Evolutionary Analysis and Functional Characterization of Different Histidinol Phosphate Phosphatases

DISSERTATION

Zur Erlangung des Doktorgrades

der Naturwissenschaften (Dr. rer. nat.)

der Fakultät für Biologie und vorklinische Medizin

der Universität Regensburg

vorgelegt von

Thomas Kinateder

aus Hutthurm

Juni 2023

Das Promotionsgesuch wurde eingereicht am:	12.06.2023
Die Arbeit wurde angeleitet von:	Prof. Dr. Reinhard Sterner
Unterschrift	……..
	Thomas Kinateder

Only a small part of scientific progress has resulted from planned search for specific objectives. A much more important part has been made possible by the freedom of the individual to follow his own curiosity.

Irving Langmuir

List of publications

This thesis is composed of four chapters. The first chapter is based on a published manuscript.

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

During this work, I also contributed to further publications that are not part of this thesis

B Kneuttinger A.C., Zwiesele S., Straub K., Bruckmann A., Busch F., **Kinateder T.**, Gaim B., Wysocki V.H., Merkl R., Sterner R. (2019) Light-Regulation of Tryptophan Synthase by Combining Protein Design and Enzymology. *Int J Mol Sci*, 20: 5106.

C Simeth N.A., **Kinateder, T.**, Rajendran, C., Nazet, J., Merkl, R., Sterner, R., König, B., Kneuttinger, A.C. (2021) Towards Photochromic Azobenzene-Based Inhibitors for Tryptophan Synthase. *Chemistry*, 27: 2439-2451.

D Maria-Solano M.A., **Kinteder T**., Iglesias-Fernández J., Sterner R., Osuna S. (2021). In Silico Identification and Experimental Validation of Distal Activity-Enhancing Mutations in Tryptophan Synthase. *ACS Catalysis*, 11: 13733-13743.

Personal contributions to publication A

The research was planned by myself, Lukas Drexler, and Reinhard Sterner. The experiments with the ancestral enzymes were performed by myself, the experiments with the extant enzymes were performed by myself and Lukas Drexler in equal parts. The ancestral sequence reconstruction was performed by Kristina Straub and the subsequent analysis of the phylogenetic trees was done by myself, Lukas Drexler, and Kristina Straub. The manuscript was drafted by myself and revised by Lukas Drexler, Reinhard Sterner, and Rainer Merkl.

Abstract

The elaborate metabolism of modern organisms raises the question of how such a complex system could have developed from a simple ancient prestage with a presumably very limited repertoire of enzymes. To answer this question various models have been put forward. A popular model assumes that the promiscuous side activities of enzymes represent a starting point for new reactions and reaction sequences from which novel enzymes can develop through gene duplication and divergent evolution. A remnant of these duplication events are the superfamilies of related enzymes, which have the same fold and may also have functional similarities but can often catalyze different reactions or catalyze the same type of reaction on different substrates. One of the most widespread superfamilies is the so-called haloacid dehalogenase (HAD) superfamily which mainly consists of phosphatases and phosphonatases that act on a wide variety of substrates. Substrate specificity is mainly achieved by highly variable cap modules which are inserted into a core Rossmann fold and cover the active site, both enabling specific interactions with the substrate and shielding the reaction center from bulk solvent. Due to the wide distribution of the HAD superfamily, it is assumed that it represents one of the oldest superfamilies. It is further assumed that various enzymes from the HAD superfamily that differed from each other by the aforementioned caps were already present in the last common ancestor of all cellular organisms (LUCA). A HAD enzyme with a comparatively original fold is the histidinol phosphate phosphatase (HolPase) from *Escherichia coli* (*ec*HisB-N). The HolPase is part of histidine biosynthesis, where it catalyzes the penultimate step, namely the dephosphorylation of histidinol phosphate to histidinol. Although this pathway is identical in all histidine-synthesizing species, which is why it is assumed that it was already present in LUCA, the HolPases of different species differ significantly and so far, HolPases from three different superfamilies have been identified. All the other enzymes from histidine biosynthesis are however conserved across different species, which means that the HolPases are probably evolutionary younger than the other enzymes of this pathway. This observation raises several questions, which are addressed in the present work.

In the first part, the question regarding the evolutionary origin of *ec*HisB-N was raised. The fact that this enzyme is not conserved in histidine-synthesizing species, while belonging to a very old protein superfamily, indicates that *ec*HisB-N evolved from a more ancestral enzyme, possibly a primordial phosphatase. In previous work, it has been argued that *ec*HisB-N and its closest homologue D-*glycero*-D-*manno*-heptose-1,7-bisphosphat-7-phosphatase (GmhB) were derived from the same promiscuous phosphatase. GmhB variants catalyze the hydrolysis of the two anomers of D-*glycero*-D-*manno*-heptose-1,7-bisphosphat (αHBP or βHBP), with one anomer usually being highly preferred by αGmhB or βGmhB, respectively. We found that *ec*HisB-N shows promiscuous activity for βHBP but not for αHBP, while the βGmhB from *Crassaminicella sp.* showed a promiscuous activity for HolP. Consistent with this, in a combined phylogenetic tree of αGmhB, βGmhB, and HisB-N sequences, HisB-N sequences formed a compact subcluster derived from βGmhBs. To analyze the properties of the precursors, several enzymes were resurrected by ancestral sequence reconstruction, and their functionality was tested *in vitro*. In this analysis, a promiscuous HolPase activity could already be detected in the ancestral enzymes belonging to nodes that predate the functional divergence of βGmhB and HisB-N. This HolPase activity was significantly increased in enzymes that belong to younger nodes and from which only modern HisB-N enzymes are derived. This increase in the catalytic efficiency of HolP turnover is reflected in the shape and electrostatics of the active site as predicted by AlphaFold. Finally, a revised model for the evolution of HisB-N from an ancestral βGmhB was developed with the help of a detailed

analysis of the phylogenetic tree. In agreement with the experimental data, a horizontal gene transfer of a promiscuous βGmhB enzyme from an ancestral δ-Proteobacterium to an ancestral γ-Proteobacterium is assumed. After the horizontal gene transfer, this βGmhB then most likely evolved into a modern HisB-N.

In the second and third part, the question is asked whether there are other HolPases that are not directly related to HolPases which were reported so far and instead developed independently.

Specifically, in the second part, a protein of the HAD superfamily from *Pseudomonas aeruginosa* (*pa*HisN) is analyzed, for which a HolPase function was suggested on the basis of *in vivo* experiments that were reported in a recent publication. An analysis of the AlphaFold-predicted structure of *pa*HisN in the present work confirmed its classification as a member of the HAD superfamily. A comparison to the HAD family HolPase *ec*HisB-N unveiled that the two structures differed significantly in their cap structures, indicating that *pa*HisN was not derived from *ec*HisB-N or *vice versa*. Instead, *pa*HisN showed considerable similarities to phosphoserine phosphatases (PSPases) both at the sequence level and regarding the protein fold, indicating a possible evolutionary relationship or overlapping functions. Subsequent characterization of the enzyme *in vitro* showed that the protein is present as a monomer in aqueous solution and has a melting temperature of 46°C. In addition, the assumed HolPase activity could be confirmed as the native function of this enzyme. Moreover, a promiscuous PSPase activity was discovered which supports the hypothesis of a distant relationship to this class of enzymes. To distinguish HolPases from PSPases among the homologues of *pa*HisN, an alanine scan of the active site was performed to identify residues critical to HolPase function. The results of this alanine scan were used in combination with a sequence logo of PSPases to derive a fingerprint, consisting of a DxD motif and a tyrosine, which is assumed to be conclusive for a HolPase function. The subsequent analysis of a sequence similarity network (SSN) of homologous proteins led to the identification of numerous non-annotated proteins in β- and γ-Proteobacteria which contained this fingerprint. The conservation of this set of critical residues suggests, that these homologues are HolPases with similar properties as *pa*HisN. This conclusion was cross-validated by a bioinformatic analysis of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database which showed that for many β- and γ-Proteobacteria there was indeed no HolPase annotated. This finding supports the proposed annotation of the homologues as HolPases most likely identifying the last missing enzyme from histidine biosynthesis in these organisms. An additional bioinformatic analysis of all phyla revealed that there generally is a considerable knowledge-gap concerning the HolPase function as for 32 % of all histidine-synthesizing organisms the enzyme that catalyzes the HolPase reaction is not known. In archaea this knowledge-gap is especially pronounced as in this domain an annotated HolPase is missing in approximately two thirds of the histidine-synthesizing species.

The third part is therefore dedicated to the search for the enzyme that catalyzes the HolPase reaction in the archaeal kingdom. In previous work, a gene between *hisC* and *hisB* that was annotated as a putative phosphatase from the HAD superfamily had been noticed in the archaeon *Nitrosopumilus maritimus*. The location of this gene within a cluster of genes from histidine biosynthesis and classification of its gene product as HAD protein made this a promising candidate for the missing HolPase. An analysis of sequence and the AlphaFold-predicted structure in the course of this work confirmed that this protein belongs to the HAD superfamily. However, the cap of this protein showed no similarity to the caps of either *pa*HisN or *ec*HisB-N, indicating that all three proteins evolved independently. The comparison to a third HolPase from the HAD superfamily which was recently discovered in the archaeon

Thermococcus onnurineus uncovered a limited sequence identity of 23.9% and a moderate similarity in the protein fold. While this indicated a distant relationship of the two proteins, the similarities were too low to infer a HolPase function for the protein from *N. maritimus.* Therefore, an *in vitro* characterization of the uncharacterized protein from *N. maritimus* was conducted which confirmed the suspected HolPase activity and showed that the protein exists as a monomer in aqueous solution and has a melting temperature of 37°C. In accordance to previously used nomenclature, *nm*HisN is therefore proposed as name for the protein from *N. maritimus*. The observed HolPase function of *nm*HisN, in combination with the distant relationship to the HolPase from *T. onnurineus*, suggested that these two proteins might be representative for a new class of significantly diverged HolPases which is widely distributed within the archaeal kingdom. To test this hypothesis, a fingerprint of HolPase defining residues was established on the basis of an alanine scan of the active site of *nm*HisN, which should allow for the identification of other archaeal HolPases. This fingerprint consisting of a DY motif, a lysine and a glutamate, which are assumed to be conclusive for a HolPase function of homologues of *nm*HisN. Afterwards, an SSN was created in which homologues of *nm*HisN were found in a wide variety of archaeal phyla. The homologues of two phyla, namely the Thaumarchaeota and the candidate phylum of Bathyarchaeota, contained a highly conserved fingerprint, strongly indicating that these proteins are HolPases. Moreover, homologues which contained three of the four fingerprint residues were identified in Euryarchaeota and, interestingly, also in bacterial δ-Proteobacteria. This indicates that these proteins also possess HolPase function. Taken together, the results of this part show that the HolPases from *N. maritimus* and *T. onnurineus* constitute a third type of HolPase from the HAD superfamily besides *pa*HisN and *ec*HisB-N. This type of HolPase seems to be widely distributed among archaea which indicates its ancient history. The proposed annotation of these proteins as HolPases also helps to close this knowledge gap on the histidine biosynthesis of archaea.

The fourth part of the work deals with functional transitions and the general evolvability of the HAD superfamily. Specifically, a possible functional transition from a PSPase to a HolPase was analyzed, as suggested in the case of the evolution of *pa*HisN. For this, the *E. coli* PSPase *ec*SerB served as a model enzyme, which showed no measurable HolPase function *in vitro*. Several active site amino acids were randomized simultaneously and then tested for their ability to mediate growth of a Δ*holPase* strain on selective medium lacking histidine. After one round of random mutagenesis and selection, two enzyme variants were identified that exhibited continuously measurable HolPase activity *in vitro* with a catalytic efficiency of 1.7 $M^{-1}s^{-1}$ and 7.8 $M^{-1}s^{-1}$, respectively. In addition, the general evolvability of HAD enzymes was investigated using the promiscuous PSPase activity of *ec*HisB-N as case study. In previous work, directed evolution improved the promiscuous PSPase activity by about an order of magnitude. In this work, the activity could be increased by another order of magnitude by further random mutagenesis and selection. Here, D58 was identified as a key residue where mutation to asparagine alone resulted in a 60-fold improvement in PSPase activity. Both lines of experiment underscore the functional versatility of the HAD superfamily which can be easily evolved to new substrates through adaptations in the variable caps, while the basic catalytic machinery is preserved.

Finally, based on the experimental data, a three-step model for the evolution of HolPases was proposed. It is assumed that HolP is hydrolyzed at a slow rate spontaneously in absence of an enzyme, mediated e.g., by Mg^{2+}. It is furthermore assumed, that during the assembly of the histidine biosynthetic pathway the HolPase function was under the lowest selection pressure. The lack of conservation within the HolPases furthermore indicates that LUCA did not yet have a specialized HolPase. In the second step of evolution, with increasing genome size, the existence of one or more phosphatases with promiscuous

HolPase activity became likely. However, specialized HolPases were most probably only established in a third step, after the three different kingdoms of life had diverged. After that, several horizontal gene transfers probably occurred, which is why homologous HolPases are now found in very distantly related organisms.

Zusammenfassung

Der ausgefeilte Stoffwechsel moderner Organismen wirft die Frage auf, wie sich ein so komplexes System aus einfachen Vorstufen mit einem mutmaßlich sehr begrenzten Repertoire von Enzymen entwickeln konnte. Um diese Frage zu beantworten, wurden verschiedene Modelle entwickelt. Ein populäres Modell geht davon aus, dass promiskuitive Nebenreaktionen von Enzymen einen Ausgangspunkt für neue Reaktionen und Reaktionssequenzen darstellen, aus dem sich durch Genduplikation und divergente Evolution neuartige Enzyme entwickeln können. Ein Relikt dieser Genduplikationen sind demnach die Superfamilien aus verwandten Enzymen, die zwar den gleichen Faltungstyp besitzen und unter Umständen auch funktionale Ähnlichkeiten aufweisen, aber häufig unterschiedliche Reaktionen katalysieren oder die gleiche Reaktion an unterschiedlichen Substraten katalysieren. Eine der am weitesten verbreiteten Superfamilien ist die sogenannte Haloacid-Dehalogenase-(HAD)-Superfamilie, die hauptsächlich Phosphatasen und Phosphonatasen beinhaltet, die eine Vielzahl verschiedener Substrate umsetzen. Substratspezifität wird dabei hauptsächlich durch hochvariable sogenannte Caps erreicht, die als Insertion in die Rossmann-Faltung eingefügt sind und das aktive Zentrum bedecken. Diese Caps ermöglichen sowohl spezifische Wechselwirkungen mit dem jeweiligen Substrat als auch eine Abschirmung des Reaktionszentrums vom Lösungsmittel. Wegen der weiten Verbreitung der HAD-Superfamilie geht man davon aus, dass sie eine der ältesten Superfamilien ist. Man geht weiterhin davon aus, dass verschiedene Enzyme der HAD Superfamilie, die sich durch die besagten Caps voneinander unterschieden, schon im letzten gemeinsamen Vorläufer aller zellulären Organismen (LUCA) vorhanden waren. Ein HAD-Enzym mit einer vergleichsweise ursprünglichen Faltung ist die Histidinolphosphatphosphatase (HolPase) aus *Escherichia coli* (*ec*HisB-N). Die HolPase ist Teil der Histidin Biosynthese, in der sie den vorletzten Schritt, nämlich die Dephosphorylierung von Histidinolphosphat zu Histidinol katalysiert. Obwohl dieser Stoffwechselweg in allen Histidin-synthetisierenden Spezies identisch abläuft - weshalb man davon ausgeht, dass er bereits im LUCA etabliert war - unterscheiden sich die HolPasen verschiedener Spezies deutlich voneinander. Bisher wurden HolPasen aus drei verschiedenen Superfamilien identifiziert. Alle anderen Enzyme der Histidin-Biosynthese sind jedoch speziesübergreifend konserviert, was bedeutet, dass die HolPasen vermutlich jünger sind als die anderen Enzyme des Stoffwechselwegs. Diese Sonderstellung der HolPasen führt zu mehreren Fragen, die in der vorliegenden Arbeit adressiert werden.

Im ersten Teil wurde der Frage nach dem evolutionären Ursprung von *ec*HisB-N nachgegangen. Die Tatsache, dass dieses Enzym in Histidin-synthetisierenden Spezies nicht konserviert ist, aber gleichzeitig zu einer sehr alten Protein-Superfamilie gehört, weist darauf hin, dass sich *ec*HisB-N aus einem ursprünglicheren Enzym, möglicherweise einer urtümlichen Phosphatase, entwickelt hat. In früheren Arbeiten wurde angenommen, dass *ec*HisB-N und das am nächsten verwandte homologe Enzym D-*glycero*-D-*manno*-heptose-1,7-Bisphosphat-7-Phosphatase (GmhB) aus derselben promiskuitiven Phosphatase hervorgegangen sind. GmhB-Varianten katalysieren die Hydrolyse des anomeren D-*glycero*-D-*manno*-heptose-1,7-bisphosphats (αHBP oder βHBP), wobei meist ein Anomer stark von αGmhB oder βGmhB bevorzugt wird. Wir konnten herausfinden, dass *ec*HisB-N promiskuitive Aktivität für βHBP, jedoch nicht für αHBP zeigt, während für βGmhB aus *Crassaminicella sp.* eine promiskuitive Aktivität für HolP nachgewiesen werden konnte. Übereinstimmend damit bildeten HisB-N Sequenzen in einem kombinierten phylogenetischen Baum aus αGmhB-, βGmhB- und HisB-N Sequenzen, einen kompakten Subcluster, der sich von βGmhBs ableitet. Um die Eigenschaften der Vorläufer zu analysieren, wurden mehrere Enzyme durch anzestrale

Sequenzrekonstruktion rekonstruiert und ihre Funktionalität *in vitro* getestet. In dieser Analyse konnte nachgewiesen werden, dass bereits Enzyme, die zu Knotenpunkten gehörten, die zeitlich vor der funktionellen Divergenz von βGmhB und HisB-N lagen, eine promiskuitive HolPase-Aktivität besaßen. Diese HolPase Aktivität war in Enzymen, die zu späteren Knotenpunkten gehören und von denen sich ausschließlich heutige HisB-N Enzyme ableiten, nochmal deutlich erhöht. Diese Steigerung der katalytischen Effizienz des HolP Umsatzes spiegelt sich in der von AlphaFold vorhergesagten Form und Elektrostatik des aktiven Zentrums wider. Abschließend wurde mit Hilfe einer detaillierten Analyse des phylogenetischen Baumes ein überarbeitetes Modell für die Evolution von HisB-N aus einem anzestralen βGmhB entworfen. Im Einklang mit den experimentellen Daten wird von einem horizontalen Gentransfer eines promiskuitiven βGmhB Enzyms von einem anzestralen δ-Proteobakterium auf ein anzestrales γ-Proteobakterium ausgegangen. Nach dem horizontalen Gentransfer entwickelte sich dieses βGmhB dann höchst-wahrscheinlich zu einem modernen HisB-N.

Im zweiten und dritten Teil wird der Frage nachgegangen, ob es noch weitere HolPasen gibt, die nicht direkt von den bisher beschriebenen HolPasen abstammen, sondern sich stattdessen unabhängig davon entwickelten.

Konkret wird im zweiten Teil ein Protein der HAD-Superfamilie aus *Pseudomonas aeruginosa* (*pa*HisN) analysiert, für das auf Basis von publizierten *in vivo* Experimenten eine HolPase Funktion vorgeschlagen wurde. Eine Analyse der von AlphaFold vorhergesagten Struktur von *pa*HisN im Rahmen dieser Arbeit bestätigte die Zugehörigkeit zur HAD-Superfamilie. Ein Vergleich mit der bereits bekannten HolPase aus der HAD-Superfamilie, *ec*HisB-N, ergab jedoch, dass sich die beiden Strukturen in ihren Caps deutlich unterschieden, was darauf hinweist, dass weder *pa*HisN von *ec*HisB-N abgeleitet ist noch umgekehrt *ec*HisB-N von *pa*HisN. Stattdessen zeigte *pa*HisN sowohl auf Sequenzebene als auch in Bezug auf die Proteinfaltung erhebliche Ähnlichkeiten zu Phosphoserin-Phosphatasen (PSPasen), was auf eine mögliche evolutionäre Verwandtschaft oder überlappende Funktionen hinweist. Eine anschließende Charakterisierung des Enzyms *in vitro* zeigte, dass das Protein in wässriger Lösung als Monomer vorliegt und einen Schmelzpunkt von 46°C besitzt. Außerdem konnte die vermutete HolPase-Aktivität als native Funktion des Enzyms bestätigt werden. Darüber hinaus wurde eine promiskuitive PSPase-Aktivität entdeckt, die die Hypothese einer entfernten Verwandtschaft zu PSPasen unterstützt. Um unter den Homologen von *pa*HisN HolPasen und PSPasen unterscheiden zu können, wurde ein Alanin-Scan des aktiven Zentrums durchgeführt, um die Reste zu identifizieren, die kritisch für die HolPase-Funktion sind. Aus dem Alanin-Scan sowie einem Vergleich mit einem Sequenzlogo von PSPasen wurde ein Fingerabdruck abgeleitet, bestehend aus einem DxD-Motiv und einem Tyrosin, von dem angenommen wird, dass er ein aussagekräftiger Klassifikator für HolPasen ist. Die anschließende Analyse eines Sequenzähnlichkeitsnetzwerks (SSN) homologer Proteine führte zur Identifikation zahlreicher nicht annotierter Proteine in β- und γ-Proteobakterien die diesen Fingerabdruck aufwiesen. Die Konserviertheit dieser kritischen Reste legt nahe, dass es sich bei diesen Homologen um HolPasen handelt, die ähnliche Eigenschaften wie *pa*HisN besitzen. Diese Schlussfolgerung wurde durch eine bioinformatische Analyse der Datenbank *Kyoto Encyclopedia of Genes and Genomes* (KEGG) validiert, die zeigte, dass für viele β- und γ-Proteobakterien tatsächlich keine HolPase annotiert war. Dieser Befund unterstützt die vorgeschlagene Annotation der Homologen als HolPasen und identifiziert damit wahrscheinlich das letzte fehlende Enzym aus der Histidin-Biosynthese dieser Organismen. Eine zusätzliche bioinformatische Analyse aller Phyla ergab, dass insgesamt eine erhebliche Wissenslücke im Hinblick auf die HolPase-Funktion besteht, da für 32% aller Histidin synthetisierenden Organismen das Enzym, das die HolPase-Reaktion katalysiert nicht bekannt

ist. In der Domäne der Archaeen ist diese Lücke besonders ausgeprägt, da hier bei etwa zwei Dritteln der Histidin-synthetisierenden Arten eine annotierte HolPase fehlt.

Im dritten Teil der Arbeit ist daher der Suche nach einem Enzym gewidmet, das die HolPase Reaktion in Archaeen katalysiert. In einer früheren Arbeit war im Archaeon *Nitrosopumilus maritimus* ein Gen zwischen *hisC* und *hisB* aufgefallen, das als putative Phosphatase aus der HAD-Superfamilie annotiert war. Die Lage innerhalb eines Genclusters aus Genen der Histidin-Biosynthese und die Klassifizierung des Genprodukts als HAD-Protein machten es zu einem vielversprechenden Kandidaten für die fehlende HolPase. Eine Analyse der Sequenz und er von AlphaFold vorhergesagten Struktur im Rahmen der vorliegenden Arbeit bestätigte die Zugehörigkeit zur HAD-Superfamilie. Allerdings zeigte die Cap-Struktur dieses Proteins keine Ähnlichkeit zu den Caps von *pa*HisN oder *ec*HisB-N, was darauf hindeutet, dass sich alle drei Proteine unabhängig voneinander entwickelten. Der Vergleich mit einer dritten HolPase aus der HAD-Superfamilie, die kürzlich im Archaeon *Thermococcus onnurineus* entdeckt wurde, ergab eine Sequenzidentität von lediglich 23.9 % und eine moderate Ähnlichkeit in der Proteinfaltung. Dies deutete zwar auf eine entfernte Verwandtschaft der beiden Protein hin, die Ähnlichkeit war jedoch zu gering, um eine HolPase-Funktion für das Protein aus *N. maritimus* ableiten zu können. Daher wurde eine *in vitro* Charakterisierung des Proteins aus *N. maritimus* durchgeführt, die die vermutete HolPase Aktivität bestätigte und zeigte, dass das Protein in wässriger Lösung als Monomer vorliegt und einen Schmelzpunkt von 37 °C besitzt. Im Einklang mit früher verwendeter Nomenklatur wird daher für das Protein aus *N. maritimus* die Bezeichnung *nm*HisN vorgeschlagen. Die beobachtete HolPase-Funktion von *nm*HisN in Kombination mit der entfernten Verwandtschaft zur HolPase aus *T. onnurineus* legt nahe, dass diesen beiden Proteinen repräsentativ für eine neue Klasse von deutlich divergierten HolPasen stehen könnten, die in der Domäne der Archaeen sehr weit verbreitet sein könnte. Um diese Hypothese zu testen, wurde auf der Grundlage eines Alanin-Scans des aktiven Zentrums von *nm*HisN ein Fingerabdruck von HolPase-definierenden Resten erstellt, der die Identifizierung anderer archaeeller HolPasen ermöglichen sollte. Dieser Fingerabdruck besteht aus einem DY-Motiv, einem Lysin und einem Glutamat, von denen angenommen wird, dass sie einen eindeutigen Hinweis auf eine HolPase Funktion darstellen. Anschließend wurde ein SSN erstellt, in dem Homologe von *nm*HisN in einer Vielzahl archaeeller Phyla gefunden wurden. In den Homologen zweier Phyla, nämlich der Thaumarchaeota und des Kandidatenphylums der Bathyarchaeota war der Fingerabdruck hoch konserviert, was stark darauf hindeutete, dass es sich bei diesen Proteinen um HolPasen handelt. Darüber hinaus wurden Homologe in Euryarchaeota und, interessanterweise, auch in bakteriellen δ-Proteobakterien identifiziert, die drei der vier Aminosäuren aus dem Fingerabdruck enthielten. Das weist darauf hin, dass auch diese Proteine HolPase-Funktion besitzen. Zusammengenommen zeigen die Ergebnisse dieses Teils, dass die HolPasen von *N. maritimus* und *T. onnurineus* neben *pa*HisN und *ec*HisB-N einen dritten HolPase-Typ aus der HAD-Superfamilie darstellen. Diese Art HolPase scheint in Archaeen weit verbreitet zu sein, was auf eine frühe Entwicklung hinweist. Die vorgeschlagene Annotation dieser Proteine als HolPasen trägt außerdem dazu bei, diese bestehende Wissenslücke über die Histidin-Biosynthese von Archaeen zu schließen.

Im vierten Teil der Arbeit wird der Frage nach funktionalen Übergängen und der allgemeinen Evolvierbarkeit der HAD-Superfamilie nachgegangen. Konkret wurde ein möglicher funktionaler Übergang von einer PSPase zu einer HolPase analysiert, wie er im Fall der Entstehung von *pa*HisN naheliegt. Als Modellenzym diente die PSPase aus *E. coli ec*SerB, die *in vitro* keine messbare HolPase zeigte. Mehrere Aminosäuren des aktiven Zentrums wurden gleichzeitig randomisiert und anschließend auf ihre Fähigkeit getestet, einem Δ*holPase* Stamm Wachstum auf Minimalmedium zu ermöglichen.

Nach einer Runde gerichteter Evolution wurden zwei Enzymvarianten gefunden, die *in vitro* kontinuierlich messbare HolPase Aktivität mit einer katalytischen Effizienz von 1,7 $M^{-1}s^{-1}$ bzw. 7,8 $M^{-1}s^{-1}$ aufwiesen. Außerdem wurde die generelle Evolvierbarkeit von HAD-Enzymen am Beispiel der promiskuitiven PSPase-Aktivität von *ec*HisB-N untersucht. In einer vorangegangenen Arbeit war die Aktivität durch gerichtete Evolution um etwa eine Größenordnung verbessert worden. In dieser Arbeit konnte die Aktivität durch Zufallsmutagenese und Selektion um eine weitere Größenordnung gesteigert werden. Dabei wurde D58 als entscheidender Rest identifiziert, bei dem die Mutation zu Asparagin allein eine 60-fache Verbesserung der PSPase-Aktivität bewirkte. Beide Versuchsreihen unterstreichen die Vielseitigkeit der HAD-Superfamilie bei der die Funktion durch Anpassungen in den variablen Caps leicht für neue Substrate angepasst werden kann, während der grundsätzliche Katalysemechanismus erhalten bleibt.

Basierend auf den experimentellen Daten wurde abschließend ein dreistufiges Modell für die Evolution von HolPasen vorgeschlagen. Es wird angenommen, dass HolP zunächst spontan und in Abwesenheit eines Enzyms langsam hydrolysiert wurde, begünstigt z.B. durch Mg^{2+}. Es wird außerdem angenommen, dass während der Assemblierung des Histidin-Biosynthesewegs die HolPase-Funktion unter dem geringsten Selektionsdruck stand. Die mangelnde Konservierung innerhalb der HolPasen, deutet darauf hin, dass LUCA noch keine spezialisierte HolPase besaß. Im zweiten Evolutionsschritt wurde mit zunehmender Genomgröße die Existenz einer oder mehrerer Phosphatasen mit promiskuitiver HolPase-Aktivität wahrscheinlich. Spezialisierte HolPasen bildeten sich jedoch wahrscheinlich erst, nachdem sich die verschiedenen Domänen des Lebens auseinanderentwickelt hatten. Danach kam es wahrscheinlich zu mehreren horizontalen Gentransfers, weshalb homologe HolPasen heute in sehr entfernt verwandten Organismen gefunden werden.

Table of contents

List of publications ..... **VI**

Personal contributions to publication A ..... **VII**

Abstract ..... **VIII**

Zusammenfassung ..... **XII**

1 General Introduction **1**

1.1 Evolution – an incremental increase in complexity 1

1.2 Evolutionary mechanisms 1

1.3 The haloacid dehalogenase superfamily – a primordial protein superfamily 4

1.4 The HolPase – the exception within histidine biosynthesis 8

1.5 Objectives of this thesis 11

2 The evolution of HisB-N from γ-Proteobacteria **12**

2.1 Introduction 12

2.2 Results and Discussion 13

2.2.1 Sequential and structural comparison of HisB-N and GmhB 13

2.2.2 Purification and functional characterization of *ec*HisB-N and *ec*GmhB 14

2.2.3 Phylogenetic analysis 17

2.2.4 Ancestral sequence reconstruction 19

2.2.5 Structural analysis of the predecessors Anc1-Anc7 22

2.2.6 A revised model for the evolution of HisB-N 24

3 Characterization of a putative HolPase from *P. aeruginosa* **27**

3.1 Introduction 27

3.2 Results and Discussion 29

3.2.1 Structural analysis and search for homologues of *pa*HisN 29

3.2.2 *In vitro* characterization of *pa*HisN 30

3.2.3 Analysis of the functionally relevant residues in *pa*HisN by alanine scanning 33

3.2.4 *In silico* analysis of the homologues of *pa*HisN 38

3.2.5 Analysis of the phylogenetic distribution of different HolPases 41

3.3 Conclusion 44

4 Characterization of a putative HolPase from *N. maritimus* **45**

4.1 Introduction 45

4.2 Results and Discussion 46

4.2.1 Analysis of the genomic neighborhood and predicted structure of *nm*HisN 46

4.2.2 *In vitro* characterization of *nm*HisN 49
4.2.3 Analysis of the functionally relevant residues in *nm*HisN by alanine scanning 52
4.2.4 *In silico* analysis of the homologues of *nm*HisN 55
4.3 Conclusion 58
5 Directed evolution of HAD enzymes 59
5.1 Introduction 59
5.2 Results and Discussion 60
5.2.1 Investigation of a possible evolutionary trajectory from PSPases to HolPases 60
5.2.2 Improvement of the promiscuous PSPase activity of *ec*HisB-N 64
6 Comprehensive Conclusion 69
6.1 The great diversity and phylogenetic scattering of HolPases 69
6.2 A model for the early evolution of the HolPase function 71
7 Materials 73
7.1 Devices 73
7.2 Consumables 75
7.3 Chemicals 76
7.4 Bacterial strains 77
7.5 Media, buffers, and solutions 78
7.5.1 Buffers for protein purification 78
7.5.2 Buffers for anion exchange chromatography 78
7.5.3 Buffers and media for microbiological methods 78
7.5.4 Buffers and solutions for molecular biological methods 80
7.5.5 Buffers and solutions for SDS-PAGE 81
7.6 Kits, enzymes, and ready-made buffers 81
7.7 Plasmids 82
7.8 Gene sequences 83
7.9 Primers 86
7.10 Software 90
7.10.1 Local software 90
7.10.2 Server based software 90
8 Methods 91
8.1 Bioinformatical methods 91
8.1.1 Analysis of the phylogenetic distribution and co-occurrence of enzymes 91
8.1.2 Sequence similarity networks 91

8.2 Microbiological methods ... 92
8.2.1 Cultivation and storage of *E. coli* strains ... 92
8.2.2 Preparation of *E. coli* knock-out strains ... 92
8.2.3 Preparation and transformation of chemically competent *E. coli* cells ... 93
8.2.4 Preparation and transformation of electrocompetent *E. coli* cells ... 93
8.2.5 Complementation of a gene knock-out ... 94
8.3 Molecular biological methods ... 95
8.3.1 Isolation and purification of plasmid DNA from E. coli ... 95
8.3.2 Measurement of DNA concentration ... 95
8.3.3 Agarose gel electrophoresis and isolation of DNA fragments ... 95
8.3.4 Enzymatic manipulation of DNA ... 95
8.4 Protein biochemical methods ... 100
8.4.1 Heterologous gene expression with subsequent analysis in analytical scale ... 100
8.4.2 Heterologous gene expression in preparative scale ... 101
8.4.3 Cell disruption and isolation of the soluble fraction ... 101
8.4.4 Immobilized metal affinity chromatography ... 101
8.4.5 Preparative size exclusion chromatography ... 102
8.4.6 Buffer exchange in analytical scale ... 102
8.4.7 Storage of purified proteins ... 102
8.4.8 Synthesis and purification of αHBP and βHBP ... 103
8.5 Analytical methods ... 104
8.5.1 Determination of protein concentration by UV/Vis spectroscopy ... 104
8.5.2 SDS-polyacrylamide gel electrophoresis (SDS-PAGE) ... 105
8.5.3 Circular dichroism (CD) spectroscopy ... 105
8.5.4 Size exclusion chromatography followed by static light scattering (SEC-SLS) ... 106
8.5.5 Steady-state enzyme kinetic experiments ... 106
8.5.6 Discontinuous enzyme kinetic experiments ... 108
9 References ... 109
Acronyms ... 118
List of Figures ... 121
List of Tables ... 122
Supplementary tables and figures ... 123
Acknowledgements ... 157

1 General Introduction

1.1 Evolution – an incremental increase in complexity

The evolution of life from the very beginning to its present form is characterized by a drastic increase in complexity and diversity. At the onset of evolution, life lacked many of its current features. It has been argued that life was confined to compartments of abiotic origin[1, 2] and consisted solely of simple self-replicating RNA molecules which were eponymous for the RNA world.[3, 4] In this scenario, the RNA served both as catalytic moiety and as carrier of the genetic information. This primordial RNA world was probably characterized by the presence of some amino acids which can be formed abiotically.[5, 6] It is assumed that this co-existence of few early amino acids and RNA allowed for the evolution of a primordial translation machinery and a preliminary stage of the genetic code.[7–10] The existence of the genetic code in turn facilitated the synthesis of longer peptides, which probably justified the requirement for a hydrophobic protein core and likely promoted an expansion of the genetic code by additional hydrophobic amino acids.[11] The driving force that led to the present day set of 20 amino acids was probably the increase in functional diversity which was provided by amino acids such as tryptophan, arginine or histidine which are energy intensive to synthesize and which are assumed to have been integrated late into the code.[12–14] The fact that this set of amino acids and the genetic code are universal in all extant organisms[15, 16] led to the conclusion that it was already present in the last universal common ancestor (LUCA) which inhabited our earth approximately 4 billion years ago.[17, 16] Comparative studies between extant organisms further suggest that the metabolism of LUCA was already rather complex and included diverse biosynthetic routes, e.g., for nucleotide and amino acid metabolism and the synthesis of membrane components.[18–20] Starting from LUCA, spatial separation must have resulted in divergent evolution and the formation of different species[21] and the split of all life forms into the three kingdoms archaea, eukaryota, and bacteria. This divergence is still ongoing and is the reason for the plethora of species which we can observe today and for their ability to inhabit the most diverse ecological niches. This brief history of the development of life showcases that the drastic increase from the most simplistic beginnings to today's diversity was in fact achieved by a series of small steps, each of which corresponded to an incremental increase in complexity.

1.2 Evolutionary mechanisms

The observation that evolution can be viewed as a series of small incremental improvements raises the question regarding the driving forces and the mechanisms that underly this process. From a metabolic perspective, the question could be rephrased to: How did sophisticated metabolic networks evolve from simple reactions or reaction modules and how are new reactions incorporated into biosynthetic routes? Different models were put forward to explain this expansion of the metabolic network.[22] The retrograde model[23] proposes that metabolic pathways evolve in a stepwise manner starting from a metabolite that is initially available from an abiotic origin. Continued demand for this metabolite leads to its depletion which creates a selective pressure that favors the synthesis of this metabolite from any available precursor molecule(s). This way, the selective pressure would facilitate the incorporation of a first enzyme which catalyzes the turnover of a precursor to the metabolite. If this precursor itself gets depleted, another reaction step would become necessary to synthesize the precursor molecule. In this

way, a reaction pathway would be assembled in a backward manner, from the last chemical step to the first.

While the retrograde model does not assume any biological relevance of the precursor molecules, the forward pathway model[24] presumes a functional benefit of these intermediates. According to the forward pathway model, the intermediates are refined by the stepwise addition of more enzymes at the end of the reaction pathway which yields an improved end-product or several different end-products with fine-tuned properties. The experimental evidence for both the retrograde and the forward pathway evolution model is however scarce and restricted to special cases.[22]

An alternative evolutionary model was proposed independently by Ycas and Jensen who both speculated that specialist enzymes arose from generalist enzymes with a broad substrate spectrum.[25, 26] Specifically, they proposed that the primitive metabolism consisted of a few generalist enzymes which were duplicated and evolved in divergent evolutionary trajectories to several specialist enzymes with a narrow substrate spectrum. This duplication-divergence model would also be applicable to reaction modules which consist of several generalist enzymes, catalyzing consecutive steps in a generic pathway until the whole module gets duplicated.

A similar hypothesis is the patchwork hypothesis which was formulated by Lazcano and Miller and which also assumes founder enzymes that catalyze more than one reaction.[27] According to the patchwork hypothesis, new pathways could evolve by combining enzymes from different pre-existing pathways. To be able to synthesize novel metabolites in this newly assembled pathway the starting enzymes are required to catalyze promiscuous side reactions in addition to their native function. Specifically, the starting enzymes either need to show reaction promiscuity i.e., catalyze different reactions on their primary substrate or show substrate promiscuity i.e., accept more than one substrate. Substrate promiscuity is indeed a widely observed phenomenon[28, 29], which supports the patchwork hypothesis and the duplication divergence model from Ycas and Jensen.

In a recent article, Noda-Garcia et al. argued that in addition to promiscuous side activities non-enzymatic activities also played an important role in the emergence of new pathways.[22] They further stated that an emerging pathway is likely composed of enzymatic reactions catalyzed by promiscuous enzymes and non-enzymatic reactions. The overall efficiency of such an emerging pathway is then likely very limited and the associated flux rather low. If this pathway comes under selective pressure an improvement of the efficiency of the rate-limiting step is most effective. Depending on whether the rate-limiting step is an enzymatic reaction or not, the biggest improvement can be achieved by enhancing a promiscuous side activity or by incorporating a new enzyme to catalyze a non-enzymatic reaction.

In addition to the presented models which consider entire reaction pathways, there is also an enzymological perspective to the question of how the increase in complexity was achieved. In his seminal work, Ohno was the first to raise the thesis that gene duplications are the main source for new enzyme functions.[30] He argued that the duplication of a gene would alleviate the selective pressure from the redundant gene copies which could therefore accumulate mutations and hence also acquire new functions. According to this line of argument, families of proteins with homologous folds that often catalyze similar reactions would be the remnants of previous duplication events.

Ohno's model was refined in the innovation-amplification-diversification (IAD) model proposed by Bergthorsson et al. (Figure 1.1).[31]

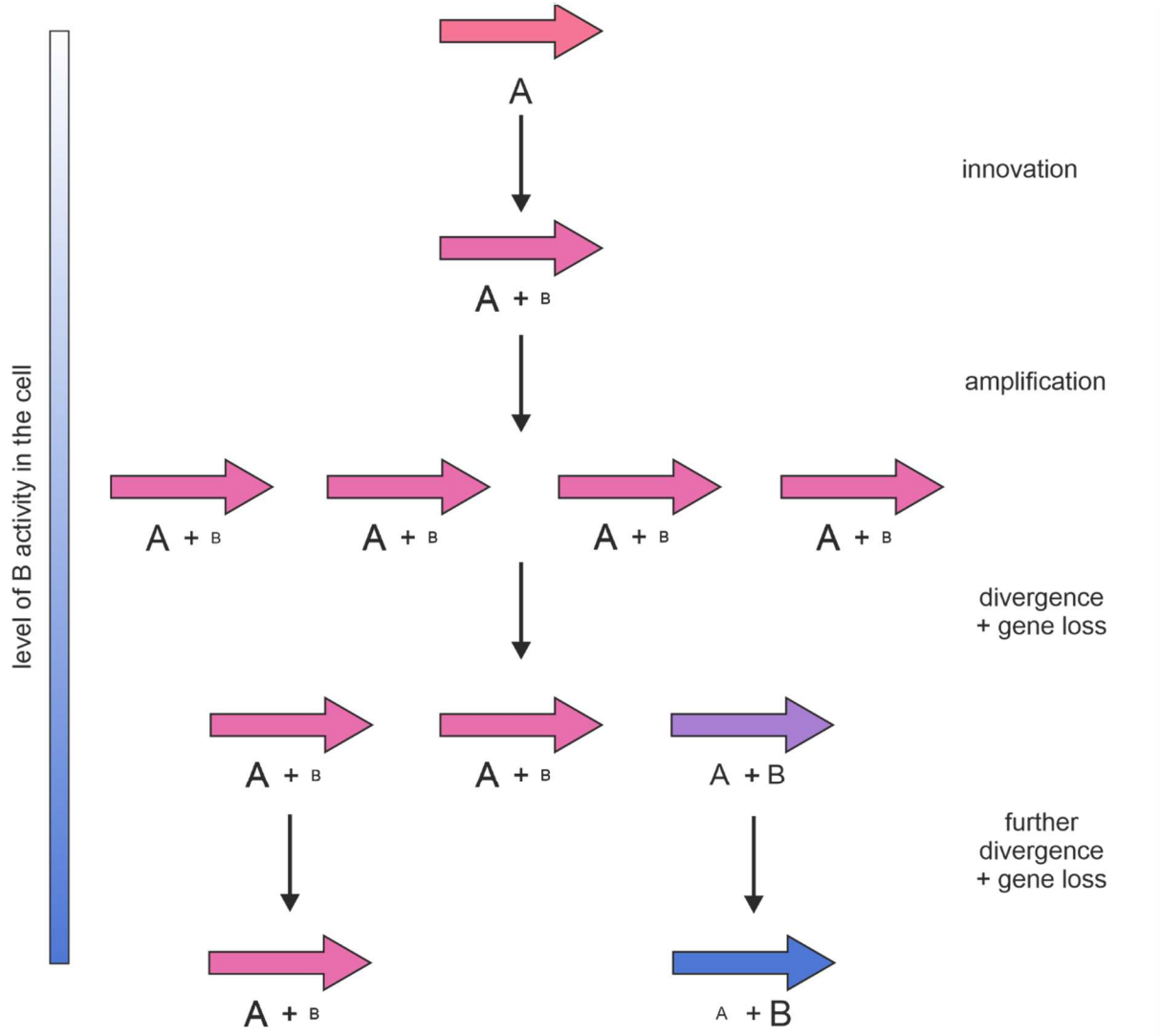

Figure 1.1: The innovation-amplification-divergence (IAD) model of evolution.

The IAD model[31] assumes a random functional innovation B against the background of a native function A on the same gene (red arrow). If this newly acquired function comes under selective pressure, amplification of the gene via several duplications can increase the level of the gene product and hence the activity level of function B. The high copy number then increases the probability for activity enhancing mutations (purple arrow), which again increase the activity level of function B and alleviate the selective pressure to retain several copies of the duplicated gene. Further functional divergence accompanied by gene loss then leads to two new genes (purple arrow, blue arrow) with distinct primary functions. The size of the letters A and B indicates the level of activity conferred by the gene product.

The IAD model postulates that a latent side activity which represents a functional innovation needs to be present prior to the duplication event. This side activity can be caused by a random mutation or simply be due to inherently limited accuracy[32, 33], but it has also been argued that a certain conformational plasticity of the protein for example mediated by flexible loop regions can lead to side activities.[34] If this side activity becomes relevant, there is a selective pressure in favor of a higher enzyme dosage which can be easily achieved by gene duplications. This selective pressure also limits the loss of the duplicated genes as this would decrease the dosage of the gene product. The parallel occurrence of multiple gene copies increases the probability for an activity enhancing mutation in one of the copies. The emergence of one gene copy with improved functionality then marks the starting point for functional divergence and the gradual improvement of the new gene abolishes the need for additional copies of the

parent gene, which is why in the end there would be only two genes retained, namely the parental gene and the new gene. Importantly, the IAD model makes no a priori assumptions on the origin of the parental gene and there is a more recent discussion which emphasizes the importance of horizontal gene transfer in the context of the evolution of new enzyme functions.[35] Between closely related species, genes are transferred with significant rates. Furthermore, horizontally transferred genes may provide new primary or promiscuous activities in the acceptor and tend to evolve faster than other genes.[35–37] Endowed with these properties, horizontally transferred genes provide starting points for the evolution of new functionalities.

Irrespective of the origin of the gene and whether the starting point for evolution was an enzyme with promiscuous side activity or a generalist enzyme as in the model of Ycas and Jensen, the evolution of a new specialist enzyme requires an increase in the turnover efficiency and specificity for the new reaction or substrate. During the evolution of life, the requirements for substrate specificity likely became more and more stringent, as the addition of new reactions also led to more reaction products that had to be distinguished from the pre-existing metabolites in order to maintain control over the different reaction pathways.[22] Therefore, an expansion of the metabolic network is closely related to the question of how substrate- and reaction-specificity can be achieved in an ever-growing network of reactions. The general requirements for substrate and reaction specificity and efficient catalysis go hand in hand and include as important factors steric and electrostatic complementarity between enzyme and substrate and optimal stabilization of the transition state.[38] In the same context, reshaping of the active site and repositioning of residues for improved substrate binding are among the most commonly observed mechanisms in enzyme evolution.[39]

But even though there is a general understanding of the factors that govern substrate specificity, it is nevertheless difficult to predict the effect of mutations within a given scaffold. The difficulties associated with the rational prediction of mutations are showcased by the success of the concept of directed evolution where randomization techniques coupled to screening or selection strategies are applied to identify enzymes with improved functions.[40, 41] Similarly, understanding the role of an amino acid in substrate binding requires close biochemical and/or bioinformatical investigation in each case. If this analysis is extended from extant enzymes, which mark the endpoints of an evolutionary process, to resurrected enzymes along an evolutionary trajectory, one can gain an in-depth understanding of the stepwise adaptation to new substrates and the evolution of a family of related proteins. If the investigated enzyme family is furthermore of ancient origin and belongs to an ancient metabolic pathway the results allow a glimpse into the early evolution of metabolism itself.

1.3 The haloacid dehalogenase superfamily – a primordial protein superfamily

Proteins are often clustered in a hierarchical manner based on their level of homology which is derived from sequence similarity and the similarity of their tertiary structure.[42–44] At a level of homology which is characterized by low sequence conservation (around 30 % sequence identity or lower), a generally conserved structure, and a similar function, proteins were combined to superfamilies. These superfamilies usually share mechanistic aspects or partial reactions even if the overall reaction and the substrates are different.[45, 46] Interestingly, all proteins can be assigned to a limited number of only about 1400 different superfamilies which indicates that gene duplications and functional divergence were the major factors that contributed to the increasing complexity of organisms.[47] Moreover, only a minor

fraction of those 1400 superfamilies were found in the majority of the organisms from each of the three kingdoms of life which suggests that this minor fraction represent protein folds of ancient origin.[48] Out of these ancient superfamilies, some are comprised of many different proteins which indicates that they were subjected to extensive duplication. A possible reason for this could be the stability of these folds and their superior tolerance of sequence divergence.[47]

One of these ancient folds that is represented by a broad array of extant proteins is the Rossman fold, which consists of a three layered αβα-sandwich[47, 48] (Figure 1.2). The Rossman fold is thought to represent one of the oldest existing folds which is underscored by the fact that proteins of ancient origin such as the aminoacyl-tRNA synthetases adopt this fold.[49, 50] Similarly, there are several protein superfamilies which adopt a Rossman fold or include domains with a Rossmann fold, as for example the NAD-dependent oxidoreductase superfamilies like the GAPDH-like superfamily, the haloacid dehalogenase (HAD) superfamily, or the class I aminoacyl tRNA synthetases. Representatives of these superfamilies can be traced back to LUCA which implies that they had already diverged previously from the most ancient founder protein of all proteins with a Rossman fold.[50]

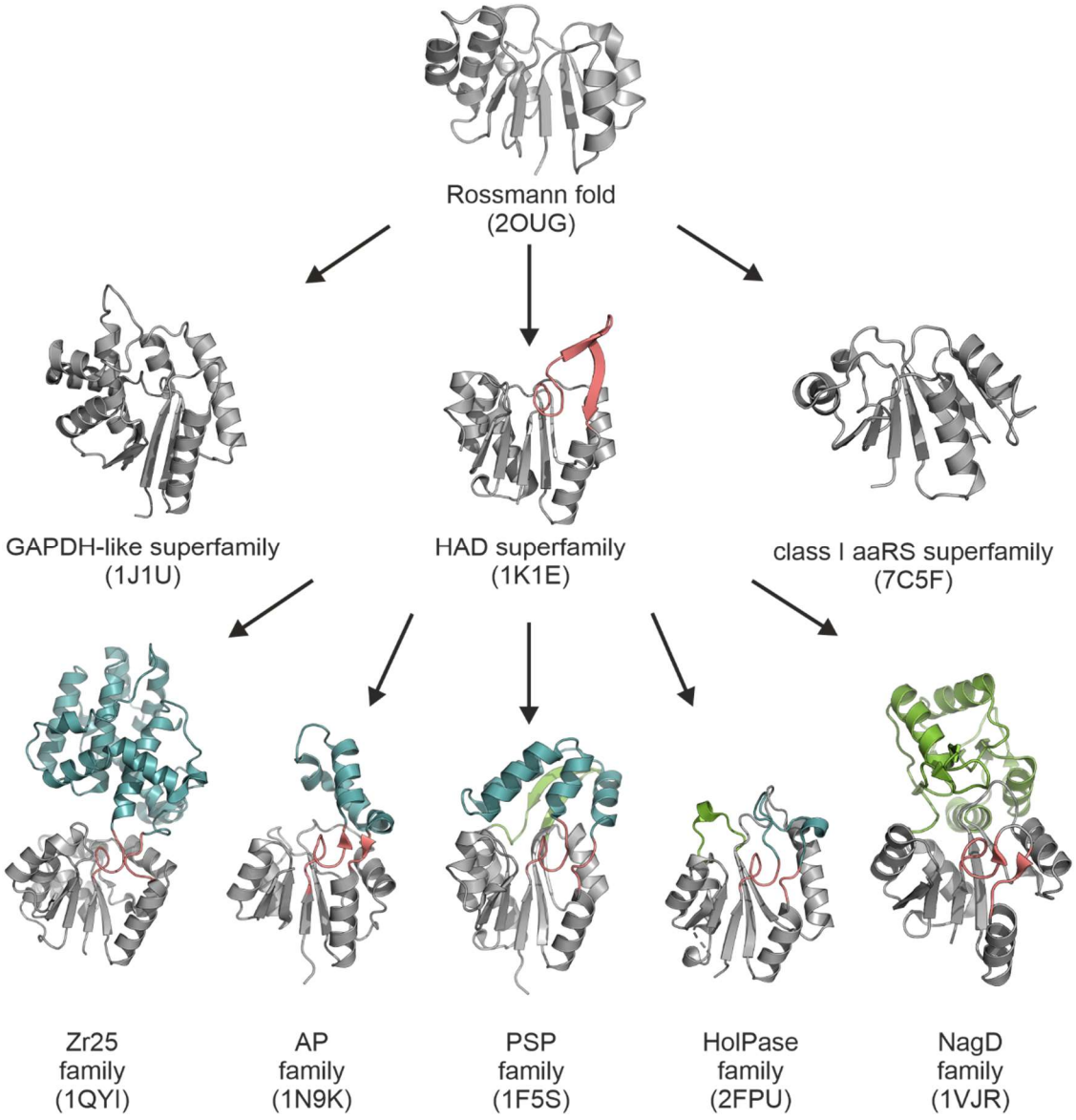

Figure 1.2: Structural diversity within the HAD Superfamily.

The basic fold of the HAD superfamily is the Rossmann fold (top), a fold which also occurs as domain in other superfamilies, for example in the GAPDH-like superfamily or in different class I aminoacyl tRNA synthetase superfamilies (aaRS). The HAD superfamily is characterized by a single helical turn followed by a β-hairpin structure (red), called squiggle and flap, which decorate the Rossmann core. The HAD superfamily gives rise to various protein families, which are exemplified by the Zr25 family, the acid phosphatase (AP) family, the phosphoserine phosphatase (PSP) family, the histidinol phosphate phosphatase (HolPase) family, and the ribonucleotide monophosphatase (NagD) family. These protein families often carry additional cap modules, which can be inserted either into the β-hairpin (cyan) or prior to a conserved lysin (green) and differ in size and structure.[51] The codes in brackets give the PDB-ID of the shown structures.[52–60]

The HAD superfamily is one of the largest and most diverse of these superfamilies[50, 51, 61]. The members of the HAD superfamily catalyze several related reactions on a vast array of substrates and include phosphatases, phosphonatases, ATPases, phosphomutases, and haloalkanoate dehalogenases which degrade xenobiotics.[51, 61, 62]

The common catalytic residue that enables all these different reactions is a conserved aspartate residue which acts as a nucleophile and forms part of a conserved DxD motif (Figure 1.3).[51, 63] In addition to the DxD motif (motif I), there are three other motifs which are characteristic for the HAD superfamily, namely a conserved threonine or serine (motif II), a conserved lysin or arginine (motif III) and a conserved DD, GDxxxD, or GDxxxxD motif (motif IV). In most enzymes, the aspartate residues of motif I and motif IV coordinate a divalent metal cation, which usually is Mg^{2+}. This magnesium ion provides a free coordination site that can be occupied by the substrate and hence aids in substrate binding and nucleophilic attack as it stabilizes the negative charge. The four conserved motifs together with the metal ion co-factor are arranged around the substrate binding pocket of the enzymes and constitute the basic catalytic machinery.[51]

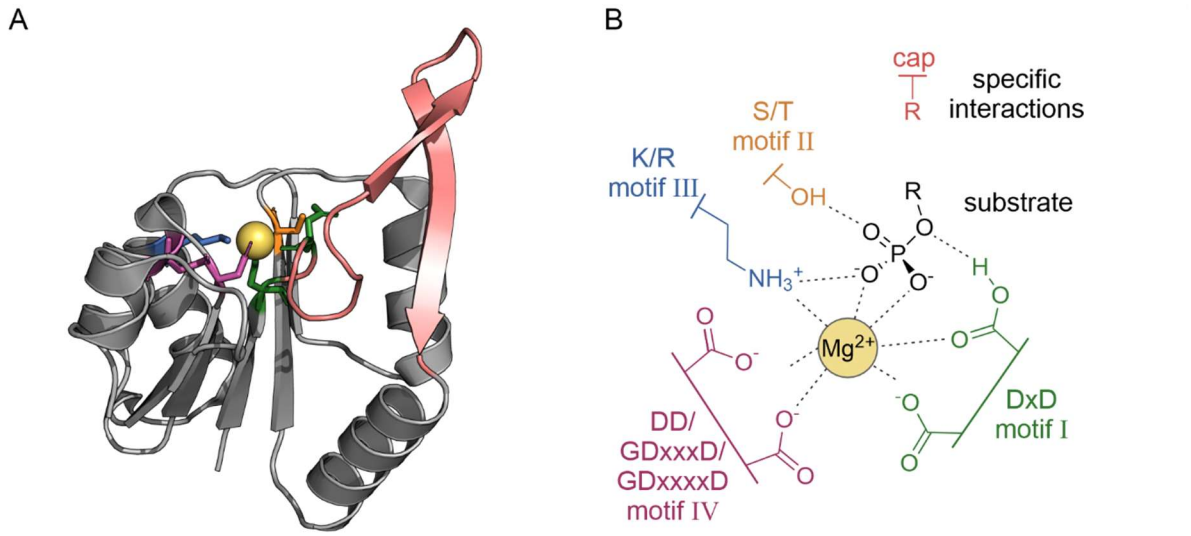

Figure 1.3: Catalytic machinery of the HAD enzymes.

(A) Structure of a representative HAD enzyme (PDB-ID 1K1E)[54] with the squiggle and flap structures colored in red. 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. Amino acids are represented by their one letter code, whereby x denotes for a random amino acid and slashes indicate variations in the sequence motifs. Dashed lines show polar, non-covalent interactions. Motifs I, III, and IV coordinate a bivalent metal ion, usually a magnesium ion, which mediates substrate binding via polar interactions. Substrate specificity is often achieved by additional residues that are part of a cap structure or the flap motif.[51]

In the most rudimentary HAD enzymes, the active site is sealed off the bulk solvent by a small lid which is located directly next to motif I and consists of two structural elements that are unique to this superfamily, namely a single helical turn and a short β-hairpin structure which were dubbed squiggle and flap.[51] Interestingly however, this small lid is often modified or complemented by additional cap modules of variable sizes and with different secondary and tertiary structures that decorate the core Rossmann fold of many HAD enzymes (Figure 1.2). These caps are either inserted into the flap or before motif III or at both positions. This means that the caps cover the active site and hence also provide additional residues that can interact with the substrate and contribute to substrate specificity.[51] One can therefore think of the HAD enzymes as modular entities with a highly conserved Rossmann fold which provides the catalytic machinery and one or two variable caps that provide the versatility that is required for the adaptation to different substrates.

Along these lines, it has been speculated, that the primordial HAD enzymes accepted nucleotides as substrates which are rather bulky and do not require extensive caps to shield the active site from the solvent.[51] It has further been hypothesized that the first specialization of some HAD family members include phosphorylated sugars, while another early specialization concerned the enzymes phosphoserine phosphatase (PSPase) and histidinol phosphate phosphatase (HolPase).[51] The latter two enzymes are involved in the biosynthesis of the amino acids serine and histidine, respectively.[64, 65] This is an interesting observation as the amino acids are the unifying building blocks of proteins and therefore cellular life in general. The evolution of the biosynthesis of such a central building block poses a chicken-and-egg dilemma, since today's amino acid biosynthesis relies on enzymes which, by circular reasoning, themselves consist of amino acids.

In the case of serine, several different pathways have been described in the literature.[66–68] It is furthermore assumed that serine can be formed abiotically and that it was probably one of the earliest amino acids to be incorporated into the genetic code.[16, 69] Taken together, these factors might complicate the evolutionary trajectory that led to the PSPase and generally blur the evolutionary picture. In contrast, for histidine there is only one biosynthetic route known.[70] The prebiotic occurrence of histidine is furthermore still questioned as it could never be identified in a Miller-Urey type experiment[71] and it is assumed that histidine was one of the last amino acids that was added to the amino acid alphabet.[16]

1.4 The HolPase – the exception within histidine biosynthesis

The HolPase catalyzes the penultimate step within the biosynthesis of histidine.[65, 72] This biosynthetic pathway consists of ten enzymatic steps which enable the synthesis of L-histidine starting from phosphoribosyl pyrophosphate (PRPP) and ATP.[70, 72] The individual reaction steps and the associated genes were first identified and characterized in *Escherichia coli* and *Salmonella typhimurium.*[70, 73–76] In these two organisms, eight different enzymes are involved in the pathway, including two fusion enzymes (HisIE and HisB), one bifunctional enzyme (HisD), and one heterodimeric enzyme complex (HisH:HisF) (Figure 1.4, upper panel).[70, 77] In both organisms, the corresponding genes are organized in a tightly regulated operon.[77]

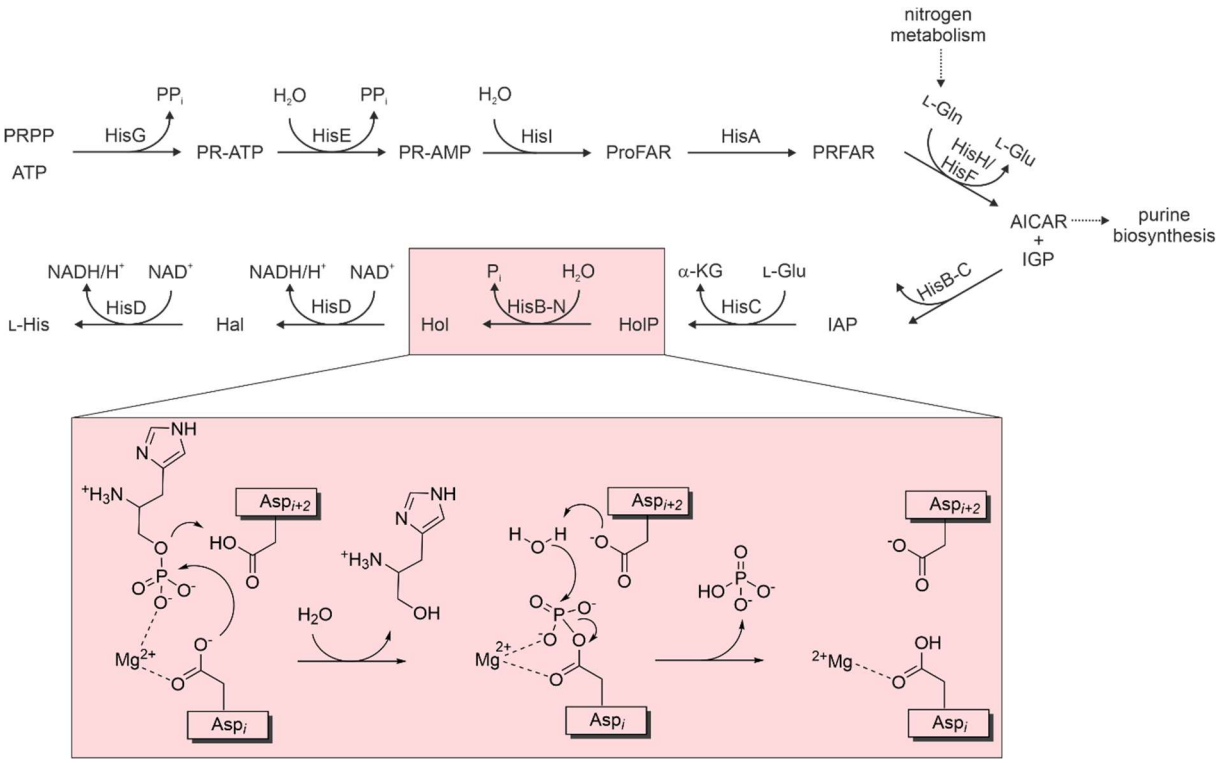

Figure 1.4: The histidine biosynthesis and the histidinol phosphate phosphatase reaction.

The upper panel shows the single chemical steps and the associated *E. coli* enzymes of histidine biosynthesis. The substrates of the pathway are 5-phosphoribosyl-α-1-pyrophosphate (PRPP) and adenosine triphosphate (ATP) which are converted to the final product L-histidine (L-His) via the intermediates N'-5'-phosphoribosyl-adenosine triphosphate (PR-ATP), N'-5'-phosphoribosyl-adenosine monophosphate (PR-AMP), N'-[(5'-phosphoribosyl)-formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR), N'-[(5'phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR), imidazole glycerol-phosphate (IGP), imidazole acetol-phosphate (IAP), histidinol phosphate (HolP), L-histidinol (Hol), and L-histidinal (Hal). As a side product of the HisH/HisF reaction 5'-phosphoribosyl-4-carboxamide-5-aminoimidazole (AICAR) is formed which is recycled in the purine biosynthesis. Side reactions include the conversion of L-glutamine (L-Glu) to L-glutamate (L-Glu), L-Glu to α-ketoglutarate (α-KG), and the reduction of nicotinamide adenine dinucleotide (NAD^+ to NADH). In *E. coli* the imidazoleglycerol-phosphate dehydratase (HisB-C) and the histidinol phosphate phosphatase (HisB-N) are encoded by a bifunctional *hisB* gene. The lower panel shows the reaction mechanism of the *E. coli* HolPase (HisB-N) which relies mostly on the two aspartate residues from motif I. The first aspartate (Asp_i) acts as a nucleophile and forms a phospho-aspartate intermediate while the second (Asp_{i+2}) functions a general acid-base catalyst.[70, 72]

Later on, comparative studies showed that the individual reactions of the pathway are identical in all organisms that synthesize histidine which includes bacteria, plants, and archaea.[70, 78] This conformity between different species also relates to the individual enzymes which are largely homologous to each other in phylogenetically diverse species from all three different kingdoms of life.[18, 72, 79] Due to this conservation of both pathway and enzymes across the tree of life, it has been argued that it was most likely assembled prior to the existence of the last universal common ancestor.[80, 81]

There is however one exception to this uniformity which is the HolPase. In the model organisms *E. coli* and *S. typhimurium*, the HolPase function is catalyzed by the N-terminal part of the bifunctional HisB enzyme and was hence termed HisB-N (Figure 1.4, lower panel), while the C-terminal part harbors the imidazole glycerol phosphate dehydratase (IGPDH).[82–86] It was however noted, that these two functions are encoded by independent genes in other organisms.[81, 87, 88] But while the monofunctional IGPDH

enzymes were homologous to the IGPDH domain of the *E. coli* enzyme [83], in many cases no homologue of the *E. coli* HolPase could be found in other organisms.[89] Consequentially, the HolPase was often the last enzyme of histidine biosynthesis that was identified[90–92] and up until now, non-homologous HolPases from three different enzyme superfamilies were discovered. Specifically, HisB-N from *E. coli* (*ec*HisB-N) exhibits a Rossmann fold and belongs to the HAD superfamily (Figure 1.5).[59]

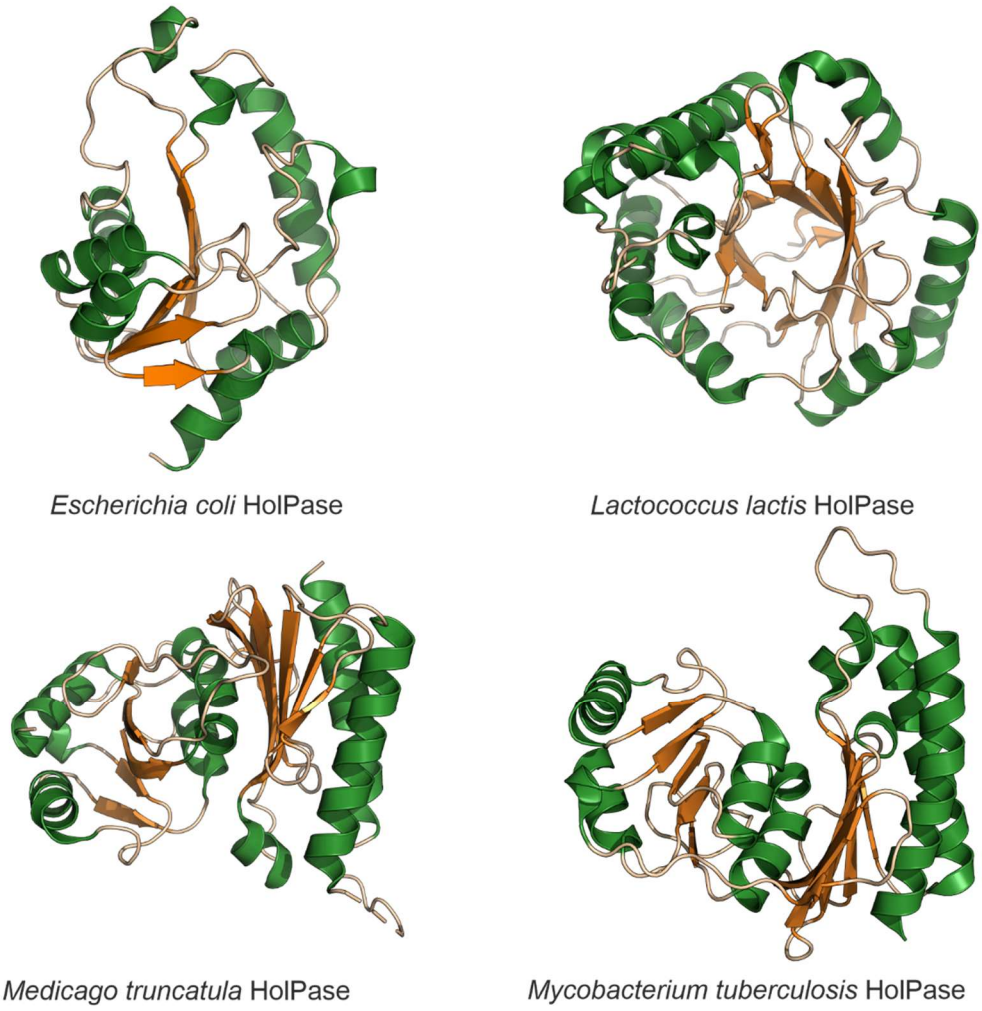

Figure 1.5: Structural diversity of different HolPases.

Shown are the crystal structures of four non-homologous bacterial HolPases. The HolPase from *E. coli* (PDB-ID 2FPU)[59] belongs to the HAD superfamily and exhibits a Rossmann fold, the HolPase from *L. lactis* (PDB-ID 4GC3)[91] belongs to the PHP superfamily and exhibits a $(\beta\alpha)_7$-barrel, The HolPases from *M. truncatula* (PDB-ID 5EQ9)[93] and *M. tuberculosis* (PDB-ID 5YHT)[94] both belong to the IMP superfamily and exhibit a αβαβα-sandwich structure.

In contrast, the monofunctional HolPase from *Lactococcus lactis* exhibits a $(\beta\alpha)_7$-barrel fold and belongs to the polymerase and histidinol phosphatase (PHP) superfamily[91], and the monofunctional HolPases from *Mycobacterium tuberculosis* and *Medicago truncatula* both adopt the fold of a αβαβα-sandwich and belong to the inositol monophosphatase (IMP) superfamily.[93, 94]

The differences between HolPases also extend to the genomic organization of their respective genes: The gene that encodes *ec*HisB-N is part of the histidine operon, whereas genes of HolPases from the PHP and IMP superfamily are typically located outside of the histidine operon.[82, 83, 90, 91, 93, 95] Taken

together, these observations indicate that HolPases differ from the other enzymes of histidine biogenesis in that they are evolutionary younger.

1.5 Objectives of this thesis

In chapter 2, the question regarding the evolutionary history of the arguably young enzyme *ec*HisB-N was addressed. In previous work, it has been argued that *ec*HisB-N and its closest homologue D-*glycero*-D-*manno*-heptose-1,7-bisphosphat-7-phosphatase (GmhB) were derived from the same promiscuous phosphatase. This hypothesis of a shared evolution between HisB-N and GmhB should be investigated by a combination of *in vivo*, *in vitro*, and *in silico* approaches.

The chapters 3 and 4 are dedicated to the search for other HolPases which have not been described yet because they may have developed independently of the currently known HolPases.

In chapter 3, this search focusses on a protein from the HAD superfamily found in the bacterium *Pseudomonas aeruginosa*, for which a HolPase function was postulated in previous work. A combination of *in vitro* and *in silico* approaches should be employed to unveil the enzymatic function. Then, the factors that are critical for the enzyme function should be determined and used to derive criteria in the form of a sequence fingerprint which can be utilized to identify iso-functional enzymes. This fingerprint should be used in combination with a sequence similarity network (SSN) to reveal the phylogenetic distribution of iso-functional enzymes. At the end of the chapter, a phylogenetic analysis should furthermore reveal phylogenetic groups of histidine-synthesizing organisms for which no HolPase is annotated.

Based on the results of this phylogenetic analysis, chapter 4 is dedicated to the search for the missing HolPase of the archaeal domain of life. A promising candidate protein from *Nitrosopumilus maritimus* which was annotated as putative phosphatase and showed little similarity to previously identified HolPases should be tested *in vitro* for its suspected HolPase function. Then, factors that are responsible for the enzymatic function should be determined, which should help to establish a fingerprint by which iso-functional enzymes should be identified. Lastly, the phylogenetic distribution of the iso-functional homologues should be revealed by means of an SSN.

Chapter 5 deals with the question of how new functions, which may be the starting point of divergent evolution, can be established. Chapter 5 is furthermore dedicated to the question of the general evolvability of a promiscuous side activity. Both questions were investigated on the scaffold of HAD proteins. In the first part of this chapter, a HolPase function should be established against the background of a PSPase, while in the second part of the chapter, a promiscuous PSPase activity in the context of a native HolPase should be enhanced by directed evolution.

2 The evolution of HisB-N from γ-Proteobacteria

2.1 Introduction

The observation that the fold and genomic organization of the HolPases are not conserved prompted several theoretical studies that aimed to rationalize the evolution of HisB-N in γ-Proteobacteria.[82, 83] Relying on multiple sequence alignments, homology, and genomic organization, it has been argued that an ancestor of HisB-N has been recruited into the histidine biosynthetic pathway after the separation of different classes of γ-Proteobacteria. It was further hypothesized that HisB-N and its closest homologue D-*glycero*-D-*manno*-heptose-1,7-bisphosphate 7-phosphatase (GmhB) were derived from the same ancestral precursor and that a gene duplication event within an ancient γ-Proteobacterium followed by divergent evolution led to the modern HisB-N and GmhB.

Indeed, both HisB-N and GmhB catalyze dephosphorylation reactions. The native substrate of HisB-N is histidinol phosphate (HolP) while GmhB enzymes catalyze the preferential dephosphorylation of one out of two anomeric sugars, namely D-*glycero*-D-*manno*-heptose-1α,7-bisphosphate (αHBP) or D-*glycero*-D-*manno*-heptose-1β,7-bisphosphate (βHBP).[96–98] While αHBP is an intermediate in the S-layer biosynthesis which is often found in gram-positive bacteria, βHBP is an intermediate in the lipopolysaccharide biosynthesis which is found in gram-negative bacteria. The preference of GmhB enzymes for one anomer over the other typically ranges from 6:1 to 1:150 (α:β).[96, 99] Here, we refer to enzymes that form part of the S-layer biosynthesis as αGmhB and to enzymes that form part of the lipopolysaccharide biosynthesis as βGmhB.

Intrigued by the above observations, we decided to investigate the hypothesis of a shared evolution between HisB-N and GmhB. Specifically, we intended to test the current model of the HisB-N evolution with a combination of *in vivo*, *in vitro*, and *in silico* approaches. First, we analyzed the extant HisB-N and GmhB enzymes from *E. coli*. To this end, we examined the structures of both enzymes and explored the substrate spectrum to check for promiscuous side activities that would indicate common ancestry. To determine the type and degree of evolutionary relationship, we calculated a phylogenetic tree based on sequences representing variants of both enzymes. Based on this tree, we additionally selected GmhB from *Crassaminicella sp*. for functional *in vitro* and *in vivo* characterization. The tree was furthermore used to reconstruct several ancestral enzymes that marked the putative functional transition between GmhB and HisB-N. The reconstructed enzymes were functionally characterized *in vitro* and the observed changes in catalytic efficiencies were rationalized by an analysis of the geometry and electrostatics of the active sites as predicted by AlphaFold.[100] The experimental data finally allowed for a detailed phylogenetic analysis, which led to a revised model for the evolution of HisB-N and GmhB in γ-Proteobacteria.

2.2 Results and Discussion

2.2.1 Sequential and structural comparison of HisB-N and GmhB

It has been postulated that HisB-N and GmhB have evolved from the same ancestral phosphatase that underwent a gene duplication event and subsequent specialization.[82, 83] This hypothesis is supported by the fact that both enzymes belong to the HAD superfamily.[99] Moreover, HisB-N and GmhB from the same organism usually show sequence identities of 26-31 %.[83] This value is significantly higher than the level of sequence identity that is normally observed for HAD enzymes with different functions[51] and indicates a pronounced structural similarity.[101] The structures of the respective *E. coli* enzymes (*ec*HisB-N and *ec*GmhB)[59, 102] indeed exhibit the same basic Rossmann fold that is characterized by an αβα-sandwich being typical for the HAD superfamily (Figure 2.1 A, B).

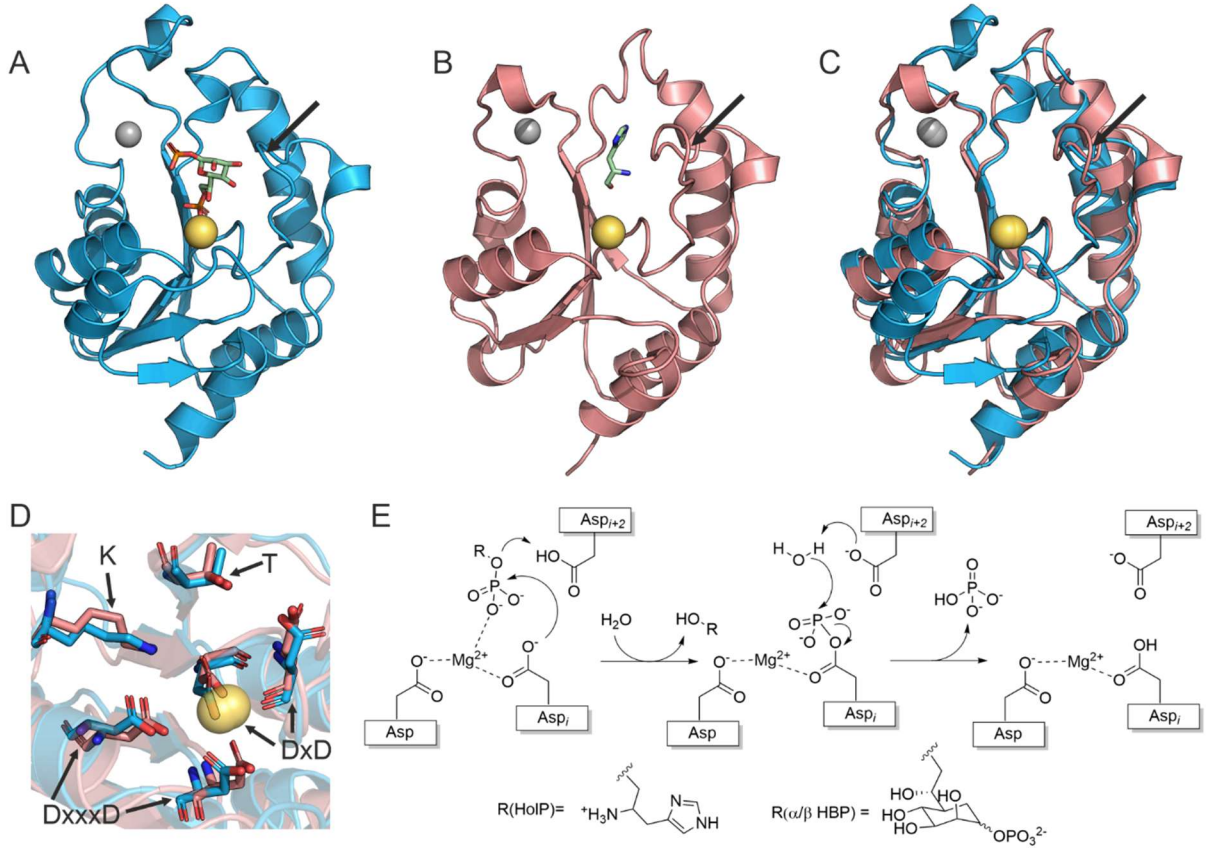

Figure 2.1: Structural and functional comparison of *ec*HisB-N and *ec*GmhB.

(A) Crystal structure of *ec*GmhB with bound substrate βHBP (PDB-ID: 3L8G) and (B) crystal structure of *ec*HisB-N with the bound product L-histidinol (PDB-ID: 2FPU).[59, 102] Substrates and products are depicted in stick representation, while yellow spheres represent a catalytic magnesium ion and grey spheres represent a zinc ion that is coordinated by a CxH and a CxC motif within a small cap-structure which is unique within the protein superfamily. Black arrows indicate the so called flap[51] which forms a lid-like loop that is involved in substrate binding. (C) Overlay of the unliganded *ec*GmhB (blue) and unliganded *ec*HisB-N (red), yielding an RMSD of 1.2 Å for 91 superimposed C_α-atoms. (D) Zoomed in view of the active sites of *ec*GmhB and *ec*HisB-N with the catalytically relevant residues and residue motifs depicted as sticks. (E) Reaction mechanisms of HisB-N and GmhB for the dephosphorylation of histidinol phosphate (HolP) and of the two anomers of D-*glycero*-D-*manno*-heptose-1,7-bisphosphate (α/β HBP) by the catalytic aspartates from the DxD motif (Asp$_i$, Asp$_{i+2}$) and one aspartate (Asp) from the DxxxD motif.[59, 100, 103]

Moreover, both enzymes exhibit a zinc-binding cap structure consisting of a CxH motif and a CxC motif which are separated by 5 and 12 residues in *ec*HisB-N and *ec*GmhB, respectively. This zinc-binding cap structure comprises a unique structural feature within the HAD superfamily.[103] The active sites of both enzymes are confined by a lid-like loop structure which has been dubbed flap and which is involved in substrate binding.[51] A superposition of the two proteins showed that the structures are highly similar and the RMSD was calculated to be 1.2 Å (Figure 2.1 C). The similarities between the two enzymes also extends to the four catalytically important motifs which are occupied by identical amino acids in *ec*GmhB and *ec*HisB-N and the respective side chains have the same geometrical orientation in the structures (Figure 2.1, D). Moreover, both HisB-N and GmhB share the same reaction mechanism, which involves substrate binding with coordinate bond formation to a magnesium ion, nucleophilic attack by an aspartate residue on the phosphorus atom, formation of a covalent phospho-aspartate intermediate, and subsequent hydrolysis to yield free phosphate and restore the aspartate residue (Figure 2.1 E).[59, 51, 103] Taken together, the sequential, structural, and mechanistic similarities indicate a close relationship between HisB-N and GmhB within the superfamily, supporting the hypothesis of a common evolution. Depending on the phylogenetic distance and degree of divergence, this could translate to a shared substrate spectrum. To clarify this issue, we tested *ec*HisB-N and *ec*GmhB for promiscuous side activities.

2.2.2 Purification and functional characterization of *ec*HisB-N and *ec*GmhB

The corresponding genes for *ec*HisB-N and *ec*GmhB were overexpressed in *E. coli* (8.4.2, 8.4.3) and the proteins were purified by affinity chromatography via N-terminal His_6-tags (8.4.4), followed by size exclusion chromatography (8.4.5, sequences of the constructs are given in Table S 1). The purity and structural integrity of both enzymes was assessed by SDS-PAGE (8.5.2, Figure S 1) and far-UV CD-spectroscopy (8.5.3, Figure S 2).

To ensure functional integrity, we first determined the steady-state enzyme catalytic parameters for *ec*HisB-N and its natural substrate HolP. To this end, we used a coupled photometric assay (8.5.5) that allowed for continuous measurement of the formation of free phosphate.[104] A hyperbolic substrate saturation curve was obtained, which yielded a turnover number (k_{cat}) of 2.8 s^{-1}, a Michaelis constant (K_M) of 48 µM, and a catalytic efficiency (k_{cat}/K_M) of 58,000 $s^{-1}M^{-1}$ (Figure 2.2 A). Both the Michaelis constant and the turnover number are in good accordance with previously reported values, which are 54 µM for the K_M of *ec*HisB-N[59] and 1-4 s^{-1} for the k_{cat} of different monofunctional HolPases[93, 94, 105].

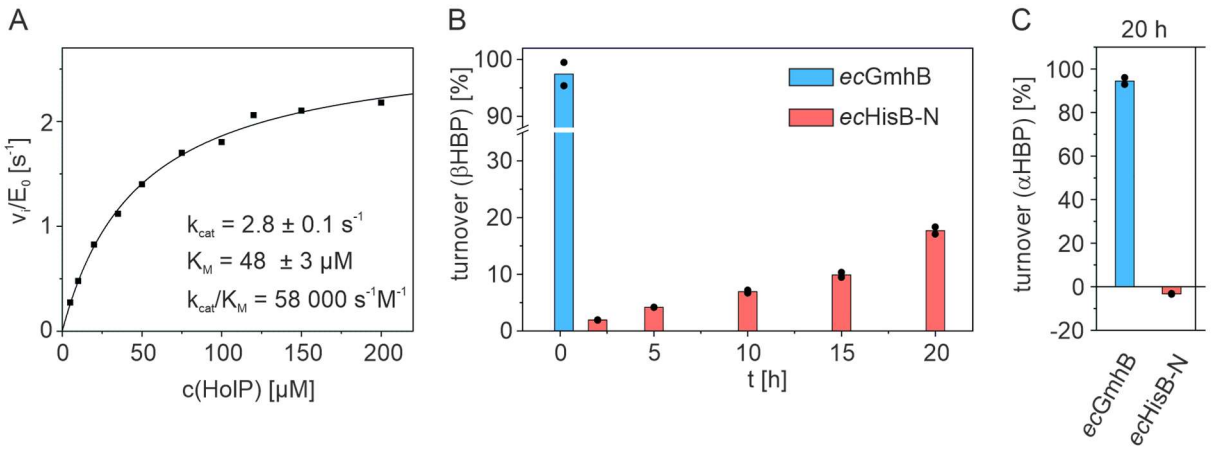

Figure 2.2: Functional characterization of *ec*HisB-N at 25 °C.

(A) Substrate saturation curve for the turnover of HolP. (B) Average percentage of product formation for the dephosphorylation of βHBP as detected by two independent measurements (black dots). Shown is the dephosphorylation of βHBP (30 µM) by *ec*HisB-N (10 µM) or *ec*GmhB (10 nM). For *ec*HisB-N a promiscuous turnover of 18 % substrate was observed within 20 h, whereas for *ec*GmhB 97 % turnover of its preferred anomer was observed within 12 min. (C) Average percentage of product formation for the dephosphorylation of αHBP as detected by two independent measurements (black dots). Shown is the turnover of αHBP (30 µM) by *ec*HisB-N (10 µM) or *ec*GmhB (10 µM). For *ec*HisB-N, no promiscuous turnover of αHBP could be detected, whereas for *ec*GmhB 94 % turnover of its non-preferred anomer was observed within 20 h.

We then checked if *ec*HisB-N also catalyzes the turnover of any of the two anomeric substrates of GmhB. To this end, we synthesized both αHBP and βHBP according to a revised version of a published protocol.[99] Then, *ec*HisB-N was incubated with either αHBP or βHBP for up to 20 h, followed by quantification of newly formed free phosphate using the aforementioned photometric assay (8.5.6). As control experiments, αHBP or βHBP were incubated with *ec*GmhB or buffer. The amount of detected phosphate in the buffer control either caused by impurities or by spontaneous hydrolysis was subtracted from all other measurements. Interestingly, a low promiscuous activity of *ec*HisB-N could be measured for βHBP (Figure 2.2 B), resulting in the hydrolysis of approximately 18 % of the substrate after 20 h. However, no turnover of the anomeric αHBP by *ec*HisB-N could be detected (Figure 2.2 C). This finding supports the hypothesis of a shared evolution between HisB-N enzymes and GmhB enzymes and suggests that HisB-N is more closely related to βGmhBs than to αGmhBs.

In the next step, we determined the steady-state enzyme catalytic parameters of *ec*GmhB for its native substrate βHBP and the anomeric αHBP (8.5.5). Hyperbolic substrate saturation curves were obtained in both cases. For βHBP, a k_{cat} of 36 s^{-1} and a K_M of 2.3 µM were determined, yielding a k_{cat}/K_M of $15.5 \cdot 10^6$ $s^{-1}M^{-1}$ (Figure 2.3 A). For αHBP, a k_{cat} of 3.5 s^{-1} and a K_M of 117 µM were determined, yielding a k_{cat}/K_M of $29.9 \cdot 10^3$ $s^{-1}M^{-1}$ (Figure 2.3 B). These values are in good accordance with previously reported results.[99]

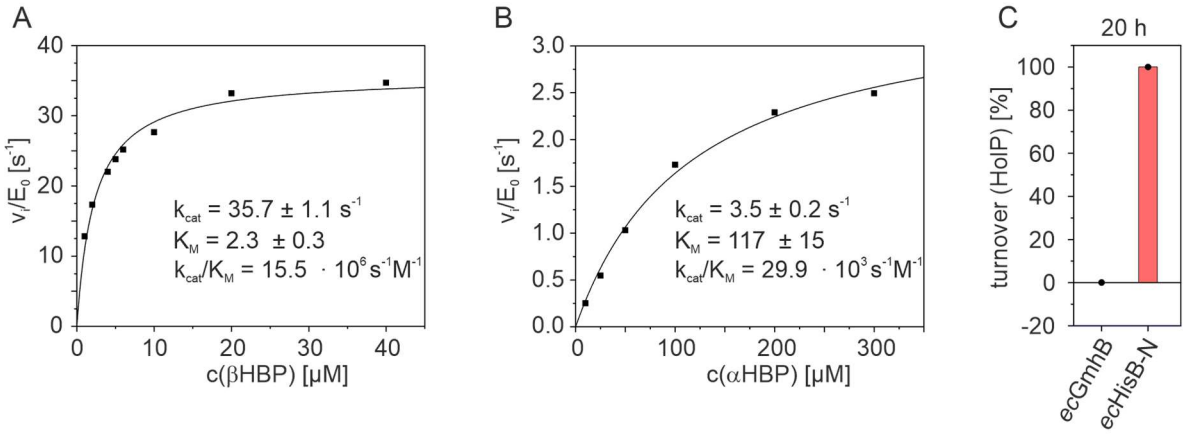

Figure 2.3: Functional characterization of *ec*GmhB at 25°C.

(A) Substrate saturation curve for the turnover of βHBP. (B) Substrate saturation curve for the turnover of αHBP. (C) Average amount of product formation for the dephosphorylation of HolP (200 µM) as detected by two independent measurements (black dots) that were performed with *ec*GmhB (10 µM) or *ec*HisB-N (10 nM). For *ec*GmhB no promiscuous turnover of HolP could be detected, whereas for *ec*HisB-N 100 % turnover was observed within 20 h.

In contrast, no product formation was detectable following the incubation of HolP with *ec*GmhB within 20 h (Figure 2.3 C). In accordance with these findings, an *E. coli* knock-out strain lacking the gene for the HolPase could not be rescued by a plasmid-encoded *ecgmhB* gene (Figure S 3). The accepted substrates and all experimentally accessible steady-state kinetic parameters of *ec*HisB-N and *ec*GmhB are summarized in Table 2.1.

Table 2.1: Activity towards HolP, βHBP, and αHBP of *ec*HisB-N, *ec*GmhB, and *cs*GmhB at 25 °C.

Enzyme	Substrate	k_{cat} [s^{-1}]	K_M [µM]	k_{cat}/K_M [s^{-1} M^{-1}]
	HolP[1]	2.8 ± 0.1	48 ± 3	58 000
*ec*HisB-N	βHBP[2]	18 % turnover (10 µM enzyme, 30 µM substrate, 20 h)		
	αHBP[2]	No detectable turnover (10 µM enzyme, 30 µM substrate, 20 h)		
	HolP[2]	No detectable turnover (10 µM enzyme, 200 µM substrate, 20 h)		
*ec*GmhB	βHBP[1]	35.7 ± 1.1	2.3 ± 0.3	15 500 000
	αHBP[1]	3.5 ± 0.2	117 ± 15	29 900
	HolP[2]	27 % turnover (10 µM enzyme, 200 µM substrate, 20 h)		
*cs*GmhB	βHBP[1]	0.5 ± 0.03	2.4 ± 0.5	210 000
	αHBP[1]	0.06 ± 0.001	10.2 ± 0.6	5 880

[1] data from steady-state kinetic experiments
[2] data from discontinuous activity assays

From an enzymological point of view, these findings were surprising; first, because βHBP is significantly larger and sterically more demanding than HolP, therefore making it unlikely that its turnover by *ec*HisB-N was a fortuitous event. Even more so, as it was shown before that *ec*HisB-N exhibits an unusually high substrate specificity within the HAD superfamily.[106, 28] On the other hand, HolP is much smaller than βHBP and should therefore fit into the binding pocket of *ec*GmhB. Hence, assuming a common promiscuous ancestor, it was unexpected that *ec*GmhB would lose any promiscuous side activity while it would be preserved in *ec*HisB-N. To help with the interpretation of these *in vitro* data, we decided to perform a phylogenetic analysis of HisB-N and GmhB.

2.2.3 Phylogenetic analysis

The promiscuous side activity of *ec*HisB-N for βHBP supports the hypothesis of a close evolutionary relationship between HisB-N and GmhB, whereas the lack of any activity of *ec*GmhB towards HolP suggests a distant relationship. To determine the degree of evolutionary relatedness and to elucidate the nature of the putative common ancestor, a comprehensive phylogenetic tree including HisB-N, αGmhB, and βGmhB sequences was deduced.

In a first step, we retrieved a comprehensive dataset comprising HisB-N, αGmhB, and βGmhB sequences from the KEGG database (8.1.1).[107] The annotation regarding substrate preference was incomplete, which rendered it difficult to discriminate between αGmhB and βGmhB. To solve this issue, we checked for the occurrence of either the lipopolysaccharide biosynthesis or S-layer biosynthesis in the corresponding host organism, which then allowed us to classify an enzyme as αGmhB or βGmhB. It is noteworthy that GmhB sequences were found for a wide variety of bacteria, amongst others in Terriglobales (previously Acidobacteriales), α-, β-, γ-, δ-, and ε-Proteobacteria, Bacilli, Bacteroidia (previously Bacteroidetes), Clostridia, Mycobacteriales (previously Corynebacteriales), Micrococcales, Negativicutes, Sphingobacteriia, Kitasatosporales (previously Streptomycetales), Synechococcales, and Thermodesulfobacteriales, whereas HisB-N sequences were only found in γ-Proteobacteria, ε-Proteobacteria and Bacteroidia. The broader phylogenetic distribution of GmhB sequences suggests that this function was established earlier than the HisB-N function. This is in line with the observation that GmhB enzymes are found in gram-positive and gram-negative bacteria, which already suggests that their precursor enzyme originated prior to the separation of these two bacterial groups.

The initially retrieved sequences were filtered to reduce the overrepresentation of preferentially sequenced phyla and to minimize other biases. The resulting set of sequences was then used to calculate a phylogenetic tree using a consensus approach based on Bayesian phylogenetics (Figure 2.4 A, Figure S 4).

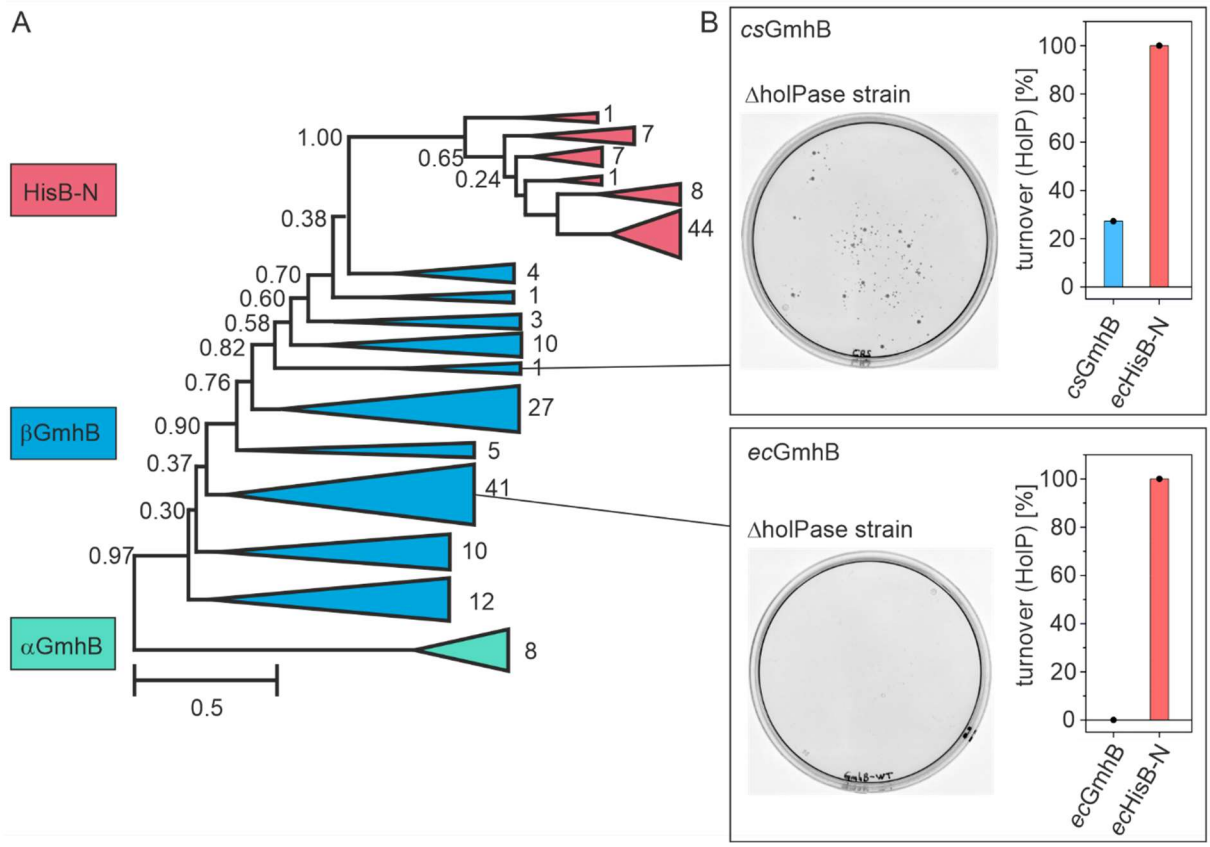

Figure 2.4: Phylogenetic analysis of αGmhB, βGmhB, and HisB-N enzymes and HolPase activities of two βGmhB representatives.

(A) Condensed phylogenetic tree deduced from a representative set of αGmhB, βGmhB, and HisB-N sequences using a consensus approach based on Bayesian phylogenetics. The putatively first functional divergence led to a separation of extant αGmhB (cyan) and βGmhB (blue) enzymes, while the second divergence gave rise to the extant HisB-N enzymes (red) in a sub-branch of the βGmhB cluster. Values to the left of each branch indicate the posterior probabilities, values to the right give the number of sequences in each cluster, and the scale bar shows the mean mutation rate per site. The clustering of αGmhB and of HisB-N sequences are both supported by highly significant posterior probabilities. Rooting was performed with the help of the αGmhB group. (B) *In vivo* and *in vitro* HolPase activities of GmhB from *Crassaminicella sp*. (*cs*GmhB) (upper panel) and *ec*GmhB (lower panel). A plasmid encoding *cs*GmhB is able to rescue an *E. coli* Δ*holPase* knock-out strain and *cs*GmhB (10 µM) shows 27 % HolP turnover within 20 h. In contrast, *ec*GmhB lacks HolPase activity both *in vivo* and *in vitro*. For *ec*HisB-N (10 nM), 100 % HolP turnover was observed.

The calculations resulted in a reliable phylogenetic tree that included all three enzyme functions. The reliability is supported by high posterior probabilities of the splits between the three enzyme functions and by a robust overall topology given different sets of sequences, which we have seen during our sequence selection process. One particular branchpoint, which is supported by a posterior probability of 0.97, discriminates between an isolated cluster of αGmhB enzymes and several clusters that include both βGmhB and HisB-N representatives. The clusters closest to this branchpoint are populated by αGmhB and βGmhB enzymes. This probably was the first functional divergence and discriminates between S-layer and lipopolysaccharide biosynthesis. Thus, this separation was used for rooting the tree. Interestingly, the HisB-N enzymes arise as a distinct cluster within the βGmhB branch. This topology indicates a closer evolutionary relationship between βGmhB and HisB-N than between αGmhB and

HisB-N. This relationship is in line with the finding that *ec*HisB-N shows promiscuous activity for βHBP but not for αHBP (Figure 2.2 B, C).

Based on these results, we tested whether βGmhBs that are located closer to the HisB-N cluster (i.e., separated by fewer nodes) than *ec*GmhB show detectable HolPase activity. Indeed, we could identify a low promiscuous HolPase activity for GmhB of *Crassaminicella sp*. (*cs*GmhB, Figure 2.4 B, upper panel). The purity and structural integrity of the enzyme was assessed by SDS-PAGE (8.5.2, Figure S 5) and far-UV CD-spectroscopy (8.5.3, Figure S 6). Moreover, steady-state enzyme kinetics showed that *cs*GmhB hydrolyzed its native substrate βHBP with high catalytic efficiency and exhibited a 36-fold preference for βHBP over αHBP (8.5.5, Figure S 7, Table 2.1).

2.2.4 Ancestral sequence reconstruction

The promiscuous HolPase activity of *cs*GmhB could have arisen by chance but it could also be an inherited property that was preserved over the course of evolution. The latter would lead to the hypothesis that HolPase activity was a common feature of ancient progenitor enzymes. With the aim to test this, we decided to perform ancestral sequence reconstruction based on the phylogenetic tree shown in Figure S 4 and to functionally characterize the resurrected enzymes. To conduct a thorough analysis of the evolutionary trajectory that led to the modern HisB-Ns, we reconstructed a set of seven enzymes which were dubbed Anc1-Anc7 from the oldest to the youngest variant (Figure 2.5 A, B, sequences of Anc1-Anc7 are given in Table S 1).

The posterior probabilities for several associated ancestral nodes are moderate, which is caused by some uncertainty in the positioning of the extant enzymes that are derived from these nodes (Figure 2.5 A). One could in principle increase the posterior probabilities by neglecting extant enzymes and working with a reduced data set. However, this would also reduce the number of ancestral nodes and raise the number of mutations between any pair of ancestral enzymes. Yet in this case, the ancestral enzymes were already separated by a considerable number of mutations (Figure 2.5 B), which implies the vast evolutionary distance that is covered by this phylogenetic tree. To balance the trade-off between robustness and evolutionary detail, we opted against a further reduction of the dataset. Despite the high number of mutations between the different ancestors, the individual sequences are supported by high marginal ancestral probabilities across all residues with median values ≥ 99.8 % for all seven reconstructed ancestors (details are given in Table S 2).

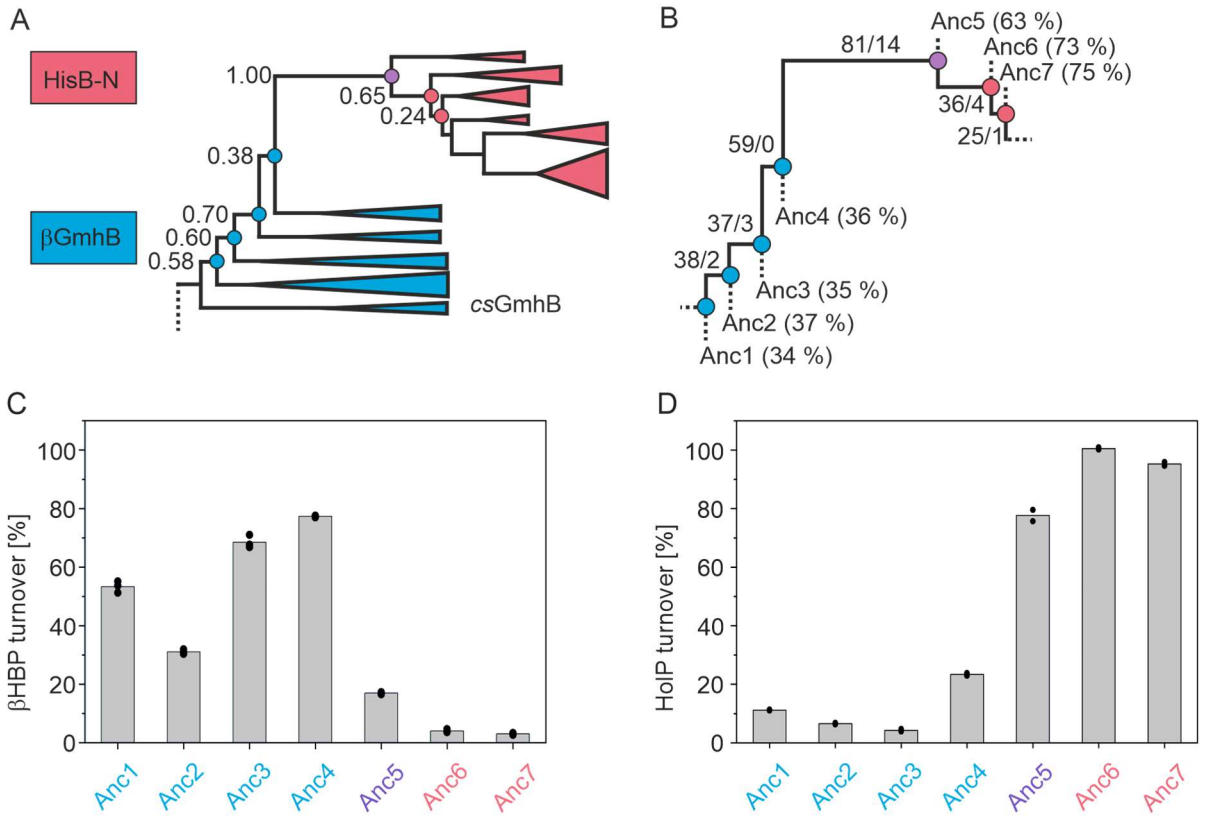

Figure 2.5: Ancestral sequence reconstruction and functional characterization of resurrected progenitor enzymes.

(A) Detailed view of the abstracted phylogenetic tree shown in Figure 2.4 A; reconstructed nodes are marked with circles and posterior node probabilities are given next to each edge. (B) Position in the phylogenetic tree of the ancestors Anc1-Anc7 that were reconstructed using a maximum likelihood approach. The first number at each branch indicates the number of mutations and the second number gives the number of insertions or deletions between two adjacent variants. Numbers in brackets give the percent sequence identity of Anc1-Anc7 as compared to *ec*HisB-N. (C) Average amount of product formation within 20 h for the dephosphorylation of βHBP (30 µM) as detected by three independent measurements (black dots) that were performed with Anc1-Anc7 (5 µM). Incubation of βHBP with Anc1-Anc4 results in conversion rates above 30 %, while for Anc5-Anc7 the conversion is decreasing to below 20 %. (D) Average amount of product formation within 20 h for the dephosphorylation of HolP (55 µM) as detected by two independent measurements (black dots) that were performed with Anc1-Anc7 (10 µM). Incubation of HolP with Anc1-Anc5 results in partial conversion of the substrate, while Anc6 and Anc7 show more than 95 % conversion.

The genes for Anc1-Anc7 were codon optimized for the expression in *E. coli* and cloned into plasmids encoding for an N-terminal His_6-tag (8.3.4.5, sequences of the constructs are given in Table S 1). The genes were overexpressed (1018.4.2, 8.4.3), and the proteins were purified by affinity chromatography (8.4.4) followed by size exclusion chromatography (8.4.5). For Anc6 and Anc7, highly pure protein was obtained (8.5.2, Figure S 8). However, for Anc2-Anc4 the protein levels as judged by SDS-Page were very low, possibly due to misfolding and subsequent degradation. Similarly, for Anc1 and Anc5 the purity of the target proteins remained limited, which obstructed further characterization. Therefore, Anc1-Anc5 were subcloned (8.3.4.5) into an expression plasmid which encoded for an N-terminally fused maltose binding protein (MBP) which served as a solubility tag. With this tag, Anc1-Anc5 could be obtained with good purity (Figure S 8). Moreover, far-UV CD-spectra indicated that all resurrected proteins were properly folded (8.5.3, Figure S 9).

In the next step, we functionally characterized Anc1-Anc7. For this purpose, proteins were incubated with either βHBP or HolP, followed by quantification of product formation (8.5.6). Remarkably, all reconstructed enzymes were able to catalyze the turnover of both substrates (Figure 2.5 C, D). Regarding βHBP hydrolysis, it is of note however, that none of the variants showed complete product formation within 20 h. This result may seem surprising at least for Anc1-Anc4, which are the immediate precursors of extant βGmhBs. A plausible explanation for incomplete product formation would be that only a sub-fraction of the proteins was properly folded and active even in the presence of MBP. The recorded far-UV CD-spectra do however not show any indication for a large fraction of disordered regions. Another possible explanation lies in the moderate posterior probabilities linking Anc1 with Anc4 (Figure 2.5 A) and the finding that the erroneous prediction of only some relevant residues might already lead to a dramatic drop in activity.[108, 109] Regarding HolP hydrolysis, Anc5-Anc7 accomplish almost complete product formation within 20 h, which is in accordance with their close phylogenetic proximity to extant HisB-Ns. To further quantify the observed HolP turnover by Anc5-Anc7, steady-state enzyme kinetic measurements were performed (8.5.5, Figure S 10). The determined catalytic parameters are listed in Table 2.2.

Table 2.2: Activity towards HolP of Anc1-Anc7 and *ec*HisB-N at 25°C.

Enzyme	Substrate	k_{cat} [s^{-1}]	K_M [µM]	k_{cat}/K_M [s^{-1} M^{-1}]
Anc1-4	HolP[2]	4 – 24 % turnover (5 µM enzyme, 55 µM substrate, 20 h)		
Anc5	HolP[1]	$(15 \pm 1.7) \times 10^{-3}$	698 ± 123	21
Anc6	HolP[1]	1.8 ± 0.15	59 ± 12	30 508
Anc7	HolP[1]	0.8 ± 0.08	122 ± 31	6 557
*ec*HisB-N	HolP[1]	2.8 ± 0.07	48 ± 3.3	57 437

[1] data from steady-state kinetic experiments
[2] data from discontinuous activity assays

While the activities of Anc1-Anc4 were too low for reliable measurements, the kinetic parameters of Anc5-Anc7 could be determined. Anc5 exhibits a k_{cat} value of 0.015 s^{-1} and a K_M of approximately 700 µM, which is a drastic improvement over Anc1-Anc4 but still constitutes a moderate activity as compared to the extant *ec*HisB-N. Anc6 and Anc7 display k_{cat} values of 1.8 and 0.8 s^{-1} and K_M values of 59 µM and 122 µM. This indicates both a drastic improvement in the turnover number and a significant improvement in affinity for HolP as compared to Anc5. Taken together, Anc6 and Anc7 showed HolPase activities that are similar to the HolPase activity of the extant *ec*HisB-N.

In summary, the *in vitro* analysis showed that the HolPase activity was already present as a side activity in the ancestors that precede the branchpoint between HisB-N and βGmhB, while high catalytic activity was established only after the separation of the HisB-N cluster in the phylogenetic tree.

2.2.5 Structural analysis of the predecessors Anc1-Anc7

The functional analysis of the ancestral sequences revealed a drastic increase in the HolPase activity during the evolution from Anc1-Anc4 to Anc5-Anc7. With the aim to rationalize the causes for this observation, the structures of Anc1-Anc7 were predicted with AlphaFold[100] and compared to the structures of *ec*HisB-N and *ec*GmhB (Figure 2.6).

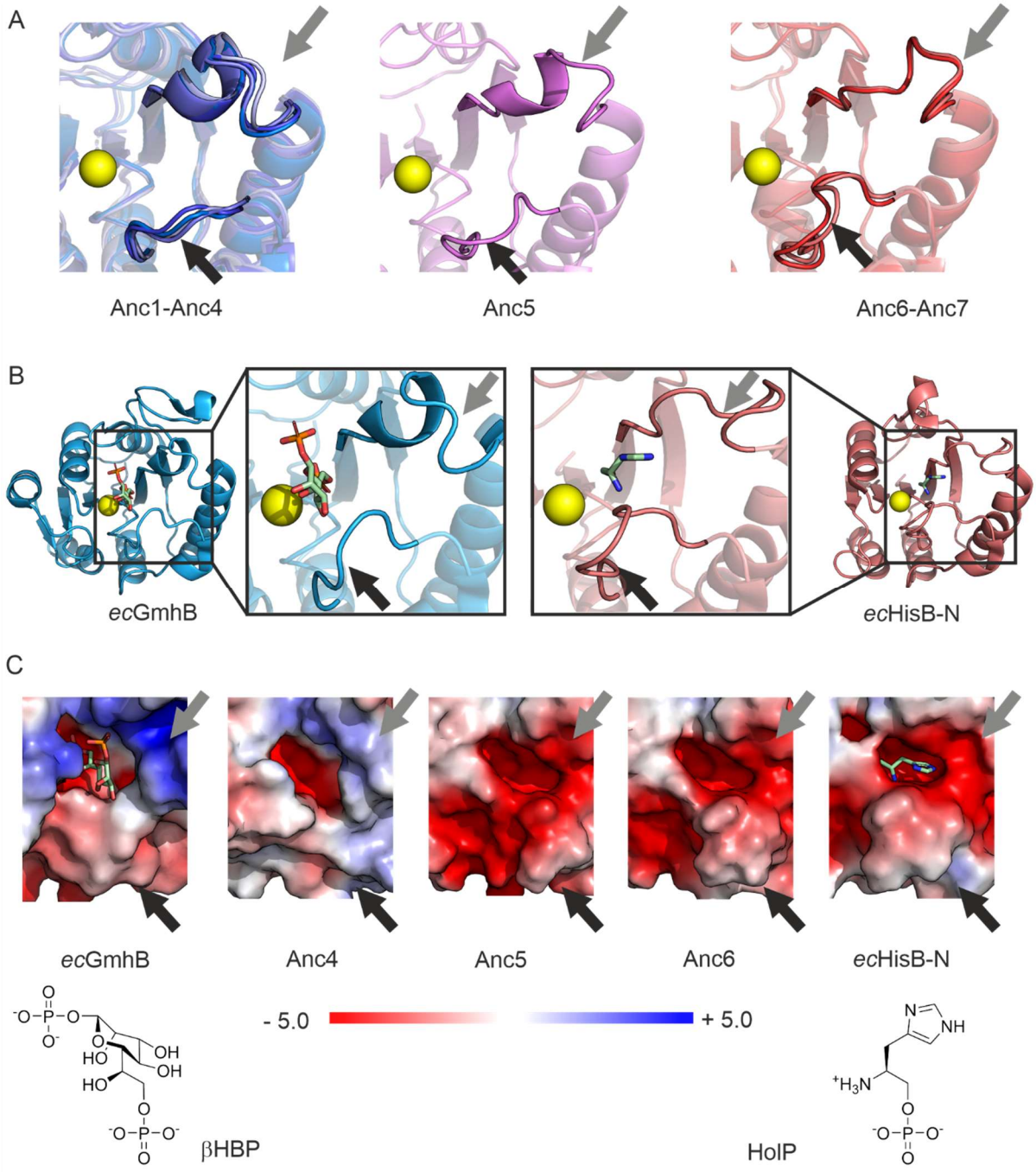

Figure 2.6: Structural analysis of fold, shape, and charge distribution of the active sites of Anc1-Anc7.

(A) Active sites of the structures of Anc1-Anc4 (superimposed in the left panel), of Anc5 (middle panel), and of Anc6-Anc7 (superimposed in the right panel) as predicted by AlphaFold.[100] The yellow spheres represent magnesium ions that were extracted from a superimposed *ec*HisB-N structure. The black arrows indicate the flap, the length of which is increased from Anc1 to Anc7. The increased length was probably necessary to ensure efficient closure of the active site while the size of the substrate decreased.

The grey arrows mark a region which folds as helix in Anc1-Anc4 and as loop in Anc5-Anc7. (B) Overall structures and zoomed-in active site views of *ec*GmhB (PDB-ID: 3L8G) with bound βHBP (green sticks) and *ec*HisB-N (PDB-ID: 2FPU) with bound histidinol (green sticks). (C) Surface shape and charge distribution of *ec*GmhB, Anc4, Anc5, Anc6, and *ec*HisB-N and the two substrates βHBP and HolP. In *ec*GmhB and Anc4, a short helix limits the space towards the right (grey arrows) whereas in Anc5, Anc6, and *ec*HisB-N a loop creates a bigger binding pocket towards the right (grey arrows). Moreover, in *ec*GmhB and Anc4 the surface exposed parts of the active sites are positively charged, allowing for the interaction with the second phosphate group and hydroxyl groups of βHBP, while they are negatively charged in Anc5-Anc6 and *ec*HisB-N allowing for the interaction with the amine group and imidazole ring of HolP.

When comparing the structures of Anc1-Anc7, it became evident that they exhibit the same overall fold with only minor differences in the distal parts of the enzymes (Figure S 11). Moreover, a close inspection of the active site revealed that the subsite closest to the phosphate binding site also showed high similarity (Figure 2.6 A, yellow spheres).

At the solvent exposed parts however, the enzymes differed significantly, which allowed for the distinction of three groups, namely Anc1-Anc4, Anc5, and Anc6-Anc7. In Anc1-Anc4, the loop that forms the flap is four residues shorter than the loop in Anc5 which is again one residue shorter than the loop in Anc6-Anc7 (Figure 2.6 A, black arrow, loop sequences are given in Table S 3). This is in accordance with the observation that in *ec*GmhB the flap structure is three residues shorter than in *ec*HisB-N (Figure 2.6 B, Table S 3). This finding fits to an earlier argument according to which a bulky substrate requires a small flap structure, whereas a small substrate requires a larger flap structure that ensures the exclusion of water from the active site.[51]

In addition to that, Anc1-Anc4 differ from Anc5 and Anc6-Anc7 in the top right part of the active site (Figure 2.6 A-C, Figure S 12, grey arrows). The secondary structure in Anc1-Anc4 was predicted as a short helix, whereas in Anc5-Anc7 there is a short loop (Figure 2.6 A, grey arrows). In *ec*GmhB the corresponding short helix fills the space thereby limiting the active site (Figure 6B). In contrast, in *ec*HisB-N the corresponding loop region creates a small binding pocket which is occupied by the imidazole ring of the substrate (Figure 2.6 B). A similar observation can be made for the ancestral enzymes. In Anc1-Anc4 the space to the right is occupied by the helix residues, whereas the loop in Anc5 and more so in Anc6-Anc7 takes up less space and creates a cavity for the substrate HolP (Figure 2.6 C).

Finally, the electrostatics of the active sites differ significantly: While the deep end of the active site is negatively charged in all enzymes, the surface-exposed parts of the active site are positively charged in the case of Anc1-Anc4 and *ec*GmhB, but negatively charged in Anc5-Anc7 and *ec*HisB-N (Figure 2.6 C). The positive charge in *ec*GmhB and Anc1-Anc4 most likely enables electrostatic interactions with the phosphate and hydroxyl groups of the substrate βHBP. In contrast, the negative charge of the surface-exposed parts in Anc5-Anc7 ensures charge complementarity with the amino group and the imidazole ring of HolP, which is protonated to a significant degree at neutral pH. Interestingly, Anc5 shares most of these properties with *ec*HisB-N but is still a catalyst with moderate efficiency. It can therefore be concluded that the significant structural adaptations observed between Anc4 and Anc5 are necessary to allow for a reasonable turnover of HolP, but only the mutations between Anc5 and Anc6 lead to an efficiency boost.

In summary, the analyses showed that the HolPase activity was already present in the early ancestors Anc1-Anc4, which share many of the structural features of modern βGmhBs. Starting with the transition

from Anc4 to Anc5 the HolPase activity was enhanced. This enhancement was accompanied by significant adaptations of the shape and electrostatics of the active site.

2.2.6 A revised model for the evolution of HisB-N

Previously, it has been hypothesized that HisB-N and GmhB were both derived from the same ancestral HAD phosphatase which underwent a gene duplication event. One of these copies was integrated into the histidine operon whereupon its gene product evolved towards a modern HisB-N.[83] Indeed, our results now show that HisB-N sequences form a sub-cluster within a bigger βGmhB cluster (Figure 2.7). This sequence relationship is additionally supported by function: *ec*HisB-N and its most recent progenitors Anc6 and Anc7 are still able to hydrolyze βHBP. At the same time, the HolPase activity could already be detected in ancestral βGmhB-like enzymes Anc1-Anc4 that existed long before the functional divergence of HisB-N and βGmhB. This refines the previous model, as it suggests that HisB-N is derived from a βGmhB and not form an αGmhB.

Next, we were interested to follow the evolution of these two enzymatic functions on an organismal level and to this end reexamined the phylogenetic tree (Figure 2.7).

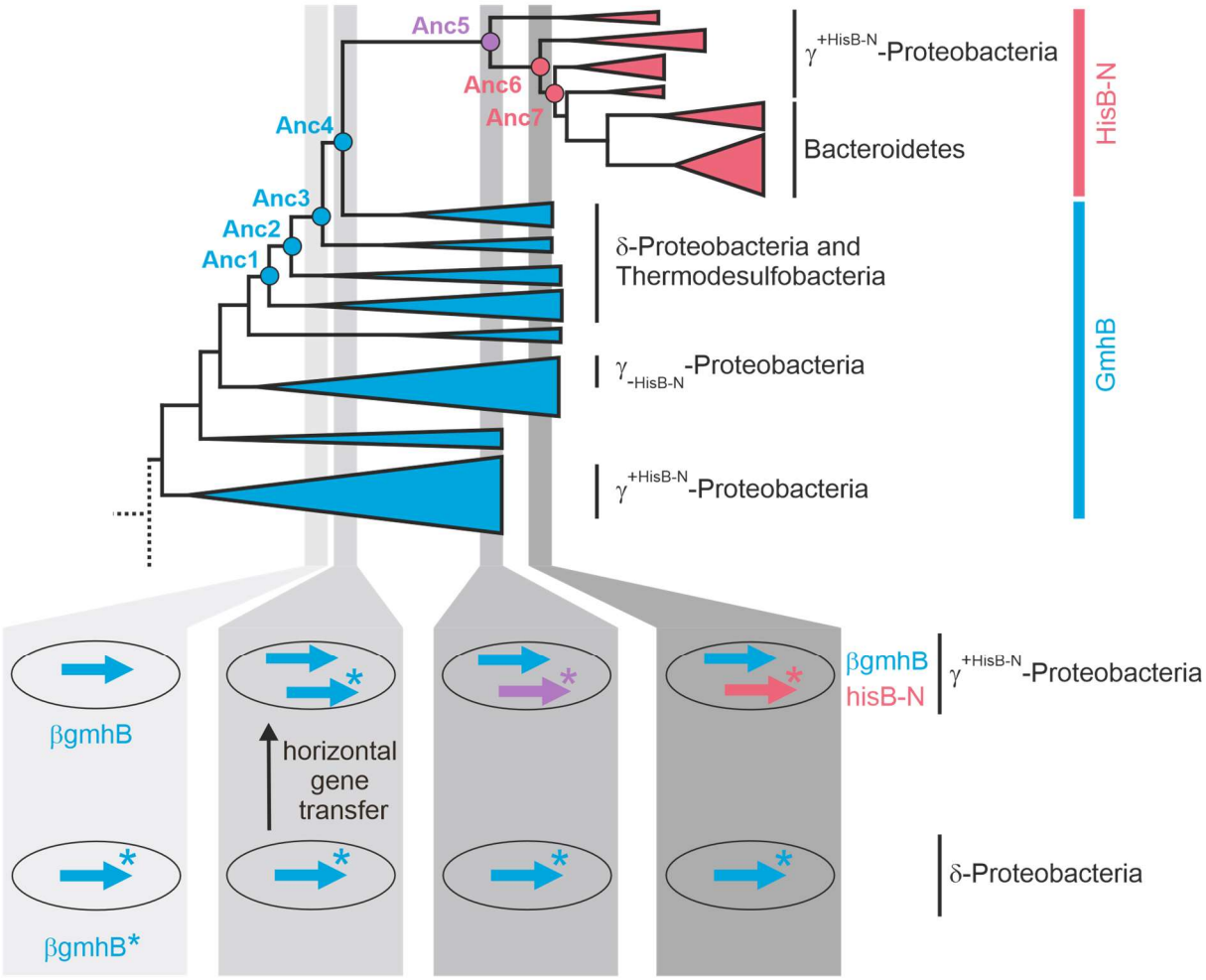

Figure 2.7: Phylogenetic distribution of *hisB-N* and *βgmhB* genes, and revised evolutionary model.

The upper panel represents parts of the condensed phylogenetic tree shown in Figure 2.4 A with clusters colored according to enzyme function (HisB-N in red, βGmhB in blue). Bacterial phyla are given on the

right of each cluster. The γ-Proteobacteria can be divided into two groups: $\gamma^{+\text{HisB-N}}$-Proteobacteria possess a β*gmhB* and a *hisB-N* gene; in contrast $\gamma_{\text{-HisB-N}}$-Proteobacteria only possess a *gmhB* gene. The panel below illustrates the proposed evolutionary scenario. At the stage of Anc3, $\gamma^{+\text{HisB-N}}$-Proteobacteria possessed a gene for a βGmhB enzyme (blue arrow), while δ-Proteobacteria had a gene for a promiscuous enzyme βGmhB* (blue arrow with asterisk). At the stage of Anc4, the horizontal gene transfer of *βgmhB** resulted in two functional *βgmhB*s in $\gamma^{+\text{HisB-N}}$-Proteobacteria. At the stage of Anc5, mutations in the *βgmhB** gene (violet arrow) led to a significant increase of the originally low HolPase activity. From stage Anc6 onward, further mutations in the *βgmhB** gene (red arrow) led to the modern HisB-N enzyme.

The most recent common precursor of HisB-N and βGmhB in this phylogenetic tree is represented by Anc4 (or in a broader sense by Anc1-Anc4). The βGmhB enzymes (blue triangles) that are directly derived from these ancestors are mostly found in Thermodesulfobacteria and δ-Proteobacteria. An ongoing debate proposes to combine δ-Proteobacteria and Thermodesulfobacteria in one phylogenetic group[110], which is in agreement with the similarity of their βGmhB sequences. However, neither δ-Proteobacteria nor Thermodesulfobacteria possess HisB-N enzymes. It is hence unlikely that a duplication event of the precursor of HisB-N and βGmhB happened in a δ-Proteobacterium or a Thermodesulfobacterium.

The HisB-N enzymes (red triangles) that are directly derived from Anc4 almost exclusively belong to γ-Proteobacteria (for a detailed discussion see Figure S 13). Interestingly, γ-Proteobacterial species form two separate βGmhB clusters and this clustering correlates with the occurrence of HisB-N in an organism. We named the γ-Proteobacteria that possess both a *gmhB* and a *hisB-N* gene $\gamma^{+\text{HisB-N}}$-Proteobacteria, and the species that only possess a *gmhB* gene $\gamma_{\text{-HisB-N}}$-Proteobacteria. A phylogenetic tree of 16S rRNA sequences revealed that the $\gamma^{+\text{HisB-N}}$- Proteobacteria form a distinct cluster and are thus closely related (Figure S 14), confirming the separation of γ-Proteobacteria into two groups. The finding that some γ-Proteobacteria lack a HisB-N like HolPase poses the question how histidine is synthesized by these bacteria. In principle these bacteria could either i) rely on histidine uptake from external sources, ii) possess a HolPase from the PHP oder IMP superfamily or iii) possess a different type of HolPase that is yet to be discovered. In summary, this strongly suggests that the first co-occurrence of HisB-N and βGmhB took place in $\gamma^{+\text{HisB-N}}$-Proteobacteria. However, the cluster of βGmhB sequences from $\gamma^{+\text{HisB-N}}$-Proteobacteria is located far apart from the HisB-N cluster. This separation of βGmhB and HisB-N from $\gamma^{+\text{HisB-N}}$-Proteobacteria is supported by a number of interjacent nodes with intermediate to high posterior probabilities. This large distance in the tree renders the previously postulated duplication event in an ancient $\gamma^{+\text{HisB-N}}$-Proteobacterium[83] unlikely.

Therefore, we propose an alternative evolutionary scenario, which is outlined in the lower part of Figure 2.7. This scenario assumes that at the timepoint presented by Anc3, βGmhBs were already present in the first $\gamma^{+\text{HisB-N}}$- and δ-Proteobacteria. The βGmhBs of these ancestral δ-Proteobacteria (βGmhB*) were promiscuous for HolP either because βGmhB* supported two metabolic pathways or due to the lack of substrate specificity. Whether this promiscuity represents a favorable degree of substrate promiscuity or if it is an evolutionary irrelevant degree of sloppiness, cannot be decided with certainty. The causes and consequences of enzyme promiscuity in general are a matter of ongoing debate.[33, 34, 111, 112] The gene of such a promiscuous βGmhB* was acquired by an ancestral $\gamma^{+\text{HisB-N}}$-Proteobacterium via horizontal gene transfer probably from an ancient Thermodesulfobacterium or δ-Proteobacterium. According to the phylogenetic tree, an ancient $\gamma_{\text{-HisB-N}}$-, or β-Proteobacterium would also be a possible donor species. Since all modern HisB-N enzymes are N-terminally fused to IGPDH and the respective bi-functional genes are located within the histidine operon, one can assume that after the horizontal gene transfer

*βgmhB** was incorporated into the histidine operon upstream of the gene that encodes for the IGPDH, as argued previously.[113, 83] The receiving $\gamma^{+\text{HisB-N}}$-Proteobacterium already possessed a βGmhB enzyme, so the selective pressure on the GmhB function of the newly transferred *βgmhB** gene was likely reduced. However, the promiscuous side activity for HolP probably conferred some evolutionary advantage which led to an increase of the primordial HolPase activity that is visible in Anc5-Anc7.

This evolutionary scenario is also supported by the phylogenetic distribution of the properties of the zinc-binding cap structure (Figure S 13). On the N-terminal end, this structure is defined by a CxH motif and on the C-terminal end by a CxC motif. In the βGmhB variants from $\gamma^{+\text{HisB-N}}$-Proteobacteria, Negativicutes, Aquificales and most α-Proteobacteria, the CxH and CxC motifs are separated by 12 or 13 residues. This long cap-structure is predominately found in phylogenetically diverse βGmhB enzymes that are close to the branchpoint with αGmhBs. The length of this cap- structure and the chemical properties of many residues are conserved, including the strictly conserved tyrosine at residue position 109 (Figure S 13). Taken together, conservation of this cap-structure and its broad phylogenetic distribution suggests that these βGmhBs correspond to the ancient state. Conversely, in the βGmhB variants from β-, $\gamma_{\text{-HisB-N}}$-, δ-Proteobacteria, Thermodesulfobacteria, and some α-Proteobacteria, the CxH and CxC motifs are separated by merely 5-6 residues. The same pattern also occurs in HisB-N variants from $\gamma^{+\text{HisB-N}}$-Proteobacteria, which again underscores the relationship between these HisB-N sequences and βGmhBs from β-, $\gamma_{\text{-HisB-N}}$-, δ-Proteobacteria, and Thermodesulfobacteria. Interestingly, the CxH and CxC motifs are missing in HisB-N from Bacteroidia and few $\gamma^{+\text{HisB-N}}$-Proteobacteria, suggesting that they were lost in the course of evolution.

According to the phylogenetic tree, the experimental data, and the discussed structural features, the proposed horizontal gene transfer seems to be the most convincing evolutionary scenario. However, one has to take into account the generally high sequence divergence of the HAD superfamily.[51] Accordingly, HisB-N and GmhB sequences, but also different GmhB sequences share not more than 30-40% identical residues. Therefore, a gene duplication event cannot be ruled out with certainty.

In conclusion, our findings nicely fit the innovation amplification and divergence (IAD) model of evolution.[31] This model assumes that a (fortuitous) promiscuous side activity of an enzyme can become relevant when posed under selective pressure, as in this case the promiscuous HolPase activity of Anc1-Anc4. It has also been shown that genes that were acquired by horizontal gene transfer from distantly related donor species make up for a significant fraction of genes in bacteria.[114] Moreover, it has been argued that horizontal gene transfer plays an underestimated role in the evolution of new enzymes, as a side activity of a newly transferred gene that contributes to the recipients fitness would be a perfect starting point for evolution.[35] The herein observed boost in HolPase activity after the horizontal transfer of Anc4 from δ-Proteobacteria to $\gamma^{+\text{HisB-N}}$-Proteobacteria is in line with this rationale.

3 Characterization of a putative HolPase from *P. aeruginosa*

3.1 Introduction

Histidine biosynthesis consists of the same sequence of reactions in all three kingdoms of life and the individual enzymes are generally conserved between different species.[70, 72, 79, 78] The only exception is the HolPase, which catalyzes the eighth reaction step of the pathway and which shows different genetic localization and structures in different species.[70, 83] In the model organism *E. coli*, the HolPase shows a Rossmann fold and is encoded by the N-terminal part of the bi-functional *hisB* gene which is located in the histidine operon.[59, 72] This bi-functional *hisB* gene is however only encountered in a narrow phylogenetic subdivision of the Proteobacteria and very few other species which probably obtained the gene via horizontal gene transfer.[115] In other phylogenetic clades non-homologous, monofunctional HolPases were identified which were encoded by genes outside the histidine operon. It was therefore argued, that the bi-functional gene is the result of a gene fusion event which occurred in γ-Proteobacteria after the divergence of *Pseudomonas*.[83] This observation immediately leads to the question of which gene encodes the HolPase function in *Pseudomonas* and related species. The question remained unanswered until recently, when Wang et al. reported that the knock-out of the *Pseudomonas aeruginosa* gene PA0335 led to a growth defect which was caused by partial histidine auxotrophy.[116] A bioinformatical analysis revealed that the associated enzyme contained sequence motifs that are typical for the HAD superfamily and the authors therefore concluded that PA0335 coded for the missing HolPase. In line with this assumption, the crude extract from BL21 with overexpressed PA0335 furthermore showed increased phosphatase activity. [116]

Since most of the research on HolPases was conducted in *E. coli* where the *hisB* gene is bifunctional, there is some ambiguity regarding the nomenclature of the gene that encodes for the HolPase. In this chapter, the nomenclature proposed by Brilli et al. will be applied, who suggested *hisN* as name for HolPase encoding genes.[83] The gene product of the *P. aeruginosa* gene PA0335 gene will consequentially be referred to as *pa*HisN.

Remarkably, *pa*HisN was not caught by functional annotation based on homology prior to the experimental work from Wang et al. although both *pa*HisN and the *E. coli* enzyme *ec*HisB-N were classified as members of the HAD superfamily. This is because the sequence similarity between *pa*HisN and *ec*HisB-N is rather limited (29.1 %) and hence does not exceed the level which is normally observed for proteins from the HAD superfamily with differing function.[51] This level of sequence similarity renders a common evolutionary origin in the recent past highly unlikely and instead implies a more distant evolutionary relationship. The results of the previous chapter moreover showed that *ec*HisB-N was the product of a relatively recent evolutionary event and was derived from an ancestral GmhB enzyme. Taken together, these findings do not support a scenario by which *ec*HisB-N was derived from *pa*HisN or *vice versa* but instead indicate that these two enzymes are the result of convergent evolution. This would mean that *pa*HisN represents an additional type of HolPase which also belongs to the HAD superfamily.

The specific functional properties of *pa*HisN are however yet to be determined. While it is probably safe to assume that it confers some level of HolPase activity, it is still unclear whether this is the primary function of a monofunctional enzyme or if *pa*HisN represents a multifunctional enzyme with a broad substrate spectrum of which the HolPase function is just one promiscuous activity. If *pa*HisN indeed

represented an additional type of HolPase, then the question of the phylogenetic distribution and evolutionary origin of this new class of HolPases would directly arise.

These questions were addressed by a combination of bioinformatical, biophysical, and biochemical methods. First, the AlphaFold2[100] predicted structure of *pa*HisN was analyzed and compared to the structure of other known HAD enzymes. A BLAST[117] search was furthermore used to identify homologues of *pa*HisN and to gain insights regarding the potential evolutionary origin. Building on this structural and sequential analysis, a thorough *in vitro* characterization of the enzyme was performed which included the measurement of the oligomerization state, folding, and thermostability of the protein, and a detailed analysis of the catalytic function and possible promiscuous side activities. Then, an alanine scan was performed with the aim to establish a fingerprint of the residues which are critical for HolPase activity. This fingerprint was used as criterion to identify other HolPases in a sequence similarity network of homologues of *pa*HisN and determine their phylogenetic distribution. Lastly, this result was cross validated by an analysis of the occurrence of the different HolPases and the missing annotation of a HolPase in histidine-synthesizing organisms.

3.2 Results and Discussion

3.2.1 Structural analysis and search for homologues of *pa*HisN

So far, the classification of *pa*HisN as a HAD enzyme was based on the occurrence of sequence motifs characteristic of the HAD superfamily and a sequence identity between 24 and 47% to other annotated HAD enzymes.[116] To validate this conclusion, the AlphaFold2[100] prediction of the three-dimensional structure was retrieved from the Uniprot database and analyzed (Figure 3.1 A).

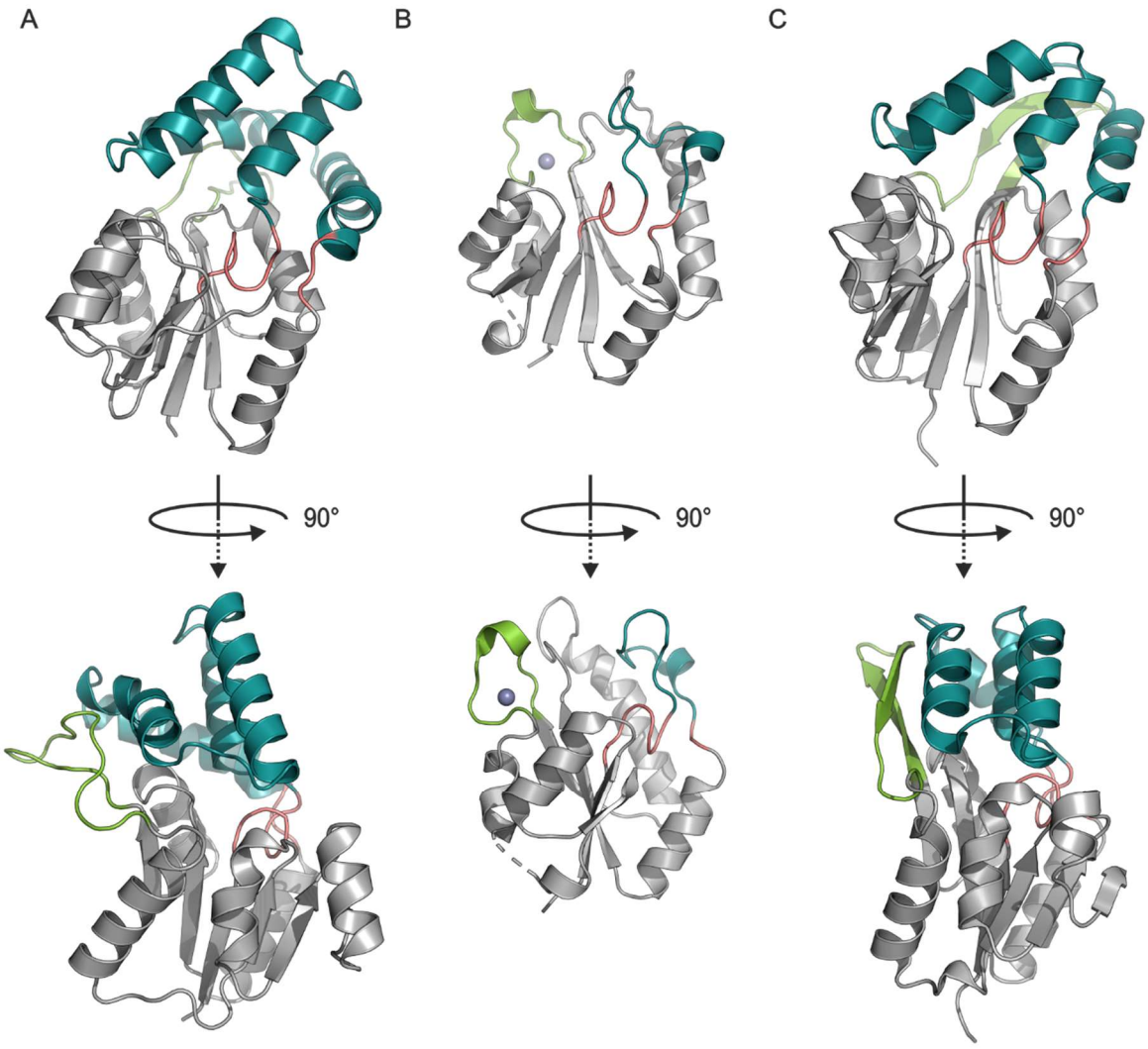

Figure 3.1: Comparison of the predicted structure of *pa*HisN with other HAD enzymes.

(A) The structure of *pa*HisN as predicted by AlphaFold2 [100] is typical for a HAD enzyme [51] and consists of a core Rossmann fold (grey) being decorated by two insertions, namely a four helical bundle (cyan) which is inserted in the flap (red) and an extended loop (green) which is inserted after the third β-strand. (B) The crystal structure of the HAD enzyme *ec*HisB-N[59] (PDB-ID: 2FPU) shows the same core Rossmann fold (grey) as *pa*HisN. Insertions into the core are located at equivalent positions but constitute different structural elements, namely a short loop (cyan) and a short turn (green) with a bound Zn^{2+} ion (grey sphere). (C) The PSPase *mj*SerB (PDB-ID: 1F5S) [58] also shows a typical Rossmann fold (grey) with two insertions (cyan, green) which are located at the same positions and show similar geometries and sizes as in *pa*HisN.

The predicted structure of *pa*HisN indeed shows the typical folding pattern of a HAD superfamily protein with a Rossmann fold core, characterized by five parallel β-sheets with the strand order 5-4-1-2-3

which are separated from each other by α-helices that surround the central β-sheet.[51] The structure also includes two insertions into the Rossmann fold core. The first insertion is located in the flap and folds as a four helical bundle (Figure 3.1 A, cyan). The second insertion is located after the third β-strand and is predicted to form an extended loop (Figure 3.1 A, green). A comparison to the crystal structure of *ec*HisB-N[59] (Figure 3.1 B) revealed that the similarities between the two structures are confined to the Rossman core. The insertions of the two enzymes differ however significantly in their secondary structure. In *ec*HisB-N the insertion that corresponds to the four helical bundle of *pa*HisN forms a short loop (Figure 3.1 B, cyan) and the insertion which corresponds to the extended loop of *pa*HisN forms a short turn that carries a conserved zinc binding motif (Figure 3.1 A, B green, Zn^{2+} is depicted as grey sphere). The two insertions in *pa*HisN are however reminiscent of the ones which are typically found in phosphoserine phosphatases (PSPases).[51] A comparison to the crystal structure of the PSPase from *Methanococcus jannaschii mj*SerB[58] indeed showed that the secondary and tertiary structure of both insertions were similar to the predicted ones in *pa*HisN (Figure 3.1 C). The insertions in both *pa*HisN and *mj*SerB furthermore consist of a similar number of residues, namely 72 for the four helical bundle and 18 for the extended loop in *pa*HisN compared to 57 and 16 in *mj*SerB. A global sequence alignment of *pa*HisN and *mj*SerB revealed 21.7 % identical and 42.5 % similar residues. This moderate sequence conservation between the two sequences suggested a distant evolutionary relationship despite the obvious structural similarities.

To check whether there are more closely related PSPase enzymes, a BLAST[117] search of the NCBI database of non-redundant protein sequences[118] was performed with *pa*HisN as query. This search retrieved more than 250 homologous sequences with more than 80 % sequence identity and a query coverage above 70 %. Most of these hits were annotated as "HAD family hydrolase" and were found in organisms closely related to *Pseudomonas aeruginosa*.

To gain additional information about the phylogenetic distribution of the *pa*HisN homologues, the search was repeated excluding any sequences from the order Pseudomonadales. This still gave more than 100 hits with sequence identities above 50 % (Table S 4). While most of these sequences were again annotated as HAD family hydrolase, 4 sequences were instead classified as phosphoserine phosphatase. Interestingly, the sequence identities between those putative PSPases and *pa*HisN ranged from 58 % to 99.5 %. There are two possible explanations for this finding. Either *pa*HisN is in fact a PSPase with a side activity for HolP or several proteins were misannotated as PSPases while actually being HolPases. To clarify this issue, in a first step, *pa*HisN should be characterized *in vitro*.

3.2.2 *In vitro* characterization of *pa*HisN

Wang et al. demonstrated that *pa*HisN could complement a gene knock-out of the HolPase encoding gene *in vivo*.[116] They furthermore confirmed that overexpression of *pa*HisN increased the phosphatase activity of the crude extract against a generic phosphorylated substrate. Information about the biophysical properties like oligomerization state, melting temperature, or enzyme kinetic parameters of the HolP turnover had not been published. Therefore, a thorough biophysical analysis of *pa*HisN should be performed and kinetic experiments with both HolP and PSer should furthermore resolve the question if *pa*HisN catalyzed the turnover of both substrates and if so, which substrate was the preferred one.

To this end, the *hisN* gene from *P. aeruginosa* was codon optimized for the expression in *E. coli*, equipped with *Bsa*I digestion sites at the N- and C-terminus, and cloned into a pUR23 expression

plasmid with an N-terminal His_6-tag by golden gate cloning (8.3.4.5). The gene was then expressed in a *ΔhisBΔserB* strain (8.4.2) to exclude any possible contamination with host cell *ec*HisB-N or host cell *ec*SerB. Afterwards, the protein was purified by affinity chromatography followed by size exclusion chromatography (8.4.3, 8.4.4, 8.4.5).

Mass and purity of the protein preparation were assessed by SDS-PAGE (8.5.2, Figure 3.2 A). The gel showed a single band which corresponded to a molecular weight of approximately 25 kDa which is in line with the theoretical monomer weight of 25.6 kDa. Next, the oligomerization state was investigated by a combination of size exclusion chromatography and static light scattering (8.5.4, Figure 3.2 B). The experimentally obtained value for the number average molar mass Mn was 25.8 kDa and the value for the mass average molar mass Mw was 25.9 kDa which are both in the range of the theoretical weight of the monomer. The ratio of Mw/Mn was 1.004, which indicates a monodisperse solution of only one molecular species.

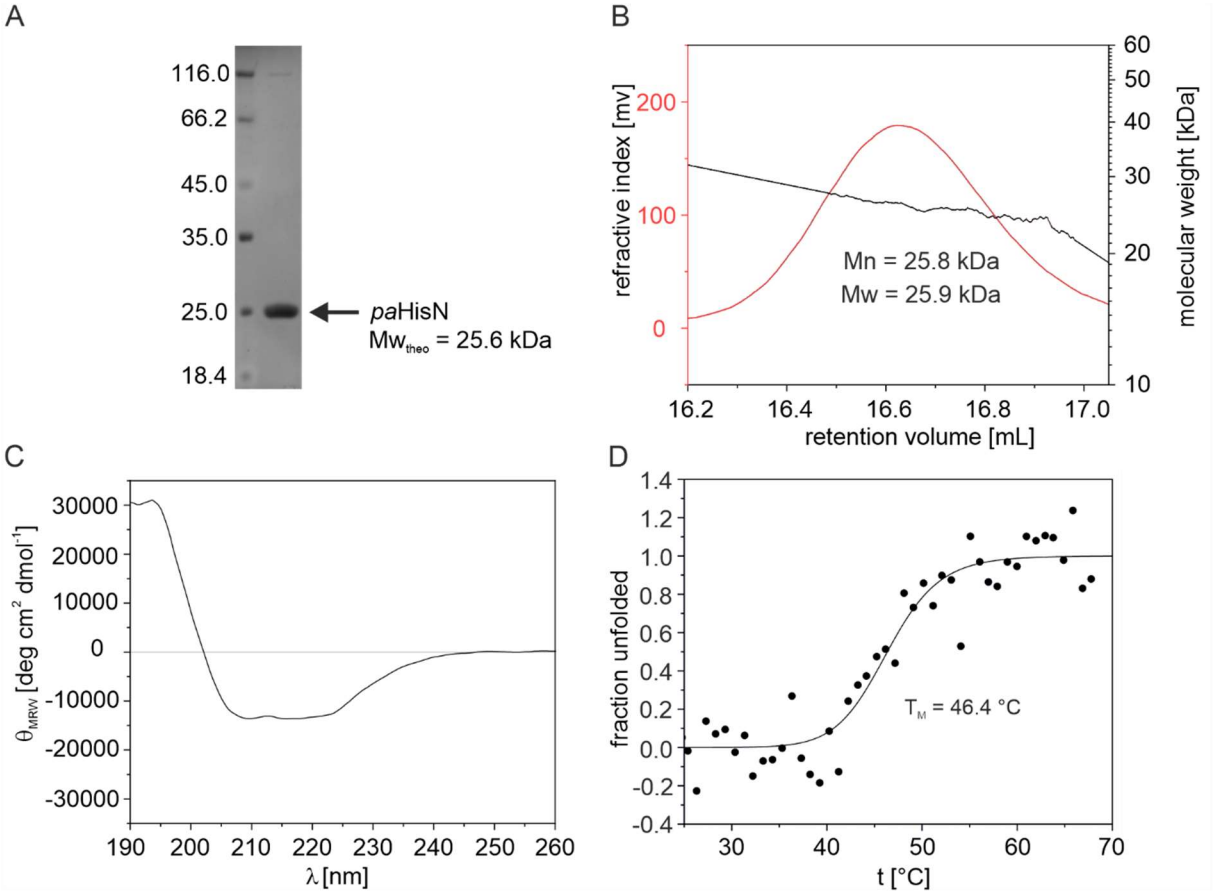

Figure 3.2: Biophysical characterization of *pa*HisN.

(A) SDS-PAGE (13.5 % acrylamide) of purified *pa*HisN (3 µg). The gel showed a single band corresponding to a molecular weight of 25 kDa which is in line with the theoretical monomer weight of 25.6 kDa. (B) Size exclusion chromatography followed by static light scattering yielded a number average molar mass (Mn) of 25.8 kDa and a mass average molar mass (Mw) of 25.9 kDa indicating that *pa*HisN formed a monomer in solution. (C) The CD spectrum of *pa*HisN (in 20 mM KP, pH 7.5) was characteristic for a well folded protein. (D) The melting curve of *pa*HisN as monitored by far-UV CD could be fit a two-state model without folding intermediates yielding a melting temperature (T_M) of 46.4 °C.

In the next step, the secondary structure of *pa*HisN was probed by CD spectroscopy (8.5.3). To prevent extensive background absorption the Tris buffer was exchanged to 20 mM KP prior to the measurements

using a NAP 5 column (8.4.6). The recorded CD spectrum (Figure 3.2 C) exhibits clear minima at around 207 nm and 220 nm, a zero crossing above 200 nm, and a maximum at around 195 nm which is characteristic for a folded protein. Next, the melting point was determined by monitoring the CD signal at 220 nm while the sample was heated at a constant rate (8.5.3). A plot of the recorded data showed a sigmoidal curve (Figure 3.1 D) which is indicative of a transition from a folded state at low temperatures to an unfolded state at high temperatures without additional folding intermediates. The data was therefore fitted according to a two-state model which gave a melting temperature T_M of 46.4 °C. This result is in line with the growth temperature of *P. aeruginosa* which ranges from 4-42 °C.[119]

After establishing that *pa*HisN was a well folded protein of reasonable purity, the enzymatic function should be examined next. To this end, the enzymatic turnover of the potential substrates HolP and PSer was investigated using a coupled enzymatic assay (8.5.5).

First, HolP was tested as substrate. Upon addition of *pa*HisN, the release of free phosphate could be detected which confirmed the HolPase function of *pa*HisN. Building on this, steady-state kinetic parameters were determined by measurement of the reaction rate for varying substrate concentrations. A saturation curve (Figure 3.3 A) was obtained and a hyperbolic fit of the data according to Michaelis-Menten yielded a k_{cat} value of 8.4 s^{-1} and a K_M value of 21 µM. The k_{cat} value was slightly higher than previously reported values of different HolPases which were between 1 s^{-1} and 4 s^{-1}.[93, 94, 105, 115] The K_M value was at the lower end of previously reported values from other HolPases which ranged from 32 µM to 400 µM.[93, 105, 120, 59] Taken together, this corresponds to a k_{cat}/K_M value of 400,000 $M^{-1}s^{-1}$ and confirms that *pa*HisN is indeed a HolPase.

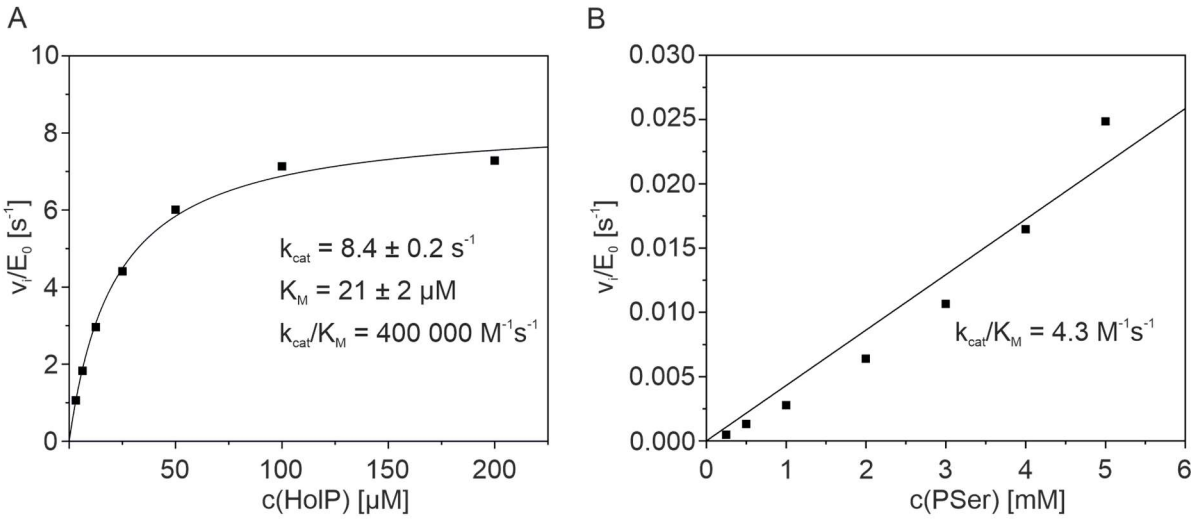

Figure 3.3: Functional characterization of *pa*HisN at 25°C.

(A) Substrate saturation curve for the turnover of HolP by *pa*HisN. (B) Linear fit for the turnover of PSer by *pa*HisN.

Second, PSer was tested as substrate. Intriguingly, enzymatic turnover could again be observed upon addition of *pa*HisN to the reaction mixture. The reaction rate, however, was low compared to the HolPase activity and substrate saturation could not be achieved (Figure 3.3, B). The data was hence fitted with a linear function which yielded a k_{cat}/K_M value of 4.3 $M^{-1}s^{-1}$which makes the turnover of HolP more than 90,000- fold more efficient than the turnover of PSer. The PSPase activity is thus a promiscuous side activity of *pa*HisN. This side activity could either be the remnant of evolutionary

relationship with PSPases or due to limited substrate specificity. The latter might be caused by strong binding of the phosphate group is a common feature of both substrates in combination with the comparatively small size of PSer which could make it hard to completely exclude PSer from the active site. The extraordinary catalytic efficiency of *pa*HisN which is evidenced by a high k_{cat} and a low K_M value for the HolPase function might then lead to erroneous turnover of PSer even though it is not the native substrate.

Given that *pa*HisN shows such a strong preference for HolP over PSer, it seems highly unlikely that sequences that share more than 90 % sequence identity with *pa*HisN function as PSPases. Instead, these entries from the NCBI database are most likely misannotated and are in fact HolPases. Based on this conclusion the question remains whether the *pa*HisN homologues which were classified as PSPases but exhibit an intermediate degree of sequence identity with *pa*HisN are PSPases or HolPases. To help with the classification of these enzymes, the features which define *pa*HisN as HolPase should be analyzed. Specifically, an alanine scan of active site residues should be performed to identify the residues that are most important for the specific binding and turnover of HolP in *pa*HisN. This set of residues should yield a fingerprint by which other HolPases could be identified.

3.2.3 Analysis of the functionally relevant residues in *pa*HisN by alanine scanning

To identify the residues within *pa*HisN which are critical for specific substrate recognition and turnover, an alanine scan of selected active site residues was performed. The residues to be exchanged in the alanine scan were selected based on the following considerations: The proteins of the HAD superfamily are characterized by a Rossmann fold which contains four highly conserved sequence motifs, namely a DxD motif, a conserved T or S, a conserved R or K and a conserved DD, GDxxxD, or GDxxxxD motif.[51] This set of residues endows the HAD proteins with the catalytic machinery that confers the phosphatase function and can also be found in *pa*HisN (Figure 3.4 A, blue sticks).

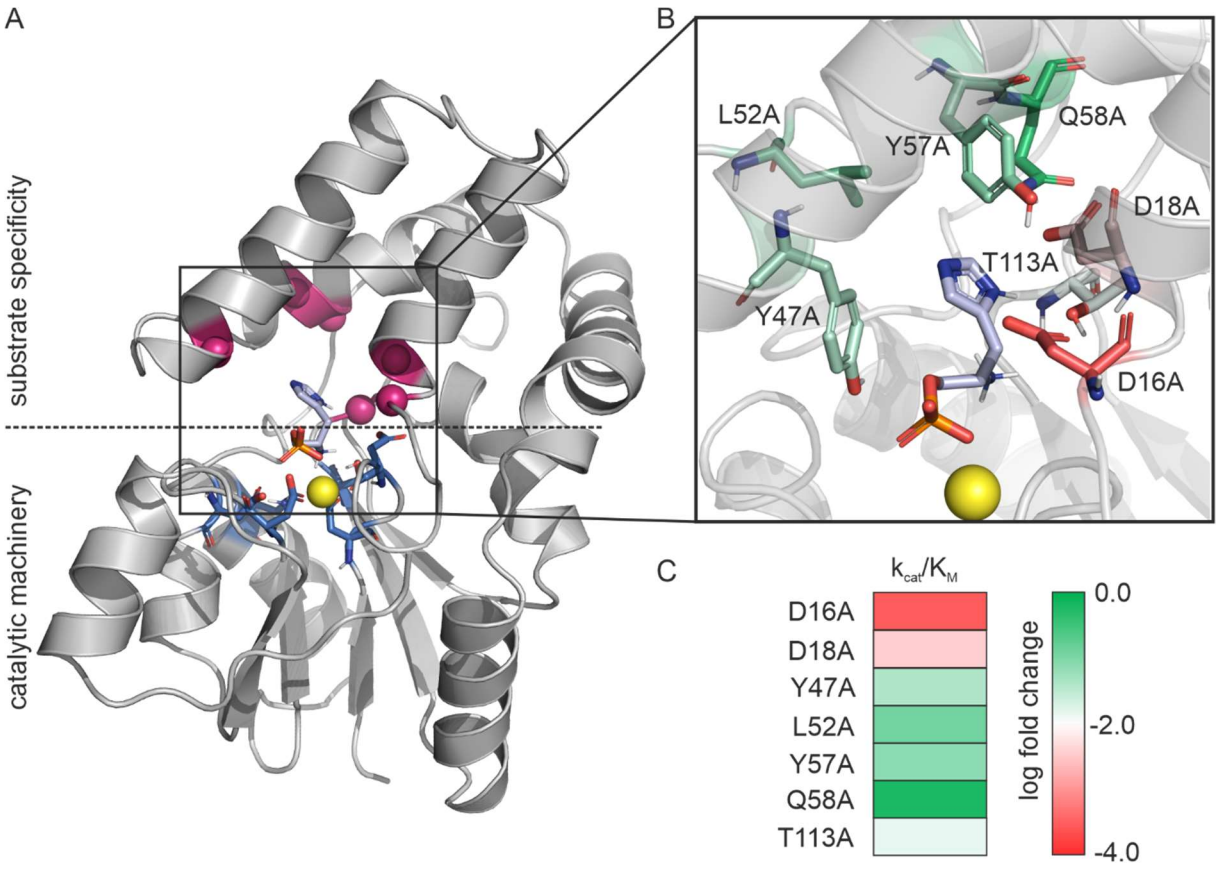

Figure 3.4: Alanine Scan of active site residues of *pa*HisN.

(A) AlphaFold[100] predicted structure of *pa*HisN with docked HolP (light blue). In HAD enzymes, the catalytic machinery is provided by conserved residues of the Rossmann core (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.

The substrate specificity is however mediated via different residues that are not needed for the chemical steps in catalysis and can therefore be adapted to meet the steric and electrostatic requirements of different substrates. In the HAD superfamily, this function is often fulfilled by the residues that form part of the cap structures.[51] In the case of *pa*HisN, the active site is covered by the insertion which is predicted to form a four helical bundle and which probably contains most of the residues that mediate substrate specificity. To identify residues of the active site and specifically of the cap which are in close proximity to the substrate, first, the position of the bound substrate was approximated, by a docking experiment which was performed by Dr. Julian Nazet (Figure 3.4 A, B, the docked substrate is colored in light blue and shown in stick representation). The ensuing visual inspection of the structure revealed seven residues which are (i) less than 4 Å away from the docked substrate and (ii) the side chains of which point towards the substrate (Figure 3.4 A, purple dots and Figure 3.4 B sticks). This set of seven residues consisted of D16, D18, Y47, L52, Y57, Q58, and T113. Except for T113, all these residues were part of the four helical cap.

Each of the seven positions was individually mutated to alanine (8.3.4.4), the resulting genes were overexpressed (8.4.2) and corresponding proteins were purified in the same manner as the wildtype

protein (8.4.3, 8.4.4, 8.4.5). As judged by SDS-PAGE (8.5.2) all proteins could be obtained with good purity (Figure S 15) and CD spectroscopy (8.5.3) confirmed that all proteins adopted a well folded conformation in solution (Figure S 16). In the next step, the influence of each mutation on the enzymatic function was probed by steady-state kinetic experiments with HolP (Figure S 17).

To illustrate the spatial arrangement, the effect of each mutation was highlighted in a zoomed-in view on the active site (Figure 3.4 B). Specifically, the color of each residue indicates the degree of activity loss upon mutation to alanine from green for a mild decrease to red for a drastic decrease in activity. The color scale was normalized to the logarithm of the fold change of the k_{cat}/K_M value as a measure for the activity decrease (Figure 3.4 C) The associated catalytic parameters are listed in Table 3.1.

Table 3.1: Catalytic parameters of *pa*HisN single mutants to alanine at 25°C.

Variant	k_{cat} [s^{-1}]	K_M [μM]	k_{cat}/K_M [$s^{-1}M^{-1}$]
wt	8.4 ± 0.2	21 ± 2	400,000
D16A	(3.0 ± 0.1) × 10^{-3}	26 ± 2	113
D18A	0.50 ± 0.06	365 ± 86	1,368
Y47A	2.6 ± 0.2	163 ± 27	15,644
L52A	6.2 ± 0.7	129 ± 38	48,217
Y57A	6.9 ± 0.6	232 ± 40	29,784
Q58A	5.8 ± 0.2	24 ± 3	238,272
T113A	3.6 ± 0.3	660 ± 84	5,454

The first mutation that was introduced concerned D16. The mutation to alanine resulted in a dramatic decrease of the k_{cat} value by four orders of magnitude whereas the K_M value remained almost unaffected. This means, while the substrate was still bound with an affinity similar as in the wildtype enzyme, the turnover rate was severely impaired which suggests that there is a significant degree of nonproductive binding. In the structure, this aspartate residue is very close to the predicted position of the imidazole ring, and one could easily imagine, that a hydrogen bond is formed between the side chain of D16 and the imidazole ring which could bind and orient the substrate. Alternatively, D16 might also form a hydrogen bond with the amino group of the substrate.

By contrast, the D18A mutation led to a decrease in the k_{cat} value and an increase in the K_M value both by a factor of approximately 17. This means that both substrate binding and the formation of a productive enzyme-substrate complex are impaired to a similar degree. Like D16, D18 is very close to the predicted location of the substrate and again, a hydrogen bond between the side chain of D18 and the imidazole ring seems plausible.

The mutation Y47A signified a drop in the k_{cat} value by a factor of 3 and an increase in the K_M value by a factor of 8, implying that this mutation mostly affects substrate binding. In principle, several modes of interaction between this residue and the substrate are possible; there could be a hydrogen bond between the terminal OH-group of tyrosine and a free electron pair of a nitrogen atom of the imidazole ring, or in the case of a protonated imidazole ring, one could think of a cation-π interaction, or simply a π-π interaction between the two aromatic rings. Even though the Y47A mutation to alanine of this

residue does not have as severe consequences as the mutation of D16 or D18, it still leads to a drop in catalytic efficiency by more than 25-fold, meaning that this residue is still relevant for HolPase function.

The mutation of either L52 or Y57 to alanine had a very similar effect in both cases which was characterized by a slight decrease in the k_{cat} value but a significant increase in the K_M value by a factor of 6 and 11, respectively. Both residues are positioned in the upper left part of the active site. The moderate effect of the exchanges to alanine could be interpreted in a way that these residues do not confer the primary interactions with the substrate. They are however still relevant, which means that a possible function of these residues could be the correct packing of the active site or the filling of cavities to ensure an optimal fit of the substrate into the active site.

The Q58A mutation has a very small effect on the k_{cat} value and no significant impact on the K_M value, which means that this residue is not essential for the enzymatic function of *pa*HisN and most likely not interacting with the substrate.

The T113A mutation had a mild effect on the k_{cat} value which was decreased by a factor of 2 but a strong effect on the K_M value which was increased more than 30-fold. This implied an important role of this residue in substrate binding. Several modes of action could be envisioned including a hydrogen bond to the imidazole ring or the amino group of the substrate. Interestingly, T113 is located in close proximity to the critical residues D16, D18 and all three residues are very close to the predicted position of the imidazole moiety of the substrate. This finding supports the docking analysis, as it seems plausible that the mutation of those residues which physically interact with the substrate is most detrimental.

The information regarding the relevance of each residue for the HolPase function should next be used for a comparison with the corresponding residues in PSPases. The goal of this comparison was to analyze if there are residues which are conserved across the two functions or if there exists a distinct pattern of amino acids in PSPases. Based on this comparison, fingerprint should be established that could classify an enzyme as HolPase or PSPase. To estimate the importance of each residue in PSPases, the conservation of each residue across different PSPases should be accounted for. For this objective, a reliable sequence logo of PSPases should be created despite the obvious annotation problem. To this end, the well characterized PSPase *mj*SerB [121, 58, 122] was used as starting point. With *mj*SerB as query, a BLAST[117] search of the NCBI database of non-redundant protein sequences [118] was performed and 211 homologous sequences with 48-100 % sequence identity and a query coverage of 93% were retrieved which were all annotated as PSPases. A threshold of about 50 % sequence identity to *mj*SerB was chosen to balance the need for diversity in the dataset with the potential problem of misannotated sequences. The resulting dataset was then used to create a sequence logo with WebLogo 3[123] which was compared to the sequence of *pa*HisN (Figure 3.5, positions from the alanine scan are marked by yellow boxes).

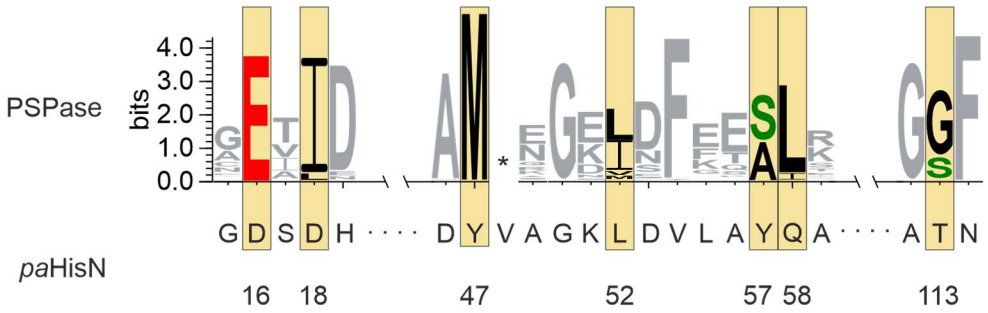

Figure 3.5: Comparison of the *pa*HisN sequence with a sequence logo of PSPases.

The residues of *pa*HisN which were analyzed by mutation to alanine were compared to the amino acids at the equivalent position in a reference data set of 211 homologues of the PSPase *mj*SerB. The upper part shows a sequence logo which is based on an MSA of the PSPases while the lower part gives the sequence of *pa*HisN, and numbers below 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 which 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. Color code: red: acidic, black: hydrophobic, green: hydroxyl group, purple: amide, blue: basic.

According to the alanine scan, D16 was by far the most important amino acid for the HolPase function in *pa*HisN. At the corresponding position of the PSPase sequence logo, there was a highly conserved glutamate. Even though glutamate and aspartate share the same chemical properties and merely differ in the length of their side chains, the strong conservation of glutamate, the complete absence of any aspartate in PSPases at this position and the crucial role of the aspartate for the HolPase activity suggest that the occurrence of an E or D at this position could nevertheless be a first criterion by which a PSPase or a HolPase could be identified.

The second most important residue in *pa*HisN was D18. In PSPases this position is occupied by a conserved isoleucine which is followed by an aspartate. In principle, this aspartate which is shifted by one position in PSPases relative to *pa*HisN could fulfill the same function as D18. However, these residues form part of a helix and close inspection of the PSPase reference structure from *mj*SerB and the predicted structure of *pa*HisN showed that these residues occupy consecutive positions within that helix. This means that the side chain of D18 in *pa*HisN is pointing towards the active site whereas the side chain of the aspartate in PSPases is pointing away from it. The relative shift by one position of the conserved aspartate hence seems to make a difference. Taken together with the occurrence of either a D or E at position 16, a DxD motif in *pa*HisN as opposed to an ExI motif in PSPases can be deduced which is assumed to be indicated of either a HolPase or a PSPase function.

The third most important residue in *pa*HisN was T113 which corresponds to a glycine or serine residue in the PSPase sequence logo. However, neither glycine nor serine are strongly conserved in PSPases. Moreover, replacing threonine by serine is a rather conservative exchange which might be compatible with the HolPase function. The nature of the amino acid at this position is hence not a reliable criterion for the distinction of HolPases and PSPases.

The fourth most important residue in *pa*HisN was Y47. The corresponding position in the PSPase sequence logo is occupied by a highly conserved methionine. The strong conservation within the reference data set of PSPases indicates functional relevance of this residue. This implies that the

occurrence of either a tyrosine or methionine at this position could serve as a third criterion to distinguish the two different enzyme functions.

The fifth most important residue in *pa*HisN was Y57. The most frequently observed residues at the equivalent position in PSPases are serine and alanine which however both show limited conservation. The weak conservation in PSPases together with the moderate effect of the Y57A mutation on the *pa*HisN activity suggests that the occurrence of a tyrosine at position 57 alone is not a reliable indicator of a HolPase function.

The residues L52 and Q58 had proven to be the least important residues for the HolPase activity among the seven tested positions. In PSPases, a weakly conserved leucin is the most frequently encountered amino acid at the position that corresponds to L52. Hence, the nature of the amino acid at position 52 is not suited for classification. The amino acid which corresponds to Q58 showed some conservation with leucin again being the most frequently observed amino acid. The minimal effect of the Q58A mutation in *pa*HisN however suggests that it is not required for the turnover of HolP and that a leucin might also be compatible with HolPase function. Thus, the amino acids at this position also does not seem to be an appropriate feature to distinguish the two enzyme functions.

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 either Y57 or T113 may support the classification of a protein as HolPase but neither Y57 nor T113 alone seem to be conclusive for a HolPase function.

3.2.4 *In silico* analysis of the homologues of *pa*HisN

The comparative analysis of the residues that define HolPase function with those amino acids that are found at corresponding positions in PSPases helped to identify three amino acids, namely two which form a DxD motif and a tyrosine, which are thought to be characteristic for *pa*HisN-type HolPases. This fingerprint should now be utilized to identify other HolPases among the homologues of *pa*HisN and uncover their phylogenetic distribution. For this objective, a sequence similarity network (SSN) was generated with *pa*HisN as query sequence using the EFI-enzyme similarity tool.[124, 125] As database, the UniRef90 was used which is derived from the Uniprot database, with the difference that in the UniRef90 close homologues are represented by one representative sequence which minimizes redundancies. Specifically, all sequences with more than 90% sequence identity to one another over more than 80 % of the sequence are clustered and only one sequence is added to the database.[126]

The resulting SSN contained approximately 2800 nodes and was further analyzed using Cytoscape[127]. In the SSN, each sequence is represented by a node and each pair of sequences with a sequence identity above a certain threshold is connected by an edge. To identify clusters of closely related proteins this threshold was increased in a stepwise manner. 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 3.6 A).

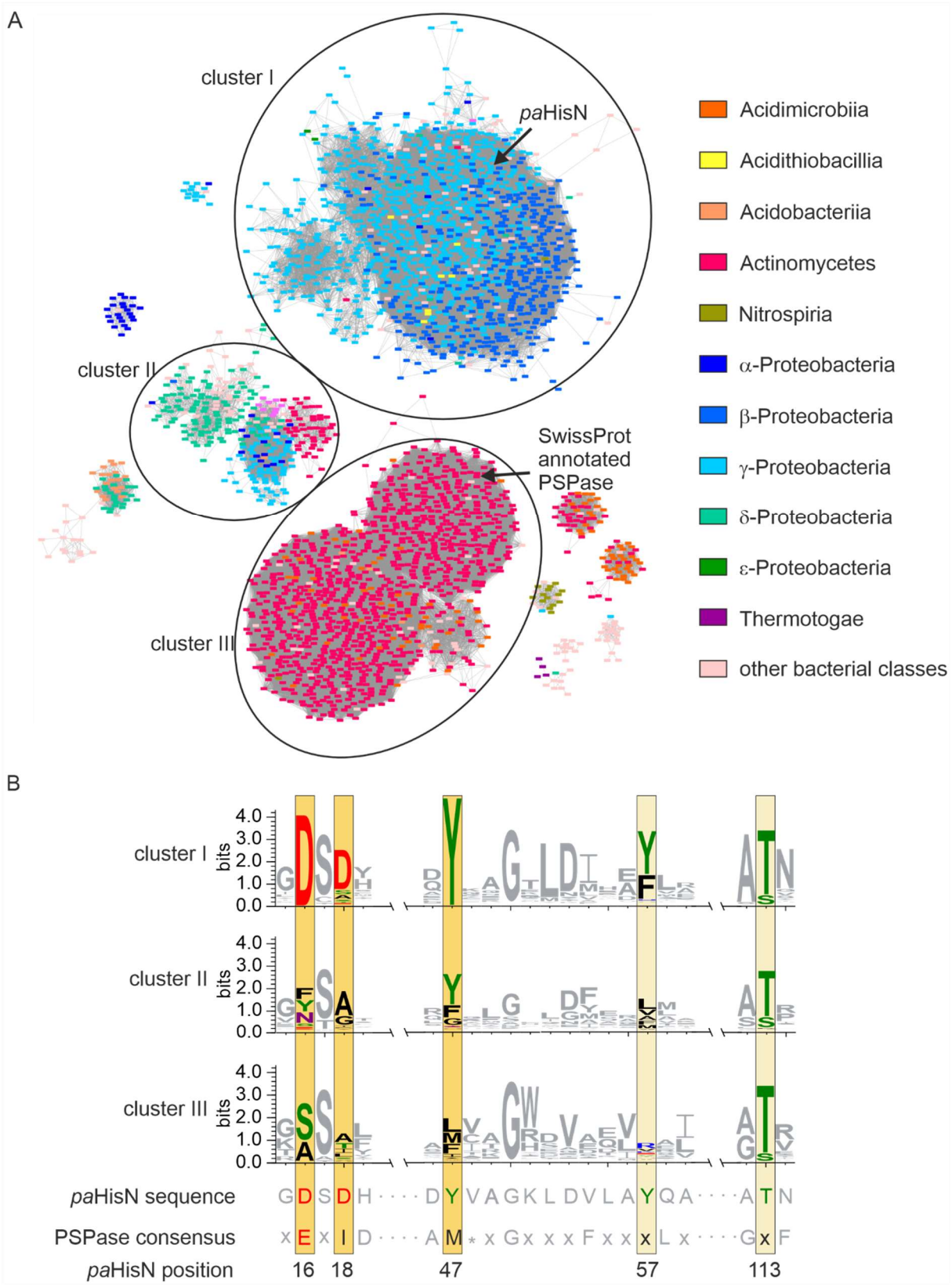

Figure 3.6: *In silico* analysis of the homologues of *pa*HisN.

(A) Sequence similarity network of homologues of *pa*HisN. In the sequence similarity network, each node represents a homologue of *pa*HisN, and the color of each node indicates the bacterial class of the corresponding organism. Nodes which share more than 45.8 % sequence identity are connected by an edge. At this sequence identity threshold, three major clusters can be distinguished, which were dubbed cluster I, II, and III. The query sequence is located in cluster I (black arrow) which is mostly composed

of β- and γ-proteobacterial sequences. Cluster II contains 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 the only sequence with a specific function assignment according to the UniProtKB/Swiss-Prot database. This sequence is annotated as PSPase (black arrow). (B) Sequence logos for the sequences of the three clusters compared to the *pa*HisN sequence and a consensus sequence of PSPases. Fingerprint residues are highlighted in bright yellow, while 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: red: acidic, black: hydrophobic, green: hydroxyl group, purple: amide, blue: basic.

At this threshold, around 100 sequences had no connections to any other sequence or were in clusters with less than 10 sequences. These sequences were excluded from further analysis and are not shown in Figure 3.6 A. To analyze whether the assignment of the sequences to the different clusters correlated with the affiliation to a phylogenetic group, nodes were colored according to the corresponding bacterial class.

Cluster I contained the query sequence *pa*HisN and was mainly populated by 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 the primary function. To test this hypothesis, the sequences from this cluster were retrieved with the help of Simon Holzinger. Then, a sequence logo was created and compared to both the sequence of *pa*HisN and a consensus sequence of PSPases which was derived from the PSPase sequence logo shown in Figure 3.5 (Figure 3.6 B). Interestingly, at two of the three fingerprint positions, namely D16 and Y47, the sequence logo of cluster I showed a strong conservation of an aspartate and a tyrosine, respectively. The third fingerprint residue, D18, showed weaker conservation but was still the most frequently encountered amino acid at this position in cluster I. Together, this led to the conclusion that most if not all sequences from cluster I were HolPases. This assertion was further strengthened by the presence of a tyrosine at the position equivalent to Y57 and a threonine at the position equivalent to T113 in approximately half of the sequences.

Cluster II contained sequences from a wide variety of phylogenetic groups such as α-, γ-, δ-Proteobacteria and Actinomycetes. The occurrence of such a wide variety of organisms in one cluster was surprising especially since for example γ-proteobacterial sequences were also present in cluster I. A close analysis of the sequences of cluster II revealed that many of them were significantly longer than *pa*HisN and approximately half of them were annotated as 1-acyl-*sn*-glycerol-3-phosphate acyltransferase. The PFAM entries of several of those annotated 1-acyl-*sn*-glycerol-3-phosphate acyltransferases revealed that these proteins consisted of two domains, namely an N-terminal HAD domain and a C-terminal acyltransferase domain, which may at least in part explain the clustering of these sequences. A subsequent analysis of the sequence logo of cluster II showed that the DxD motif was absent and that Y47 was the only one of the three fingerprint residues which showed some degree of conservation in cluster II. Additionally, only T113 was also conserved in cluster II. Even though the fusion of two domains complicates the situation, the absence of two out of three HolPase defining residues indicates that the sequences in cluster II are no HolPases. At the same time, the sequence logo of cluster II is also not consistent with the PSPase consensus sequence as neither the ExI motif nor the conserved methionine could be found. The enzymatic function of this cluster remains therefore unclear and would require additional *in vitro* experiments with representatives.

Cluster III was mostly composed of sequences from Actinomycetes and Acidimicrobiia. This cluster also contained the only sequence within the SSN with a specific function annotated according to the UniProtKB/Swiss-Prot database. The Swiss-Prot section of the UniProt database contains manually annotated sequences which suggests a higher reliability of the annotations compared to computationally annotated databases.[128] However, the publication which is cited for this specific function assignment does not contain any *in vitro* data but deduced the PSPase function of this sequence from homology[129] which clearly reduces the reliability of this annotation. In the sequence logo, the positions which correspond to the fingerprint residues in *pa*HisN all showed a low level of sequence conservation. Apart from a conserved threonine which is equivalent to T113 there is no indication of a HolPase function. The low conservation also means that the sequence logo is inconclusive regarding a potential PSPase activity of cluster III. The enzymatic function of the proteins from cluster III therefore remains elusive and requires additional *in vitro* experiments.

In summary, the *in silico* analysis strongly indicated that the sequences from cluster I which were mostly found in β- and γ-Proteobacteria are HolPases.

3.2.5 Analysis of the phylogenetic distribution of different HolPases

The results from the previous section strongly suggested that the homologues of *pa*HisN represent the HolPases of β- and γ-proteobacterial organisms. To verify this conclusion, these phylogenetic classes should be searched for histidine-producing organisms without a known HolPase. To this end, a list with all organisms that are able to produce histidine should be generated first. To decide whether an organism is able to produce histidine, the KEGG database[107] was searched for organisms with annotated IGPDH, as this enzyme from histidine biosynthesis is highly conserved.[113] In a second step, another list was generated which contained all organisms from the KEGG database with an annotated HolPase that was homologous to the HolPase either from *M. tuberculosis* (KEGG identifier K05602, IMP superfamily), from *M. truncatula* (K18649, IMP), from *L. lactis* (K04486, PHP), or from *E. coli* (K01089, HisB-N-like, HAD). The corresponding data was downloaded from the KEGG database with the help of Dr. Julian Nazet.

Then, both lists were filtered for organisms that belonged to the phylum of Pseudomonadota -formally known as Proteobacteria[130]- and compared to one another. To condense the results, the individual species were allocated to their phylogenetic class. The frequency at which each HolPase occurred within a particular phylogenetic class was then depicted in a stacked bar plot. (Figure 3.7 A). Likewise, a missing HolPase was also represented in the same bar plot (Figure 3.7, grey bars).

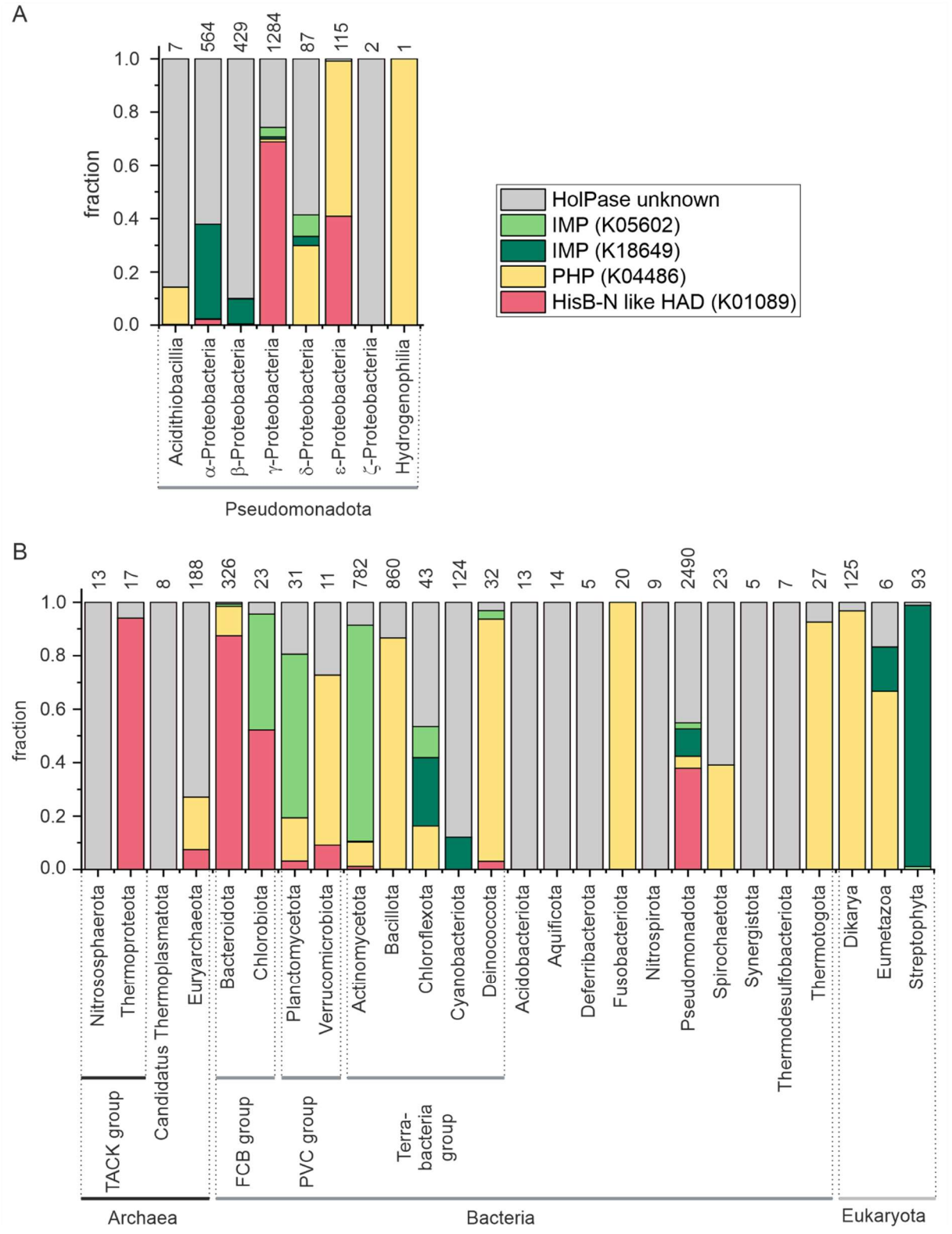

Figure 3.7: Phylogenetic distribution of the different non-homologous HolPases.

The stacked bar plots show the occurrence of the different types of HolPases for each phylogenetic group. The occurrence of a HolPase from the IMP superfamily homologous to the *M. tuberculosis* HolPase is indicated by a light green bar, of a HolPase from the IMP superfamily homologous to the *M. truncatula* HolPase by a dark green bar, of a HolPase from the PHP superfamily homologous to the *L. lactis* HolPase by a yellow bar, and of a HolPase from the HAD superfamily homologous to the *E. coli* HolPase by a red bar. The presence of IGPDH at the concurrent absence of a HolPase is indicated by a grey bar. Numbers above each bar represent number of species for each phylogenetic group. (A) The

diagram shows the fraction of each type of HolPase or of a missing HolPase for all classes of the phylum Pseudomonadota (B) The diagram shows the fraction of each type of HolPase or of a missing HolPase for different phyla, which are grouped into clades and domains according to the NCBI taxonomy browser.[131]

This comparison unveiled, that for a significant fraction of species from the Pseudomonadota, the enzyme which fulfills the HolPase function is not known. This includes data of Acidithiobacillia, ζ-Proteobacteria, and Hydrogenophilalia which was based on very few sequences and might therefore not be conclusive. However, in more than half of the species from α-Proteobacteria, β-Proteobacteria, and δ-Proteobacteria, the HolPase was also unknown, signifying that there was a very pronounced knowledge gap regarding the HolPases from these bacterial classes. Interestingly, the HolPase was also not know in about 25 % of the histidine synthesizing organisms from γ-Proteobacteria.

The missing HolPase from β- and γ-proteobacterial species fits well to the proposed HolPase function of the homologues of *pa*HisN from these two phylogenetic clades, further supporting this functional annotation and most likely closing the knowledge gap in many of those species.

Intrigued by the fact that the HolPase was missing for so many species, the analysis was extended to all phylogenetic classes (Figure 3.7 B). Interestingly, an annotated HolPase was missing in approximately 32 % of all species that did contain an IGPDH, which means that the missing HolPase in Pseudomonadota was not an exception but could also be observed for other phylogenetic groups. This finding also indicates that there are probably more types of HolPases which are yet to be discovered and which likely show little or no homology to the presently known HolPases. When comparing the three kingdoms of life, the biggest knowledge gap could be observed in archaea where only for one third of all histidine synthesizing species a HolPase was annotated suggesting that the search for a HolPase in this group might be most rewarding.

A second observation concerns the phylogenetic distribution of the different types of HolPases. Remarkably, the different types of HolPases do not seem to be confined to closely related phylogenetic groups. Instead, each type of HolPase appears to be scattered across the tree of life, as for example the *E. coli* type HisB-N which is found in Pseudomonadota, in bacterial species from the FCB group, but also in some archaea. The HolPase from the PHP superfamily is even more widely distributed and can be found in all three kingdoms of life. This observation indicates that there must have been extensive horizontal transfer of the different HolPase genes at the early stages of evolution.

3.3 Conclusion

An analysis of the structure and sequence of the gene product of PA0335, herein referred to as *pa*HisN, confirmed that it belongs to the HAD superfamily and revealed a significant similarity to PSPases which indicates common ancestry. Subsequent *in vitro* experiments confirmed the postulated HolPase function of *pa*HisN as its native function and a PSPase function as promiscuous side activity. Taken together, this indicates a distant evolutionary relationship between *pa*HisN and PSPases.

An alanine scan furthermore provided a fingerprint which consisted of a DxD motif together with a tyrosine which is postulated to be conclusive for a HolPase function in homologues of *pa*HisN. In search for related HolPases, a sequence similarity network was constructed which helped to identify many homologues in different bacterial species. Among those homologues, putative HolPases could be identified in β- and γ-Proteobacteria utilizing the previously established fingerprint. The function of the distantly related sequences remained however obscure, as they neither contained the sequence motif which is assumed to be conclusive for HolPase function nor the sequence motifs that are conserved in PSPases. This also means that no sequences could be identified which might bridge the functional space between PSPases and HolPases and show increased promiscuity or even bifunctionality. This most likely means that the presumed common ancestry between *pa*HisN-type HolPases and PSPases is confined to early stages of evolution.

The deduced HolPase function of the sequences from β- and γ-Proteobacteria was cross validated by a phylogenetic analysis of the occurrence of HolPases which showed that in β- and γ-Proteobacteria, there was indeed a significant fraction of histidine synthesizing organisms without an annotated HolPase. Moreover, it became apparent, that the HolPase encoding gene was obscure in many species. Specifically, only about one third of the species from the archaeal domain possessed an annotated HolPase. For this reason, the following chapter was dedicated to the search for a potential HolPase in archaea.

Lastly, the apparent lack of any pattern regarding the occurrence of the different types of HolPases suggests that the evolution of this function was characterized by several events of horizontal gene transfer, even between species that today belong to very different phylogenetic groups.

4 Characterization of a putative HolPase from *N. maritimus*

4.1 Introduction

Within the biosynthesis of histidine, the HolPase forms an exception inasmuch as it is not conserved between different species whereas the remaining enzymes are largely conserved.[70, 82] Specifically, non-homologous HolPases from three different protein superfamilies have been identified so far[59, 94, 91, 93] which complicates the identification of the HolPase of an organism by automated annotation. The results of chapter 3 additionally showed that there is a significant fraction of organisms which synthesize histidine but lack an annotated HolPase. This knowledge gap of the HolPase function is especially pronounced in the archaeal kingdom. In a previous study, an annotated HolPase was missing for all analyzed archaeal organisms[79] and the recently conducted search of the KEGG database revealed that for two thirds of all organisms the HolPase continues to be unknown (3.2.5). This is especially surprising for the archaea *Thermococcus onnurineus*, *Thermococcus kodakarensis*, *Thermococcus gammatolerans*, *Pyrococcus furiosus*, and *Picrophilus torridus* which, regarding their genes of histidine biosynthesis, exhibit many similarities to *E. coli*.[79] In these organisms, the genes of histidine biosynthesis are organized in an operonic structure with a gene order that is very similar to the one from *E. coli*[79, 72] and they possess a bi-functional *hisIE* gene which is also observed in *E. coli.* In a phylogenetic tree of concatenated *his* genes these five archaeal species moreover formed part of a larger cluster of bacterial species. It was therefore concluded that the histidine operon was horizontally transferred from an ancestor of these five archaeal species to an ancestor of *E. coli* or vice versa.[79] Despite this postulated gene transfer, no HolPase had been identified in these five archaea. [79]

In search for the missing HolPase, Lee et al. noted, that immediately downstream of the histidine operon of *Thermococcus onnurineus*, *Thermococcus kodakarensis*, and *Pyrococcus furiosus* there was a gene which was predicted to encode for a hydrolase from the HAD superfamily.[132] A subsequent *in vitro* characterization of the corresponding protein from *Thermococcus onnurineus* showed that this was indeed the missing HolPase. In accordance with the previously used nomenclature the *T. onnurineus* gene will therefore be called *hisN* and the corresponding protein *to*HisN.

Due to the putative acquisition of their histidine genes via horizontal gene transfer and the positioning of these organisms among bacteria in the phylogenetic tree, these three archaeal species are probably not representative for the archaeal kingdom. This also suggests that their HolPases might not be representative for the HolPases from other archaeal organisms but instead, yet another protein might fulfill the HolPase function.

So far, three of the five reported types of HolPases were enzymes from the HAD superfamily [132, 116, 59] which implies that uncharacterized proteins from this superfamily are probable HolPase candidates. The fact that the histidine genes are often encountered in operon-like structures[82, 79] further indicates that uncharacterized genes in the vicinity of annotated histidine genes are very promising HolPase candidates. Intriguingly, Fondi et al. reported that in *Nitrosopumilus maritimus* there was an uncharacterized gene located between other histidine genes which was furthermore predicted to be a phosphatase.[79]

To find out, whether the uncharacterized gene from *N. maritimus* indeed encoded for a HolPase, first, the genomic context was analyzed. Then, its AlphaFold2[100] predicted structure was compared to previously reported HolPases. This comparison revealed that the function-determining cap structure of the *N. maritimus* protein was not homologous to the cap of either *pa*HisN or *ec*HisB-N. A comparison

to *to*HisN showed moderate similarities which were however inconclusive regarding the function of the *N. maritimus* protein. Therefore, a functional *in vitro* characterization of the protein was performed which corroborated the suspected HolPase function of the *N. maritimus* enzyme. This led to the conclusion that the *N. maritimus* HolPase and its distant homologue *to*HisN may be representative for a type of HAD enzyme which evolved independently from *pa*HisN and *ec*HisB-N and fulfills the missing HolPase function in archaea. To test this hypothesis and identify other archaeal HolPases among the homologues of the *N. maritimus* HolPase, an alanine scan was performed to establish a fingerprint which should be a conclusive criterion for a HolPase that shared the same fold as the *N. maritimus* HolPase. This fingerprint was finally used in combination with a sequence similarity network to find putative HolPases among the homologues of the *N. maritimus* HolPase and uncover their phylogenetic distribution.

4.2 Results and Discussion

4.2.1 Analysis of the genomic neighborhood and predicted structure of *nm*HisN

The work from Fondi et al. indicated that the missing HolPase from *N. maritimus* might be encoded by a uncharacterized gene which was located between *hisC* and *hisB*.[79] To verify this, the genomic neighborhood of the *hisB* gene was examined utilizing the STRING database.[133] In the three archaeal species *Nitrosopumilus maritimus*, *Nitrosopumilus sp.* AR2, and *Thaumarchaeota sp.* SCGC AB629123 an uncharacterized gene, called Nmar_1556, NSED_00725, and ARWQ01000001_gene1227, respectively, was indeed found between the two genes *hisC* and *hisB* (representatively schown for *N. maritimus* in Figure 4.1). This uncharacterized gene which was predicted to encode for a HAD domain protein with a hydrolase function according to the PFAM database[134] confirming the previously reported observations.[79] It is of note that the histidine genes form an operon like gene cluster in *N. maritimus* and that the *hisE* gene was not encountered in this gene cluster.

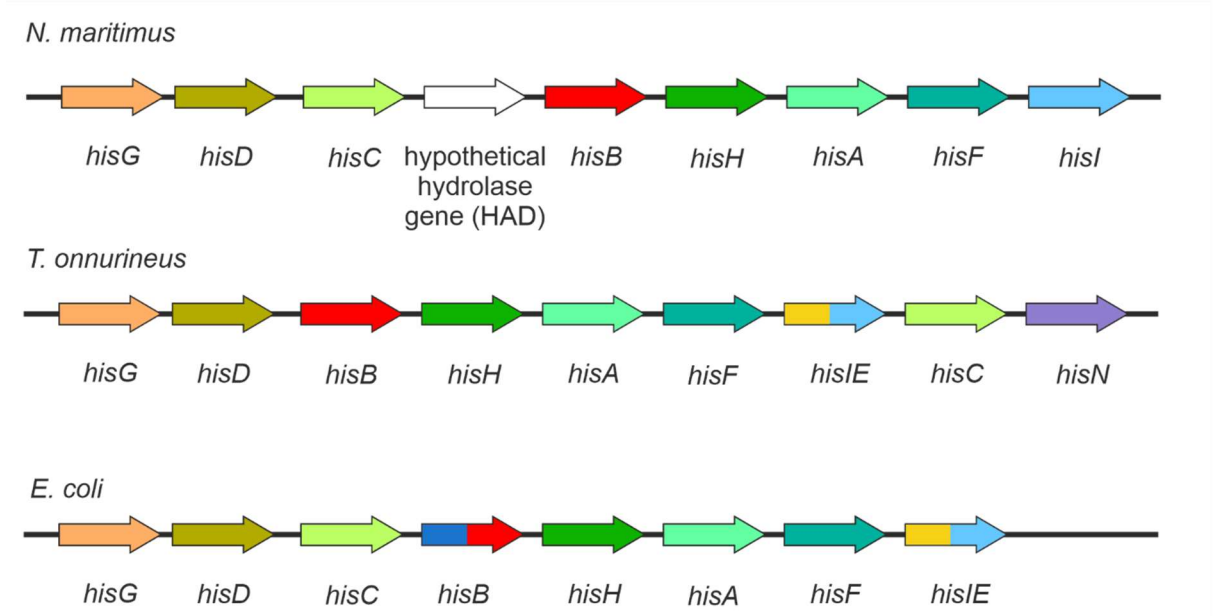

Figure 4.1: Comparison of gene loci of histidine biosynthesis of different species.

Shown is the genomic organization of the histidine synthesizing genes from the two archaea *N. maritimus* and *T. onnurineus* and from the bacterium *E. coli*.[79] Each gene is represented by an arrow

which indicates the directionality of the coding sequence and gene names are given below each arrow. Homologues are represented in the same color and bifunctional genes are indicated by arrows with two different colors. In *N. maritimus*, the gene between *hisC* and *hisB,* called Nmar_1556, lacks a specific functional annotation and is predicted to encode for a hydrolase which belongs to the HAD superfamily. Despite the high variability which is normally observed in the gene order of the histidine pathway, the relative arrangement is identical for *N. maritimus* and *E. coli*, and still very similar in *T. onnurineus*.

Consistent with the previously used nomenclature, the uncharacterized gene Nmar_1556 from *N. maritimus* and its gene product will be referred to as *hisN* and *nm*HisN, respectively. A comparison of the genomic neighborhood in *N. maritimus* with the genomic loci from *T. onnurineus* and *E. coli* showed that, surprisingly, the similarities in the gene order between *N. maritimus* and *E. coli* were even more pronounced than between *N. maritimus* and *T. onnurineus.* This was unexpected, considering the vast evolutionary distance between *N. maritimus* and *E. coli* and the generally moderate conservation of the genomic organization of the histidine genes.[79] The location of *hisN* in *N. maritimus* was moreover exactly the same as the HolPase encoding part of the bi-functional *hisB* gene in *E. coli*. However, the *N. maritimus* gene was not fused to the ensuing *hisB* gene and the putative HolPase had not been caught by automated functional annotation, which implied low homology and a convergent evolution rather than common ancestry. The equal location of the (putative) HolPase genes upstream of the *hisB* gene might therefore be a mere coincidence.

To gain more insights on *nm*HisN, the AlphaFold2[100] predicted structure was retrieved from the Uniprot database and compared to the structures of previously discovered HolPases of the HAD superfamily, *pa*HisN, *ec*HisB-N and *to*HisN (Figure 4.2).

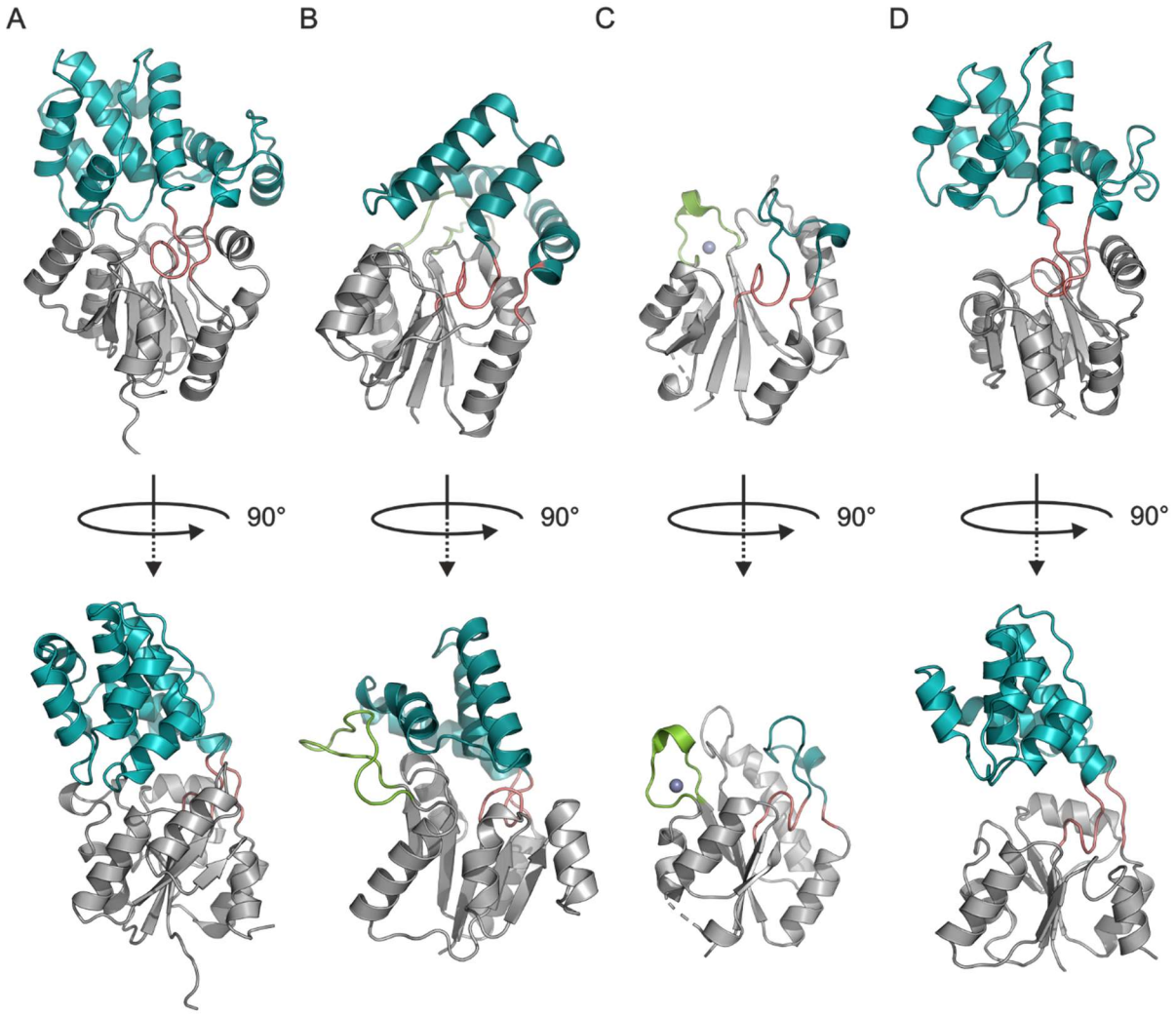

Figure 4.2: Comparison of the structure of *nm*HisN to other HolPases of the HAD superfamily.

(A) The structure of *nm*HisN as predicted by AlphaFold2[100] is typical for a HAD enzyme and consists of a core Rossmann fold (grey) that is modified by a large insertion (cyan) which is inserted in the flap structure (red) and consists of nine helices. (B, C) Shown are the structures of the HAD superfamily HolPases *pa*HisN, as predicted by AlphaFold[100], and the crystal structure of *ec*HisB-N[59] (PDB-ID: 2FPU). Both enzymes consist of a core Rossmann fold (grey) as *nm*HisN. However, the insertions in the flap structure differ significantly from the one in *nm*HisN and fold as four helical bundle (cyan) in *pa*HisN and short loop (cyan) in the case of *ec*HisB-N. Moreover, both structures contain an additional insertion (green), which folds as extended loop in the case of *pa*HisN and a turn with a bound zinc ion (grey sphere) in the case of *ec*HisB-N. (D) The structure of *to*HisN as predicted by AlphaFold[100] also exhibits a core Rossmann fold (grey) which, similar to *nm*HisN, contains only one large insertion which consists of seven helices and is inserted at the equivalent position as the cap in *nm*HisN. The RMSD value for *nm*HisN and *to*HisN was calculated to be 2.73 Å over 1162 atoms.

The analysis of the structure of *nm*HisN (Figure 4.2 A) unveiled that it consists of a core Rossmann fold with six parallel β-sheets with the strand order 6-5-4-1-2-3, corroborating its classification as HAD superfamily protein.[51] The core Rossmann fold is decorated by one large insertion in the flap structure (red) which folds as nine helices (cyan). This cap structure is very different from the caps of *pa*HisN or *ec*HisB-N, which fold as four helical bundle and short loop, respectively (Figure 4.2 B, C, cyan). Moreover, *pa*HisN and *ec*HisB-N contain a second insertion, which is absent in *nm*HisN. These differences indicated that *nm*HisN was neither closely related to *pa*HisN nor to *ec*HisB-N. However, the

structure of the archaeal HolPase *to*HisN (Figure 4.2 D) showed some similarities to *nm*HisN, as it is also characterized by only one large helical insertion in the flap structure. Unlike the cap of *nm*HisN, the cap of toHisN only forms seven helices and is composed of 114 amino acids instead of 145 amino acids as in *nm*HisN. This means, that apart from an analogous overall composition of structural elements, the similarities are limited, which is underscored by an RMSD value of 2.73 Å over 1162 atoms. A global sequence alignment furthermore unveiled a sequence identity of 23.9 % and a sequence similarity of 41.9 %. These values are in the same range as the values for the two proteins HAD superfamily proteins HisB-N and GmhB which are distantly related but catalyze different reactions.[83] The moderate structural similarities in combination with the limited sequence similarity therefore indicate a distant evolutionary relationship of *nm*HisN and *to*HisN but were considered too low to deduced a HolPase function of *nm*HisN.

4.2.2 *In vitro* characterization of *nm*HisN

So far, the only data concerning the catalytic function of an archaeal HolPase were reported for *to*HisN. The moderate sequence identity between *nm*HisN and *to*HisN does however not allow for a reliable prediction of the function of *nm*HisN. Therefore, a detailed *in vitro* characterization of *nm*HisN was performed.

To this end, the *hisN* gene from *N. maritimus* was codon optimized for the expression in *E. coli*, equipped with *Bsa*I digestion sites at the N- and C-terminus, and cloned into a pUR23 expression plasmid with an N-terminal His_6-tag by golden gate cloning (8.3.4.5). The gene was expressed in a *ΔhisB* strain (8.4.2) to exclude any possible contamination with host cell *ec*HisB-N. The protein was then purified by affinity chromatography followed by size exclusion chromatography (8.4.3, 8.4.4, 8.4.5).

Mass and purity of the protein preparation were assessed by SDS-PAGE (8.5.2, Figure 3.2 A) which showed a single band with an estimated molecular weight of approximately 34 kDa which fits well to the theoretical monomer weight of 35.6 kDa. Next, the oligomerization state was investigated by a combination of size exclusion chromatography and static light scattering (8.5.4, Figure 4.3 B). The experimentally obtained value for the number average molar mass Mn was 34.8 kDa and the value for the mass average molar mass Mw was 35.0 kDa which are both in the range of the theoretical weight of the monomer. A Mw/Mn ratio of 1.006 further indicated a monodisperse solution of only one molecular species.

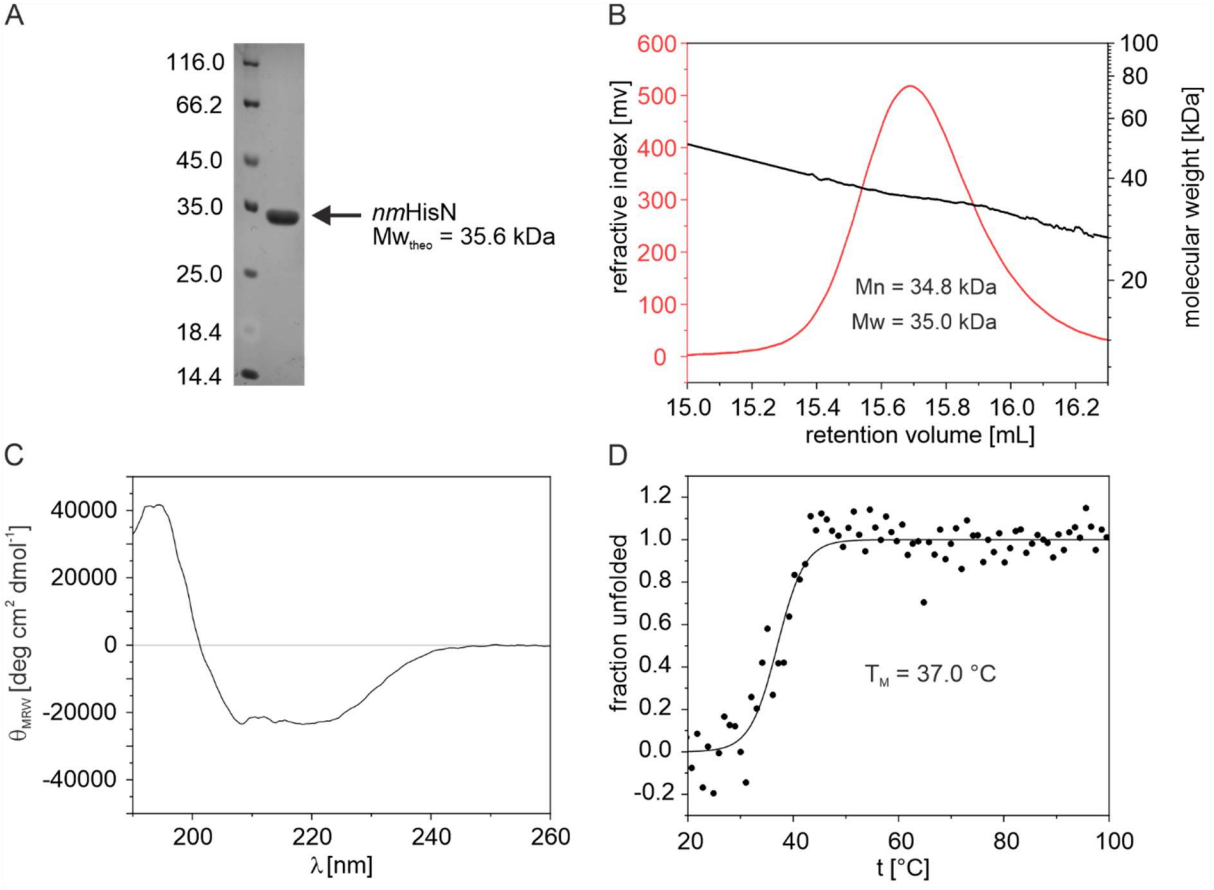

Figure 4.3: Biophysical characterization of *nm*HisN.

(A) SDS-PAGE (13.5 % acrylamide) of purified *nm*HisN (3 µg). The gel shows a single band corresponding to a molecular weight of 34 kDa which is in line with the theoretical monomer weight of 35.6 kDa. (B) Size exclusion chromatography followed by static light scattering yielded a number average molar mass (Mn) of 34.8 kDa and a mass average molar mass (Mw) of 35.0 kDa indicating that *nm*HisN formed a monomer in solution. (C) The CD spectrum of *nm*HisN (in 20 mM KP, pH 7.5) was characteristic for a folded protein. (D) The melting curve of *nm*HisN as monitored by far-UV CD could be fit to a two-state model without folding intermediates yielding a melting temperature (T_M) of 37.0 °C.

Next, the secondary structure of *nm*HisN was tested by CD spectroscopy (8.5.3). To prevent extensive background absorption the Tris buffer was exchanged to 20 mM KP prior to the measurements using a NAP 5 column (8.4.6). The recorded CD spectrum (Figure 4.3 C) exhibits minima at around 207 nm and 220 nm, a zero crossing above 200 nm, and a maximum at around 195 nm which is characteristic for a folded protein. Next, the melting point was determined by monitoring the CD signal at 220 nm while the sample was heated at a constant rate (8.5.3). A plot of the recorded data showed a sigmoidal curve (Figure 4.3 D) which is indicative of a transition from a folded state to an unfolded state. The data was therefore fitted according to a two-state model which gave a T_M value of 37.0 °C. This result is in line with the growth temperature of *N. maritimus* which ranges from 9-29 °C.[135]

The biophysical characterization confirmed that *nm*HisN was a well folded protein in solution with no detectable impurities, which laid the foundation for the subsequent functional analysis. To examine the putative HolPase function, the turnover of HolP was measured utilizing a coupled photometric enzyme assay (8.5.5). Turnover of HolP could indeed be detected which confirmed the presumed HolPase activity of *nm*HisN. To quantify the detected HolPase activity, steady-state kinetic experiments were

performed and a plot of the reaction rate against the substrate concentration gave a saturation curve (Figure 4.4 A).

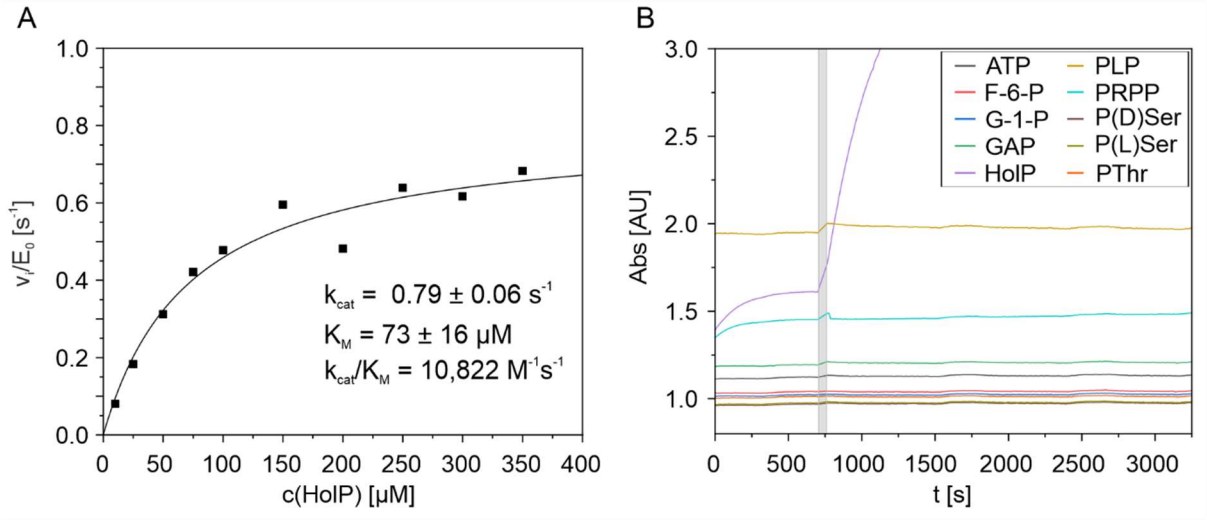

Figure 4.4: Functional characterization of *nm*HisN at 25°C.

(A) Substrate saturation curve for the turnover of HolP by *nm*HisN. (B) Substrate screen with ten different substrates (500 mM); the grey bar indicates the addition of the *nm*HisN (5 µM). Abbreviations: adenosine triphosphate: ATP, fructose-6-phosphate: F-6-P, glycerol-1-phosphate: G-1-P, glyceraldehyde-3-phosphate: GAP, pyridoxal phosphate: PLP, phosphoribosyl pyrophosphate: PRPP, *o*-phospho-D-serine: P(D)Ser, *o*-phospho-L-serine: P(L)Ser, *o*-phospho-L-threonine: PThr.

The data was fitted with a hyperbolic function according to Michaelis-Menten which yielded a k_{cat} value of 0.79 s^{-1} and a K_M value of 73 µM. The k_{cat} value is at the lower end of the range of previously reported k_{cat} values from other HolPases which were between 1 s^{-1} and 4 s^{-1}.[93, 94, 105, 115] The K_M value is in the same order of magnitude as the K_M values of other HolPases which are in the range of 32 µM to 400 µM.[93, 105, 120, 59] The catalytic efficiency k_{cat}/K_M for *nm*HisN was calculated to be 10,822 $M^{-1}s^{-1}$ which is about five times lower than the k_{cat}/K_M of *ec*HisB-N, which is 57,437 $M^{-1}s^{-1}$, and almost 40 times lower than the k_{cat}/K_M of *pa*HisN which was calculated to be 400,000 $M^{-1}s^{-1}$. This comparison demonstrates that the chemical steps of the HolPase reaction are compatible with higher reaction rates and a higher catalytic efficiency than observed for *nm*HisN. It is of note, that the assay temperature of 25 °C was probably close to the optimal reaction temperature, given that it corresponds well to the growth conditions of *N. maritimus* and taking into account that the T_M of *nm*HisN is 37°C. A possible explanation for the moderate activity would be other non-ideal experimental conditions, as for example suboptimal pH, high or low salt concentrations, or a missing cofactor like Co^{2-} or Mn^{2-} which might be preferred over Mg^{2+}. An alternative explanation would be, that the HolPase activity was a side activity of *nm*HisN. For this reason, ten phosphorylated compounds of different shapes and sizes were tested as alternative substrates. However, no other activity could be detected at and enzyme concentration of 5 µM and substrate concentrations of 500 µM (Figure 4.4 B) which is surprising, given that the HAD superfamily is well-known for its pronounced substrate promiscuity.[28, 106] What is more, the homologous *to*HisN showed promiscuous side activities for adenosine monophosphate, fructose-6-phosphate, and *o*-phospho-L-serine[79], however neither fructose-6-phosphate nor *o*-phospho-L-serine were dephosphorylated by *nm*HisN.

In conclusion, the functional experiments confirmed the presumed HolPase activity of *nm*HisN which is most likely the primary function of this enzyme.

Encouraged by the results of the study on *pa*HisN, we wanted to deepen our understanding of the HolPase function of *nm*HisN and establish a fingerprint by which other HolPases should be identified among the homologues of *nm*HisN.

4.2.3 Analysis of the functionally relevant residues in *nm*HisN by alanine scanning

With the goal to identify the residues which are critical for substrate binding and the turnover of HolP, an alanine scan was performed on residues of the active site of *nm*HisN. Parallel to the alanine scan that was previously performed on *pa*HisN, residues of the catalytic machinery which are highly conserved among all HAD enzymes were excluded from the alanine scan. In the case of *nm*HisN, the catalytic machinery is represented by D12 and D14 which correspond to the DxD motif, T189 and K224, which correspond to the conserved T/S and R/K residues, and D247 and D251 which form part of the DxxxD motif (Figure 4.5 A, blue sticks). [51]

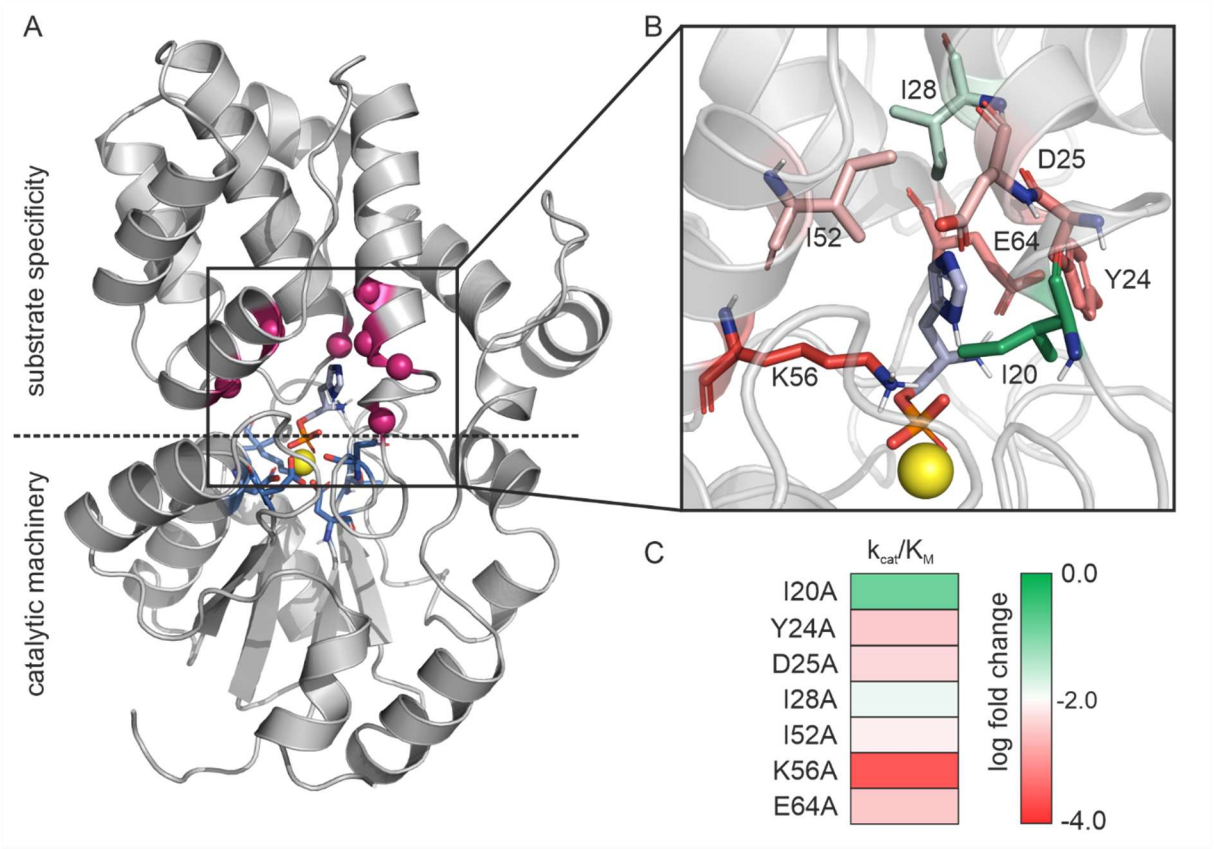

Figure 4.5: Alanine Scan of active site residues of *nm*HisN.

(A) AlphaFold[100] predicted structure of *nm*HisN with docked HolP (light blue). In HAD enzymes, 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 of the k_{cat}/K_M value to a color on the scale from green over white to red.

Instead, residues that were unique for this type of HolPase and which were specifically adapted to the substrate HolP should be mutated to alanine. To help with the identification of candidate residues in the vicinity of the bound substrate, a docking experiment with HolP was performed by Dr. Julian Nazet. An analysis of the active site resulted in the identification of seven residues which were less than 4 Å away from the substrate and the side chains of which were oriented towards the substrate (Figure 4.5A, purple spheres). This set of seven residues consisted of I20, Y24, D25, I28, I52, K56, and E64 which all formed part of the helical cap structure that covered the active site (Figure 4.5 B).

Then, all selected positions were individually mutated to alanine (8.3.4.4), the resulting genes were expressed (8.4.2) and the corresponding proteins were purified according to the same scheme as the wildtype enzyme (8.4.3, 8.4.4, 8.4.5). As judged by SDS-PAGE (8.5.2) all proteins could be obtained with good purity (Figure S 18) and CD spectroscopy (8.5.3) confirmed that all proteins adopted a well folded conformation in solution (Figure S 19). Next, the influence of each mutation on the enzymatic function was analyzed by steady-state kinetic experiments (Figure S 20).

To illustrate the spatial arrangement, the effect of each mutation was highlighted in a zoomed in view on the active site (Figure 4.5 B). Specifically, the color of each residue indicates the degree of activity loss upon mutation to alanine from green for a mild decrease to red for a drastic decrease in activity. The color scale was normalized to the logarithm of the fold change of the k_{cat}/K_M value as a measure for the activity decrease (Figure 4.5 C) The associated catalytic parameters are listed in Table 4.1.

Table 4.1: Catalytic parameters of *nm*HisN single mutants to alanine at 25°C.

Variant	k_{cat} [10^{-2} s^{-1}]	K_M [µM]	k_{cat}/K_M [$s^{-1}M^{-1}$]
wt	79 ± 6	73 ± 16	10,822
I20A	8.8 ± 0.4	58 ± 7	1,520
Y24A	1.0 ± 0.1	276 ± 61	36
D25A	2.9 ± 0.1	582 ± 28	49
I28A	13.2 ± 2.4	910 ± 231	145
I52A	4.1 ± 0.8	506 ± 152	82
K57A	0.17 ± 0.02	607 ± 107	2.8
E64A	0.22 ± 0.01	63 ± 7	34

The first mutation of the alanine scan was I20A which led to a 9-fold decrease in k_{cat} and slight improvement of the K_M value. Thus, while substrate affinity is retained, the turnover number is diminished. This might be explained by an increase in non-productive binding of HolP which may for example lead to improper alignment for the nucleophilic attack and therefore impact the chemical step. However, the overall decrease in activity was moderate, which identified I20 as the least important residue of the tested ones.

The next mutations concerned Y24 and D25. In both cases, the mutation to alanine led to a significant decrease in the k_{cat} value by a factor of 79 and 27 and a moderate increase in K_M by a factor of approximately 4 and 8, respectively. Again, substrate affinity is largely retained, indicating, that the substrate is still bound whereas turnover is compromised. Since the catalytic mechanism is mostly

conserved across the HAD superfamily[51] an immediate involvement in the catalytic cycle, as for example by protonation of the substrate, seems unlikely. Therefore, an increased fraction of non-productive binding of HolP seems to be the most plausible explanation for the observed effect.

The fourth probed mutation was I28A. This mutation signified a decrease in the k_{cat} value by a factor of 6 and an increase in the K_M value by a factor of 12 which was the strongest reduction in binding affinity for HolP that was observed. This was surprising, because (i) according to the docking I28 is further away from HolP than other residues from the alanine scan and (ii) the hydrophobicity of the sidechain prevents strong polar interactions with the substrate. A possible explanation for the increased K_M might be an involvement of I28 in the correct packing of the cap module, thus influencing substrate binding in an indirect manner.

The mutation I52A led to a decrease in the k_{cat} value by a factor of 27 and to an increase in the K_M value by a factor of 7. As in the case of I28, an involvement of this residue in the correct packing of the cap and preorganization of the active site seems more plausible than a direct interaction with the substrate.

The next mutation was K57A which led to a drastic reduction of the k_{cat} value by more than 460-fold and an increase in the K_M value by a factor of 8, which made this residue the most critical one of all tested residues. According to the docking analysis, the side chain of K56 is in close proximity to the phosphate group suggesting a potential role in the binding of the phosphate group or stabilization of the negative charge in the transition state, similar to the function of the conserved lysine from the catalytic machinery (K224). Depending on the pK_a of the amino group, K57 might also protonate the nascent hydroxyl group of the product, which is normally done by the second aspartate of the DxD motif.[51] However, due to the length and flexibility of the side chain of lysine, alternative conformations cannot be ruled out and different modes of actions as for example an interaction with the imidazole moiety might also be possible.

The last mutation which was tested was E64A. This mutation led to a 360-fold reduction of the k_{cat} value while the K_M value remained unaffected which resembles the effect of the mutations of D24 and Y25 which are located in close vicinity to E64. This suggests that the imidazole ring is probably pointing towards the back of the active site, interacting with Y24, D25, I52, and E64. The three residues D24, Y25, and E64 are reminiscent of the fingerprint residues from *pa*HisN, D16, D18, and Y47. Different from *pa*HisN however, D24 and Y25 form part of the same helix and E64 is located at the same side of the active site, whereas in *pa*HisN D16 and D18 form part of the same helix and Y47 is located at the opposite side of the active site. The importance of two acidic residues both in *pa*HisN and *nm*HisN nevertheless implies a similar mode of action which, is most likely the product of convergent evolution.

In summary, the alanine scan identified four charged residues, namely D24, Y25, K57, and E64 which were most critical for the HolPase function of *nm*HisN. These residues most likely define a fingerprint, which could classify homologues of *nm*HisN as HolPases. Moreover, the mutation of either of the two hydrophobic residues I28 and I52 also proved detrimental for the HolPase function. The additional occurrence of I28 or I52 may therefore support the classification of a protein as HolPase but neither amino acid alone seems to represent a conclusive criterion for a HolPase function.

4.2.4 *In silico* analysis of the homologues of *nm*HisN

The alanine scan of the active site of *nm*HisN revealed four charged residues which were critical for the HolPase function and two hydrophobic residues which could support the classification of a protein as HolPase. This fingerprint of functionally relevant residues should be used to identify other HolPases among the homologues of *nm*HisN and reveal their phylogenetic distribution. Like in the case of *pa*HisN, this question was addressed by creating an SSN, this time with *nm*HisN as query sequence, and by analyzing individual clusters by means of sequence logos. The SSN was again generated utilizing the EFI-enzyme similarity tool and the UniRef90 as sequence database which was also used previously (3.2.4).[124, 125] This search resulted in an SSN that consisted of 506 homologous sequences and which was further analyzed using Cytoscape[127]. Cluster formation started at a sequence identity threshold of about 37 % and at a threshold of 45.0 % two main clusters had formed (Figure 4.6 A, cluster I and cluster II). At this threshold, 82 sequences were either isolated or formed part of a small cluster consisting of less than five sequences. These sequences were excluded from further analysis which reduced the dataset to 434 sequences.

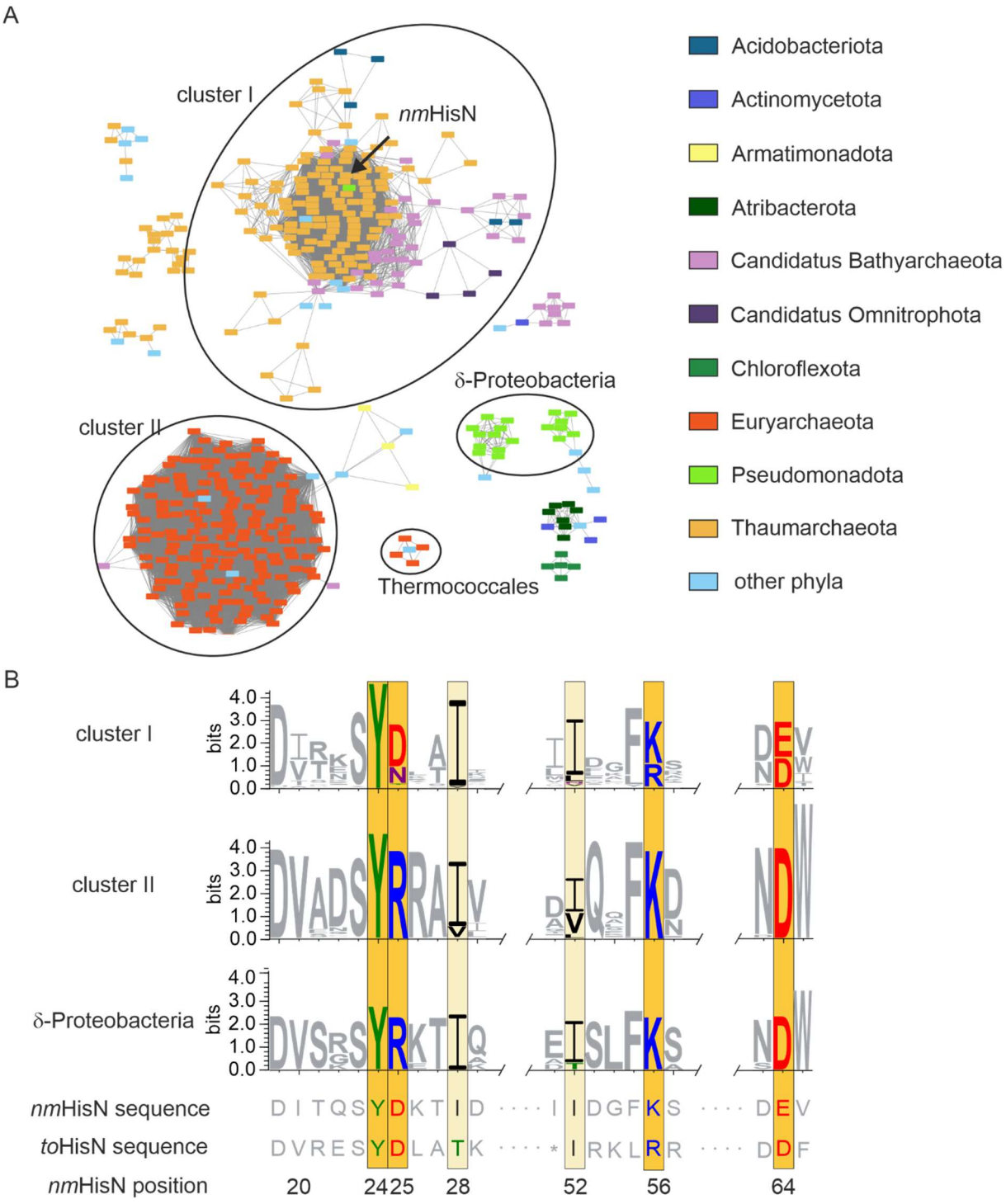

Figure 4.6: *In silico* analysis of the homologues of *nm*HisN.

(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 of each node indicates the phylum to which it belongs. Nodes which share more than 45.0 % sequence identity are connected by an edge. At this threshold, there are two major clusters, namely cluster I and cluster II. The node which represents *nm*HisN is located in cluster I which mostly contains sequences from Thaumarchaeota and the candidate phylum Bathyarchaeota. Cluster II almost exclusively comprises 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 to the sequences of *nm*HisN and *to*HisN. Fingerprint residues are highlighted in bright yellow, while 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.

The sequence of *nm*HisN was located in cluster I, which mainly contained sequences from Thaumarchaeota, the phylogenetic group to which *N. maritimus* belongs, and from the candidate phylum Bathyarchaeota. Both Thaumarchaeota and Bathyarchaeota belong to the so called TACK-Superphylum[136], which most likely explains the clustering of these two phyla. To find out, whether the sequences of this cluster were HolPases, their sequences were retrieved from the UniProt with the help of Simon Holzinger and a sequence logo was created and compared to the fingerprint residues of *nm*HisN and *to*HisN (Figure 4.6, B). In the sequence logo, a strictly conserved tyrosine was found at the position which corresponded to Y24 from *nm*HisN. Moreover, in most sequences there was an aspartate at the position equivalent to D25, a lysine or arginine was found at the position which corresponds to K56, and a glutamate or aspartate occupied the positions which corresponds to E62. Interestingly, arginine instead of K56 and aspartate instead of E62 are also found in *to*HisN, indicating that these conservative exchanges are compatible with the HolPase function. Remarkably, the two hydrophobic residues I28 and I52 were also largely conserved, or replaced by similar amino acids, like valine or leucine. In conclusion, these results strongly suggest that the sequences from cluster I represent HolPases.

Cluster II mainly contained sequences from Euryarchaeota. Interestingly, the sequences from the phylogenetic order of the Thermococcales, which belong to the phylum Euryarchaeota[131] were not contained in cluster II but formed an independent cluster, which suggests that the HolPase from the Thermococcales diverged from the one from other Euryarchaeota. To find out, if this divergence also concerned the function of these proteins, a sequence logo of cluster II was created. In the sequence logo, the amino acids tyrosine, lysine and aspartate were highly conserved at the positions which corresponded to the fingerprint residues Y24, K56, and E62, respectively, indicating HolPase function. Interestingly however, a highly conserved arginine was found at the position that was equivalent to D25. This drastic change in the chemical properties of the amino acid at this position put the HolPase function into question. To resolve this matter, the amino acids at the positions equivalent to I28 and I52 were analyzed which revealed a conserved isoleucine at the first position and either an isoleucine, leucine, or valine at the second position. Taken together, this conservation at five out of six positions supports the classification as HolPases even though this conclusion is less reliable than for cluster I due to the discussed replacement of an aspartate by an arginine.

Interestingly, the SSN also contains two small clusters which are 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. Specifically, all fingerprint positions show strong conservation and all except for one of these positions are identical either to *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, even though the data are not as conclusive as for cluster I.

A definitive answer regarding the function of cluster II and the δ-proteobacterial cluster however requires additional *in vitro* testing of representatives.

4.3 Conclusion

An analysis of the genomic neighborhood of the hisB gene in *N. maritimus* showed that there is an uncharacterized gene called Nmar_1556 that is located between *hisC* and *hisB.* The corresponding gene product (*nm*HisN) was annotated as putative phosphatase from the HAD superfamily, making it a promising candidate for the missing HolPase. A comparison to the only archaeal HolPase which was characterized so far, the HolPase from *Thermococcus onnurineus*[132] (*to*HisN), revealed a sequence identity of 24 % and moderate structural similarities, both being inconclusive regarding the function of *nm*HisN. Subsequent *in vitro* experiments on *nm*HisN confirmed its suspected HolPase function, thus closing this knowledge gap in the histidine biosynthesis of *N. maritimus*. An alanine scan of active site residues of *nm*HisN furthermore unveiled a fingerprint which consisted of a DY motif together with a lysine and a glutamate residue. This fingerprint is postulated to be conclusive for a HolPase function in homologues of *nm*HisN. A sequence similarity network then identified homologues which mainly fall into two different groups. The first group of enzymes is found in other Thaumarchaeota and the candidate phylum Bathyarchaeota and shows high sequence similarity to *nm*HisN. Additionally, these proteins also show a strong conservation of the fingerprint residues, suggesting that these proteins are HolPases. The second group consisted of a cluster of proteins from Euryarchaeota and a cluster of proteins from δ-Proteobacteria. In the sequences of both clusters, all fingerprint residues except for one were highly conserved. Even though the only difference concerned a conserved arginine instead of an aspartate, a HolPase function nevertheless seems likely for these enzymes. Interestingly, the SSN also contained sequences from Thermococcales which did not cluster with other Euryarchaeota but formed a separate small cluster which is in line with the previous results from Fondi et al. which showed that the genomic organization of the histidine genes in Thermococcales is unique within the Euryarchaeota and instead resembles the organization which is also found in bacterial species.[79]

A comparison of the sequence of *nm*HisN and *to*HisN showed that the fingerprint residues were largely conserved even though the sequence identity between the two enzymes was only about 24 %. This low overall sequence conservation in combination with the high conservation of the fingerprint residues implies that most sequences of the SSN are probably HolPases which have undergone significant changes due to divergent evolution which might be caused by evolutionary drift or adaptation to secondary selective pressure like temperature, pH, or salt concentration. The significant sequence divergence together with the broad phylogenetic distribution of the homologues across different superphyla suggests an ancient origin of this type of HolPase.

This conclusion might be useful when it comes to the evolution of the whole histidine biosynthetic pathway. The similarities in the genomic organization of the histidine genes between *N. maritimus*, *T. onnurineus*, and *E. coli* strongly suggest a horizontal gene transfer between the progenitors of those organisms as previously postulated for the progenitors of *T. onnurineus* and *E. coli*.[79] The fact that *E. coli* and the archaeal organisms differ in their HolPase suggest that the remaining histidine genes were most likely horizontally transferred prior to the acquisition of the HolPase in the donor and the acceptor species. This can be rationalized as follows: The presence of a HolPase in the donor species would lead to the question why every gene except for the one which codes for the HolPase was transferred. At the same time, the presence of a HolPase in an acceptor species makes no sense at the concomitant absence of other histidine genes. Therefore, the putative horizontal gene transfer most likely happened at the early stages of evolution, prior to the divergence of the archaeal phyla.

5 Directed evolution of HAD enzymes

5.1 Introduction

The evolution of a complex metabolism from a rudimentary prestage which was most likely characterized by a limited set of enzymes required some mechanism by which new reactions or new compounds were added. Such a mechanism was proposed by Ycas and Jensen who argued that gene duplication followed by functional diversification provides a resource from which new enzymes can be derived.[25, 26] Jensen based this hypothesis on the observation that in different pathways reaction sequences can be found which consist of analogous chemical mechanisms but act on different substrates. To explain this, Jensen envisioned a model by which an erroneous side reaction proved beneficial under certain circumstances. Following gene duplication, this side reaction was improved in one of the two gene copies whereas the second copy maintained the native function. As a result, two homologous enzymes with similar catalytic mechanisms but different substrate specificities are obtained. In support of this claim, several sets of analogous reactions from different pathways were cited.[26] One of these examples concerned the three steps of the canonical biosynthesis of L-serine and the last three reactions of the biosynthesis of L-histidine (Figure 5.1).

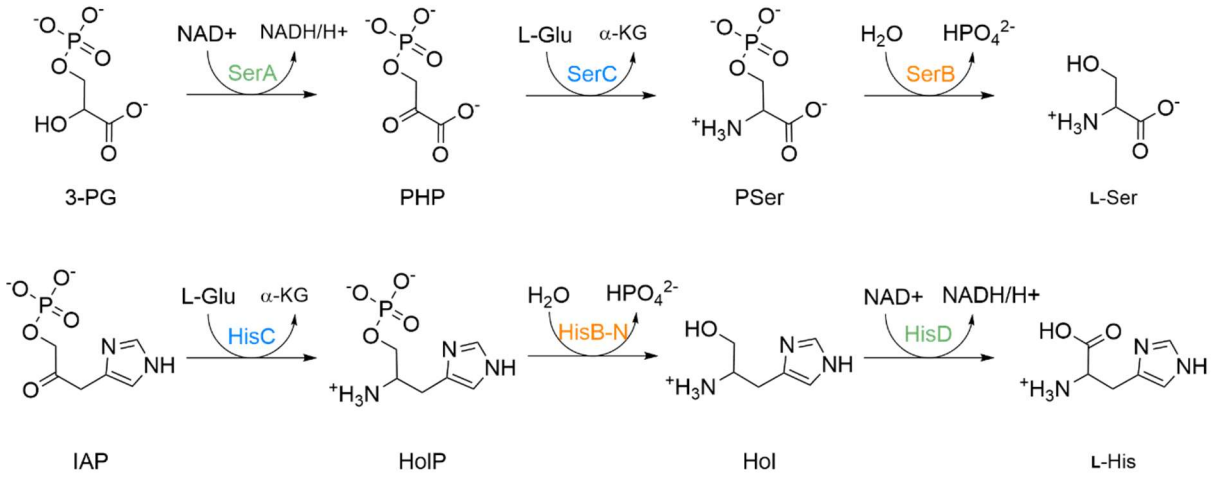

Figure 5.1: Analogous reactions in the biosynthesis of L-serine and L-histidine.

Shown are the three steps of the canonical biosynthetic route of L-serine and the last three steps of the biosynthesis of L-histidine. Enzyme names correspond to the nomenclature in *E. coli* and enzymes that catalyze analogous reactions are highlighted in the same color. Abbreviations: 3-PG: 3-phosphogylcerate, PHP: phosphohydroxypyruvate, PSer phosphoserine, L-Ser: L-serine, IAP: imidazole acetol phosphate, HolP: histidinol phosphate, Hol: histidinol, L-His: L-histidine.

In both cases the amino acid is produced from a precursor through a sequence of three analogous reactions, namely an NAD^+-dependent oxidation (SerA, HisD), a transamination whereby an amino group is transferred from L-glutamate to the L-Ser or L-His precursor (SerC, HisC), and a dephosphorylation (SerB, HisB-N). It was hence suggested that each enzyme pair should be examined for homology which would strengthen the claim of a shared evolution according to a duplication-divergence mechanism.[26] Intriguingly, as discussed in section 3.2.1, the HolPase *pa*HisN does exhibit considerable structural and sequential similarities with the PSPase *mj*SerB and shows promiscuous activity for PSer. Considering the narrow phylogenetic distribution of iso-functional homologues of *pa*HisN in β-Proteobacteria, γ-Proteobacteria and Acidithiobacillia, it seems plausible that this type of HolPase is evolutionary young. One could therefore imagine a potential evolutionary trajectory which

starts at a PSPase and leads to *pa*HisN-like HolPases. The feasibility of such a functional conversion was tested and the results from these experiments form the basis of the first half of this chapter. Specifically, *ec*SerB was used as a model PSPase on which a HolPase function should be established by means of directed evolution.

While analogous mechanisms and a conserved fold help to identify the products from divergent evolution in hindsight, substrate promiscuity or the erroneous turnover of a non-native substrate is believed to be the starting point of divergent evolution.[26, 31] Such a promiscuous side activity was discovered for the *E. coli* HolPase *ec*HisB-N which was able to rescue an auxotrophic Δ*serB* strain.[137] In a follow-up study, this latent side activity could be enhanced by error-prone PCR and subsequent selection for growth of the Δ*serB* deletion strain on minimal medium.[138] Albeit the evolution of PSPase from HisB-N does most likely not represent the actual course of evolution because *ec*HisB-N is evolutionary young whereas SerB shows a broad phylogenetic distribution, this nevertheless presents an interesting case-study regarding the evolvability of this scaffold towards new substrates. The second part of this chapter will therefore be focused on the improvement of the latent PSPase activity of *ec*HisB-N.

5.2 Results and Discussion

5.2.1 Investigation of a possible evolutionary trajectory from PSPases to HolPases

The structural similarities between *pa*HisN and PSPases (3.2.1), the shared set of catalytically relevant residues, the fact that *pa*HisN accepts PSer as substrate (3.2.2), and the relatively narrow phylogenetic distribution of *pa*HisN (3.2.4) indicate that *pa*HisN might be derived from an ancient PSPase. This potential evolutionary trajectory from a PSPase to a HolPase should be followed by means of laboratory evolution. To avoid any problems regarding gene expression, protein folding, or protein solubility in the cytoplasm during the selection in *E. coli*, the native PSPase from *E. coli* (*ec*SerB) was selected as starting enzyme.

To set the basis for all further experiments, the wildtype *ec*SerB should first be expressed and tested for both its native PSPase activity and a potential HolPase side activity. An expression plasmid encoding for *ec*SerB with a C-terminal His_6-tag was obtained from previous work by Fabian Ruperti.[139] This plasmid was used to transform a Δ*serB*Δ*hisB E. coli* strain from previous work by Dr. Bettina Rohweder[140] (8.2.3), the *serB* gene was expressed (8.4.2), cells were disrupted (8.4.3), and the target protein was purified by affinity chromatography (8.4.4) followed by size exclusion chromatography (8.4.5). Afterwards, the purity of the target protein was confirmed by SDS-PAGE (8.5.2, Figure S 21). To ascertain that *ec*SerB was properly folded and active, the turnover of the native substrate PSer was tested in a coupled enzymatic assay (8.5.5). As expected, a substrate saturation curve could be observed for increasing PSer concentrations. Upon fitting with a hyperbolic function according to Michaelis-Menten, a k_{cat} value of 12.6 s^{-1} and a K_M value of 382 µM were obtained, which corresponds to a k_{cat}/K_M value of 33,000 $M^{-1}s^{-1}$ (Figure S 22). Next, the turnover of HolP was tested using the same coupled assay (8.5.5). However, no turnover of HolP could be detected for concentrations of up to 1 mM HolP and of 5 µM *ec*SerB.

We speculated that the absence of any HolPase activity in *ec*SerB might be caused by the differences in size and charge of the two substrates. In detail, PSer is small compared to HolP and the phosphate group forms an ester with the side chain oxygen atom which upon dephosphorylation becomes the hydroxyl

group of the side chain. Since the phosphate group is bound by the catalytic machinery, the opposite end of the molecule consequentially faces the cap structure. In PSer, this part is defined by the negatively charged carboxylate group. In contrast, HolP is sterically more demanding than PSer, and the phosphate moiety forms an ester with an oxygen which is part of the carboxylate group in L-His. Therefore, the part of the molecule facing the cap structure is defined by the imidazole ring which is protonated to a significant degree at neutral pH. So instead of a small negatively charged moiety as in PSer, a larger, positively charged aromatic ring has to be bound by the cap resides in the case of HolP. These considerations implied that the active site needed to be reshaped and the electrostatic interactions had to be adjusted to allow for the turnover of HolP. To facilitate the parallel analysis of different combinations of mutations, concerted randomization of several active site residues followed by *in vivo* selection was chosen as strategy.

In preparation of the *in vivo* selection, the selection system was set up first. To allow for the selection of the best HolPase, the gene encoding for the HolPase had to be deleted from the *E. coli* genome and thus a Δ*hisB* deletion strain was constructed and tested for growth on M9 selective agar (8.2.2, Figure S 23). The *hisB* gene from *E. coli* is however bifunctional and codes for both the IGPDH and the HolPase. Since the *ec*SerB library should only be selected for its HolPase functionality, the missing IGPDH function had to be recovered. To this end, the monofunctional *hisB* gene from *Bacillus subtilis* which encodes only for a IGPDH was cloned into a plasmid equipped with a constitutively active promotor and used to transform the Δ*hisB* strain effectively yielding a HolPase deficient strain (Δ*holPase*). To test the selection system, Δ*holPase* cells were plated onto minimal medium. As expected, in absence of a HolPase encoding gene no growth could be observed within 8 days (Figure S 3). Cell growth could however be restored by transformation with a second plasmid which contained the *hisB-N* gene from *E. coli* that was also put under the control of a constitutively active promotor (Figure S 3).

After establishing a selection system, suitable active site residues of *ec*SerB should be selected for randomization. To this end, the crystal structure of the *ec*SerB homologue *mj*SerB with bound substrate[121] was analyzed (Figure 5.2 A). Based on the structural analysis, four residues in close proximity to the bound substrate PSer (Figure 5.2 B) which also showed high conservation in the sequence logo of representative PSPases were chosen (3.2.3, Figure 5.2 C).

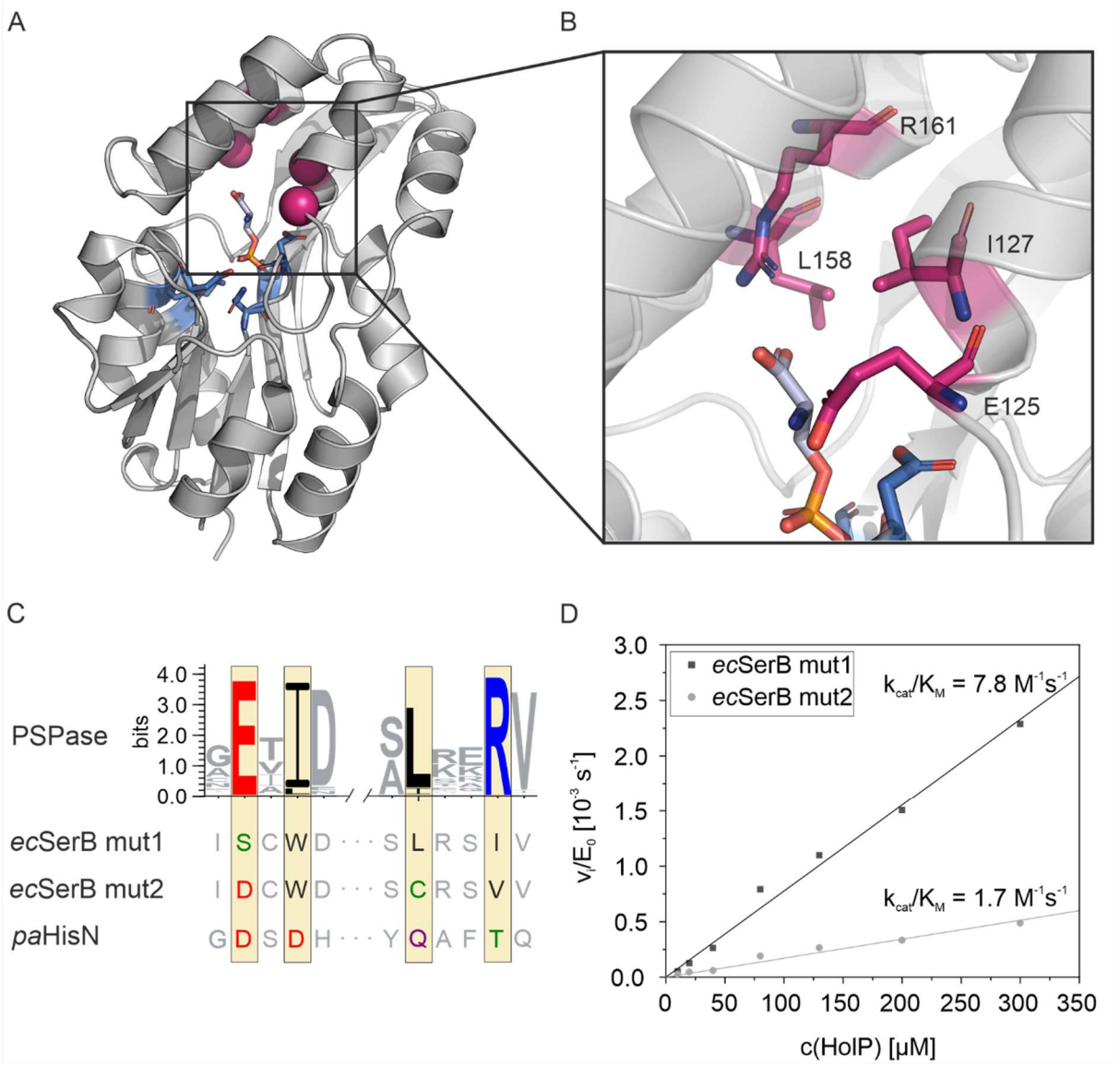

Figure 5.2: Establishing HolPase activity on *ec*SerB.

(A) Crystal structure of *mj*SerB[121] (PDB-ID: 1L7P) and (B) zoomed in view of the active site of *mj*SerB. The bound substrate PSer is represented as light blue sticks while residues of the catalytic machinery are represented as dark blue sticks and residues selected for randomization are marked by purple spheres (panel A) or shown as purple sticks (panel B). (C) Comparison of a PSPase sequence logo with the sequence of the two *ec*SerB mutants with the highest HolPase activity (mut1 and mut2) and with the sequence of the homologous HolPase *pa*HisN. Randomized positions as marked by yellow boxes. Color code: red: acidic, black: hydrophobic, green: hydroxyl group, purple: amide, blue: basic. (D) Linear fit for the turnover of HolP by *ec*SerB mut1 and *ec*SerB mut2 at 25°C.

These four residues were all randomized simultaneously using degenerate NNK primers (8.3.4.6). To assess the quality of the resulting repertoire, cells were transformed with the plasmid library (8.2.4) and the gene sequence of 25 individual colonies was determined by Sanger sequencing (8.3.1, Figure S 24). The sequencing data showed that, every amino acid except for threonine was represented at least once at one randomized position and the observed frequency of each amino acid was similar to the one expected for an NNK scheme. Based on the sequencing data and accounting for premature stop codons or fragmented genes, and based on the calculated number of transformed colonies, the library size was estimated to be around $8 \cdot 10^7$. The simultaneous randomization of 4 positions introducing 32 possible

codons at every position results in $1.05 \cdot 10^6$ possible codon combinations. This means that the necessary library size was oversampled by a factor of approximately 80 indicating that most possible codon combinations should be actually represented in the library.[141]

Encouraged by these results, the Δ*holPase* strain was transformed with the *ec*SerB library, streaked on petri dishes with M9 minimal agar and incubated at 37 °C. After several days, individual colonies became visible. These colonies were picked and analyzed by Sanger sequencing (8.3.1).

Out of 14 sequenced colonies, 11 contained full length genes with an unambiguous read which in turn encoded for 8 different *ec*SerB mutants (Table S 5). To minimize the number of false positive results, the plasmids were used to retransform the Δ*holPase* selection strain and to repeat the selection experiments. The genes which enabled the fastest complementation of the HolPase deficiency (Figure S 25) were subcloned into expression vectors and the corresponding proteins were produced and purified in the same way as wildtype *ec*SerB. After verifying the purity of the protein preparations by SDS-PAGE (Figure S 21), the HolPase activity of the mutant proteins was tested in a coupled photometric assay (8.5.5). Intriguingly, two of the purified mutants, designated *ec*SerB mut1 and *ec*SerB mut2 showed HolPase activity at a level which allowed for steady-state kinetic experiments. For HolP concentrations up to 300 µM, no saturation could be observed for the reaction rate. Significantly higher substrate concentrations could not be tested because a slight phosphate contamination of the purchased HolP led to a very high starting absorption for HolP concentrations higher than 300 µM. Therefore, the k_{cat}/K_M value was determined from the initial slope of the reaction rate plotted against substrate concentration and was calculated to be 7.8 $M^{-1}s^{-1}$ for *ec*SerB mut1 and 1.7 $M^{-1}s^{-1}$ for *ec*SerB mut2 (Figure 5.2 D).

An analysis of the amino acids at the randomized positions in *ec*SerB mut1 and *ec*SerB mut2 unveiled that the glutamate at position 125 was replaced by a smaller, polar residue in both mutants, namely by a serine in mut1 and by an aspartate in mut2. Interestingly, an aspartate was also found at the corresponding position in *pa*HisN which was furthermore identified as a fingerprint residue for HolPase function (3.2.3). This finding supports the previously formulated hypothesis, that the minor change from glutamate to aspartate might be critical at this position, but it also shows that serine is compatible with a low HolPase activity.

At position 127 which is occupied by an isoleucine in *ec*SerB, both mutants contained a tryptophan which is surprising, given that in *pa*HisN, this position is occupied by the second aspartate from the DxD motif. However, none of the sequenced *ec*SerB mutants featured an aspartate at this position, instead tryptophan or other mostly hydrophobic residues were found (Table S 5), indicating, that the binding mechanism in the mutant enzymes was different from the on in *pa*HisN. The possible explanation for the absence of aspartate in any of the mutants could be some epistatic effect which may influence the effect of a mutation at this position.

Position 158 which is occupied by a leucine *ec*SerB, is occupied by a cysteine in *ec*SerB mut2. However, the leucine was retained in *ec*SerB mut1 together with most of the other mutants (Table S 5), indicating, that this residue was not relevant for the HolPase function of *ec*SerB mut1 and mut2.

At position 161 there is an arginine in *ec*SerB which is strongly conserved in other PSPases (Figure 5.2 C). In the two mutants, this amino acid was replaced by an isoleucine or valine, respectively, both of which were frequently found in other mutants that were able to rescue the Δ*holPase* deletion. In the crystal structure of *mj*SerB, the side chain of this arginine is very close to the carboxylate group of PSer, indicating its importance for substrate binding. Exchanging this arginine by a smaller, more hydrophobic

residue makes sense, given that histidinol phosphate is larger than phosphoserine and given that the amino acid at this position most likely is in close contact to the imidazole ring. In *pa*HisN, the corresponding position is occupied by a threonine, which is in line with the argument, that a smaller residue is favoring the binding of histidinol phosphate.

In summary, establishing a HolPase function on the scaffold of *ec*SerB was achieved by one round of directed evolution, showcasing that a trajectory which starts at a PSPase and converts it to a HolPase is indeed possible and that only a few mutations are required to establish a starting activity high enough for continuous measurement and high enough to enable growth of a Δ*holPase* strain when overexpressed. Interestingly, only one mutation corresponded to the amino acid from *pa*HisN. The limited catalytic efficiency of the mutants however prevents any speculations on whether a different ensemble of residues would be compatible with high HolPase activity on this scaffold.

5.2.2 Improvement of the promiscuous PSPase activity of *ec*HisB-N

Previous work has shown that *ec*HisB-N possesses a low promiscuous side activity for PSer.[137] In another work, Yip et al. were able to increase this side activity by a factor of 23.7 in the triple mutant L41Q L96R T144I.[138] The numbering used by Yip et al. is shifted relative to the one which was used by Rangarajan et al. who previously solved the crystal structure of *ec*HisB-N.[59] In this work, the numbering will be used in accordance with Rangarajan et al. which means that the mutations introduced by Yip et al. concern the positions 42, 97, and 145. Interestingly, the *ec*HisB-N isoform that was crystallized contains an asparagine at position 145 instead of a threonine[59], indicating that there might be some sequence variability at this position in the protein variants of different *E. coli* strains. In this study, the *ec*HisB-N enzyme of the *E. coli* strain BW27786 was used which contains a threonine at position 145 (Table S 1).

To verify the reported results, first the wildtype *ec*HisB-N should be tested for its ability to rescue a Δ*serB* knock-out and for its promiscuous PSPase activity *in vitro*. For the complementation experiment, a corresponding plasmid with a constitutively active promotor and a Δ*serB E. coli* strain were taken from previous work by Dr. Bettina Rohweder.[140] The knock-out strain was transformed with the *hisB-N* encoding plasmid and streaked on M9 selective agar plates. After an incubation of 2-3 days at room temperature, colonies became visible (Figure S 28), confirming the previously reported observations.[137] To increase the time window for the subsequent selection of improved variants, a sequence coding for an SsrA degradation tag was introduced at the 3'-end of the *hisB-N* gene which should decrease the intracellular protein level.[142] An ensuing complementation experiment with this new construct showed that the colony size was significantly reduced compared to the construct without degradation tag (Figure S 28). Next, the PSPase activity of *ec*HisB-N should be tested *in vitro*. To this end, a Δ*serB*Δ*hisB* strain was transformed with an expression plasmid encoding for *ec*HisB-N which was also obtained from previous work by Dr. Bettina Rohweder.[140] The corresponding gene was expressed (8.4.2) and the target protein was purified from the crude extract (8.4.3) by affinity chromatography (8.4.4) followed by size exclusion chromatography (8.4.5). Subsequent SDS-PAGE (8.5.2) confirmed the purity of *ec*HisB-N (Figure S 1) and steady-state kinetic experiments (8.5.5) with PSer verified the PSPase activity (Figure S 27). For PSer concentrations up to 4 mM, no substrate saturation could be achieved which indicated a poor affinity of *ec*HisB-N for PSer. Therefore, the values for k_{cat} and K_M could not be determined independently and instead the k_{cat}/K_M value was calculated from a linear fit of the data which gave a value of 0.32 $M^{-1}s^{-1}$.

In the next step, the triple mutant should be generated and tested for its PSPase activity. To this end, the three mutations L42Q L97R T145I were introduced by side directed mutagenesis and the resulting plasmid was used to transform a Δ*serB*Δ*hisB E. coli* strain. The target protein was then produced in the same way as the wild-type protein and the success of the purification was verified by SDS-PAGE (8.5.2, Figure S 26). Then, steady-state kinetic experiments with PSer were performed. Like for the wildtype, no substrate saturation could be achieved, and a calculation of k_{cat}/K_M yielded a value of 3.3 $M^{-1}s^{-1}$ which corresponds to a 11-fold increase in the catalytic efficiency compared to the wildtype *ec*HisB-N and confirms the beneficial effect of the three mutations on the PSPase activity.[138] According to the data provided by Yip et al., the enhancement of the catalytic efficiency however was 26-fold which is approximately 2-fold higher than what was measured in this work. The authors also reported K_M values of 649 µM and 360 µM for the wildtype and the triple mutant, respectively, indicating that substrate saturation was observed for PSer concentrations one order of magnitude smaller than assayed here.[138] A substrate saturation of *ec*HisB-N by PSer is surprising due to two reasons. First, PSer and HolP differ significantly in their electrostatic properties and size, which makes a relatively high affinity towards PSer improbable. Secondly, in a previous work on the substrate specificities of HAD enzymes by Kuznetsova et al., no PSPase activity was detected for *ec*HisB-N which could be easily explained by a high K_M value and a catalytic efficiency of 0.3 $M^{-1}s^{-1}$ as measured in his study. A K_M value of 360 µM and a k_{cat}/K_M value of 180 $M^{-1}s^{-1}$ for the triple mutant as reported by Yip et al. however pose the question, why this activity was missed in the study by Kuznetsova et al.[106] In general, differences in the apparent K_M value can be caused by an erroneous substrate concentration, which however in this case would mean that there was a big discrepancy between the actual and the assumed substrate concentration. Another factor that might explain the different results concerns the assays which were used to determine the amount of formed product. Yip et al. used a discontinuous assay which allows for the detection of free phosphate by the formation of a complex with malachite green dye.[138] In contrast, in this work, an enzyme coupled assay was used which allowed for the continuous measurement of phosphate release.[104] Using this continuous assay, we noted, that both the commercially available PSer and even more so HolP were contaminated with free phosphate. To account for this, substrate and auxiliary enzymes and substrates were preincubated until a constant absorption was established before enzyme was added to initiate the reaction. Taken together, the values obtained in this study appear to be in line with previous results by Kuznetsova et al.[106] and the use of a continuous instead of a discontinuous assay in combination with a potential phosphate contamination of the substrate might explain the different results compared to Yip et al. regarding the PSPase activity of *ec*HisB-N wildtype and the triple mutant.[138]

The fact that no substrate saturation could be achieved indicated sub-optimal substrate binding and a low affinity of *ec*HisB-N for PSer. Moreover, mapping the three mutations L42Q L97R T145I onto the crystal structure showed that none of the residues was part of the residues of the first shell around the active site and are in immediate contact with the substrate (Figure 5.3 A). This analysis suggested that the substrate binding could be optimized which might result in an improved K_M value and therefore also an enhanced catalytic efficiency. For this objective, several active site residues should be selected for a first round of randomization.

.

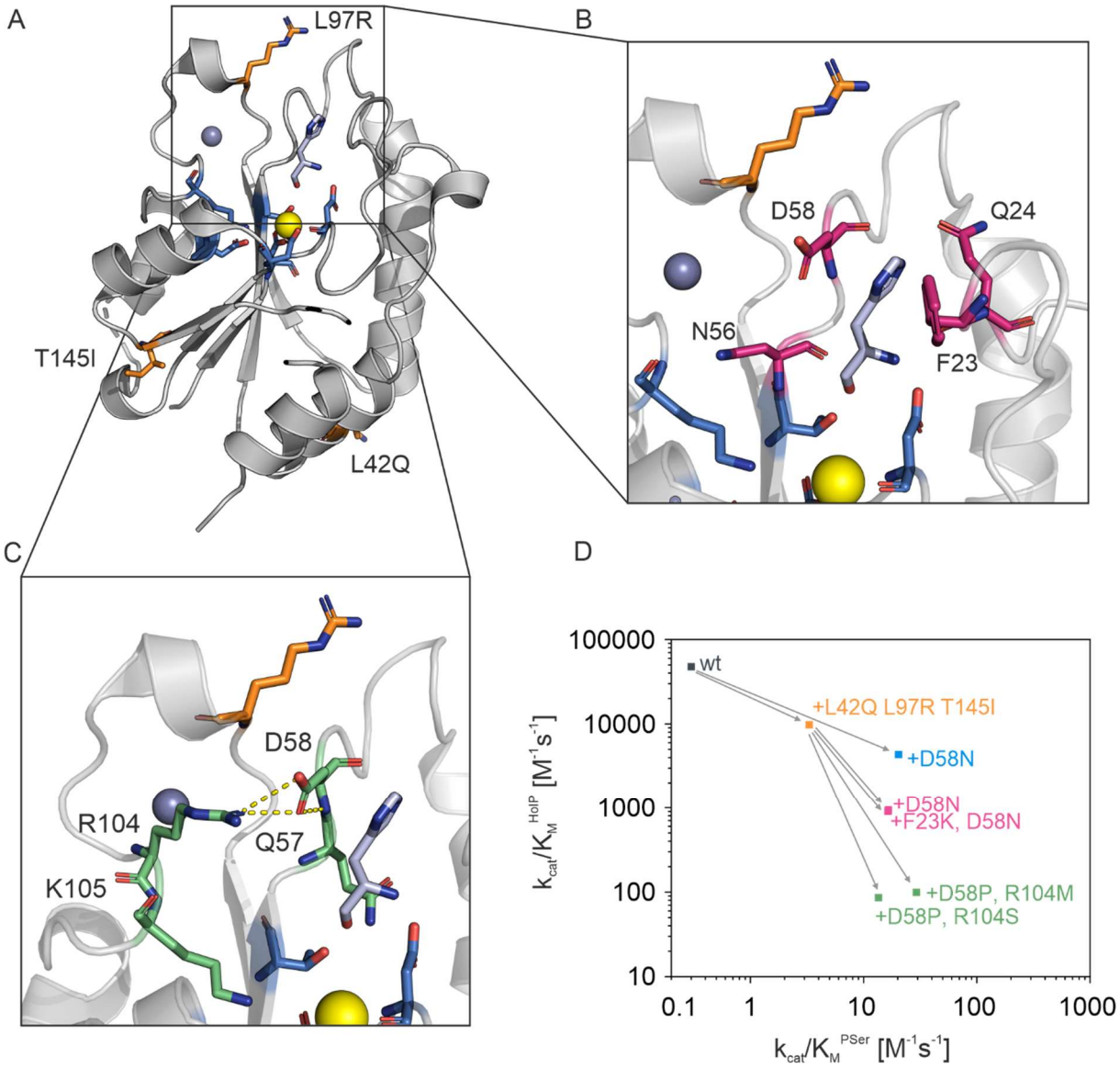

Figure 5.3: Improving the promiscuous PSPase activity of *ec*HisB-N.

(A) Crystal structure of *ec*HisB-N[59] (PDB-ID: 2FPU) with bound product Hol (light blue sticks). Residues of the catalytic machinery are shown in dark blue, Mg^{2+} and Zn^{2+} are represented as yellow and grey sphere, respectively, and HolPase enhancing mutations identified by Yip et al.[138] were introduced by the mutagenesis tool of PyMOL and are shown as orange sticks. (B) Zoomed in view of the active site of *ec*HisB-N with residues selected for the first round of randomization depicted as purple sticks. (C) Zoomed-in view of the active site of *ec*HisB-N with residues selected for the second round of randomization depicted as green sticks. (D) Plot of the catalytic efficiency of the HolPase activity against the catalytic efficiency of the PSPase activity of selected *ec*HisB-N mutants at 25°C. The arrows indicate from which variant each mutant was derived and additional mutations are indicated. Color code: Orange: variant from Yip et al.[138], magenta: first round of randomization, blue: rational mutation based on the first library, green: second round of randomization.

Binding PSer instead of HolP translates to binding a carboxylate group instead of an imidazole ring. Therefore, residues in the immediate vicinity of the imidazole moiety were selected for randomization, based on a crystal structure with bound product (Figure 5.3 B).[59] Specifically, the four residues F23, Q24, N56, and D58 were selected and randomized in a concerted manner in order to sample the available sequence space of the active site and account for possible epistatic effects. Starting point for the randomization was the triple mutant gene which was put under the control of a constitutively active

promotor and featured a sequence coding for the SsrA degradation tag at the 3'-end. The randomization of this construct was conducted with NNK primers (8.3.4.6). Afterwards, the quality of the resulting library was assessed by Sanger sequencing of 18 individual clones (8.3.1) which revealed that at least one codon for every amino acid except for glutamine was found (Figure S 29). The amino acid distribution was furthermore similar to the expected one for a randomization with NNK primers and based on the sequencing data, the library size was calculated to be approximately $2.5 \cdot 10^7$.

To identify enzyme variants with improved PSPase activity, Δ*serB* cells were transformed with the library and selection experiments were performed on M9 minimalagar plates at room temperature (8.3.4.6). Afterwards, plasmids were isolated from fast growing colonies and, to exclude false positives, the isolated plasmids were used to individually transform Δ*serB* cells and repeat the selection experiments (Figure S 30). After this second round of selection, 16 gene variants that promoted increased growth rates were sequenced and subcloned in expression plasmids. Then, these genes were expressed, and the corresponding proteins were purified in the same manner as the wildtype *ec*HisB-N. Afterwards, the PSPase activity of these 16 proteins was measured in vitro (Figure S 31). These experiments demonstrated that 15/16 proteins indeed possessed an increased PSPase activity with a up to 6-fold elevated k_{cat}/K_M value compared to the triple mutant, confirming that the selection system was indeed suited to find enzymes with enhanced PSPase activity. Analyzing the amino acids at the randomized positions revealed that the mutants differed in the amino acids at position 23 and 24, which suggested, that this position was of minor importance for the PSPase activity. At position 56 all mutants showed a reversion to the wildtype amino acid asparagine, suggesting that no other amino acid was tolerated at this position. At position 58 however, 8 mutants, among them the 5 mutants with the highest PSPase activity, exhibited the substitution D58N (Figure S 31). This strongly indicated that this exchange was crucial for the increased PSPase activity. To corroborate this conclusion, the D58N mutation was introduced into the wildtype enzyme which resulted in an enhancement of the HolPase activity by more than 60-fold from 0.32 $M^{-1}s^{-1}$ for the wildtype to 20.6 $M^{-1}s^{-1}$ for the D58N single mutant (Figure S 26, Figure S 32), corroborating the functional importance of the amino acid at this position. In search for further potential to improve the HolPase function, we therefore focused our attention on this residue. The inspection of the crystal structure of *ec*HisB-N wildtype unveiled that D58 formed two hydrogen bonds with R104 (Figure 5.3 C), indicating that the nature of the amino acid at position 104 might restrict the amino acid which can be incorporated at position 58. Therefore, a second NNK library based on the triple mutant was constructed which included the positions 104 and 58 and the neighboring residues Q57 and K105, respectively (Figure 5.3 C). Again, a library size of about $2 \cdot 10^7$ was achieved and in the following selection experiments on M9 minimal agar, 51 fast growing colonies were identified and sequenced (8.3.4.7). From these 51 colonies, 40 possessed unambiguous reads at all 4 randomized sites (Table S 6). Interestingly, the residues Q57 and K105 were retained in all enzyme variants whereas at positions 58 and 104 a greater variability was observed. The most frequently observed combinations of mutations were D58P and R104M, which was found 7 times, and D58P and R104S, which was found 4 times. These two enzyme variants were therefore selected for *in vitro* testing and produced in the same manner as the wildtype enzyme (Figure S 26). Steady-state kinetic experiments showed a k_{cat}/K_M value of 13 $M^{-1}s^{-1}$ for the variant which contained the D58P and R104S mutations and of 29.2 $M^{-1}s^{-1}$ for the mutant which contained the D58P and R104M mutations (Figure S 34). The activity of the triple mutant with the additional D58P R104M is thus almost two orders of magnitude higher than the PSPase activity of the wildtype enzyme. This result confirmed the assumption that due to the close spatial proximity of the residues at position 58 and R104 the tolerated amino acids at position 58 were defined by the amino acid at position 104. It also showed, that concerted mutation of R104 and D58 could further improve the PSPase activity. Interestingly, substrate saturation with PSer could not be observed for any of the assayed variants, neither from the first library, nor from the second library or the D58N single mutant. Thus, it stays unclear whether the improved catalytic efficiencies are due an improved k_{cat} or K_M value.

The discussed improvements of the PSPase activity all required the mutation of D58 either to asparagine or proline. According to the crystal structure, the amino acid at position 58 is in very close proximity to the imidazole ring of the bound product histidinol which raises the question of how the introduced mutations influence the native HolPase activity. This question was addressed by steady-state kinetic experiments for the two best variants of each library, the triple mutant from Yip et al.[138] and the D58N single mutant (Figure S 32-34).

These experiments showed, that for all mutants, the HolPase activity was significantly reduced compared to the *ec*HisB-N wildtype. In the triple mutant L42Q L97R T145I, the 10-fold increase in PSPase activity was accompanied by a 6-fold decrease in HolPase activity from 58,000 $M^{-1}s^{-1}$ to 9800 $M^{-1}s^{-1}$. The additional D58N mutation increased the PSPase activity by almost one order of magnitude to approximately 20 $M^{-1}s^{-1}$, but at the same time, the HolPase was reduced by the same factor to around 1000 $M^{-1}s^{-1}$. Interestingly, the single mutant D58N showed the same level of PSPase activity as the mutants with the additional L42Q L97R T145I substitution, but the HolPase activity of the single mutant was roughly 4-fold higher. This indicated that there is no positive epistasis regarding the PSPase activity between D58N and the three mutations L42Q L97R T145I, whereas the negative effects on the HolPase activity caused by the D58N mutation and the triple mutation L42Q L97R T145I are additive. The D58P mutation let to the highest observed PSPase activity of approximately 30 $M^{-1}s^{-1}$ but it also greatly reduced the HolPase activity to around 100 $M^{-1}s^{-1}$. This means that for all mutants a trade-off between the two functions was observed and greater improvements of the PSPase activity came at the cost of a greater reduction in HolPase activity (Figure 5.3). This observation indicates that the HolPase activity and PSPase activity require different adaptations of the active site which are not compatible to one another. It seems therefore unlikely, that an enzyme can be designed or evolved which is both characterized by a high HolPase activity and a high PSPase activity.

The marked effect of the single mutation D58N on the other hand showcases the high evolvability of the HAD superfamily, where a conserved catalytic machinery together with just one adaptive point mutation can lead to a significantly altered reaction profile.

It is furthermore of note, that only one out of five randomized positions of the first shell conferred a strong effect on the PSPase activity, which prompts the question of how further improvements of the PSPase activity might be achieved. One possibility would be a modification of the flap structure next to the randomized residues F23 and Q24. In the native PSPase SerB, the four-helical cap is inserted into this flap structure, whereas in GmhB, the flap structure is significantly smaller than in HisB-N. These examples represent a general trend according to which the size of the substrate is anti-correlated with the size of the flap or the cap that occludes the active site.[51] Thus, inserting additional residues into this region seems to be the most promising approach as it could (i) provide additional interaction sites and (ii) occlude the active site more efficiently from the bulk solvent in presence of such a small substrate as PSer.

6 Comprehensive Conclusion

6.1 The great diversity and phylogenetic scattering of HolPases

In this work, the HolPases of the two organisms *P. aeruginosa* and *N. maritmius* could be clearly identified. These two enzymes both belong to the HAD superfamily yet differ in their structure from each other and the previously analyzed of HAD superfamily HolPase of *E. coli* (Figure 3.1, Figure 4.2). These differences between the three types of HolPases are also reflected in their pairwise sequence identity, which is below 20 % in all three cases (Table 6.1).

Table 6.1: Pairwise sequence alignments for the different HolPases from the HAD superfamily.

Listed are the percentage of identical residues, the percentage of similar residues, and the percentage of gaps with respect to the length of the pairwise alignment.

Enzyme 1	Enzyme 2	% sequence identity	% sequence similarity	% gaps
*ec*HisB-N	*pa*HisN	19.4	29.1	44.9
*ec*HisB-N	*nm*HisN	13.4	20.2	59.9
*pa*HisN	*nm*HisN	13.1	21.7	54.7

These differences in sequence and fold provide a strong argument against a shared evolution of *ec*HisB-N, *pa*HisN, and *nm*HisN. Instead, the results from chapter 2 strongly suggest, that *ec*HisB-N was derived from an ancient βGmhB. In the case of *pa*HisN, the narrow phylogenetic distribution, the structural similarities with *mj*SerB and a promiscuous PSPase activity suggest that it might be derived from an ancient PSPase. The archaeal *nm*HisN finally seems to represent a class of HolPases which is widely distributed within the archaeal kingdom, which implies that this type of HolPase might be older than the *ec*HisB-N-type and the *pa*HisN-type HolPases.

Taken together, these results indicated a convergent evolution of the different HolPases which is equivalent to the notion, that the HolPase activity was “invented” on several occasions independently. This also means, the HolPase activity was probably established three times on the scaffold of the HAD superfamily. This is surprising given that there are about 1400 protein superfamilies [47] and suggests, that the HAD superfamily provides a well-suited backbone on which this enzymatic function can be easily established. This conclusion is supported by the directed evolution experiments on *ec*SerB which showed that only three mutations were required to establish a significant degree of HolPase activity and rescue a Δ*holPase* strain from starvation. This reasoning is furthermore in line with the previously reported high substrate promiscuity within the HAD superfamily, which provides a reservoir for the evolution of new enzymatic functions.[28, 106]

Taken together, there are now six different types of HolPases known, represented by the HolPases from *L. lactis* (PHP superfamily)[91], from *M. truncatula* (IMP)[93], *M. tuberculosis* (IMP)[94], *E. coli* (HAD)[59], *P. aeruginosa* (HAD), and *N. maritimus* (HAD). In section 3.2.5 the phylogenetic distribution of the previously known types of HolPases was analyzed. These results were combined with the findings on the homologues of *pa*HisN and *nm*HisN from the respective sequence similarity networks and used to map the different types of HolPases onto an abstracted representation of the tree of life based on a recently published phylogenetic tree by Hug et al. (Figure 6.1).[143]

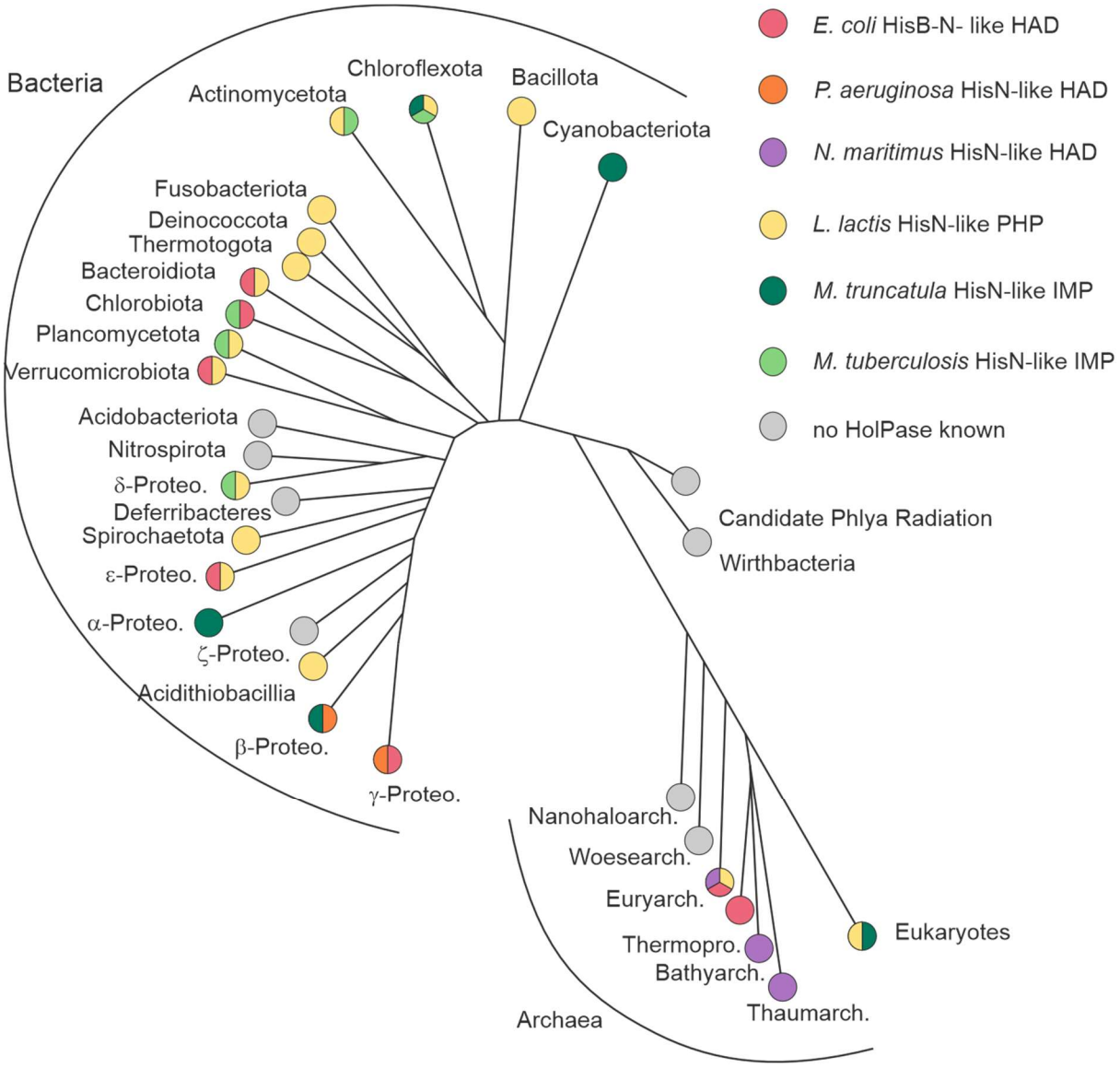

Figure 6.1: Mapping of the different HolPases on an abstracted tree of life.

Shown is an abstracted representation of the tree of life based on comprehensive phylogenetic tree by Hug et al. [143] For each phylogenetic group, the type of observed HolPase is indicated, if the respective type is found in at least 5 % of the organisms of this phylogenetic group, according to the data from section 3.2.5, or if the occurrence of this type is strongly suggested based on the analysis in section 3.2.4 and 4.2.4.

This representation reiterates the observation that there is no clear pattern in the phylogenetic distribution of the different types of HolPases, and instead the different types of HolPases seem to be scattered across the tree of life, which leads to two conclusions:

1. Most HolPases were probably recruited late. Since otherwise, if a HolPase was integrated early into the pathway, for example in the last common ancestor of all bacteria one would assume to find the same type of HolPase in the great majority of all extant bacterial species.
2. There was probably extensive horizontal gene transfer of HolPase encoding genes, because this is the most plausible explanation for the occurrence of HolPases of the same type in unrelated organisms from different kingdoms.

Based on these conclusions, an integrative model for the evolution of the HolPase function is proposed.

6.2 A model for the early evolution of the HolPase function

Noda-Garcia et al. argued, that during the assembly of a new metabolic pathway when the overall flux is low, non-enzymatic reactions probably play an important role.[22] One can therefore think of an early pathway as a series of (slow promiscuous) enzymatic and non-enzymatic steps. The highest selective pressure within such a pathway is on the slowest reaction step which limits the rate at which the product is formed.[22] If this reaction is already catalyzed by an enzyme the selective pressure favors the improvement of the catalytic properties of this enzyme. But if the rate-limiting reaction proceeds non-enzymatically the integration of an enzyme into the pathway which accelerates this particular reaction step yields the highest advantage.[22] Histidine biosynthesis is generally characterized by a high energy cost of 41 molecules of ATP for each molecule of histidine[144] and features a number of complex reactions like the Amadori rearrangement catalyzed by HisA.[72] Moreover, the abiotical synthesis of histidine under supposedly prebiotic conditions reported by Shen et al. required several reaction steps with temperatures of up to 105°C for up to 24 h and the yield of the final product was only 3.5 %[145] which is why the prebiotic existence of histidine is still questioned.[71] This leads to the conclusions that without enzymatic catalysis, most reactions steps within histidine biosyntheisis likely proceed very slowly and some might not occur at a relevant rate at all. Therefore, the selective pressure probably forced an early incorporation of enzymes into the reaction sequence to catalyze these reactions. This assumption is supported by the homology of most of the enzymes across different species which indicates common ancestry.[70, 82] The six different types of HolPases that are known until now form an exception to this uniformity, indicating that the HolPases were recruited late to the pathway. Interestingly, experiments by Lukas Drexler revealed a tendency of HolP to decompose to Hol and free phosphate at elevated temperatures and after 20 h at 85°C 30% product had formed compared to the catalyzed reaction.[146] Taken together, this indicates that the reaction rate of the non-enzymatic hydrolysis of HolP was probably higher than the reaction rate of any other reaction of the histidine biosynthesis in absence of enzymatic catalysis. The associated selective pressure on the incorporation of an enzyme to catalyze this reaction was therefore probably lower than for all the other reaction steps.

It is therefore assumed that at the first stage (stage I) of the assembly of histidine biosynthesis, when the set of available enzymes was likely very limited, the hydrolysis of HolP was not catalyzed by an enzyme but instead this reaction step was abiotic (Figure 6.2, upper panel). There are several factors which may have aided in the spontaneous hydrolysis of HolP, like the presence of Mg^{2+} or Mn^{2+} on which the enzyme catalyzed HolPase reactions also depend[59, 147, 94, 91, 48], high temperatures, or high concentrations of either OH^- or H^+, which may have promoted the hydrolysis according to the classical acidic or basic ester hydrolysis.

Intriguingly, Wang et al. reported, that the deletion of the HolPase encoding gene PA0335 of *P. aeruginosa* resulted in an incomplete histidine auxotrophy.[116] One possible explanation for this finding might be a promiscuous side activity for HolP of other phosphatases (Figure 6.2, middle panel). Based on this assumption, it is further assumed that in the stage II of the evolution of the HolPase function, an increasing genome size and a rising number of available enzymes enhanced the probability for one or more phosphatases with promiscuous HolPase activity. At the same time, it seems unlikely that LUCA possessed a dedicated HolPase because - apart from extensive gene loss - homologues of this primordial HolPase would be found in most extant histidine-synthesizing organisms.

It is hence further assumed, that promiscuous HolPases from stage II were recruited in progenitor species of extant phylogenetic groups during a further stage III, after the three kingdoms of life had diverged

(Figure 6.2, lower panel). As discussed previously, this third stage was probably characterized by extensive horizontal gene transfer of the different HolPase encoding genes. The investigations on the evolution of *ec*HisB-N further indicate that in some cases phosphatases with a different primary function evolved into HolPases after they were horizontally transferred into a new host.

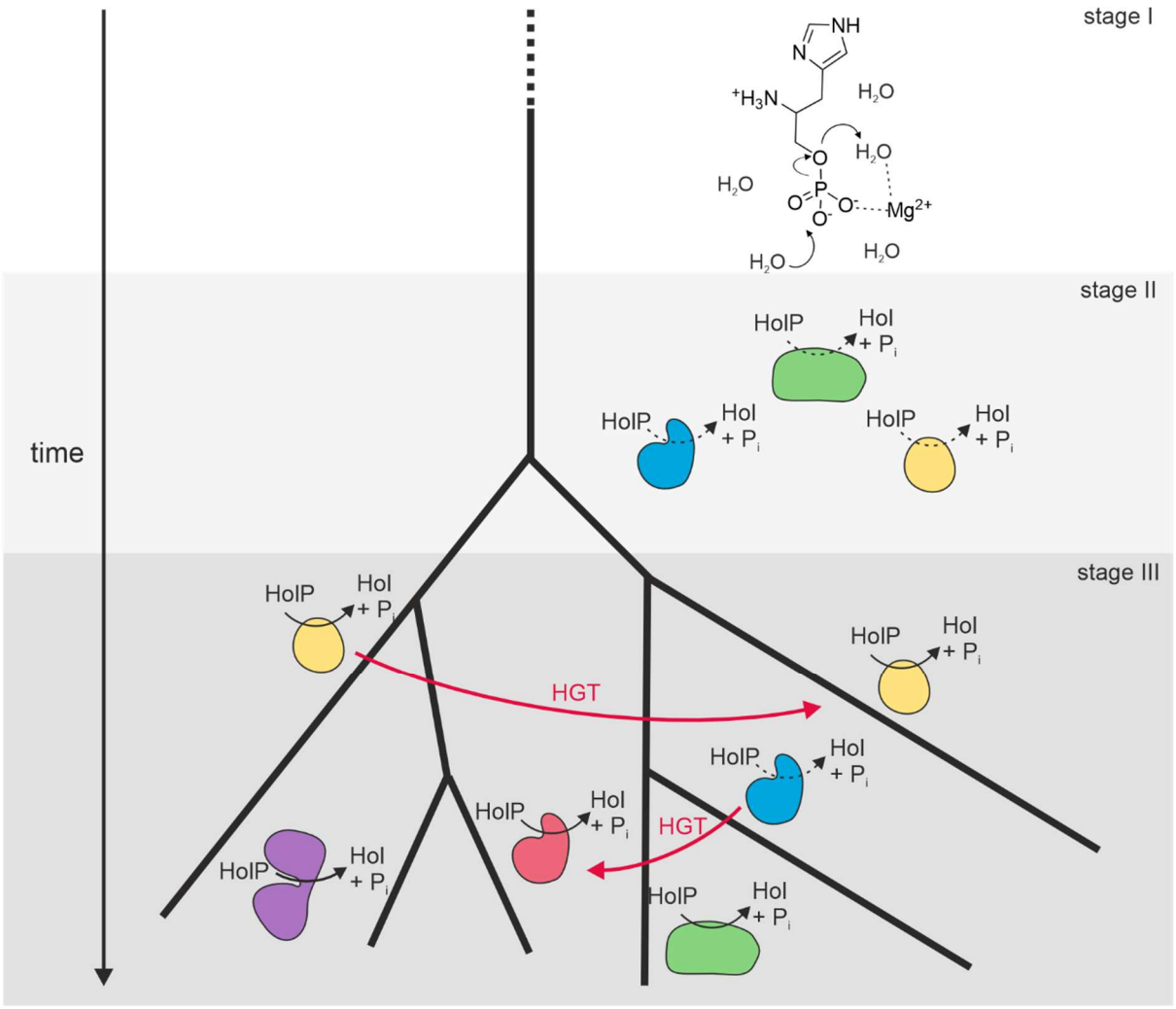

Figure 6.2: Evolutionary model for the early evolution of the HolPase function.

It is assumed, that in stage I, when the histidine biosynthesis was assembled, spontaneous hydrolysis of HolP, e.g., aided by Mg^{2+}, contributed significantly to the low flux within the pathway. In stage II, the growing number of proteins increased the probability for the existence of phosphatases with low promiscuous HolPase function (dashed reaction arrows). In a third stage, different HolPases were probably recruited by different organisms. This stage was likely also characterized by horizontal gene transfer (HGT) of genes coding for specific HolPases (yellow enzyme, solid reaction arrow) or genes coding for enzymes with latent HolPase activity (blue enzyme, dashed reaction arrow) which only evolved into specific HolPases (red enzyme) after horizontal gene transfer.

7 Materials

7.1 Devices

Absorption spectrophotometers	
JASCO V-650	JASCO GmhB, Groß-Umstadt
JASCO V-750	JASCO GmhB, Groß-Umstadt
Microplate reader Infinite M200 Pro	TECAN, Austria
NanoDrop One	THERMO FISHER SCIENTIFIC, Waltham, USA
CD spectrometer	
JASCO J-815	JASCO GmhB, Groß-Umstadt
Centrifuges and rotors	
Avanti J–26 S XP	BECKMAN COULTER, Krefeld
JLA–8.1000 rotor	BECKMAN COULTER, Krefeld
Centrifuge 5810	Centrifuge 5810 EPPENDORF, Hamburg
A–4–81 rotor	EPPENDORF, Hamburg
Sorvall RC 2B	DU PONT, Bad Homburg
SS34 rotor	DU PONT, Bad Homburg
Centrifuge 5810 R	EPPENDORF, Hamburg
Centrifuge 5415 R	EPPENDORF, Hamburg
Centrifuge 5418 R	EPPENDORF, Hamburg
Centrifuge 5301 R	EPPENDORF, Hamburg
FPLC columns	
HisTrap FF crude 5 mL	GE HEALTHCARE, München
HisTrap excel 5 mL	CYTIVA, Freiburg im Breisbau
Superdex S75 HiLoad 26/600	GE HEALTHCARE, München
MonoQ HR 16/10	GE HEALTHCARE, München
Static light scattering	
Viscotec TDA 305	Malvern Panalytical, Kassel

FPLC systems

Amersham FPLC 900 system	AMERSHAM, Little Chalfont, UK
ÄKTA prime/prime plus	CYTIVA, Freiburg im Breisgau
ÄKTA purifier 10	CYTIVA, Freiburg im Breisgau
ÄKTA start	CYTIVA, Freiburg im Breisgau
ÄKTA pure	CYTIVA, Freiburg im Breisgau
ÄKTA micro	CYTIVA, Freiburg im Breisgau
ALIAS™ autosampler	CYTIVA, Freiburg im Breisgau

Gel electrophoresis

Gel electrophoresis chamber	HOEFER PHARMACIA BIOTECH, USA
Power supply EPS 301 AMERSHAM, Little Chalfont, UK	AMERSHAM, Little Chalfont, UK
Multi gel caster assembling gel apparatus	GE HEALTHCARE, Chicago, USA

Heating blocks

HBP–2 131	HEAP LABOR CONSULT, Bovenden

Incubators

Binder (9010-0086)	BINDER, Tuttlingen

pH-meter

766 Calimatic	KNICK, Berlin

Shakers

Ceromat BS-1	B. BRAUN, Melsungen
Certomat H	B. BRAUN, Melsungen
Multitron	Infors HT, BOTTMINGEN, Switzerland
Vortex Genie 2	SCIENTIFIC IND., Bohemia, USA

Sonifier

Branson 450 Digital Sonifier	HEINEMANN, Schwäbisch Gmünd

Thermocycler	
Mastercycler personal	EPPENDORF, Hamburg
Mastercycler gradient	EPPENDORF, Hamburg
Ultrapure water systems	
MilliQ Q–POD	MILLIPORE GmbH, Schwalbach
Biopak Polisher filter	MILLIPORE GmbH, Schwalbach
Millipak filter	MILLIPORE GMBH, Schwalbach
UV-imagers	
GelDoc Go System	BIO-RAD LABORATORIES, München
GelDoc–It Imaging System	UVP, Upland, USA
Vacuum Pump	
Laboport® N820 FTP	VWR, Leuven, Belgium

7.2 Consumables

Centrifugal Filter Device	
Amicon Ultra-15 (MWCO: 10 kDa)	Millipore, Bedford, USA
Buffer exchange columns	
NAP-5, NAP-10	CYTIVA, Freiburg im Breisgau
Centrifuge tubes 15 mL/50 mL	SARSTEDT, Nürnbrecht
Cuvettes	
UV Cuvettes	SARSTEDT, Nürnbrecht
Electroporation cuvettes (2 mm)	MOLECULAR BIOPRODUCTS, Sand Diego, USA
Quartz cuvettes 1 mm	HELLMA, Jena
Dialysis tubes Visking type 27/32 (14 kDA)	ROTH, Karlsruhe
Membrane filters ME24	WHATMAN, Dassel
(47 mm, 0.2 µm pore size)	

Microtiter plates	
UV-STAR® 96-well	GREINER Bio-One, Frickenhausen
CELLSTAR® 96-well	
Micro tubes 1.5 mL/2.0 mL	SARSTEDT, Nürnbrecht
Parafilm „M“ Laboratory film	PECHINEY, Menasha, USA
PCR tubes 0.2 mL	SARSTEDT, Nürnbrecht
	PEQLAB, Erlangen
Petri dish	GREINER Bio-One, Frickenhausen
Pipette tips 10 µL, 200 µL, 1000 µL	SARSTEDT, Nürnbrecht
Syringe filters Filtropur 0.2 µM/ 0.45 µM pores	SARSTEDT, Nürnbrecht

7.3 Chemicals

All chemicals which were used in this work were at least graded p.a. and were purchased from the following companies:

APPLICHEM, Darmstadt

BIO-RAD LABORAsTORIES, München

BIOZYM, Hessisch Oldendorf

BOEHIRNGER MANNHEIM, Mannheim

GE HEALTHCARE, München

GIBCO BRL, Eggstein

FLUKA, Buchs, Switzerland

MERCK, Darmstadt

MP BIOCHEMICALS, Illkirch, France

NATIONAL DIAGNOSTICS, Simerville, NJ, USA

RIEDEL-DE HAEN, Seelze

ROCHE DIAGNOSTICS, Mannheim

ROTH, Karlsruhe

SERVA, Heidelberg

SIGMA-ALDRICH, St. Louis, USA

THERMO FISHER SCIENTIFIC, Waltham, USA

VWR INTERNATIONAL, Radnor, USA

7.4 Bacterial strains

name	genotype	origin	description
E. coli BL21 Gold (DE3)	*B F⁻ ompT hsdS(r_B^- m_B^-) dcm⁺ Tet^r gal λ(DE3) endA Hte*	Agilent Technologies, Santa Clara, USA	Expression strain with increased transformation efficiency and T7 DNA polymerase
E. coli NEB Turbo	F' *proA⁺B⁺ $lacI_q$ Δ(lacZ)M15/fhuA2 Δ(lac-proAB) glnV galK16 galE15 R(zgb-210::Tn10)* Tets *endA1 thi-1 Δ(hsdS-mcrB)5*	New England Biolabs, Ipswich, USA	Cloning strain with T1-phage resistance
E. coli BW25113 (K-12 derivative)	*rrnB3 ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1*	Datsenko and Wanner (2000)[148]	Strain which is accessible for phage transduction
E. coli BW25113 *ΔserB ΔhisB*	*rrnB3 ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1 ΔserB:cam^R ΔhisB:kana^R*	Bettina Rohweder[140]	*ΔserB ΔhisB* knock-out strain used for gene expression
E. coli BW25113 *ΔserB*	*rrnB3 ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1 ΔserB:cam^R*	Bettina Rohweder[140]	*ΔserB* knock-out strain used for complementation experiments
E. coli BW25113 *ΔhisB*	*rrnB3 ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1 ΔhisB:kana^R*	This work	*ΔhisB* knock-out strain used for complementation experiments

7.5 Media, buffers, and solutions

All buffers were filter-sterilized prior to storage. Buffers for SEC were additionally degassed.

7.5.1 Buffers for protein purification

Equilibration buffer for IMAC

50 mM	Tris/HCl (pH 7.5)
400 mM	NaCl
2 mM	$MgCl_2$
10 mM	Imidazole

Elution buffer for IMAC

50 mM	Tris/HCl (pH 7.5)
400 mM	NaCl
2 mM	$MgCl_2$
500 mM	Imidazole

Equilibration buffer for SEC

50 mM	Tris/HCl (pH 7.5)
400 mM	NaCl
2 mM	$MgCl_2$

7.5.2 Buffers for anion exchange chromatography

Equilibration buffer

50 mM NH_4HCO_3 (pH 9.0), the pH was adjusted by titration with NaOH

Elution buffer

1 M NH_4HCO_3 (pH 7.8), the pH was adjusted by titration with NaOH

7.5.3 Buffers and media for microbiological methods

Media were autoclaved for 20 min at 121°C and 2 bar for sterilization. Antibiotics were added after cooling when applicable.

LB medium and LB agar

0.5 % (w/v)	Yeast extract
1.0 % (w/v)	NaCl

1.0 % (w/v)	Tryptone
1.5 % (w/v)	Agar (only in the case of LB agar)

M9 minimal medium and M9 minimal agar

12.8 g/L	Na_2HPO_4 x 7 H_2O
3.0 g/L	KH_2PO_4
0.5 g/L	NaCl
1.0 g/L	NH_4Cl
2 mM	Mg_2SO_4
100 µM	$CaCl_2$
0.4 % (w/v)	Glucose
15 g/L	Agar (only in the case of M9 minimal agar)

SOB medium

0.5 % (w/v)	Yeast extract
0.05 % (w/v)	NaCl
2.0 % (w/v)	Tryptone
10 mM	$MgCl_2$ (after autoclaving, filter-sterilized)
10 mM	$MgSO_4$ (after autoclaving, filter-sterilized)
2.5 mM	KCl (after autoclaving, filter-sterilized)

Transformation buffer I (TFB I)

100 mM	KCl
50 mM	$MnCl_2$
30 mM	KOAc
10 mM	$CaCl_2$
15 %	Glycerol
pH	6

Transformation buffer II (TFB II)

10 mM	Tris
10 mM	KCl

75 mM	$CaCl_2$
15 %	Glycerol
pH	7

Saline for P1 transduction

145 mM	NaCl
50 mM	Trisodium citrate

Ampicillin stock solution

150 mg/mL ampicillin (sodium salt) dissolved in water, stored at -20 °C.

Chloramphenicol stock solution

30 mg/mL chloramphenicol dissolved in 10 % ethanol, stored at -20 °C.

Kanamycin stock solution

75 mg/mL kanamycin dissolved in water, stored at -20 °C.

Glucose (20 %)

20 % (w/v) glucose dissolved in water, stored at -20 °C.

IPTG stock solution

0.5 M IPTG dissolved in water, stored at -20 °C.

NaCl stock solution (0.9 %)

0.9 % (w/v) dissolved in water, stored at room temperature.

7.5.4 Buffers and solutions for molecular biological methods

TBE buffer

89 mM	Tris
89 mM	Boric acid
2.5 mM	EDTA

Agarose (0.8-1.5 %)

0.8-1.5% (w/v) dissolved in 0.5x TBE, boiled and stored at 60 °C.

Ethidium bromide solution

10 mg/mL ethidium bromide in water.

dNTPs

solution of dNTPs (2 mM of each A, C, G, and T)

7.5.5 Buffers and solutions for SDS-PAGE

Ammonium persulfate (APS) solution	10 % (w/v) APS solution in water, filter-sterilized, stored at -20 °C.
Coomassie staining solution	
Stacking gel buffer	0.4 % (w/v) SDS, 0.5 M Tris/HCl pH 6.8
Resolving gel buffer	0.4 % (w/v) SDS, 0.5 M Tris/HCl pH 8.8
Electrophoresis buffer	0.1 % (w/v) SDS, 0.025 M Tris, 0.2 M glycine, pH 8.5
SDS Sample buffer (5x)	5 % (w/v) SDS, 25 % (w/v) glycerol, 12.5 % (v/v) β-mercaptoethanol, 0.025 % (w/v) bromphenol blue, 1.25 M Tris/HCl pH 6.8

7.6 Kits, enzymes, and ready-made buffers

DNA extraction	
GeneJET™ Gel Extraction Kit	THERMO FISHER SCIENTIFIC, Waltham, USA
GeneJET™ Plasmid Miniprep Kit	THERMO FISHER SCIENTIFIC, Waltham, USA
Standards	
GeneRulerTM 1kb Plus DNA Ladder	THERMO FISHER SCHIENTIFIC, Waltham, USA
Protein standard (LMW)	THERMO FISHER SCIENTIFIC, Waltham, USA
Pierce™ Unstained Protein MW Marker	
Polymerases and buffers	
Q5® HF DNA polymerase (2000 U/mL)	NEW ENGLAND BIOLABS, Ipswich, USA
Q5® reaction buffer (5x)	NEW ENGLAND BIOLABS, Ipswich, USA
Phusion HF	NEW ENGLAND BIOLABS, Ipswich, USA
Phusion reaction buffer (5x)	NEW ENGLAND BIOLABS, Ipswich, USA
GC-enhancer	NEW ENGLAND BIOLABS, Ipswich, USA
Restriction enzymes, ligases and buffers	
*Bsa*I-HFv2 (20000 U/mL)	NEW ENGLAND BIOLABS, Ipswich, USA
*Bsa*I buffer (10x)	NEW ENGLAND BIOLABS, Ipswich, USA
*Dpn*I	NEW ENGLAND BIOLABS, Ipswich, USA
CutSmart Buffer (10x)	NEW ENGLAND BIOLABS, Ipswich, USA
T4-DNA-ligase (5000 U/mL)	THERMO FISHER SCIENTIFIC, Waltham, USA

T4-DNA-ligase buffer (10x)	THERMO FISHER SCIENTIFIC, Waltham, USA
Kinase and buffer	
T4-polynucleotide-kinase (PNK, 10000 U/mL)	NEW ENGLAND BIOLABS, Ipswich, USA
T4 PNK Reaction buffer (10x)	NEW ENGLAND BIOLABS, Ipswich, USA
Other enzymes	
Xanthine oxidase	SIGMA-ALDRICH, St. Louis, USA
Purine nucleoside phosphorylase	SIGMA-ALDRICH, St. Louis, USA
Other buffers	
Agarose gel loading dye (6x)	NEW ENGLAND BIOLABS, Ipswich, USA
Protogel™ 30 % (w/v) acrylamide, 0.8% (v/v) bisacrylamide	THERMO FISHER SCIENTIFIC, Waltham, USA

7.7 Plasmids

The plasmids pExp, pUR22, and pUR23 were obtained from previous work by Rohweder et al.[149]

Table 7.1: Plasmids used in this work.

name	resistance	Restriction site	Description
pExp	Amp	*Bsa*I	Plasmid with constitutively active promotor for complementation experiments
pExp_*ssrA*	Amp	*Bsa*I	Plasmid with constitutively active promotor encoding for a C-terminal degradation tag for complementation experiments
pUR22	Amp	*Bsa*I	Expression plasmid encoding for a C-terminal His_6-tag
pUR23	Amp	*Bsa*I	Expression plasmid encoding for an N-terminal His_6-tag
pMal_*Bsa*I	Kana	*Bsa*I	Expression plasmid encoding for an N-terminal maltose binding protein
pKD3	Kana	-	Template for kanamycin resistance gene

7.8 Gene sequences

Table 7.2: Gene sequences of experimentally characterized proteins.

Residues which are cleaved off by digestion with *Bsa*I are highlighted in blue.

Name	Nucleotide sequence
echisB-N	ATGAGTCAGAAGTATCTTTTTATCGATCGCGATGGAACCCTGATTAGCGAA CCGCCGAGTGATTTTCAGGTGGACCGTTTTGATAAACTCGCCTTTGAACCG GGCGTGATCCCGGAACTGCTGAAGCTGCAAAAAGCGGGCTACAAGCTGGT GATGATCACTAATCAGGATGGTCTTGGAACACAAAGTTTCCCACAGGCGGA TTTCGATGGCCCGCACAACCTGATGATGCAGATCTTCACCTCGCAAGGCGT ACAGTTTGATGAAGTGCTGATTTGTCCGCACCTGCCCGCCGATGAGTGCGA CTGCCGTAAGCCGAAAGTAAAACTGGTGGAACGTTATCTGGCTGAGCAAG CGATGGATCGCGCTAACAGTTATGTGATTGGCGATCGCGCGACCGACATTC AACTGGCGGAAAACATGGGCATTACTGGTTTACGCTACGACCGCGAAACCC TGAACTGGCCAATGATTGGCGAGCAACTCACCAGACGTGACTAA
ecgmhB	GTGGCGAAGAGCGTACCCGCAATTTTTCTTGACCGTGATGGCACCATTAAT GTCGATCACGGCTATGTCCATGAGATCGACAACTTTGAATTTATCGACGGT GTTATTGACGCCATGCGCGAGCTAAAAAAAATGGGCTTTGCGCTGGTGGTA GTAACCAACCAGTCTGGCATTGCTCGCGGTAAATTTACCGAAGCACAGTTT GAAACGCTGACCGAGTGGATGGACTGGTCGCTGGCGGACCGAGATGTCGA TCTGGATGGTATCTATTATTGCCCGCATCATCCGCAGGGTAGTGTTGAAGA GTTTCGCCAGGTCTGCGATTGCCGCAAACCACATCCGGGGATGCTTTTGTC AGCACGCGATTATTTGCATATTGATATGGCCGCTTCTTATATGGTGGGCGAT AAATTAGAAGATATGCAGGCAGCGGTTGCGGCGAACGTGGGAACAAAAGT GCTGGTGCGTACGGGTAAACCTATTACGCCTGAAGCAGAAAACGCGGCGG ATTGGGTGTTAAATAGCCTGGCAGACCTGCCGCAAGCGATAAAAAAGCAG CAAAAACCGGCACAATGA
csgmhB	AAAAAAGGTCTCACATGACCAAACGTGCAGTTTTTCTGGATCGTGATGGCA CCCTGATTGTTGATCATGGCTATATTCATAAACCGAGCCAGGTTGAACTGC TGCCTGGTGTTATTGAAGCACTGATTAAACTGAAAACCTTTGGCTTTGAGCT GATCATTATTAGCAATCAGAGCGGTATTGGTCGTGGTTTCTTTACCAAGAA AGAAGTGGACCACGTTAACCAGCATCTGTATAATCTGCTGATTAGCCACAA AATCAAACTGACCGGCATCTATTATTGTCCGCATCATCCTGATGATAAATG TACCTGTCGTAAACCGGAACCGGGTCTGCTGCTGCAGGCACTGAGCGAACA TAAAATTGATGCAAAGAAAAGCTATTTTGTGGGCGATAAACTGACAGATGT TCAGGCAGCAATTGCAGCCGGTGTTCAGCCGGTTCTGCTGAGCCGTGATAA TGTTAGCACCCATACCATTCCGATTATTATCGTTGATAGCCTGCTGAAATTC ACCAAGGTGATCAAACAAGAGGACTTTTAAGAGAGACCAAAAAA
anc1	AAAAAAGGTCTCACATGCGTCGTTATGTTTTTCTGGATCGTGATGGCACCA TTAACGAAGATGTTTATCTGAGCGATCCGGAAAAAGTTCGTCTGCTGCCTG GTGCAGCAGAAGGTCTGAAACTGCTGCAAGAAGCAGGTTTTGCACTGGTTG TTGTTACCAATCAGAGCGGTATTGCACGTGGTTATTTTAGCGAAGAAAGCG TTCATGCAGTTCATGCCCGTATGCAAAAACTGCTGGCACCGCATGGTGTTC

	AGATTGATGGTATCTATTTTTGTCCGCATCATCCGGAAGAAAATTGCGATT GTCGTAAACCGAGTCCTGGTATGGTTCTGCAGGCAGCACGTGAACTGGGTA TTGATCCGAGCCGTAGCTATGTTATTGGTGATCGTGAAAGCGATATTGAAC TGGCACGTAATGTTGGTGCAAAAAGCGTTCTGGTTCTGACCGGTAAAGGTC AAGAACAGCCGGATCTGGTTGCCGAAGATCTGCTGGATGCAGCACGTCTGA TTCTGAAATAAGAGAGACCAAAAAA
anc2	AAAAAAGGTCTCACATGCCGGTTGTTTTTCTGGATCGTGATGGCACCATTA ACGAAGAAGTGTATCTGAATAATCCGGAAAAAGTTCGTCTGCTGCCTGGTG TTGCAGAAGCACTGCGTCTGCTGAAAGAAGCAGGTTTTAAACTGGTTGTGA TTACCAATCAGAGCGGTATTGCACGTGGTTATTTTAGCGAAGAAGAAGTGC ACGCAGTTCATCAGCGTATGATGAAACGTCTGGCAACCCATGGTGTTCAGA TTGATGATATCTATTTTTGTCCGCATCACCCGGAAGAAAATTGTGATTGTCG TAAACCGAAACCTGGTCTGGTTCTGAAAGCAGCACAGAAATGGAATATTG ATCTGAGCCGTAGCTATGTGATTGGTGATCGTGATACCGATATTGAACTGG CATTTAATGCAGGTTGTAAAGGTGTTCTGGTGCTGACCGGTTATGGTAAAC AGCTGCCGGATGTTGCAAAAGATCTGCTGGATGCAGCAAAACTGATTCTGA AATAAGAGAGACCAAAAAA
anc3	AAAAAAGGTCTCACATGCAGCCGGTTGTTTTTCTGGATCGTGATGGCACCA TTAACGAAGAAGTGTATCTGAATAATCTGAGCCGTCTGCGTCTGCTGCCTG GTGTTGCAGAAGCAATTCGTCTGCTGAAAAAGGCAGGTTTTGCAGTTGTTG TTATTACCAATCAGAGCGGTCTGGCACGTGGTTATTTTCCGGAAGATACCC TGCATGCCGTTCATCAGAAAATGCTGAAACGTCTGAGCACCCGTGGTGCAA CCATTGATGGTATTTATGTTTGTCCGCATCATCCGGAAGAAAATTGTGAAT GTCGTAAACCGAAACCTGGTCTGGTTCTGAAAGCAAGCCAAGAACTGAAA CTGGATCTGAGTCGTGCATATGTTATTGGTGATCGTGATACCGATATTCAGC TGGCCAAAAACATTGGTGCAAAAGGTGTTCTGGTGCTGACCGGTTATGGTG AAGGTGAACTGCCGGATCTGGTTGCAAAAGATCTGCTGAGCGCAGCAAAG AAGATTCTGAAATAAGAGAGACCAAAAAA
anc4	AAAAAAGGTCTCACATGCGTCCGGTTGTTTTTCTGGATCGTGATGGCACCA TTAACGAAGAGATGTATATTAACCATCTGAGCCGTCTGCGTCTGCTGCCTG GTGTTGCACAGGCACTGCGCCTGCTGCAAGAAGCAGGTTTTAAACTGGTTA TTGTTACCAATCAGAGCGGTCCGGCACGTGGTTATTTTCCGGAAGAACTGG TTCATGAAACCCATCAGATGCTGCAGCGTCGTCTGGCAGCACAGGGTGTTA AACTGGATGATCTGTATGTTTGTCTGCATCATCCTGAAGAAGGTTGTAGCT GTCGTAAACCGCGTCCTGGTCTGGTTCTGCAGGCACTGGAAGAATATAGCA TTGATCTGGAACGTAGCTATGTGATTGGTGATCGTTGGGTTGATTTAGAACT GGCACGTAATATGGGTATTCGTGGTATTCTGGTGCTGACCGGTTATGGTCG TGGTGAACTGGAACCGGATGTTGCAAAAGATTGGAAAAGCGCAGCAGAAC TGATTCTGAAATAAGAGAGACCAAAAAA
anc5	AAAAAGGTCTCACATGCAGCCGATTCTGTTTATTGATCGTGATGGCACCCT GGTTGAAGAACCGCCTACCGATCAGGTTGATAGCCTGGATAAACTGGAATT TGAACCGGCAGTTATTCCGGCACTGCGTAAACTGCAGAATGCAGGTTATCG TCTGGTTATGGTTAGCAATCAGGATGGTCTGGGCACCCCGAGCTTTCCGCA

	AGAAAGCTTTGAAGCACCGCATAATATGATGATGGATATTTTTGCAAGCCA GGGCATCACCTTTGATGAAGTGTATATTTGTCCGCACTTTCCGGAAGATAA TTGCGATTGTCGTAAACCGCGTACCGGTCTGGTTACCAATTATCTGCGTGA ACGTCGTTTTGATCCGCAGCGTAGCTATGTTATTGGTGATCGTGAAACCGA TATGCAGCTGGCAGAAAATATGGGCACCGAAGGTATTCAGTATCGTCCTGG TGGTCTGGATTGGCCTGCAATTGCCGAACGTCTGCTGTAAGAGAGACCAAA AAA
anc6	AAAAAAGGTCTCACATGAGCCAGAAAATCCTGTTTATTGATCGTGATGGCA CCCTGATTGAAGAACCGCCTACCGATTTTCAGGTTGATAGCCTGGAAAAAC TGAAATTTGAACCGGCAGTTATTCCGGCACTGCTGAAACTGCAGAATGCAG GTTATCGTCTGGTTATGGTTACCAATCAGGATGGTCTGGGCACCGATAGCT TTCCGCAGGCAGATTTTGATCCGCCTCATAATCTGATGATGCAGATTTTTGA AAGCCAGGGCATTCGTTTTGATGAAGTTCTGATTTGTCCGCACTTTCCGGAA GATAATTGTAGCTGTCGTAAACCGAAAACCGGTCTGGTTACCCGTTATCTG AAAGAAGGCAAAATTGATAAAGAACGCAGCTATGTGATTGGTGATCGTGA AACCGATATGCAGCTGGCAGAAAATATGGGTATTCGTGGTCTGCAGTATAA TCCGAATCTGAATTGGGAAGCAATTACCGAACAGCTGACCAATTAAGAGA GACCAAAAAA
anc7	AAAAAAGGTCTCACATGAGCCAGAAAATCCTGTTTATTGATCGTGATGGCA CCCTGATTACCGAACCGCCTACCGATTTTCAGGTTGATAGCCTGGAAAAAC TGGCATTTGAACCGGCAGTTATTCCGGCACTGCTGAAACTGCAGAAAGCAG GTTATCGTCTGGTTATGGTTACCAATCAGGATGGTCTGGGCACCGATAGCT TTCCGCAGGCAGATTTTGATCCGCCTCATAATCTGATGATGCAGATTTTTGA AAGCCAGGGCATTCGTTTTGATGAAGTTCTGATTTGTCCGCATTTTCCGGCA GATAATTGTAGCTGTCGTAAACCGAAAACCGGTCTGGTTACCCATTATCTG CAAGAAGGCAAAATTGATTTTGAGCGCAGCTATGTTATTGGTGATCGTGAA ACCGATATTCAGCTGGCAGAAAACATGGGTATTCAGGGTCTGCGTTATAGT CCGGAACTGGATTGGGCAGCAATTACCCATCAGCTGACCTAAGAGAGACC AAAAAA
pahisN	AAAAAGGTCTCACATGCGTCTGGCACTGTTTGATCTGGATAATACCCTGCT GGCAGGCGATAGCGATCATTCATGGGGTGAATGGCTGTGTCAGCGTGGTCT GGTTGATGCAGCAGAATATCAGGCACGTAATGATGCATTCTATGCAGATTA TGTTGCCGGTAAACTGGATGTTCTGGCCTATCAGGCATTTACCCAGGCAAT TCTGGGTCGTACCGAAATGGCACAGCTGGAAACCTGGCATCGTCAGTTTAT GCAAGAAGTTATTGAACCGATTGTTCTGGCAAAAGGTGAAGCACTGCTGGC CGAACATCGTGCAGCCGGTGATCGTCTGGTGATTATTACCGCAACCAATCG TTTTGTTACCGGTCCGATTGCAGAACGTCTGGGTGTTGAAACCCTGATTGCA ACCGAATGTGAAATGCGTGATGGTCGTTATACCGGTCAGACCTTTGATGTT CCGTGTTTTCAAGGTGGTAAAGTTGTTCGTCTGCAGCGTTGGCTGGATGAA AATGGTCTGGATCTGGAAGGTGCAAGCTTTTATAGCGATAGCCTGAATGAT CTGCCGCTGCTGGAAAAAGTTAGCCGTCCGGTTGCAGTTGATCCTGATCCG CGTCTGCGTGCCGAAGCAGAAAAACGTGGTTGGCCGATTATTAGCCTGCGT TAAGAGAGACCAAAAAA

nmhisN	AAAAAAGGTCTCACATGGATAGCCAGAAATTCGATAGCATCATTTTTGATT GTGATGGCGTTCTGGTTGATATTACCCAGAGCTATGATAAAACCATCGATA AAACCTGCCGCTATGTGCTGAAAGAATTTGCCAAAATTGATTCCATCACCA TCGACCACAAAATCATCGATGGCTTTAAAAGCAGCGGTGGCTTTAATGATG AAGTGGATCTGGTTTATGCAGCAATTCTGAGCCTGTATACCGCAAATAAAC TGAATAAAAAGCCGAGCGAGTTCATCTATGATGTGATTAGCAATACCGACA AAACCGGTATTCGTAGCGTTCAGAGCTATCTGGAAAGCATTTATGATGTTA GCGAGTTCCTGAGCAAACTGGGTAGCCTGGGTGATCGTCATAATAATCCGG TTTATAGCATCTTTGATCAGTTTTTCTTTGGCCCTGAACTGTATGGCAAACT GTTTGATAAACAGAGCAAATTTAGCGAAGAGGGCATGATCAGCAACGATA AAGTTATTCTGAGCGTTAGCCTGCTGGAAACCCTGCAGAAAGAATTCGGTA AAAAGATTGCAGTTGTTACCGGTCGTGGTATTGAAAGCATTCGTTATAGCC TGAAAGACATGATGGATTACTTCGACACCAAAAATAGCGCCTTTCTGGAAG ATGAACCGCGTGAACTGGCAAAACCGAATCCGGCAACACTGATTCGTGCA ATTCAGAGCATGGAAAGCAAAAATTGTCTGTATGTTGGCGACAGCATGGA AGATTATATGATGGCAAAAGATGCAGCACAGGCAGGTCATAGCACCACCT TTTGTGCAATTGTTGGCACCAGCACCAATCCGGAAGATCGTCGTAAACTGT TCGCAGATAGCGGTGTTGAAATGATTCTGGAATCCATTAACGATATCCCGA AAGTTCTGAATCTGGTGTAAGAGAGACCAAAAAA
ecserB	ATGCCTAACATTACCTGGTGCGACCTGCCTGAAGATGTCTCTTTATGGCCG GGTCTGCCTCTTTCATTAAGTGGTGATGAAGTGATGCCACTGGATTACCAC GCAGGTCGTAGCGGCTGGCTGCTGTATGGTCGTGGGCTGGATAAACAACGT CTGACCCAATACCAGAGCAAACTGGGTGCGGCGATGGTGATTGTTGCCGCC TGGTGCGTGGAAGATTATCAGGTGATTCGTCTGGCAGGTTCACTCACCGCA CGGGCTACACGCCTGGCCCACGAAGCGCAGCTGGATGTCGCCCCGCTGGG GAAAATCCCGCACCTGCGCACGCCGGGTTTGCTGGTGATGGATATGGACTC CACCGCCATCCAGATTGAATGTATTGATGAAATTGCCAAACTGGCCGGAAC GGGCGAGATGGTGGCGGAAGTAACCGAACGGGCGATGCGCGGCGAACTCG ATTTTACCGCCAGCCTGCGCAGCCGTGTGGCGACGCTGAAAGGCGCTGACG CCAATATTCTGCAACAGGTGCGTGAAAATCTGCCGCTGATGCCAGGCTTAA CGCAACTGGTGCTCAAGCTGGAAACGCTGGGCTGGAAAGTGGCGATTGCCT CCGGCGGCTTTACTTTCTTTGCTGAATACCTGCGCGACAAGCTGCGCCTGAC CGCCGTGGTAGCCAATGAACTGGAGATCATGGACGGTAAATTTACCGGCA ATGTGATCGGCGACATCGTAGACGCGCAGTACAAAGCGAAAACTCTGACT CGCCTCGCGCAGGAGTATGAAATCCCGCTGGCGCAGACCGTGGCGATTGGC GATGGAGCCAATGACCTGCCGATGATCAAAGCGGCAGGGCTGGGGATTGC CTACCATGCCAAGCCAAAAGTGAATGAAAAGGCGGAAGTCACCATCCGTC ACGCTGACCTGATGGGGGTATTCTGCATCCTCTCAGGCAGCCTGAATCAGA AGCTCGAGCACCACCACCACCACCACTAA

7.9 Primers

Name	Sequence	Description

*nm*HisN_I19A_fo	CGACCCAGAGCTATGATAAAACC	Alanine Scan of *nm*HisN
*nm*HisN_I19A_re	CATCAACCAGAACGCCATC	
*nm*HisN_Y24A_fo	GGATAAAACCATCGATAAAACCTG	
*nm*HisN_Y24A_re	GCGCTCTGGGTAATATCAACC	
*nm*HisN_D25A_fo	GAAAACCATCGATAAAACCTGC	
*nm*HisN_D25A_re	GCATAGCTCTGGGTAATATCAAC	
*nm*HisN_I28A_fo	CCGATAAAACCTGCCGC	
*nm*HisN_I28A_re	CGGTTTTATCATAGCTCTGGG	
*nm*HisN_I53A_fo	CCGATGGCTTTAAAAGCAG	
*nm*HisN_I53A_re	CGATTTTGTGGTCGATGGTG	
*nm*HisN_K57A_fo	GAGCAGCGGTGGC	
*nm*HisN_K57A_re	GCAAAGCCATCGATGATTTTGTG	
*nm*HisN_F62A_fo	CGAATGATGAAGTGGATCTGGTTTATGC	
*nm*HisN_F62A_re	CGCCACCGCTGCTTTTAAAG	
*nm*HisN_E64A_fo	GGTGGATCTGGTTTATGCAG	
*nm*HisN_E64A_re	GCATCATTAAAGCCACCGC	
*pa*HisN_D16A_fo	GAGCGATCATTCATGGGG	Alanine Scan of *pa*HisN
*pa*HisN_D16A_re	GCGCCTGCCAGC	
*pa*HisN_D18A_fo	GCATTCATGGGGTGAATGG	
*pa*HisN_D18A_re	GCGCTATCGCCTGC	
*pa*HisN_Y47A_fo	GGTTGCCGGTAAACTGG	
*pa*HisN_Y47A_re	GCATCTGCATAGAATGCATCATTAC	
*pa*HisN_L52A_fo	CGGATGTTCTGGCCTATC	
*pa*HisN_L52A_re	CTTTACCGGCAACATAATCTG	
*pa*HisN_T113A_fo	CAATCGTTTTGTTACCGGTCCG	
*pa*HisN_T113A_re	GCTGCGGTAATAATCACCAGACG	
*pa*HisN_Y57_fo	CACAGGCATTTACCCAGG	
*pa*HisN_Y57_re	CGGCCAGAACATCCAGTTTAC	
*pa*HisN_Q58_fo	CGGCATTTACCCAGGC	
*pa*HisN_Q58_re	CATAGGCCAGAACATCCAG	
pKD3_*hisB*_fo	AAGAAAGCCAGCGCGTCATTGACGCCTTA CGTGCGGAGCAAGTTTGATGAGTGTAGGC TGGAGCTGCTTC	Add a sequence complementary to *hisB*

pKD3_*hisB*_re	TTCACCGAGTTCAGGTTGGCGCAGCCGGTATCAAGGATCACCACGTTCATCATATGAATATCCTCCTTAG	to the kanamycin resistance gene
*hisB*_fo	CGTGAAGAAAGCCAGCGCGT	Primer to test for genomic deletion of *hisB*
*hisB*_re	CTTCACCGAGTTCAGGTTGG	
pExp_tag_fw	CATAGTTTTCATCGTTCGCCGCTAAGCTTAATTAGCTGAGCT	Introduce SsrA-degadation tag
pExp_tag_rv	CGCTGGCGGCGTAACTCGAGTGAGACCTTAGGATCCTCGGTC	
HisB_Yip_fw1	AAAAGGTCTCAACTGCAGAAGCTGCAAAAAGCG	Create HisB-N triple mutant
HisB_Yip_rv1	AAAAGGTCTCACGGGGCGGTGCGGACAAATCAG	
HisB_Yip_fw2	AAAGGTCTCACCCGCCGATGAG	
HisB_Yip_rv2	AAAAGGTCTCAAACCAATAATGCCCATGTTTTCCG	
HisB_Yip_fw3	AAAAGGTCTCAGGTTTACGCTACGACC	
HisB_Yip_rv3	AAAAGGTCTCACAGTTCCGGGATCAC	
SerB_temp_fw1	GATGAAATTGCCAAACTGG	Generate a template for *ec*SerB randomization
SerB_temp_rv1	GCTGGCGGTAAAATCG	
SerB_temp_fw2	GTGGCGACGCTGAAAGG	
SerB_temp_rv2	AATCTGGATGGCGGTG	
SerB_rand_fw1	AAAAGGTCTCAGATTNNKTGTNNKGATGAAATTGCCAAACTGG	Randomize *ec*SerB
SerB_rand_rv1	AAAAGGTCTCATGCGMNNGCTGGCGGTAAAATC	
SerB_rand_fw2	AAAAGGTCTCACGCAGCNNKGTGGCGACGCTGAAAGG	
SerB_rand_rv2	AAAAGGTCTCAAATCTGGATGGCGGTGGAG	
SerB_seq_fw	ATGCCTAACATTACCTGGTGC	Sequencing primer for *ec*SerB
SerB_seq_rv	TTACTTCTGATTCAGGCTGCC	
HisBN_temp1_rv2	AATCACTCGGCGGTTC	
HisBN_temp1_fw2	GGTCTTGGAACACAAAGTTTC	

HisBN_temp1_rv1	AGTGATCATCACCAGCTTG	Generate template for *ec*HisB-N randomization (libaray 1)
HisBN_temp1_fw1	GTGGACCGTTTTGATAAACTC	
HisBN_rand1_fw1	AAAAGGTCTCATGATNNKNNKGTGGACCGTTTTGATAAAC	Randomize *ec*HisB-N (library 1)
HisBN_rand1_rv1	AAAAGGTCTCAAGTGATCATCACCAGC	
HisBN_rand1_fw2	AAAAGGTCTCACACTNNKCAGNNKGGTCTTGGAACACAAAG	
HisBN_rand1_rv2	AAAAGGTCTCAATCACTCGGCGGT	
HisBN_temp2_fo1	GGTCTTGGAACACAAAGTTTCC	Generate template for *ec*HisB-N randomization (library 2)
HisBN_temp2_rv1	GCAGTCGCACTCATCG	
HisBN_temp2_fo2	CCGAAAGTAAAACTGGTGGAAC	
HisBN_temp2_rv2	ATTAGTGATCATCACCAGCTTG	
HisBN_rand2_fo1	AAAAGGTCTCAGGTCTTGGAACACAAAGTTTC	Randomize *ec*HisB-N (library 2)
HisBN_rand2_rv1	AAAAGGTCTCATCGGMNNMNNGCAGTCGCACTCATCGGC	
HisBN_rand2_fo2	AAAAGGTCTCACCGAAAGTAAAACTGGTGGAACGTTATC	
HisBN_rand2_rv2	TTTTGGTCTCAGACCMNNMNNATTAGTGATCATCACCAGCTTGTAGCC	
HisBN_D58N_rv	CTGATTAGTGATCATCACCAG	Introduce D58N mutation
HisBN_D58N_fo	AATGGTCTTGGAACACAAAG	

7.10 Software

7.10.1 Local software

ÄKTA Unicorn Version 7.6.0	© GE HEALTHCARE
AliView Version 1.27	© Larsson A., 2014
Citavi 6	© Swiss Academic Software
ChemOffice 2020	© PerkinElmer
CLC main workbench 2021	© Quiagen
CorelDraw 2021	© Corel Corporation
Cytoscape 3.9.1	© Cytoscape Consortium
MS Office Professional 2020	© Microsoft Corporation
OmniSec	© Malvern Panalytical
Origin 2021	© OriginLab Corporation
Pymol Version 2.4.0	© Schrödinger LLC
Spectra Analysis 2.15.09	© JASCO Germany GmbH
Spectra Manager 2.14.06	© JASCO Germany GmbH

7.10.2 Server based software

BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
EFI-enzyme similarity tool	https://efi.igb.illinois.edu/efi-est/
Emboss Needle	https://www.ebi.ac.uk/Tools/psa/emboss_needle/
Emboss Water	https://www.ebi.ac.uk/Tools/psa/emboss_water/
ExPASy ProtParam tool	https://web.expasy.org/protparam/
NEB T_M calculator	https://tmcalculator.neb.com/#!/main
MAFFT	https://www.ebi.ac.uk/Tools/msa/mafft/
WebLogo 3	https://weblogo.berkeley.edu/logo.cgi

8 Methods

8.1 Bioinformatical methods

8.1.1 Analysis of the phylogenetic distribution and co-occurrence of enzymes

Different phylogenetic analyses were conducted using the orthology database (KO) of the Kyoto Encyclopedia of Genes and Genomes (KEGG). In the KO database, each group of orthologous enzymes is assigned with a KO number which can be used to access all members of the group with their associated sequences and host organisms.

First, the KO database was used to identify phylogenetic patterns in the occurrence of HolPases and 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 which are designated as K01089, the orthologs of the *L. lactis* HolPase from the PHP superfamily which are designated as K04486, the orthologs of the *M. truncatula* HolPase which are designated as K18649, and the orthologs of the *M. tuberculosis* HolPase which are designated as K05602. To identify additional organisms that most likely produce histidine, the occurrence of the conserved IGPDH[113] was used which is annotated as K01693. The organisms associated with each KO number were retrieved and grouped according to their phylogenetic class, which gave the occurrence of each type of HolPase for each phylogenetic class. Organisms with annotated IGPDH but without an annotated HolPase represent candidates for organisms where the HolPase is either missing or yet to be discovered.

The KO database was furthermore used to retrieve a comprehensive set of GmhBs. GmhB enzymes are denoted as K03273 irrespective of their preferred substrate (αHBP or βHBP). However, there are two distinct groups of nucleotidyltransferases, namely the D-*glycero*-α-D-*manno*-heptose-1-phosphate guanylyltransferases which are denoted as K15669 and the D-*glycero*-α-D-*manno*-heptose-1-phosphate adenylyltransferases which are denoted as K21345. The co-occurrence of either of the two enzymes indicated the existence of either the S-layer biosynthesis (K15669) or the lipopolysaccharide biosynthesis (K21345) and was used to classify GmhB enzymes as αGmhB or βGmhB.

8.1.2 Sequence similarity networks

Sequence similarity networks (SSNs) were used to find homologous enzymes, to explore their phylogenetic distribution, and to cluster closely related enzymes which are probably iso-functional. Homologous sequences of a query sequence were retrieved using the BLAST option of the enzyme similarity tool on the Enzyme Function Initiative (EFI) website.[124] As reference database, the UniRef90 was used which is derived from the Uniprot database. However, in the UniRef90, all sequences with more than 90% sequence identity to one another over more than 80 % of the sequence are clustered and only one sequence is added to the database. [126] Full SSNs were downloaded and further analyzed using Cytoscape (Version 3.9.1). To this end, nodes were colored according to the phylogenetic class of the corresponding organism and the sequence identity threshold for two sequences to be connected by an edge was increased in a stepwise manner. Isolated nodes without a connecting edge to another node were removed from the network after each step.

8.2 Microbiological methods

8.2.1 Cultivation and storage of *E. coli* strains

E. coli strains were cultivated in LB medium which was supplemented with at least one antibiotic (150 µg/mL for ampicillin, 30 µg/mL for chloramphenicol, 75 µg/mL for kanamycin) that corresponded to a plasmid-encoded or genome-encoded resistance gene of the respective strain. These cell suspensions were incubated at 37 °C while shaking at 120-140 rpm. Any deviations from these standard parameters are explicitly stated. To obtain single colonies, the cell cultures were plated on LB agar plates supplemented with at least one antibiotic and incubated for 1-3 days at 21-37 °C.

For long time storage, bacterial culture media were mixed with the same volume of glycerol (87 % in aqueous solution) and stored at –70 °C or –80 °C.

8.2.2 Preparation of *E. coli* knock-out strains

E. coli knock-out strains were constructed by P1 phage mediated homologous recombination. [148, 150, 151] The protocol relies on the heat inducible expression of the lambda phage derived genes *β*, *γ* and *exo* in the *E. coli* strain DY329.[152] The corresponding gene products slow the degradation of linear double stranded DNA and mediate homologous recombination. To avoid premature expression of *β*, *γ* and *exo*, DY329 cells were cultivated at 30 °C.

The first step of the protocol consisted in the PCR amplification (8.3.4.1) of an antibiotic resistance gene from a donor plasmid (kanamycin from pKD3, chloramphenicol from pKD4). The primers which were used during the PCR contained 50 bp overhangs at their 5' ends that were homologous to the genomic regions upstream and downstream of the target gene. The PCR products were then separated by agarose gel electrophoresis, the band with the correct size was excised and the amplified resistance gene was isolated from the gel (8.3.3). The purified product was then used to transform freshly prepared DY329 cells. To this end, 50 mL LB_0 medium were inoculated with 500 µL of an overnight culture of DY329 which was further cultivated until it reached an OD_{600} of 0.4-0.6. To prepare the cells for DNA uptake, the expression of *β*, *γ* and *exo* was triggered by a heat step of 15 min at 42 °C to which the bacterial suspension was subjected. Afterwards, the cells were shaken for 5 min on ice, prepared for electroporation according to the standard protocol (8.2.4), and transformed with the previously synthesized PCR product (8.2.4). The transformed cells were then streaked on LB agar plates supplemented with the appropriate antibiotic and incubated overnight at 30 °C. To identify colonies with the desired gene knock-out, colonies were analyzed by colony PCR (8.3.4.3). For the colony PCR, two sets of primers were used. One set consisted of a primer that bound outside of the newly introduced resistance gene and a primer that bound within the coding sequence of the resistance gene. The second set consisted of primers both bound to the flanking regions upstream and downstream of the resistance gene. A colony that yielded two amplicons of the correct sizes was then selected as donor strain of the gene knock-out. To establish a genetically stable cell line with the desired genetic modification, the gene knock-out was transferred from DY329 to the *E. coli* strain BW25113 by means of P1 phage transduction. For that purpose, 5 mL LB medium supplemented with 0.2 % glucose and 5 mM $CaCl_2$ were inoculated with an overnight culture of DY329 and shaken at 30 °C until a slight turbidity was visible. Then, P1 phage lysate was added in varying amounts (0-200 µL) and the cell suspension was further shaken until diminishing turbidity indicated phage induced cell lysis. The solution with the

lowest phage concentration that still became clear was selected and remaining *E. coli* cells were lysed by addition of three drops of chloroform followed by vigorous shaking using a vortex mixer. Afterwards, the solution was incubated for an additional 15 min at 37 °C. To remove cell debris, the suspension was centrifuged (10 min, 3220 g) and the resulting supernatant with P1 phages was transferred into a new reaction tube and stored at 4 °C. To prepare the acceptor strain BW25113 for P1 infection, an overnight culture was supplemented with $CaCl_2$ at a final concentration of 2.5 mM. For P1 infection, 0.8 mL of the cell suspension were mixed with 0.4 mL of either the undiluted or a diluted (1:10 in ddH_2O) phage lysate and incubated for 25 min at 37 °C without shaking. As a negative control, the phage lysate was replaced by 0.4 mL LB_0 medium. The infection was stopped by the addition of 5 mL P1 saline (145 mM NaCl, 50 mM trisodium citrate, autoclaved) which complexes free calcium ions that are required for the adsorption of the phages to the bacterial cells. The infected cells were then cultivated for 1 h at 37 °C to allow for expression of the newly acquired resistance genes. Afterwards, cells were pelleted by centrifugation (5-10 min, 3220 g), resuspended in P1 saline (100 µL) and streaked onto LB agar plates that were supplemented with the appropriate antibiotic. Colonies of BW25113 which carried the desired gene knock-out were then identified by means of colony PCR (8.3.4.3) with the same sets of primers as were used in the case of DY329.

8.2.3 Preparation and transformation of chemically competent *E. coli* cells

Chemically competent *E. coli* cells were prepared following a protocol that was established by Inoue et al.[153] First, 500 mL SOB or LB medium were inoculated to an OD_{600} of 0.1 with an overnight culture of the desired *E. coli* strain. The medium was then incubated at 37 °C while shaking until an OD_{600} of 0.6-0.8 was reached. Growth was arrested by incubation on ice for 15 min and cells were subsequently pelleted by centrifugation (10 min, 3220 g). To wash the cell pellet, it was resuspended in 100 mL of cooled TFB I buffer and again pelleted by centrifugation (10 min, 3220 rpm). Afterwards, the pellet was resuspended in 10 mL of cooled TFB II buffer, split into 100 µL aliquots, and either used for transformation immediately or stored at –80 °C for later usage.

To transform chemically competent cells, either a frozen aliquot was thawed on ice (5-10 min) and then mixed with plasmid DNA (70-130 ng), or a fresh aliquot was directly mixed with plasmid DNA (70-130 ng). Thereafter, the mixture was incubated for another 5 minutes on ice. The uptake of plasmid DNA was then triggered by a heat shock (45 s, 42 °C) followed by another incubation on ice (5-10 min). To allow for cell recovery and expression of the resistance genes, 900 µL LB_0 medium were added, and the cell suspension was incubated for 30-60 min at 37 °C while shaking. To select for transformed cells, the medium was then plated on LB agar plates that was supplemented with an appropriate antibiotic.

8.2.4 Preparation and transformation of electrocompetent *E. coli* cells

Electrocompetent cells were prepared immediately prior to usage to ensure maximum transformation efficiency. For that purpose, 50 mL LB medium supplemented with an appropriate antibiotic were inoculated with an overnight culture to a final OD_{600} of 0.1 and grown at 37 °C while shaking to a final OD_{600} of 0.6-0.8. The cells were then harvested by centrifugation (10 min, 3220 g) and washed four times with decreasing volumes of ice-cold ddH_2O (50 mL, 30 mL, 1.5 mL, 1.5 mL). The final cell suspension was split into aliquots (100 µL each) and used immediately for transformation.

To transform electrocompetent cells, freshly prepared aliquots of electrocompetent cells were mixed with 1-10 ng plasmid DNA and incubated on ice (5-10 min). Each cell suspension was then transferred into a precooled electroporation cuvette and subjected to an electric pulse (2500 V, 25 µF). To allow for cell recovery, 900 µL LB_0 medium were added immediately, and the cell suspension was incubated for 30-60 min at 37 °C while shaking. For subsequent complementation experiments of a gene knock-out, the cell suspension was plated onto agar plates with M9 minimal medium supplemented with an appropriate antibiotic. To cultivate and afterwards harvest cells that contained gene libraries, cells transformed with the respective library were plated onto agar plates with LB medium supplemented with an appropriate antibiotic.

8.2.5 Complementation of a gene knock-out

To identify enzymes that could complement a missing HolPase activity or to identify enzyme variants with improved HolPase activity, complementation experiments with a HolPase deficient strain were conducted. For that purpose, a Δ*hisB::kana*^{R} strain which was based on the *E. coli* strain BW25113 was constructed (8.2.2). To complement for the missing IGP dehydratase activity of the bifunctional HisB enzyme, the Δ*hisB::kana*^{R} strain was transformed (8.2.3) with a plasmid that contained a gene encoding for the monofunctional IGP dehydratase of *Bacillus subtilis* under the control of a constitutively active promoter and a chloramphenicol resistance gene (pExp_*igpdh_cam*^{R}). Genes encoding for the target enzymes that should be tested for their HolPase activity were subcloned into a vector with a constitutively active promotor and an ampicillin resistance gene (pExp_*gene_of_interest_amp*^{R}). The HolPase deficient cells (Δ*hisB::kana*^{R} + pExp_*igpdh_cam*^{R}) were then cultivated in $LB_{Kana,Cam}$, prepared for electroporation, transformed with the genes of interest (8.2.4), and streaked on M9 minimal medium plates that were supplemented with ampicillin, chloramphenicol, and kanamycin and incubated at 21-37 °C for up to 10 days.

To identify *ec*HisB-N variants that could complement a missing phosphoserine phosphatase, complementation experiments with a SerB deficient strain were conducted. To this end, a Δ*serB::kana*^{R} strain which was based on the E. coli strain BW25113 from previous work by Dr. Bettina Rohweder was employed. Genes of interest were subcloned into a vector with a constitutively active promoter and a C-terminal degradation tag (pExp_*gene_of_interest_ssrA_tag_amp*^{R}) which allowed for very low protein levels. Cells were transformed with these plasmids (8.2.4) and the ensuing complementation experiments were performed in the same fashion as for the complementation of the HolPase deficiency.

8.3 Molecular biological methods

8.3.1 Isolation and purification of plasmid DNA from E. coli

Plasmid DNA from *E. coli* was isolated and purified from 5 mL of an overnight culture using the GeneJET Plasmid Miniprep kit. The experimental procedures were performed according to the manual provided by the supplier except for the elution step, where 35-40 mL ddH_2O was used instead of the provided elution buffer. Isolated plasmid DNA was then stored at –20 °C.

8.3.2 Measurement of DNA concentration

The concentrations of DNA solutions were measured by means of UV/Vis spectroscopy. To this end the absorption was determined at 260 nm using a NanoDrop One spectrophotometer and concentrations were calculated using Lambert-Beer's law. The extinction coefficient for double stranded DNA is 0.02 $cm^2 \mu g^{-1}$.

8.3.3 Agarose gel electrophoresis and isolation of DNA fragments

Agarose gel electrophoresis was employed to separate DNA molecules according to their size. To visualize DNA bands in the gel, ethidium bromide was used.[154] Agarose solutions were prepared at a final agarose concentration of 0.8-2.5 % (w/v) in 0.5 x TBE buffer by heating the initial suspensions in the microwave until the agarose was completely dissolved. To prevent degradation of ethidium bromide or premature polymerization, the solution was cooled to 60 °C and kept at this temperature until use. Immediately prior to pouring a gel, 0.2 µL ethidium bromide (10 mg/mL) were added per 30 mL of agarose solution. After polymerization of the gel, it was covered with 0.5 x TBE buffer. DNA samples were mixed with loading dye when necessary and loaded onto the gel. As a reference, the GeneRuler 1 kb Plus DNA Ladder was used. Electrophoresis was performed by applying an electric potential of 190 V for 20-25 min. Bands were visualized under UV light (λ = 302 nm) with a GelDoc Go or GelDoc-It imaging system, fragments with the correct size were excised from the gel and isolated by means of the GeneJET gel extraction kit. The experimental procedures were performed according to the manual provided by the supplier except for the elution step, where 33-40 mL ddH_2O were used instead of the provided elution buffer. Isolated DNA was either used immediately or stored at –20 °C.

8.3.4 Enzymatic manipulation of DNA

8.3.4.1 Polymerase chain reaction (PCR)

Polymerase chain reaction (PCR) was applied to amplify DNA fragments from genomic or plasmid DNA.[155] For PCR reactions that were intended to introduce changes in the sequence of a target gene, the commercially available high-fidelity polymerases Q5 HF or Phusion HF together with their specific reaction buffers were employed. However, for PCR reactions that were intended to verify the correct size of a DNA fragment, the commercially available GoTaq polymerase was used together with 5x Green GoTaq buffer. The standard reaction solution is given in Table 8.1.

Table 8.1: Standard PCR reaction mix.

Volume [µL]	Component	Concentration
0.5	Forward primer	
0.5	Reverse primer	
0.5	dNTPs	
5.0	Reaction buffer	
17.1	ddH_2O	
1.0	Template DNA	(0.5 - 5 ng/µL)
0.4	Polymerase	

PCR reactions were performed as hot-start reactions in thermocyclers according to the standard protocol given in Table 8.2.

Table 8.2: Standard protocol for PCR reactions.

step	temperature	duration
1. Initial denaturation	98 °C	1 min 30 s
2. Denaturation	98 °C	10-15 s
3. Annealing	T_A	20-30 s
4. Elongation	72 °C	30 s/kb (Q5, Phusion) 1 min/kb (GoTaq)
5. Final elongation	72 °C	10 min
6. Hold	4 °C	∞

The steps 2. - 4. were repeated in this order for 35 cycles. The lid temperature of the thermocycler was set to 105 °C and the annealing temperature T_A was calculated using the NEB T_M calculator. In the case of hard-to-amplify DNA targets, different annealing temperatures in a range of ±5 °C of the calculated T_A were tested using the gradient function of the thermocycler. For GC rich target sequences, High GC Enhancer was added to the reaction mixture.

8.3.4.2 Digestion of template DNA by *Dpn*I

The restriction enzyme *Dpn*I specifically binds and digests methylated DNA[156] and was used to digest template DNA after PCR amplification. For this purpose, 1 µL of commercially available *Dpn*I (10 U/µL) together with 5 µL CutSmart buffer were added to the PCR products and incubated for 30 min at 37 °C. Afterwards, the PCR products were separated by agarose gel electrophoresis and the target DNA was isolated (8.3.3).

8.3.4.3 Colony PCR

Colony PCR was used to identify colonies that had a correctly sized insert in their genomic DNA or in their plasmid DNA. For that purpose, a single colony was picked from a selective agar plate, mixed with 30-50 µL ddH_2O and incubated for 5 min at 95 °C to disrupt the cells. The resulting cell debris was pelleted by centrifugation (30 s, 12000 g) and 1-2 µL of the DNA containing supernatant were added as DNA template to a standard PCR mix with GoTaq polymerase and Green GoTaq buffer (8.3.4.1).

8.3.4.4 Site directed mutagenesis

The site directed mutagenesis protocol is based on a protocol from Finnzymes (THERMO FISHER) and was used to introduce point mutations, insertions, and deletions into double stranded plasmid DNA. Desired point mutations or insertions were introduced at the 5'-end of the PCR primers which were designed in such a way that they would bind to a contiguous DNA stretch in an end-to-end orientation. If more than one base was mutated or inserted, both primers contained a similar number of mutated or inserted bases. In the case of deletions, the primers were designed to bind immediately upstream and downstream of the DNA stretch that should be deleted. Both primers were designed in such a way that the calculated annealing temperatures deviated by less than 2 °C and were usually in the range of 57 °C to 62 °C. If a desired amino acid exchange could be achieved by different nucleotide exchanges, codons with higher frequency were preferred and adenine or thymine bases at the 5'-ends were avoided if possible. This was done because in the final plasmid, these adenine or thymine bases were often found to be missing. PCR reactions were performed with the Q5 polymerase or Phusion polymerase (8.3.4.1) and the final PCR amplicons were purified by means of agarose gel electrophoresis and subsequent gel extraction (8.3.3). The linear product DNA was cyclized by a coupled phosphorylation-ligation reaction and the remaining template DNA was removed by *Dpn*I digestion (8.3.4.2). Finally, NEB Turbo cells were transformed with the ligation product and plated on LB agar plates with the appropriate antibiotic (8.2.3).

8.3.4.5 Golden Gate cloning

Golden gate cloning[157] was used to subclone synthetic genes and PCR products into plasmids. Different plasmids were used, depending on the purpose of the final construct. For gene expression and subsequent protein purification (8.4.2), IPTG inducible expression plasmids with His_6-tags at the N- or C-terminal ends (pUR23, pUR22) or with a combination of N-terminal His_6-tag and N-terminal maltose binding protein (pMal_*Bsa*I) were utilized.[149] For complementation experiments of gene knock-outs (8.2.5), target genes were cloned into a plasmid with a tryptophanase operator (pExp) that mediates low constitutive expression.[149] For an even lower protein level, a pExp plasmid which was modified with a C-terminal SsrA degradation tag[158] (pExp_*ssrA*) was employed. To allow for the ligation of the *Bsa*I digested plasmid and the gene of interest, *Bsa*I cleavage sites together with a stretch of four bases, that would lead to single stranded overlaps complementary to the single stranded overlaps of the digested plasmid, were added to the target gene. This was either done by adding the corresponding bases at the 3'-end and the 5'-end of a synthetic gene (8.3.4.8) or by a PCR reaction (8.3.4.1). A standard reaction mixture for the coupled *Bsa*I digestion and T4 DNA ligase ligation is given in Table 8.3.

Table 8.3: Standard reaction mix for coupled digestion-ligation reactions.

Amount	Component
2 µL	*Bsa*I reaction buffer
2 µL	Ligase buffer
1 µL	T4 DNA Ligase
1 µL	*Bsa*I
100 ng	Plasmid DNA
300 ng	Insert
Ad 20 µL	ddH_2O

The coupled digestion/ligation reaction was then performed in a thermocycler according to the protocol given in Table 8.4.

Table 8.4: Standard reaction conditions for coupled digestion-ligation reactions

Reaction step	Temperature	Duration
1. Initial restriction	37 °C	10 min
2. Ligation	16 °C	5 min
3. Restriction	37 °C	5 min
4. Final restriction	37 °C	5 min
5. Heat inactivation	65 °C	10 min
6. Hold	4 °C	∞

Steps 2. - 3. were repeated in this order for 50 cycles and the lid temperature of the thermocycler was set to 50 °C. The ligated plasmid was then used to transform chemically competent cells (8.2.3) or, in the case of gene libraries, to transform freshly prepared electrocompetent cells (8.2.4). Transformed cells were then plated onto LB agar plates with an appropriate antibiotic.

8.3.4.6 Construction of focused gene libraries

Focused gene libraries were constructed by coupling a randomization reaction with degenerate primers[159] with the high-fidelity golden gate cloning method (8.3.4.5). This allowed for the simultaneous randomization of several distant positions. To this end, two rounds of PCR reactions were employed (Figure 8.1).

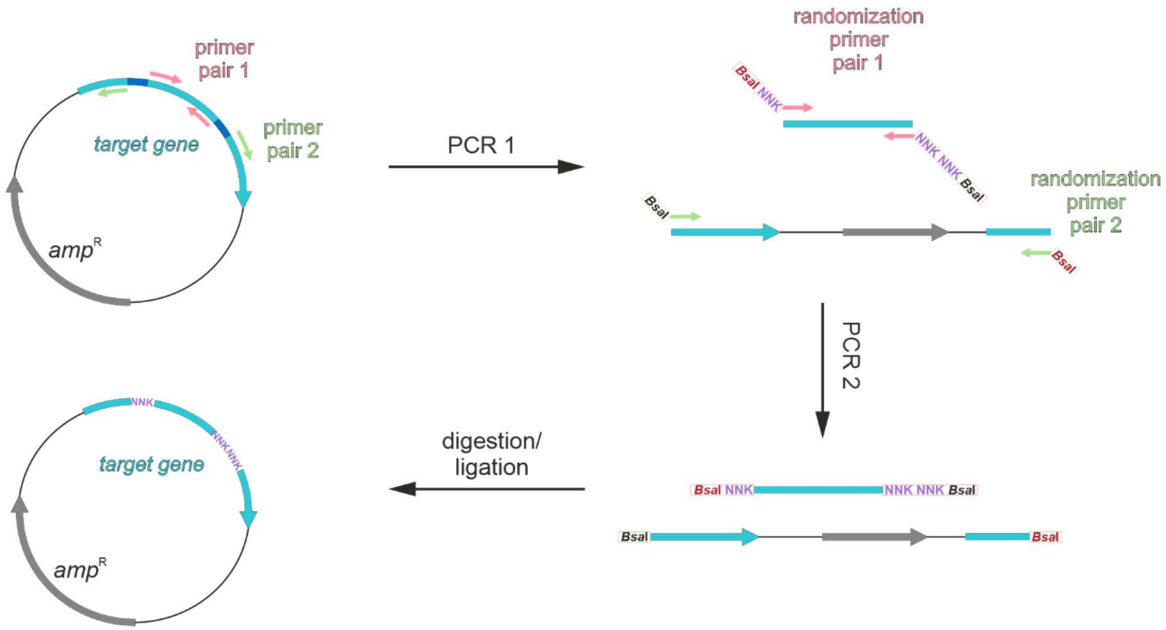

Figure 8.1: Three-step construction of focused gene libraries.

Focused gene libraries were constructed by a three-step process. In the first step, two independent PCR reactions were performed, using either primer pair 1 or primer pair 2. The two primer pairs were designed in such a way that the codons which should be randomized, and any codons in between, (dark blue) were missing in the two linear PCR products. The two PCR products were used as templates for the two PCR reactions of step 2. In step two, degenerate primers with randomized codons and BsaI digestion sites were used. Since the codons which should be randomized were missing in the PCR template, the bias which is introduced by preferential binding of some randomized codons over others was minimized. In the third step, the two linear PCR products with flanking BsaI digestion sites and randomized codons were digested and ligated to give a plasmid-encoded, randomized target gene.

The primers of the first PCR were designed in such a way that the positions that should be randomized would be omitted in the amplification. This step served to prepare the template DNA for the ensuing second PCR. The primers for the second PCR consisted of a 3'-end that corresponded to the primers from the first round and bound to the template DNA and a 5'-overlap that served to introduce randomized codons. This 5'-end contained one or two NNK codons, a *Bsa*I recognition site and four (coding) bases that formed the sticky overlap after *Bsa*I digestion. This 5' end did not bind to the template DNA which should lead to equal binding of all degenerate primers to the template DNA and hence reduce any wildtype bias. To minimize the bias that is caused by different annealing temperatures of different NNK codons in later cycles of the PCR reaction and to ensure that enough genetic material was produced, up to 12 reactions were performed in a parallel fashion with varying annealing temperatures. After successful amplification, the PCR products were separated by agarose gel electrophoresis, extracted from the gel (8.3.3), and ligated in a coupled digestion/ligation reaction in 8-12 parallel reactions of 20 µL each (8.3.4.5). Afterwards, residual template DNA was removed by *Dpn*I digestion (8.3.4.2). To remove excess salt and increase the transformation efficiency, the ligated plasmids were pipetted onto membrane filters (0.025 µM pore size, Millipore) and dialyzed against 30 mL ddH_2O for 4 h. The dialyzed plasmid DNA was then used to transform electrocompetent Δ*hisB*Δ*serB E. coli* cells in 10-20 parallel transformation reactions (8.2.4). Afterwards, cells were plated onto 10-15 LB_{Amp} agar plates (16 cm diameter) and incubated overnight at 37 °C. Additionally, a dilution series with dilutions up to 10^{-6} was prepared in LB medium and plated on LB_{Amp} agar plates (10 cm diameter) and incubated overnight at 37 °C. Afterwards, the gene libraries generated this way were

harvested from the plates. For this purpose, 3-4 mL LB_{Amp} medium were added to each plate and the densely grown colonies were scraped off the agar, pooled in a final volume of 50 mL LB_{Amp} medium and mixed thoroughly. Several aliquots of the cell suspension were then utilized to isolate the plasmid library (8.3.1). The dilution series was used to calculate the total number of transformed cells which is equal to the maximum theoretical library size. The plasmids of 15-20 colonies of the dilution series were additionally purified (8.3.1) and sequenced (8.3.4.7) to determine the fraction of introduced stop codons and fragmented genes. These numbers were then used to estimate the size of the gene library.

8.3.4.7 DNA sequencing

The correct coding sequence of all plasmids was verified by means of DNA sequencing. This was done using the Sanger sequencing service from Microsynth Seqlab. Samples that were sent for sequencing contained an aqueous solution of 600-700 ng plasmid DNA and 2 µM sequencing primer in a final volume of 15 µL. The sequencing data were analyzed with CLC main workbench (Qiagen).

8.3.4.8 Gene synthesis

Gene sequences, that correlate with genes from ancestral sequence reconstruction, genes from organisms other than *E. coli*, or gene sequences which should be altered by several mutations were ordered from GeneArt (THERMO FISHER). The respective coding sequences were codon optimized for the expression in *E. coli* and *Bsa*I recognition sites were attached to the 5'- and 3'-ends of the sequences to enable golden gate cloning (8.3.4.5).

8.4 Protein biochemical methods

8.4.1 Heterologous gene expression with subsequent analysis in analytical scale

Genes were expressed in analytical scale with the aim to improve protein solubility or protein stability in cases where no overexpression was observed or where the corresponding protein was predominantly found in the insoluble fraction of the fragmented cells. To this end, different cells (BL21 gold DE3, BW25113 Δ*serB*Δ*hisB*), different temperatures during gene expression (16 °C - 37 °C) and different constructs (N-terminal or C-terminal His_6-tags and N-terminal maltose binding protein) were tested. An overnight culture of the transformed cells was then used to inoculate 50 mL of LB medium supplemented with an appropriate antibiotic to an OD_{600} of 0.1. Then, the inoculated bacteria suspension was shaken at 37 °C. When an OD_{600} of 0.6-0.8 was reached, gene expression was induced by addition of IPTG (0.5 mM), and cells were grown overnight at temperatures ranging from 16 °C to 37 °C. Afterwards, cells were harvested by centrifugation (3200 g, 5-10 min, 4 °C), resuspended in protein buffer and disrupted by sonication (10 % amplitude, 2 s on, 2 s off). Resulting cell debris was pelleted by centrifugation (23,700g, 15 min, 4 °C), the soluble fraction was removed, and the pellet was resuspended in protein buffer. Both the pellet fraction and the soluble fraction were then analyzed by SDS-PAGE (8.5.2) and expression conditions that led to prominent bands of the target protein in the soluble fraction were used for gene expression in preparative scale (8.4.2).

8.4.2 Heterologous gene expression in preparative scale

Strains for gene expression were selected in such a way that contamination with host cell enzymes of the same function was avoided, especially when low promiscuous activities of the target proteins should be measured. Therefore, genes encoding for GmhB variants were expressed in a Δh*isB* knockout strain whereas genes encoding for different PSPases and HolPases were expressed in a Δ*serB*Δ*hisB* strain. First, LB medium (2-10x 1 L), supplemented with the antibiotic that corresponded to the plasmid encoded resistance gene, was inoculated with an overnight culture to an OD_{600} of 0.1 and cultivated at 37 °C while shaking. When an OD_{600} of 0.6-0.8 was reached, gene expression was induced by addition of IPTG (0.5 mM), and cells were grown overnight at 20-25 °C.

8.4.3 Cell disruption and isolation of the soluble fraction

Cells from expression cultures (8.4.2) were harvested by centrifugation (3220 g, 20 min, 4 °C), resuspended in resuspension buffer, and disrupted by sonication (60% amplitude, 2 min 30 sonication time, 2 s on, 2 s off). Cell debris was removed by centrifugation (23,700g, 40–45 min, 4 °C) and target proteins were purified from the soluble fraction.

8.4.4 Immobilized metal affinity chromatography

Immobilized metal affinity chromatography was used as the first step in the purification of His_6-tagged target proteins from the soluble fraction after cell disruption and centrifugation (8.4.3). To this end, Äkta chromatography systems were used together with either a HisTrap™ FF crude column (5 ml, GE Healthcare) or a HisTrap™ excel column (5 mL, Cytiva). The chromatography system was first equilibrated with resuspension buffer, then the sample was applied to the column and host cell proteins were removed by a wash step. Bound target proteins were then eluted from the column by a linear imidazole gradient. The exact instrument settings can be found in Table 8.5.

Table 8.5: Standard settings for immobilized metal affinity chromatography.

step	Setting	Flow rate [ml/min]
Pre-equilibration	2 CV resuspension buffer	5
Sample application	Soluble fraction in resuspension buffer	3-4
Wash	10 CV resuspension buffer	3-4
Elution	15 CV linear gradient	3-4
	0-75 % elution buffer	
Wash-out	2 CV 100 % elution buffer	5
Re-equilibration	3 CV resuspension buffer	5

The elution process was monitored by absorbance measurement at 260 nm and 280 nm and fractions that contributed to absorbance peaks at 280 nm were further analyzed by SDS-PAGE (8.5.2). Fractions

with a band that corresponded to the expected molecular weight of the target protein were pooled and further purified by preparative size exclusion chromatography (8.4.5).

8.4.5 Preparative size exclusion chromatography

Size exclusion chromatography was used as the second chromatography step in protein purification. To this end, Äkta chromatography systems were used with a HiLoad 26/600 Superdex 75 pg column at 4 °C. Prior to sample application, the column was equilibrated with the buffer in which the protein sould be stored. Then, the sample was loaded, and the elution process was monitored by absorbance measurement at 280 nm. Fractions that contributed to absorbance peaks were then analyzed by SDS-PAGE (8.5.2). Fractions that showed a protein band that corresponded to the expected molecular weight of the target protein and minimal contamination by other proteins were pooled and concentrated by means of a centrifugal filter (molecular weight cut-off 10 kDa).

8.4.6 Buffer exchange in analytical scale

To exchange the buffer of small volumes of protein solutions for analytical purposes, size exclusion columns with gravity flow were used. This was done with NAP-5 or NAP-10 dialysis tubes with a molecular weight cut-off of 14 kDa.

8.4.7 Storage of purified proteins

Purified proteins were dripped into liquid nitrogen, and stored at -70 °C or -80 °C.

8.4.8 Synthesis and purification of αHBP and βHBP

8.4.8.1 Enzymatic synthesis of αHBP and βHBP

The two anomeric substrates D-*glycero*-D-*manno*-heptose-1β,7-bisphosphate (βHBP) and D-*glycero*-D-*manno*-heptose-1α,7-bisphosphate (αHBP) were synthesized enzymatically according to a modified protocol of Wang et al. (Figure 8.2).[99]

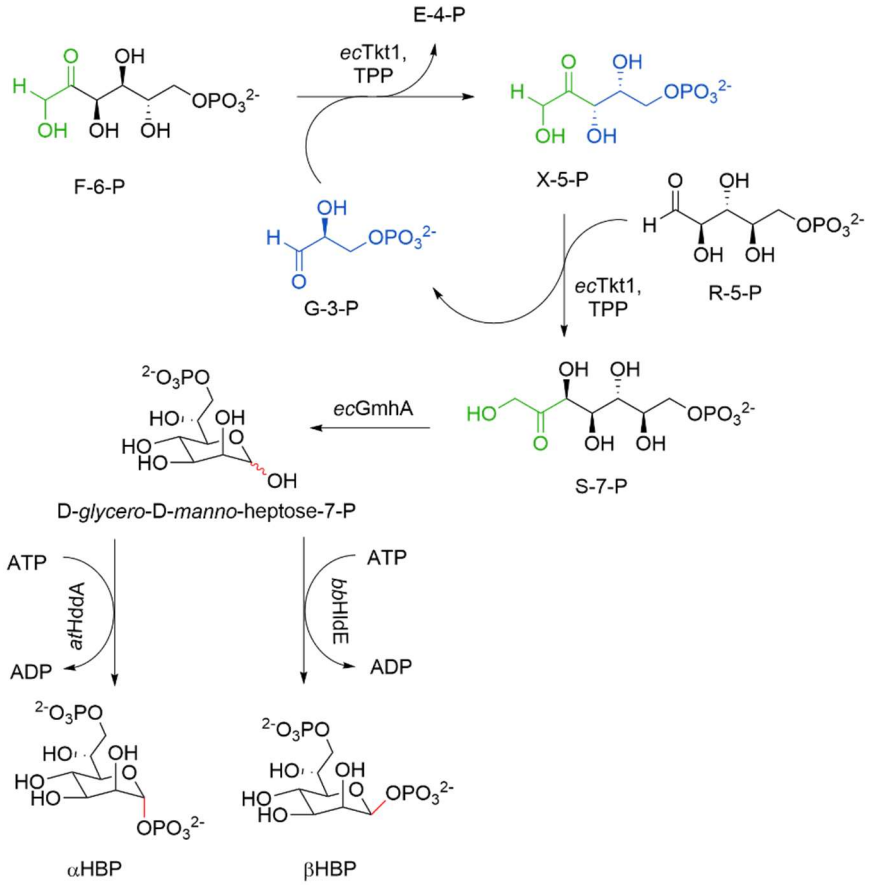

Figure 8.2: Enzymatic synthesis of αHBP and βHBP.

The two anomeric substrates αHBP and βHBP were synthesized enzymatically according to a modified protocol of Wang et al. [99]. The reaction was performed as one-pot reaction and the solution was incubated for 24 h at 37°C. Abbreviations: fructose-6-phosphate: F-6-P, thiamine pyrophosphate: TPP, G-3-P: glyceraldehyde-3-phosphate, erythrose-4-phosphate: E-4-P, xylulose-5-phosphate: X-5-P, ribose-5-phosphate: R-5-P, sedoheptulose-7-phosphate S-7-P, D-*glycero*-D-*manno*-heptose-7-phosphate: D-*glycero*-D-*manno*-heptose-7-P, D-*glycero*-D-*manno*-heptose-1α,7-bisphosphate: αHBP, D-*glycero*-D-*manno*-heptose-1β,7-bisphosphate: βHBP, transketolase from *E. coli*: *ec*Tkt1, ecGmhA, D-*glycero*-D-*manno*-heptose-7-phosphate kinase from *Bordetella bronchiseptica*: *bb*HldE, D-*glycero*-D-*manno*-heptose-7-phosphate kinase from *Aneurinibacillus thermoaerophilus*: *at*HddA.

The synthesis required the auxiliary enzymes transketolase (*ec*Tkt1) from *E. coli*, D-sedoheptulose-7-phosphate isomerase (*ec*GmhA) from *E. coli*, a β-C(1)OH specific D-*glycero*-D-*manno*-heptose-7-phosphate kinase from *Bordetella bronchiseptica* (*bb*HldE), an α C(1)OH specific D-*glycero*-D-*manno*-heptose-7-phosphate kinase from *Aneurinibacillus thermoaerophilus* (*at*HddA), and adenylate kinase from *E. coli* (*ec*ADK). The corresponding genes were codon optimized for *E. coli* (8.3.4.8), cloned into pUR23 (8.3.4.5), and the proteins were purified (8.4.2, 8.4.3, 8.4.4, 8.4.5). The commercially available substrates fructose-6-phosphate (97 mg), ribose-5-phosphate (110 mg), ATP (300 mg), thiamine pyrophosphate (50 mg), and D-glyceraldehyde-

3-phosphate (80 µL, 10 mM) were incubated with the enzymes *ec*Tkt1, *ec*GmhA, and *bb*HldE or *at*HddA at 37 °C for 24 h. Afterwards, *ec*ADK was added, to reduce the amount of ADP and thereby increase the purity of the final product and the reaction mixture incubated at 37°C for 16h. All enzymes were subsequently removed with a centrifugal filter (molecular weight cut-off 10 kDa) and the product was purified from the flowthrough by anion exchange chromatography (8.4.8.2).

8.4.8.2 Purification of αHBP and βHBP by anion exchange chromatography

The two anomeric sugars αHBP and βHBP were purified by anion exchange chromatography exploiting the fact that αHBP, βHBP, and ADP were the only compounds of the reaction mixture that carried two phosphate moieties. This was done with a Äkta Chromatography system, using a MonoQ HR 16/10 column. First, the chromatography system was equilibrated with sample buffer (50 mM NH_4HCO_3, pH 9.0), the pH of the reaction mixture was adjusted to 9.0 and the reaction mixture was subsequently applied to the column. After a wash step (3 CV, 50 mM NH_4HCO_3, pH 9.0) the final product was eluted by increasing concentrations of elution buffer (1.0 M NH_4HCO_3, pH 7.8). The concentration of the elution buffer was first increased in a linear gradient (2 CV, 0-25 % elution buffer), then maintained for 2 CV, and then increased in a second linear gradient (10 CV, 25-100 %). The final product showed no distinct absorbance spectrum therefore the elution process was monitored by absorbance measurement at 254 nm to detect the elution of AMP, ADP, and ATP. Fractions that contained the final product αHBP or βHBP were then identified by a coupled enzymatic assay (8.5.5) and the respective concentrations were determined by total turnover with *ec*GmhB. Fractions with high product concentrations and little contamination with ADP were pooled and stored at -20 °C.

8.5 Analytical methods

8.5.1 Determination of protein concentration by UV/Vis spectroscopy

Protein concentrations were determined by means of UV/Vis spectroscopy. The theoretical extinction coefficients (ε_{280}) were calculated form the protein sequence using the web tool ProtParam (Expasy) which relies on the formula shown in equation (1).

$$\varepsilon_{280} = 5500 \cdot \Sigma\,(Trp) + 1490 \cdot \Sigma\,(Tyr) + 125 \cdot \Sigma(Cystines)\ [M^{-1}cm^{-1}] \quad (1)$$

In absence of structural information from crystallography, all cysteine residues were assumed to be in their reduced state. Absorbance spectra were then recorded between 230 nm and 350 nm either with a spectrophotometer JASCO V-650, or JASCO V-750 or a NanoDrop One spectrophotometer. To account for the contribution of light scattering, the measured A_{280} value was corrected by subtracting double the A_{333} value according to equation (2) and the protein concentration was calculated using Lambert-Beer's law according to equation (3).

$$A_{280,corrected} = A_{280,measured} - 2 \cdot A_{333,measured} \quad (2)$$

$$c = \frac{A_{280,corrected}}{\varepsilon_{280} \cdot d} \quad (3)$$

8.5.2 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate proteins according to their molecular weight[160], estimate the abundancy of target proteins based on their molecular weight, and assess the purity of target proteins. The gels consisted of a resolving gel with 13.5 % acrylamide and a stacking gel on top of the resolving gel with 6 % acrylamide. The exact composition of stacking and resolving gel for a total of 13 gels is given in Table 8.6.

Table 8.6: Composition of standard gels for SDS PAGE

component	Stacking gel	Resolving gel
0.5 M Tris/HCl (pH 6.8)	7.38 mL	-
0.4 % SDS (w/v)		
1.5 M Tris/HCl (pH 8.8)	-	19.5 mL
0.4 % SDS (w/v)		
Acrylamide (30 %)	5.9 mL	26.2 mL
H_2O	15.95 mL	31.58 mL
TEMED	0.029 mL	0.089 mL
APS (10 %)	0.089 mL	0.195 mL

Protein samples were mixed in a ratio of 1:4 with 5x SDS sample buffer and incubated for 5 minutes at 95 °C. The samples were then loaded into the gel pockets and the electrophoresis was conducted at 50 mA and 300 V for 20-30 min. To visualize the resulting protein bands, gels were left for 5-10 minutes in a staining solution with Coomassie Brilliant Blue G-250. Afterwards, unbound dye was removed by several washing steps with boiling water until the protein bands became clearly visible.

8.5.3 Circular dichroism (CD) spectroscopy

Far-UV CD spectroscopy was used to assess the structural integrity of purified proteins and to determine melting temperatures of folded proteins. All measurements were performed with a J-815 CD spectrometer. Prior to measurements, the buffer was exchanged to 20 mM KP (pH 7.5) using NAP column (8.4.6) to minimize interfering buffer absorption.

To assess the structural integrity of a protein, spectra were recorded between 260 and 190 nm using a quartz cuvette (pathlength: 0.2-2 mm) in 5 to 8 replicas at 25 °C. All spectra were corrected for the signal that was caused by the buffer and smoothened using the Savitzky-Golay algorithm[161] with a convolution width of 7 to 11. 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 NA, according to equation (4).

$$\theta_{MRW} = \frac{\theta_{obs} \cdot 10^5}{c \cdot d \cdot N_A} \quad (4)$$

To determine the melting temperature of a protein, the ellipticity at 220 nm was monitored while the temperature was gradually increased from 25 °C to 95 °C at a rate of 1 °C per minute. When cooperative unfolding with constant θ_{obs} values for the folded and unfolded state were observed, the raw data was normalized in such a way that the mean θ_{obs} value for the folded protein was corresponded to 0.0, while the mean θ_{obs} value for the unfolded protein corresponded to 1.0. To obtain the melting temperature T_M, the fraction of unfolded protein f_u plotted against the temperature t was fit with logistic fit according to equation (5).

$$f_u = \frac{1}{1 + \left(\frac{t}{T_M}\right)^p} \tag{5}$$

8.5.4 Size exclusion chromatography followed by static light scattering (SEC-SLS)

Size exclusion chromatography followed by static light scattering (SEC-SLS) was used to estimate the molecular weights of proteins and deduce the oligomerization state of a target protein. For that purpose, an ÄKTA micro system with a Superdex 75 10/300 GL column and an ALIAS autosampler (Spark Holland) was coupled to a Viscotec TDA 305 detector (Malvern). Prior to sample application, the whole system was equilibrated with SEC buffer. All protein samples were diluted to a final concentration of 50 µM. As a reference, bovine serum albumin (BSA, 50 µM) was used which was diluted in SEC buffer. Then, the samples and the BSA reference were filled into glass vials, loaded onto the autosampler, and applied automatically onto the SEC column. The elution of the proteins from the column was monitored by absorption measurement at 280 nm. With the Viscotec TDA, small-angle light scattering, right-angle light scattering, and changes in the refractive index were recorded. Afterwards, the data was analyzed using the OmniSec software. First, the exact concentration of the BSA reference was calculated from the changes in the refractive index according to the instructions of the manufacturer. Then, the light scattering data that was recorded for the BSA reference in combination with the calculated concentration and the molecular weight of BSA were used to calibrate the software. After the calibration, the light scattering data of the different samples was used to calculate of the respective molecular weights.

8.5.5 Steady-state enzyme kinetic experiments

Steady-state kinetic experiments were conducted to determine the turnover number k_{cat} and the Michaelis constant K_M of an enzymatic reaction. The turnover of substrates was monitored by UV/Vis spectroscopy using a spectrophotometer (V-650 or V-750, JASCO) or a microplate reader (Infinite M200 Pro). In this work, all analyzed enzymes were phosphatases and therefore, all reactions could be monitored by quantification of free phosphate. To this end, a coupled enzymatic assay was employed which couples the dephosphorylation reaction to the oxidation of hypoxanthine to uric acid and thus yields an absorption signal at 293 nm (ε_{293} = 12600 $M^{-1}cm^{-1}$, Figure 8.3). [104]

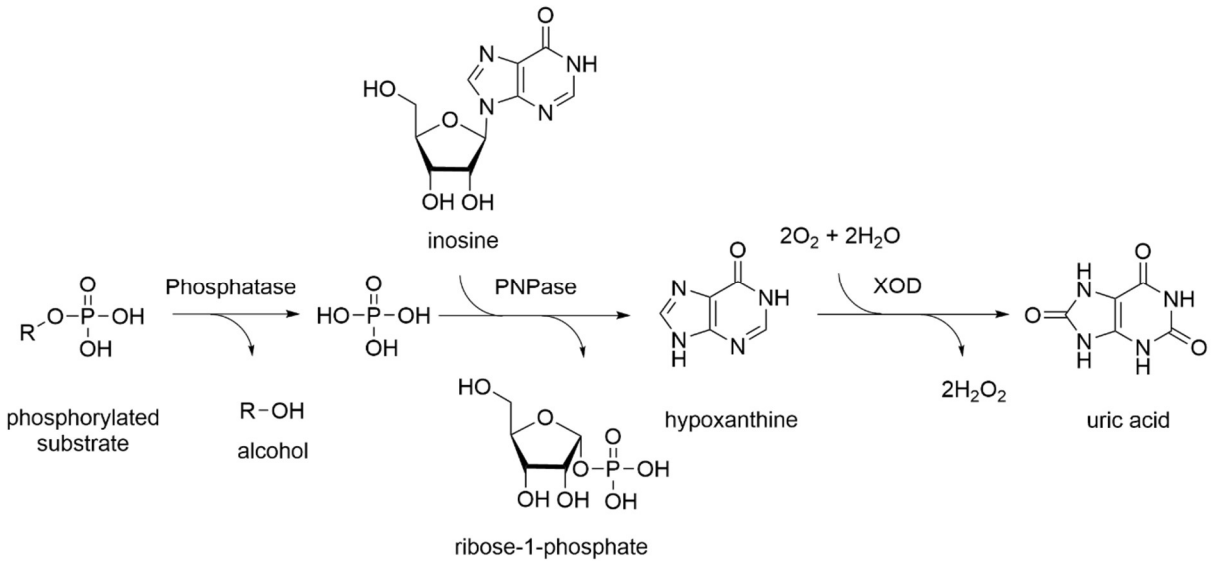

Figure 8.3: Coupled enzyme assay for the continuous measurement of phosphatase reactions.

The assay couples the dephosphorylation of a substrate to the oxidation of hypoxanthine to xanthine Which yields an absorption signal at 293 nm. This is achieved by the use of purine nucleoside phosphorylase (PNPase) and xanthine oxidase (XOD) as auxiliary enzymes and inosine as auxiliary substrate.[104]

The enzyme assay was measured in Tris/HCl buffer (100 mM, pH 7.8) at 25 °C. The reaction solution furthermore included the cofactor Mg^{2+} which was added as $MgCl_2$ (5 mM), the auxiliary substrate inosine (0.5 mM), the auxiliary enzymes purine nucleoside phosphorylase (0.25 U/mL) and xanthine oxidase (2.5 U/mL), and varying concentrations of substrate. To remove any traces of free phosphate from the reaction mixture which would obstruct an accurate measurement of the reaction rate, the reaction mixtures were pre-incubated prior to the addition of the enzyme of interest until a plateau in the absorption value was reached. Only then the reactions were started by enzyme addition. The final concentration of the target enzyme was chosen in such a way that the reaction rate of the target enzyme was at least 5-fold lower than the reaction rate of the coupled enzymes. This was done to ensure linearity of the assay and to ensure that the reaction rate of the coupled enzymes was not limiting. At the same time, the target enzyme concentration was chosen in such a way that the measured slopes at substrate saturation were several orders of magnitude higher than the slope of the baseline to ensure robust measurements and minimize the effect of random error of the instrumentation. Then, the reaction rate v_i was calculated from the initial ascending slopes ($\Delta A_{293}/\Delta t$), the extinction coefficient (ε_{293}), and the width of the cuvette according to Lambert Beer's law (equation 6).

$$v_i = \frac{\Delta c}{\Delta t} = \frac{\Delta A_{293}}{\Delta t \cdot \varepsilon_{293} \cdot d} \tag{6}$$

The reaction rate v_i was divided by the enzyme concentration E_0 and plotted against the substrate concentration $[S_i]$. Then, the kinetic parameters k_{cat} and K_M were calculated by means of curve fitting with the Michaelis-Menten equation (equation 7, Origin 2019, Origin 2021, OriginLab).

$$\frac{v_i}{E_0} = k_{cat} \cdot \frac{[S_i]}{[S_i] + K_M} \tag{7}$$

8.5.6 Discontinuous enzyme kinetic experiments

The discontinuous enzyme kinetic experiments were performed to detect very low enzymatic activities which were too low for reliable quantification with the continuous assay (8.5.5) or when potential side activities towards contaminating substances led to an overestimation of the primary activity. First, the substrate was incubated with the enzyme of interest (10 nm to 10 µM) for up to 20 h at 37 °C in Tris/HCl buffer (100 mM, pH 7.8) supplemented with $MgCl_2$ (5 mM). As a negative control, the substrate was incubated in the same buffer but in absence of enzyme. As positive control, the substrates were incubated with a wildtype enzyme that catalyzed its turnover, i.e., HolP was incubated with *ec*HisB-N while αHBP and βHBP were incubated with *ec*GmhB. After the incubation, the reactions were stopped by removal of the enzyme by means of filtration (pore size: 0.22 µM). The amount of free phosphate in the filtered reaction mixture was then determined by UV/Vis spectroscopy by adding inosine, purine nucleoside phosphorylase, and xanthin oxidase, applying the same assay as in the steady-state kinetic experiments (8.5.5). The difference in the plateaus in absorption prior and after the addition of inosine, purine nucleoside phosphorylase, and xanthin is then proportional to the amount of free phosphate of each sample. To correct for impurities or spontaneous hydrolysis of the substrate, the amount of free phosphate of the negative control experiment was subtracted from the other samples.

In the case of Anc1-Anc4, a side activity toward a contamination of βHBP was detected. Therefore, we chose to determine the amount of residual βHBP instead of the amount of free phosphate. To this end, βHBP was incubated with Anc1-Anc7 overnight followed by removal of the enzymes by filtration. Then, free phosphate that was formed by hydrolysis of βHBP and contaminant was removed from the reaction mix by adding inosine, purine nucleoside phosphorylase, and xanthine oxidase. The remaining βHBP concentration was finally determined by addition of *ec*GmhB, which was added as soon as plateau in the absorption was reached. The difference in the plateaus prior and after the addition of *ec*GmhB is proportional to the amount of residual βHBP. Subtraction of the amount of residual βHBP from the starting amount yields the amount which was hydrolyzed by Anc1-Anc7.

9 References

1. Koonin EV, Martin W (2005): On the origin of genomes and cells within inorganic compartments, *Trends Genet.* 21: 647–654.
2. Russell MJ, Hall AJ (1997): The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front, *J. Geol. Soc. London* 154: 377–402.
3. Woese CR (1967): The genetic code- The molecular basis for genetic expression. *Harper & Row.*
4. Gilbert W (1986): Origin of life: The RNA world, *Nature* 319: 618.
5. Miller SL (1953): A production of amino acids under possible primitive earth conditions, *Science* 117: 528–529.
6. Burton AS, Stern JC, Elsila JE, Glavin DP, Dworkin JP (2012): Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites, *Chem. Soc. Rev.* 41: 5459–5472.
7. Koonin EV, Novozhilov AS (2017): Origin and evolution of the universal genetic code, *Annu. Rev. Genet.* 51: 45–62.
8. Eigen M, Winkler-Oswatitsch R (1981): Transfer-RNA, an early gene?, *Sci. Nat.* 68: 282–292.
9. Crick FH (1968): The origin of the genetic code, *J. Mol. Biol.* 38: 367–379.
10. Woese CR (1965): On the evolution of the genetic code, *Proc. Natl. Acad. Sci. U. S. A.* 54: 1546–1552.
11. Bernhardt HS, Patrick WM (2014): Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids, *J. Mol. Evol.* 78: 307–309.
12. Akashi H, Gojobori T (2002): Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*, *Proc. Natl. Acad. Sci. U. S. A.* 99: 3695–3700.
13. Francis BR (2013): Evolution of the genetic code by incorporation of amino acids that improved or changed protein function, *J. Mol. Evol.* 77: 134–158.
14. Stouthamer AH (1973): A theoretical study on the amount of ATP required for synthesis of microbial cell material, *Antonie Van Leeuwenhoek* 39: 545–565.
15. Woese CR (1964): Universality in the genetic code, *Science* 144: 1030–1031.
16. Koonin EV, Novozhilov AS (2009): Origin and evolution of the genetic code: the universal enigma, *IUBMB Life* 61: 99–111.
17. Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D (2018): Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin, *Nat. Ecol. Evol.* 2: 1556–1562.
18. Kyrpides N, Overbeek R, Ouzounis C (1999): Universal protein families and the functional content of the last universal common ancestor, *J. Mol. Evol.* 49: 413–423.
19. Penny D, Poole A (1999): The nature of the last universal common ancestor, *Curr. Opin. Genet. Dev.* 9: 672–677.
20. Kim KM, Caetano-Anollés G (2011): The proteomic complexity and rise of the primordial ancestor of diversified life, *BMC Evol. Biol.* 11: 140.
21. Safran RJ, Nosil P (2012): Speciation: The Origin of New Species., *Nature Education Knowledge* 3: 17.
22. Noda-Garcia L, Liebermeister W, Tawfik DS (2018): Metabolite-Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks, *Annu. Rev. Biochem.* 87: 187–216.
23. Horowitz NH (1945): On the Evolution of Biochemical Syntheses, *Proc. Natl. Acad. Sci. U. S. A.* 31: 153–157.

24. Granick S (1965): Evolution of Heme and Chlorophyll. In: *Evolving Genes and Proteins*: 67–88
25. Ycas M (1974): On earlier states of the biochemical system, *J. Theor. Biol.* 44: 145–160.
26. Jensen RA (1976): Enzyme recruitment in evolution of new function, *Annu. Rev. Microbiol.* 30: 409–425.
27. Lazcano A, Miller SL (1999): On the origin of metabolic pathways, *J. Mol. Biol.* 49: 424–431.
28. Huang H, Pandya C, Liu C, Al-Obaidi NF, Wang M, Zheng L, Toews Keating S, Aono M, Love JD, Evans B, Seidel RD, Hillerich BS, Garforth SJ, Almo SC, Mariano PS, Dunaway-Mariano D, Allen KN, Farelli JD (2015): Panoramic view of a superfamily of phosphatases through substrate profiling, *Proc. Natl. Acad. Sci. U. S. A.* 112: E1974-83.
29. Keenan T, Parmeggiani F, Malassis J, Fontenelle CQ, Vendeville J-B, Offen W, Both P, Huang K, Marchesi A, Heyam A, Young C, Charnock SJ, Davies GJ, Linclau B, Flitsch SL, Fascione MA (2020): Profiling substrate promiscuity of wild-type sugar kinases for multi-fluorinated monosaccharides, *Cell Chem. Biol.* 27: 1199-1206.e5.
30. Ohno S (1970): Evolution by gene duplication. *Springer Verlag.*
31. Bergthorsson U, Andersson DI, Roth JR (2007): Ohno's dilemma: evolution of new genes under continuous selection, *Proc. Natl. Acad. Sci. U. S. A.* 104: 17004–17009.
32. Tawfik DS (2010): Messy biology and the origins of evolutionary innovations, *Nat. Chem. Biol.* 6: 692–696.
33. Copley SD (2020): The physical basis and practical consequences of biological promiscuity, *Phys. Biol.* 17: 51001.
34. Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN (2014): Enzyme promiscuity: engine of evolutionary innovation, *J. Biol. Chem.* 289: 30229–30236.
35. Glasner ME, Truong DP, Morse BC (2020): How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation, *FEBS J.* 287: 1323–1342.
36. Hao W, Golding GB (2006): The fate of laterally transferred genes: Life in the fast lane to adaptation or death, *Genome Res.* 16: 636–643.
37. Marri PR, Hao W, Golding GB (2007): The role of laterally transferred genes in adaptive evolution, *BMC Evol. Biol.* 7 Suppl 1: S8.
38. Yang G, Miton CM, Tokuriki N (2020): A mechanistic view of enzyme evolution, *Protein Sci.* 29: 1724–1747.
39. Copley SD (2020): Evolution of new enzymes by gene duplication and divergence, *FEBS J.* 287: 1262–1283.
40. Zeymer C, Hilvert D (2018): Directed evolution of protein catalysts, *Annu. Rev. Biochem.* 87: 131–157.
41. Lutz S, Iamurri SM (2018): Protein engineering: Past, present, and future, *Methods Mol. Biol.* 1685: 1–12.
42. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995): SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* 247: 536–540.
43. Knudsen M, Wiuf C (2010): The CATH database, *Hum. Genomics* 4: 207–212.
44. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997): CATH—a hierarchic classification of protein domain structures, *Structure* 5: 1093–1108.
45. Das S, Dawson NL, Orengo CA (2015): Diversity in protein domain superfamilies, *Curr. Opin. Genet. Dev.* 35: 40–49.
46. Chiang RA, Sali A, Babbitt PC (2008): Evolutionarily conserved substrate substructures for automated annotation of enzyme superfamilies, *PLoS Comput. Biol.* 4: E1000142.

47. Orengo CA, Thornton JM (2005): Protein families and their evolution - a structural perspective, *Annu. Rev. Biochem.* 74: 867–900.
48. Lee D, Grant A, Marsden RL, Orengo C (2005): Identification and distribution of protein families in 120 completed genomes using Gene3D, *Proteins* 59: 603–615.
49. Moras D (1992): Structural and functional relationships between aminoacyl-tRNA synthetases, *Trends Biochem. Sci.* 17: 159–164.
50. Aravind L, Mazumder R, Vasudevan S, Koonin EV (2002): Trends in protein evolution inferred from sequence and structure analysis, *Curr. Opin. Struct. Biol.* 12: 392–399.
51. Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L (2006): Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes, *J. Mol. Biol.* 361: 1003–1034.
52. Belogurov GA, Vassylyeva MN, Svetlov V, Klyuyev S, Grishin NV, Vassylyev DG, Artsimovitch I (2007): Structural basis for converting a general transcription factor into an operon-specific virulence regulator, *Mol. Cell* 26: 117–129.
53. Kobayashi T, Nureki O, Ishitani R, Yaremchuk A, Tukalo M, Cusack S, Sakamoto K, Yokoyama S (2003): Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion, *Nat. Struct. Biol.* 10: 425–432.
54. Parsons JF, Lim K, Tempczyk A, Krajewski W, Eisenstein E, Herzberg O (2002): From structure to function: YrbI from *Haemophilus influenzae* (HI1679) is a phosphatase, *Proteins* 46: 393–404.
55. Zhang L, Liu MR, Yao YC, Bostrom IK, Wang YD, Chen AQ, Li JX, Gu SH, Ji CN (2020): Characterization and structure of glyceraldehyde-3-phosphate dehydrogenase type 1 from *Escherichia coli*, *Acta Crystallogr., Sect. F: Struct. Biol. Commun.* 76: 406–413.
56. Kuzin AP, Edstrom W, Ma LC, Shin L, Xiao R, Acton TB, Montelione GT, Hunt JF, Tong L (2003): X-ray structure of Q8NW41 northeast structural genomics consortium target Zr25
57. Calderone V, Forleo C, Benvenuti M, Cristina Thaller M, Rossolini GM, Mangani S (2004): The first structure of a bacterial class B acid phosphatase reveals further structural heterogeneity among phosphatases of the haloacid dehalogenase fold, *J. Mol. Biol.* 335: 761–773.
58. Wang W, Kim R, Jancarik J, Yokota H, Kim SH (2001): Crystal structure of phosphoserine phosphatase from *Methanococcus jannaschii*, a hyperthermophile, at 1.8 A resolution, *Structure* 9: 65–71.
59. Rangarajan ES, Proteau A, Wagner J, Hung M-N, Matte A, Cygler M (2006): Structural snapshots of *Escherichia coli* histidinol phosphate phosphatase along the reaction pathway, *J. Biol. Chem.* 281: 37930–37941.
60. Joint Center for Structural Genomics (2004): Crystal structure of 4-nitrophenylphosphatase (TM1742) from *Thermotoga maritima* at 2.40 A resolution
61. Koonin EV, Tatusov RL (1994): 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.* 244: 125–132.
62. Aravind L, Galperin MY, Koonin EV (1998): The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold, *Trends Biochem. Sci.* 23: 127–129.
63. Collet JF, Stroobant V, Pirard M, Delpierre G, van Schaftingen E (1998): A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif, *J. Biol. Chem.* 273: 14107–14112.
64. Schramm M (1958): *O*-phosphoserine phosphatase from Baker's yeast, *J. Biol. Chem.* 233: 1169–1171.

65. Ames BN (1957): The biosynthesis of histidine; L-histidinol phosphate phosphatase, *J. Biol. Chem.* 226: 583–593.
66. Umbarger HE, Umbarger MA, Siu PM (1963): Biosynthesis of serine in *Escherichia coli* and *Salmonella typhimurium*, *J. Bacteriol.* 85: 1431–1439.
67. Makino Y, Sato T, Kawamura H, Hachisuka S-I, Takeno R, Imanaka T, Atomi H (2016): An archaeal ADP-dependent serine kinase involved in cysteine biosynthesis and serine metabolism, *Nat. Commun.* 7: 13446.
68. Ravnikar PD, Somerville RL (1987): Genetic characterization of a highly efficient alternate pathway of serine biosynthesis in Escherichia coli, *J. Bacteriol.* 169: 2611–2617.
69. Doig AJ (2017): Frozen, but no accident - why the 20 standard amino acids were selected, *FEBS J.* 284: 1296–1305.
70. Winkler ME, Ramos-Montañez S (2009): Biosynthesis of histidine, *EcoSal Plus* 3.6.1.9
71. Vázquez-Salazar A, Becerra A, Lazcano A (2018): Evolutionary convergence in the biosyntheses of the imidazole moieties of histidine and purines, *PLoS One* 13: e0196349.
72. Alifano P, Fani R, Liò P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB (1996): Histidine biosynthetic pathway and genes: structure, regulation, and evolution, *Microbiol. Rev.* 60: 44–69.
73. Ames BN, Garry B, Herzenberg LA (1960): The genetic control of the enzymes of histidine biosynthesis in *Salmonella typhimurium*, *J. Gen. Microbiol.* 22: 369–378.
74. Ames BN, Martin RG, Garry BJ (1961): The first step of histidine biosynthesis, *J. Biol. Chem.* 236: 2019–2026.
75. Smith DW, Ames BN (1964): Intermediates in the early steps of histidine biosynthesis, *J. Biol. Chem.* 239: 1848–1855.
76. Hartman PE, Loper JC, Serman D (1960): Fine structure mapping by complete transduction between histidine-requiring *Salmonella* mutants, *J. Gen. Microbiol.* 22: 323–353.
77. Carlomagno MS, Chiariotti L, Alifano P, Nappo AG, Bruni CB (1988): Structure and function of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons, *J. Mol. Biol.* 203: 585–606.
78. Stepansky A, Leustek T (2006): Histidine biosynthesis in plants, *Amino Acids* 30: 127–142.
79. Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R (2009): The evolution of histidine biosynthesis in archaea: insights into the his genes structure and organization in LUCA, *J. Mol. Evol.* 69: 512–526.
80. Fani R, Mori E, Tamburini E, Lazcano A (1998): Evolution of the structure and chromosomal distribution of histidine biosynthetic genes, *Orig. Life Evol. Biosph.* 28: 555–570.
81. Fani R, Liò P, Lazcano A (1995): Molecular evolution of the histidine biosynthetic pathway, *J. Mol. Evol.* 41: 760–774.
82. Fani R, Brilli M, Liò P (2005): The origin and evolution of operons: The piecewise building of the proteobacterial histidine operon, *J. Mol. Evol.* 60: 378–390.
83. Brilli M, Fani R (2004): Molecular evolution of *hisB* genes, *J. Mol. Evol.* 58: 225–237.
84. Chiariotti L, Nappo AG, Carlomagno MS, Bruni CB (1986): Gene structure in the histidine operon of *Escherichia coli*. Identification and nucleotide sequence of the *hisB* gene, *Mol. Gen. Genet.* 202: 42–47.
85. Brady DR, Houston LL (1973): Some properties of the catalytic sites of imidazoleglycerol phosphate dehydratase-histidinol phosphate phosphatase, a bifunctional enzyme from *Salmonella typhimurium*, *J. Biol. Chem.* 248: 2588–2592.

86. Chumley FG, Roth JR (1981): Genetic fusions that place the lactose genes under histidine operon control, *J. Mol. Biol.* 145: 697–712.
87. Fink GR (1964): Gene-enzyme relations in histidine biosynthesis in yeast, *Science* 146: 525–527.
88. Limauro D, Avitabile A, Cappellano C, Puglia AM, Bruni CB (1990): Cloning and characterization of the histidine biosynthetic gene cluster of *Streptomyces coelicolor* A3(2), *Gene* 90: 31–41.
89. Delorme C, Ehrlich SD, Renault P (1992): Histidine biosynthesis genes in *Lactococcus lactis* subsp. *lactis*, *J Bacteriol* 174: 6571–6579.
90. Kulis-Horn RK, Rückert C, Kalinowski J, Persicke M (2017): Sequence-based identification of inositol monophosphatase-like histidinol-phosphate phosphatases (HisN) in *Corynebacterium glutamicum*, Actinobacteria, and beyond, *BMC Microbiol.* 17: 161.
91. Ghodge SV, Fedorov AA, Fedorov EV, Hillerich B, Seidel R, Almo SC, Raushel FM (2013): Structural and mechanistic characterization of L-histidinol phosphate phosphatase from the polymerase and histidinol phosphatase family of proteins, *Biochemistry* 52: 1101–1112.
92. Petersen LN, Marineo S, Mandalà S, Davids F, Sewell BT, Ingle RA (2010): The missing link in plant histidine biosynthesis: *Arabidopsis* myoinositol monophosphatase-like2 encodes a functional histidinol-phosphate phosphatase, *Plant Physiol.* 152: 1186–1196.
93. Ruszkowski M, Dauter Z (2016): 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.* 291: 9960–9973.
94. Jha B, Kumar D, Sharma A, Dwivedy A, Singh R, Biswal BK (2018): Identification and structural characterization of a histidinol phosphate phosphatase from *Mycobacterium tuberculosis*, *J. Biol. Chem.* 293: 10102–10118.
95. Mormann S, Lömker A, Rückert C, Gaigalat L, Tauch A, Pühler A, Kalinowski J (2006): Random mutagenesis in *Corynebacterium glutamicum* ATCC 13032 using an IS6100-based transposon vector identified the last unknown gene in the histidine biosynthesis pathway, *BMC Genom.* 7: 205.
96. Valvano MA, Messner P, Kosma P (2002): Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides, *Microbiology (Reading, Engl.)* 148: 1979–1989.
97. Kneidinger B, Marolda C, Graninger M, Zamyatina A, McArthur F, Kosma P, Valvano MA, Messner P (2002): Biosynthesis pathway of ADP-L-*glycero*-β-D-*manno*-heptose in *Escherichia coli*, *J. Bacteriol.* 184: 363–369.
98. Kneidinger B, Graninger M, Puchberger M, Kosma P, Messner P (2001): Biosynthesis of nucleotide-activated D-*glycero*-D-*manno*-heptose, *J. Biol. Chem.* 276: 20935–20944.
99. Wang L, Huang H, Nguyen HH, Allen KN, Mariano PS, Dunaway-Mariano D (2010): Divergence of biochemical function in the HAD superfamily: D-*glycero*-D-*manno*-heptose-1,7-bisphosphate phosphatase (GmhB), *Biochemistry* 49: 1072–1081.
100. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021): Highly accurate protein structure prediction with AlphaFold, *Nature* 596: 583–589.
101. Sander C, Schneider R (1991): Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins* 9: 56–68.

102. Nguyen HH, Wang L, Huang H, Peisach E, Dunaway-Mariano D, Allen KN (2010): Structural determinants of substrate recognition in the HAD superfamily member D-*glycero*-D-*manno*-heptose-1,7-bisphosphate phosphatase (GmhB), *Biochemistry* 49: 1082–1092.
103. Taylor PL, Sugiman-Marangos S, Zhang K, Valvano MA, Wright GD, Junop MS (2010): Structural and kinetic characterization of the LPS biosynthetic enzyme D-α,β-D-heptose-1,7-bisphosphate phosphatase (GmhB) from *Escherichia coli*, *Biochemistry* 49: 1033–1041.
104. Suárez ASG, Stefan A, Lemma S, Conte E, Hochkoeppler A (2012): Continuous enzyme-coupled assay of phosphate- or pyrophosphate-releasing enzymes, *BioTechniques* 53: 99–103.
105. Nourbakhsh A, Collakova E, Gillaspy GE (2014): Characterization of the inositol monophosphatase gene family in *Arabidopsis*, *Front. Plant Sci.* 5: 725.
106. Kuznetsova E, Proudfoot M, Gonzalez CF, Brown G, Omelchenko MV, Borozan I, Carmel L, Wolf YI, Mori H, Savchenko AV, Arrowsmith CH, Koonin EV, Edwards AM, Yakunin AF (2006): Genome-wide analysis of substrate specificities of the *Escherichia coli* haloacid dehalogenase-like phosphatase family, *J. Biol. Chem.* 281: 36149–36161.
107. Kanehisa M, Goto S (2000): KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.* 28: 27–30.
108. Bonthron DT, Jaskólski M (1997): Why a "benign" mutation kills enzyme activity. Structure-based analysis of the A176V mutant of *Saccharomyces cerevisiae* L-asparaginase I, *Acta Biochim. Pol.* 44: 491–504.
109. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros P-A, Tenaillon O (2013): Capturing the mutational landscape of the beta-lactamase TEM-1, *Proc. Natl. Acad. Sci. U. S. A.* 110: 13067–13072.
110. Waite DW, Chuvochina M, Pelikan C, Parks DH, Yilmaz P, Wagner M, Loy A, Naganuma T, Nakai R, Whitman WB, Hahn MW, Kuever J, Hugenholtz P (2020): Proposal to reclassify the proteobacterial classes δ-Proteobacteria and Oligoflexia, and the phylum Thermodesulfobacteria into four phyla reflecting major functional capabilities, *Int. J. Syst. Evol. Microbiol.* 70: 5972–6016.
111. Babtie A, Tokuriki N, Hollfelder F (2010): What makes an enzyme promiscuous?, *Curr. Opin. Chem. Biol.* 14: 200–207.
112. Peracchi A (2018): The limits of enzyme specificity and the evolution of metabolism, *Trends Biochem. Sci.* 43: 984–996.
113. Del Duca S, Chioccioli S, Vassallo A, Castronovo LM, Fani R (2020): The role of gene elongation in the evolution of histidine biosynthetic genes, *Microorganisms* 8: 732.
114. Ochman H, Lawrence JG, Groisman EA (2000): Lateral gene transfer and the nature of bacterial innovation, *Nature* 405: 299–304.
115. Kinateder T, Drexler L, Straub K, Merkl R, Sterner R (2023): Experimental and computational analysis of the ancestry of an evolutionary young enzyme from histidine biosynthesis, *Protein Sci.* 32: E4536.
116. Wang Y, Wang L, Zhang J, Duan X, Feng Y, Wang S, Shen L (2020): PA0335, a gene encoding histidinol phosphate phosphatase, mediates histidine auxotrophy in Pseudomonas aeruginosa, *Appl. Environ. Microbiol.* 86: E02593-19.
117. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990): Basic local alignment search tool, *J Mol Biol* 215: 403–410.
118. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T,

Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST (2022): Database resources of the national center for biotechnology information, *Nucleic Acids Res.* 50: D20-26.

119. LaBauve AE, Wargo MJ (2012): Growth and laboratory maintenance of *Pseudomonas aeruginosa*, *Curr. Protoc. Microbiol.* Chapter 6: Unit 6E.1.
120. Wiater A, Krajewska-Grynkiewicz K, Klopotowski T (1971): Histidine biosynthesis and its regulation in higher plants, *Acta Biochim. Pol.* 18: 299–307.
121. Wang W, Cho HS, Kim R, Jancarik J, Yokota H, Nguyen HH, Grigoriev IV, Wemmer DE, Kim S-H (2002): Structural characterization of the reaction pathway in phosphoserine phosphatase: Crystallographic "snapshots" of intermediate states, *J. Mol. Biol.* 319: 421–431.
122. Cho H, Wang W, Kim R, Yokota H, Damo S, Kim SH, Wemmer D, Kustu S, Yan D (2001): 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.* 98: 8525–8530.
123. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004): WebLogo: a sequence logo generator, *Genome Res.* 14: 1188–1190.
124. Zallot R, Oberg N, Gerlt JA (2019): The EFI web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways, *Biochemistry* 58: 4169–4182.
125. Oberg N, Zallot R, Gerlt JA (2023): EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme function initiative (EFI) web resource for genomic enzymology tools, *J. Mol. Biol.*: 168018.
126. Steinegger M, Söding J (2018): Clustering huge protein sequence sets in linear time, *Nat. Commun.* 9: 2542.
127. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003): Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome Res.* 13: 2498–2504.
128. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A (2007): UniProtKB/Swiss-Prot, *Methods Mol. Biol.* 406: 89–112.
129. Arora G, Tiwari P, Mandal RS, Gupta A, Sharma D, Saha S, Singh R (2014): High throughput screen identifies small molecule inhibitors specific for *Mycobacterium tuberculosis* phosphoserine phosphatase, *J. Biol. Chem.* 289: 25149–25165.
130. Panda A, Islam ST, Sharma G (2022): Harmonizing prokaryotic nomenclature: Fixing the fuss over phylum name flipping, *mBio* 13: E0097022.
131. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I (2020): NCBI taxonomy: A comprehensive update on curation, resources and tools, *DATABASE-OXFORD*
132. Lee HS, Cho Y, Lee J-H, Kang SG (2008): Novel monofunctional histidinol-phosphate phosphatase of the DDDD superfamily of phosphohydrolases, *J. Bacteriol.* 190: 2629–2632.
133. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, Mering C von (2023): The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest, *Nucleic Acids Res.* 51: D638-646.
134. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A (2021): Pfam: The protein families database in 2021, *Nucleic Acids Res.* 49: D412-419.

135. Walker CB, La Torre JR de, Klotz MG, Urakawa H, Pinel N, Arp DJ, Brochier-Armanet C, Chain PSG, Chan PP, Gollabgir A, Hemp J, Hügler M, Karr EA, Könneke M, Shin M, Lawton TJ, Lowe T, Martens-Habbena W, Sayavedra-Soto LA, Lang D, Sievert SM, Rosenzweig AC, Manning G, Stahl DA (2010): *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea, *Proc. Natl. Acad. Sci. U. S. A.* 107: 8818–8823.
136. Castelle CJ, Banfield JF (2018): Major new microbial groups expand diversity and alter our understanding of the tree of life, *Cell* 172: 1181–1197.
137. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I (2007): Multicopy suppression underpins metabolic evolvability, *Mol. Biol. Evol.* 24: 2716–2722.
138. Yip SH-C, Matsumura I (2013): Substrate ambiguous enzymes within the *Escherichia coli* proteome offer different evolutionary solutions to the same problem, *Mol. Biol. Evol.* 30: 2001–2012.
139. Ruperti F (2018): Characterization and comparison of homologous enzymes of serine and histidine biosynthesis. Master thesis, University of Regensburg.
140. Rohweder B, Lehmann G, Eichner N, Polen T, Rajendran C, Ruperti F, Linde M, Treiber T, Jung O, Dettmer K, Meister G, Bott M, Gronwald W, Sterner R (2019): Library Selection with a Randomized Repertoire of (βα)8-Barrel Enzymes Results in Unexpected Induction of Gene Expression, *Biochemistry* 58: 4207–4217.
141. Reetz MT, Kahakeaw D, Lohmer R (2008): Addressing the numbers problem in directed evolution, *ChemBioChem* 9: 1797–1804.
142. Neuenschwander M, Butz M, Heintz C, Kast P, Hilvert D (2007): A simple selection strategy for evolving highly efficient enzymes, *Nat. Biotechnol.* 25: 1145–1147.
143. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF (2016): A new view of the tree of life, *Nat Microbiol* 1: 16048.
144. Brenner M, Ames BN (1971): The histidine operon and its regulation. In: *Metabolic Regulation*: 349–387
145. Shen C, Yang L, Miller SL, Oro J (1990): Prebiotic synthesis of histidine, *J. Mol. Evol.* 31: 167–174.
146. Drexler L (2021): Biochemical analysis of related enzymes from the HAD superfamily. Master thesis, University of Regensburg.
147. Marineo S, Cusimano MG, Limauro D, Coticchio G, Puglia AM (2008): The histidinol phosphate phosphatase involved in histidine biosynthetic pathway is encoded by SCO5208 (*hisN*) in *Streptomyces coelicolor* A3(2), *Curr. Microbiol.* 56: 6–13.
148. Datsenko KA, Wanner BL (2000): One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products, *Proc. Natl. Acad. Sci. U. S. A.* 97: 6640–6645.
149. Rohweder B, Semmelmann F, Endres C, Sterner R (2018): Standardized cloning vectors for protein production and generation of large gene libraries in *Escherichia coli*, *BioTechniques* 64: 24–26.
150. Murphy KC (1998): Use of bacteriophage λ recombination functions to promote gene replacement in *Escherichia coli*, *J. Bacteriol.* 180: 2063–2071.
151. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL (2000): An efficient recombination system for chromosome engineering in *Escherichia coli*, *Proc. Natl. Acad. Sci. U. S. A.* 97: 5978–5983.

152. Sawitzke JA, Thomason LC, Costantino N, Bubunenko M, Datta S, Court DL (2007): Recombineering: In vivo genetic engineering in *E. coli*, *S. enterica*, and beyond, *Methods Enzymol.* 421: 171–199.
153. Inoue H, Nojima H, Okayama H (1990): High efficiency transformation of *Escherichia coli* with plasmids, *Gene* 96: 23–28.
154. Sharp PA, Sugden B, Sambrook J (1973): Detection of two restriction endonuclease activities in *Haemophilus parainfluenzae* using analytical agarose—ethidium bromide electrophoresis, *Biochemistry* 12: 3055–3063.
155. Mullis KB, Faloona FA (1987): Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction, *Methods Enzymol.* 155: 335–350.
156. Vovis GF, Lacks S (1977): Complementary action of restriction enzymes endo R-*Dpn*I and endo R-*Dpn*II on bacteriophage f1 DNA, *J. Mol. Biol.* 115: 525–538.
157. Engler C, Kandzia R, Marillonnet S (2008): A one pot, one step, precision cloning method with high throughput capability, *PLoS One* 3: e3647.
158. Keiler KC, Waller PR, Sauer RT (1996): Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA, *Science* 271: 990–993.
159. Reetz MT, Wang L-W, Bocola M (2006): Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space, *Angew. Chem. Int. Ed. Engl.* 45: 1236–1241.
160. Laemmli UK (1970): Cleavage of structural proteins during the assembly of the head of bacteriophage T4, *Nature* 227: 680–685.
161. Savitzky A, Golay MJE (1964): Smoothing and differentiation of data by simplified least squares procedures, *Anal. Chem.* 36: 1627–1639.

Acronyms

Å	Ångström
Amp	ampicillin
Anc1	first resurrected ancestral enzyme
AP	Acid phosphatase family
ATP	Adenosine triphosphate
ATPase	Adenosine 5'-triphosphatase
Cam	Chloramphenicol
CD	Circular dichroism
cs	*Crassaminicella sp.* enzyme
CV	Column volume(s)
ddH_2O	Double-distilled water
E. coli	Escherichia coli
ec	*E. coli* enzyme
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GmhB	D-*glycero*-D-*manno*-heptose-1,7-bisphosphate phosphatase
h	hour
HAD superfamily	Haloacid dehalogenase superfamily
HBP	D-*glycero*-D-*manno*-heptose-1,7-bisphosphate
HisB	imidazoleglycerol-phosphate dehydratase
HisB-N	*E. coli* type HolPase
Hol	Histidinol
HolP	Histidinol phosphate
HolPase	Histidinol phosphate phosphatase
IAD model	Innovation-amplification-divergence model
IAP	imidazole acetol-phosphate
IGP	imidazole glycerol-phosphate
IGPDH	imidazoleglycerol-phosphate dehydratase
IMP superfamily	Inositol-monophosphate phosphatase superfamily
Kana	Kanamycin
k_{cat}	Turnover number
k_{cat}/K_M	Kinetic efficiency

K_M	Michaelis constant
L. lactis	*Lactococcus lactis*
LB	Luria Bertani medium
LB_0	Luria Bertani medium without antibiotic
LUCA	Last universal common ancestor
M	mol/L
M. truncatula	*Medicago truncatula*
M. tuberculosis	*Mycobacterium tuberculosis*
Mn	Number average molar mass
Mw	Mass average molar mass
n	10^{-9}
NAD^+	Oxidized form of nicotinamide adenine dinucleotide
NADH	Reduced form of nicotinamide adenine dinucleotide
PAGE	Polyacrylamide gel electrophoresis
PHP superfamily	Polymerase and histidinol phosphatase superfamily
Pi	phosphate
PPi	pyrophosphate
PSP	*o*-Phospho-L-serine
PSPase	Phosphoserine phosphatase
RMSD	Root mean square deviation
RNA	Ribonucleic acid
rpm	Rotations per minute
s	second
SDS	Sodium dodecyl sulfate
T_M	Melting temperature
tRNA	Transfer ribonucleic acid
wt	wildtype
αGmhB	GmhB with αHBP as native substrate
αGmhB, βGmhB	GmhB with αHBP or βHBP as native substrate
αHBP	D-*glycero*-D-*manno*-heptose-1α,7-bisphosphate
αHBP	D-*glycero*-D-*manno*-heptose-1α,7-bisphosphate
α-KG	A-ketoglutarate

βGmhB	GmhB with βHBP as native substrate
βHBP	D-*glycero*-D-*manno*-heptose-1β,7-bisphosphate
βHBP	D-*glycero*-D-*manno*-heptose-1β,7-bisphosphate
$\gamma^{+\text{HisB-N}}$-Proteobacteria	Proteobacteria with β*gmhB* and *hisB-N* genes
$\gamma_{\text{-HisB-N}}$-Proteobacteria	Proteobacteria with a β*gmhB* gene but without a *hisB-N* gene
μ	10^{-6}

List of Figures

Figure 1.1: The innovation-amplification-divergence (IAD) model of evolution. 3
Figure 1.2: Structural diversity within the HAD Superfamily. 6
Figure 1.3: Catalytic machinery of the HAD enzymes. 7
Figure 1.4: The histidine biosynthesis and the histidinol phosphate phosphatase reaction. 9
Figure 1.5: Structural diversity of different HolPases. 10
Figure 2.1: Structural and functional comparison of *ec*HisB-N and *ec*GmhB. 13
Figure 2.2: Functional characterization of *ec*HisB-N at 25 °C. 15
Figure 2.3: Functional characterization of *ec*GmhB at 25°C. 16
Figure 2.4: Phylogenetic analysis of αGmhB, βGmhB, and HisB-N enzymes and HolPase activities of two βGmhB representatives. 18
Figure 2.5: Ancestral sequence reconstruction and functional characterization of resurrected progenitor enzymes. 20
Figure 2.6: Structural analysis of fold, shape, and charge distribution of the active sites of Anc1-Anc7. 22
Figure 2.7: Phylogenetic distribution of *hisB-N* and *βgmhB* genes, and revised evolutionary model. 24
Figure 3.1: Comparison of the predicted structure of *pa*HisN with other HAD enzymes. 29
Figure 3.2: Biophysical characterization of *pa*HisN. 31
Figure 3.3: Functional characterization of *pa*HisN at 25°C. 32
Figure 3.4: Alanine Scan of active site residues of *pa*HisN. 34
Figure 3.5: Comparison of the *pa*HisN sequence with a sequence logo of PSPases. 37
Figure 3.6: *In silico* analysis of the homologues of *pa*HisN. 39
Figure 3.7: Phylogenetic distribution of the different non-homologous HolPases. 42
Figure 4.1: Comparison of gene loci of histidine biosynthesis of different species. 46
Figure 4.2: Comparison of the structure of *nm*HisN to other HolPases of the HAD superfamily. 48
Figure 4.3: Biophysical characterization of *nm*HisN. 50
Figure 4.4: Functional characterization of *nm*HisN at 25°C. 51
Figure 4.5: Alanine Scan of active site residues of *nm*HisN. 52
Figure 4.6: *In silico* analysis of the homologues of *nm*HisN. 56
Figure 5.1: Analogous reactions in the biosynthesis of L-serine and L-histidine. 59
Figure 5.2: Establishing HolPase activity on *ec*SerB. 62
Figure 5.3: Improving the promiscuous PSPase activity of *ec*HisB-N. 66
Figure 6.1: Mapping of the different HolPases on an abstracted tree of life. 70
Figure 6.2: Evolutionary model for the early evolution of the HolPase function. 72
Figure 8.1: Three-step construction of focused gene libraries. 99
Figure 8.2: Enzymatic synthesis of αHBP and βHBP. 103
Figure 8.3: Coupled enzyme assay for the continuous measurement of phosphatase reactions. 107

List of Tables

Table 2.1: Activity towards HolP, βHBP, and αHBP of *ec*HisB-N, *ec*GmhB, and *cs*GmhB at 25 °C. 16
Table 2.2: Activity towards HolP of Anc1-Anc7 and *ec*HisB-N at 25°C. 21
Table 3.1: Catalytic parameters of *pa*HisN single mutants to alanine at 25°C. 35
Table 4.1: Catalytic parameters of *nm*HisN single mutants to alanine at 25°C..... 53
Table 6.1: Pairwise sequence alignments for the different HolPases from the HAD superfamily..... 69
Table 7.1: Plasmids used in this work. 82
Table 7.2: Gene sequences of experimentally characterized proteins. 83
Table 8.1: Standard PCR reaction mix. 96
Table 8.2: Standard protocol for PCR reactions. 96
Table 8.3: Standard reaction mix for coupled digestion-ligation reactions. 98
Table 8.4: Standard reaction conditions for coupled digestion-ligation reactions 98
Table 8.5: Standard settings for immobilized metal affinity chromatography. 101
Table 8.6: Composition of standard gels for SDS PAGE..... 105

Supplementary tables and figures

Table S 1: Experimentally characterized wildtype proteins.

The table lists the designation of the characterized proteins and their amino acid sequences. The His_6-tag and the sequence of the maltose binding protein (MBP) are printed in blue.

Construct	Protein	Amino acid sequence
pUR23_ *echisB-N*	*ec*HisB-N	MHHHHHHLDMSQKYLFIDRDGTLISEPPSDFQVDRFDKLAFEP GVIPELLKLQKAGYKLVMITNQDGLGTQSFPQADFDGPHNLM MQIFTSQGVQFDEVLICPHLPADECDCRKPKVKLVERYLAEQA MDRANSYVIGDRATDIQLAENMGITGLRYDRETLNWPMIGEQL TRRD
pCA24N_ *ecgmhB*	*ec*GmhB	MRGSHHHHHHTDPALRAAKSVPAIFLDRDGTINVDHGYVHEID NFEFIDGVIDAMRELKKMGFALVVVTNQSGIARGKFTEAQFET LTEWMDWSLADRDVDLDGIYYCPHHPQGSVEEFRQVCDCRKP HPGMLLSARDYLHIDMAASYMVGDKLEDMQAAVAANVGTK VLVRTGKPITPEAENAADWVLNSLADLPQAIKKQQKPAQGLCG R
pUR23_ *csgmhB*	*cs*GmhB	MHHHHHHLDMTKRAVFLDRDGTLIVDHGYIHKPSQVELLPGVI EALIKLKTFGFELIIISNQSGIGRGFFTKKEVDHVNQHLYNLLISH KIKLTGIYYCPHHPDDKCTCRKPEPGLLLQALSEHKIDAKKSYF VGDKLTDVQAAIAAGVQPVLLSRDNVSTHTIPIIIVDSLLKFTKV IKQEDF
pMal_*anc1*	Anc1	MHHHHHHMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIK VTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLA EITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDL LPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAAD GGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMN ADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLP TFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEA VNKDKPLGAVALKSYEEELVKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNN NNNNPGLVPRGSHMRRYVFLDRDGTINEDVYLSDPEKVRLLPG AAEGLKLLQEAGFALVVVTNQSGIARGYFSEESVHAVHARMQ KLLAPHGVQIDGIYFCPHHPEENCDCRKPSPGMVLQAARELGID PSRSYVIGDRESDIELARNVGAKSVLVLTGKGQEQPDLVAEDLL DAARLILK
pMal_*anc2*	Anc2	MHHHHHHMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIK VTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLA EITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDL LPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAAD GGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMN

		ADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLP TFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEA VNKDKPLGAVALKSYEEELVKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNN NNNNPGLVPRGSHMPVVFLDRDGTINEEVYLNNPEKVRLLPGV AEALRLLKEAGFKLVVITNQSGIARGYFSEEEVHAVHQRMMKR LATHGVQIDDIYFCPHHPEENCDCRKPKPGLVLKAAQKWNIDL SRSYVIGDRDTDIELAFNAGCKGVLVLTGYGKQLPDVAKDLLD AAKLILK
pMal_*anc3*	Anc3	MHHHHHHMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIK VTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLA EITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDL LPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAAD GGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMN ADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLP TFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEA VNKDKPLGAVALKSYEEELVKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNN NNNNPGLVPRGSHMQPVVFLDRDGTINEEVYLNNLSRLRLLPG VAEAIRLLKKAGFAVVVITNQSGLARGYFPEDTLHAVHQKML KRLSTRGATIDGIYVCPHHPEENCECRKPKPGLVLKASQELKLD LSRAYVIGDRDTDIQLAKNIGAKGVLVLTGYGEGELPDLVAKD LLSAAKKILK
pMal_*anc4*	Anc4	MHHHHHHMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIK VTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLA EITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDL LPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAAD GGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMN ADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLP TFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEA VNKDKPLGAVALKSYEEELVKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNN NNNNPGLVPRGSHMRPVVFLDRDGTINEEMYINHLSRLRLLPG VAQALRLLQEAGFKLVIVTNQSGPARGYFPEELVHETHQMLQR RLAAQGVKLDDLYVCLHHPEEGCSCRKPRPGLVLQALEEYSID LERSYVIGDRWVDLELARNMGIRGILVLTGYGRGELEPDVAKD WKSAAELILK
pMal_*anc5*	Anc5	MHHHHHHMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIK VTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLA EITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDL LPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAAD GGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMN ADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLP

		TFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEA VNKDKPLGAVALKSYEEELVKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNN NNNNPGLVPRGSHMQPILFIDRDGTLVEEPPTDQVDSLDKLEFE PAVIPALRKLQNAGYRLVMVSNQDGLGTPSFPQESFEAPHNMM MDIFASQGITFDEVYICPHFPEDNCDCRKPRTGLVTNYLRERRF DPQRSYVIGDRETDMQLAENMGTEGIQYRPGGLDWPAIAERLL
pUR23_ *anc6*	Anc6	MHHHHHHLDMSQKILFIDRDGTLIEEPPTDFQVDSLEKLKFEPA VIPALLKLQNAGYRLVMVTNQDGLGTDSFPQADFDPPHNLMM QIFESQGIRFDEVLICPHFPEDNCSCRKPKTGLVTRYLKEGKIDK ERSYVIGDRETDMQLAENMGIRGLQYNPNLNWEAITEQLTN
pUR23_ *anc7*	Anc7	MHHHHHHLDMSQKILFIDRDGTLITEPPTDFQVDSLEKLAFEPA VIPALLKLQKAGYRLVMVTNQDGLGTDSFPQADFDPPHNLMM QIFESQGIRFDEVLICPHFPADNCSCRKPKTGLVTHYLQEGKIDF ERSYVIGDRETDIQLAENMGIQGLRYSPELDWAAITHQLT
pUR23_ *pahisN*	*pa*HisN	MHHHHHHLDMRLALFDLDNTLLAGDSDHSWGEWLCQRGLVD AAEYQARNDAFYADYVAGKLDVLAYQAFTQAILGRTEMAQL ETWHRQFMQEVIEPIVLAKGEALLAEHRAAGDRLVIITATNRFV TGPIAERLGVETLIATECEMRDGRYTGQTFDVPCFQGGKVVRL QRWLDENGLDLEGASFYSDSLNDLPLLEKVSRPVAVDPDPRLR AEAEKRGWPIISLR
pUR23_ *nmhisN*	*nm*HisN	MHHHHHHLDMDSQKFDSIIFDCDGVLVDITQSYDKTIDKTCRY VLKEFAKIDSITIDHKIIDGFKSSGGFNDEVDLVYAAILSLYTAN KLNKKPSEFIYDVISNTDKTGIRSVQSYLESIYDVSEFLSKLGSL GDRHNNPVYSIFDQFFFGPELYGKLFDKQSKFSEEGMISNDKVI LSVSLLETLQKEFGKKIAVVTGRGIESIRYSLKDMMDYFDTKNS AFLEDEPRELAKPNPATLIRAIQSMESKNCLYVGDSMEDYMMA KDAAQAGHSTTFCAIVGTSTNPEDRRKLFADSGVEMILESINDIP KVLNLV
pUR22_ *ecserB*	*ec*SerB	MPNITWCDLPEDVSLWPGLPLSLSGDEVMPLDYHAGRSGWLL YGRGLDKQRLTQYQSKLGAAMVIVAAWCVEDYQVIRLAGSLT ARATRLAHEAQLDVAPLGKIPHLRTPGLLVMDMDSTAIQIECID EIAKLAGTGEMVAEVTERAMRGELDFTASLRSRVATLKGADA NILQQVRENLPLMPGLTQLVLKLETLGWKVAIASGGFTFFAEY LRDKLRLTAVVANELEIMDGKFTGNVIGDIVDAQYKAKTLTRL AQEYEIPLAQTVAIGDGANDLPMIKAAGLGIAYHAKPKVNEKA EVTIRHADLMGVFCILSGSLNQKLEHHHHHH

Table S 2: Median marginal ancestral probabilities calculated across all residues.

	Anc1	Anc2	Anc3	Anc4	Anc5	Anc6	Anc7
Median	1.0	0.999	0.999	0.998	1.0	1.0	1.0

Table S 3: Alignment of the lid-like loop sequences following the DxD motif.

Acidic residues are shown in red, residues with a hydroxyl-group in green, and aromatic residues in blue. Compared to Anc5, this loop is four residues shorter in the most primordial predecessors Anc1-Anc4 and one residue longer in the successors Anc6 and Anc7.

	Position in the MSA								
	15	16	17	18	19	20	21	22	23
*ec*GmhB	V	D	-	-	-	H	G	Y	V
Anc1	E	D	-	-	-	-	V	Y	L
Anc2	E	E	-	-	-	-	V	Y	L
Anc3	E	E	-	-	-	-	V	Y	L
Anc4	E	E	-	-	-	-	M	Y	I
Anc5	E	E	P	P	T	D	-	Q	V
Anc6	E	E	P	P	T	D	F	Q	V
Anc7	T	E	P	P	T	D	F	Q	V
*ec*HisB-N	S	E	P	P	S	D	F	Q	V

Table S 4: Result of a BLAST search with *pa*HisN as query.

Shown are the results as retrieved by a BLAST search with *pa*HisN as query sequence sorted by E-value. The column "Per. ident." gives the percentage of identical residues between hit and query sequence and "Acc. Len." gives the length of the hit sequence.

#	Description	Scientific Name	E value	Per. ident	Acc. Len
1	HAD hydrolase	family IB [Acinetobacter baumannii]	2,00E-157	100	217
2	PA0335	synthetic construct	3,00E-157	100.00	218
3	**phosphoserine phosphatase**	Klebsiella pneumoniae	7,00E-157	99.54	217
4	haloacid dehalogenase	Acinetobacter baumannii	7,00E-157	99.54	217
5	haloacid dehalogenase	Acinetobacter baumannii	1,00E-156	99.54	217
6	HAD-superfamily hydrolase	Enterobacter cloacae	3,00E-156	99.08	217
7	**phosphoserine phosphatase**	Streptococcus dysgalactiae subsp. equisimilis	4,00E-156	99.08	217

8	HAD-IB family hydrolase	Starkeya novella	5,00E-131	84.33	217
9	**phosphoserine phosphatase**	Prochlorothrix hollandica PCC 9006 = CALU 1027	3,00E-125	80.65	217
10	HAD-IB family hydrolase	Gammaproteobacteria bacterium	5,00E-121	78.34	217
11	HAD family phosphatase	Gammaproteobacteria bacterium	3,00E-120	74.19	217
12	HAD-IB family hydrolase	Gammaproteobacteria bacterium	6,00E-119	74.19	217
13	HAD-IB family hydrolase	Gammaproteobacteria bacterium HGW-Gammaproteobacteria-13	3,00E-117	72.35	217
14	HAD-IB family hydrolase	Gammaproteobacteria bacterium	4,00E-117	72.81	217
15	HAD-IB family hydrolase	Gammaproteobacteria bacterium	6,00E-117	72.81	217
16	phosphoserine phosphatase	Streptococcus pneumoniae	6,00E-116	73.73	218
17	HAD-IB family hydrolase	Gammaproteobacteria bacterium	5,00E-114	73.27	217
18	HAD-IB family hydrolase	Gammaproteobacteria bacterium	6,00E-114	73.27	217
19	HAD family hydrolase	Pseudomonas cremoris	8,00E-114	71.89	218
20	HAD family hydrolase	Gammaproteobacteria bacterium	1,00E-113	72.81	218
21	HAD family hydrolase	Pseudomonas brassicae	2,00E-113	72.35	218
22	HAD-IB family hydrolase	Gammaproteobacteria bacterium	6,00E-113	72.81	217
23	HAD-IB family hydrolase	Gammaproteobacteria bacterium HGW-Gammaproteobacteria-9	9,00E-113	72.81	217
24	HAD-IB family hydrolase [Gammaproteobacteria bacterium]	Gammaproteobacteria bacterium	1,00E-112	71.89	217
…					
30	HAD family hydrolase	Bacillus sp. TH86	7,00E-111	71.43	218
…					
35	HAD-IB family hydrolase	Escherichia coli	4,00E-107	66.82	218
…					
91	HAD family hydrolase	Oceanospirillaceae bacterium	1,00E-86	58.33	219
92	HAD family hydrolase	Larsenimonas suaedae	1,00E-86	60.83	219
93	HAD-IB family hydrolase	Gammaproteobacteria bacterium	2,00E-86	57.14	217
94	HAD-IB family hydrolase	Gammaproteobacteria bacterium	2,00E-86	58.53	218
95	HAD-IB family hydrolase	Spongiibacter sp.	2,00E-86	58.99	217
96	HAD family hydrolase	Motiliproteus sediminis	2,00E-86	59.07	217
97	**phosphoserine phosphatase**	Cellvibrio sp. PSBB006	2,00E-86	58.06	231

98	HAD-IB family hydrolase	Thiomicrospira sp. R3	3,00E-86	57.14	217
99	HAD family hydrolase	Spongiibacter thalassae	4,00E-86	60.37	217
100	HAD family hydrolase	Spongiibacter marinus	4,00E-86	58.53	217

Table S 5: Sequencing result of SerB mutants that rescue a Δ*holPase* strain.

Color code: red: acidic, black: hydrophobic, green: hydroxyl or sulfhydryl group, purple: amide, blue: basic

	Codons at randomized positions				Amino acids at randomized positions			
Clone #	125	127	158	161	125	127	158	161
wt	GAA	ATT	CTG	CGT	E	I	L	R
1	Ambiguous read							
2	Ambiguous read							
3	GAG	TGT	CTT	GTT	E	C	L	V
4	AGT	TGG	ATT	GTG	S	W	I	V
5	GAG	CTG	CTG	AGT	E	L	L	S
6	GAG	AGA ?	CAA ?	GTT	E	R	Q	V
7	GAG	GTT	CTG	TGT	E	V	L	C
8	AGT	TGG	CTT	ATT	S	W	L	I
9	GAT	TGG	TGT	GTG	D	W	C	V
10	GAG	TGT	CTT	GTT	E	C	L	V
11	Ambiguous read							
12	AGT	TGG	CTT	ATT	S	W	L	I
13	ATT	GTG	GCT	TGG	I	V	A	W
14	AGT	TGG	CTT	ATT	S	W	L	I

Table S 6: Sequencing result of the second library of HisB-N mutants that rescue a ΔserB strain.

Color code: red: acidic, black: hydrophobic, green: hydroxyl or sulfhydryl group, purple: amide, blue: basic.

	Codons at randomized positions				Number of clones
Clone #	57	58	104	105	
1	Q	P	M	K	7
2	Q	P	S	K	4
3	Q	P	A	K	3
4	Q	D	R	K	2
5	Q	P	E	K	2
6	Q	T	R	K	2
7	Q	P	H	K	2
8	Q	P	K	K	1
9	Q	P	Q	K	1

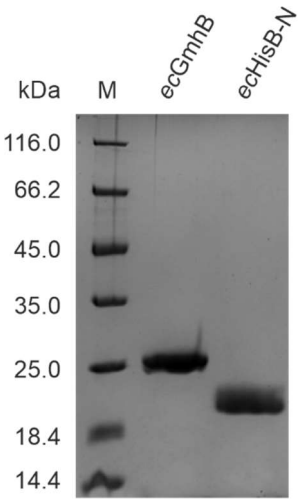

Figure S 1: SDS-PAGE (13.5 % acrylamide) of purified His_6-tagged proteins *ec*HisB-N and *ec*GmhB (3 µg each).

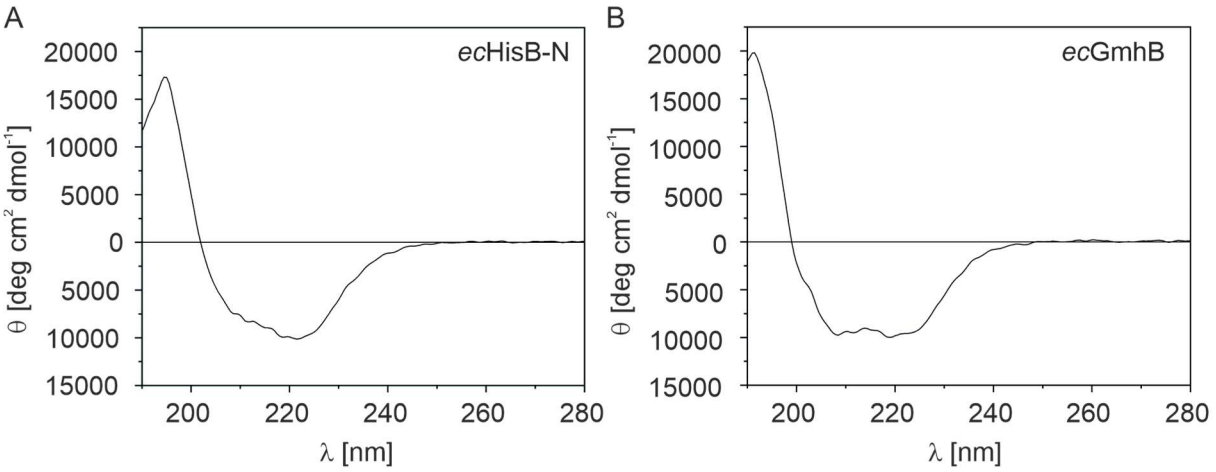

Figure S 2: Far-UV CD spectra of purified *ec*HisB-N (A) and *ec*GmhB (B).

CD spectra were recorded in 20 mM potassium phosphate buffer (pH 7.5).

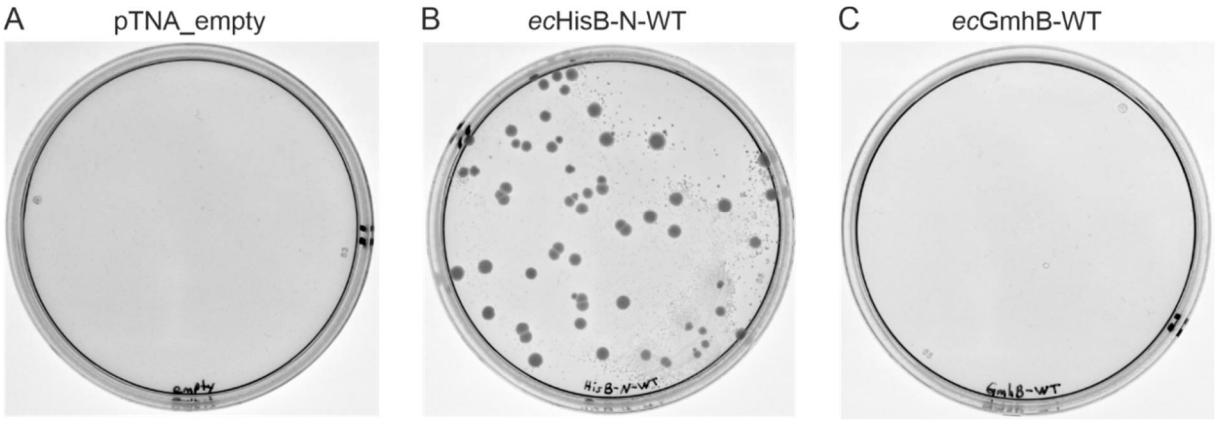

Figure S 3: *In vivo* complementation experiments with an auxotrophic *E. coli* Δ*holPase* knock-out strain.

The strain was transformed with (A) the empty vector (B) a vector encoding *ec*HisB-N or (C) a vector encoding *ec*GmhB, which are both the wild types (WT). The transformants were plated on M9 minimal medium and incubated at 37 °C for 8 days.

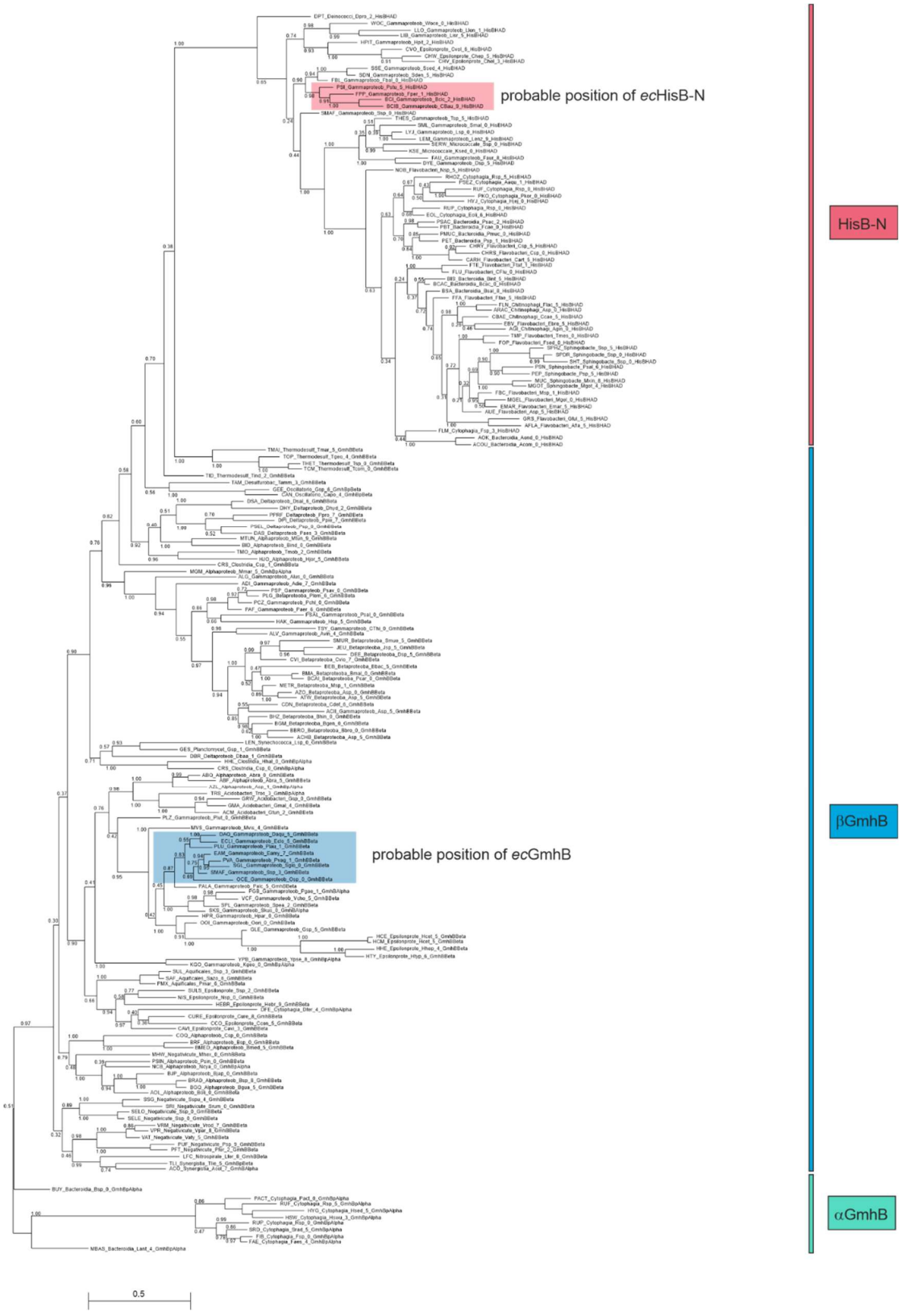

Figure S 4: Phylogenetic tree including sequences of αGmhB, βGmhB, and HisB-N proteins.

This consensus tree was calculated using Bayesian phylogenetics. The numbers at each branch represent the corresponding posterior probabilities. Due to filtering processes the data set is lacking the sequences of ecHisB N and ecGmhB. Their most probable positions are indicated by rectangles marking their expected neighbors (red rectangle for ecHisB-N and blue rectangle for ecGmhB). The length of the bar at the bottom corresponds to 0.5 mutations per site.

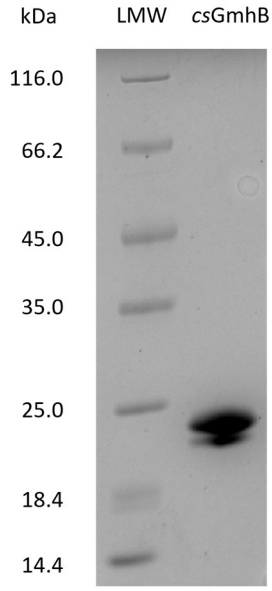

Figure S 5: SDS-PAGE (13.5 % acrylamide) of purified His_6-tagged *cs*GmhB (3 µg).

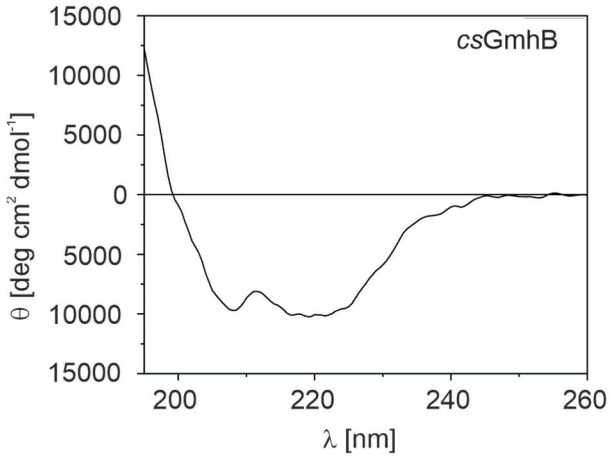

Figure S 6: Far-UV CD-spectrum of purified *cs*GmhB.

CD spectra were recorded in 20 mM potassium phosphate buffer (pH 7.5).

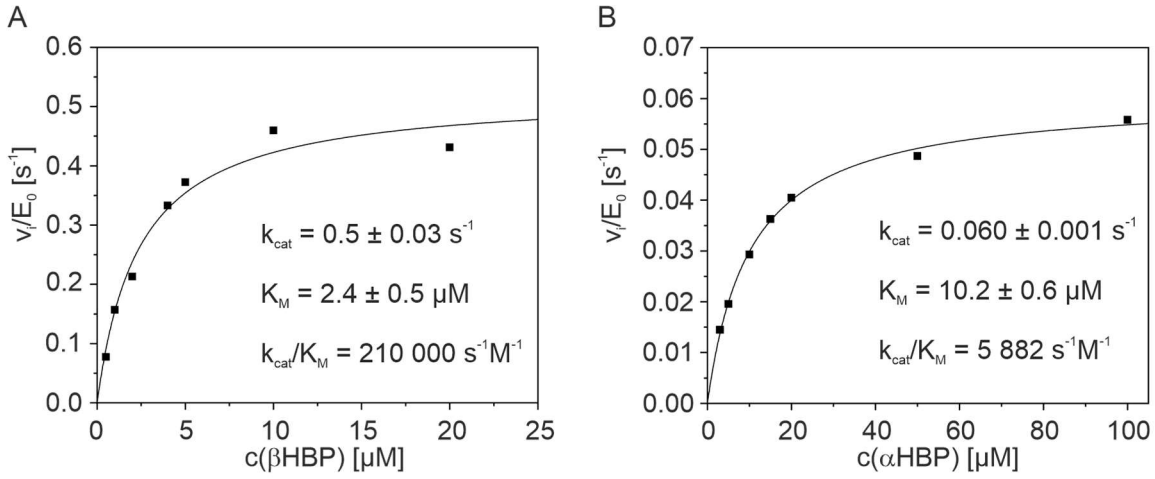

Figure S 7: Substrate saturation curve for the turnover of βHBP and αHBP by *cs*GmhB at 25 °C.

Reaction conditions included 100 mM Tris/HCl buffer (pH 7.8), 5mM $MgCl_2$, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

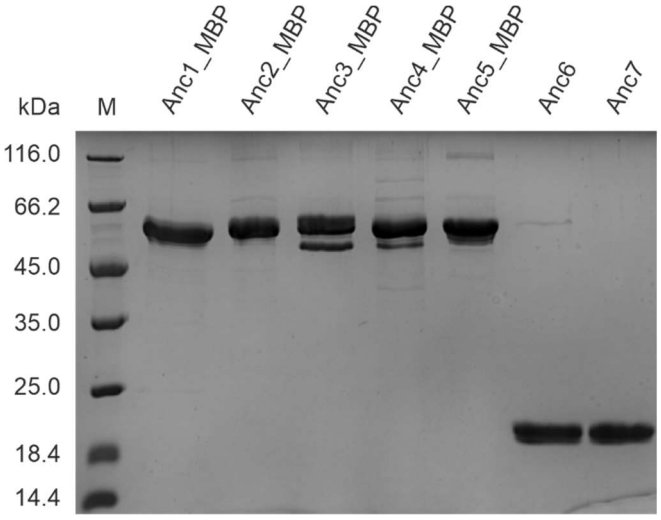

Figure S 8: SDS-PAGE (13.5 % acrylamide) of purified resurrected proteins (3 µg each).

Anc1, Anc2, Anc3, Anc4, and Anc5 contain an N-terminal MBP tagged protein, whereas Anc6 and Anc7 contain an N-terminal His6-tag.

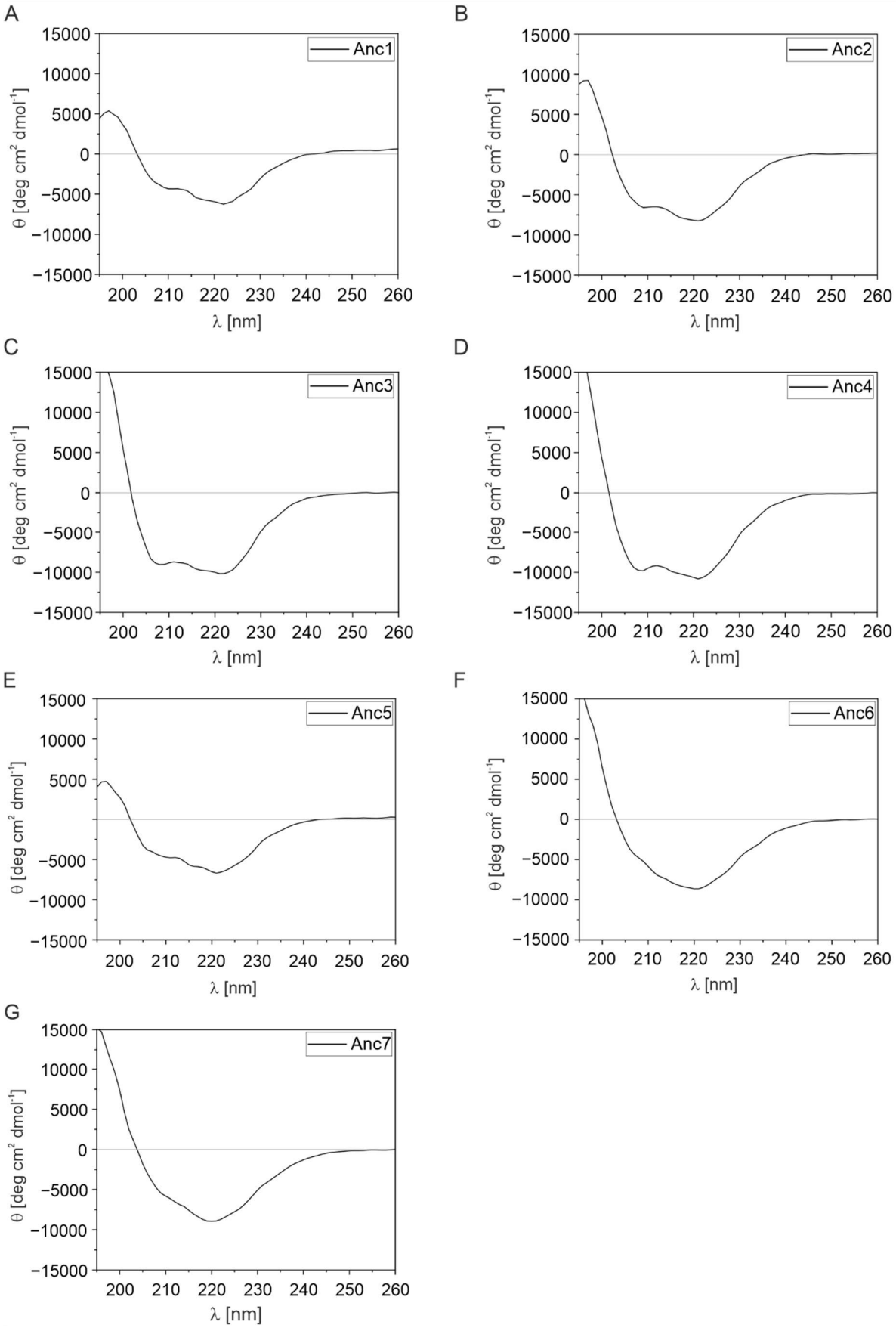

Figure S 9: Far-UV CD-spectra of purified resurrected proteins Anc1-Anc7.

Anc1-Anc4 contain an N-terminal MBP tag, whereas Anc6 and Anc7 contain an N-terminal His_6-tag. Spectra were recorded in 20 mM potassium phosphate buffer (pH 7.5).

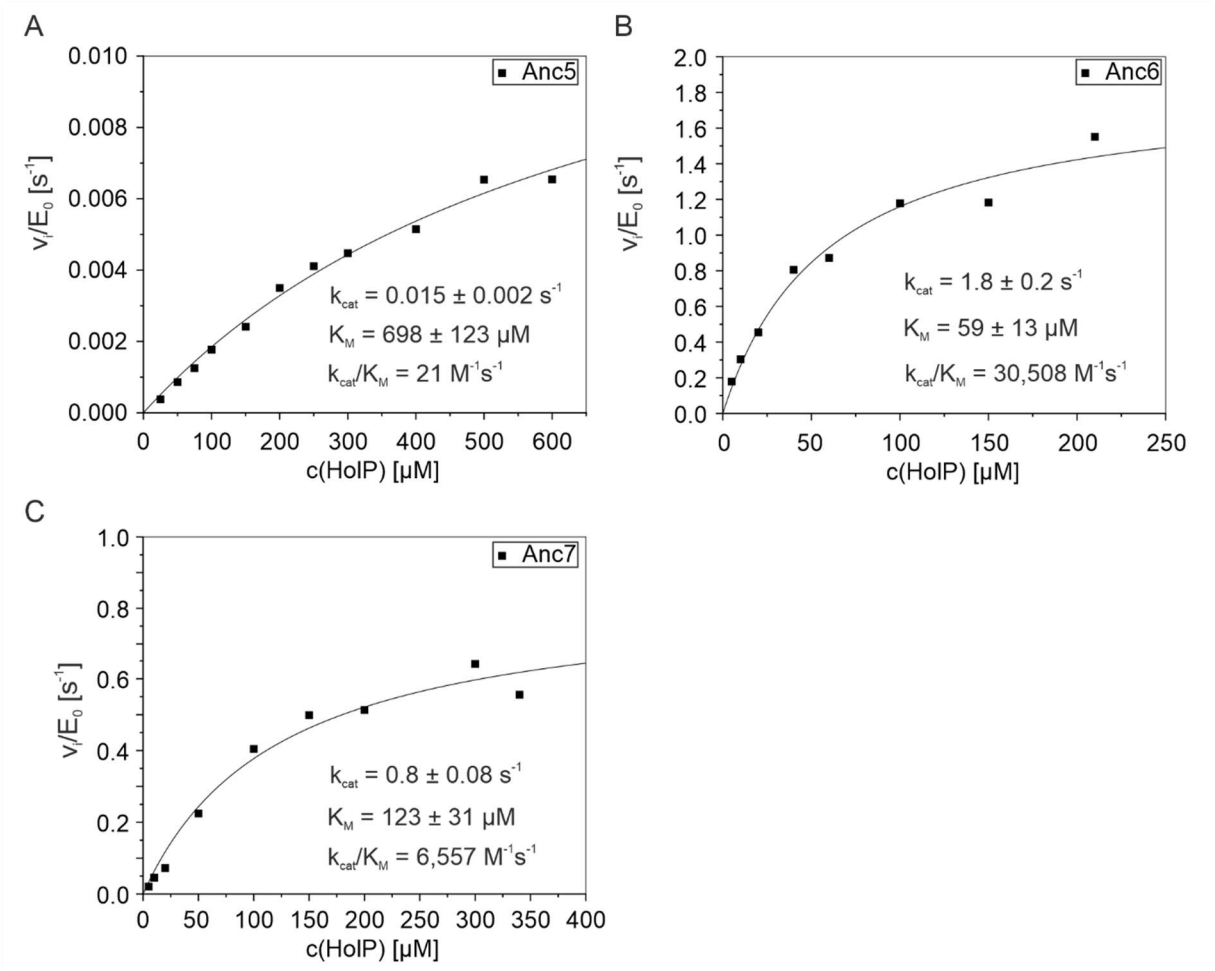

Figure S 10: Substrate saturation curve for the HolPase activity of Anc5-Anc7 at 25 °C.

Reaction conditions included 100 mM Tris/HCl buffer (pH 7.8), 5mM $MgCl_2$, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

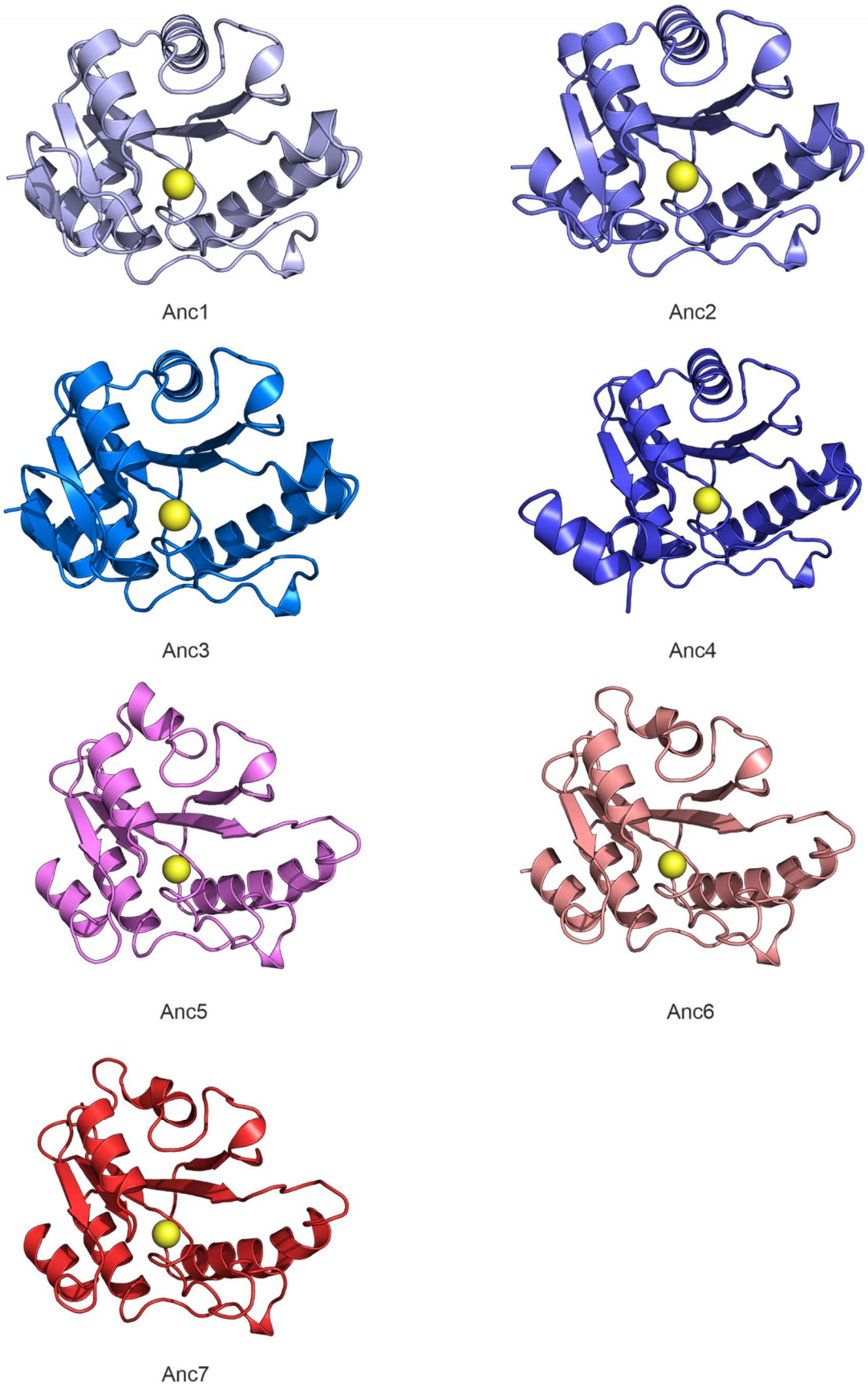

Figure S 11: Folds of Anc1-Anc7 as predicted by AlphaFold.[100]

The yellow spheres represent magnesium ions that were extracted from a superimposed *ec*HisB-N structure (PDB ID: 2FPU).[59]

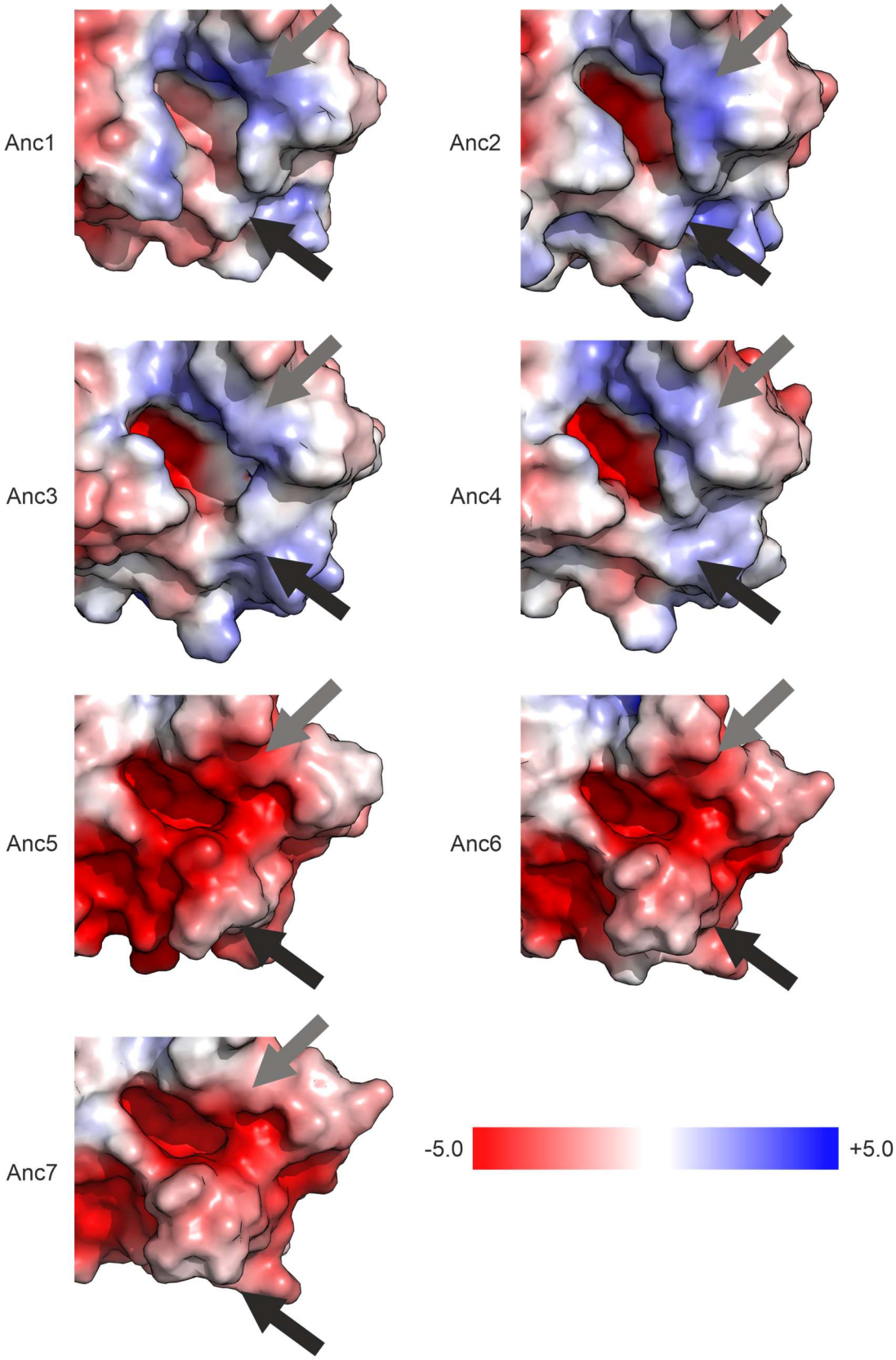

Figure S 12: Structural analysis of the shape and charge distribution of the active sites of Anc1-Anc7.

Grey arrows indicate the part of the active site that is gradually extended towards the right and black arrows indicate the lid whose size is increasing from Anc1 to Anc7. The scale bar indicates the color that is associated with the electrostatic surface potential normalized to a range from -5 to +5.

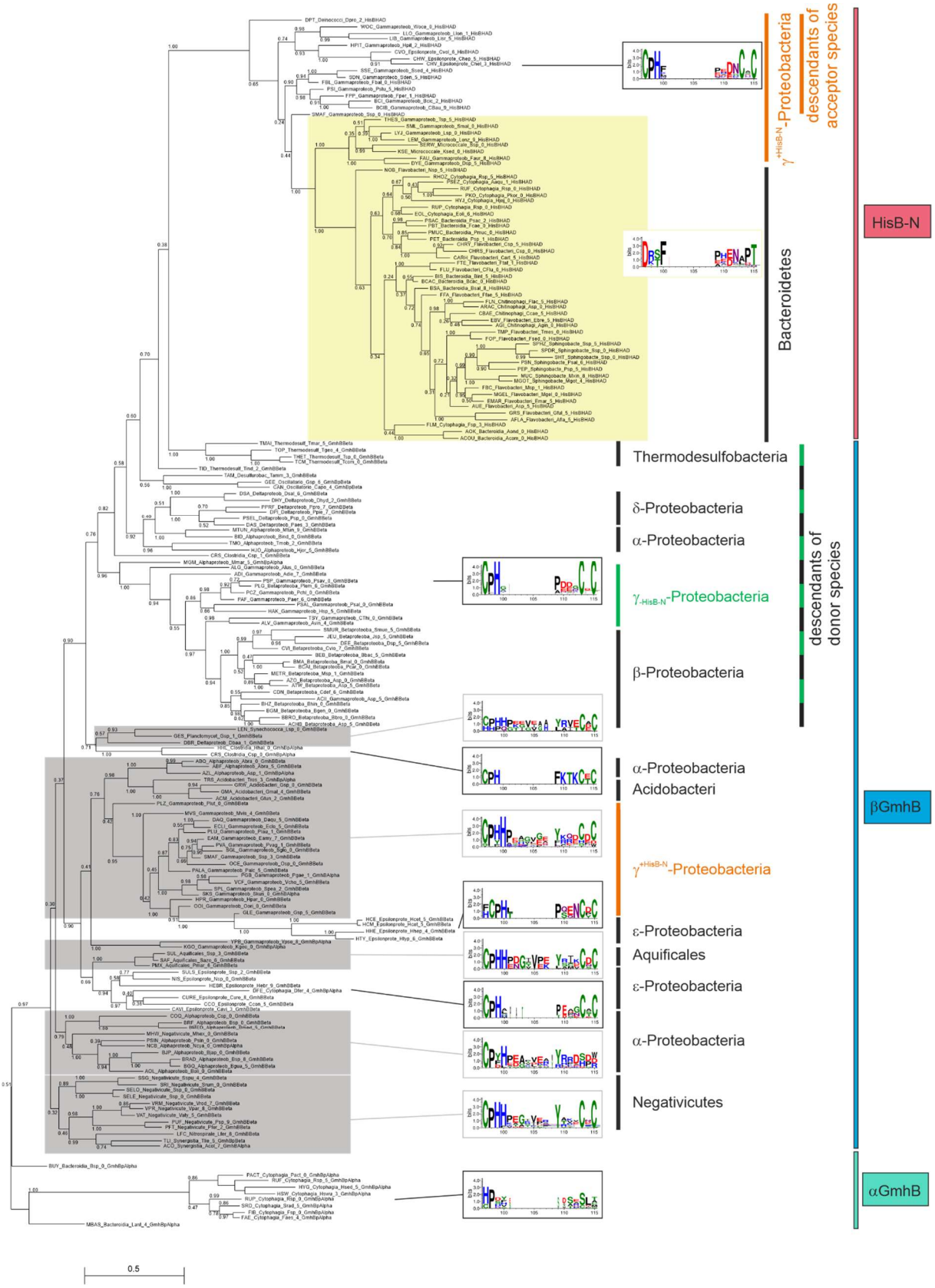

Figure S 13: Phylogenetic tree including sequences of αGmhB, βGmhB, and HisB-N proteins.

The numbers at each branch represent the posterior probabilities. Grey boxes indicate βGmhB sequences where the zinc binding CxH and CxC motifs are separated by 12-13 residues with the corresponding sequence logos depicted to the right. The x-axis of the logos gives the residue-positions of the multiple

alignment containing all sequences. In HisB-N and βGmhB sequences that are not marked, the CxH and CxC motifs are separated by 5-6 residues. The yellow box indicates HisB-N sequences which lack the CxH and CxC motifs. Bacterial phyla are listed on the right of each cluster and the descendants of the probable donor and acceptor species of the horizontal gene transfer are marked. The length of the bar at the bottom corresponds to 0.5 mutations per site. The immediate descendants of the most plausible acceptor species are γ-Proteobacteria, with the exception of three ε-proteobacterial species from the Campylobacterales order and one from the Deinococcales order. The occurrence of the *hisB-N* gene in Campylobacterales was previously attributed to a horizontal gene transfer of most of the *his* biosynthetic genes from a γ-Proteobacterium to a Campylobacterium.[83] We did not find a HisB-N gene in any other ε-Proteobacterium or any other organism that is related to *Deinococcus* and hence also concluded that the occurrence of HisB-N is most likely due to another horizontal gene transfer.

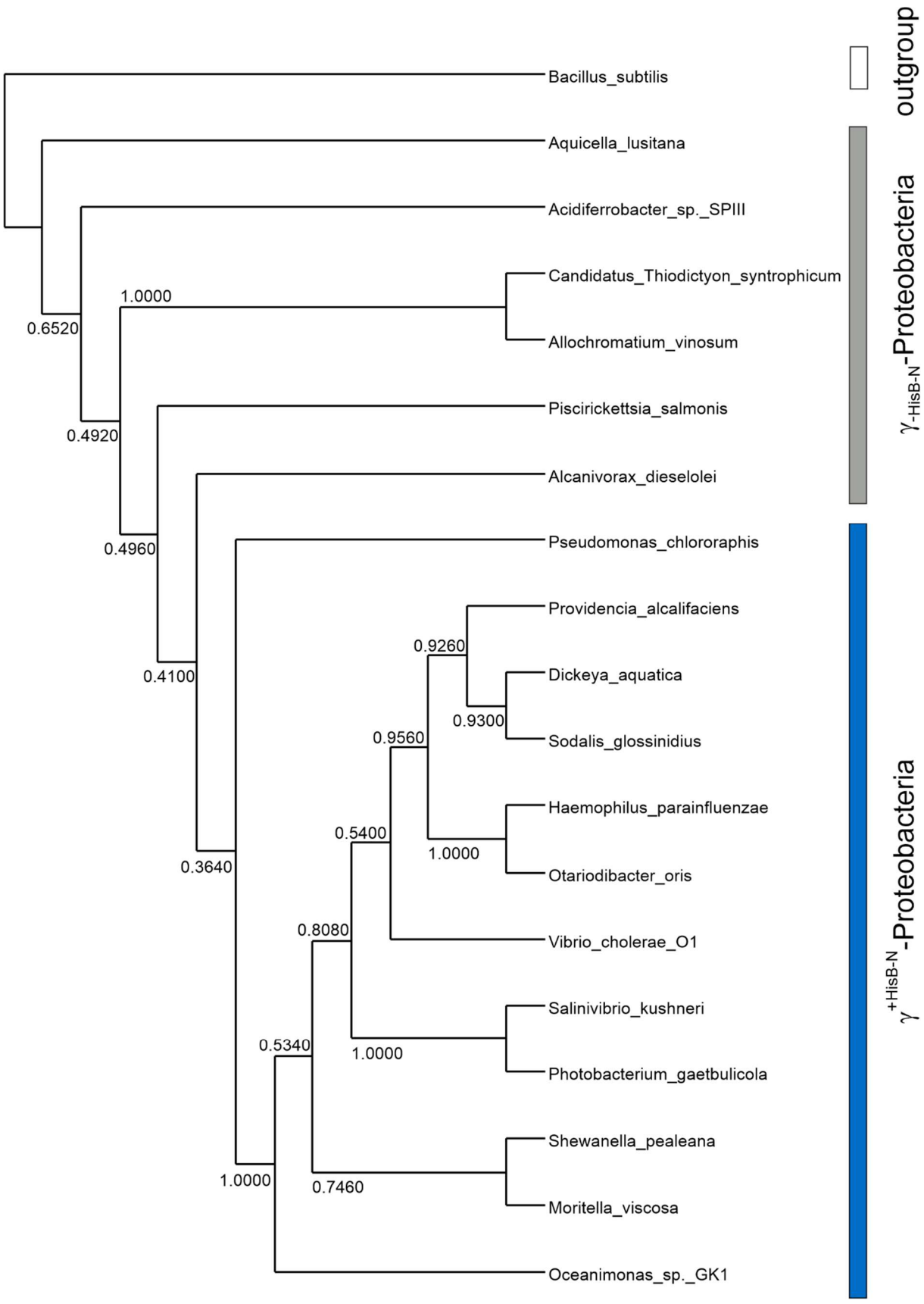

Figure S 14: Phylogenetic relationship of γ-Proteobacteria on the basis of their 16S r-RNA sequences.

Numbers indicate bootstrap values and bars the $\gamma^{+HisB\text{-}N}$ (blue) and $\gamma_{-HisB\text{-}N}$ (grey) clades. The 16S r-RNA sequence of *Bacillus subtilis* serves as an outgroup.

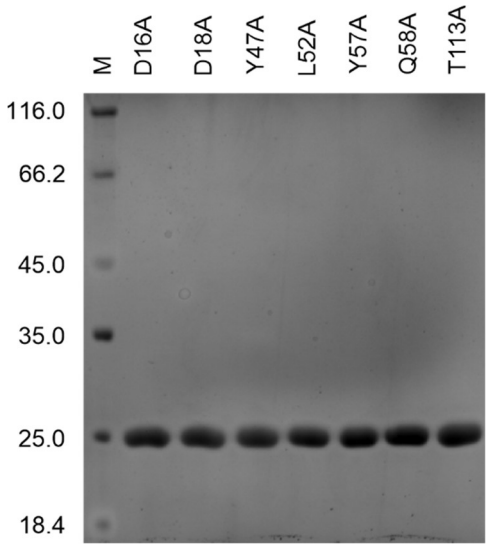

Figure S 15: SDS-PAGE (13.5 % acrylamide) of purified proteins from the *pa*HisN alanine scan (3 µg each).

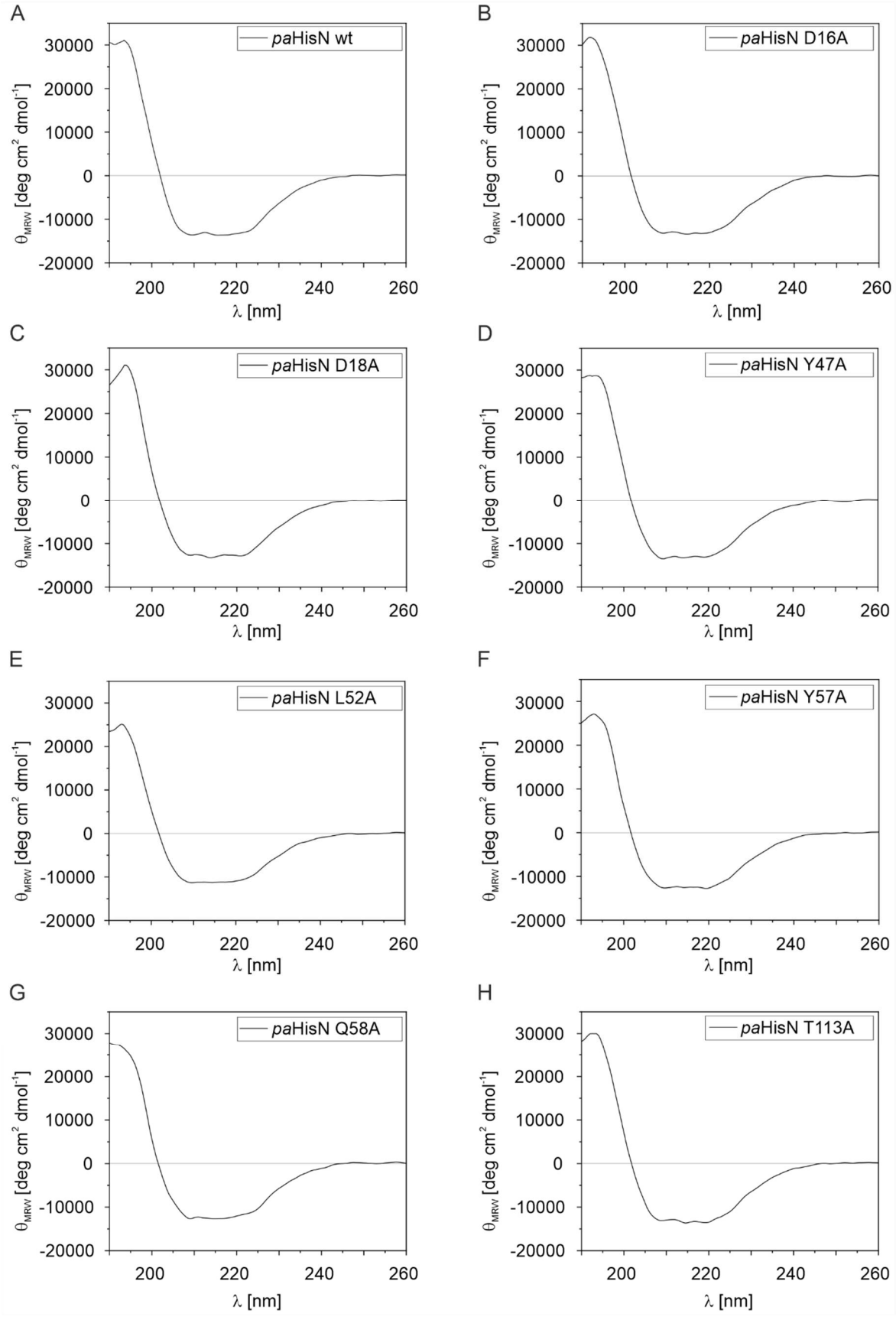

Figure S 16: Far-UV CD-spectra of purified proteins from the *pa*HisN alanine scan.

spectra were recorded in 20 mM potassium phosphate buffer (pH 7.5).

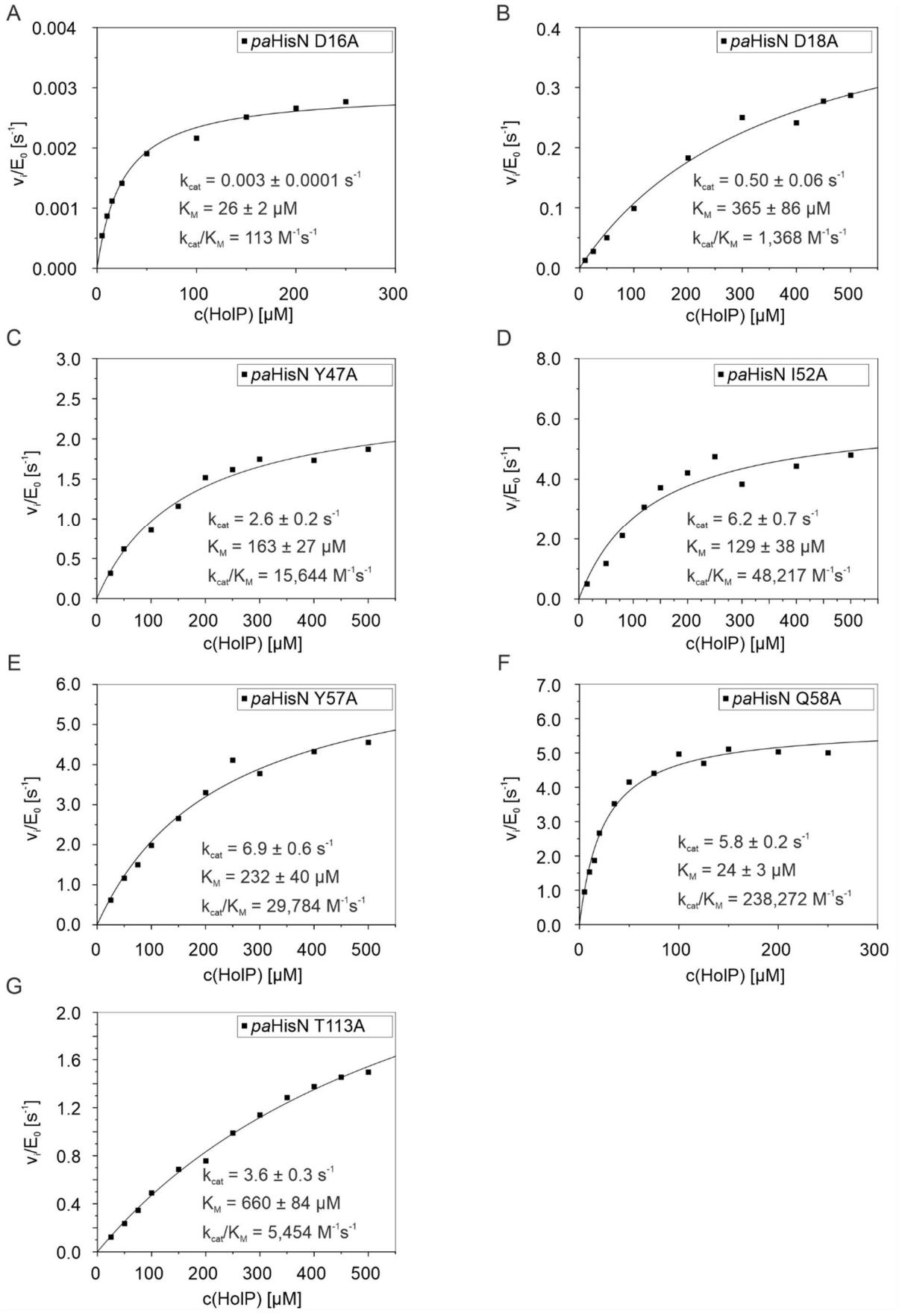

Figure S 17: Saturation curves for the turnover of HolP by the *pa*HisN alanine scan mutants.

Reaction conditions included 100 mM Tris/HCl buffer (pH 7.8), 5mM $MgCl_2$, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

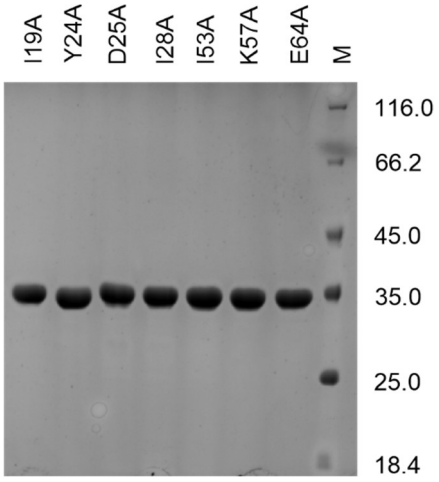

Figure S 18: SDS-PAGE (13.5 % acrylamide) of purified proteins from the *nm*HisN alanine scan (3 µg each).

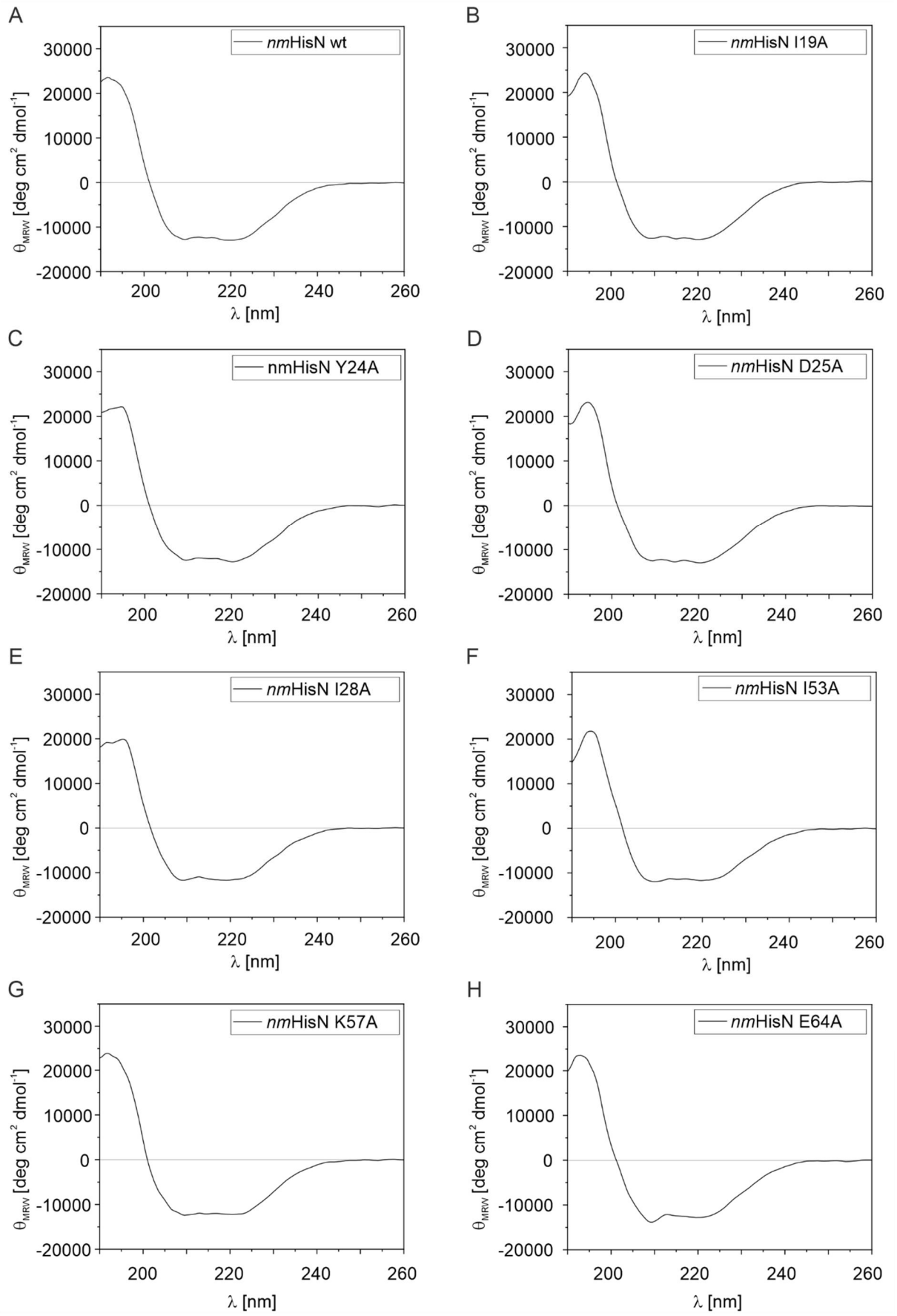

Figure S 19: Far-UV CD-spectra of purified proteins from the *nm*HisN alanine scan.

spectra were recorded in 20 mM potassium phosphate buffer (pH 7.5).

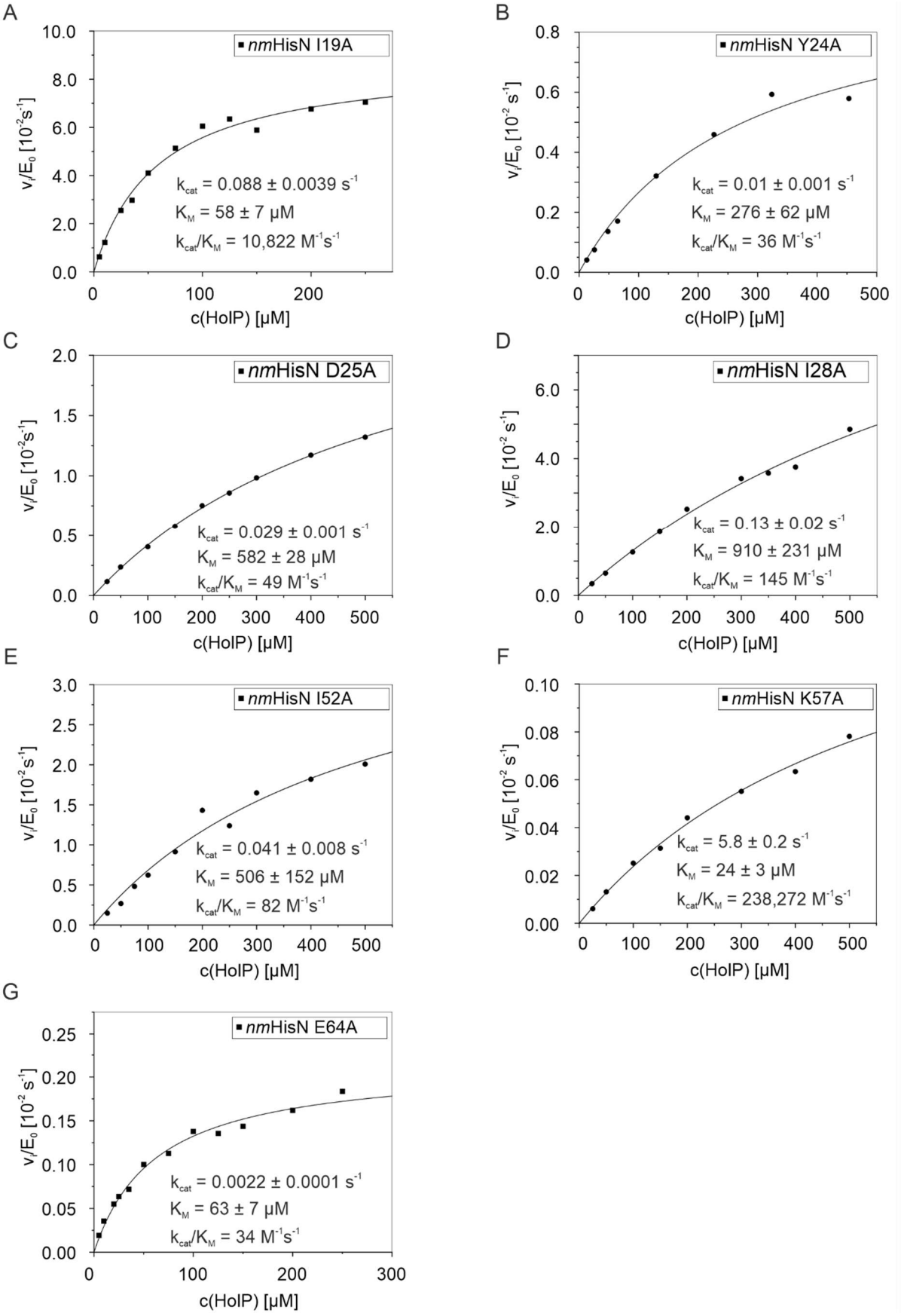

Figure S 20: Saturation curves for the turnover of HolP by the *nm*HisN alanine scan mutants.

Reaction conditions included 100 mM Tris/HCl buffer (pH 7.8), 5mM $MgCl_2$, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

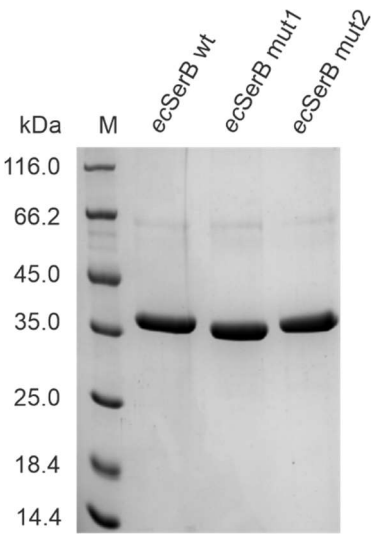

Figure S 21: SDS-PAGE of *ec*SerB mutants (3 µg each).

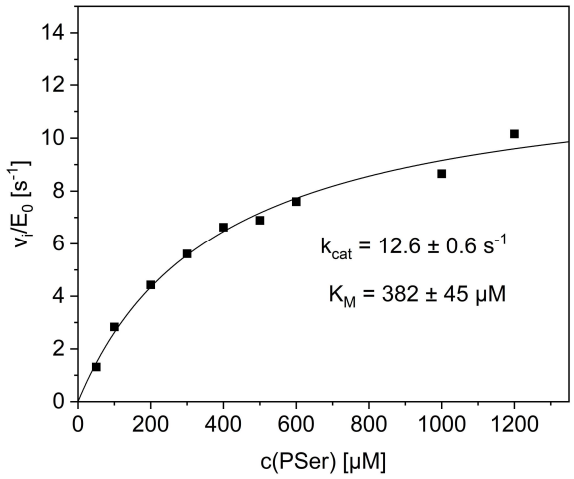

Figure S 22: Saturation curve for the turnover of PSer by *ec*SerB at 25°C.

Reaction conditions included 100 mM Tris/HCl buffer (pH 7.8), 5mM $MgCl_2$, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

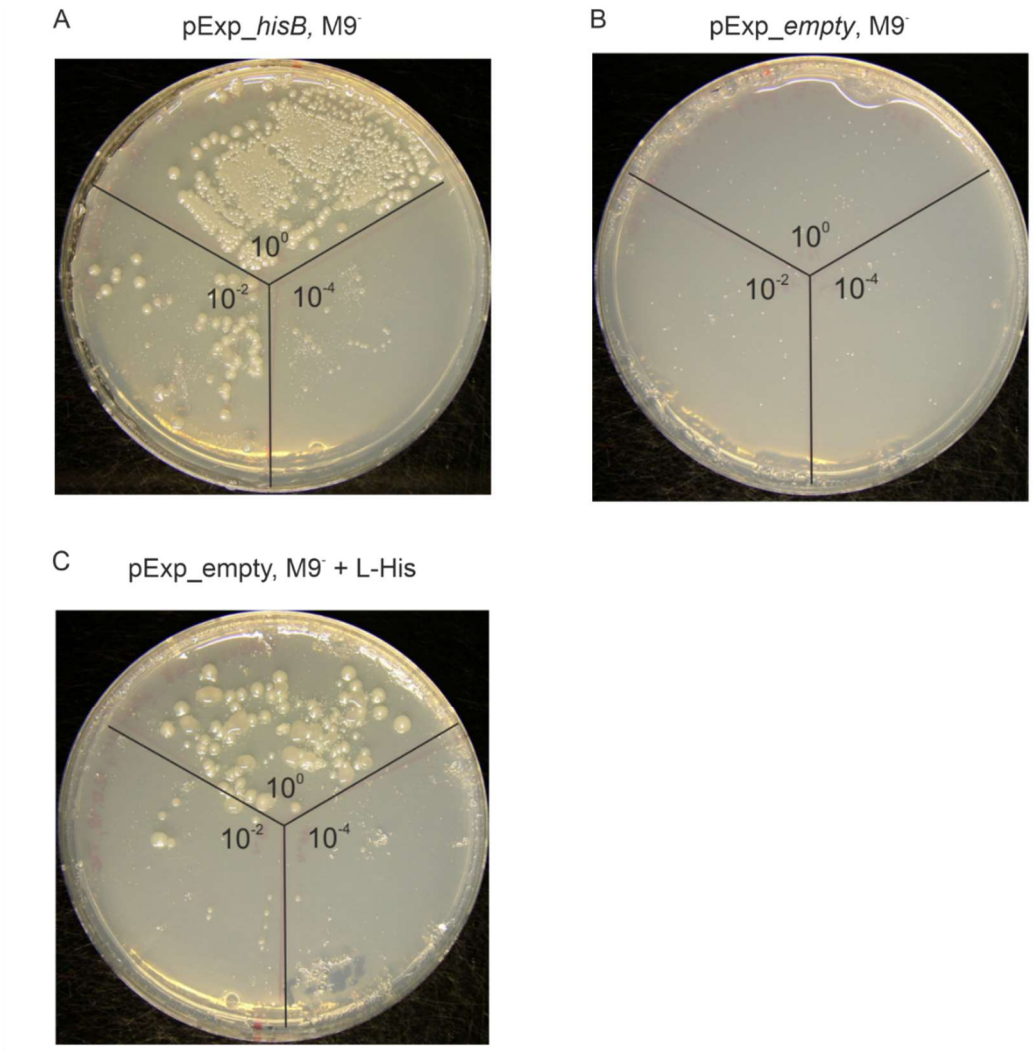

Figure S 23: Verification of a Δ*hisB::kana*^{R} knock-out.

E. coli cells with the genomic Δ*hisB::kana*^{R} knock-out were transformed with a plasmid encoding for a hisB gene (A), or an empty vector (B, C), streaked onto M9 minimal medium plates (A,B) or M9 minimal medium plates supplied with L-histidine (C). Pictures were taken after five days of incubation at 37°C.

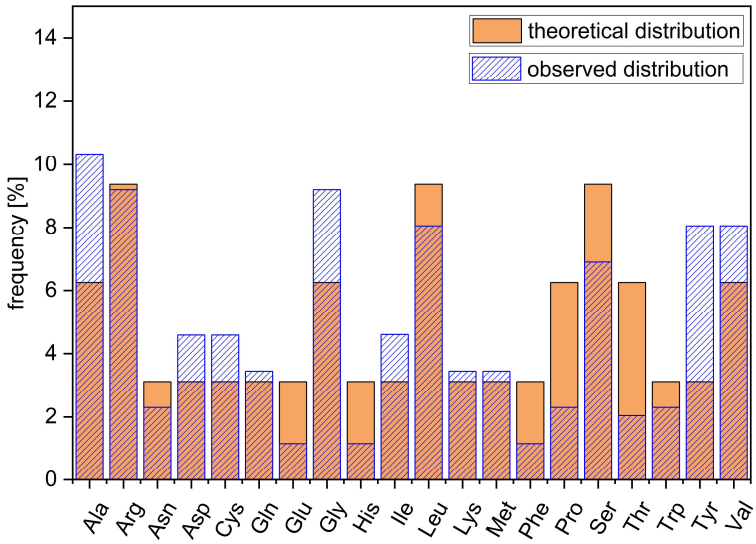

Figure S 24: Amino acid distribution of the gene library on *ec*SerB.

Amino acid distribution of 4 randomized sites from 25 individually picked clones as determined by Sanger sequencing (blue bars), compared to the expected distribution (orange bars) based on the number of codons with which each amino acid is represented in an NNK randomization scheme.

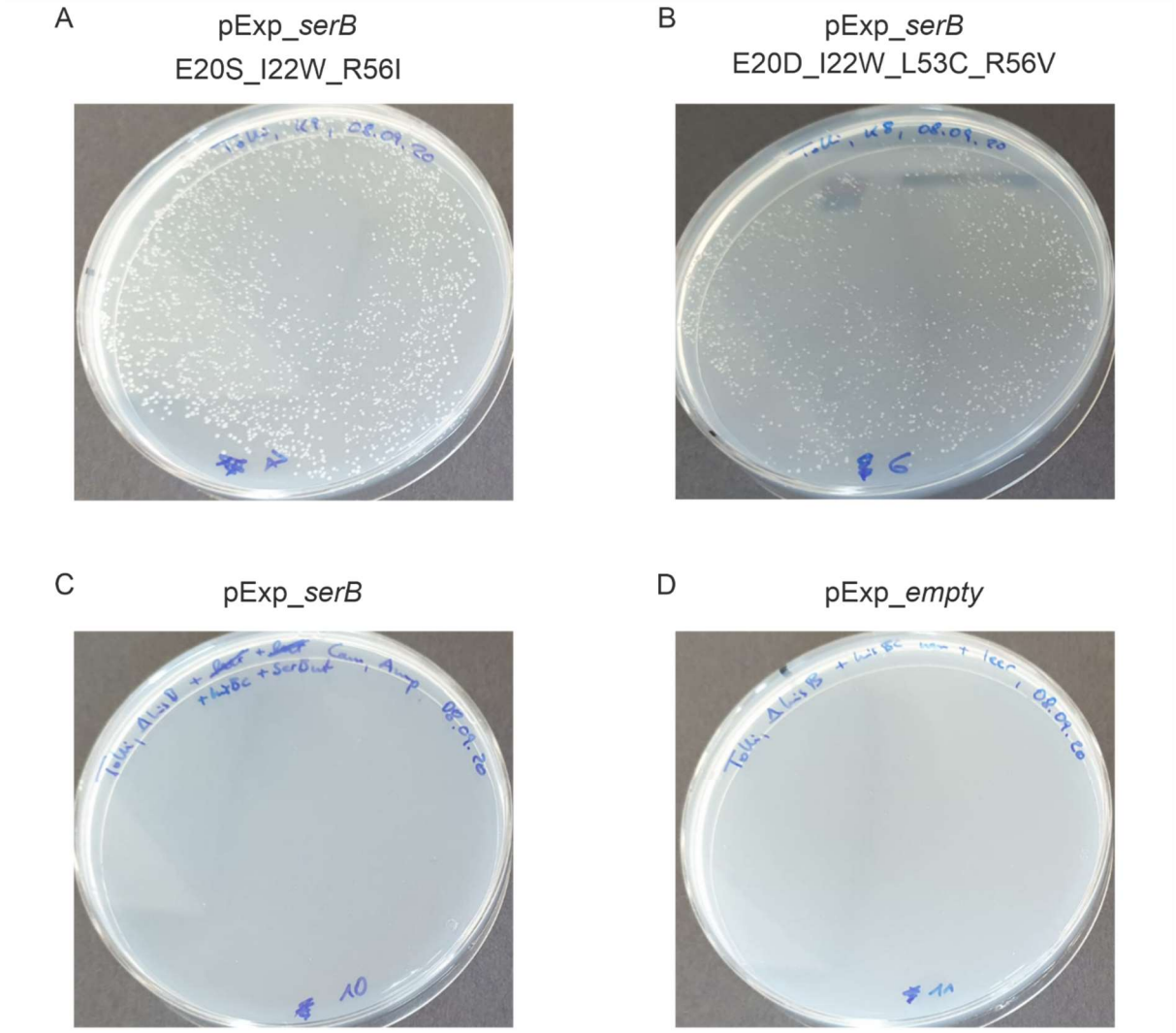

Figure S 25: Re-transformation experiment with the *serB* gene library.

E. coli cells with the genomic Δ*holPase* knock-out were transformed with a plasmid encoding for an *ecserB* mutant gene which was isolated from a fast-growing colony (A, B), with a plasmid encoding for the *ecserB* wildtye gene (C), or with an empty vector (D) and streaked onto M9 minimal medium plates. Pictures were taken after three days of incubation at 37°C.

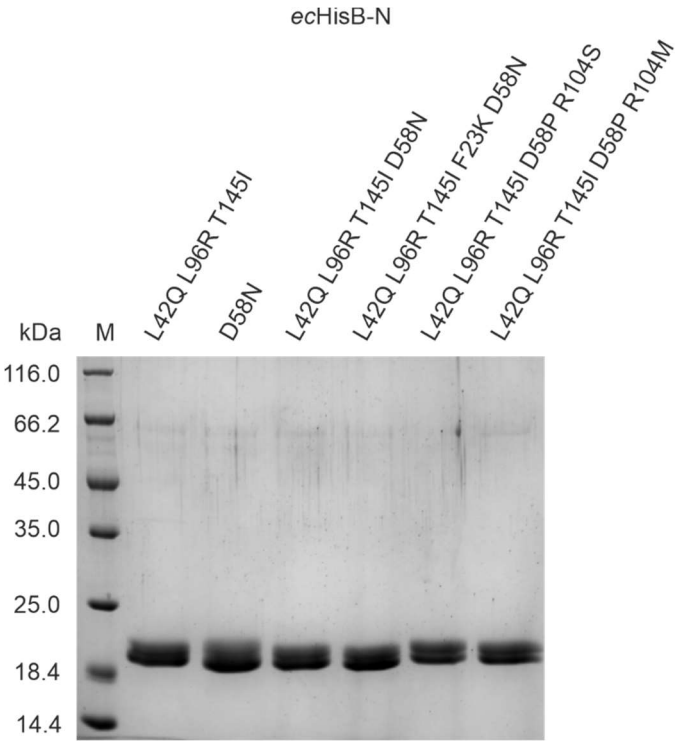

Figure S 26: SDS-PAGE of *ec*HisB-N mutants (3 µg each).

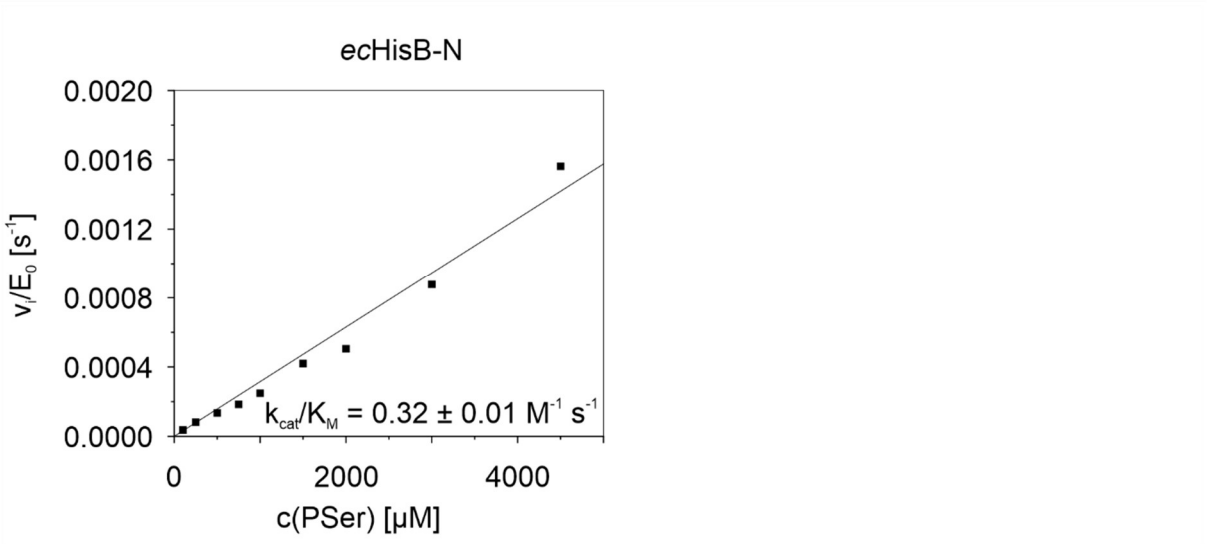

Figure S 27: Steady-state experiment for PSer by ecHisB-N at 25°C.

Reaction conditions included 100 mM Tris/HCl buffer (100 mM, pH 7.8), 5mM MgCl2, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

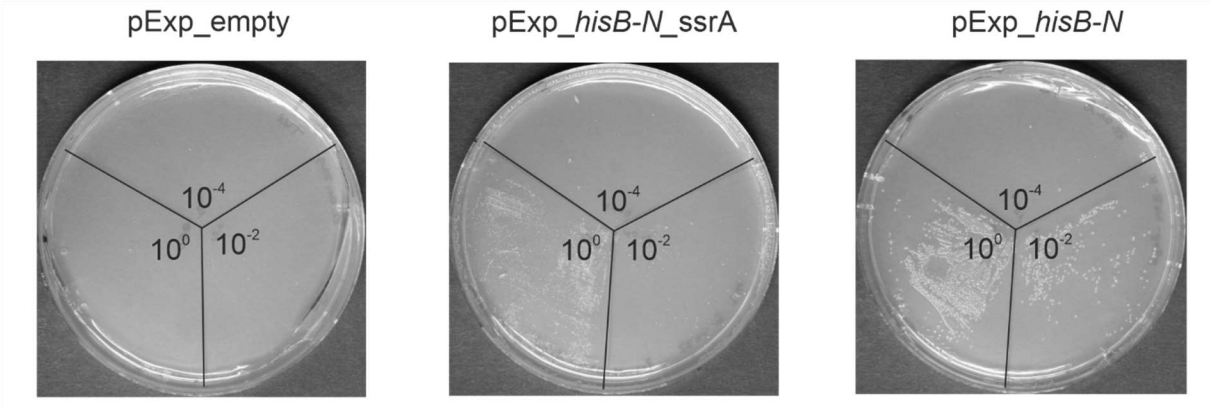

Figure S 28: Test of the SsrA degradation tag.

E. coli cells with the genomic Δ*serB* knock-out were transformed with an empty vector (A), with a plasmid encoding for the *ecHisB* with a C-terminal SsrA degradation tag (B), of with a plasmid encoding for the *ecHisB* wildtype gene (C) and streaked onto M9 minimal medium plates. Pictures were taken after two days of incubation at room temperature.

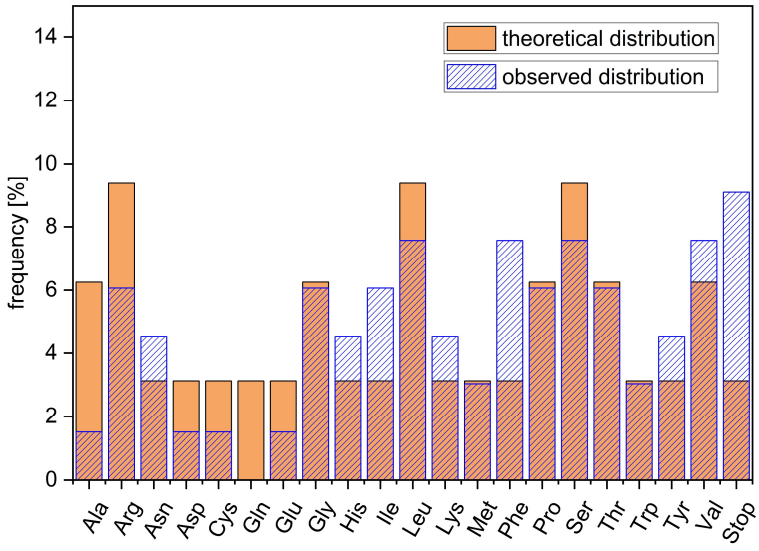

Figure S 29: Amino acid distribution of the gene library on *ec*HisB-N.

Amino acid distribution of 4 randomized sites from 18 individually picked clones as determined by Sanger sequencing (blue bars), compared to the expected distribution (orange bars) based on the number of codons with which each amino acid is represented in an NNK randomization scheme.

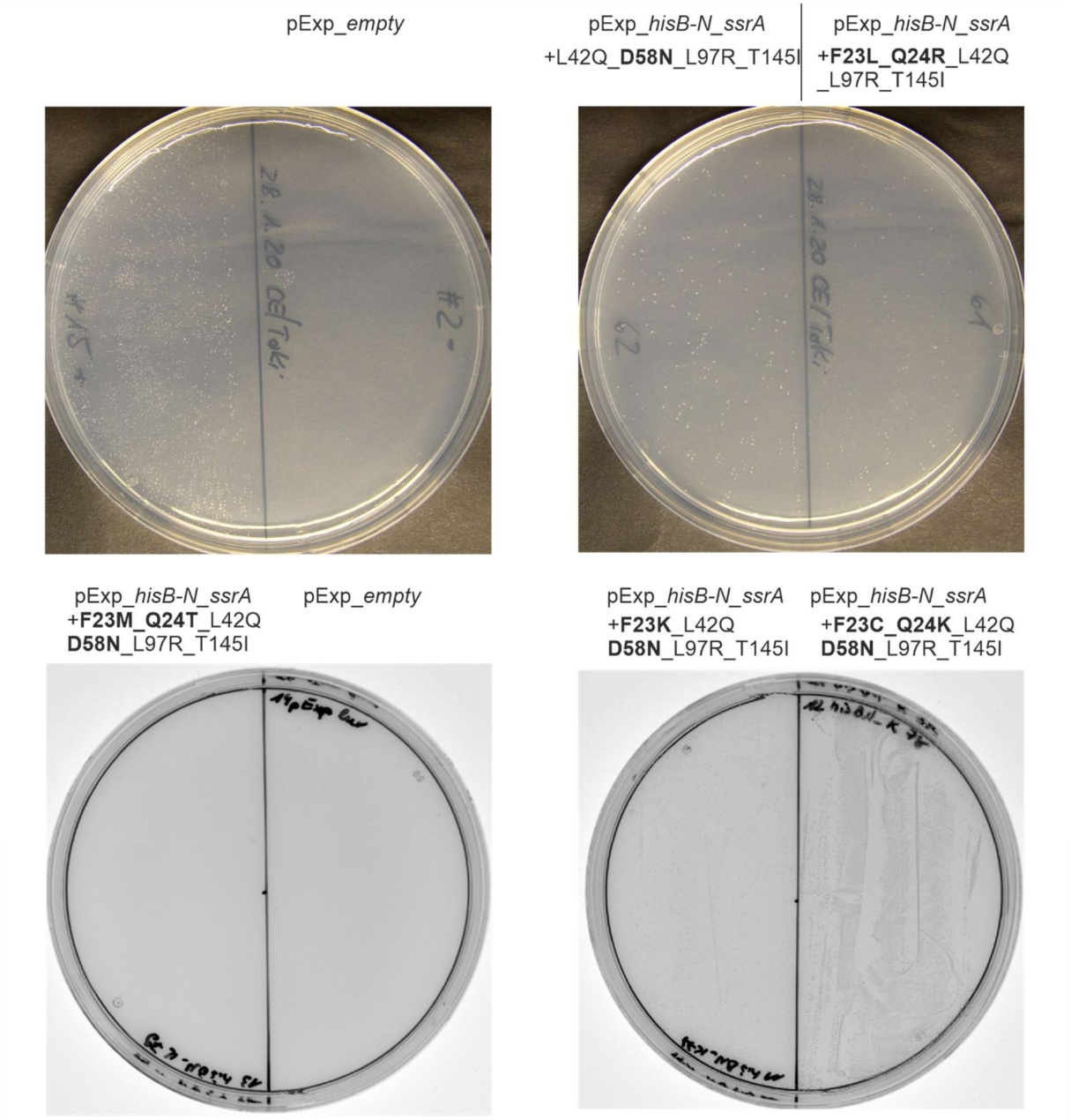

Figure S 30: Re-transformation experiment with the *hisB-N* gene library.

E. coli cells with the genomic Δ*serB* knock-out were transformed with a plasmid coding for an *ecHisB-N* mutant gene which was isolated from a fast-growing colony or with an empty vector (pExp_*empty*) and streaked onto M9 minimal medium plates. The plasmids contained a C-terminal SsrA degradation tag. Mutations relative to the wildtype gene are indicated and mutations at randomized positions are highlighted in bold print. Pictures were taken after 20 h of incubation at 37°C.

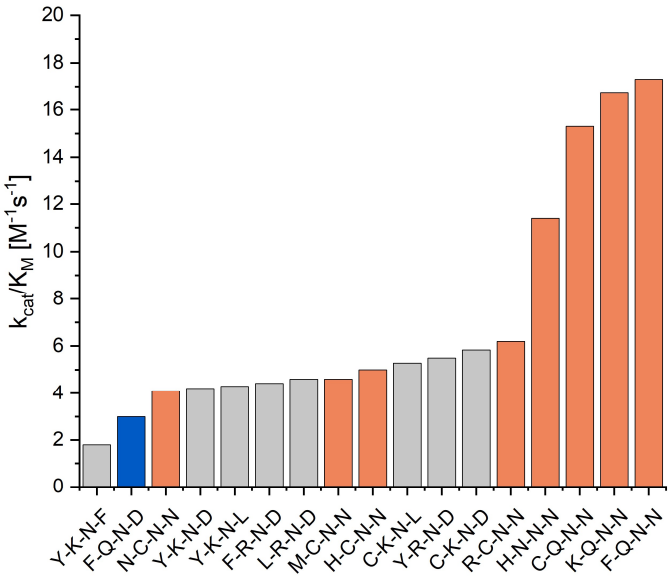

Figure S 31: PSPase activities of selected *ec*HisB-N mutants at 25°C.

The bar plots indicate the kcat/KM value from steady state kinetic experiments on different *ec*HisB-N mutants from library 1. The amino acids at the randomized positions F23, Q24, N56, and D58 are given below. Reaction conditions included 100 mM Tris/HCl buffer (100 mM, pH 7.8), 5mM MgCl2, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase. Color code: blue: wildtype, orange: D58N containing mutant, grey: D58x mutant.

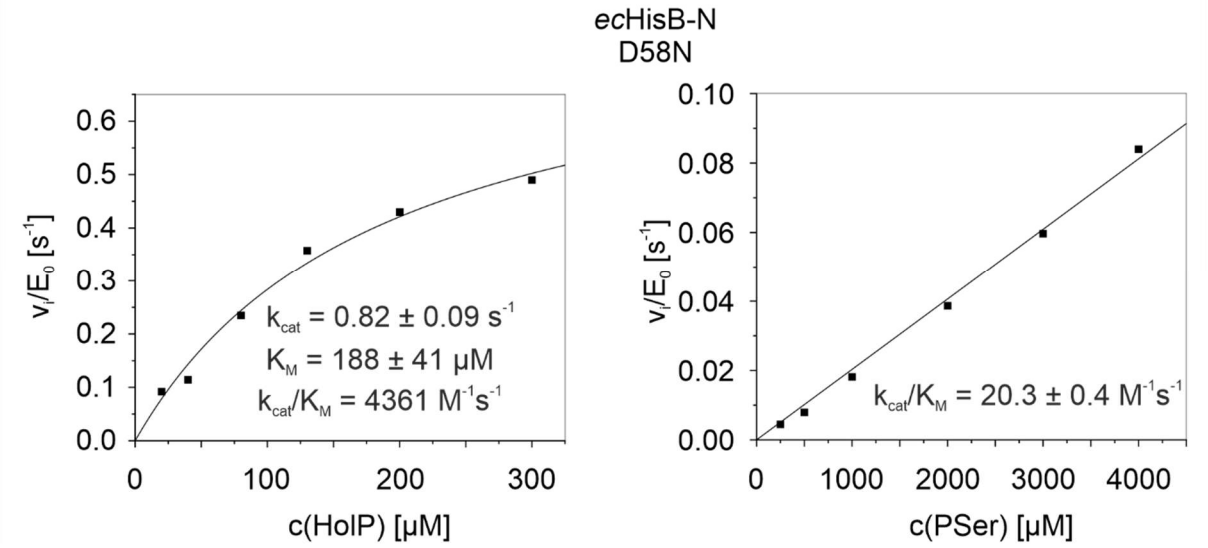

Figure S 32: D58N single mutant activity at 25°C.

Reaction conditions included 100 mM Tris/HCl buffer (100 mM, pH 7.8), 5mM MgCl2, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

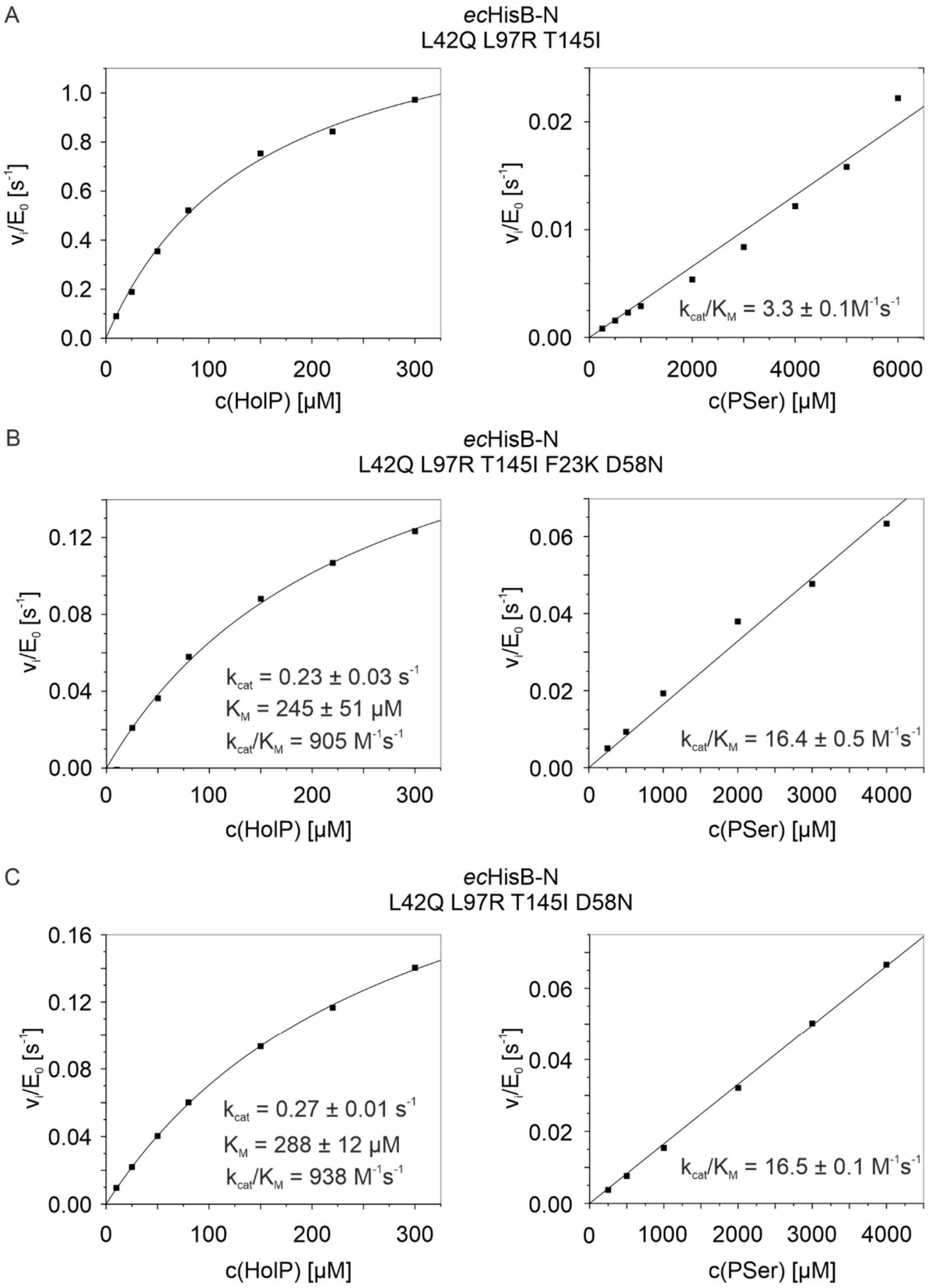

Figure S 33: Steady-state kinetic experiments for the turnover of HolP and PSer by selected *ec*HisB-N mutants at 25°C.

Reaction conditions included 100 mM Tris/HCl buffer (100 mM, pH 7.8), 5mM MgCl2, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

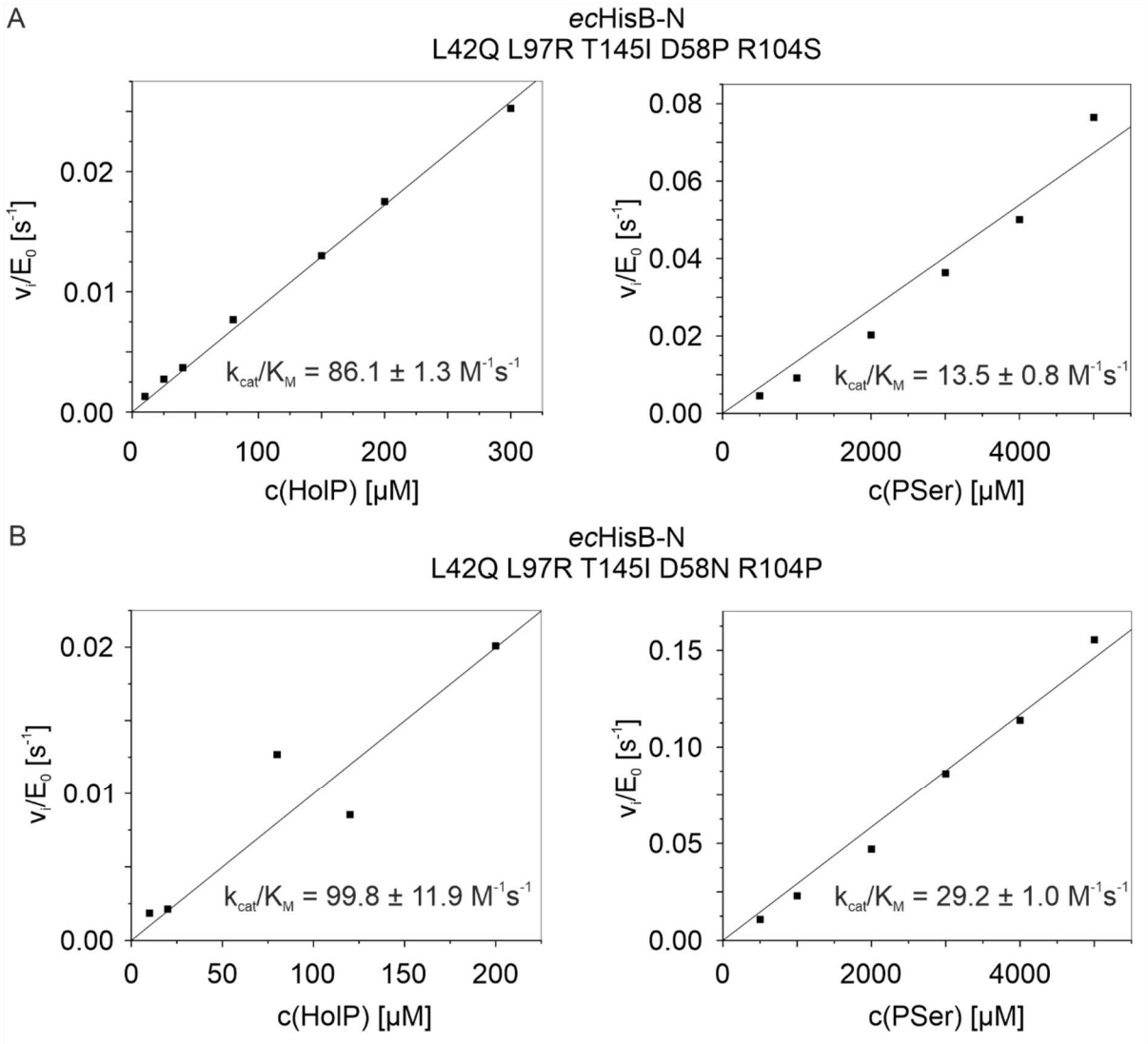

Figure S 34: Steady-state kinetic experiments for the second library mutants at 25°C.

Reaction conditions included 100 mM Tris/HCl buffer (100 mM, pH 7.8), 5mM MgCl2, 0.5 mM inosine 0.25 U/mL purine nucleoside phosphorylase, and 2.5 U/mL xanthine oxidase.

Acknowledgements

First, I want to cordially thank my supervisor Prof. Reinhard Sterner for his constant support, for always having an “open ear” for any problems, for creating a friendly work atmosphere, and for giving me the freedom to pursue my scientific interests.

Next, I want to thank Prof. Christine Ziegler and Prof. Markus Jeschek for acting as second and third examiners of this thesis.

Then, I also want to thank Dr. Patrick Babinger and Prof. Gernot Längst for acting as chairperson and substitute of the PhD defense, respectively.

A special thank you goes to Lukas Drexler, Sandra Schlee, Franzi Funke, and Dominik Wagner who provided feedback on the initial manuscript and figures of this thesis. Then I also would like to thank Prof. Rainer Merkl for his expert revision of our Protein Science paper.

For their support regarding bioinformatic questions, I want to thank Julian Nazet, Simon Holzinger, and Kristina Straub.

I also want to express my gratitude to all colleagues with whom I shared the office over the years, namely Michi Schupfner, Sandra Schlee, Fabian Ruperti, Flo Busch, Julian Nazet, Lenni Heizinger, Kristina Straub, Max Plach, Patrick Löffler, Mona Linde, Cosimo Kropp, Basti Pirner, Regina Hertle, Alisa Ruisinger, Sonja Fuchs, Anna-Lena Schmidt, Mona Wieland, Carina Mayer, Michi Bartl, and Barbara Hufnagl. It was nice to share the working space with every single one of you and all contributed to a warm welcoming working environment.

My thanks also go to all present and former colleagues and friends, especially Flo Semmelmann for mentoring me in the early stages of my PhD and for sharing my road cycling enthusiasm, Thomas Klein for epic chess battles, Caro Hiefinger and Franzi Funke for always taking time for a chat, Lenni Heitzinger for his relaxed attitude, Lukas Drexler for being so relaxed and smart at the same time, Enrico Hupfeld for asking me every day during my Bachlor thesis how things were, Cosimo Kropp for sharing my enthusiasm for espresso and making group meetings during COVID much more enjoyable, to Michi Bartl for preparing coffee every morning and for his incredible politeness, Laure Gauthier-Manuel for her positive attitude, and Povilas Uzdavinys for his dark sense of humor.

I also would like to thank Sandra Schlee, Jeannette Ückert, Andrea Kneuttinger, Michi Schupfner, Julian Nazet, and Max Plach for supervising my internships and theses at the Sterner lab and teaching me most of the things I know about enzymes.

A big thank you also goes to Jeannette Ückert, Sonja Fuchs, Sabine Laberer, and Christiane Endres for their amazing support in the lab, to Klaus Tiefenbach for his trouble shooting of our instrumentation, and to Claudia Pauer for organizing everything way in advance.

Then, I want to express my gratitude to all students who contributed to my research, namely Lukas Drexler and Carina Mayer during their Master theses, Lisa Schrag during her Bachelor thesis, and again Lukas Drexler, Anja Hoffmann, Johanna Rein, and Tamara Specht for the work during the respective research internships.

Besides the people from the lab, I also want to thank my friends, Simon Weishäupl, Simon Holzinger, Michi Bauer, Christian Lang, Kevin Heizler, Domi Zahnweh, Georg Heydenreich, Vero Seitz, and Kathl Fließer who accompanied me during this journey, moaned my failures and cheered my breakthroughs.

A big thank you also goes to my parents for nurturing my curiosity, for believing in my skills, and for always supporting me.

Lastly and most importantly, I want to thank Babsi for her amazing support, coaching in difficult situations, advice on big decisions, for her ability to cheer me up or calm me down, for sharing the excitement of a successful experiment and for being the most amazing partner.