Received: 19 September 2023 | Revised: 13 December 2023 | Accepted: 1 January 2024

DOI: 10.1002/pro.4899

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

Improving enzyme functional annotation by integrating in vitro and in silico approaches: The example of histidinol phosphate phosphatases

Thomas Kinateder | Carina Mayer | Julian Nazet | Reinhard Sterner

Institute of Biophysics and Physical Biochemistry & Regensburg Center for Biochemistry, University of Regensburg, Regensburg, Germany

Correspondence
Reinhard Sterner, Institute of Biophysics and Physical Biochemistry & Regensburg Center for Biochemistry, University of Regensburg, Regensburg, Germany.
Email: reinhard.sterner@ur.de

Funding information
Deutsche Forschungsgemeinschaft

Review Editor: Nir Ben-Tal

Abstract

Advances in sequencing technologies have led to a rapid growth of public protein sequence databases, whereby the fraction of proteins with experimentally verified function continuously decreases. This problem is currently addressed by automated functional annotations with computational tools, which however lack the accuracy of experimental approaches and are susceptible to error propagation. Here, we present an approach that combines the efficiency of functional annotation by in silico methods with the rigor of enzyme characterization in vitro. First, a thorough experimental analysis of a representative enzyme of a group of homologues is performed which includes a focused alanine scan of the active site to determine a fingerprint of function-determining residues. In a second step, this fingerprint is used in combination with a sequence similarity network to identify putative isofunctional enzymes among the homologues. Using this approach in a proof-of-principle study, homologues of the histidinol phosphate phosphatase (HolPase) from *Pseudomonas aeruginosa*, many of which were annotated as phosphoserine phosphatases, were predicted to be HolPases. This functional annotation of the homologues was verified by in vitro testing of several representatives and an analysis of the occurrence of annotated HolPases in the corresponding phylogenetic groups. Moreover, the application of the same approach to the homologues of the HolPase from the archaeon *Nitrosopumilus maritimus*, which is not related to the HolPase from *P. aeruginosa* and was newly discovered in the course of this work, led to the annotation of the putative HolPase from various archaeal species.

KEYWORDS

alanine scan, functional annotation of enzymes, haloacid dehalogenase superfamily, histidinol phosphate phosphatase

1 | INTRODUCTION

In the last decade, the number of protein sequences in public databases has exploded which was mostly caused by advances in sequencing technologies (Katz et al., 2022; Kodama et al., 2012). This trend is exemplified by the universal protein database knowledgebase (UniProtKB)/Translated EMBL Nucleotide Sequence Data Library (TrEMBL) database, which contained around 25 million entries in the year 2012 but almost 250 million entries in 2023 according to the 2023_02 release statistics (https://www.ebi.ac.uk/uniprot/TrEMBLstats). This rapid increase in sequencing data promises to be a valuable resource, for example, for investigators in search for enzymes with novel functionalities (Hon et al., 2020). However, the number of experimentally tested proteins is growing at a much slower rate which is evidenced by the fact that the UniProtKB/Swiss-Prot database that relies on experimental evidence and manual annotation by experts only contains about 570,000 entries which make up a mere 0.2% of the entries from the UniProtKB/TrEMBL database. Consequently, there is an increasing proportion of proteins without experimentally determined function. This can be viewed as a growing problem (Furnham et al., 2009), for example, in the context of proteomics studies where the function of an upregulated or downregulated gene product may help to understand a biological phenomenon. At the same time, improving the annotation also provides a great opportunity to explore uncharted regions in the sequence space and, for example, harness the functionality of enzymes with altered properties or novel functions, which are yet to be discovered in sequence databases. To this end, several automated annotation tools have been developed. The most simplistic approaches to infer for example the function of an uncharacterized enzyme are based on close homology to another enzyme of known function (Abascal & Valencia, 2003; Das & Orengo, 2016). The implementations of this rationale often rely on the hierarchical clustering of enzymes into superfamilies that share the same topology and basic reaction mechanism and families that perform the same reaction (Das & Orengo, 2016). In this way and based on homology, a new query sequence can be assigned to a superfamily and inherit the functional annotation (Das & Orengo, 2016; Furnham et al., 2009). More sophisticated methods additionally rely on information on evolutionary relationship (Engelhardt et al., 2005; Radivojac et al., 2013), genomic context (Enault et al., 2005), or data on the protein structure (Loewenstein et al., 2009). Despite the fact that considerable progress has been made in the field of automated functional annotation, these methods still struggle with the accurate prediction of enzyme function (Furnham et al., 2009) and the proportion of misannotated enzymes even seems to be increasing over time (Radivojac et al., 2013). Superfamilies that contain multiple enzyme families catalyzing different reactions appear to be particularly susceptible to mis-annotation due to their functional diversity (Schnoes et al., 2009). Along these lines, the most commonly observed annotation errors concern over-annotations (Schnoes et al., 2009), which means that the enzyme is assigned to the correct superfamily and overall reaction type but to the wrong family and hence the wrong substrate or the wrong specific reaction. Unsurprisingly, the degree of mis-annotation is not uniform but differs between both protein superfamilies and families. This is evidenced by an analysis of different superfamilies which revealed an average level of misannotation in the UniProtKB/TrEMBL database which ranged from 8% for the terpene cyclase superfamily to 65% for the haloacid dehalogenase superfamily (HAD) (Schnoes et al., 2009). In contrast, in the manually curated UniProtKB/Swiss-Prot database, the misannotation level of the HAD superfamily was limited to 15% (Schnoes et al., 2009). This highlights the quality of expert annotations and demonstrates the importance of biochemical testing as a high percentage of experimentally characterized enzymes in a database clearly increase the overall accuracy of functional annotation.

However, experimental testing is labor intensive and generally includes significant hands-on time. Therefore, we propose a workflow which combines the time and cost efficiency of in silico annotation with the thorough in vitro analysis of one representative from a family of homologous enzymes with unclear annotation status. Then, a focused alanine scan of the active site pocket of this representative is used to establish a fingerprint of the function-determining residues. In the next step, the members of a sequence similarity network (SSN) containing hundreds of homologues of the analyzed enzyme are searched for the occurrence of the characteristic fingerprint. This will allow for the identification of the function of hundreds of enzymes based on the in vitro characterization of a single example.

Due to the previously discussed poor annotation status of proteins from the HAD superfamily, we decided to focus on this superfamily in our study. The HAD superfamily is characterized by a Rossmann fold, which consists of a three-layered αβα-sandwich (Lee et al., 2005; Orengo & Thornton, 2005) (Figure 1a). Members of this superfamily catalyze several related reactions on a vast array of substrates and include phosphatases, phosphonatases, adenosine triphosphatases (ATPases), phosphomutases, and haloalkanoate dehalogenases, which degrade xenobiotics (Aravind et al., 1998; Burroughs et al., 2006; Koonin & Tatusov, 1994). The common catalytic residue

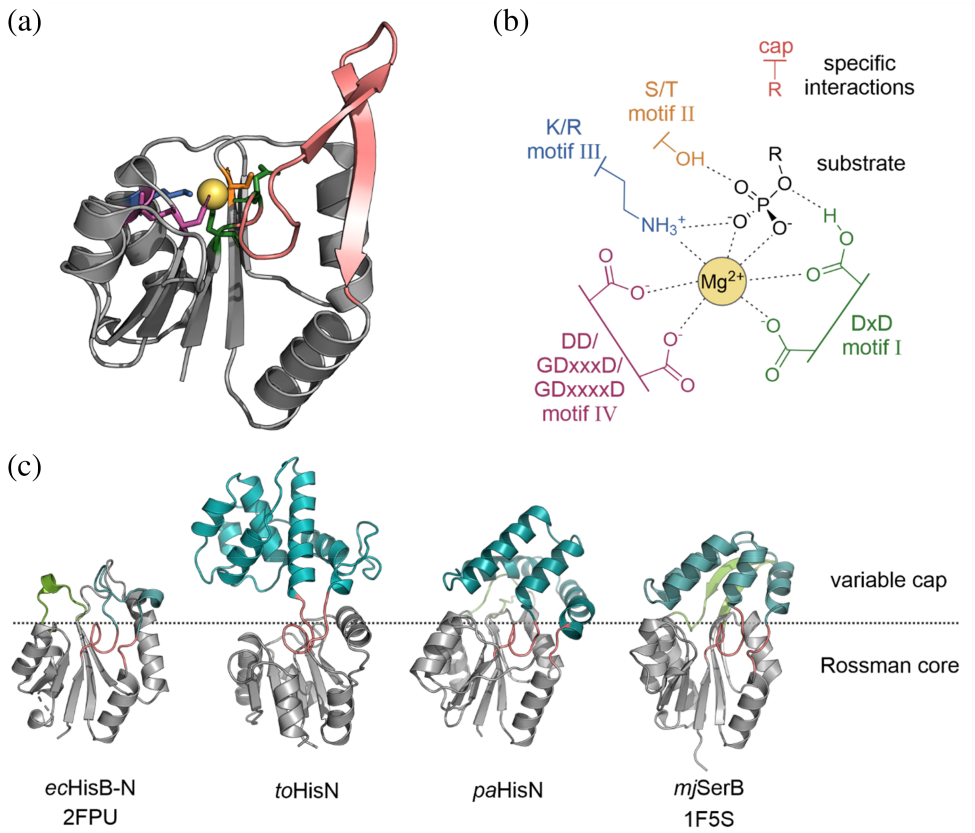

FIGURE 1 Structure and catalytic machinery of haloacid dehalogenase superfamily (HAD) enzymes. (a) Structure of an archetypical HAD enzyme (PDB-ID 1K1E) (Parsons et al., 2002) with a core Rossmann fold (gray) and the characteristic single helical turn and β-hairpin structures (red) but without additional modifications. Conserved active site residues are shown as colored sticks, and a bound bivalent metal ion is represented as yellow sphere. (b) Schematic representation of the four sequence motifs that define the active site of the HAD superfamily phosphatases. Amino acids are represented by the one letter code, whereby *x* denotes for a random amino acid and slashes variations in the sequence motifs. Dashed lines show polar, noncovalent interactions. Motifs I, III, and IV coordinate a bivalent metal ion, usually Mg^{2+}, which mediates substrate binding via polar interactions. Substrate specificity is often achieved by additional residues that are part of a cap structure or the β-hairpin (Burroughs et al., 2006). (c) AlphaFold and crystal structures of representative proteins from the HAD superfamily with a Rossmann core (gray) and variable caps inserted into the β-hairpin (cyan) or before motif III (green) (Burroughs et al., 2006). Shown are the crystal structure of the histidinol phosphate phosphatase (HolPase) from *Escherichia coli* (*ec*HisB-N) (Rangarajan et al., 2006), the AlphaFold predicted structure of the HolPase from *Thermococcus onnurineus* (*to*HisN) (Jumper et al., 2021; Lee et al., 2008), the AlphaFold predicted structure of the HolPase from *Pseudomonas aeruginosa* (*pa*HisN) (Jumper et al., 2021; Wang et al., 2020), and the crystal structure of the phosphoserine phosphatase from *Methanococcus jannaschii* (*mj*SerB) (Wang et al., 2001). Codes below *ec*HisB-N and *mj*SerB give the PDB-ID of the crystal structures.

that enables all these different reactions is a conserved aspartate, which acts as a nucleophile. In phosphatases and phosphomutases, this aspartate forms part of a conserved DxD motif (motif I), whereas in phosphonatases or haloalkanoate dehalogenases, different residues occur two positions downstream of the catalytic aspartate (Burroughs et al., 2006; Collet et al., 1998; Morais et al., 2000). Additionally, there are three other motifs which are characteristic for the HAD superfamily, namely a conserved threonine or serine (motif II), a conserved lysine or arginine (motif III), and two conserved aspartate residues which are part of a DD, a GDxxxD, or a GDxxxxD stretch (motif IV) (Figure 1b). In most enzymes, the aspartate residues of motif I and motif IV coordinate an Mg^{2+} ion, which aids in substrate binding and nucleophilic attack as it stabilizes the additional negative charge of the transition state (Figure S1). This invariable catalytic machinery is complemented by different insertions of variable sizes and structures, termed caps, which decorate the core Rossmann fold. These insertions can typically be found at two different positions in the protein sequence. First, insertions often occur downstream of motif I, where the protein fold exhibits a single helical turn followed by a short β-hairpin structure. Moreover, insertions are regularly detected upstream of motif III. Members of the HAD-superfamily can have an

insertion at one of the two positions or at both positions. The caps cover the active site and thus provide additional residues for specific interactions with different substrates (Baker et al., 1998; Burroughs et al., 2006; Kurihara et al., 1995; Olsen et al., 1988). This modular structure of the HAD enzymes with an invariable catalytic core and function-defining cap in combination with the broad substrate scope and high adaptability (Burroughs et al., 2006; Huang et al., 2015) might be one of the reasons for the poor annotation of the HAD superfamily.

To validate the above-described in vitro–in silico workflow for functional annotation in a proof-of-principle study, we decided to focus on the bona fide histidinol phosphate phosphatase (HolPase) from *Pseudomonas aeruginosa* (*pa*HisN), which is a HAD enzyme that was recently functionally annotated based on metabolic complementation (Wang et al., 2020). HolPases catalyze the dephosphorylation of histidinol phosphate to histidinol in the context of the biosynthesis of histidine (Alifano et al., 1996; Ames, 1957), which consists of 10 reaction steps that are identical in bacteria, plants, and archaea (Stepansky & Leustek, 2006; Winkler & Ramos-Montañez, 2009).

We started our workflow by first characterizing *pa*HisN in vitro and confirming its primary catalytic function as HolPase. Following the generation of a fingerprint of function-determining residues, we identified bona fide isofunctional homologues out of an SSN. The in vitro characterization of several representatives of different sequence clusters from the SSN confirmed our predictions. In an extension of our approach, we searched the Kyoto Encyclopedia of Genes and Genomes (KEGG) database for the occurrence of annotated HolPases. This analysis unveiled that there is a significant knowledge gap regarding the HolPase function, especially in the archaeal superkingdom. For this reason, we set out to find missing HolPases from archaea, which led to the identification of the HolPase from *Nitrosopumilus maritimus* (*nm*HisN), a hitherto unannotated protein. Interestingly, *nm*HisN shared little similarity to previously described HolPases, indicating independent evolution of this enzyme. Applying our workflow on *nm*HisN finally led to the identification of the missing HolPase of numerous archaeal organisms.

2 | RESULTS AND DISCUSSION

2.1 | Characterization of *pa*HisN

The HolPase from *P. aeruginosa* (*pa*HisN) was annotated based on the phenotype of a knockout of the gene PA0335, which showed an incomplete histidine auxotrophy that could be complemented by other HolPase genes or an overexpression of PA0335 (Wang et al., 2020). According to the AlphaFold predicted structure, *pa*HisN belongs to the HAD superfamily, just like the two previously described HolPases: The HolPase from *Escherichia coli* (*ec*HisB-N) and the HolPase from *Thermococcus onnurineus* (*to*HisN). However, the structure of the inserted caps in *pa*HisN differs significantly from both *ec*HisB-N and *to*HisN (Figure 1c). Specifically, the caps in *pa*HisN fold as four helices and an extended turn, whereas in *ec*HisB-N, the corresponding caps fold as a short loop and a slightly longer loop with a conserved zinc binding motif, and in *to*HisN there is only one large insertion which consists of seven helices. Interestingly, the structure of the caps of *pa*HisN closely resembles the canonical cap structures of phosphoserine phosphatases (PSPases, Figure 1c). In line with this finding, the sequence identities and similarities are significantly higher between the PSPase from *Methanococcus jannaschii* (*mj*SerB) and the HolPase *pa*HisN than between the HolPases *ec*HisB-N and *pa*HisN or *to*HisN and *pa*HisN (Table S1). This similarity between the HolPase *pa*HisN and PSPases presents a challenge to the automated annotation of the homologues of *pa*HisN. Indeed, a BLAST search with *pa*HisN against the nonredundant protein database revealed that many of its close homologues with sequence identities ranging from 58.0% to 99.5% were annotated as PSPases, whereas the remaining homologues mostly possessed unspecific functional annotations like "HAD family hydrolase." Similarly, in an SSN of homologues of *pa*HisN, which was retrieved from the UniProtKB and consisted of around 2700 protein sequences, there were approximately 400 sequences classified as PSPases, whereas the remaining sequences were mostly annotated as hydrolases (Figure S2). Furthermore, only one protein possessed an annotation as PSPase according to the UniProtKB/SwissProt database, which suggests higher reliability. However, the validated HolPase activity of *pa*HisN strongly indicates that there are more HolPases among the homologues which are non-specifically annotated as HAD hydrolases. Additionally, the close sequence and structural similarity between *pa*HisN and *mj*SerB suggests that some of these proteins might be erroneously annotated as PSPases while they in fact represent HolPases. In the absence of kinetic data for *pa*HisN, an alternative explanation would of course be that the previously detected HolPase function of *pa*HisN is only a side activity. In this scenario, *pa*HisN itself would be mis-annotated as HolPase while in fact having a different native function. To clarify this issue, we decided to perform a thorough functional characterization of *pa*HisN in vitro.

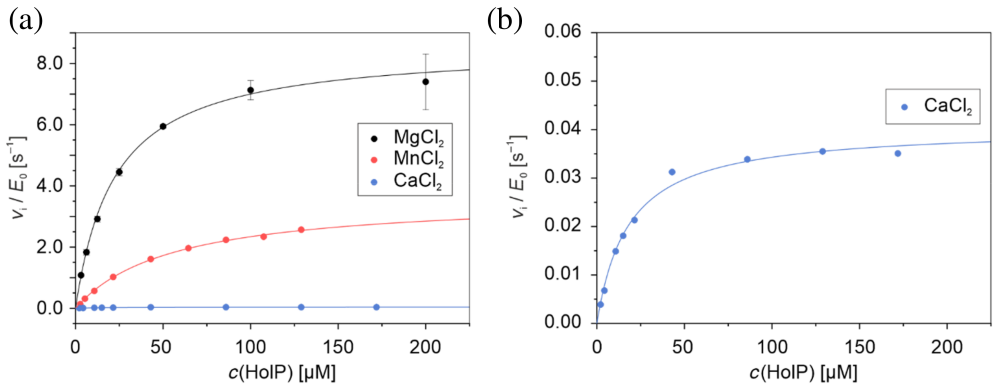

FIGURE 2 Steady-state kinetic characterization of histidinol phosphate (HolP) phosphatase (HolPase) from *Pseudomonas aeruginosa* (*pa*HisN) at 25°C. (a) Substrate saturation curves for the turnover of HolP by *pa*HisN with different metal cofactors. The standard purification was conducted in presence of 5 mM $MgCl_2$. To test other metal ions, several washing steps with EDTA were performed and then, 5 mM $MnCl_2$ or $CaCl_2$ was added, and kinetic assays were conducted with HolP. Measurements with Mg^{2+} were performed as triplicates and measurements with Mn^{2+} as duplicates. (b) Zoomed-in saturation curve for the turnover of HolP by *pa*HisN with Ca^{2+} as cofactor.

TABLE 1 Steady-state kinetic parameters of *pa*HisN at 25°C in the presence of 5 mM of different divalent cofactors.

Substrate	Cofactor	k_{cat} [s^{-1}]	K_M [µM]	k_{cat}/K_M [M^{-1} s^{-1}]
HolP	Mg^{2+}	8.6 ± 0.2	22 ± 1	390,000 ± 20,000
HolP	Mn^{2+}	3.6 ± 0.1	56 ± 3	20,000 ± 3620
HolP	Ca^{2+}	(4.0 ± 0.1) × 10^{-2}	18 ± 2	2200 ± 231
PSer	Mg^{2+}	n.d.	n.d.	3.7 ± 0.03

Abbreviations: HolP, histidinol phosphate; n.d., not detectable; *pa*HisN, histidinol phosphate phosphatase from *Pseudomonas aeruginosa*; PSer, *o*-phospho-L-serine.

For this purpose, *pa*HisN was heterologously produced in *E. coli* and purified by affinity chromatography followed by size exclusion chromatography. The purification of the protein was validated by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), and a biophysical characterization was conducted which consisted of the determination of the oligomerization state by static light scattering, a test of the protein folding by circular dichroism (CD) spectroscopy, and thermal unfolding coupled with CD spectroscopy (Figure S3). The results indicated a highly pure protein that formed a monodisperse solution of a monomer with a melting temperature (T_M) of 46.4°C, which is in line with the growth temperature of *P. aeruginosa* being between 4 and 42°C (LaBauve & Wargo, 2012).

After establishing that *pa*HisN was a well-folded protein of reasonable purity, the enzymatic function was examined. To this end, steady-state kinetic experiments were performed with histidinol phosphate (HolP) using a coupled enzymatic assay (Suárez et al., 2012). The resulting substrate saturation curves are shown in Figure 2.

In the presence of the bona fide native cofactor $MgCl_2$, the catalytic parameters were determined to be 8.6 s^{-1} for k_{cat} and 22 µM for K_M (Table 1) Both values are similar to previously reported values for other HolPases which ranged from 1 to 4 s^{-1} for k_{cat} (Jha et al., 2018; Kinateder et al., 2023; Nourbakhsh et al., 2014; Ruszkowski & Dauter, 2016) and from 32 to 400 µM for K_M (Nourbakhsh et al., 2014; Rangarajan et al., 2006; Ruszkowski & Dauter, 2016; Wiater et al., 1971). Taken together, the catalytic parameters of *pa*HisN correspond to a respectable k_{cat}/K_M value of 400,000 M^{-1} s^{-1}, which confirms that *pa*HisN is indeed a HolPase.

Although Mg^{2+} is the most common cofactor in HAD superfamily proteins (Aravind et al., 1998; Burroughs et al., 2006; Koonin & Tatusov, 1994), it is not the preferred cofactor for every enzyme of this superfamily. For example, *to*HisN—albeit being highly active in presence of Mg^{2+}—showed an even higher k_{cat}/K_M value in presence of Mn^{2+} (Lee et al., 2008). For this reason, we tested the activity of *pa*HisN in the presence of other bivalent metal cofactors. To this end, first the purified protein was washed several times with EDTA-containing buffer. Then, the buffer was exchanged by a buffer with an excess of either $MnCl_2$ or $CaCl_2$. Afterward, steady-state kinetic parameters were determined which revealed that in the presence of Mn^{2+}, the k_{cat} value was decreased by

a factor of 2.3, whereas the K_M value was increased by a factor of 2.7. In the presence of Ca^{2+}, the k_{cat} value was reduced by a factor of 210, whereas the K_M value remained unchanged (Table 1). These results confirm Mg^{2+} as native cofactor of *pa*HisN. The observed deleterious effect of Ca^{2+} was similar to the previously noted effect of Ca^{2+} on the HolPase *ec*HisB-N (Rangarajan et al., 2006) and provides another example for the sensitivity of HAD enzymes for subtle changes in the radius of the metal cofactor.

Because of the structural similarity to PSPases and since many close homologues of *pa*HisN were annotated as PSPases, *o*-phospho-L-serine (PSer) was next tested as substrate. Intriguingly, enzymatic turnover could be observed, and steady-state experiments were conducted. However, substrate saturation could not be achieved (Figure S4A), and a linear fit of the data yielded a k_{cat}/K_M value of 3.7 M^{-1} s^{-1}, meaning that the catalytic efficiency for PSer is more than 100,000-fold lower than for HolP (Table 1). The low promiscuous side activity for PSer could either be the remnant of evolutionary relationship with PSPases or be caused by an inherently limited substrate specificity.

To obtain a more complete picture of the activities of *pa*HisN, the enzyme was additionally tested for the promiscuous turnover of the phosphorylated compounds phosphoribosyl pyrophosphate, glyceraldehyde-3-phosphate (G3P), glycerol-1-phosphate (G1P), fructose-6-phosphate (F6P), ATP, *o*-phospho-D-serine (P-D-Ser), and phosphothreonine. This experiment showed that G3P, G1P, and F6P could also be dephosphorylated by *pa*HisN, whereas P-D-Ser, the enantiomer of PSer, was not dephosphorylated (Figure S4B). This indicates that small substrates are more likely to be accepted, probably due to steric constraints, while at the same time, there seem to be specific requirements for the relative spatial arrangement of charged groups, which allow for the turnover of only one of the two enantiomers of phosphoserine. Taken together, this presents another example of the well-documented substrate promiscuity within the HAD superfamily (Huang et al., 2015; Kuznetsova et al., 2006).

In summary, we could confirm *pa*HisN as HolPase with Mg^{2+} as native cofactor and a low promiscuous side activity for the hydrolysis of several other phosphorylated compounds. The strong preference for HolP over PSer makes it highly unlikely that close homologues of *pa*HisN are PSPases, further supporting our initial hypothesis that many homologues of *pa*HisN were misannotated.

To help with the classification of these homologues, the features that define *pa*HisN as HolPase were analyzed. Specifically, an alanine scan of active site residues was conducted to identify those residues in *pa*HisN that are most important for the specific binding and turnover of HolP. This set of residues should then yield a fingerprint by which other HolPases could be identified.

2.2 | Identification of the functionally relevant residues in *pa*HisN by alanine scanning

Residues that should be exchanged to alanine were selected based on the following considerations: The basic catalytic function of the HAD proteins is conferred by four conserved sequence motifs I–IV which form part of the core Rossmann fold, whereas substrate specificity is mostly mediated via residues from the caps (Baker et al., 1998; Burroughs et al., 2006; Kurihara et al., 1995; Olsen et al., 1988) (Figure 1). To identify residues of the cap of *pa*HisN that are in close proximity to the substrate, the position of bound HolP was approximated by a docking experiment (Figure 3a,b).

An analysis of the structure with docked substrate revealed a total of eight residues that are less than 4 Å away from the carbon or nitrogen atoms of the imidazole ring of HolP. The sidechain of one of the seven residues, namely A112, was pointing away from the substrate, and this residue was therefore not considered further. Consequentially, seven residues, namely D16, D18, Y47, L52, Y57, Q58, and T113, were identified, which we deemed as potential key residues for substrate specificity (Figure 3a, purple dots, and Figure 3b, sticks). Except for T113, these residues were all part of the cap.

Each of the seven positions was individually mutated to alanine, and corresponding proteins were purified in the same manner as the wild-type protein. As judged by SDS-PAGE, all proteins could be obtained with good purity (Figure S5), and CD spectroscopy confirmed proper folding (Figure S6). In the next step, the influence of each mutation on enzyme activity was probed by steady-state kinetic experiments (Figure S7). The determined kinetic parameters are listed in Table 2, and the effect of each mutation is highlighted in Figure 3b,c.

The introduced mutations can be separated into three groups, depending on their effect on the HolPase activity.

The first group contains the D16A and D18A mutations, which both had a strong effect on the HolPase function. Mutation of D16A resulted in a dramatic decrease of the k_{cat} value by four orders of magnitude, whereas the K_M value remained almost unaffected which means that this mutation significantly affected catalysis. Mutation of D18A, however, affected both the k_{cat} and the K_M value, suggesting that both the substrate binding and the formation of a productive enzyme–substrate complex are impaired. The importance of both residues

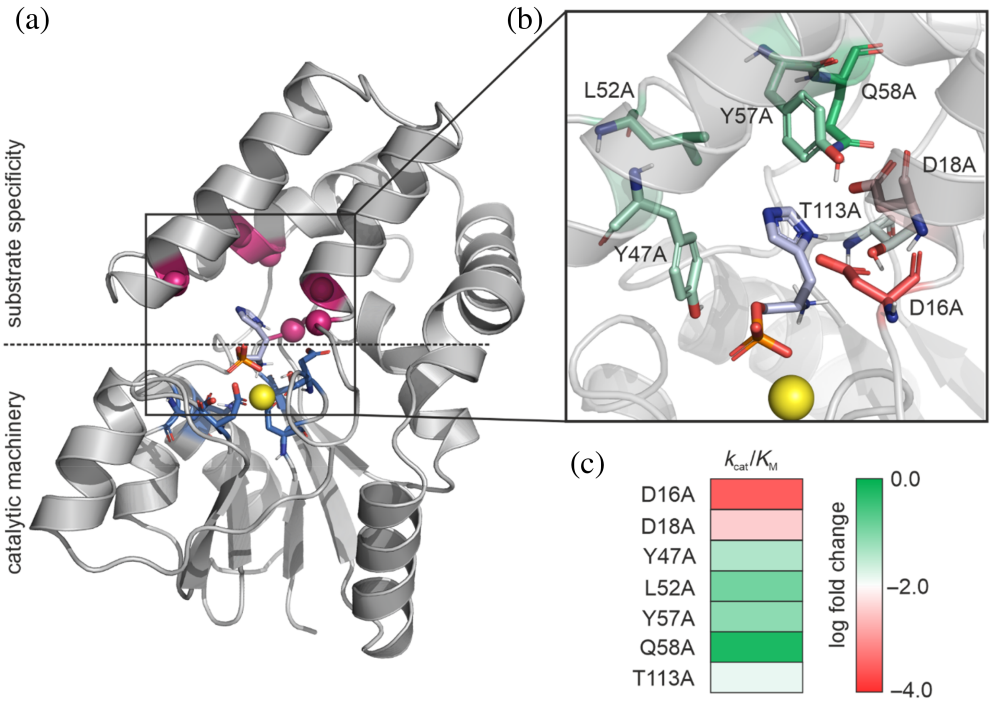

FIGURE 3 Alanine scan of cap residues of histidinol phosphate (HolP) phosphatase (HolPase) from *Pseudomonas aeruginosa* (*pa*HisN). (a) AlphaFold predicted structure of *pa*HisN with docked HolP (light blue) and Mg^{2+} (yellow shere). In HAD superfamily enzymes, the catalytic machinery is provided by conserved residues of the Rossmann core (dark blue sticks), whereas substrate specificity is mostly mediated by residues of the different caps. Seven cap residues (magenta) were individually mutated to alanine and the corresponding mutant enzymes were functionally characterized. (b) Zoomed-in view on the catalytic site of *pa*HisN. Colors of the residues indicate the effect on the k_{cat}/K_M value upon mutation to alanine from weak effects (green) to strong effects (red). (c) Graphic representation of the impact of each point mutation, which relates the log fold change on the k_{cat}/K_M value to a color on the scale from green over white to red.

TABLE 2 Steady-state kinetic parameters at 25°C of *pa*HisN alanine mutants in presence of 5 mM $MgCl_2$.

Variant	k_{cat} [s^{-1}]	K_M [μM]	k_{cat}/K_M [s^{-1} M^{-1}]
wild-type	8.4 ± 0.2	21 ± 2	400,000 ± 39,300
D16A	(3.0 ± 0.1) × 10^{-3}	26 ± 2	115 ± 9.7
D18A	0.50 ± 0.06	365 ± 86	1370 ± 360
Y47A	2.6 ± 0.2	163 ± 27	15,900 ± 2920
L52A	6.2 ± 0.7	129 ± 38	48,100 ± 15,200
Y57A	6.9 ± 0.6	232 ± 40	29,700 ± 5700
Q58A	5.8 ± 0.2	24 ± 3	241,000 ± 31,300
T113A	3.6 ± 0.3	660 ± 84	5460 ± 830

Abbreviation: *pa*HisN, histidinol phosphate phosphatase from *Pseudomonas aeruginosa*.

can be rationalized by the close proximity of the side chains of D16 and D18 to the imidazole ring and the amino group of the substrate, which may allow for direct interaction via hydrogen bonds.

The second group consists of the mutations Y47A, L52A, Y57A, and T113A, which all led to an intermediate decrease in HolPase activity. All four mutations mostly affected the K_M value, which was increased by a factor between 6 and 31, whereas the k_{cat} value was only slightly decreased by up to a factor of 3. According to the docking analysis, only Y47 and T113 are close enough to the substrate for direct interactions with the imidazole moiety, for example, via hydrogen bonds or, in the case of Y47, a π–π interaction. Interestingly, T113 is in close proximity to the crucial residues D16 and D18, which supports the docking analysis, as it seems plausible that mutations of residues that physically interact with the substrate are most detrimental. According to the docking,

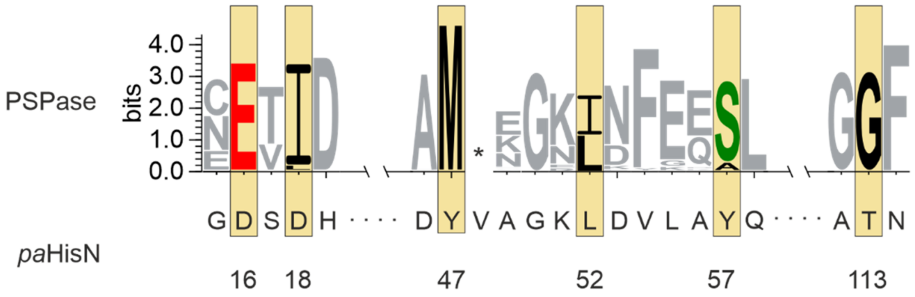

FIGURE 4 Comparison of the HolPase from *Pseudomonas aeruginosa* (*pa*HisN) sequence with a sequence logo of PSPases. The relevant residues of *pa*HisN from the alanine scan (marked by yellow boxes) were compared with the amino acids at the equivalent position in a reference data set of 42 homologues of the PSPase from *Methanococcus jannaschii*. The upper part shows a sequence logo, which is based on a multiple sequence alignment of the PSPases, whereas the lower part gives the sequence of *pa*HisN. Numbers indicate the residue number in *pa*HisN. The most pronounced differences concern the functionally relevant residues D16, D18, and Y47 from *pa*HisN, which correspond to the highly conserved residues E, I, and M in PSPases. The amino acids that are equivalent to the remaining *pa*HisN residues showed weaker conservation in PSPases. Therefore, the simultaneous occurrence of D16, D18, and Y47 provide a fingerprint, which suggests HolPase activity. Points denote for sequence stretches, which are not part of the active site and the asterisk in the sequence logo denotes for an inserted residue in *pa*HisN. Color code: red, acidic; black, hydrophobic; green, hydroxyl group.

the two residues L52 and Y57 are further away from the substrate than Y47 or T113. The distance to the substrate in combination with the moderate effect of the mutations L52A and Y57A, makes immediate interactions with the substrate unlikely. Instead, L52 and Y57 might contribute in an indirect manner to the HolPase activity, for example, by influencing the correct packing of the active site or by filling cavities.

The last group only contains the mutation Q58A, which has a negligible effect on the HolPase function of *pa*HisN and therefore was excluded from further analysis.

The information on the functional relevance of each residue should next be used for a comparison with the corresponding residues in the closest homologues, the PSPases. The goal of this comparison was to find out if the targeted residues are conserved across the two families or if the functionally important amino acids in HolPases and PSPases differ from each other. Based on this comparison, a fingerprint should be established that could classify an enzyme as HolPase and at the same time exclude that it is a PSPase. For this objective, a sequence logo of PSPases was created. To achieve this despite the obvious annotation problem, the well characterized PSPase *mj*SerB (Cho et al., 2001; Wang et al., 2001; Wang et al., 2002) was used as starting point, and homologues of *mj*SerB were retrieved by a BLAST search of the National Center for Biotechnology Information (NCBI) database of nonredundant protein sequences (Altschul et al., 1990). The 42 closest homologues exhibited a sequence identity between 60% and 100%, a query coverage of at least 96%, and were all annotated as PSPases. A threshold of 60% sequence identity to *mj*SerB was chosen to ensure high accuracy in the annotation as PSPase in spite of the potential problem of misannotated sequences (Tian & Skolnick, 2003) and balance it with the need for diversity in the data set (for details see Figure S8). Then, a sequence logo was created and compared with the sequence of *pa*HisN (Figure 4).

To distinguish HolPases from PSPases, we were looking for residues that were (i) different between PSPases and *pa*HisN, (ii) of high relevance for the HolPase function of *pa*HisN, and (iii) highly conserved in PSPases. The rationale for the latter being that a high conservation in PSPases indicated functional relevance of this residue for PSPase function.

The two positions in *pa*HisN at which the mutation to alanine led to the most pronounced drop in HolPase activity were D16 and D18. At the position that corresponds to D16, there is a highly conserved glutamate in the PSPase sequence logo. Although the exchange of an aspartate by a glutamate is rather conservative, the paramount importance of the aspartate for HolPase activity and the strong conservation of glutamate in PSPases suggest that the occurrence of an aspartate at this position could nevertheless be a first criterion by which a HolPase can be identified. At the position corresponding to D18, there is a highly conserved isoleucine in PSPases, which is equivalent to a significant change, in terms of both sterics and electrostatics. This isoleucine in PSPases is followed by a conserved aspartate, which in principle could replace D18. However, both the critical residues D16 and D18 in HolPases and the corresponding E20 and I22 in PSPases form part of a helix. Close inspection of the predicted structure from *pa*HisN and the PSPase reference structure from *mj*SerB showed that the side chains of

D16 and of E20 from PSPases point in the same direction (Figure S9). The same is true for D18 from *pa*HisN and I22 from PSPases, which are both pointing toward the active site. Consequently, the aspartate residue D23 that follows the I22 in PSPases is rotated relative to D18 and points away from the active site, which makes it unlikely that this residue could fulfill the same function as D18. Taken together, the residues D16 and D18 form a DxD motif, which is indicative for a HolPase, whereas an ExI motif indicates a PSPase.

From the group of residues with an intermediate effect on HolPase activity, Y47, L52, Y57, and T113, the residue corresponding to Y47 is strongly conserved within PSPases as a methionine. Due to the functional relevance of Y47 and the high conservation of the rare amino acid methionine, we concluded that this position provides a third criterion to discriminate between HolPases and PSPases. The amino acid at the position equivalent to L52 is however not suited for functional discrimination because in PSPases, this position was often occupied by the same or very similar amino acids as in *pa*HisN, namely isoleucine, leucine, or valine. At the position corresponding to Y57, the conservation within PSPases was rather limited, which is why we suggest that the occurrence of a tyrosine at position 57 is a reliable criterion by itself. At the position corresponding to T113, glycine was the most frequently encountered amino acid in PSPases. Because of the lacking sidechain and due to the fact that glycine is a relatively common amino acid (Bogatyreva et al., 2006; King & Jukes, 1969), we considered this position not conclusive for a PSPase or HolPase function. However, because both Y57 and T113 are still relevant for the HolPase function of *pa*HisN, the occurrence of either of the two residues might nevertheless support the classification of an enzyme as HolPase.

In summary, the sequence motif DxD together with the tyrosine at position 47 are most likely to define a fingerprint, which could classify a protein as HolPase. The additional occurrence of the accessory amino acids Y57 or T113 may support this classification.

2.3 | In silico and in vitro analysis of *pa*HisN homologues

The fingerprint consisting of a DxD motif and tyrosine, which is assumed to be a reliable classification for *pa*HisN-type HolPases should next be utilized to identify other HolPases among the homologues of *pa*HisN and uncover their phylogenetic distribution. For this objective, an SSN was generated with *pa*HisN as query sequence using the Enzyme Function Initiative (EFI)-enzyme similarity tool (Oberg et al., 2023; Zallot et al., 2019) and the UniRef90 as database.

The resulting SSN contained approximately 2800 nodes. To identify clusters of potentially isofunctional proteins, this threshold was increased in a stepwise manner, and at a threshold of about 40%, distinct clusters started to emerge, and at 45.8%, three separate main clusters (designated as cluster I–III) and several smaller clusters without interconnecting edges had formed (Figure 5a).

Cluster I contained the query sequence *pa*HisN and was mainly populated by 1151 unique sequences from β- and γ-Proteobacteria and few sequences from Acidithiobacillia. The sequence similarity to *pa*HisN and the narrow phylogenetic distribution of the species implied that these sequences might be derived from a common ancestor and that the hydrolysis of HolP is their primary function. To further substantiate this hypothesis, a sequence logo was created and compared with both the sequence of *pa*HisN and a consensus sequence of PSPases (Figure 5b). Interestingly, and in line with our hypothesis, the PSPase-defining EIM motif could not be found in any of the sequences from cluster I (Figure S10). A PSPase activity of any protein in this cluster is thus highly unlikely. Conversely, at two of the three fingerprint positions from *pa*HisN, namely D16 and Y47, the sequence logo of cluster I showed the strong conservation of an aspartate, which was found in 96.4% and a tyrosine, which was found in 97.3% of all sequences (Figure S10). The third fingerprint residue, D18, showed weaker conservation but was still encountered at this position in 69.2% of all sequences from cluster I. The complete DDY motif was found in 793 (68.9%) sequences from cluster I. Based on the criteria from the alanine scan, we conclude that all sequences from cluster I that contain all three fingerprint residues are HolPases. For the minor fraction of sequences that lack D18, a HolPase function cannot be deduced with the same degree of certainty. The clustering of these sequences with other HolPases and the co-occurrence of a tyrosine at the position equivalent to Y57 and a threonine at the position equivalent to T113 in approximately half of the sequences from cluster I nevertheless indicated a HolPase function for these proteins.

Cluster II contained sequences from a wide variety of phylogenetic groups such as α-, γ-, δ-Proteobacteria, and Actinomycetes. A close analysis of the sequences of cluster II revealed that many of them were fusion proteins with an N-terminal domain annotated as HAD protein and a C-terminal domain annotated as 1-acyl-*sn*-glycerol-3-phosphate acyltransferase, which may explain the clustering of these sequences (Figure S2). The logo of cluster II showed that of the three fingerprint residues D16, D18,

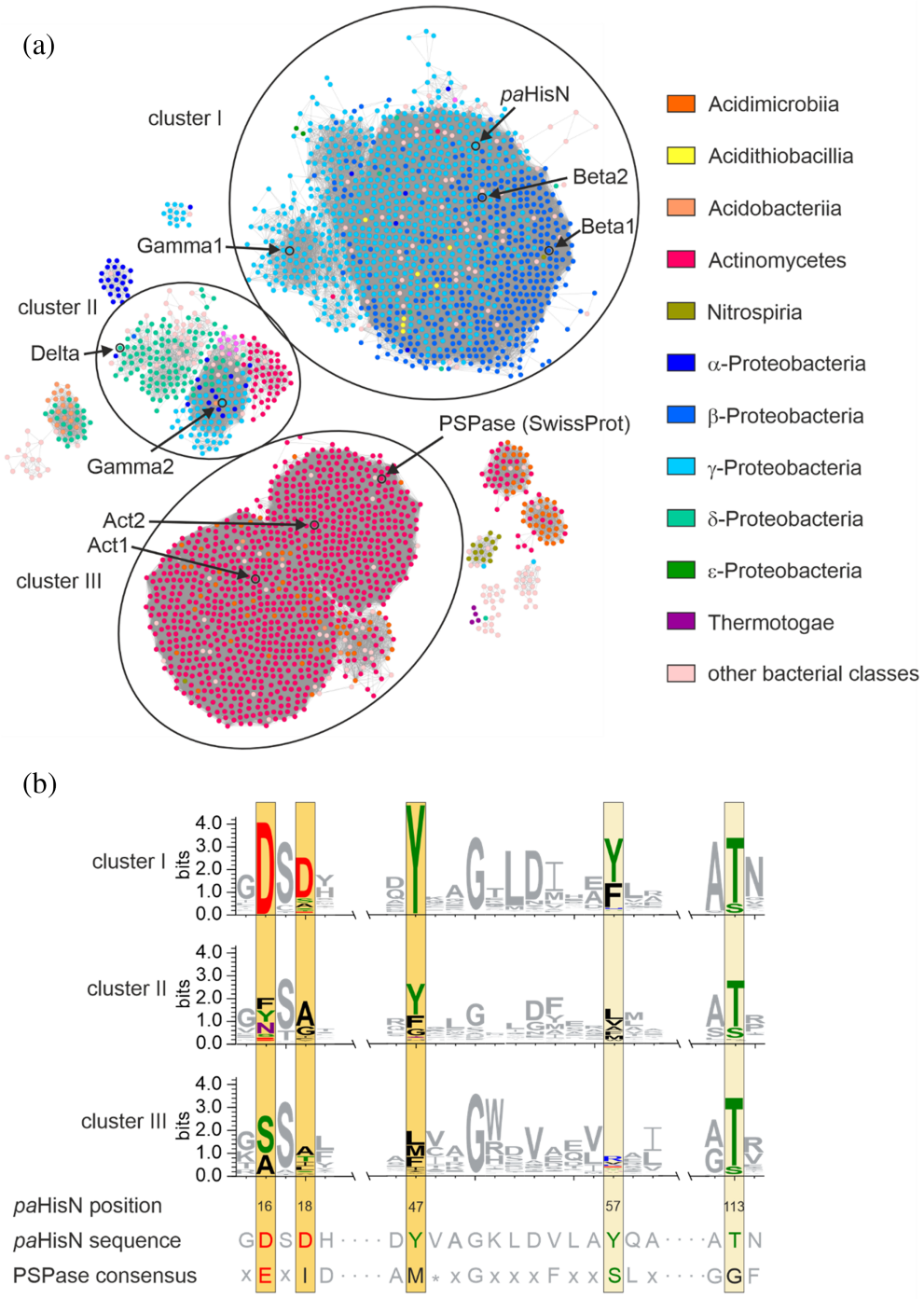

FIGURE 5 In silico analysis of the homologues of histidinol phosphate phosphatase (HolPase) from *Pseudomonas aeruginosa* (*pa*HisN). (a) Sequence similarity network (SSN) of homologues of *pa*HisN. In the SSN, each node represents a homologue of *pa*HisN, and the color indicates the bacterial class to which the corresponding organism belongs. Nodes that share more than 45.8% sequence identity are connected by an edge. At this sequence identity threshold, three major clusters (I, II, and III) can be distinguished. The query sequence is located in cluster I (black arrow), which is mostly composed of β- and γ-proteobacterial sequences. Cluster II contains unannotated sequences from a variety of bacterial classes such as α-, γ-, and δ-Proteobacteria, and Actinomycetes. Cluster III mostly comprises sequences from Actinomycetes and Acidimicrobiia and contains only one sequence with a specific functional assignment as phosphoserine phosphatase (PSPase), according to SwissProt. Sequences marked with an arrow were selected for subsequent in vitro testing. (b) Sequence logos for the three clusters compared to the *pa*HisN sequence and a consensus sequence of PSPases (cf. Figure 4). Fingerprint residues are highlighted in bright yellow, whereas positions that could indicate HolPase function but are not conclusive are highlighted in faint yellow. Sequences of cluster I showed strong conservation of the fingerprint residues, suggesting HolPase function of these sequences. In contrast, in cluster II only Y47 and T113 and in cluster III only T113 showed some degree of conservation. Color code of residues: red, acidic; black, hydrophobic; green, hydroxyl group; purple, amide; blue, basic.

and Y47, only Y47 occurred in a relevant fraction of the sequences. Moreover, from the accessory residues Y57 and T113, only T113 was conserved. Due to the absence of two out of three HolPase-defining residues, a HolPase function could not be deduced. At the same time, the sequence logo of cluster II also does not match the PSPase consensus sequence as neither the ExI motif nor the conserved methionine could be found. Therefore, the sequence logo was also not conclusive regarding a possible PSPase activity.

Cluster III was mostly composed of sequences from Actinomycetes and Acidimicrobiia and contained the only sequence within the SSN with a specific function annotated according to the UniProtKB/Swiss-Prot database. The publication that is cited for this specific functional assignment does, however, not contain any in vitro data for this protein. Instead, the PSPase function was deduced from homology (Arora et al., 2014) which clearly reduces the reliability of this annotation. Moreover, the sequence logo indicated a very low sequence conservation at the three HolPase fingerprint positions and neither of the function-defining amino acids was represented at a relevant frequency. From the accessory residues Y57 and T113, again only T113 was found in a relevant fraction of sequences. Like in the case of cluster II, the sequence logo was therefore inconclusive regarding a HolPase or a PSPase function.

In summary, the in silico analysis strongly indicated that the sequences from cluster I were mostly HolPases, whereas the sequences from clusters II and III fulfill the criteria for a reliable classification neither as HolPase nor as PSPase.

To clarify the primary function of each cluster, several proteins were selected for experimental testing. To achieve a good coverage of the variability within each cluster, we selected sequences from different subclusters of the main clusters I–III (Figures S11 and S12). Additionally, the sequences were selected in such a way that they were representative of the variations of each cluster at the fingerprint positions (Table S2) and did not contain extensive insertions or deletions compared to other sequences from the subcluster. From cluster I, we selected three sequences for experimental testing. The first sequence (Beta 1) possessed all three fingerprint residues and all accessory residues. The second sequence (Beta 2) contained all three fingerprint residues but deviated from *pa*HisN by the mutation Y57F. The corresponding amino acid combination DDYFT at the fingerprint and accessory positions was the second most frequently observed motif in cluster I (Figure S11). The third sequence deviated in the fingerprint residue D18 and instead possessed an alanine at this position (Gamma1). From cluster II, we selected one sequence from the subcluster, which was mostly comprised by sequences from Actinomycetes and γ-Proteobacteria (Gamma 2) and one sequence (Delta) from the subcluster, which was mainly populated by δ-proteobacterial sequences (Figure S12). Gamma2 differed from *pa*HisN in all three fingerprint positions and the two accessory positions, whereas Delta contained the fingerprint residue Y47 and the accessory amino acid T113. In cluster III, the overall conservation at the HolPase-defining positions was very low; therefore, we selected one sequence of each of the two discernible lobes of cluster III (Act1, Act2; Figure S12). Neither Act1 nor Act2 contained any of the fingerprint or of the accessory amino acids. The residues at the HolPase-defining positions of each sequence are given in Table S2, and the pairwise sequence identities of all selected sequences are given in Table S3.

All candidate proteins were produced heterologously in *E. coli* and purified following the same protocol as used for *pa*HisN. The candidate proteins Beta1, Beta2, and Act1 could be readily produced and purified to near homogeneity as judged by SDS-PAGE (Figure S13). Despite a lower expression level, Gamma1 could still be produced with reasonable purity, whereas Act2, Gamma2, and Delta exhibited a low expression level and were mostly found in the insoluble fraction. Therefore, the genes coding for Act2, Gamma2, and Delta were subcloned into a vector coding for an N-terminally fused maltose-binding protein (Rohweder et al., 2018), which should serve as a solubility tag. In this way, Act2 and Delta could be obtained in reasonable pure form (Figure S13), whereas Gamma2 still proofed to be poorly soluble which prevented further characterization. CD spectra that were recorded for each purified variant were indicative of well-folded proteins (Figure S14). To test for HolPase function, each variant was assayed by steady-state kinetic measurements (Figure S15). The determined kinetic parameters are listed in Table 3.

Beta1 had a sequence identity with *pa*HisN of around 48% but contained all three fingerprint and the two accessory residues (Table S2). In line with our predictions, the kinetic parameters of Beta1 for both HolP and PSer were very close to the ones from *pa*HisN, confirming the annotation of Beta1 as HolPase. Beta2 deviated from *pa*HisN in the accessory position Y57 where it possessed a phenylalanine (Table S2). Compared with *pa*HisN, Beta2 possesses a relatively high K_M value of 340 μM but a very similar k_{cat} value. In line with this finding, the Y57A mutation in *pa*HisN led to a significant increase in the K_M value from 21 to 232 μM, whereas the k_{cat} value remained almost unchanged (Table 2), which implies that an exchange of Y57 by a different residue leads to a similar effect, irrespective of the background. Despite the relatively high K_M value of Beta2, both kinetic

TABLE 3 Annotated function, predicted native function from sequence fingerprint, and kinetic parameters at 25°C of the tested homologues of *pa*HisN.

Enzyme (cluster)	Annotation (UniProt)	Predicted function	Substrate	k_{cat} [s^{-1}]	K_M [μM]	k_{cat}/K_M [s^{-1} M^{-1}]
*pa*HisN (I)	-	-	HolP	8.6 ± 0.2	22 ± 1	390,000 ± 20,300
			PSer	n.d.	n.d.	3.7 ± 0.03
Beta1 (I)	HAD family hydrolase	HolPase	HolP	8.4 ± 0.5	21 ± 5	400,000 ± 98,000
			PSer	n.d.	n.d.	9.8 ± 0.3
Beta2 (I)	PSPase	HolPase	HolP	6.2 ± 0.7	340 ± 63	18,300 ± 3960
			PSer	n.d.	n.d.	< 1.0
Gamma1 (I)	HAD family hydrolase	HolPase	HolP	1.1 ± 0.05	69 ± 10	15,500 ± 2420
			PSer	n.d.	n.d.	< 1.0
Delta (II)	HAD family hydrolase	?	HolP	$(1.7 \pm 0.1) \times 10^{-2}$	52 ± 5	327 ± 36
			PSer	n.d.	n.d.	n.d.
Act1 (III)	HAD family hydrolase	?	HolP	$(3.9 \pm 0.1) \times 10^{-2}$	18 ± 1	2167 ± 132
			PSer	n.d.	n.d.	n.d.
Act2 (III)	HAD family hydrolase	?	HolP	$(8.8 \pm 0.2) \times 10^{-2}$	18 ± 1	4889 ± 293
			PSer	n.d.	n.d.	n.d.

Abbreviations: HAD, haloacid dehalogenase superfamily; HolP, histidinol phosphate; HolPase, histidinol phosphate phosphatase; n.d., not detectable; *pa*HisN, HolPase from *Pseudomonas aeruginosa*; PSPase, phosphoserine phosphatase.

parameters were still in the range of previously reported parameters of other HolPases (Jha et al., 2018; Kinateder et al., 2023; Nourbakhsh et al., 2014; Ruszkowski & Dauter, 2016), and the PSPase activity was too low for reliable measurement, implying that the hydrolysis of HolP is its primary function. Gamma1 contains an alanine at the position corresponding to D18 in *pa*HisN (Table S2). Compared with *pa*HisN, Gamma1 exhibits a relatively low k_{cat} value of 1.1 s^{-1}. Strikingly, the equivalent mutation D18A in *pa*HisN also led to a reduced k_{cat} value of 0.5 s^{-1} compared with 8.6 s^{-1} for the wild type (Table 2). This again indicates that the mutational effects are largely independent of the background. As in the case of Beta2, the reduced catalytic activity of Gamma1 compared with *pa*HisN is still in the range of reported values, and the PSPase activity was again too low for reliable measurement, suggesting that Gamma1 is also a native HolPase. The HolPase activity of Beta2 and Gamma1 furthermore indicates that the other proteins from cluster I that deviate from *pa*HisN by the Y57F mutation (which is the most commonly observed mutation in cluster I at position 57) or the D18A mutation (which is the second most commonly observed mutation in cluster I at position 18) are also HolPases.

The protein from cluster II, Delta, shared Y47 and T113 with *pa*HisN (Table S2). The enzyme possessed a k_{cat} value of 0.017 s^{-1}, which is two orders of magnitude lower than normally observed k_{cat} values of other HolPases (Jha et al., 2018; Kinateder et al., 2023; Nourbakhsh et al., 2014; Ruszkowski & Dauter, 2016) and, surprisingly, a K_M value in the low micromolar range. Due to the low k_{cat} value, it seems unlikely that the hydrolysis of HolP proceeds at a relevant rate under native conditions.

The two proteins from cluster III, Act1 and Act2, contain none of the fingerprint or the accessory residues (Table S2). The k_{cat} values of the two enzymes are 0.039 and 0.088 s^{-1}, which is one order of magnitude below the normally observed range for HolPases (Jha et al., 2018; Kinateder et al., 2023; Nourbakhsh et al., 2014; Ruszkowski & Dauter, 2016) but again both enzymes exhibit surprisingly low K_M values of 18 μM. In summary, the absence of the critical fingerprint residues leads to a drastic reduction of the HolPase activities and it, therefore, seems likely that the hydrolysis of HolP is not the primary function of these two enzymes but rather a promiscuous side activity. Nevertheless, because the k_{cat}/K_M values are only 3–8 times lower than in Beta2 or Gamma1 and still 6–15 times higher than for Delta, we cannot rule out that the remaining HolPase activities are relevant under native conditions.

In search for an alternative native function, we tested the turnover of PSer by Delta, Act1, or Act2, but the PSPase activities were too low for reliable measurements even at PSer concentrations of 400–500 μM and enzyme concentrations of 0.9 μM for Act1, 0.5 μM for Act2 and

4.3 μM for Delta. This may be surprising especially for Act1 and Act2, which are part of the same cluster as the SwissProt annotated PSPase; it is however in line with the absence of the ExI motif and the conserved methionine in the sequence logo of the clusters II and III, which we deem indicative of a PSPase. Therefore, the function of Delta remains unclear while a potential HolPase function of Act1 and Act2 is at least improbable. Assuming that those three genes are expressed and code for functional enzymes, further testing of candidate substrates is required to uncover their true function. Alternatively, the three enzymes could also represent pseudogenes, which may have lost their primary HolPase function, for example, because their host organisms adapted to new environments and supply their need for histidine by uptake from external sources.

In summary, the functional assessment of representative proteins by in vitro characterization showed that all enzymes that were predicted as HolPases indeed showed a level of HolPase activity, which is compatible with previous reports, thus proving the reliability of our functional predictions. Moreover, the results showed that the effect of deviations at functionally relevant positions in homologous sequences can be predicted to some extent by the mutational effect in the homologous *pa*HisN. This predictive power is even more remarkable given that the sequence identity between *pa*HisN and the tested homologues of cluster I was only around 30%–50% (Table S3). In absence of the fingerprint, a HolPase activity cannot be ruled out, but a significant reduction of the catalytic activity was observed in all cases, further proving the reliability of our method.

To cross validate the annotation of the proteobacterial sequences from cluster I as HolPases, we performed a bioinformatic analysis to estimate the proportion of organisms from Pseudomonadota (formally denoted as Proteobacteria) which (i) are most likely able to synthesize histidine and (ii) lack an annotated HolPase. To assess the ability of an organism to synthesize histidine, we used the occurrence of the conserved imidazole glycerol phosphate dehydratase (IGPDH) as an indicator and searched these organisms for the co-occurrence of an annotated HolPase. Interestingly, the HolPases form an exception with the histidine biosynthesis in as much as they are not conserved across species. Instead, HolPases from three different protein superfamilies have been identified so far (Figure S16). The HolPases from *E. coli* is represented by the N-terminal part of the bifunctional HisB enzyme and belongs to the HAD superfamily (*ec*HisB-N; Figure 1c) (Brilli & Fani, 2004; Rangarajan et al., 2006), the *to*HisN (Lee et al., 2008), which is monofunctional, also belongs to the HAD superfamily but most likely evolved independently of *ec*HisB-N, the HolPase from *Lactococcus lactis* exhibits a $(\beta\alpha)_7$-barrel fold and belongs to the polymerase and histidinol phosphatase (PHP) superfamily (Ghodge et al., 2013), and the HolPases from *Medicago truncatula* and *Mycobacterium tuberculosis* both adopt the fold of a αβαβα-sandwich and belong to the inositol monophosphatase superfamily (Jha et al., 2018; Ruszkowski & Dauter, 2016). Except for the HolPase from the archaeon *T. onnurineus*, all other types of HolPases are listed in the KEGG database and were included in our search.

In line with our functional annotation of the homologues of *pa*HisN as proteobacterial HolPases, for approximately 50% of the histidine synthesizing Pseudomonadota, no HolPase has been annotated so far (Figure S17). Intrigued by this observation, we expanded this KEGG analysis to other phylogenetic groups which showed that for around 32% of all organisms that were presumed to synthesize histidine, the HolPase was still unknown (Figure S17). This knowledge gap was especially pronounced in the archaeal superkingdom, where an annotated HolPase was missing in two thirds of all histidine-producing organisms. Therefore, we set out to find this missing function and use our proposed methodology to identify isofunctional homologues.

2.3.1 | In vitro characterization of a putative HolPase from the archaeon *N. maritimus*

A thorough literature search reveled that HolPases from the three closely related archaeal organisms *T. onnurineus*, *T. kodakarensis*, and *Pyrococcus furiosus* are known (Lee et al., 2008). A phylogenetic analysis of the remaining *his* genes of these three organisms, however, showed that they have most likely been obtained by horizontal gene transfer from bacterial species (Fondi et al., 2009). Hence, it seemed unlikely that the HolPases of *T. onnurineus*, *T. kodakarensis*, and *P. furiosus* were representative for the HolPases of other archaea. Therefore, we went on to search for other clues for the missing archaeal HolPases. Interestingly, there was a report on an uncharacterized gene located between *hisC* and the *hisB* in Thaumarchaeota, which was predicted to code for a hydrolase (Fondi et al., 2009). An inquiry of the Search tool for the retrieval of interacting genes/proteins (STRING) database confirmed the occurrence of an uncharacterized gene between *hisB* and *hisC*, denoted Nmar_1556, in the Thaumarchaeon *N. maritimus* (Figure S18). Due to the predicted hydrolase function and the location within an operon-like cluster of *his* genes, we deemed this the most promising candidate gene for the missing HolPase.

Close inspection of the AlphaFold predicted structure of the gene product of Nmar_1556 (Figure S19) showed that it belonged to the HAD superfamily like the previously identified HolPases from *E. coli*, *P. aeruginosa*, and *T. onnurineus* (Lee et al., 2008; Rangarajan et al., 2006; Wang et al., 2020). However, the function-defining cap of the *N. maritimus* protein differed significantly from the previously identified HolPases *ec*HisB-N and *pa*HisN (Figure S19). Specifically, in the *N. maritimus* protein, there was only one large insertion, which was predicted as nine helices, whereas *pa*HisN and *ec*HisB-N each contain two insertions, which were much smaller and showed different secondary structures. The remaining HolPase *to*HisN showed some similarities to the *N. maritimus* protein as it also contained only one insertion which was, however, significantly smaller and was predicted to fold as seven helices. What is more, the sequence identity between these two proteins was calculated to be 23.9%, which is too low to conclude that the gene product of Nmar_1556 and *to*HisN fulfill the same function (Todd et al., 2001).

To determine the function of Nmar_1556, the protein was heterologously produced in *E. coli* and purified by affinity chromatography followed by size exclusion chromatography. The purification of the protein was validated by SDS-PAGE, and a biophysical characterization was conducted which consisted of the determination of the oligomerization state by static light scattering, a test of proper protein folding by CD spectroscopy, and thermal unfolding coupled with CD spectroscopy (Figure S20). The results indicated a highly pure protein that forms a monodisperse solution of a monomer and possesses a melting temperature of 37°C. This is in line with the growth temperature of *N. maritimus*, which ranges from 9 to 29°C (Walker et al., 2010).

Subsequently, the HolPase function of Nmar_1556 was confirmed by steady-state kinetic experiments in the presence of Mg^{2+}, which yielded a k_{cat} value of 0.79 s^{-1} and a K_M value of 73 μM (Figure S21A and Table 4), which are both in the range of previously reported values from other HolPases (Jha et al., 2018; Kinateder et al., 2023; Nourbakhsh et al., 2014; Ruszkowski & Dauter, 2016), and the catalytic efficiency k_{cat}/K_M for *nm*HisN was calculated to be 10,822 M^{-1} s^{-1}. The distantly related *to*HisN from *T. onnurineus* was reported to be slightly more active in presence of Mn^{2+} compared with Mg^{2+}; therefore, we replaced $MgCl_2$ by $MnCl_2$ and repeated the kinetic measurements with HolP. Under those conditions, the k_{cat} value was determined to be 0.05 s^{-1}, and the K_M value was 28 μM, which translates to an overall reduction of the catalytic efficiency k_{cat}/K_M by a factor of 6 compared with the measurements with Mg^{2+} (Figure S21B and Table 4).

To test for possible alternative functions, nine phosphorylated compounds of different shapes and sizes were tested as substrates, namely ATP, fructose-6-phosphate, glycerol-1-phosphate, glyceraldehyde-3-phosphate, pyridoxal phosphate, phosphoribosyl pyrophosphate, *o*-phospho-D-serine, *o*-phospho-L-serine, and phosphothreonine. However, none of the substrates was dephosphorylated (Figure S21C), and we hence concluded that the product of Nmar_1556 was the missing HolPase from *N. maritimus*. Consistent with previously used nomenclature, we thus propose to name this protein *nm*HisN. The proven HolPase activity of *nm*HisN together with the structural similarities to *to*HisN suggests that these two proteins might be representative of archaeal HolPases, which underwent significant evolutionary divergence. Due to the absence of any homology in the function-defining cap structures between the two homologous archaeal HolPases *nm*HisN/*to*HisN and the bacterial HolPases *ec*HisB-N and *pa*HisN, these archaeal HolPases seem to have evolved independently of either *ec*HisB-N or *pa*HisN. This means that, although the phosphatase activity was implemented early on in the HAD superfamily, the specific proteins from which the three evolutionary trajectories started were most likely different superfamily members, presumably already decorated with the distinct caps that are found in *nm*HisN, *ec*HisB-N, and *pa*HisN.

To verify our hypothesis that *nm*HisN is representative for a new type of archaeal HolPases, we again applied our method to find isofunctional proteins.

2.3.2 | Identification of the functionally relevant residues in *nm*HisN by alanine scanning

To help with the classification of isofunctional homologues within the Archaea, we applied the same methodology for *nm*HisN as used with *pa*HisN and first

Substrate	Cofactor	k_{cat} [10^{-2} s^{-1}]	K_M [μM]	k_{cat}/K_M [M^{-1} s^{-1}]
HolP	Mg^{2+}	79 ± 6	73 ± 16	10,822 ± 2500
HolP	Mn^{2+}	5.1 ± 0.1	28 ± 1	1821 ± 74

Abbreviations: HolP, histidinol phosphate; HolPase, histidinol phosphate phosphatase; *nm*HisN, HolPase from *Nitrosopumilus maritimus*.

TABLE 4 Steady-state kinetic parameters of *nm*HisN at 25°C in the presence of 5 mM of different divalent cofactors.

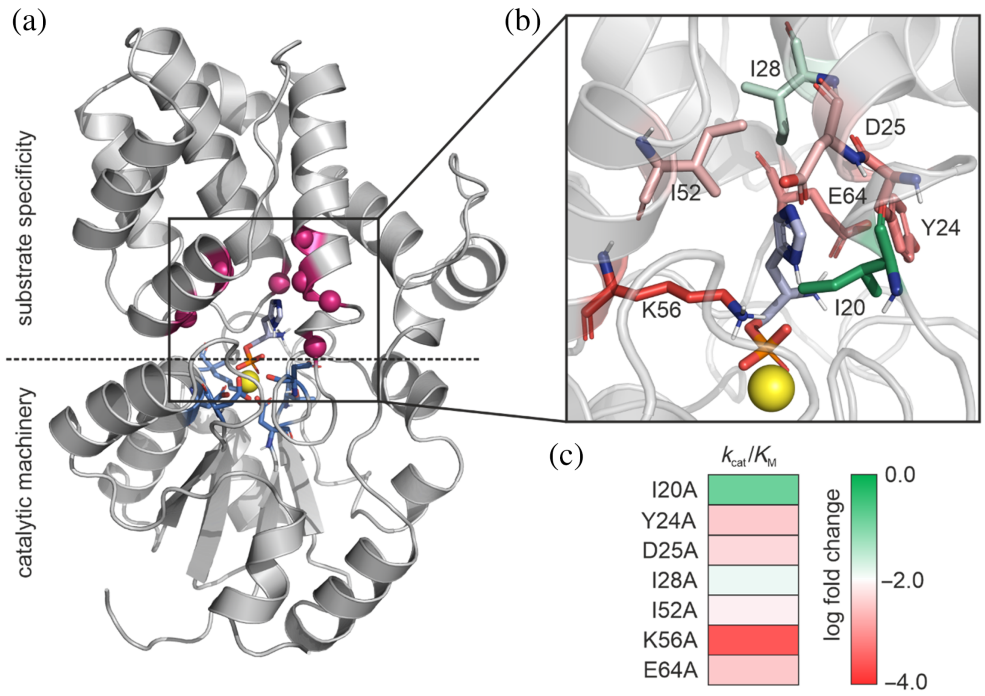

FIGURE 6 Alanine scan of cap residues of homologues of HolPase from *Nitrosopumilus maritimus* (*nm*HisN). (a) AlphaFold predicted structure of *nm*HisN with docked HolP (light blue) and Mg^{2+} (yellow sphere). As generally observed with HAD superfamily enzymes, the catalytic machinery is provided by conserved residues of the Rossmann core (blue sticks), whereas substrate specificity is mostly mediated by the different caps. Seven cap residues (magenta) were individually mutated to alanine, and the corresponding mutant enzymes were functionally characterized. (b) Zoomed-in view on the catalytic site of *nm*HisN. Colors of the residues indicate the effect on the k_{cat}/K_M value upon mutation from weak effects (green) to strong effects (red). (c) Graphic representation of the impact of each point mutation to alanine which relates the log fold change on the k_{cat}/K_M value to a color on the scale from green over white to red.

approximated the position of HolP in the active site by a docking experiment (Figure 6a,b).

Visual inspection of the structure revealed seven residues which are (i) less than 4 Å away from HolP and (ii) the side chains of which were pointed towards the substrate, namely I20, Y24, D25, I28, L52, K56, and E64. These residues all form part of the helical cap structure that covers the active site.

To determine which of the seven residues was critical for substrate binding, all selected positions were individually mutated to alanine and the corresponding proteins were purified in the same way as the wild-type enzyme. As judged by SDS-PAGE, all proteins could be obtained with good purity (Figure S22), and CD spectroscopy confirmed proper folding (Figure S23). Then, the influence of each mutation on enzyme activity was probed by steady-state kinetic experiments (Figure S24). The determined kinetic parameters are listed in Table 5, and the effect of each mutation is highlighted in Figure 6b,c.

The introduced mutations can be separated into four groups, depending on their effect on the HolPase activity.

The first group only contains the mutation K56A which had the strongest effect on the HolPase activity and led to a reduction of the k_{cat} value by more than 460-fold and an increase of the K_M value by more than 8-fold. Possible functions of this residue could be the stabilization of the additional negative charge of the transition state or protonation of the product OH-group. Due to the drastic effect on the HolPase function, the occurrence of this residue is deemed the first criterion to identify a HolPase.

The second group contains the mutations Y24A, D25A, and E64A, which all led to a significant decrease of the k_{cat} value by a factor of between 27 and 360, whereas the K_M was increased only moderately. This importance of the three residues is supported by the substrate docking, according to which, all three residues are positioned in a way that facilitates a polar interaction with the imidazole ring. The severe reduction of the k_{cat} in all three mutants suggests that these residues are also critical for the HolPase function of *nm*HisN.

The third group contains the mutations of the two hydrophobic residues I28 and I52. The mutation of each residue to alanine had similar but moderate effects on the k_{cat} and the K_M values. The hydrophobic nature of these two residues makes a direct interaction with the substrate unlikely; however, they are still relevant for the HolPase activity and could for example aid in the correct orientation of other, more critical residues.

TABLE 5 Steady-state kinetic parameters at 25°C of *nm*HisN alanine mutants in presence of 5 mM $MgCl_2$.

Variant	k_{cat} [10^{-2} s^{-1}]	K_M [μM]	k_{cat}/K_M [s^{-1} M^{-1}]
wild-type	79 ± 6	73 ± 16	10,822 ± 2500
I20A	8.8 ± 0.4	58 ± 7	1520 ± 200
Y24A	1.0 ± 0.1	276 ± 61	36 ± 9
D25A	2.9 ± 0.1	582 ± 28	49 ± 3
I28A	13.2 ± 2.4	910 ± 231	145 ± 45
I52A	4.1 ± 0.8	506 ± 152	82 ± 29
K56A	(1.7 ± 0.2) × 10^{-1}	607 ± 107	2.8 ± 0.6
E64A	(0.2 ± 0.1) × 10^{-1}	63 ± 7	34 ± 4

Abbreviation: *nm*HisN, histidinol phosphate phosphatase from *Nitrosopumilus maritimus*.

The last group only contains the mutation I20A, which only affected the k_{cat} value. Compared with the other mutations, however, this residue seemed to be least important for the HolPase function, which is why we deemed this residue not suited for the classification of a HolPase.

In summary, the alanine scan identified four charged residues, namely Y24, D25, K56, E64, which were most critical for the HolPase function of *nm*HisN. We think that these residues define a fingerprint, which classifies a homologue of *nm*HisN as HolPase. The two hydrophobic residues I28 and I52 were still relevant for HolPase function, and the additional occurrence these two residues may therefore support the classification of a protein as HolPase.

2.4 | In silico analysis of *nm*HisN homologues

To identify groups of HolPases among the homologues of *nm*HisN, an SSN was generated which consisted of 506 homologous sequences. At a sequence identity threshold of 45.0%, two main clusters (designated as clusters I and II) and several small clusters became visible, and 82 sequences were either isolated or formed part of a small cluster consisting of four or less sequences (Figure 7a). These sequences were excluded from further analysis.

Interestingly, the SSN also contained a small, isolated cluster with sequences from Thermococcales. This is the phylogenetic group to which *T. onnurineus*, which contains the previously characterized archaeal HolPase *to*HisN, belongs to. Although *to*HisN itself was not part of the cluster, the sequences of the cluster exhibited a global sequence identity to *to*HisN ranging from 49.4% to 71.7%, indicating high structural similarity. Because the HolPase activity of *to*HisN was previously established by in vitro experiments (Lee et al., 2008) and closely related sequences were included in the SSN, the *to*HisN sequence was utilized as second functional reference together with *nm*HisN.

The sequence of *nm*HisN was located in cluster I, which contained 154 unique sequences mainly from Thaumarchaeota, the phylogenetic group to which *N. maritimus* belongs, and from the candidate phylum Bathyarchaeota. These two phylogenetic groups are related as they both belong to the so called TACK-Superphylum (Castelle & Banfield, 2018). This relationship likely explains the clustering of proteins from Thaumarchaeota and Candidatus Bathyarchaeota. In the sequence logo of cluster I, at the position corresponding to Y24, there was a strictly conserved tyrosine that was found in 98.1% of all sequences from this cluster (Figure S25). The remaining fingerprint positions D25, K56, and E64 still showed reasonable conservation of an aspartate which was found in 65.6%, a lysine which was found in 63.6%, and a glutamate which was found in 53.9% of all sequences from cluster I. Interestingly, *to*HisN exhibits two conservative exchanges relative to *nm*HisN, namely K56R and E64D, indicating that these residues are compatible with a HolPase function. The two variations R56 and D64 were also found in 32.3% and 45.8% of the sequences from cluster I. Along these lines, the positions corresponding to the two accessory residues I28 and I52 were mostly occupied by the hydrophobic residues valine, leucine, or isoleucine. Overall, a total of 65.6% of cluster I possess the fingerprint from *nm*HisN (YDKE, 50.0%), the fingerprint from *to*HisN (YDRD, 14.3%), or a combination of the two fingerprints (YDKD, 1.2%, Figure S25). These results strongly suggest that these 65.6% of sequences from cluster I represent HolPases. The remaining sequences mostly contain the variation D25N, but it cannot be decided at this point if this mutation is compatible with a HolPase function.

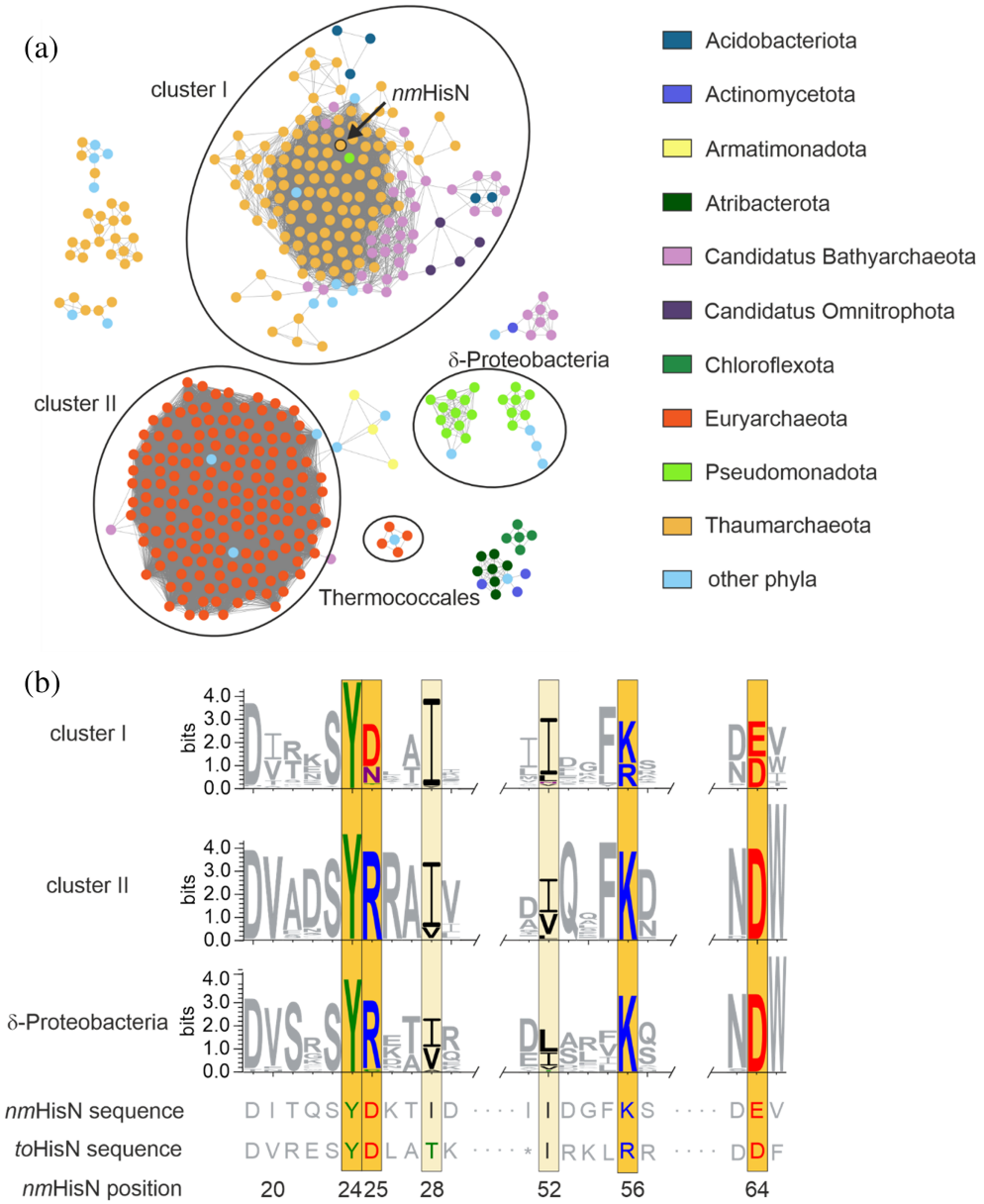

FIGURE 7 In silico analysis of the homologues of HolPase from *Nitrosopumilus maritimus* (*nm*HisN). (a) Sequence similarity network of homologues of *nm*HisN. In the sequence similarity network, each node represents a homologue of *nm*HisN, and the color indicates the phylum to which the corresponding organism belongs. Nodes that share more than 45.0% sequence identity are connected by an edge. At this sequence identity threshold, two major clusters (I and II) can be distinguished. The query sequence is located in cluster I (black arrow), which mostly contains sequences from Thaumarchaeota and the candidate phylum Bathyarchaeota. Cluster II comprises only sequences from Euryarchaeota. Besides, there is a small cluster of sequences from the order of Thermococcales and two small clusters of sequences from δ-Proteobacteria. (b) Sequence logos for the sequences of the two main clusters and the δ-proteobacterial sequences compared with the sequences of *nm*HisN and HolPase from *Thermococcus onnurineus* (*to*HisN). Fingerprint residues are highlighted in bright yellow, whereas positions that could indicate HolPase function but are not conclusive are highlighted in faint yellow. Except for position 25, where cluster II and the δ-proteobacterial sequences exhibit a conserved arginine instead of an aspartate, there is a broad consensus at the positions from the alanine scan between all sequence logos and the two experimentally verified HolPases *nm*HisN and *to*HisN. Color code: red, acidic; black, hydrophobic; green, hydroxyl group; purple, amide; blue, basic.

Cluster II mainly contained 195 nonredundant sequences from Euryarchaeota. Interestingly, the sequences from the phylogenetic order of the Thermococcales, which belong to the phylum Euryarchaeota (Schoch et al., 2020), were not contained in cluster II but formed an independent cluster, which suggests that the enzymes from the Thermococcales had diverged significantly from their homologues from other Euryarchaeota.

In the sequence logo, the residues at the fingerprint positions Y24, K56, and D62 were strictly conserved, indicating HolPase function (for details see Figure S25B). However, a highly conserved arginine was found at the position corresponding to D25, which puts the HolPase function of this cluster into question. At the two accessory positions I28 and I52, isoleucine was the most frequently observed amino acid, occasionally replaced by valine or leucine, again indicating a HolPase function. Taken together, the conservation at five out of six positions supports the classification as HolPases, even though this conclusion is less reliable than in the case of cluster I.

Interestingly, the SSN also contains two small clusters with a total of 22 sequences, which are mostly populated by sequences from δ-Proteobacteria. According to a comprehensive sequence logo of these two clusters, the contained sequences closely resemble the sequences from cluster II (for details see Figure S25C). Specifically, all fingerprint positions show strong conservation and all—except for one—are identical to either *nm*HisN or *to*HisN. The only exception is again a conserved arginine at the position equivalent to D25 in *nm*HisN. Following the same line of argument as for cluster II, these proteins are likely HolPases.

Despite the fact that similar residues are included in the HolPase-defining fingerprints of *pa*HisN and *nm*HisN, it is of note that the fingerprint motifs differ in their order on the peptide chain (D16-D18-Y47 in *pa*HisN and Y24-D25-K56-E64 in *nm*HisN) and in their relative spatial arrangement (Figure S26). This observation further supports the conclusion from structural and sequential comparisons that the two enzymes evolved independently from one another, despite the fact that both proteins belong to the HAD superfamily. This in turn again highlights the great evolvability of this superfamily toward new substrates (Huang et al., 2015), which is most likely due to the modular structure consisting of a core fold which provides the residues which are required for the chemical reaction and the necessary stability to tolerate various cap insertions which can be adapted to various substrates (Baker et al., 1998; Burroughs et al., 2006; Kurihara et al., 1995; Olsen et al., 1988).

3 | CONCLUSION

In the present study, we tackle the annotation problem by combining functional in vitro and in silico methods. This four-step approach is summarized in Figure 8.

To prove the feasibility of this approach, we first analyzed the HolPase *pa*HisN from *P. aeruginosa*, which is very similar to PSPases both in terms of sequence similarity and regarding the protein fold. This in turn leads to an unclear annotation status of many homologues. Experimental characterization of the wild-type enzyme confirmed the HolPase activity and unveiled Mg^{2+} as the preferred metal cofactor. A subsequent alanine scan combined with a comparison of all functionally relevant residues to a sequence logo of PSPases revealed a fingerprint consisting of the three residues D16, D18, and Y47, which we deemed most suited for the classification of a homologue as HolPase. Additionally, the two residues Y57 and T113 were also found to be relevant for the HolPase function, and the occurrence of these two residues was therefore considered an accessory criterion, which supports the classification of a protein as HolPase. An SSN with *pa*HisN as query revealed a cluster of 1151 homologues from β- and γ-Proteobacteria, 793 of which contained the fingerprint and for which we therefore deduced a HolPase function. These predictions could be confirmed by experimental testing of a set of six representative proteins. An additional analysis of the KEGG database showed that in 90% of the histidine synthesizing β-Proteobacteria and 25% of the histidine synthesizing γ-Proteobacteria, a HolPase was indeed missing, further corroborating our function annotation.

In the second part of this work, we identified the missing HolPase from *N. maritimus*, which was dubbed *nm*HisN. Then, we applied our workflow to the homologues of this newly identified enzyme and established a fingerprint consisting of the four residues Y24, D25, K/R56, and D/E64 that were highly relevant for the HolPase activity and the two accessory residues I28 and I52 which were of intermediate functional relevance. An SSN revealed a phylogenetically diverse cluster of 154 homologues from the phylogenetic groups of Thaumarchaeota and Candidatus Bathyarchaeota, 101 of which contained the fingerprint and most likely represent HolPases. Moreover, homologues with one amino acid exchange at a fingerprint position were found in Euryarchaeota and δ-Proteobacteria. Due to the conservation of most of the HolPase-defining residues, it seems likely that these sequences are also HolPases.

In conclusion, our proposed workflow focusses the attention onto those parts of the protein that are most critical for the enzyme function and therefore most informative when it comes to the functional annotation of an enzyme. The experimental effort of this approach is limited as, for example, in our case, the information gathered from seven mutants sufficed to deduce a fingerprint of function-defining residues and predict the most likely function of almost 800 β- and γ-proteobacterial sequences, which contained this fingerprint. At the same time the rate of erroneous functional annotation following our approach is presumed

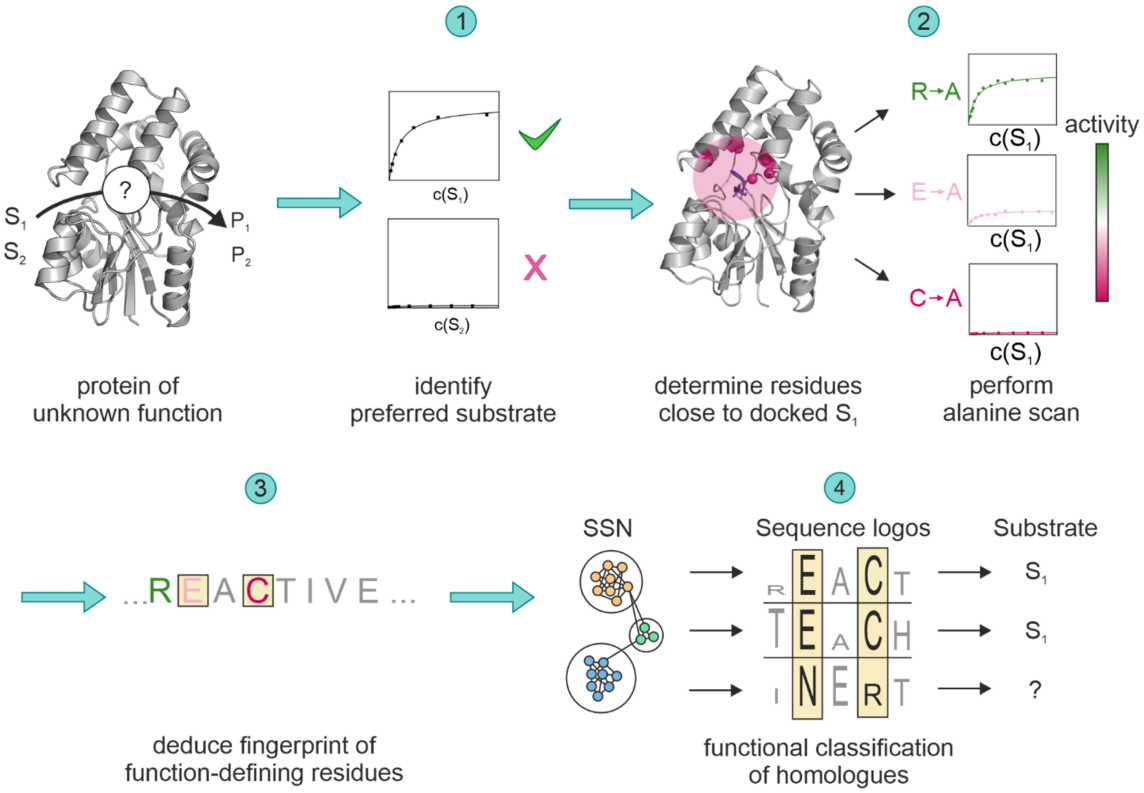

FIGURE 8 Applied workflow for optimizing the functional annotation of protein family members. In the first step, the native function of a representative protein from the family is determined by in vitro measurement of the turnover of several candidate substrates (S_1, S_2) to their corresponding products (P_1, P_2). In the second step, the preferred substrate is docked to the active site and residues (Cα-atoms as red dots), which are located within a certain cutoff distance to the substrate (transparent red sphere), are mutated to alanine. The importance of each residue is then determined by in vitro measurement of the enzymatic activity of each mutant. In the third step, a "fingerprint" of function-defining residues is deduced from the experimental data. In the fourth step, the occurrence of the fingerprint is used to identify groups of isofunctional homologues within the protein family. SSN, sequence similarity network.

to be very low, an assessment that is supported by the successful experimental validation of the homologues of *pa*HisN.

We believe that our workflow may be especially helpful in the case of enzymes that are hard to distinguish because they share certain sequential or structural features but still catalyze different reactions. The approach also promises to be applicable in the case of homologues that are hard to handle as the functional properties can be derived from a reference protein. On a more general note, predictive algorithms tend to suffer from over-annotation, which means that normally, the general reaction type—like for example oxidation or hydrolysis—is correctly identified, but the specific reaction is wrong (Schnoes et al., 2009). This trend can also be observed in our data set, where many enzymes were over-annotated as PSPases. In contrast, if little variation in the decisive fingerprint residues is allowed, our approach can be easily tuned to be rather restrictive or even under-annotate proteins, which ultimately will lead to a highly accurate annotation.

4 | MATERIALS AND METHODS

4.1 | Cloning

To produce target proteins heterologously in *E. coli*, the corresponding genes were codon optimized, and *BsaI* cleavage sites were added. The genes were then subcloned into isopropyl-β-D-1-thiogalactopyranoside (IPTG)-inducible expression plasmids, which coded for N-terminal His_6-tags. In the case of Act2 and Gamma2, the genes were subcloned in a pMal_BsaI plasmid, which coded for an N-terminal His_6-tag followed by an N-terminal maltose-binding protein (pMal_BsaI) (Rohweder et al., 2018).

4.2 | Protein purification

Genes encoding target proteins were expressed in a Δ*his*B knockout strain that was derived from *E. coli* strain BW25113. Transformed cells were grown at 37°C to an

OD_{600} of 0.6–0.8. Then, gene expression was induced by the addition of IPTG (0.5 mM), and cells were grown over night at 20°C. Afterward, cells were harvested by centrifugation (relative centrifugal force (rcf) 4000*g*, 20 min, 4°C), resuspended in buffer (50 mM Tris/HCl, pH 7.5, 400 mM NaCl, 2 mM $MgCl_2$, 10 mM imidazole), and disrupted by sonication (Heinemann Branson sonifier, 60% amplitude, 2 min 30 s sonication time, 2 s on, 2 s off). Cell debris was removed by centrifugation (rcf 23,700*g*, 40–45 min, 4°C), and the target proteins were purified from the soluble fraction by affinity chromatography (HisTrap FF crude column, 5 mL, GE Healthcare; buffer: 50 mM Tris/HCl, pH 7.5, 400 mM NaCl, 2 mM $MgCl_2$ with linear gradient of imidazole from 10 to 500 mM) and subsequent size exclusion chromatography (HiLoad™ 26/600 Superdex™ S75 or S200, GE Healthcare, buffer: 50 mM Tris/HCl, pH 7.5, 400 mM NaCl, 2 mM $MgCl_2$). Purified proteins were dripped into liquid nitrogen and stored at −70°C. Protein purity was assessed by SDS-PAGE, and protein concentrations were determined by absorbance spectroscopy using molar extinction coefficients at 280 nm that were calculated with the ProtParam tool (Wilkins et al., 1999).

4.3 | Cofactor exchange

To exchange the Mg^{2+} cofactor for a different metal ion, the protein solution (in 50 mM Tris/HCl, 400 mM NaCl, 2 mM $MgCl_2$) was added to a centrifugal filter tube (Amicon Ultra) with a molecular weight cutoff of 10 kDa. The protein solution was then concentrated by centrifugation by a factor of around 4 and diluted again by addition of EDTA-containing buffer (50 mM Tris/HCl, pH 7.5, 400 mM NaCl, 5 mM EDTA). This procedure was repeated three times. Then, the same procedure was repeated at least three times with buffer containing the new metal cofactor (50 mM Tris/HCl, pH 7.5, 400 mM NaCl, 5 mM $MnCl_2$ or $CaCl_2$).

4.4 | Steady-state enzyme kinetics

Enzyme activities were measured under steady-state conditions using a coupled photometric assay, which allowed for the continuous measurement of the released phosphate (Suárez et al., 2012). The reaction mixture contained 100 mM Tris/HCl (pH 7.8), 5 mM $MgCl_2$, 0.5 mM inosine, 0.25 U/mL purine nucleoside phosphorylase (Sigma-Aldrich), 2.5 U/mL xanthine oxidase (Sigma-Aldrich), and variable concentrations of the tested phosphorylated substrates. The reactions were started by addition of the respective enzyme. Kinetic parameters were calculated from the initial ascending slopes by curve fitting with the Michaelis–Menten equation (Origin 2019 and Origin 2021, OriginLab).

4.5 | Circular dichroism spectroscopy

Structural integrity of the proteins was assessed by far-UV CD spectroscopy. Prior to measurements, buffer was exchanged to 20 mM potassium phosphate (pH 7.5), to avoid absorption by Tris in the far UV region. Spectra were then recorded with a CD spectrophotometer (J-815; JASCO) between 260 and 190 nm using a quartz cuvette (0.2–1 mm) and measuring five to eight replicas. All spectra were corrected for buffer absorption and smoothed using the Savitzky–Golay algorithm (Arnold et al., 2003) with a convolution width of 7. The mean molar ellipticity per residue θ_{MRW} (deg cm^2 $dmol^{-1}$) was calculated from the observed ellipticity θ_{obs} (mdeg), the width of the cuvette d (cm), the protein concentration c (μM), and number of residues N_A, according to the following equation:

$$\theta_{MRW} = \frac{\theta_{obs} \times 10^5}{c \times d \times N_A}.$$

The melting temperature was then measured by monitoring the CD signal at 220 nm while heating the sample at a constant rate of 1°C per minute. Constant ellipticity values of the folded and unfolded protein were set to 0.0 and 1.0, respectively, and the remaining data were normalized accordingly, which gave the fraction of unfolded protein for each temperature. To obtain the melting temperature T_M, the fraction of unfolded protein f_u plotted against the temperature t was fit with a logistic fit:

$$f_u = \frac{1}{1 + \left(\frac{t}{T_M}\right)^p}.$$

4.6 | Static light scattering

Size exclusion chromatography followed by static light scattering was used to estimate the molecular weights of proteins and deduce the oligomerization state of a target protein. This was realized with an ÄKTA micro system with a Superdex 75 10/300 GL column and an ALIAS autosampler (Spark Holland) coupled to a Viscotec TDA 305 detector (Malvern). All samples were diluted to a final concentration of 50 μM, and measurements were performed in 50 mM Tris/HCl, pH 7.5, 400 mM NaCl, and 2 mM $MgCl_2$. As a reference, bovine serum albumin

(BSA, 50 μM) was used. Changes in small-angle light scattering, right-angle light scattering, and in the refractive index were recorded and analyzed using the OmniSec software. The light scattering signal was calibrated using the exact concentration of the BSA reference as calculated from the changes in the refractive index and the molecular weight of BSA.

4.7 | Sequence similarity networks

SSNs were used to find homologous enzymes, to explore their phylogenetic distribution, and to cluster closely related enzymes using the enzyme similarity tool on the EFI website (Zallot et al., 2019). Specifically, *pa*HisN or *nm*HisN were used as query sequences to conduct a BLAST (Altschul et al., 1990) search for homologues in the UniRef90 database (Steinegger & Söding, 2018). In the UniRef90, any sequences that show a sequence identity of more than 90% are clustered and represented by only one sequence from the cluster. For the BLAST search, the default *E*-value of 10^{-5} was used, and the maximum number of sequences was set to 5000. Taxonomical filters were not applied. For the initial calculation of the edges of the SSN, the default *E*-value of 10^{-5} was utilized. Then, the full SSN was downloaded in each case and further analyzed using Cytoscape (Version 3.9.1). To discriminate between clusters of sequences from different phylogenetic groups, nodes were colored according to the phylogenetic class (for *pa*HisN) or phylum (for *nm*HisN) of the corresponding organism, and the sequence identity threshold for two sequences to be connected by an edge was increased in a stepwise manner until distinct clusters had formed. Isolated nodes without a connecting edge to another node were removed from the network after each step.

4.8 | Sequence logos

Sequence logos were used to analyze the conservation at function defining residues. In the case of the homologues of *pa*HisN and *nm*HisN, the amino acid sequences for each cluster were downloaded from the UniProt database, and a multiple sequence alignment (MSA) was generated using AliView (Larsson, 2014). In the case of the sequence logo of PSPases, a BLAST search (Altschul et al., 1990) of the nonredundant protein sequence database was performed with *mj*SerB as query sequence. For the search, default parameter settings were used. The sequences of the 42 closest homologues which showed a sequence identity level of 60%–100% were downloaded, and an MSA was created using AliView. Based on the MSA, a sequence logo was generated using WebLogo 3.7.12 (Crooks et al., 2004; Schneider & Stephens, 1990). The sequence logos were created by defining the input sequences as amino acid, using the chemistry color code, and showing no error bars. Scaling of the *y*-axis was set to default, which corresponds to information bits according to Shannon entropy.

4.9 | KEGG database analysis

The orthology database KEGG orthology database (KO) of the KEGG was used to identify histidine-producing organisms that lack an annotated HolPase. Four different groups of HolPases were previously described and annotated in the KO, namely the orthologs of the *E. coli* HolPase from the HAD superfamily (K01089), the orthologs of the *L. lactis* HolPase from the PHP superfamily (K04486), the orthologs of the *M. truncatula* HolPase (K18649), and the orthologs of the *M. tuberculosis* HolPase (K05602). We suspected that this list might not be comprehensive due to unidentified HolPases. Therefore, we used the occurrence of the conserved IGPDH which catalyzes the sixth step of histidine biosynthesis (Del Duca et al., 2020) (K01693) as a criterion to identify organisms that are most likely capable of histidine biosynthesis. By comparing the list of all organisms with an annotated IGPDH with the list of all organisms with an annotated HolPase, we could identify those organisms that lacked an annotated HolPase. These organisms could either lack a HolPase or the HolPase could be yet to be discovered.

4.10 | Docking of HolP

To approximate the position of the substrate in the binding pocket and identify those residues, which might be important for substrate specificity, the substrate was docked to *pa*HisN and *nm*HisN, respectively. The docking of HolP was conducted using VINA Auto Dock (Trott & Olson, 2010) as implemented in YASARA, utilizing the YASARA2 forcefield (Krieger & Vriend, 2014). The active site residues were kept flexible, whereas the remaining residues were fixed in place. Then, different substrate conformations were docked 100 times to the active site. The results were sorted by their estimated binding energy. The conformation with the best score for the binding energy was then utilized for further analysis.

AUTHOR CONTRIBUTIONS

Reinhard Sterner: Conceptualization; writing – review and editing; funding acquisition; supervision; validation. **Thomas Kinateder:** Conceptualization; investigation; writing – original draft; methodology; visualization. **Carina Mayer:** Investigation; methodology; visualization. **Julian Nazet:** Investigation; methodology; formal analysis.

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

ORCID

Reinhard Sterner https://orcid.org/0000-0001-8177-8460

REFERENCES

Abascal F, Valencia A. Automatic annotation of protein function based on family identification. Proteins. 2003;53(3):683–92.

Alifano P, Fani R, Liò P, Lazcano A, Bazzicalupo M, Carlomagno MS, et al. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev. 1996;60(1):44–69.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

Ames BN. The biosynthesis of histidine; L-histidinol phosphate phosphatase. J Biol Chem. 1957;226(2):583–93.

Aravind L, Galperin MY, Koonin EV. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. Trends Biochem Sci. 1998;23(4):127–9.

Arnold DN, Santosa F, Rosenthal J, Gilliam DS, editors. Mathematical systems theory in biology, communications, computation, and finance. The IMA Volumes in Mathematics and its Applications. New York, NY: Springer New York; 2003.

Arora G, Tiwari P, Mandal RS, Gupta A, Sharma D, Saha S, et al. High throughput screen identifies small molecule inhibitors specific for *Mycobacterium tuberculosis* phosphoserine phosphatase. J Biol Chem. 2014;289(36):25149–65.

Baker AS, Ciocci MJ, Metcalf WW, Kim J, Babbitt PC, Wanner BL, et al. Insights into the mechanism of catalysis by the P-C bond-cleaving enzyme phosphonoacetaldehyde hydrolase derived from gene sequence analysis and mutagenesis. Biochemistry. 1998;37(26):9305–15.

Bogatyreva NS, Finkelstein AV, Galzitskaya OV. Trend of amino acid composition of proteins of different taxa. J Bioinform Comput Biol. 2006;4(2):597–608.

Brilli M, Fani R. Molecular evolution of *hisB* genes. J Mol Evol. 2004;58(2):225–37.

Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L. Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. J Mol Biol. 2006;361(5):1003–34.

Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell. 2018;172(6):1181–97.

Cho H, Wang W, Kim R, Yokota H, Damo S, Kim SH, et al. BeF(3)(−) acts as a phosphate analog in proteins phosphorylated on aspartate: structure of a BeF(3)(−) complex with phosphoserine phosphatase. Proc Natl Acad Sci U S A. 2001;98(15):8525–30.

Collet JF, Stroobant V, Pirard M, Delpierre G, van Schaftingen E. A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif. J Biol Chem. 1998;273(23):14107–12.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.

Das S, Orengo CA. Protein function annotation using protein domain family resources. Methods. 2016;93:24–34.

Del Duca S, Chioccioli S, Vassallo A, Castronovo LM, Fani R. The role of gene elongation in the evolution of histidine biosynthetic genes. Microorganisms. 2020;8(5):732.

Enault F, Suhre K, Claverie J-M. Phydbac "gene function predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005;6:247.

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol. 2005;1(5):e45.

Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R. The evolution of histidine biosynthesis in archaea: insights into the his genes structure and organization in LUCA. J Mol Evol. 2009;69(5):512–26.

Furnham N, Garavelli JS, Apweiler R, Thornton JM. Missing in action: enzyme functional annotations in biological databases. Nat Chem Biol. 2009;5(8):521–5.

Ghodge SV, Fedorov AA, Fedorov EV, Hillerich B, Seidel R, Almo SC, et al. Structural and mechanistic characterization of L-histidinol phosphate phosphatase from the polymerase and histidinol phosphatase family of proteins. Biochemistry. 2013;52(6):1101–12.

Hon J, Borko S, Stourac J, Prokop Z, Zendulka J, Bednar D, et al. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Res. 2020;48(W1):W104–9.

Huang H, Pandya C, Liu C, Al-Obaidi NF, Wang M, Zheng L, et al. Panoramic view of a superfamily of phosphatases through substrate profiling. Proc Natl Acad Sci U S A. 2015;112(16):E1974–83.

Jha B, Kumar D, Sharma A, Dwivedy A, Singh R, Biswal BK. Identification and structural characterization of a histidinol phosphate phosphatase from *Mycobacterium tuberculosis*. J Biol Chem. 2018;293(26):10102–18.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.

Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The sequence read archive: a decade more of explosive growth. Nucleic Acids Res. 2022;50(D1):D387–90.

Kinateder T, Drexler L, Straub K, Merkl R, Sterner R. Experimental and computational analysis of the ancestry of an evolutionary young enzyme from histidine biosynthesis. Protein Sci. 2023;32(1):E4536.

King JL, Jukes TH. Non-Darwinian evolution. Science. 1969;164(3881):788–98.

Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40(Database issue):D54–6.

Koonin EV, Tatusov RL. Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. J Mol Biol. 1994;244(1):125–32.

Krieger E, Vriend G. YASARA view—molecular graphics for all devices—from smartphones to workstations. Bioinformatics. 2014;30(20):2981–2.

Kurihara T, Liu JQ, Nardi-Dei V, Koshikawa H, Esaki N, Soda K. Comprehensive site-directed mutagenesis of L-2-halo acid dehalogenase to probe catalytic amino acid residues. J Biochem. 1995;117(6):1317–22.

Kuznetsova E, Proudfoot M, Gonzalez CF, Brown G, Omelchenko MV, Borozan I, et al. Genome-wide analysis of substrate specificities of the *Escherichia coli* haloacid dehalogenase-like phosphatase family. J Biol Chem. 2006;281(47):36149–61.

LaBauve AE, Wargo MJ. Growth and laboratory maintenance of *Pseudomonas aeruginosa*. Curr Protoc Microbiol. 2012;Chapter 6:Unit 6E.1.

Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30(22):3276–8.

Lee D, Grant A, Marsden RL, Orengo C. Identification and distribution of protein families in 120 completed genomes using Gene3D. Proteins. 2005;59(3):603–15.

Lee HS, Cho Y, Lee J-H, Kang SG. Novel monofunctional histidinol-phosphate phosphatase of the DDDD superfamily of phosphohydrolases. J Bacteriol. 2008;190(7):2629–32.

Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, et al. Protein function annotation by homology-based inference. Genome Biol. 2009;10(2):207.

Morais MC, Zhang W, Baker AS, Zhang G, Dunaway-Mariano D, Allen KN. The crystal structure of *Bacillus cereus* phosphonoacetaldehyde hydrolase: insight into catalysis of phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily. Biochemistry. 2000;39(34):10385–96.

Nourbakhsh A, Collakova E, Gillaspy GE. Characterization of the inositol monophosphatase gene family in *Arabidopsis*. Front Plant Sci. 2014;5:725.

Oberg N, Zallot R, Gerlt JA. EFI-EST, EFI-GNT, and EFI-CGFP: enzyme function initiative (EFI) web resource for genomic enzymology tools. J Mol Biol. 2023;168018:168018.

Olsen DB, Hepburn TW, Moos M, Mariano PS, Dunaway-Mariano D. Investigation of the *Bacillus cereus* phosphonoacetaldehyde hydrolase. Evidence for a Schiff base mechanism and sequence analysis of an active-site peptide containing the catalytic lysine residue. Biochemistry. 1988;27(6):2229–34.

Orengo CA, Thornton JM. Protein families and their evolution—a structural perspective. Annu Rev Biochem. 2005;74:867–900.

Parsons JF, Lim K, Tempczyk A, Krajewski W, Eisenstein E, Herzberg O. From structure to function: YrbI from *Haemophilus influenzae* (HI1679) is a phosphatase. Proteins. 2002;46(4):393–404.

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221–7.

Rangarajan ES, Proteau A, Wagner J, Hung M-N, Matte A, Cygler M. Structural snapshots of *Escherichia coli* histidinol phosphate phosphatase along the reaction pathway. J Biol Chem. 2006;281(49):37930–41.

Rohweder B, Semmelmann F, Endres C, Sterner R. Standardized cloning vectors for protein production and generation of large gene libraries in *Escherichia coli*. Biotechniques. 2018;64(1):24–6.

Ruszkowski M, Dauter Z. Structural studies of *Medicago truncatula* histidinol phosphate phosphatase from inositol monophosphatase superfamily reveal details of penultimate step of histidine biosynthesis in plants. J Biol Chem. 2016;291(19):9960–73.

Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–100.

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009;5(12):e1000605.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. Database Oxford. 2020;2020.

Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9(1):2542.

Stepansky A, Leustek T. Histidine biosynthesis in plants. Amino Acids. 2006;30(2):127–42.

Suárez ASG, Stefan A, Lemma S, Conte E, Hochkoeppler A. Continuous enzyme-coupled assay of phosphate—or pyrophosphate—releasing enzymes. Biotechniques. 2012;53(2):99–103.

Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003;333(4):863–82.

Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol. 2001;307(4):1113–43.

Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61.

Walker CB, de La Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, et al. *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proc Natl Acad Sci U S A. 2010;107(19):8818–23.

Wang W, Cho HS, Kim R, Jancarik J, Yokota H, Nguyen HH, et al. Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic "snapshots" of intermediate states. J Mol Biol. 2002;319(2):421–31.

Wang W, Kim R, Jancarik J, Yokota H, Kim SH. Crystal structure of phosphoserine phosphatase from *Methanococcus jannaschii*, a hyperthermophile, at 1.8 a resolution. Structure. 2001;9(1):65–71.

Wang Y, Wang L, Zhang J, Duan X, Feng Y, Wang S, et al. PA0335, a gene encoding histidinol phosphate phosphatase, mediates histidine auxotrophy in *Pseudomonas aeruginosa*. Appl Environ Microbiol. 2020;86(5):E02593-19.

Wiater A, Krajewska-Grynkiewicz K, Klopotowski T. Histidine biosynthesis and its regulation in higher plants. Acta Biochim pol. 1971;18(3):299–307.

Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol. 1999;112:531–52.

Winkler ME, Ramos-Montañez S. Biosynthesis of histidine. EcoSal Plus. 2009;3(2).

Zallot R, Oberg N, Gerlt JA. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry. 2019;58(41):4169–82.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kinateder T, Mayer C, Nazet J, Sterner R. Improving enzyme functional annotation by integrating in vitro and in silico approaches: The example of histidinol phosphate phosphatases. Protein Science. 2024;33(2):e4899. https://doi.org/10.1002/pro.4899