# Dual-wield NTPase: 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**

**Contact: K.S. (ksakuma@i.nagoya-u.ac.jp), M.O. (mota@i.naogoya-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 NTPase proteins, noticeable two-fold symmetry, and independent two pairs of putative active sites. Our finding proves structural data mining is a powerful approach to identifying undiscovered protein families.**

Since protein structures are essential to investigate protein function, structural biologists have determined individual protein structures through experiments and deposited them in the Protein Data Bank (PDB)[1,2]. When the protein shows a novel structure that appeared at the first time, the finding is usually reported by the researchers who determined it. However, public structural databases produced by modern state-of-the-art structure predictions, such as AlphaFold protein structure database (AlphaFold DB) and ESM metagenomic Atlas (ESM Atlas), are changing the situation[3,4]. The databases are about three orders of magnitudes larger than the current PDB and contain a vast number of experimentally unsolved protein structures. These databases must hold structural models truly unwatched by human-being since the models were generated automatically by artificial intelligences and deposited without any human curations. The overwhelming number of predicted models provides opportunities for finding novel proteins based only on the structural information *in-silico*[5].

Dedicated data mining requires a clearly stated working hypothesis. Therefore, we set a simple question to be answered by the database search: Are there monomeric proteins that contain multiple phosphate-binding loops (P-loops) on a single continuous β-sheet? P-loop, so-called Walker-A motif, is a local functional motif to recognize phosphate groups that are shared among P-loop NTPase proteins such as ATPases, GTPases, and nucleotide kinases (NKs)[6–9]. In general, one P-loop resides on a single continuous β-sheet of three-layered α/β/α sandwich architecture. Our preliminary search against the PDB supported this observation: no experimentally determined structure has multiple P-loops in a β-sheet. However, there is no reason to rule out the possibility that a single β-sheet possesses multiple P-loops. We hypothesized that such unobserved multiple-P-loop structures should exist in AlphaFold DB and hence can be discovered via systematic data mining.

By the computational scanning of more than 0.2 billion entries in AlphaFold DB version 4[3,10], we extracted 15,977 single-chained structures possessing multiple P-loops. Then we performed hydrogen-bond network analysis and extracted 839 structures with multiple P-loops on a single continuous β-sheet[11]. By their structural similarity, the structures were grouped into 11 clusters[12]. As a result, we found an uncharacterized family of P-loop proteins, dual-wield P-loop NTPase (dwNTPase), as the largest cluster owning 711 members. All of their structural models were predicted with high confidence scores, i.e., the predicted Local Distance Difference Test (pLDDT) was 94.27 on average, meaning that predictions were reliable (Extended Data Fig 1)[13].

The overall architecture of dwNTPase was highly novel and showed noticeable two-fold symmetry. Fig. 1a shows the structure of a representative one from *Bacillus thuringiensis* (Bt. UniProt accession A0A1Y0TWD8). Two P-loop domains are tightly packed and surrounded by two long bridging α-helices and two framing α-helices. Regarding to the secondary structural elements (SSEs) in a large sheet, each of the domains comprises six β-strands, and twelve strands in total form a continuous single β-sheet, which shows 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 can be taken as tunnels rather than pockets (Fig. 1c). Structure search against the PDB clarified no similar structures had been experimentally solved[12,14]. Similarly, SwissProt subset of AlphaFold DB contained no similar structures, meaning that dwNTPase family has no reliable annotations manually verified by UniProt curators yet.

We found that even the P-loop domain of dwNTPase was structurally atypical for a P-loop NTPase protein by searches against the PDB (Fig. 2a)[12,14]. A crystal structure of mutual gliding-motility protein MglAa from *Myxococcus xanthus* (PDB 6h35), a type of bacterial small and monomeric GTPase, was the only known P-loop NTPase protein that showed relevant structural similarity to dwNTPase P-loop domain[15]. When compared to MglAa structure, the P-loop domain has an additional β-strand on the N-terminus (strand 0 in 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 SSEs around them. These unique arrangements of SSEs lead to the atypical topology that does not resemble other P-loop NTPase proteins (Extended Data Fig. 2 and 3)[8,9,16]. Furthermore, the P-loop domain has a long loop in place of a helix conserved in others (Extended Data Fig. 4), which we named the switch loop (Fig. 2a) and is discussed later. These atypical features of the P-loop domain make it difficult to assign dwNTPase to known classes of P-loop NTPase proteins.

Despite these novel features of dwNTPase, 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 (Extended Data Fig. 5 and Table 1)[3,12]. Similar searches against ESM Atlas culled by 30% sequence identity found 748 similar structures with more than 0.5 TM-score (Extended Data Table 2)[4,17]. We observed several subtypes of dwNTPase structure and classified them into six sub-classes by their conservation of motifs and domains (Extended Data Fig. 6). The bonafide dwNTPase structures with both two P-loops intact (class 1) were most abundant, which suggests that there are functional constraints to conserve two active P-loops. Blast search against the non-redundant database revealed that dwNTPase had been classified as PRK06851 family protein in NCBI conserved domain database, but tblastn search against eukaryotic genomes yielded no relevant hit[18]. From these observations, we concluded that dwNTPase constitutes a conserved protein family among bacteria.

Functions of dwNTPase are unknown; we sought to infer the molecular functions by analyzing the conserved residues (Extended Data Fig. 7). Although the sequence similarities between the symmetric halves of dwNTPase were about 23% on average, we found that the most symmetric class of dwNTPase (class 1 in Extended Data Figure 4) possesses several conserved residues shared among both halves (Fig. 2c). Notably, the cluster of conserved residues Cys66/Cys248 (the residue number follows that of Bt. dwNTPase), Asp74/Asp256, Asp87/Asp269, and His92/His274 formed putative metal binding sites. Molecular dynamics (MD) simulations of Bt. dwNTPase structure complexed with two ATPs, two Mg2+ ions, and two Zn2+ ions showed that the Zn2+ ions were stably coordinated by two aspartates (Asp74/Asp256 and Asp87/Asp269) and the γ-phosphate group of ATPs (Extended Data

Fig. 6 a and b), which resembles the active site structure of metal-dependent nucleotidyl-transfer enzymes (Extended Data Fig. 8 c)[19]. The sidechain of Cys66/Cys248 and His92/His274 stayed unoccupied (Extended Data Fig. 8 d), suggesting that they may have roles other than metal-binding. As the pair of cysteine and histidine residues reminds us of the catalytic triad/dyad in cysteine proteases (Extended Data Fig. 8 e), we speculate that dwNTPase may have additional hydrolase/ligase enzymatic activity[20].

In addition to these conserved residues, we identified other regions characteristic of dwNTPase. First, each P-loop domain has conserved lysine residues (Lys36/Lys218) which precede the P-loops and interact with two switch loops. Since the switch loop partially seals the ligand binding tunnels (Extended Data Fig. 9a) and highly fluctuated in MD simulations (Extended Data Fig. 9b), the conserved lysine residues may play sensor-like roles to trigger NTPase activity depending on the binding of other ligands to the tunnels. Second, P-loops are surrounded by several charged or polar residues that support the recognition of NTPs and Mg2+ ion (Extended Data Fig. 9c). To our knowledge, they are not conserved in known P-loop NTPase proteins.

The biological backgrounds of dwNTPases remain elusive since their structures appear too novel to gain insights by analogy. However, the structures of dwNTPase are strongly inspiring, and it is attempting to speculate that the family is responsible for unrecognized biological functions. The clear two-fold symmetry implies that the substrates of dwNTPases may also be two-fold symmetric molecules such as double-stranded DNAs, or that the cleft between two P-loop domains may recognize ligand molecules in similar manners to periplasmic heme-binding proteins (Extended Data Fig. 10)[21]. Since the left-half part (residue 1–139 and 321–369) of the structure in Fig. 1a was more positively charged than the right-half part (residue 140–320), each half may play different functional roles (Fig. 1 c and Extended Data Fig. 11). As for the ligand binding tunnels, the ability to recognize NTPs by P-loops alongside with another putative metal-binding sites imply functions like nucleotidyl-transfer enzymes or kinases. Since dwNTPases distribute among various Firmicutes, their biological functions may be related to spore-formation, which is characteristic of the species. Another mystery about dwNTPase is its evolutional origin. Although it is reasonable to assume that dwNTPase gained the two-fold symmetry via gene duplication, the origin of the P-loop domains remains obscure. Since dwNTPase constitutes a novel class of P-loop NTPase proteins, detailed phylogenetic analysis might impact the understanding of the evolution of P-loop NTPase protein[8,9]. Structural and biochemical studies are highly desirable to be more conclusive for the biological significance of dwNTPase family. In summary, we demonstrated that structural data mining can discover an uncharacterized protein family and is a powerful approach to exploring the dark proteomes, the unwatched region of the protein universe[22,23], which will help and encourage to design experimental studies.

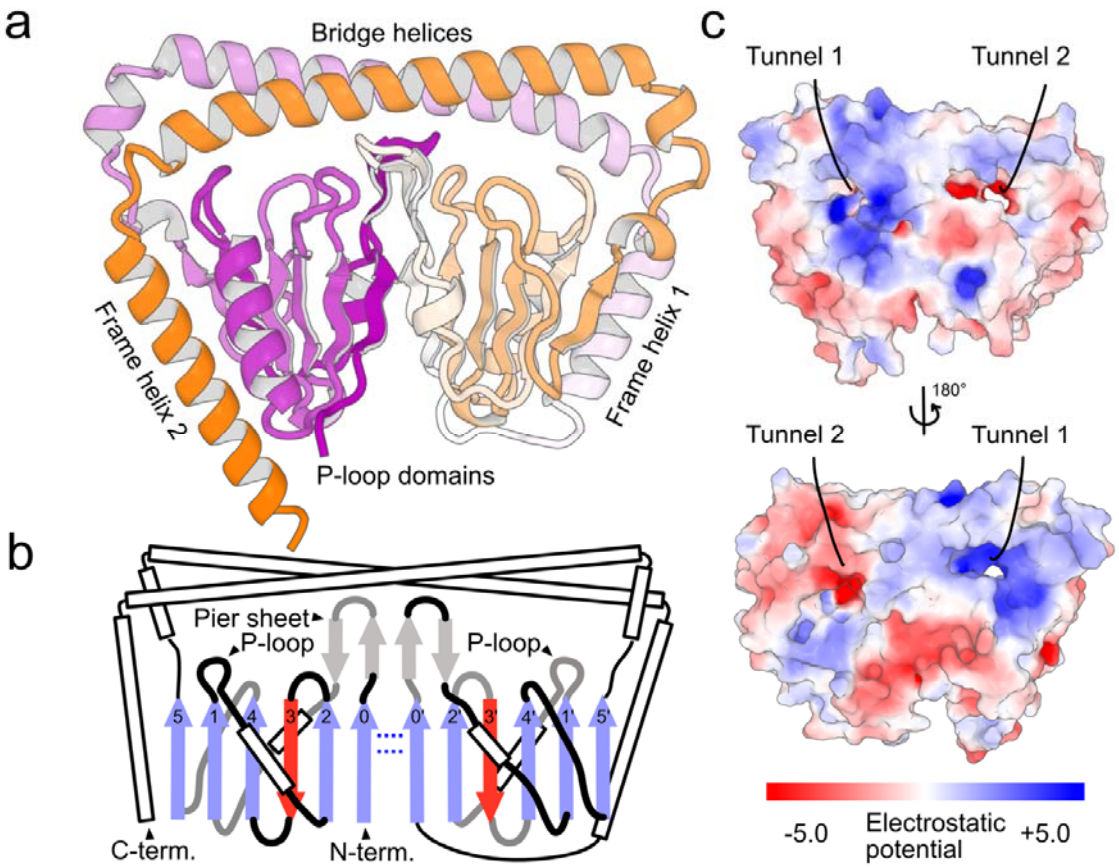



Fig. 1: **AlphaFold2 prediction model of dual-wield NTPase structure (AF-A0A1Y0TWD8-F1-model_v4).** (a) Cartoon representation of overall dwNTPase structure colored in purple-white-orange gradient from N- to C-termini. (b) Topology diagram of dwNTPase. Blue and red arrows represent □-strands pointing up and down that form the large β-sheets in the P-loop domains. The strands are numbered from N- to C-termini. Light-grey 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 the hydrogen bonds connecting the two halves of the large β-sheet. (c) The location and shape of ligand binding tunnels. Surface representation is colored according to the surface electrostatic potential as calculated by PyMOL APBS plugin.

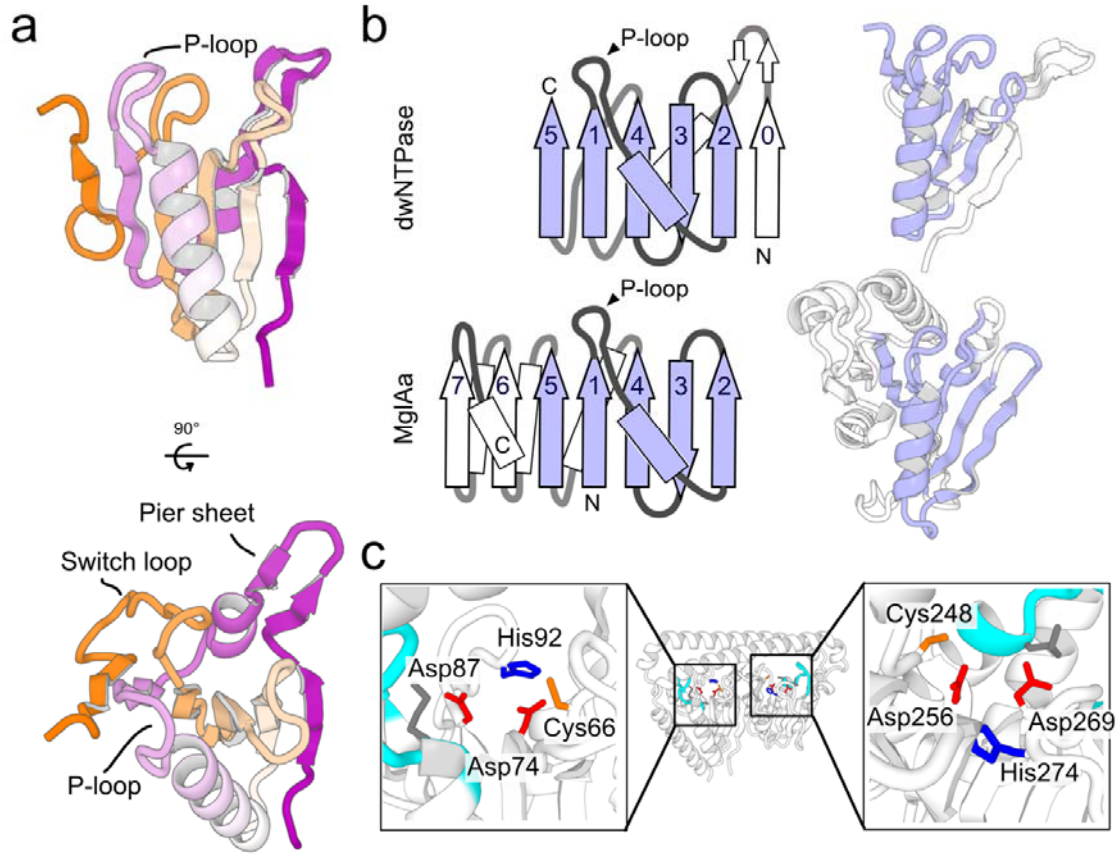



Fig. 2: **P-loop domain and its conserved residues.** (a) Front and top views of dwNTPase P-loop domain in cartoon representation colored in purple-white-orange gradient from N- to C-termini. P-loop, switch loop, and pier-sheet are indicated by labels. (b) Topology diagrams and cartoon representations of dwNTPase P-loop domains and MglAa structures. Arrows and rectangles represent β-strands and α-helices. Secondary structures that are aligned between two structures are colored blue. Grey and black lines indicate junctions between strands projecting behind or out of the alignable β-sheet.. Two topology diagrams are aligned with respect to their pre-P-loop strands (strand 1). (c) Conserved residues in the putative ligand binding tunnels. The conserved residues are shown in sticks. 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 Foldcomp database[3,10]. To extract the P-loop NTPase protein structures, we converted the models into the sequences of ABEGO, where A, B, E, and G respectively denote dihedral angles (phi, psi) for α, β, left-handed β, left-handed α on the Ramachandran plot[24]. O denotes other conformations unassignable on Ramachandran plot, typically 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 (chain A and chain D of PDB entry 1BMF)[25]. Since P-loop is known to serve as 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 match. Then, we calculated the C-α root-mean-square deviations (RMSD) of the matched substructures against the reference P-loop fragment (residue 166–179 of 1BMF chain A) and filtered out the ones with RMSD larger than 2.0 Å. We obtained 15,977 proteins containing multiple P-loop-like fragments and built a custom Foldcomp database for the following procedures.

### Identification of dual-wield NTPases

By the visual inspection, we found that most of the structures owning multiple P-loop-like fragments within a single chain were tandem repeats of known P-loop NTPase domains connected by flexible linkers. To exclude such proteins, we analyzed structures by an in-house program that detects the β-sheet hydrogen-bond network and extracted the structures possessing two P-loops-like fragments on a single continuous β-sheet[11]. We obtained 839 structures possessing two P-loop-like fragments on a single β-sheet. These structures were clustered by TM-score calculation with Foldseek into 11 clusters[12]. The largest cluster contained 711 members, which corresponded to dual-wield NTPase proteins. 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 (AF-A0A1Y0TWD8-F1-model_v4)[14].

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

To enumerate as many structures that resemble dwNTPase as possible, we performed iterative structure searches using FoldSeek[12]. In the first stage, we performed a structure search using all of the 711 structures mined from AlphaFold DB as queries. After removing overlapping structures, we obtained 1377 structures. Using these structures as the seeds, we again performed a Foldseek search and obtained 135 new non-overlapping structures. The third iteration of Foldseek searches resulted in some non-specific hits. Therefore, we stopped the iteration, manually selected similar structures, and discarded the rest. Consequently, we obtained 2219 structures of dwNTPase from AlphaFold DB. Similarly, we performed structure searches against highquality_clust30 subset of ESM-atlas using 711 structures found in AlphaFold DB as queries and obtained 748 structures with TM-score larger than 0.5[4,17].

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

To assess the novelty of dwNTPase structure and gain insights into the function, we performed structural searches against PDB100 and Swiss-Prot subset of AlphaFold DB (version 4) by Foldseek server in TM-align mode using the representative structure as the query[3,12]. No relevant (TM-score >=0.5) hit was found among these databases[17]. To be more rigorous, we used MICAN to perform one-against-all searches without pre-filtering, but no

similar (TM-score >= 0.5) structures were found among the PDB as of 2023-01-09 and Swiss-Prot subset of AlphaFold DB (version 2)[14].

**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 (residue 1–110) against the PDB100 and Swiss-Prot subset of AlphaFold DB by Foldseek server, but no relevant hit was found[12]. To be more rigorous, we used MICAN to perform structure search without pre-filtering against the PDB as of 2023-09-01 and Swiss-Prot subset of AlphaFold DB (version 2)[3,14]. 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 35% and obtained 15 and 137 clusters[26]. For all the cluster representatives, we checked the alignments by visual inspections. 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, all of which are annotated as GTPase or GTP-binding proteins[15]. The rest of hits showed RecA-like topology and were not topologically identical to dwNTPase, since RecA-like topology has all-parallel β-sheet, whereas dsNTPase has anti-parallel-containing β-sheets[8,9]. Similarly, we performed structural comparison 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[14,16]. 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 Figure 2).

**Calculation of sequence identities between two halves of dwNTPase structures**

We selected 1,903 structures owning more than 340 residues from the set of dwNTPases extracted from AlphaFold DB. A structure is self-aligned by MICAN with the rewiring mode, which ignores the sequential order of SSEs. The sequence identity was calculated based on the second-best alignment since it realizes the superposition of the halves of the symmetric structure[14].

**Identification of putative catalytic residues and side-chain patterns search against the PDB**

To gain insight into the function of the dwNTPase, we performed a sequence search and alignment to identify the conserved residues by HHblits against UniRef30_2022_02[27,28]. After 3 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 the aligned sequences with more than 10 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 determine the conserved residues, and the top 10 residues around two tunnels are 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, residue 244, 246, 247, 248, 252, 255, 256, 261, 263, and 274 were identified. According to the orientation of sidechain toward the tunnel, we regarded residues 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 regarded as the clusters of functional residues for tunnels 1 and 2, respectively. To examine whether known protein structures possess similar sidechain configurations, we performed sidechain pattern search against the PDB by

strucmotif-search program[29]. 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 as of the 2022-28-12 with the threshold for the structure similarity 1.0 Å. The side-chain pattern search resulted in no hit and suggested the putative catalytic residues have novel configurations 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 (6j18) was superposed to the P-loop of the representative structure by MICAN on PyMOL and the ATP and Mg2+ models were extracted[30]. Similarly, His125 from zinc finger motif (2hgh) was superposed to His92 and His274 and the coordinating Zn2+ ions were extracted[31]. The extracted ligand molecules were merged to the representative structure.

**MD simulations**

MD simulations were performed by Gromacs version 2022.04 with charmm36 force field[32,33]. 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 TIP3P water model with 0.1 mol/L of 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 steepest descent and equilibrated by 100 ps NVT and NPT simulations with harmonic restraints on the non-hydrogen atoms. The system's temperature and pressure was controlled to 300 K and 1 bar by V-rescale thermostat and Parrinello-Rahman barostat. The electrostatic interaction was computed by particle mesh Ewald method and bonds involving the hydrogen atoms were constrained by LINCS algorithm. For each docked model, we performed 20 trajectories of 100 ns simulations with 2 fs time step.

**Figure preparation**

The images of molecular structures were created by PyMOL and Mol* viewer[34,35]. The surface electrostatic potential was calculated by PyMOL APBS plugin[36]. Hydrogen bonds were detected by HBPLUS and visualized by PyMOL[34,37]. Secondary structure elements were assigned by DSSP and illustrated by ESPript[38,39].

**Author contributions**

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

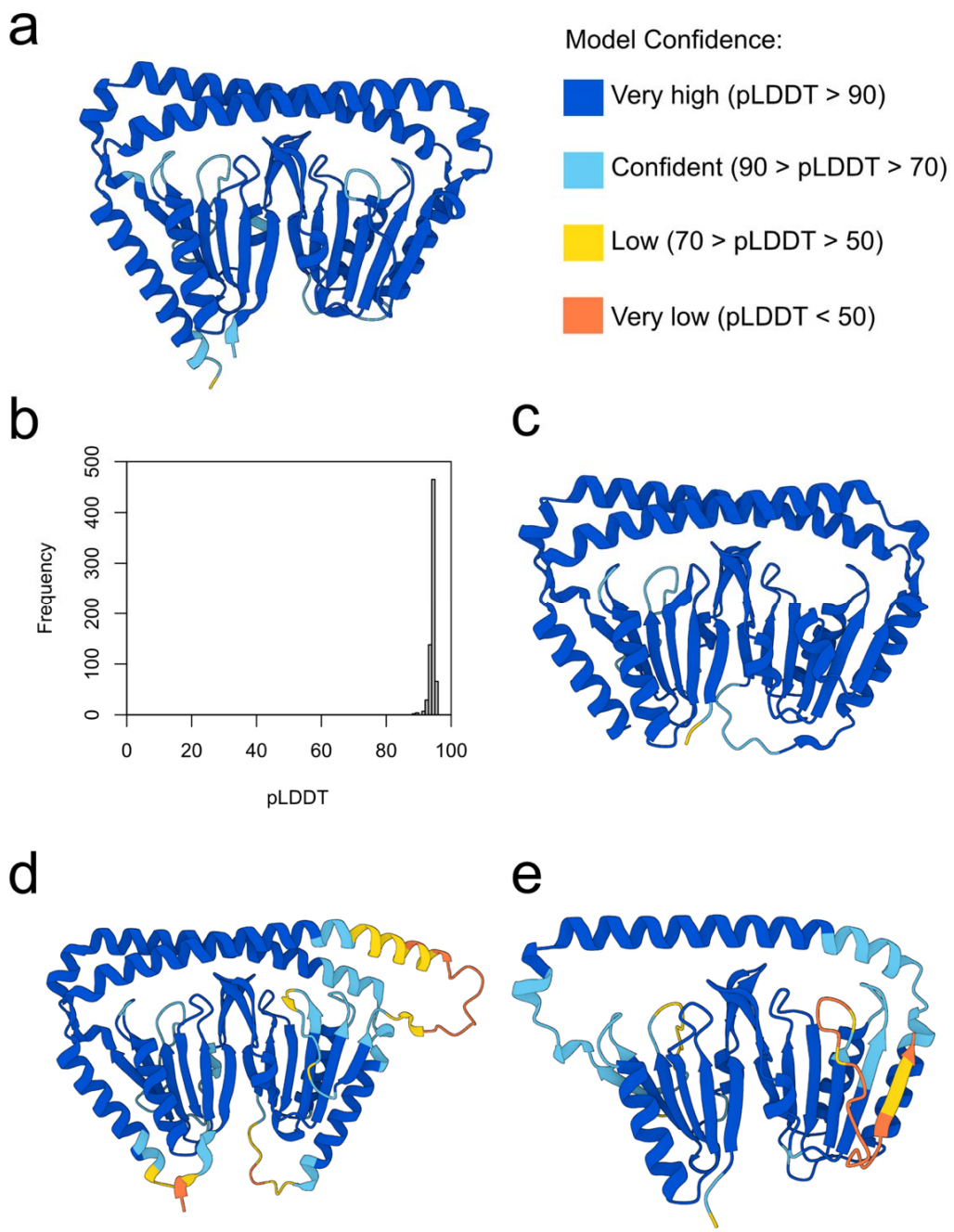
**Acknowledgements**

K.S. would like to thank Dr. George Chikenji for providing the program to analyze hydrogen-bonding patterns in β-sheets. K.S would like to thank Dr. Milot Mirdita for instructing the procedure to build custom Foldcomp databases. K.S. would like to thank Dr. Naoya Kobayashi and Dr. Shintaro Minami for their suggestions on extracting functionally relevant residues to infer the functions from structures. K.S. would like to thank Dr. Shigeo S. Sugano and Mr. 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., 20H05932 to M.O.

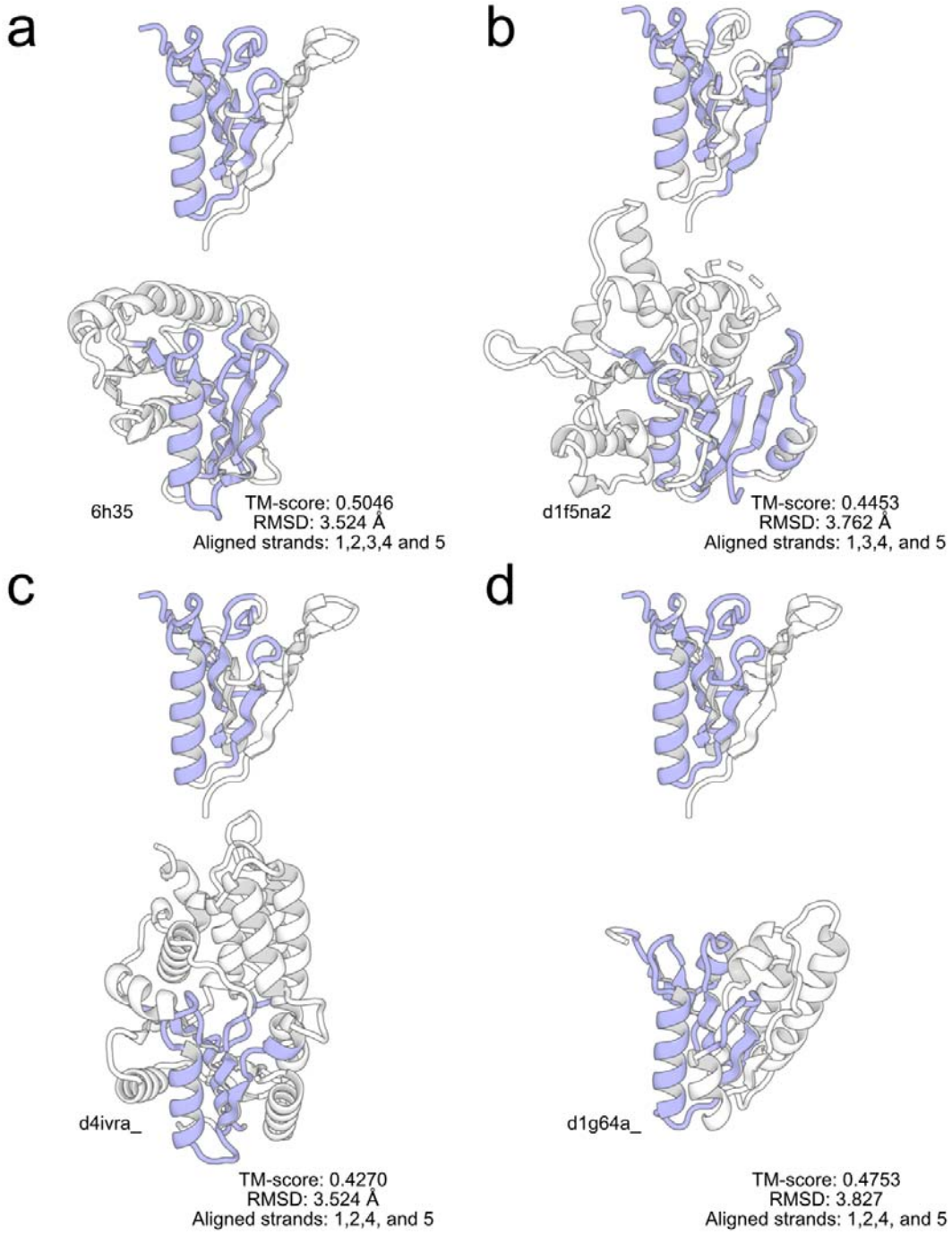
References


1. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
2. 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).
3. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).
4. 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.
5. Bordin, N. *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol* **6**, 160 (2023).
6. 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. *The EMBO Journal* **1**, 945–951 (1982).
7. Saraste, M., Sibbald, P. R. & Wittinghofer, A. The P-loop — a common motif in ATP- and GTP-binding proteins. *Trends in Biochemical Sciences* **15**, 430–434 (1990).
8. Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. Classification and evolution of P-loop GTPases and related ATPases. *Journal of Molecular Biology* **317**, 41–72 (2002).
9. Leipe, D. D., Koonin, E. V. & Aravind, L. Evolution and Classification of P-loop Kinases and Related Proteins. *Journal of Molecular Biology* **333**, 781–815 (2003).
10. Kim, H., Mirdita, M. & Steinegger, M. *Foldcomp: a library and format for compressing and indexing large protein structure sets*. http://biorxiv.org/lookup/doi/10.1101/2022.12.09.519715 (2022) doi:10.1101/2022.12.09.519715.
11. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* **23**, 566–579 (1995).
12. van Kempen, M. *et al. Foldseek: fast and accurate protein structure search*. http://biorxiv.org/lookup/doi/10.1101/2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.
13. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
14. Minami, S., Sawada, K., Ota, M. & Chikenji, G. MICAN-SQ: a sequential protein structure alignment program that is applicable to monomers and all types of oligomers. *Bioinformatics* **34**, 3324–3331 (2018).
15. Galicia, C. *et al.* MglA functions as a three-state GTPase to control movement reversals of Myxococcus xanthus. *Nat Commun* **10**, 5300 (2019).
16. Chandonia, J.-M. *et al.* SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research* **50**, D553–D559 (2022).
17. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
18. Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Research* **51**, D384–D388 (2023).
19. Yang, W. An equivalent metal ion in one- and two-metal-ion catalysis. *Nat Struct Mol Biol* **15**, 1228–1231 (2008).
20. Dodson, G. & Wlodawer, A. Catalytic triads and their relatives. *Trends in Biochemical Sciences* **23**, 347–352 (1998).
21. Mattle, D., Zeltina, A., Woo, J.-S., Goetz, B. A. & Locher, K. P. Two Stacked Heme Molecules in the Binding Pocket of the Periplasmic Heme-Binding Protein HmuT from Yersinia pestis. *Journal of Molecular Biology* **404**, 220–231 (2010).
22. Taylor, W. R., Chelliah, V., Hollup, S. M., MacDonald, J. T. & Jonassen, I. Probing the “Dark Matter” of Protein Fold Space. *Structure* **17**, 1244–1252 (2009).
23. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15898–15903 (2015).

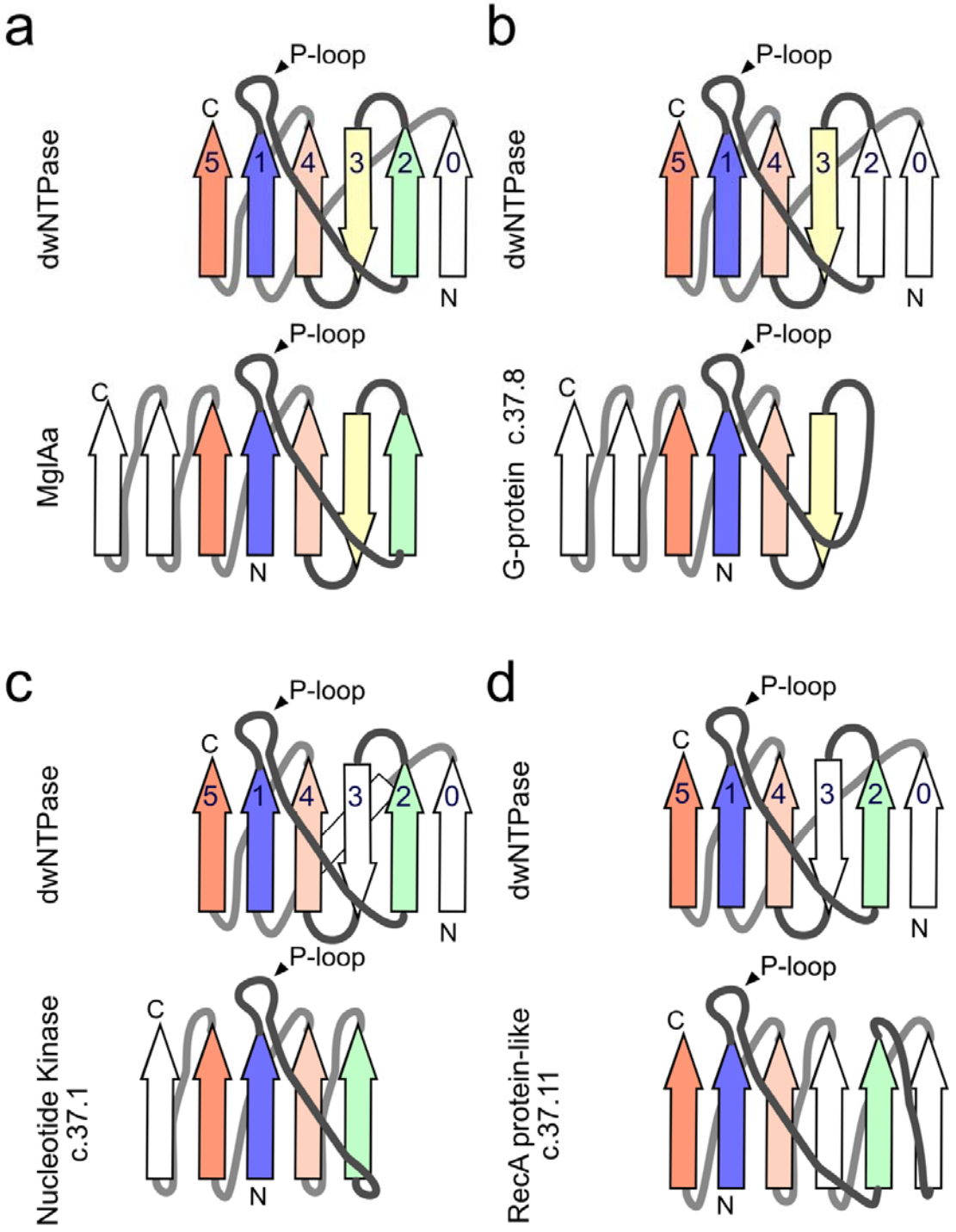
24. Wintjens, R. T., Rooman, M. J. & Wodak, S. J. Automatic Classification and Analysis of aa-Turn Motifs in Proteins.
25. Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. Structure at 2.8 Â resolution of F1-ATPase from bovine heart mitochondria. *Nature* **370**, 621–628 (1994).
26. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
27. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173–175 (2012).
28. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
29. Bittrich, S., Burley, S. K. & Rose, A. S. Real-time structural motif searching in proteins using an inverted index strategy. *PLOS Computational Biology* **16**, e1008502 (2020).
30. Wang, S. *et al.* Structural insights into substrate recognition by the type VII secretion system. *Protein & Cell* **11**, 124–137 (2020).
31. Lee, B. M. *et al.* Induced Fit and "Lock and Key" Recognition of 5S RNA by Zinc Fingers of Transcription Factor IIIA. *Journal of Molecular Biology* **357**, 275–291 (2006).
32. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
33. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73 (2017).
34. Schrodinger. The PyMOL Molecular Graphics System, Version 1.8. (2015).
35. Sehnal, D. *et al.* Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research* **49**, W431–W437 (2021).
36. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Science* **27**, 112–128 (2018).
37. McDonald, I. K. & Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology* **238**, 777–793 (1994).
38. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
39. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research* **42**, W320–W324 (2014).



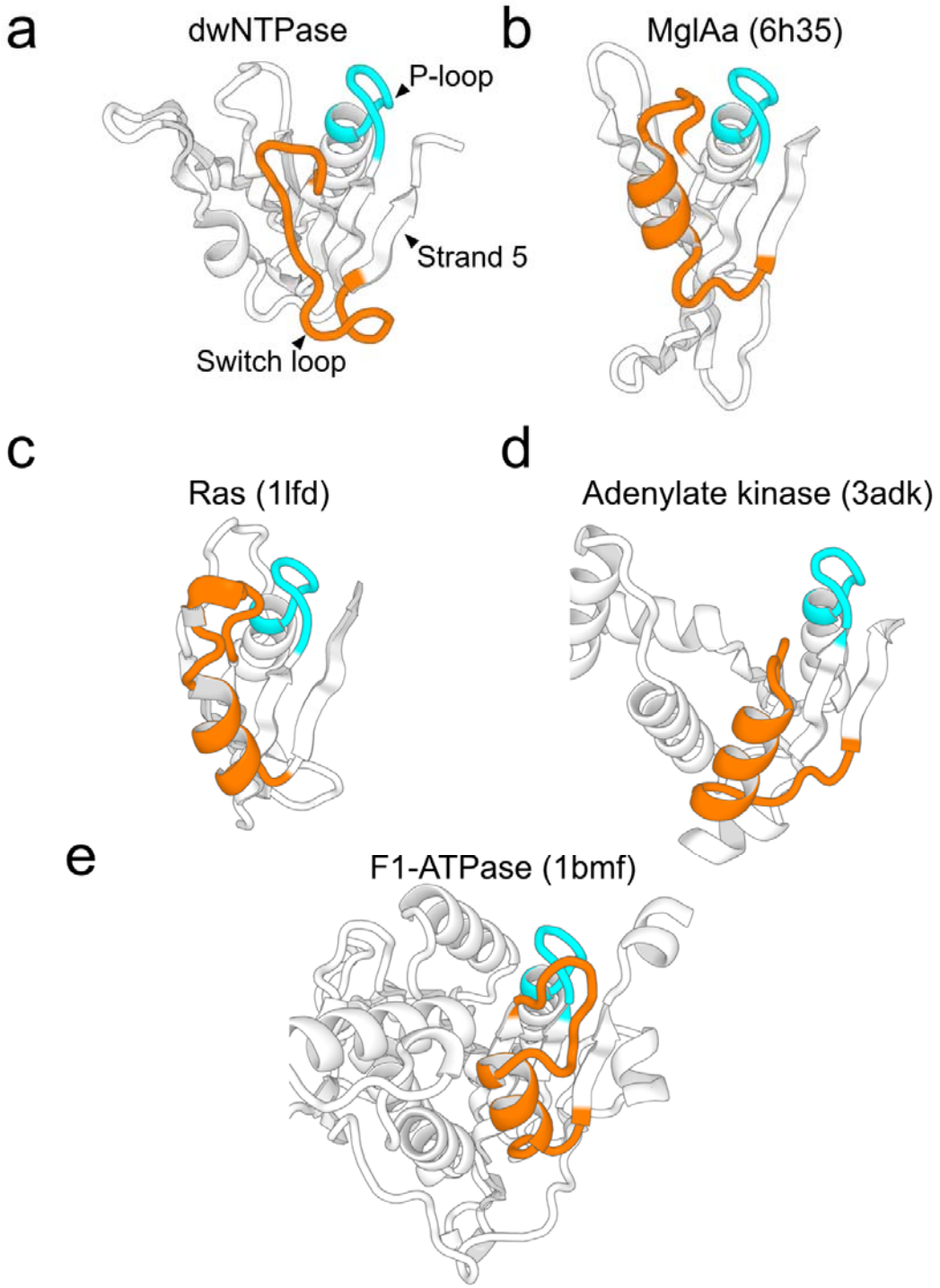

**Extended Data Figure 1: Model confidence of initially mined dwNTPase structures.** (a) The cartoon representation of the representative dwNTPase structure (UniProt accession A0A1Y0TWD8). The cartoon is colored according to the residue-wise values of pLDDT. The coloring scheme is shown on the right of the structure, which is the same as displayed in AlphaFold DB. (b) Distribution of total pLDDT values among 711 dwNTPase structures initially mined from AlphaFold DB. (c)(d)(e) The most confident, second worst, and worst predictions among 711 dwNTPase structures (A0A7V6TLZ4, A0A7C6J2V5, and W4RL53) showed the averaged pLDDTs 95.67, 88.43, and 88.01 respectively. The images of structure models were created by Mol* viewer.




**Extended Data Figure 2: Structure of dwNTPase P-loop domain compared with other representative P-loop NTPase protein structures.** Aligned structures of dwNTPase P-loop domain and known NTPase. (a) MglAa. (b) G-proteins. (c) NKs. (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 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 dwNTPase P-loop domain are summarized at the bottom of the panels.



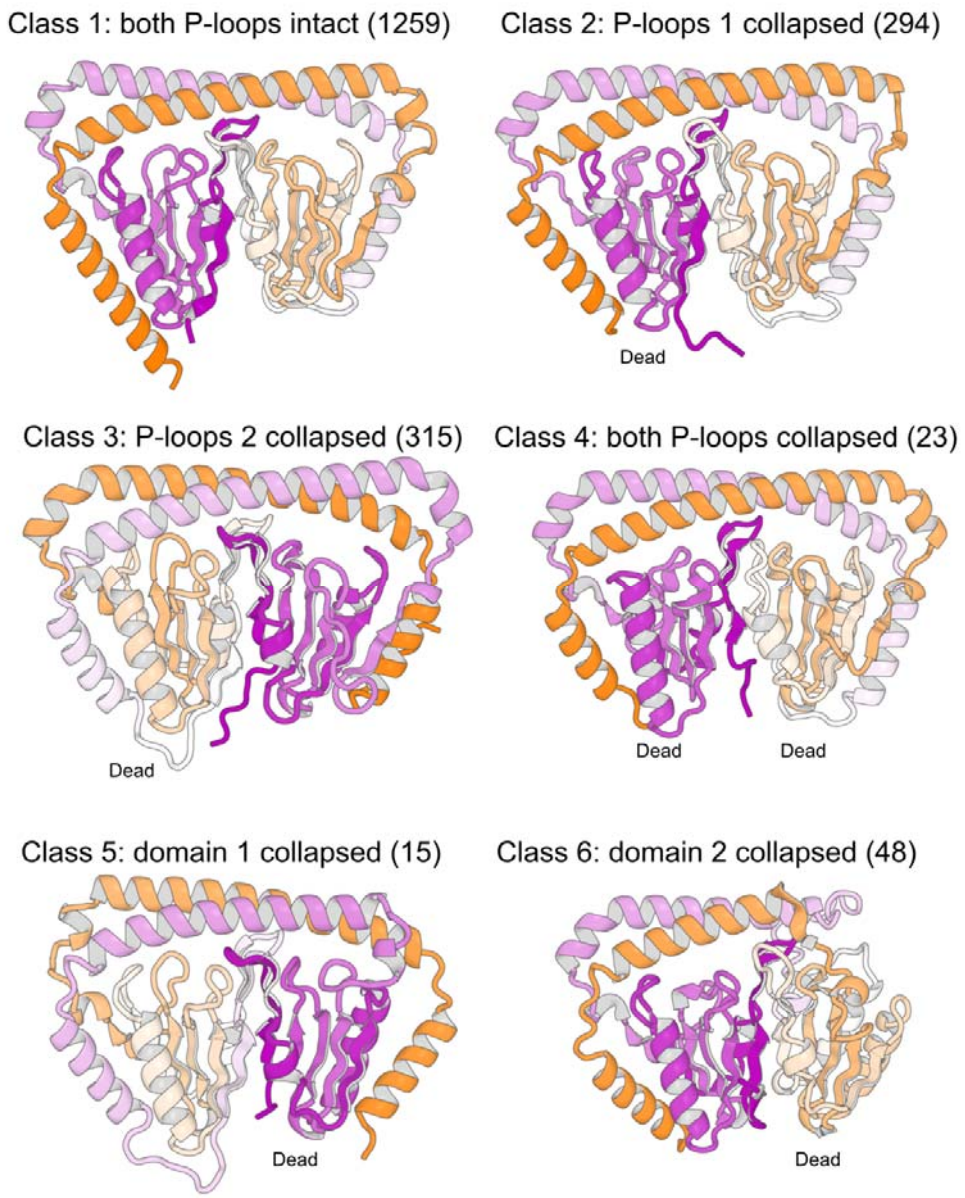

**Extended Data Figure 3: Topology diagram of dwNTPase P-loop domain compared with other representative P-loop NTPase proteins.** (a) Comparisons with MglAa, (b) G-proteins, (c) NKs, and (d) RecA-like proteins. Arrows represent strands. Grey and black lines indicate junctions projecting behind and out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



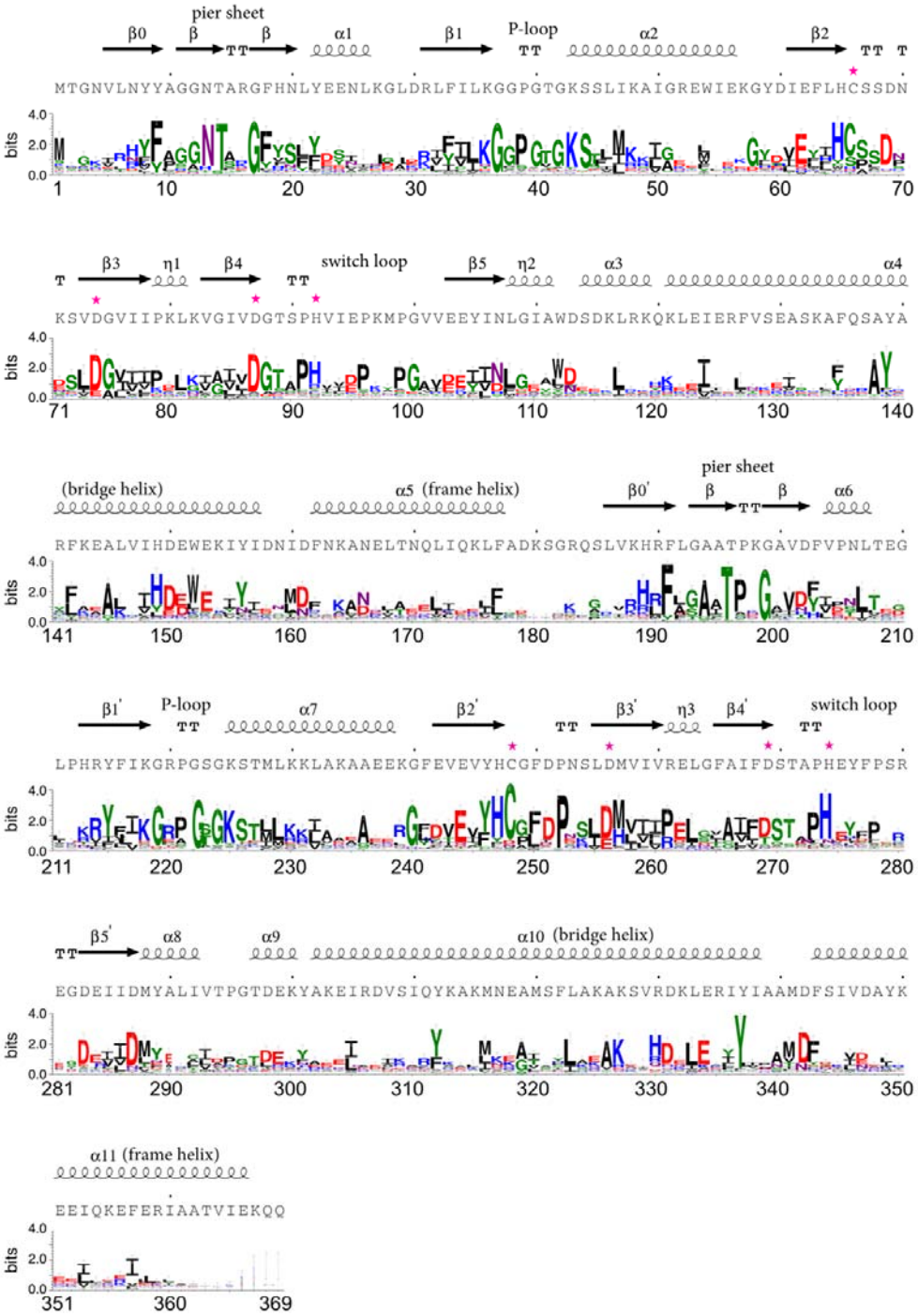

**Extended Data Figure 4: Comparison of switch loop with the helical region conserved among P-loop NTPase proteins.** Typical P-loop NTPase structures are compared to dwNTPase P-loop domain (a). MglAa (b), Ras (c), adenylate-kinase (d), and F1-ATPase (e) are selected from G-proteins (SCCS: c.37.8), NKs (c.37.1), and Rec-A like protein (c.37.11) from SCOPe. Switch loop or corresponding helical regions are colored orange. P-loops are colored 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.

| 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 |

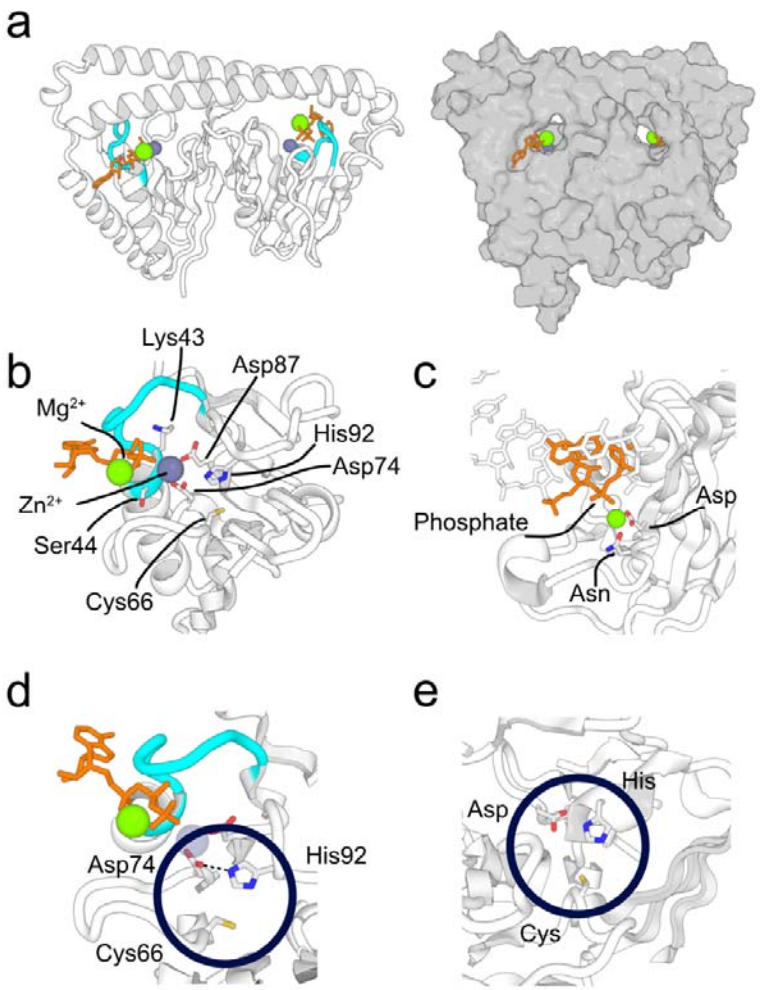
**Extended Data Figure 5: Distribution of dwNTPase among species. W**e performed structural alignment of all 2219 structures against the representative dwNTPase structure and selected 1843 structures showing TM-score more than 0.85 to exclude the fragmented structures. The structures were classified by their species. The entries that have no phylogenetic information available in UniProt were ignored. Others include environmental samples, metagenomes, unclassified bacteria, and Firmicutes from environmental samples.



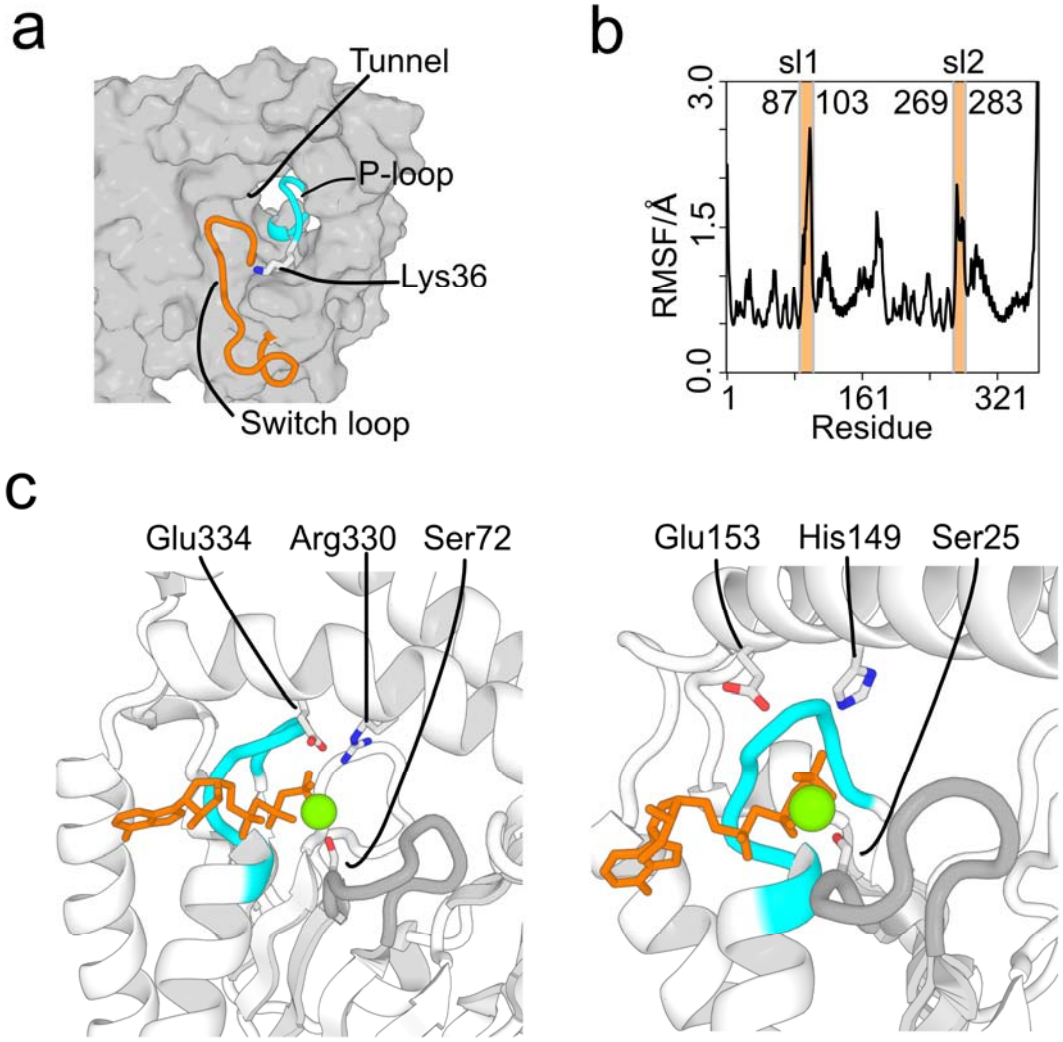

**Extended Data Figure 6: Variations of dwNTPase structure.** Depending on the conservation of P-loops and domains, dwNTPase family can be classified into six classes. Class 1 structure (e.g. Bt. dwNTPase, Uniport accession A0A1Y0TWD8) preserves both P-loops intact. Class 2 structure (A0A1T4Y6S3) has P-loop 1 degenerated and P-loop2 intact. Class 3 structure (A0A1Y4J6T6) has P-loop 1 intact and P-loop 2 degenerated. Class 4 structure (A0A1I3I7D1) has both P- loop degenerated. Class 5 structure (A0A1Y4EVI7) has domain 1 degenerated and domain 2 intact. Class 6 structure (A0A101VWI5) has domain 1 intact and domain 2 degenerated. The numbers of the structures that belong to respective classes are shown in parentheses. The structures are shown as cartoon colored in purple-white-orange gradient from N-terminus to C-terminus. Note that the structures were rotated to show the collapsed region in front.



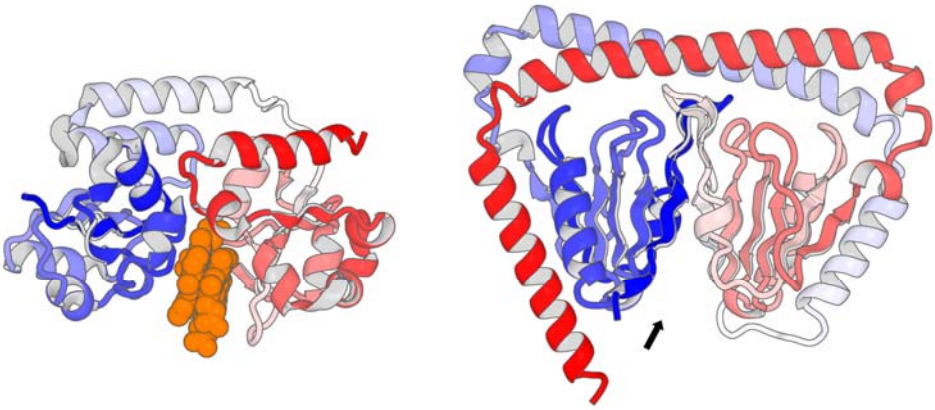

**Extended Data Figure 7: Sequence logo of dwNTPase.** Using HHblits, we performed sequence search against UniRef30 database and constructed multiple sequence alignments. Black, blue, green, and red represent hydrophobic, positively charged, polar and negatively charged residues. Amino-acid sequence of the representative structure (Bt. dwNTPase) and the secondary structures assigned by DSSP are shown. The vertical axis represents the bit score, and horizontal axis represents the residue index. The putative functional residues discussed in the main text are marked by purple stars placed on the characters.



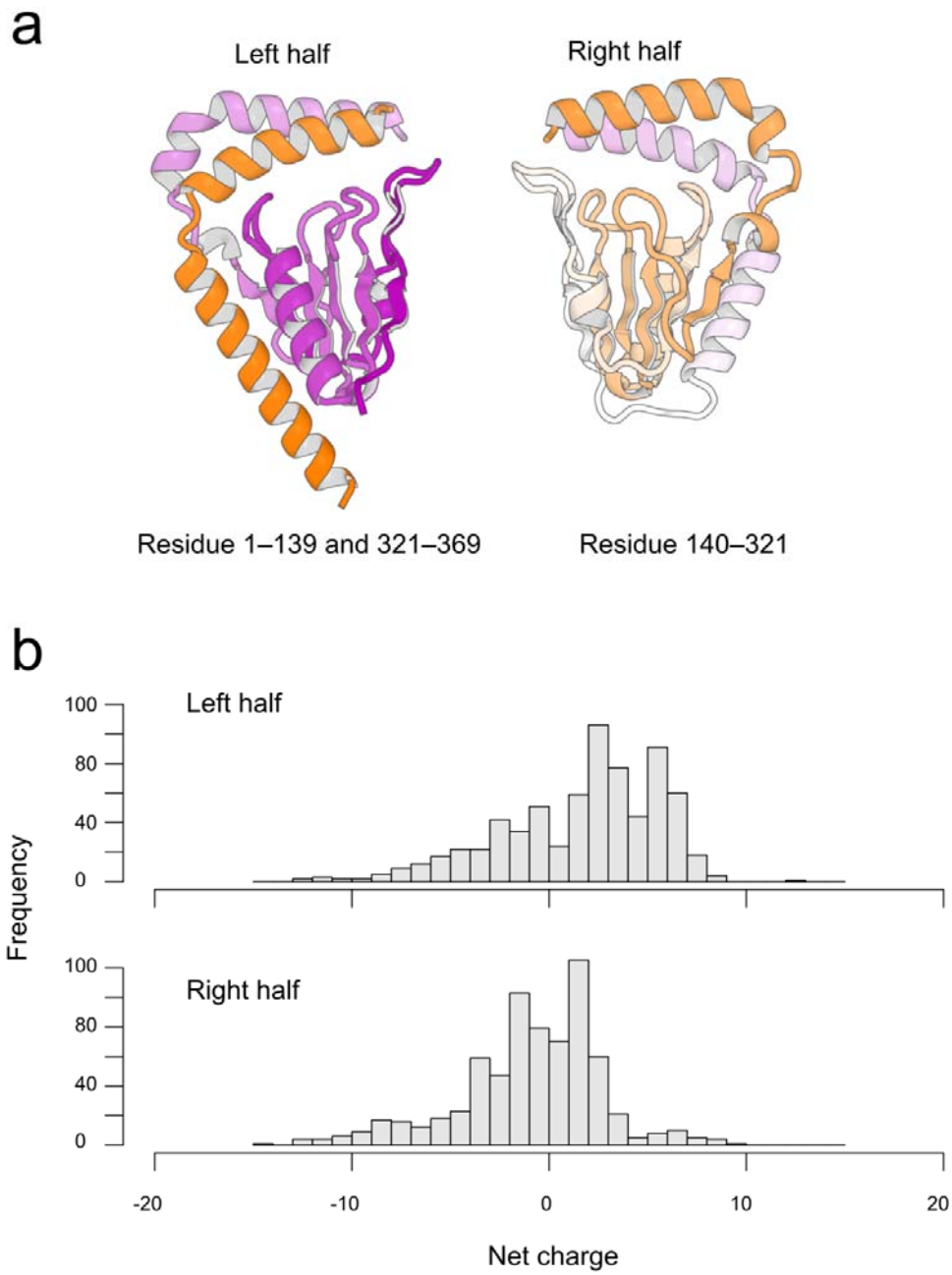

**Extended Data Figure 8: Putative functionally relevant residues and comparison to the active sites of known enzymes.** (a) Overall structure of dwNTPase complexed with two pairs of ATPs, Mg2+, and Zn2+ ions after 100-ns MD simulation. The structure of dwNTPase is shown as white cartoon or grey surface. Mg2+ and Zn2+ are shown as green or grey spheres. ATP is shown as orange sticks. (b) Coordination of metal ions by two aspartate sidechains observed in MD simulations. ATP is shown as orange sticks. Sidechain moieties of relevant residues are shown in stick and colored in CPK scheme. Green and grey spheres represent Mg2+ and Zn2+ ions. P-loop is colored cyan. (c) The active site structure of RNase H (PDB 1zbl). Sidechain of metal coordinating amino-acid residues Asp and Asn are shown as sticks and colored in CPK scheme, where Asn is a mutation from Asp. Mg2+ ions are shown as sphere. The Mg2+ ion occupying the sidechain of Asp and Asn is colored green. Nucleic acid residues that contact the Mg2+ ion are colored orange. (d) Catalytic triad-like sidechain configuration observed during the MD simulations. Triad-like sidechain cluster is circled. Black dotted line indicates the hydrogen bond between sidechain of H92 and D74 detected by HBPLUS. (e) The active site structure of TEV protease (PDB 1lvm). Sidechain of the Cys-His-Asp catalytic triad residues are shown as sticks, colored in CPK scheme, and circled.




**Extended Data Figure 9: Other characteristic residues and substructures found in dwNTPase.** (a) Location of switch loop relative to the tunnel and hydrogen bond formed between conserved lysine and glutamate residues. Switch loop and P-loop are shown as orange and cyan cartoon, and the rest of the structure is shown as grey surface. The sidechain of lysin36 is shown as sticks. (b) Root-mean-square fluctuation (RMSF) of the C-α atom during 20 trajectories of 100 ns MD simulations. Vertical axis represents the RMSF computed relative to the initial conformation and horizontal axis represents the residue index. The positions of switch loops are indicated by orange bars and labels reading sl1/sl2. The number on top of the plot shows the residue index of the starts/ends of switch loops. (c) Conserved residues around P-loops supporting the recognition of ATP molecules. The sidechain atoms of conserved residue recognizing ATPs or Mg2+ ions are shown as sticks. P-loops, ATPs, and Mg2+ ions are colored cyan, orange, and green. Additional loops unobserved in other P-loop proteins are colored grey.



**Extended Data Figure 10: Comparison with periplasmic heme binding proteins.** (Left) A crystal structure of periplasmic heme binding protein HmuT (3nu1) liganded with two heme molecules represented in orange spheres. (Right) Structure of dwNTPase shown in the same scale. The cleft between two P-loop domains found in dwNTPase is indicated by an arrow. The chain is represented by cartoon colored in blue-white-red gradient from N-terminus to C-terminus.




**Extended Data Figure 11: Distribution of net charge in left- and right-half of dwNTPase structure. (a)** The representative structure was divided into left and right halves and used as template to find corresponding halves for other structures. The structures are colored in purple-white-orange gradient. (b) From the initially mined 711 dwNTPase structures, 707 structures with all the secondary structures intact were selected. The structures were aligned to the left half of the representative structure and the aligned regions was assigned as left halves. The rest regions of the structures were assigned as the right halves. In the calculation of net charges, Arg, His, and Lys residues were counted to have +1 charge, and Asp and Glu residues were counted to have -1 charge.