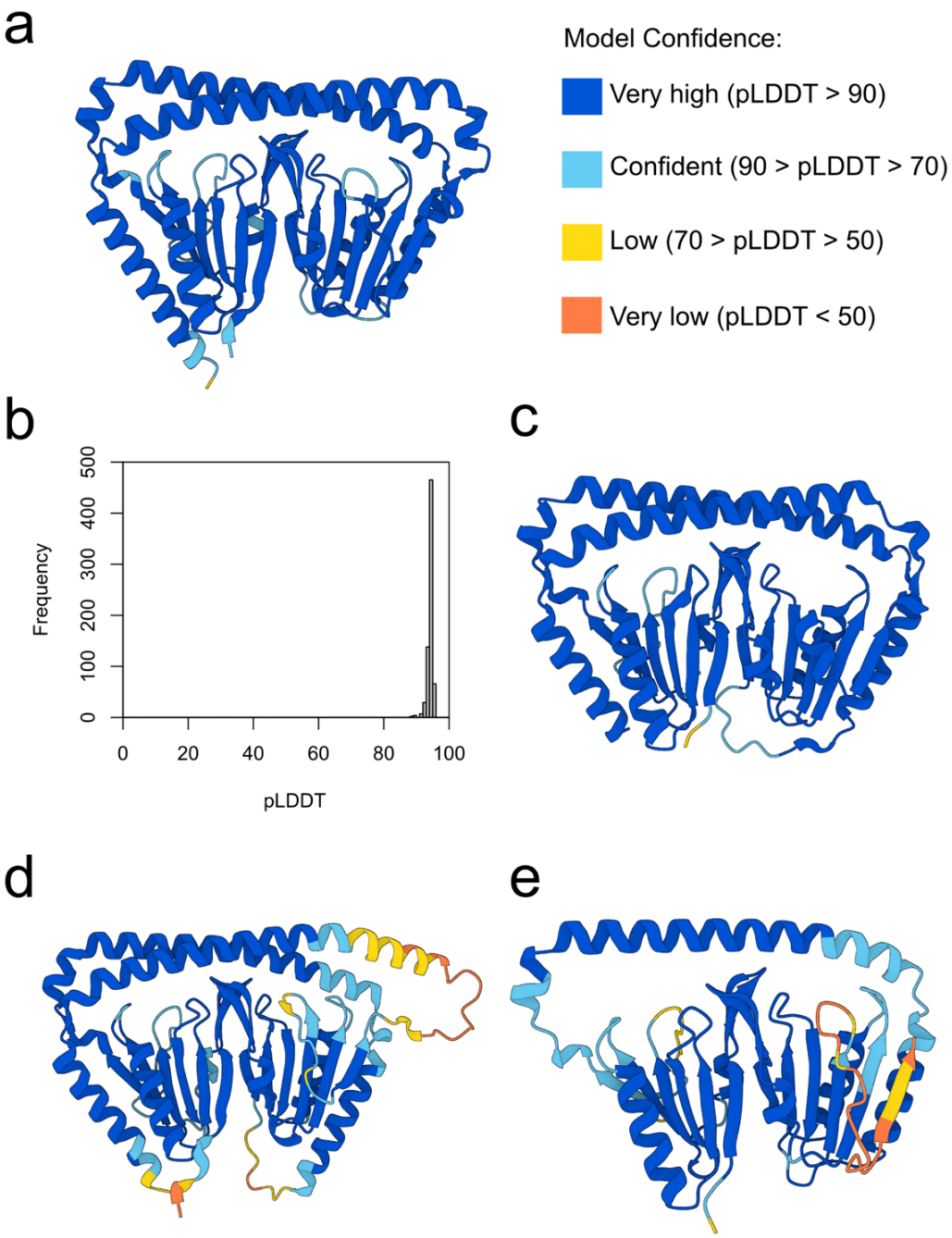
The images of molecular structures were created by PyMOL and Mol* viewer[27,31]. The surface electrostatic potential was calculated by the PyMOL APBS plugin[32]. Hydrogen bonds were detected by HBPLUS and visualized by PyMOL[33]. Secondary structure elements were assigned by DSSP and illustrated by ESPript[34,35].

Author contributions

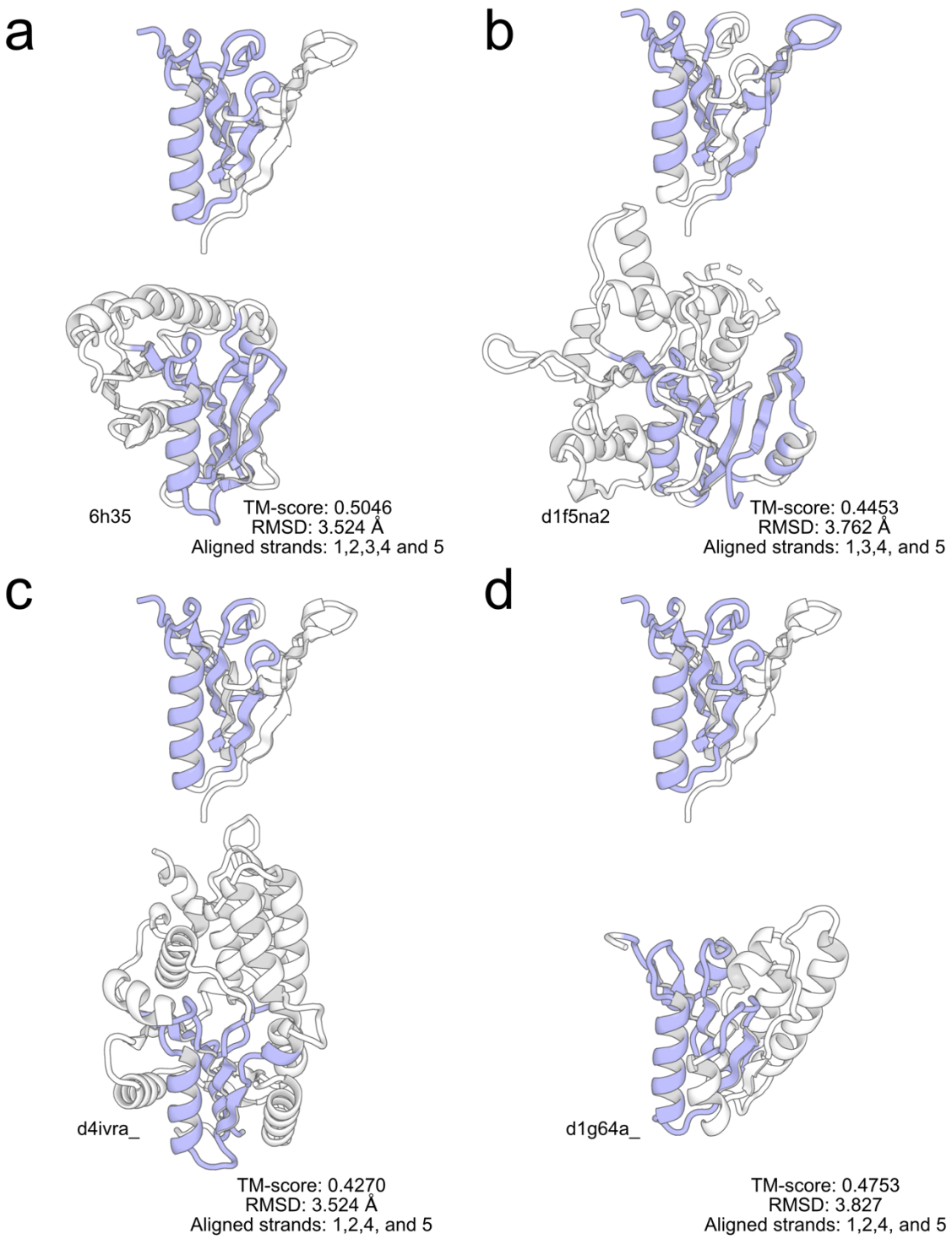
K.S. designed and initiated the study, performed structure mining, structure searches, structure comparisons, complex modeling, MD simulations, interpreted data and prepared and wrote the manuscript. R.K. and M.O. interpreted data and wrote the manuscript.

Acknowledgements

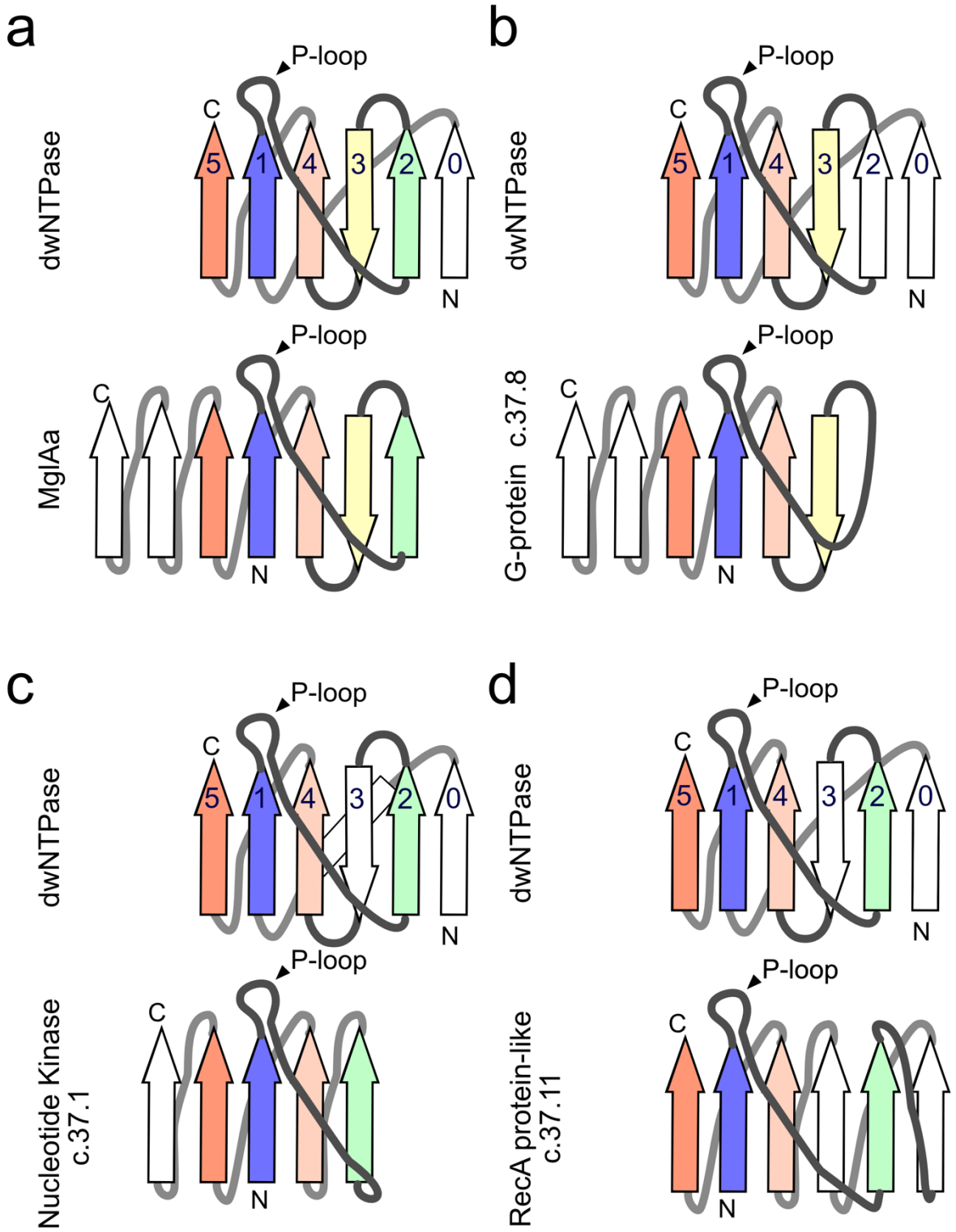
K.S. would like to thank George Chikenji for providing the program to analyze hydrogen-bonding patterns in β-sheets, Milot Mirdita for providing instructions on building the custom Foldcomp databases, Naoya Kobayashi and Shintaro Minami for their suggestions on extracting functionally relevant residues to infer functions from structures, Nobu C. Shirai for his constructive comments on the manuscript, and Shigeo S. Sugano and Ryo Ozuka for discussing the feasibility and significance of structure-based protein mining, which inspired K.S. to design this study. Structure searches by Foldseek and MD simulations were carried out on the supercomputer "Flow" at the Information Technology Center, Nagoya University. This study was supported by KAKENHI grant numbers JP21H00394 to R.K. and 20H05932 to M.O. We thank Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.



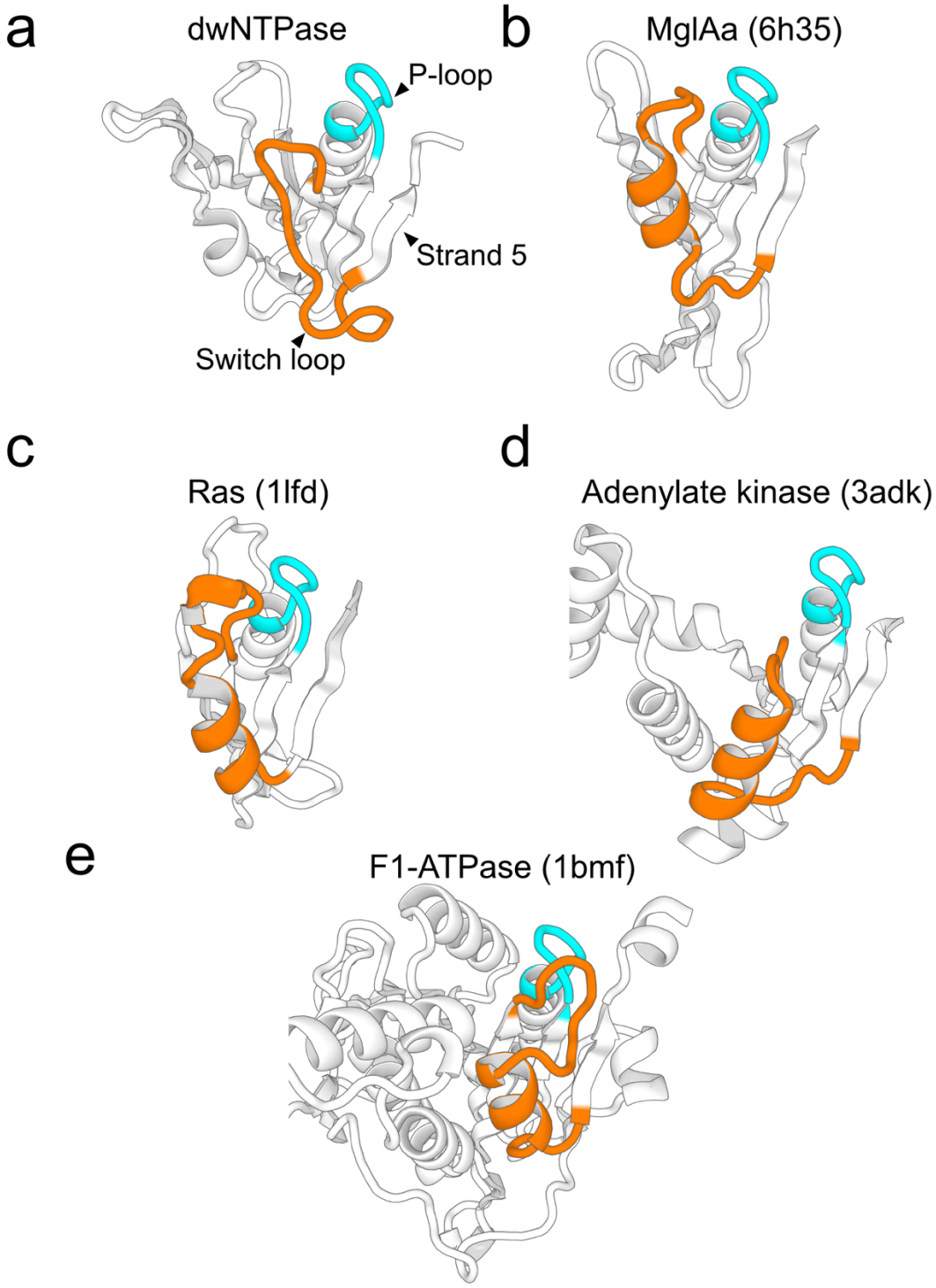
Extended Data Fig. 1 | Model confidence of initially mined dwNTPase structures. a, The cartoon shows the representative dwNTPase structure (UniProt accession No.: A0A1Y0TWD8). The cartoon is colored according to the residue-wise values of pLDDT. The coloring scheme is shown on the right of the structure, which is the same as displayed in AlphaFold DB. **b**, Distribution of total pLDDT values among the 711 dwNTPase structures initially mined from AlphaFold DB. **c–e**, The most confident, second worst and worst predictions among the 711 dwNTPase structures (A0A7V6TLZ4, A0A7C6J2V5 and W4RL53) gave average pLDDTs of 95.67, 88.43 and 88.01, respectively. Structures were created by Mol* viewer.



Extended Data Fig. 2 | Structure of the dwNTPase P-loop domain compared with other representative P-loop NTPase protein structures. Aligned structures of the dwNTPase P-loop domain and known NTPases. **a**, MglAa, (**b**) G-proteins, (**c**) NKs and (**d**) Rec-A like. The P-loop domain of dwNTPase and known NTPase are shown on the top and bottom of the panels, respectively. The latter three structures were selected from c.37.8, c.37.1 and c.37.11 SCCSs in SCOPe that showed the highest similarity to the dwNTPase P-loop domain. The aligned regions are colored blue. The PDB or SCOPe IDs are shown below the structures. The RMSD, TM-score and list of aligned strands in the dwNTPase P-loop domain are summarized at the bottom of the panels.



Extended Data Fig. 3 | Topology diagram of the dwNTPase P-loop domain compared with other representative P-loop NTPases. a, Comparisons with MglAa, (**b**) G-proteins, (**c**) NKs and (**d**) RecA-like proteins. Arrows represent strands. Grey and black lines indicate junctions projecting behind and out of the β-sheet. The pair of aligned strands are colored in the same colors. Unaligned strands are colored white. Helices are omitted for clarity.



Extended Data Fig. 4 | Comparison of the switch loop with the helical region conserved among P-loop NTPases. Typical P-loop NTPase structures are compared to the dwNTPase P-loop domain (**a**). MglAa (**b**), Ras (**c**), adenylate-kinase (**d**) and F1-ATPase (**e**) are selected from G-proteins (SCCS: c.37.8), NKs (c.37.1) and Rec-A like protein (c.37.11) from SCOPe. The switch loop or corresponding helical regions are orange, and the P-loops are cyan. C-terminal portions of the structures that follow strand 5 are omitted for clarity. Protein names and PDB IDs are shown above the panels.

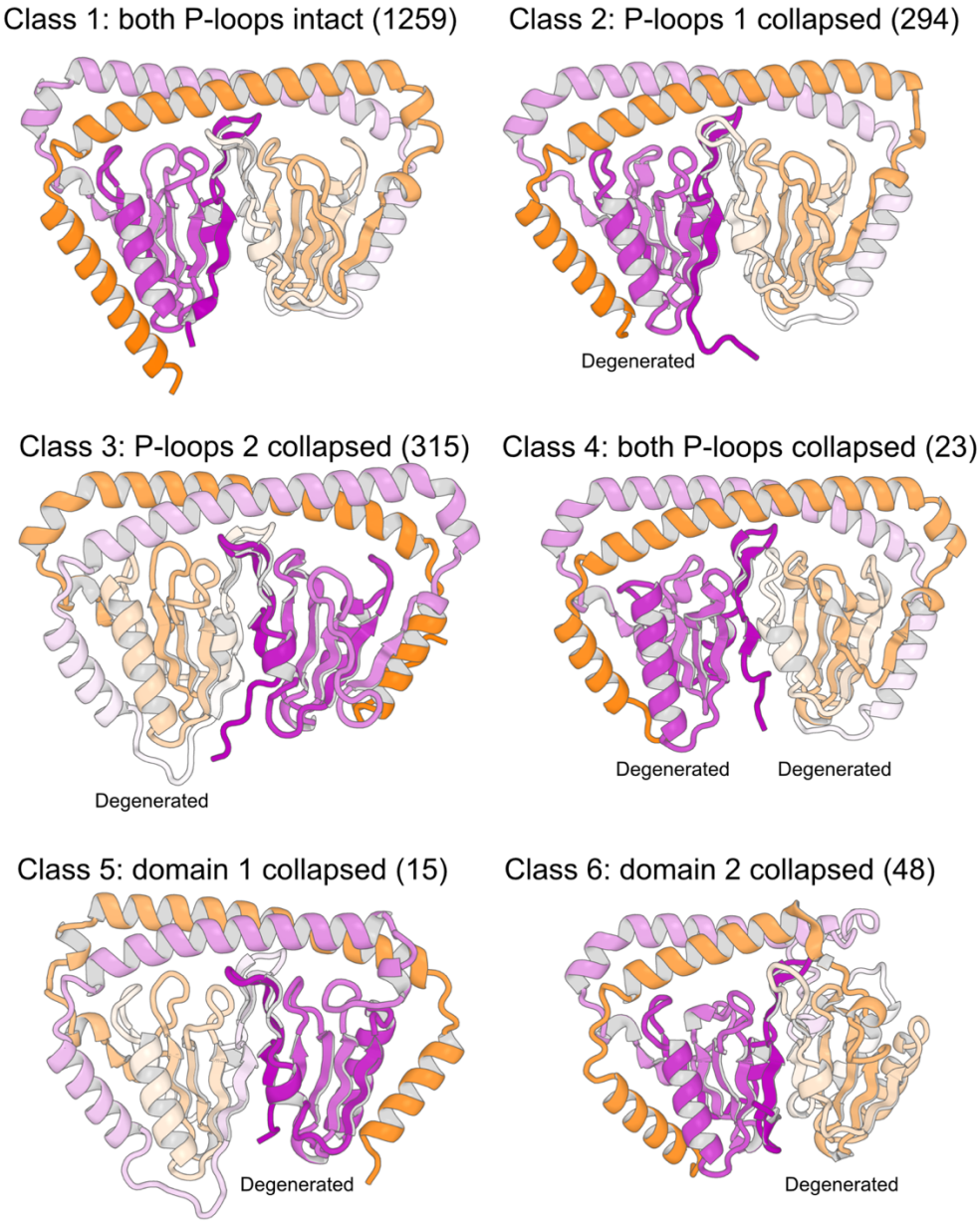
329

Domain	Phylum	Class	Count
Bacteria	Firmicutes	Bacilli	877
		Clostridia	705
		Desulfuribacillia	1
		Erysipelotrichia	5
		Negativicutes	13
		Tissierellia	32
		Unclassified Firmicutes	66
	Proteobacteria	Deltaproteobacteria	8
	Chloroflexi	Anaerolineae	1
	Tenericutes	Mollicutes	1
Archaea	Euryarchaeota	Methanomada group	1
Others			17

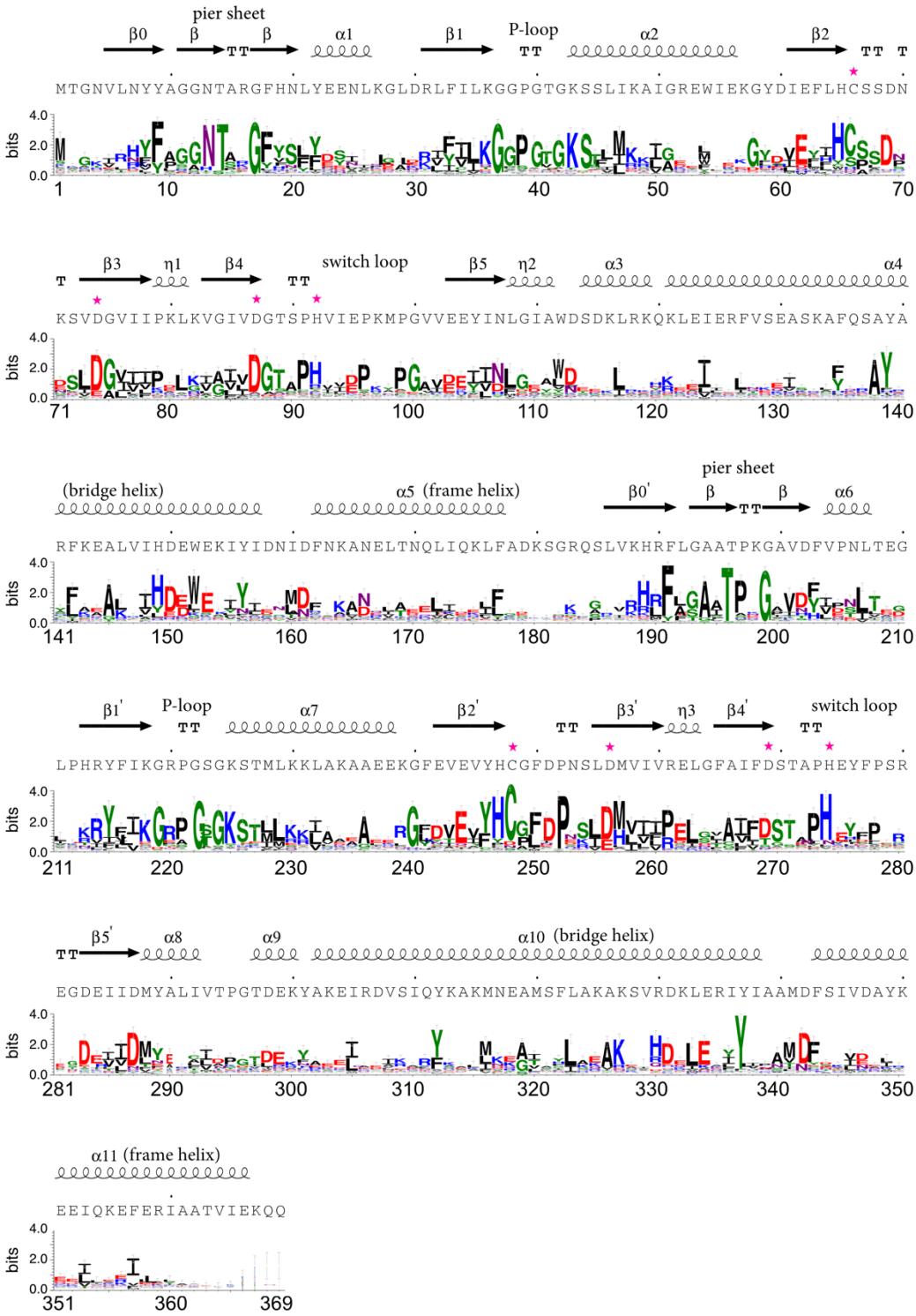
330

331 **Extended Data Fig. 5 | Distribution of dwNTPases among species.** We performed structural
332 alignment of all 2219 structures against the representative dwNTPase structure and selected
333 1843 structures showing TM-scores > 0.85 to ensure fragmented structures were excluded.
334 The structures were classified by their species. Entries with no phylogenetic information
335 available in UniProt were ignored. Others include environmental samples, metagenomes,
336 unclassified bacteria and Firmicutes from environmental samples.

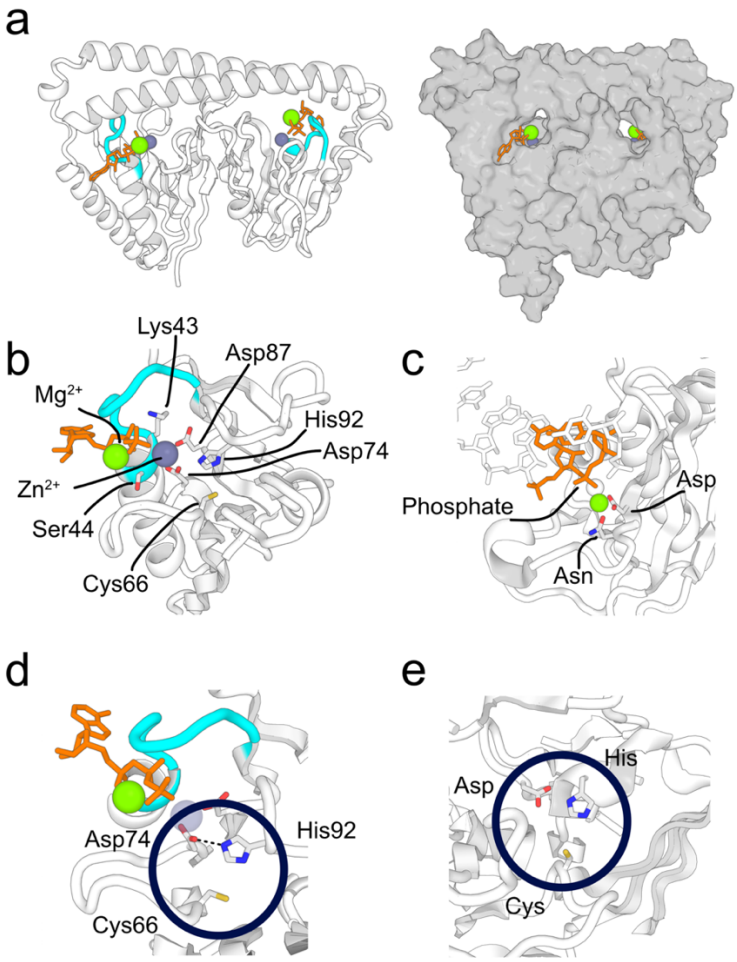
337



Extended Data Fig. 6 | Variations in the dwNTPase structure. Depending on the conservation of P-loops and domains, the dwNTPase family can be classified into six classes. The class 1 structure (e.g., Bt. dwNTPase, Uniport accession No. A0A1Y0TWD8) preserves both P-loops intact. The class 2 structure (A0A1T4Y6S3) has P-loop 1 degenerate and P-loop 2 intact. The class 3 structure (A0A1Y4J6T6) has P-loop 1 intact and P-loop 2 degenerate. The class 4 structure (A0A1I3I7D1) has both P-loops degenerate. The class 5 structure (A0A1Y4EVI7) has domain 1 degenerate and domain 2 intact. The class 6 structure (A0A101VWI5) has domain 1 intact and domain 2 degenerate. The number of structures that belong to respective classes is shown in parentheses. The structures are colored according to a purple-white-orange gradient from the N- to C-terminus. Note that the structures were rotated to show the collapsed region in front.



Extended Data Fig. 7 | Sequence logo of dwNTPase. Using HHblits, we performed a sequence search against the UniRef30 database and constructed multiple sequence alignments. Black, blue, green and red represent hydrophobic, positively charged, polar and negatively charged residues. The amino acid sequence of the representative structure (Bt. dwNTPase) and the secondary structures assigned by DSSP are shown. The vertical axis represents the bit score, and the horizontal axis represents the residue index. Putative functional residues discussed in the main text are indicated by purple stars.



Extended Data Fig. 8 | Putative functionally relevant residues and comparison to the active sites of known enzymes. a, Overall structure of the dwNTPase complexed with two pairs of ATPs, Mg^{2+} and Zn^{2+} ions after a 100-ns MD simulation. The structure of dwNTPase is shown as a white cartoon or grey surface representation. Mg^{2+} and Zn^{2+} are shown as green and grey spheres, respectively. ATP is shown in orange stick representation. **b**, Coordination of metal ions by two aspartate side chains observed in MD simulations. ATP is shown in orange stick representation. Side chains of relevant residues are shown as sticks and CPK coloring. Green and grey spheres represent Mg^{2+} and Zn^{2+} ions, respectively. The P-loop is colored cyan. **c**, The active site structure of RNase H (PDB ID: 1zbl). The side chain of metal coordinating amino acid residues Asp and Asn are shown as sticks and CPK coloring, where Asn is a mutation from Asp. Mg^{2+} ions are shown as spheres. The Mg^{2+} ion coordinating with the side chain of Asp and Asn is colored green. Nucleic acid residues that contact the Mg^{2+} ion are shown in orange. **d**, The catalytic triad-like side chain configuration observed during the MD simulations. The triad-like side chain cluster is circled. The black dotted line indicates the hydrogen bond between the side chains of H92 and D74, which HBPLUS detected. **e**, The active site structure of TEV protease (PDB ID: 1lvm). Side chains of the Cys-His-Asp catalytic triad are shown as sticks, CPK coloring and circled.

References (full)

1. Burley, S. K. *et al.* Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students. *Biomolecules* **12**, 1425 (2022).
2. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
3. Lin, Z. *et al. Evolutionary-scale prediction of atomic level protein structure with a language model*. http://biorxiv.org/lookup/doi/10.1101/2022.07.20.500902 (2022) doi:10.1101/2022.07.20.500902.
4. Bordin, N. *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun. Biol.* **6**, 160 (2023).
5. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951 (1982).
6. Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.* **317**, 41–72 (2002).
7. Leipe, D. D., Koonin, E. V. & Aravind, L. Evolution and Classification of P-loop Kinases and Related Proteins. *J. Mol. Biol.* **333**, 781–815 (2003).
8. Kim, H., Mirdita, M. & Steinegger, M. *Foldcomp: a library and format for compressing and indexing large protein structure sets*.

400 http://biorxiv.org/lookup/doi/10.1101/2022.12.09.519715 (2022)
401 doi:10.1101/2022.12.09.519715.
402 9. van Kempen, M. *et al. Foldseek: fast and accurate protein structure search.*
403 http://biorxiv.org/lookup/doi/10.1101/2022.02.07.479398 (2022)
404 doi:10.1101/2022.02.07.479398.
405 10. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*
406 **596**, 583–589 (2021).
407 11. Minami, S., Sawada, K., Ota, M. & Chikenji, G. MICAN-SQ: a sequential protein
408 structure alignment program that is applicable to monomers and all types of
409 oligomers. *Bioinformatics* **34**, 3324–3331 (2018).
410 12. Galicia, C. *et al.* MglA functions as a three-state GTPase to control movement reversals
411 of Myxococcus xanthus. *Nat. Commun.* **10**, 5300 (2019).
412 13. Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Res.* **51**, D384–
413 D388 (2023).
414 14. Yang, W. An equivalent metal ion in one- and two-metal-ion catalysis. *Nat. Struct.*
415 *Mol. Biol.* **15**, 1228–1231 (2008).
416 15. Dodson, G. & Wlodawer, A. Catalytic triads and their relatives. *Trends Biochem. Sci.*
417 **23**, 347–352 (1998).
418 16. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.*
419 **112**, 15898–15903 (2015).
420 17. Wintjens, R. T., Rooman, M. J. & Wodak, S. J. Automatic Classification and Analysis
421 of aa-Turn Motifs in Proteins.

422 18. Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. Structure at 2.8 Â
423 resolution of F1-ATPase from bovine heart mitochondria. *Nature* **370**, 621–628
424 (1994).
425 19. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment.
426 *Proteins Struct. Funct. Bioinforma.* **23**, 566–579 (1995).
427 20. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score =
428 0.5? *Bioinformatics* **26**, 889–895 (2010).
429 21. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for
430 the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
431 22. Chandonia, J.-M. *et al.* SCOPe: improvements to the structural classification of
432 proteins – extended database to facilitate variant interpretation and machine learning.
433 *Nucleic Acids Res.* **50**, D553–D559 (2022).
434 23. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative
435 protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175
436 (2012).
437 24. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for
438 improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
439 25. Bittrich, S., Burley, S. K. & Rose, A. S. Real-time structural motif searching in proteins
440 using an inverted index strategy. *PLOS Comput. Biol.* **16**, e1008502 (2020).
441 26. Wang, S. *et al.* Structural insights into substrate recognition by the type VII secretion
442 system. *Protein Cell* **11**, 124–137 (2020).
443 27. Schrodinger. The PyMOL Molecular Graphics System, Version 1.8. (2015).

444 28. Lee, B. M. *et al.* Induced Fit and "Lock and Key" Recognition of 5S RNA by Zinc
445 Fingers of Transcription Factor IIIA. *J. Mol. Biol.* **357**, 275–291 (2006).
446 29. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through
447 multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
448 30. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically
449 disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
450 31. Sehnal, D. *et al.* Mol* Viewer: modern web app for 3D visualization and analysis of
451 large biomolecular structures. *Nucleic Acids Res.* **49**, W431–W437 (2021).
452 32. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite.
453 *Protein Sci.* **27**, 112–128 (2018).
454 33. McDonald, I. K. & Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins.
455 *J. Mol. Biol.* **238**, 777–793 (1994).
456 34. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern
457 recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637
458 (1983).
459 35. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new
460 ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
461