# Resurrected ancestral proteins as scaffolds for enzyme engineering and evolution

Luis Ignacio Gutiérrez Rus

Doctoral Thesis

PhD Program in Chemistry 2023

UNIVERSIDAD
DE GRANADA

Supervisors

José Manuel Sánchez Ruiz

Valeria Alejandra Risso

Editor: Universidad de Granada. Tesis Doctorales
Autor: Luis Ignacio Gutiérrez Rus
ISBN: 978-84-1117-967-6
URI: https://hdl.handle.net/10481/84376

# INDEX

ABSTRACT ..... 1

RESUMEN ..... 3

INTRODUCTION ..... 5

The emergence of catalytic proteins in the origin of life ..... 5

The Dayhoff's hypothesis ..... 5

The emergence of enzymatic catalysis in the origin of life ..... 9

*De novo* enzyme design ..... 10

Computational rational design of active sites ..... 13

Minimalist design of active sites ..... 15

Design of *de novo* metal/cofactor-based catalytic sites ..... 17

Directed evolution – Understanding catalytic optimization ..... 20

Resurrected ancestral proteins as scaffolds for protein evolution and engineering ... 24

Evolutionary implications of resurrected ancestral proteins ..... 26

Biotechnological and protein engineering implications of resurrected ancestral proteins ..... 28

HYPOTHESIS AND OBJECTIVES ..... 33

RESULTS ..... 38

Chapter 1 – Engineering and optimization of a *de novo* enzymatic activity in a resurrected ancestral scaffold ..... 38

Introduction and background ..... 38

Publication I - Enhancing a De Novo Enzyme Activity by Computationally-Focused Ultra-Low-Throughput Screening ..... 44

Supporting information for: Enhancing a De Novo Enzyme Activity by Computationally-Focused Ultra-Low-Throughput Screening ..... 45

Publication II - Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme ..... 46

Supporting information for: Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme ..... 47

Chapter 2 – Characterization of resurrected ancestral TIM-barrels as scaffolds to engineer and understand the emergence of *de novo* catalysis ..... 48

Introduction and background ..... 48

Publication III - Heme-Binding Enables Allosteric Modulation in an Ancient TIM-Barrel Glycosidase ........ 52

Supporting information for: Heme-Binding Enables Allosteric Modulation in an Ancient TIM-Barrel Glycosidase ........ 53

Publication IV (Preprint) - Protection of Catalytic Cofactors by Polypeptides as a Driver for the Emergence of Primordial Enzymes ........ 54

DISCUSSION ........ 55

Ancestral proteins as scaffolds for the engineering and evolution of artificial *de novo* active sites ........ 55

Ancestral proteins as scaffolds to reveal unexpected combinations of properties in the unexplored sequence space ........ 60

Ancestral proteins as scaffolds to understand the emergence of enzymatic catalysis ........ 63

CONCLUSIONS ........ 66

CONCLUSIONES ........ 68

BIBLIOGRAPHY ........ 70

# ABSTRACT

Enzymes are extraordinary efficient natural molecular machines that catalyze chemical reactions and transformations that sustain life in all organisms. Decades of intensive research have led to significant advances in the study of enzymes. Researchers have developed sophisticated methodologies and approaches to extensively study and gain an in-depth understanding of the molecular basis of enzyme structure, dynamics, function, and regulation. As a result, it is now possible to accurately describe the physiochemical implications of every element involved in almost every enzyme's active site during the specific molecular processes that drive the catalytic reactions. Moreover, the extensive knowledge about enzymatic catalysis has allowed us to understand how enzymes have evolved during billions of years of natural selection to catalyze chemical reactions with proficient efficiency, specificity and selectivity towards the chemical transformations and their substrates. Overall, the study of enzymes has provided a fascinating window into the molecular machinery of life. But also, it has allowed researchers to accumulate a solid scientific knowledge base that enables to engineer and tune the molecular architecture of enzymes towards designing efficient artificial and modified versions of tailored enzymes for catalyzing chemical reactions of biotechnological and biomedical interest.

However, despite our deep knowledge and advanced understanding about the fundamentals and evolution of enzymatic catalysis, one elemental question remains unanswered – How enzymatic catalysis firstly emerged and evolved at the origin of proteins and enzymes? Understanding the molecular mechanisms underlying the evolutionary emergence of new enzymatic catalysis would not only be essential to understand the birth of enzymes and its implications in the origins of life. But also, it would be critical to design new biotechnological approaches inspired in these molecular mechanisms to efficiently design and generate novel enzymes to catalyze artificial unnatural chemical reactions of interest. Yet, the study of modern enzymes with the aim to shed some light on this fundamental question has not provided significant advances.

In this thesis, we propose the hypothesis that resurrected ancestral proteins might be better scaffolds than their modern counterparts to study and understand the emergence of enzymatic catalysis. Ancestral active sites and their molecular architectures would be more useful to reveal and study the minimal requirements for catalysis. But also, ancestral proteins might be better starting points for engineering novel active sites to catalyze artificial unnatural chemical reactions. Advances in both directions may help us to reveal the molecular processes that drive the emergence of new catalysis in nature. Ancestral proteins then show the potential to have a profound impact in our understanding about enzyme catalysis, with critical implications in our knowledge about the origins of life and our capacity to develop new artificial enzymes. In order to validate our hypothesis, we have performed several experiments with different resurrected ancestral protein systems aiming to evolve a *de novo* artificial active site, as well as to understand how primordial levels of cofactor-dependent catalysis are promoted in an unevolved ancestral molecular scaffold.

In the first part of the thesis, we describe the evolution of an artificial *de novo* active site, previously engineered in a resurrected ancestral β-lactamase scaffold, by means of computational and experimental low-throughput screenings. As a result, we have demonstrated how mutations in residues directly involved in a *de novo* active site or how the introduction of new additional residues in the protein sequence may improve the geometrical preorganization of the active site and generate new interactions that enhance the stabilization of the reaction transition state and promote low initial levels of activity to reach an efficient enzymatic catalysis comparable to natural enzymes. These results have direct implications in protein engineering and *de novo* enzyme design. But also, it provides new insights about the evolutionary processes that may led the early optimization of novel active sites during the emergence of enzymatic catalysis.

In the second part of the thesis, we have resurrected an ancestral glycosidase protein with a typical TIM-barrel fold that displays unusual biochemical and biophysical features. Mainly, our ancestral TIM-barrel shows the ability to bind a molecule of the redox cofactor heme in a highly flexible region of the barrel architecture. Upon heme binding, the ancestral TIM-barrel displays a general rigidification of its structure, an allosteric modulation of its natural enzymatic activity and an unnatural novel peroxidase activity based on the redox catalytic power of heme. As a result, the ancestral heme binding TIM-barrel protein demonstrates the potential of resurrected proteins as scaffolds to harbor unusual combinations of properties of evolutionary and biotechnological interest. Additionally, the study of our redox active TIM-barrel provides new insights about the role cofactor protection in the emergence of proteins and enzymatic catalysis during the origin of life.

Overall, the results presented in this thesis support the hypothesis that resurrected ancestral proteins may serve as superior scaffolds for enzyme engineering and evolutionary studies, aimed to better understand the emergence of enzymatic catalysis during the origin of life.

# RESUMEN

Las enzimas son máquinas moleculares naturales extraordinariamente eficientes que catalizan las reacciones y transformaciones químicas que sustentan la vida de todos los organismos. Décadas de investigación intensiva han logrado avances significativos en el estudio de las enzimas. Los investigadores han desarrollado metodologías y aproximaciones sofisticadas para estudiar de manera extensiva y conseguir un conocimiento en profundidad sobre sobre las bases moleculares de la estructura, dinámica, función y regulación de las enzimas. Como resultado, hoy en día es posible describir de forma precisa las implicaciones fisicoquímicas de cada elemento involucrado en el sitio activo de prácticamente cualquier enzima durante los procesos moleculares específicos que permiten las reacciones catalíticas. Además, el conocimiento extensivo sobre la catálisis enzimática nos ha permitido entender cómo las enzimas han evolucionado durante miles de millones de años de selección natural para catalizar reacciones químicas con eficiencia, especificidad y selectividad excelentes con respecto a las reacciones de transformación y sus sustratos. En general, el estudio de las enzimas ha abierto una fascinante ventana a la maquinaria molecular de la vida. Pero, además, ha permitido a los investigadores acumular una sólida base de conocimiento científico que podemos aplicar en el diseño, modificación y optimización de la arquitectura molecular de las enzimas con el objetivo de diseñar versiones artificiales y modificadas de enzimas a medida para catalizar reacciones químicas de interés biotecnológico y biomédico.

A pesar del extenso y avanzado conocimiento sobre los fundamentos y la evolución de la catálisis enzimática, sigue habiendo una pregunta elemental sin respuesta con respecto al estudio de las enzimas: ¿Cómo emergió y evolucionó la catálisis enzimática por primera vez durante el origen de las proteínas y las enzimas? La capacidad de entender los mecanismos moleculares que subyacen a la emergencia evolutiva de nuevas capacidades catalíticas en enzimas no solo es fundamental para entender el nacimiento de las enzimas y sus implicaciones en el origen de la vida. Además, es crítica para diseñar nuevas aproximaciones biotecnológicas inspiradas en estos mecanismos moleculares con el objetivo de diseñar y generar de forma eficiente nuevas enzimas para catalizar reacciones químicas artificiales no naturales de interés. Sin embargo, el estudio basado en enzimas modernas, encontradas en los organismos actuales, con el objetivo de responder a esta pregunta fundamental no logrado avances significativos.

En esta tesis proponemos la hipótesis de que las proteínas ancestrales resucitadas podrían funcionar como mejores “andamios moleculares”, en comparación con sus homologas modernas, para estudiar y comprender la emergencia de la catálisis enzimática. El estudio de sitios activos ancestrales y sus arquitecturas moleculares podría ser más útil para revelar y estudiar los requerimientos mínimos necesarios para la catálisis enzimática. Además, las proteínas ancestrales podrían ser mejores puntos de inicio para el diseño de sitios nuevos sitios activos para catalizar reacciones químicas no naturales artificiales. En este sentido, lograr avances en ambas direcciones tendría

importantes implicaciones para entender y revelar los procesos moleculares que dirigen la emergencia de nuevas catálisis enzimáticas en la naturaleza. Por lo tanto, las proteínas ancestrales muestran de potencial de tener un profundo impacto en nuestra comprensión sobre la catálisis enzimática, con implicaciones críticas en nuestro conocimiento sobre el origen de la vida y la capacidad de desarrollar nuevas enzimas artificiales. Para validar nuestra hipótesis, hemos realizado diferentes experimentos con diferentes sistemas de proteínas ancestrales resucitadas con el objetivo de evolucionar un sitio activo artificial *de novo* y de entender cómo niveles primordiales de catálisis dependiente de un cofactor son mejorados en un andamio molecular ancestral sin evolucionar.

En la primera parte de la tesis describimos la evolución de un sitio activo artificial *de novo*, previamente diseñado en el andamio molecular de una β-lactamasa ancestral mediante cribados computacionales y experimentales de bajo número. Como resultado, hemos demostrado cómo mutaciones en residuos directamente involucrados en el sitio activo artificial o cómo la introducción de nuevos residuos adicionales en la secuencia de la proteína puede mejorar la preorganización geométrica del sitio activo y generar nuevas interacciones que aumentan la estabilización del estado de transición y mejoran los bajos niveles de actividad enzimática para llegar a una catálisis enzimática eficiente comparable a la de enzimas naturales. Estos resultados tienen implicaciones inmediatas en ingeniería de proteínas y el diseño *de novo* de enzimas. Pero, adicionalmente, aporta nuevo conocimiento sobre los mecanismos evolutivos que pudieron dar lugar a la optimización temprana de sitios activos nuevos durante la emergencia de la catálisis enzimática.

En la segunda parte de esta tesis, hemos resucitado una glicosidasa ancestral que presenta un plegamiento típico en forma de barril TIM y que muestra unas propiedades bioquímicas y biofísicas inusuales. Principalmente, nuestro barril TIM ancestral muestra la capacidad de unir una molécula del cofactor redox hemo en una región excepcionalmente flexible de arquitectura del barril. La unión del hemo da lugar a un aumento general de la rigidez de la estructura de la proteína, a una modulación alostérica de la actividad natural de la enzima y a la generación de una actividad peroxidasa nueva no natural basada en el poder catalítico redox intrínseco del hemo. Como resultado, nuestra proteína ancestral con estructura de barril TIM y con la capacidad de unir hemo demuestra el potencial de la resurrección ancestral de proteínas como andamios que muestran combinaciones inusuales de propiedades con interés biotecnológico. Adicionalmente, el estudio de nuestro barril TIM con actividad redox aporta nuevos puntos de vista sobre el papel de la protección de los cofactores durante la emergencia de las proteínas y la catálisis enzimática durante en origen de la vida.

En general, los resultados presentados en esta tesis apoyan la hipótesis de que las proteínas ancestrales resucitadas pueden servir como mejores andamiajes moleculares en la ingeniería y estudios evolutivos de enzimas, dirigidos a lograr un mayor conocimiento sobre la emergencia de la catálisis enzimática durante el origen de la vida.

# INTRODUCTION

## The emergence of catalytic proteins in the origin of life

Life is powered and sustained by an intricate network of chemical reactions orchestrated by enzymes, sophisticated molecular machines that act as catalysts of biochemical reactions and exhibit proficient catalytic capabilities. Over the course of billions of years of Darwinian evolution, enzymes have been optimized by natural selection to perform a wide range of biochemical reactions with maximum efficiency, specificity, and selectivity through the precise positioning of functional groups in their active sites and the fine tuning of their scaffold dynamics. This makes one of the most beautiful testaments to how the power of natural evolution shapes biology at the molecular scale. The study of enzyme structural and functional evolution involves a variety of complex questions, some of which are fundamentally linked to the origin of life. These questions aim to understand from how well-structured and functional proteins emerged in the primitive biochemical systems, to how new catalytic activities were generated and optimized through natural evolution in the primordial enzymes. Understanding the molecular mechanisms and driving forces of protein evolution has been a longstanding goal in evolutionary biochemistry. However, critical knowledge gaps persist unsolved. This constitutes a stimulating research field which not only aims to reveal critical insights about the origins and evolution of biochemistry and life, but also serves as a source of inspiration for bioengineers who seek to imitate molecular evolution to design new proteins with different biotechnological and biomedical applications.

### The Dayhoff's hypothesis

One of the most fundamental and primordial questions in the study of enzyme evolution is how proteins emerged in the origin of life and evolved into efficient biological catalysts. There is a growing body of evidence that supports the well-accepted hypothesis that proteins emerged from short primordial peptide fragments with at least marginal stability, foldability and biological/chemical functionality, which were selected and recruited by natural selection from a random pool of polypeptides[1]. These primordial peptides could be seen as minimalist representations of functional proteins that were later subjected to evolutionary processes such as duplication, fusion, recombination or augmentation that eventually led to more complex and functional globular proteins incorporated in the most primordial proteomes (Figure 1). This hypothesis was originally proposed by Eck and Dayhoff in their 1966 seminal paper[2]. In this work the authors described the evolutionary history of ferredoxin, a protein they described as a "living fossil" that contains an internal sequence symmetry that could explain its evolutionary origin. This observation led them to postulate that ferredoxin likely evolved through consecutive duplications of a shorter protein, which itself could have emerged through repetition of even shorter and simpler peptides. Hence, they proposed that this evolutionary mechanism for the origin of proteins could be valid in different protein families, so that more modern proteins could still harbor sequence fingerprints from their primordial structures.

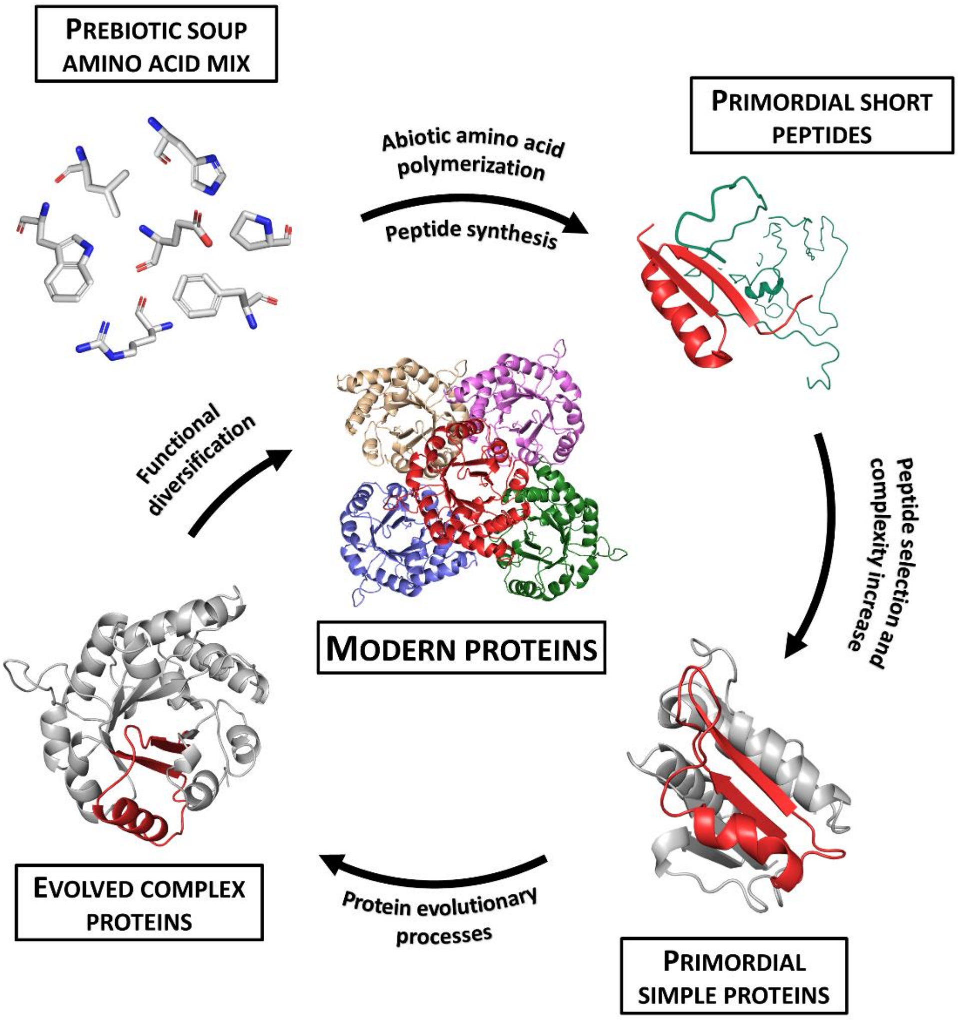

PREBIOTIC SOUP
AMINO ACID MIX
Abiotic amino acid polymerization
Peptide synthesis
PRIMORDIAL SHORT PEPTIDES
Peptide selection and complexity increase
PRIMORDIAL SIMPLE PROTEINS
Protein evolutionary processes
EVOLVED COMPLEX PROTEINS
Functional diversification
MODERN PROTEINS

**Figure 1**. Graphical representation of Dayhoff's hypothesis. First, primordial short peptides are synthesized abiotically from the condensation and polymerization of amino acids present in the prebiotic soup. Second, some representatives from the random pool of abiotically generated polypeptides are selected in terms of their stability, solubility, foldability and/or functionality. Then, by following events of duplication, repetition or fusion, primordial simple proteins with rudimentary functionalities are formed. Third, primordial simple proteins keep evolving and increase their complexity, forming more complex and evolved proteins. Fourth and last, the evolved complex proteins further evolve leading to the functional diversification observed in modern proteomes.

Eck and Dayhoff laid the groundwork for a hypothesis that nowadays is well supported as a result of the exponential growth in protein sequence databases. Numerous primordial short peptide sequences have been identified to co-occur in unrelated protein families independently of their structural or functional context. This discovery has established completely new evolutionary connections between proteins that go beyond the structural-based classifications in superfamilies and folds[3–10]. Therefore, modern analyses performed in massive protein sequence databases support the Dayhoff's hypothesis and point to the plausibility of proteins emerging in the origin of life as a result of the recruitment and further evolution of primordial functional short peptides.

However, the significance of the Dayhoff's hypothesis goes far beyond the evolutionary implications about the chemical emergence of proteins. The last universal common ancestor or LUCA is an evolutionary theoretical construct that represents a plausible living organism at the base of the phylogenetic tree that encompasses all the living forms that exist and have existed, ranging from prokaryotes, eukaryotes, and archaea, and which is inferred to have existed prior to 3.5 billion years ago[11]. Understanding and deciphering the biochemical and physiological conditions in LUCA is an extremely complex task, still unsolved, but central to understand the emergence and early evolution of life[12]. Yet, the Dayhoff's hypothesis provides important insights about how the emergence of proteins in a prebiotic scenario facilitated the emergence of the first forms of life (Figure 2). Common primordial peptides spread across extant proteomes are also easily identified in extremely old proteins folds that can be traced back to LUCA, and which are involved in core rudimentary biological and metabolic functions shared by all forms of life[7]. But most importantly, the majority of these rudimentary functions are also linked to the use of metal ions and/or (in)organic cofactors in order to perform the biochemical functions. The most common examples include the nucleotide binding P-loop β-α motif, the nucleotide binding Rossmann β-α-β motif, and iron-sulfur cluster binding motifs[7,13]. Therefore, the most relevant implication of this observation is that these cofactor dependent primordial peptides hosting fundamental biochemical functionalities also constitutes a solid evidence that primordial peptides and proteins likely promoted the transition from prebiotic chemistry and proto biochemistry by incorporating the chemical functionalities of metals and cofactors into the early proteome inferred in LUCA[14]. This functional interplay of proteins with metals and cofactors suggests a plausible role of cofactors in the chemical emergence of proteins. But also, it helps to bridge the conceptual gap between the geochemical prebiotic chemistry and the nature of the first cells at the dawn of life.

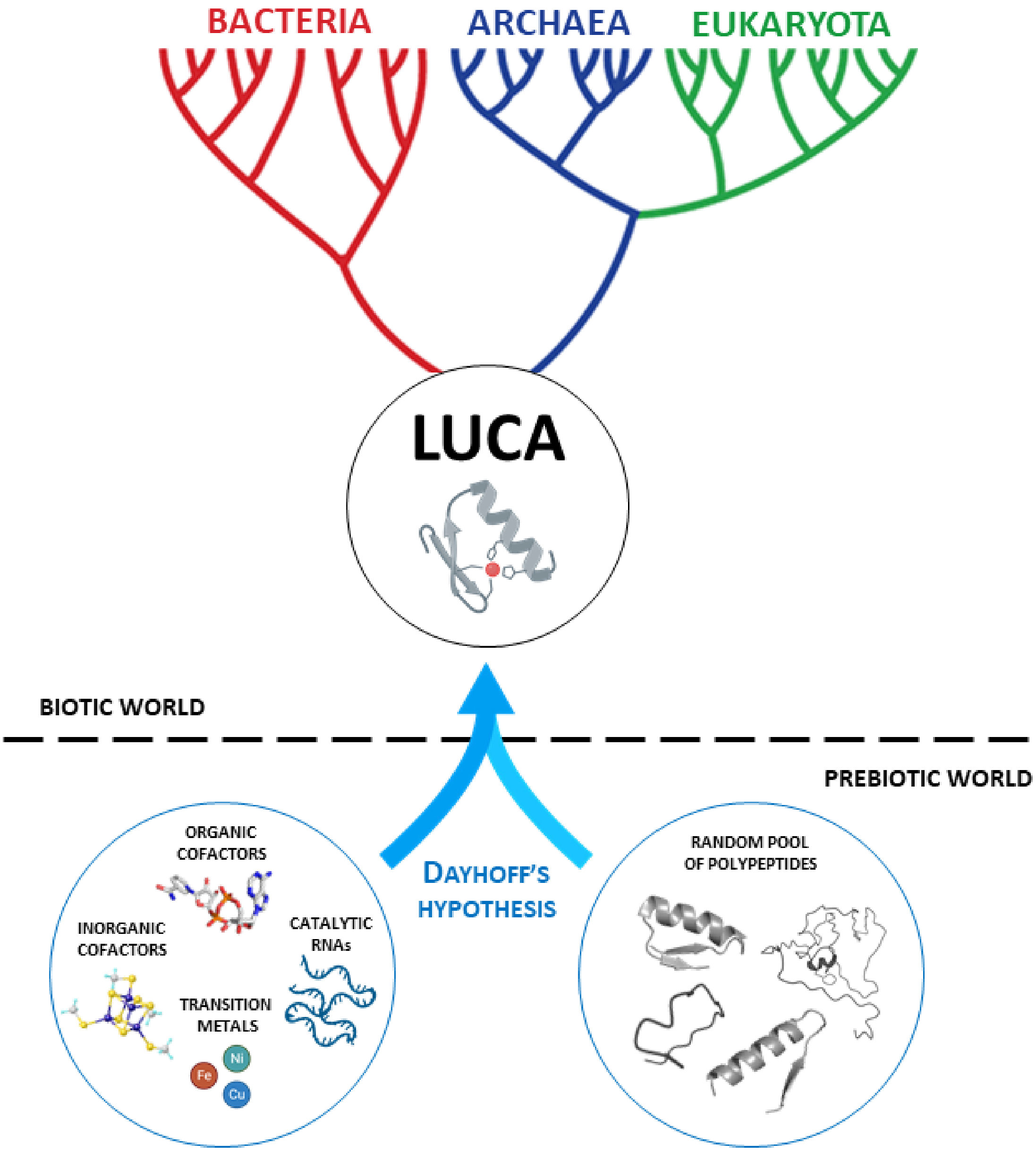


**Figure 2**. The Dayhoff's hypothesis helps to explain the transition between prebiotic chemistry and LUCA. Primordial peptides were likely selected from a random pool based on their physico-chemical properties and their (bio)chemical functionalities. The majority of the most ancient and widespread primordial peptides found across proteomes display the ability to bind cofactors or metals with (bio)chemical functionalities. Selection of cofactor/metal binding polypeptides in the context of the Dayhoff's hypothesis likely facilitated the transition between prebiotic chemistry to the primordial biochemistry of LUCA, that contained abundant cofactor/metal dependent enzymes. In this context, metals and abiotically generated cofactors likely played a key role in the chemical emergence of proteins, acting as elements with a direct influence in the early selection and evolution of primordial peptides.

## The emergence of enzymatic catalysis in the origin of life

The Dayhoff's hypothesis provides a plausible explanation for the chemical emergence of proteins and establishes a link to the origin of life. However, it does not provide answers to probably the most fundamental unsolved question in protein science: **How the catalytic power of enzymes firstly emerged and evolved in the origin of life?** A fundamental paradox in protein evolution is that nothing evolves unless it already exists, or in words of DeVries: "*Natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest*"[15]. One plausible solution to this paradox, in the context of Dayhoff's hypothesis, would consist in assuming the existence of primordial catalytic peptides and their role as minimal evolutionary modules that were selected by natural evolution and incorporated in the structure of primordial enzymes along with their pre-enzymatic chemical reactivities[16,17]. However, this does not explain how biochemical catalysis firstly emerged and evolved in the primordial peptides. An alternative plausible solution would be that once primordial peptides were selected and evolved into primordial proteins, enzymatic functionalities then emerged in the more complex protein scaffolds. In any case, it is conceivable that efficient **mechanisms for the emergence and evolution of completely novel enzymatic functionalities must have existed at some point of the primordial protein stage**, a notion that is backed up by other observations as outlined below.

Firstly, it has been demonstrated that many enzymes have arisen during evolution via divergence from pre-existing enzymes with similar catalytic machineries[18] or displaying latent promiscuous activities or broad specificity for its substrates[19–21]. Mutations accumulated in (or near to) the active sites can introduce changes in the catalytic machinery of the enzymes that modulate enzymatic activity and lead to shifts in the overall chemistry or just in the substrate specificity. This mechanism for evolution of catalytic machineries and actives sites is the basis for the emergence of novel functionalities in modern enzymes and a source of quick innovation in nature to, for example, evolve new enzymes with the capability to degrade human-made chemicals. An estimation made from the number of unique enzymes and domain superfamilies revealed that around 87% of all known enzyme functions have evolved from previous enzymatic activities[22]. However, most interestingly this estimation suggests that around 13% of all enzymatic functionalities are plausible candidates for having emerged completely *de novo* at some point during evolution from non-catalytic precursors and with no upcycling of previously existing catalytic machineries.

Secondly, genomic and phylogenetic studies aimed to reconstruct the minimal gene content in LUCA suggest the plausibility of ancestral proteomes that already contained catalytic enzymes likely involved in many different biochemical reactions[12,23]. This notion was firstly presented by Jensen in his 1976 highly influential paper[24], where he proposed that ancestral enzymes were already present in LUCA acting as multifunctional generalists with the ability to catalyze different chemical reactions on a range of substrates. Then, divergence would give rise to more specialized and active enzymes that gradually evolved into the current specialist modern enzymes. In this context, the ancestral enzymes present in the primordial proteomes could not have evolved from previously existing catalytic enzymes, but through molecular mechanisms of *de novo* enzymatic catalysis emergence.

Finally, it has been described that most biochemical reactions are extremely slow in the absence of enzymes, with half-times in the same order of magnitude as the age of the Earth[25]. The extreme inefficiency of uncatalyzed biochemical reactions indicates the necessity of enzymatic catalysis emerging in the origin of life to increase the turnover rate of biochemical reactions. This would have been critical to allow biochemical reactions to occur at the biological timescale and to promote the transition from prebiotic chemistry to proto biochemistry.

Altogether, these observations provide a convincing argument that efficient mechanisms for *de novo* emergence of enzymatic functions likely existed during the early stages of primordial proteins in the origin of life. However, they do not offer any insight into the plausible underlying molecular principles for the *de novo* emergence of functionalities, emphasizing the need for further experimentation in order to elucidate such mechanisms. Achieving a deeper knowledge about these emergence mechanisms would be key to understand key aspects of biochemistry in the origin of life. Particularly, because catalytically active proteins with the ability to catalyze biochemical functions must have been critical in the emergence of primordial metabolisms. But also, it would reveal fundamental principles of catalytic innovation in proteins that could be easily imitated in the laboratory to design and engineer completely new enzymes with artificial activities not found in nature.

# *De novo* enzyme design

As quoted by Richard Feynman in his blackboard by the time of his death[26], "*What I cannot create, I do not understand*," successful design is a great way to prove a true comprehension of the design concept. The best example of this idea applied in the field of protein science is *de novo* enzyme design, a thrilling research field that aims to generate completely new enzymes from scratch to catalyze specific chemical reactions. The main idea of this approach is to use the knowledge about enzyme structure, function, and catalysis in order to apply the same fundamental principles to efficiently design new active sites. Usually, the goals of *de novo* enzyme design are set towards the generation of enzymes that display catalytic activities with applications in biotechnology or biomedicine. However, the importance of this concept goes far beyond the direct applications of the designed enzymes. Designing enzymes from scratch is likely the most rigorous way to test our comprehension about how evolution generated and optimized the highly efficient active sites of natural enzymes. Therefore, being able to design new active sites and catalytic functions successfully and comprehensively in proteins would be critical to shed light on the molecular evolutionary mechanisms that likely promoted the emergence of the first enzymatic activities in the origin of life. In other words, if we truly understand how enzymes work, we should be able to routinely design them from scratch – But at the same time, exploring new methodologies and approaches to make *de novo* design a feasible strategy in protein science will push towards increasing our knowledge about the elementary processes that explain how enzymatic catalysis is fundamentally achieved in protein scaffolds. *De novo* enzyme design is then not just a biotechnological approach with a great potential, but it also presents important and profound evolutionary implications.

Conceptually speaking, designing *de novo* enzymatic activities in protein scaffolds requires navigating the inactive protein sequence space in a fitness landscape in order to find a combination that confers catalytic activity. This task constitutes a challenging process that heavily relies on computational methodologies based on the fundamental principles of transition state (TS) stabilization. The importance of the transition state in enzymatic catalysis was first proposed in 1946 by Linus Pauling in his seminal paper[27]. Despite lacking substantial knowledge about the structure-function relationships at the atomic level in enzymes, Pauling speculated that the active sites of enzymes should be structurally complementary to the substrate molecule in a strained configuration corresponding to the activated state during the reaction catalyzed by the enzyme. Since then, several advances in structural and functional enzymatic catalysis have confirmed Pauling's hypothesis by showing that the three-dimensional scaffold of the enzymes displays a specific configuration in which the functional groups create a well-defined pre-organized active site. This pre-organization is critical for enzymatic catalysis, as it favors the stabilization of charges in the TS and, therefore, promotes the catalytic transformation of substrates into products by lowering the activation energy barrier[28,29]. Several factors have been described to contribute for the stabilization of the TS, ranging from proximity and orientation effects (restricted motion, loss of degrees of freedom, orbital steering…) to other non-covalent stabilization interactions (electrostatic effects, desolvation, hydrogen bonding…). Additionally, in some enzymes, during the catalytic cycle, residues involved in the active site may react covalently with the substrate to generate an enzyme-substrate covalent intermediate. This covalent interaction directly alters the TS and changes the free energy profile from what it is in water, favoring the catalytic transformation[30]. In other words, the active sites of enzymes have evolved to discriminate between the ground state and the transition state of the substrate, leading to the proficient catalytic levels found in natural enzymes. If these fundamental physicochemical principles for TS stabilization are taken literally, it should be to theoretically possible to arrange a set of residues in a specific configuration and produce an active site for a particular reaction. This constitutes the basis of most modern computational methodologies aimed to design and engineer completely new active sites.

Three main computational methodologies can be identified for *de novo* generation of enzymatic actives sites, each of them relying on different approaches: rational design of active sites based in "*theozymes"*, minimalist design of active sites based in single specific mutations, and design of metal/cofactor-based catalytic sites (Figure 3). Once the new active sites are generated in the protein scaffolds, further optimization of the catalytic efficiencies can be achieved by applying different approaches. First steps of optimization usually rely on computational methodologies to rationally design new mutations in the active site in order to improve the stability and interactions of the transition state in the protein scaffold. However, optimizing the novel enzymatic activities usually involves alterations in the backbone conformation or motion that could depend on distal mutations. Therefore, optimization of the activity is better approached by using experimental directed evolution, with the aim to introduce random mutations and navigate the protein sequence space towards new maxima in the fitness landscape. All of these approaches are not just useful to generate and evolve novel enzymatic functionalities. Indeed, success in applying the principles underlying each methodology and generating an efficient novel active site leads to a deeper knowledge about how nature performed the same tasks in proteins during evolution and evolved the extremely efficient enzymes that we find in extant organisms. The details about each methodology and the evolutionary principles that can be learnt through them are discussed in the following sections.

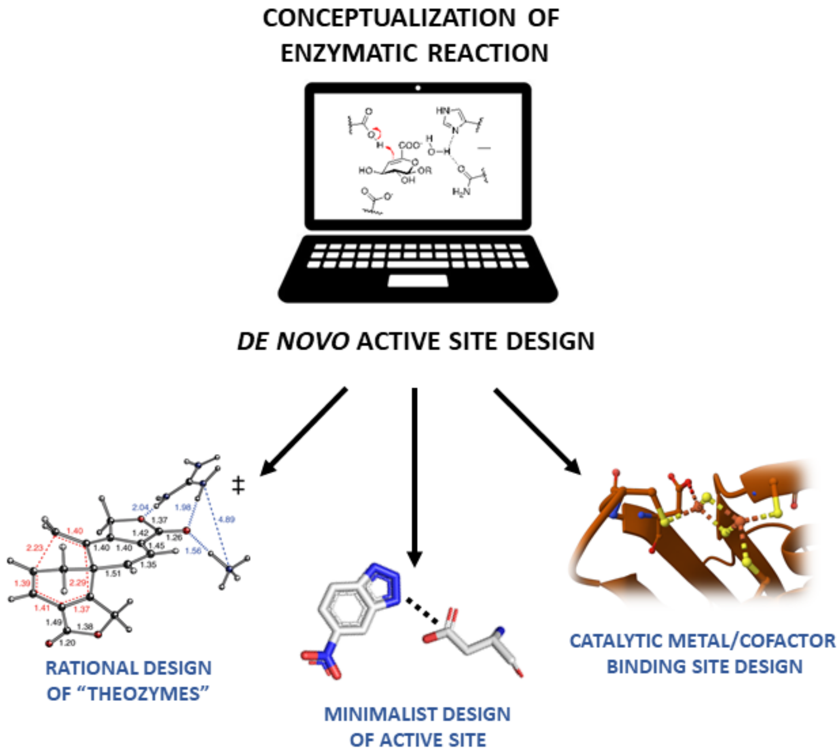


**Figure 3**. Artificial *de novo* active site design workflow. The first step requires a proper conceptualization of the enzymatic reaction to be performed to define minimal requirements for catalysis. The, a *de novo* active site is designed by following three different approaches that require different levels of computational input. (I) Rational design of "theozymes", that heavily relies on computational approaches to design an *in silico* complex constellation of binding and catalytic groups. (II) Minimalist design of active sites, based on following a rational simplistic mechanistic about enzyme catalysis that requires very low computational input to design a minimal combination of residues involved in the desired catalysis. (III) Design of metal/cofactor binding sites computationally designed to efficiently bind and capture the chemical catalytic properties of the metal or cofactors.

## Computational rational design of active sites

Computational rational design of active sites is the most common approach for *de novo* design of enzymes. This methodology is based on the fundamental principles of transition state stabilization proposed by Pauling and extended in the following decades. The main aim of this methodology consists on generating a completely new active site with a specific geometry designed to stabilize the transition state (TS) of the targeted chemical reaction, allowing to discriminate between the TS and the ground state[31,32]. Achieving an effective high degree of precision in the design is crucial for the generation of novel active sites displaying enzymatic catalysis with substantial efficiencies. In the first step of this design protocol, quantum mechanics calculations are carried out in order to design a theoretical enzyme called theozyme. Theozymes are computationally designed three-dimensional arrangements that contain the functional groups involved in the catalysis of the target reaction and placed with a high degree of precision to stabilize the TS of the reaction[33]. Once the theozyme motif is designed, the functional groups are placed in a suitable protein scaffold in which the desired catalytic reaction will be carried out. Suitable protein scaffolds have to be identified through some form of geometric filtering and ranking so that the backbone positions of the structure can accommodate the three-dimensional designed arrangement of the theozyme into a suitable structural environment[34,35]. Once the theozyme has been placed in the protein scaffold, optimization is required to achieve the extreme degree of precision needed to generate efficient active sites[36,37]. First, rotamers are generated in the chemical groups surrounding the theozyme to optimize interactions with the catalytic residues and maintaining the designed geometry. Finally, additional residues are introduced to optimize the packing, generate new interactions with the catalytic residues, tune pKa values or optimize TS binding. Iterative rounds of computational optimization are performed to sequentially optimize the theozyme by exploring the conformational space of the TS and catalytic groups to minimize the energy of each generated combination.

There are different examples of successful computational design of *de novo* enzymes to catalyze a variety of chemical reactions, including Kemp elimination[38,39], retro-aldol reactions[40], ester hydrolysis[41], Diels-Alder reactions[42] and Morita-Baylis-Hillman reactions[43]. However, the success rate of this design methodology is low, considering that computational enzyme design started as a feasible methodology around two decades ago and the number of successfully designed enzymes is very limited. Also, the catalytic efficiencies of the *de novo* designed enzymes are very low, in the range of 0.1-100 $M^{-1}$ $s^{-1}$, in comparison to the average natural enzymes[44], which display average catalytic efficiencies of $10^5$ $M^{-1}$ $s^{-1}$ (Figure 4). One of the most important limitations in computational enzyme design is that conformational sampling methods and energy functions used in the design algorithms do not achieve the sub-Angstrom scale, required to reach the extremely high degree of precision needed for an effective discrimination of the transition state from the ground state. As a result, the actual active site engineered into the protein scaffold usually differs from the one designed, displaying inaccuracies in the side chain positioning and leading to a detrimental impact on protein catalysis.

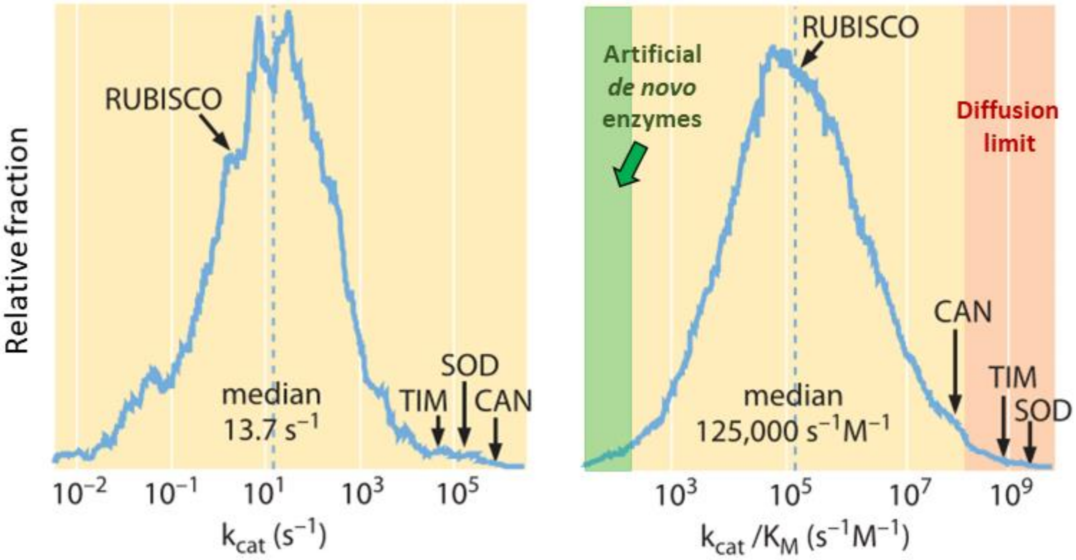


**Figure 4**. Distribution of kinetic parameters (left) $k_{cat}$ values (N = 1942) and (right) $k_{cat}/K_M$ (N = 1882) extracted from the literature about artificial *de novo* enzymes and in the BRENDA database. Only values referring to the natural substrates of enzymes were used for the distributions. Some well-studied enzymes are highlighted in the graph: CAN, carbonic anhydrase; SOD, superoxide dismutase; TIM, triosephosphate isomerase; Rubisco, ribulose-1,5-bisphosphate carboxylase oxygenase. The green zone refers to the catalytic efficiency of reported de novo artificial enzymes. The red zone refers to the catalytic efficiencies of natural enzymes in the diffusion limit. A huge gap can be immediately identified between the catalytic efficiency of natural and artificial enzymes. Adapted with permission from Bar-Even *et al*, Biochemistry, 50(21):4402-4410, 2011. Copyright 2011 American Chemical Society.

From the evolutionary point of view, the computational design of active sites has revealed critical aspects about the role of the TS stabilization in enzyme catalysis and evolution. In particular, the importance of a preorganized environment in the active site to stabilize the TS demonstrated that the catalytic power of enzymes resides in the intrinsic structural organization of the protein scaffold, rather than in the substrate-enzyme interaction. In other words, TS stabilization constitutes a chemical force that drives enzymatic reactions and promotes the chemical transformation of substrates into products in a particular enzymatic environment. Therefore, it is plausible to assume that the fundamental mechanisms of *de novo* active site generation and optimization in nature must happen following the same principle. Therefore, computational design reveals the importance of structural preorganization and precision to stabilize the TS as critical factors to understand the emergence of active sites in primordial enzymes and their optimization during evolution that led to the highly efficient levels of catalysis found in modern enzymes. Further development of these computational methodologies will enable the better characterization of the complex catalytic geometrical constraints between the TS and catalytic residues and the generation of more accurate electrostatic environments in designed enzymes. This would help to describe in more detail the fine points about how nature likely performed the same tasks naturally during evolution.

However, it is important to note that emergence mechanisms of enzymatic catalysis likely relied on single specific mutations, rather than a combination of different coupled mutations, in specific environments of the protein structures that generated catalytic minimal active sites with chemical functionalities. Therefore, computational design of *de novo* enzymes based on theozymes, which usually required a constellation of multiple coupled specific mutations that may have no effect individually, is of little utility to understand the fundamentals of this minimalistic mechanism of catalytic emergence. This idea indicates the necessity of more minimalistic methodologies to generate *de novo* active sites that better represent the plausible mechanisms for *de novo* emergence of enzymatic catalysis in the origin of life.

## Minimalist design of active sites

A different approach for the generation of *de novo* enzymatic activities is the minimalist design of active sites based on single mutations. In this case, computational approaches are also used to identify suitable structural regions in the protein scaffold and introduce single highly reactive residues to generate a new active site[45,46]. The main difference between the minimalist-based and the theozyme-based design is that only a single mutation is introduced in the protein scaffold, which confers the catalytic power. In contrast, in the theozyme-based design multiple computational calculations and introduction of many mutations must be performed in the protein scaffold to generate an efficient active site. Therefore, the minimalist design of active sites constitutes a more simplistic and generalist approach for the generation of *de novo* enzymatic functionalities in inert protein scaffolds.

In terms of the computational and experimental methodology, minimalist design is inherently rational and relies directly on the previous knowledge about the mechanism of reaction of the targeted enzymatic reaction[46]. In the first step, the catalytic site to be designed is rationally devised in order to identify the minimal requirements for catalysis to happen in terms of a specific catalytic mechanism. A suitable protein scaffold has to be identified *in silico*, aiming to find a (preferably) buried structural region in which the substrate of the reaction can bind to the protein and where the functional catalytic group has to be introduced. Docking experiments are performed in order to test *in silico* the feasibility of the targeted region and if the substrate matches the physicochemical requirements to associate with the protein at the targeted site. Then, suitable positions around the docked substrate are identified and targeted to be mutated with the catalytic active residue. The impact of the individual proposed mutations on the protein scaffold is evaluated *in silico* to identify those substitutions that do not disrupt the protein fold and do not preclude the binding of the substrate. Finally, before testing *in vitro* the final designs, the feasibility of catalysis has to be evaluated by analyzing the correct positioning and geometry of the catalytic residues introduced in the active site in order to identify the most suitable interactions with the Michaelis complex and for the stabilization of the transition state of the targeted reaction. This constitutes a much more simplistic methodology to generate a new active site in comparison to the computational rational design of theozymes, which usually require multiple specific mutations designed with a high degree of precision.

In general, the main idea of this approach is that the catalytic residue can be introduced directly with minimal computational analysis and no previous engineering on the scaffold, leading to the generation of a novel catalytic site. Successful examples of this procedure include the generation of artificial minimalistic active sites that catalyze Kemp elimination[47], retro aldolase reaction[48] or ester hydrolysis[49], all of them using calmodulin as the initial scaffold, and the generation of a *de novo* Kemp eliminase in a β-lactamase scaffold. However, in other cases, some previous subtle engineering needs to be performed in the protein scaffold in order to generate a suitable buried cavity where the catalytic residue can be placed through the standard methodology. Generally, the previous engineering efforts are aimed to mutate specific positions to replace a polar or larger residue for an apolar or smaller residue and generate the hydrophobic cavity. Then, the reactive catalytic residue is specifically placed into the hydrophobic cavity to generate the novel active site. Some examples of previously engineered hydrophobic cavities and subsequent introduction of the catalytic residue include the design of a Kemp elimination in the T4 lysozyme[50] and ester hydrolysis in a thioredoxin[51].

Overall, strictly from the engineering point of view, the simplicity of minimalist design allows for a quick and easier generation and screening of potential catalyst candidates based on the docking procedures and the transition state geometry requirements. Additionally, this methodology constitutes an extremely valuable source of evolutionary information that helps us to understand the plausible mechanisms for the emergence of enzymatic catalysis in the origin of life. First, the mechanisms and principles underlying the generation of minimalistic active sites are likely more similar to the evolutionary mechanisms used by natural evolution to drive the generation of catalysis in primordial enzymes. This is an obvious observation if one assumes that these emergence mechanisms should have relied on single point mutations in an inert protein scaffold, rather than on multiple coupled mutations that would be inactive individually. Second, a very important lesson that can be drawn from the efficient generation of minimalistic active sites is the fact that inert protein scaffolds can be just one mutation away from generating a minimal enzymatic activity and becoming entry points for subsequent evolution. This is particularly relevant when considering that minimalist design looks for the mere possibility of catalysis, rather than trying to identify highly efficient catalysts, by navigating the protein sequence space in a fitness landscape searching for a single mutation that confers a minimum amount of enzymatic catalysis. Therefore, more examples of successful minimalist design of active sites would be critical to discover fundamental principles, relying on individual mutations occurring in particular protein environments, that would help to explain the origin of enzymatic catalysis. One of the fundamental principles that can be clearly drawn from the current successful designs described in the literature is the emergence of novel active sites through the hydrophobic-to-ionizable substitution of a buried residue in a hydrophobic cavity. It is well known that (partial) burial of ionizable residues in hydrophobic regions of proteins leads to the modulation and perturbation of physicochemical properties such as the pKa value, nucleophilicity or interaction with substrates, directly linked to biological and chemical functionalities[52,53]. Additionally, is has been demonstrated that some protein scaffolds with high stability can tolerate even more than one disruptive hydrophobic-to-ionizable substitutions[54]. Therefore, correct placement of ionizable residues with perturbed physicochemical properties inside hydrophobic cavities of the protein scaffold through hydrophobic-to-ionizable substitutions would likely represent a plausible mechanism for the emergence of *de novo* catalysis in protein scaffolds.

To sum up, the major take-away from the evolutionary standpoint of minimalistic design is its significance in comprehending the molecular evolutionary mechanisms that led to the emergence of enzymatic catalysis in the origin of life. Minimalist design is based on the minimum requirements to achieve catalytic activity in an inert protein scaffold. Therefore, successful examples of minimalist design more accurately represent plausible one-step trajectories in the protein sequence space and molecular mechanisms of evolution that primordial proteins could have followed to yield completely novel enzymatic catalysis. Additionally, minimalist design constitutes a more straightforward and simpler methodology than the computational design of theozymes for the engineering of new enzymatic activities in protein scaffolds. However, the success rate in terms of correctly designed active enzymes is comparable between both approaches and the catalytic efficiencies of the designed active sites are similar, but still far from the average of natural enzymes. Consequently, this bottleneck in the number of successful designs and the low catalytic efficiencies achieved has been addressed by developing alternative *de novo* design strategies that do not solely rely on navigating the protein sequence space.

## Design of *de novo* metal/cofactor-based catalytic sites

The design of metal and cofactor binding proteins has gained popularity in the field of *de novo* enzyme design due to the wide diversity of chemical reactions that metalloenzymes have access to, including transformations not performed by natural enzymes. The main advantage in comparison to the computational design or minimalistic design of active sites is that the catalytic activity is not achieved by navigating the complex protein sequence space. Instead, the catalytic power of the *de novo* metal/cofactor dependent enzymes resides fundamentally in the chemical properties and reactivity of the metal or cofactor. The reactivity of the metal/cofactor is modulated as a consequence of the interplay between the intrinsic properties of the metal/cofactor and the surrounding protein scaffold, which can be subsequently tuned or improved through optimization methodologies based on introducing mutations in the protein scaffold.

*Design of de novo metal binding catalytic sites*

Metal ions are present in about one third of all known proteins playing key roles in structural, catalytic and electron transfer functionalities[55]. Since the beginning of enzyme design, metals have been used to generate catalytic functionalities into protein scaffolds by combining their chemical features and inherent catalytic functionalities with the outstanding control of their electronic and steric properties achieved by the protein scaffold. In particular, metal ions can be considered super (Lewis) acids, as they can display charge densities higher than +1 in neutral environments, where protons can only reach lower concentrations, and coordinate with more than just one donor atom. Metal ions are recruited in enzyme actives sites as electrostatic catalysts that contribute to the rate acceleration of enzymatic reactions by (I) charge shielding negatively charged compounds during catalysis facilitating subsequent attack of other chemical groups, or (II) by stabilizing negative charges in the reaction intermediates functioning as electrophilic catalysts. Additionally, metal ions can exist in more than one oxidation state in solution, facilitating their role in electron transfer (redox) reactions and showing reduction potentials that can be modulated depending on the nature of the coordinating residues from the protein scaffold[56]. In fact, metal ions usually display intrinsic redox catalytic power that can be transferred to a protein scaffold by designing metal binding sites.

Therefore, engineering efforts for the efficient generation of *de novo* metalloenzymes are set towards the goal of generating efficient artificial enzymes with metal dependent active sites to catalyze specific reactions. One of the most successful approaches for this purpose is the generation of completely new artificial metal binding sites in proteins, leading to the acquisition of new catalytic functionalities in an inert protein scaffold. This approach exploits and tests fundamental principles of biophysics and biochemistry with the aim to establish the criteria that allows for the generation of artificial metalloenzymes[57]. Additionally, as it happens in the rational and minimalistic design of active sites, success in designing new metalloenzymes and understanding the underlying design principles contributes to improve our knowledge about the fundamental role of the protein scaffold in binding metal ions and tuning their catalytic properties. In fact, most of the examples of designed metalloenzymes rely on minimalistic designs that introduce a single catalytic metal center by mutating just one or a few residues in the protein scaffold with minimal optimization of its properties[45]. This demonstrates that metal binding and novel catalysis (specially with redox active metals) can be easily achieved and modulated in inert protein scaffolds with just a few mutations, revealing fundamental evolutionary mechanisms by which nature could have recruited metal ions in the origin of life and generated some primordial metalloenzymes. Successful examples of *de novo* metalloenzyme design includes the generation of a wide diversity of redox and non-redox artificial enzymes, including oxygen-reactive enzymes[58], heme-copper oxidases[59], nitric oxide reductases[60], sulfite reductases[61] and glycosidases[62] in natural protein scaffolds; and phenol oxidases[63], hydrolases[64], hydroxylases[65], lyases[66], superoxide dismutases[67] and esterases[68] in *de novo* designed artificial protein scaffolds.

### *Design of de novo cofactor binding catalytic sites*

In a similar way to metal ions, the idea behind designing cofactor binding sites in proteins seeks to exploit the inherent catalytic features of cofactors in order to generate new enzymatic functionalities in the inert protein scaffold. It has been estimated that more than half of all the known enzymatic reaction mechanisms depend on at least one cofactor[69], which provides the protein scaffold with the ability to perform chemical transformations that would be inaccessible using only amino acid residues. Cofactors bear the advantage of intrinsic activity, usually extremely low in solution, that is increased and modulated when cofactors are bound in a protein scaffold that provides a different environment for catalysis. However, the design of completely *de novo* cofactor binding sites in proteins is a much more challenging task in comparison to the design of metal binding sites. The main reason of this higher difficulty is because cofactor binding in a protein environment requires more specific supramolecular interactions to achieve a high binding affinity. Computational methods to design binding sites and correctly place cofactors in protein scaffolds have been developed and applied with relative success[70–72]. However, when trying to provide the cofactor binding proteins with enzymatic functionalities the task becomes even more complex. Mainly because, not only must the cofactors be bound specifically with a particular orientation and properties, but also the interactions with the protein environment must be specifically tuned to promote catalysis. Most of the successful designs of artificial cofactor dependent enzymes rely on exploiting serendipitous cofactor binding of natural proteins, using artificial cofactors with modified properties in natural cofactor-binding proteins, or covalently anchoring reactive cofactors into proteins[70–72]. However, further progress is required for the development of cofactor dependent *de novo* active sites and, therefore, to fully understand the evolutionary mechanisms by which primordial proteins associated with cofactors in the origin of life to generate primordial cofactor dependent enzymes.

Successful examples of artificially designed cofactor binding sites include the generation of scaffolds that bind flavins[73], iron-sulfur clusters[74], chlorophylls[75], ATP[76], chlorins[77] and porphyrins[78]. However, to our knowledge, the examples of designed cofactor binding sites from its first principles in proteins and displaying relevant enzymatic functionalities are strictly limited to the design of heme binding proteins. These heme binding artificial proteins exploit the intrinsic redox catalytic power of the metalloporphyrin to provide the protein scaffold with many different natural and non-natural catalytic activities[79–84].

In conclusion, metal ions and cofactors display exceptional catalytic functionalities involved in most of the protein catalyzed biochemical reactions. Therefore, the ability to successfully design *de novo* enzymes based in metal or cofactor dependent artificial active sites, and their subsequent engineering and modulation, hold promise for the development of multiple new catalysts which mimic and/or extend the natural diversity of enzymatic reactions. However, much progress is still needed to efficiently design catalytic active sites able to bind metals or cofactors and to exploit their intrinsic catalytic powers. This will not just be critical for the efficient design of artificial enzymes with biotechnological interest. But it will also be essential to understand the evolutionary mechanisms that led to the functional association between primordial proteins and metals/cofactors.

## Directed evolution – Understanding catalytic optimization

Computational design of rational, minimalistic or metal/cofactor-based active sites usually lead to the generation of novel enzymes that usually display poor catalytic efficiencies and do not reach performance targets. Computational and rational optimization of the artificial *de novo* enzymatic activities can be effective if there is sufficient knowledge about the designed enzyme and the reaction mechanism. However, this is not the case in most of the designs where less is known or when large catalytic efficiency increases are needed to reach the levels of natural enzymes.

Imitating natural evolution to introduce random mutations in the protein scaffolds and select the best variants according to different features through experimental directed evolution has proven to be an extremely efficient strategy to improve and optimize artificial *de novo* enzymes[85–93]. Directed evolution was developed more than three decades ago as an iterative approach to engineer enzymes by incorporating random mutations in the protein sequence[94,95]. The typical directed evolution cycle relies on a general four-step methodology that presents many different variations (Figure 5): (I) generation of molecular diversity by introducing and combining random mutations in the coding DNA in order to generate mutant libraries, (II) transform a suitable host with the mutant libraries DNA in order to express the mutant variants of the protein to evolve, (III) screening of the mutant libraries to identify variants with improved phenotypes (enhanced catalytic activity for enzymes), and (IV) selection and characterization of the best candidates for a second cycle of evolution[96]. This evolution cycle is repeated iteratively in order to accumulate mutations and improve the properties of the targeted enzyme. For the case of *de novo* active sites optimization, usually multiple rounds of mutagenesis and screening are needed to achieve large increases of the catalytic efficiencies.

The main advantages of directed evolution for enzyme engineering is that requires no detailed knowledge about the enzyme structure or catalytic mechanism, and it can be used to identify improving mutations throughout the enzyme sequence, not just close to the active site. However, one of the main limitations is that usually just a tiny fraction of the vast protein sequence space is explored, skewing and restricting the potential of this methodology. In order to solve this issue, hybrid directed evolution approaches have been developed in order to exploit available structural and functional information about the enzyme and the benefits of sampling mutations throughout the enzyme sequence for designing smaller higher quality libraries[97]. Furthermore, ultra-high throughput screening methods that combine cell-surface display, microfluidics and fluorescence-activated cell sorting methodologies[98–100], and *in vivo* continuous evolution methods[101] have been developed to increase the sampled sequence space. In addition to the *in vitro* experimental directed evolution approaches, quantum mechanics/molecular mechanics (QM/MM) and molecular dynamics (MD) simulations[102–107], as well as the recent development of machine-learning algorithms[108–112], have become invaluable tools for the analysis and prediction of the impact of mutations on enzyme structure, stability, and activity.

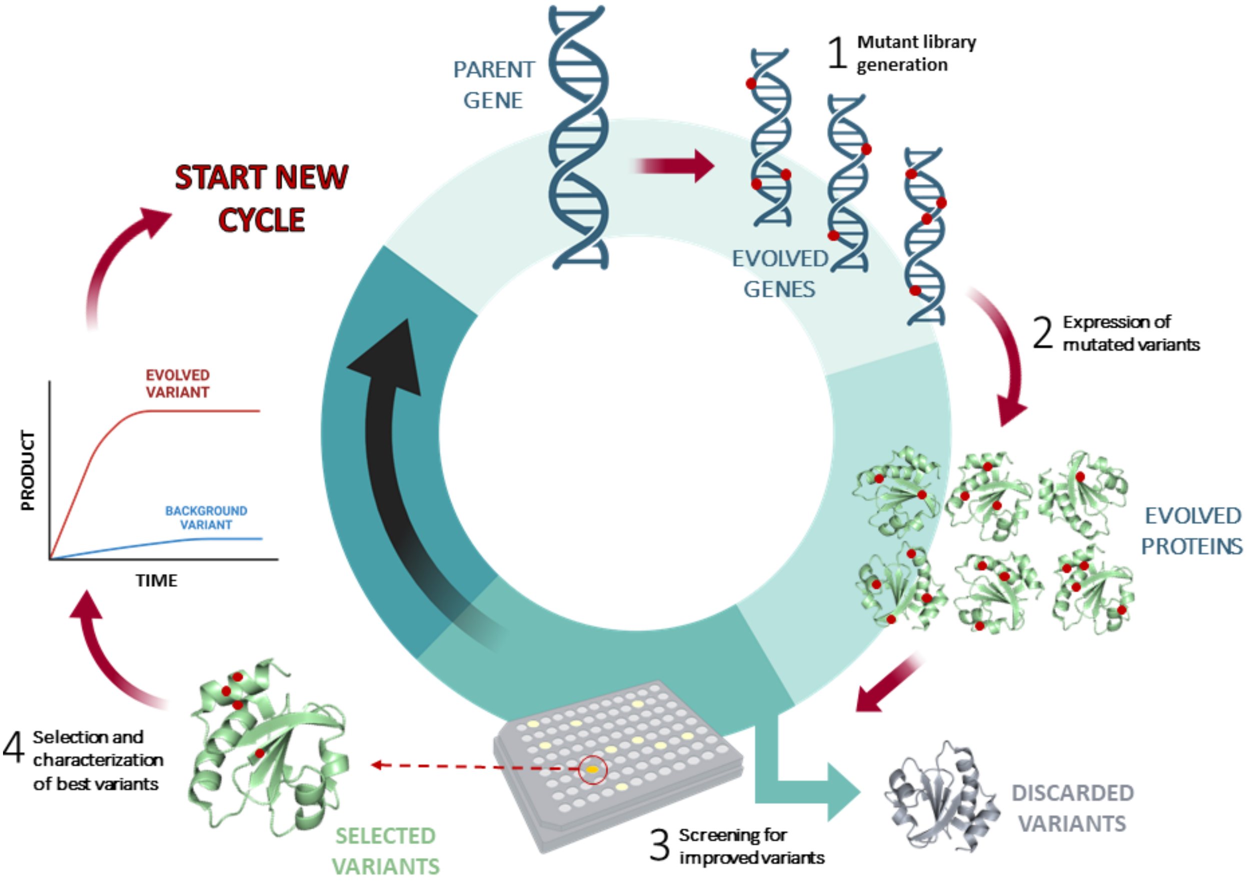

PARENT GENE
1 Mutant library generation
START NEW CYCLE
EVOLVED GENES
2 Expression of mutated variants
EVOLVED VARIANT
PRODUCT
BACKGROUND VARIANT
TIME
EVOLVED PROTEINS
4 Selection and characterization of best variants
DISCARDED VARIANTS
SELECTED VARIANTS
3 Screening for improved variants

**Figure 5.** The workflow cycle for the directed evolution of enzymes. The directed evolution of enzymes generally relies on four different steps repeated iteratively in order to improve the catalytic activity (or other properties) of the targeted enzyme. Step 1, generation of mutant libraries by introducing random mutations in the parental gene. Step 2, expression of the evolved enzyme variants harboring mutations that may affect its functional properties. Step 3, *in vitro* screening of mutant enzymes in order to identify enhanced variants. Step 4, selection of improved variants and confirmation by purifying and characterizing the hits from the directed evolution round. Positive hits are then subjected to the next round of evolution in order to accumulate enhancing mutations and further improve its functional properties.

In summary, advances in the methodology of directed evolution have allowed for the development of new strategies to design novel efficient biocatalysts. While most of the studies are set towards the evolution of natural enzymes to improve or optimize their catalytic properties, directed evolution has proven to be an efficient methodology to optimize artificially designed active sites. This constitutes the first step to eventually unlock the possibility to routinely create state-of-the-art protein catalysts to perform reactions unknown in nature, starting either by exploiting natural promiscuous reactivities or by designing completely novel active sites and evolving the initial catalysts to reach maximum efficiencies. However, further work is yet to be done particularly to develop more advanced methodologies that allows for the efficient and rapid generation of novel active sites, targeted for tailored artificial chemical reactions[113–115].

### *Evolutionary implications of directed evolution*

In addition to the important advances achieved from the standpoint of protein engineering, directed evolution is also a valuable source of evolutionary information that allows researchers to explore and study in a short scale of time the evolutionary processes that drive protein evolution in nature. This is possible because natural evolution and experimental directed evolution are based on the same fundamental principle: the impact of mutations in structure and function evolution. Over the last decades, protein engineers have performed hundreds of directed evolution experiments to improve the properties of many different enzymes. The results of these experiments offer empirical lessons and substantial insights about how proteins evolve in nature, revealing fundamental mechanisms and principles of molecular evolution that explain the possible pathways of adaptive protein evolution[116]. Understanding these mechanisms and principles are not just critical to understand how the catalytic power of enzymes emerged and evolved at the primordial protein stages, but they can also be experimentally mimicked to drive the evolution of artificial or natural enzymes towards different purposes and applications. Below, the most relevant insights according to the thesis' scope are outlined.

Directed evolution of natural proteins and enzymes has shed some light about the evolutionary pathways that drive the evolution of proteins and explain the generation and optimization of functionalities. The first and most obvious lesson that we can learn from experimental evolution is the fact that many functional properties can be easily improved incrementally through single random mutations of the protein sequence. Directed evolution experiments usually classify mutations as beneficial, neutral or deleterious depending on how they affect the targeted property. It has been determined that the vast majority of mutations are neutral or deleterious, while just a small fraction of them is beneficial[116]. However, while single mutations easily lead to improvements of particular properties, it usually fails to generate new catalytic functionalities in inert scaffolds[117] or on new substrates[118]. Such functional jumps are usually simply too big for single mutations and require more incremental paths that rely on previous neutral mutations, which leads to the second lesson. Many beneficial effects of single mutations are contingent to the existence of previous neutral or even slightly deleterious mutations. This directly points to the existence of epistatic coupling between mutations, which has been demonstrated to be ubiquitous in protein evolution. Such epistasis usually occurs through simple mechanisms that can be easily understood. The most relevant example is the stability-mediated epistasis, where stabilizing neutral mutations that do not affect the target property increase the protein's robustness to other subsequent beneficial mutations that may decrease the protein stability[119–121]. These results directly indicate that stabilizing mutations increase evolvability in a protein scaffold by the same mechanism that they increase mutational robustness[121]. In a natural context, it appears that during protein evolution stabilizing mutations can be accumulated through neutral drift that eventually facilitates the generation of a new functionality through a destabilizing beneficial single mutation.

Additionally, directed evolution has been widely applied to *de novo* artificial enzymes in order to evolve the designed active sites and increase their catalytic efficiencies[122]. Analysis of the evolutionary trajectories of *de novo* enzymes can enhance our understanding about how new catalytic functionalities and the architecture of active sites evolve in natural evolution. One of the most surprising observations is the fact that in some *de novo* artificial enzymes directed evolution led to a complete remodeling of the original active site by introducing new sets of coordinated catalytic residues[88], and even to the emergence of a completely new catalytic machinery that substitutes the original one[87]. These results suggest that analyzing the architecture of modern natural active sites may not be a valid approach to study their molecular origins, as changes through natural evolution could have completely replaced the original architectures and machineries. Another important observation about how novel active sites evolve involves the role of evolution in reshaping protein flexibility and dynamics through proximal and distal mutations from the active site. Specifically, some *de novo* enzymes increase their catalytic efficiencies by rigidifying catalytic residues and backbones in the active sites, which leads to a gradual narrowing of the conformational ensemble to favor and stabilize conformations that facilitate catalytically productive sub-states[90,123–126]. The ultimate outcome of the evolution of the enzyme dynamics is that *de novo* active sites evolve from flexible to more rigid architectures that improves the preorganization in the active site. All of these results exemplify the link between protein dynamics with catalysis and evolvability.

Despite the abundant information that can be learnt from directed evolution experiments about the natural evolution of proteins, it is important to note that these conclusions should be taken with care. *In vitro* directed evolution experiments offer empirical lessons about how proteins evolved in the face of controlled and defined laboratory selection pressures, but it is unclear if similar selection pressures apply in both natural and laboratory conditions. Experimental assays are relatively insensitive to some functional properties that are not measured in the assay but could have important impact on the biological fitness. In other words, proteins evolve in nature under pleiotropic constrains which are not present in experimental directed evolution, as they must function in a cellular context while minimizing negative impacts on other cellular components or pathways. Additionally, experimental directed evolution imposes a strong selection for a targeted property of the protein, such that mutations that benefit the selected property may be selected in detriment to other functional properties. Therefore, these issues will always to some degree raise speculation about whether the lessons learnt from directed evolution can be directly applied to understand natural protein evolution.

# Resurrected ancestral proteins as scaffolds for protein evolution and engineering

The term “ancestral proteins” directly refers to proteins from extinct organisms that existed in the past. The study of ancestral proteins constitutes an enriching source of evolutionary information that provides new insights about the evolution of living beings, but also about the natural history of Earth. The field of paleoproteomics approaches this issue by recovering ancient proteins from fossils, biological remains or other residues and analyzing them to determine their sequences[127]. However, even though the oldest recovered proteins from millions of years ago outlast the oldest surviving DNA, their longevity is just enough to cover and study a tiny part of the natural history.

Ancestral sequence reconstruction (ASR) is a methodology that relies on the analysis of modern protein sequences to infer plausible approximations to ancestral protein sequences and allows for their further preparation in the laboratory for functional and structural characterizations. The possibility of inferring the plausible sequences of ancestral proteins was first proposed by Linus Pauling and Emile Zuckerkandl in their 1963 seminal paper[128]. These authors proposed that by comparing the sequences of homologous proteins in modern organisms, reasonable estimations of sequences from proteins in common ancestors could be obtained. However, their proposal remained only a theorical possibility mainly because of the lack of protein sequences from extant organisms. ASR started to become a feasible methodology from the 1990s[129,130], when advances in bioinformatics methods, affordable costs of DNA sequencing and synthesis, and the development of models to understand protein molecular evolution started to become available. Furthermore, the development and advances in molecular biology techniques allowed the actual preparation of the encoded proteins in the ancestral sequences to be prepared and experimentally characterized in the laboratory.

The procedure of ancestral sequence reconstruction and subsequent ancestral protein resurrection comprises several steps and involves different methodologies ranging from pure bioinformatic analysis to extensive molecular biology work (Figure 6). Details about the computational methodologies, software and bioinformatic analysis used in the process of ASR has been detailed in relevant publications[131] and will not be discussed in this thesis. In general, ASR methodology is similar to the process of reconstructing words from ancient extinct languages by comparing their modern descendant words with suitable models of language evolution[132], as proteins can be considered as words made up with an alphabet of 20 letters (amino acids). The first step of any ASR study is the generation of a multiple sequence alignment (MSA) with extant sequences. Homologous sequences from modern organisms are retrieved from protein sequence databases, representing a diverse set of homologues from different evolutionary lineages or domains of life, and aligned using MSA algorithms to establish a full comparison of their sequences in each position. Phylogenetic relationships between the homologous sequences are then established in form of a phylogenetic tree by means of a distance-based maximum likelihood or Bayesian approach. The alignment and phylogenetic tree are manually fine-tuned in an iterative way by identifying and purging sequences that contain insertions or deletions, incomplete sequences, or sequences that do not represent true homologs, in order to guarantee the quality of the MSA and phylogenetic tree. Once the MSA and phylogenetic tree are correctly built, different analysis software can be used to target specific internal ancestral nodes of the phylogenetic tree and infer their most probable ancestral sequences by means of different probabilistic models of molecular evolution. The most probable sequence for each ancestor is obtained by assigning to each position of the alignment the amino acid inferred with the highest posterior probability for that position. Finally, the most probable sequences (or alternative sequences if needed) of the ancestral nodes in the phylogenetic tree are encoded in a synthetic DNA which is introduced into a suitable overexpression plasmid. Transformation of common protein overexpression hosts, such as *E. coli* or *S. cerevisiae*, and further overexpression and purification of the encoded ancestral proteins enable the production of the proteins in the lab for their experimental characterization. Therefore, the process of comparing modern sequences and inferring the ancestral sequences is what is commonly called ***ancestral sequence reconstruction***, while the production of the encoded ancestral proteins in the lab is named in the jargon as ***ancestral protein resurrection***.

The importance of ASR in the field of protein evolution is best represented by the fact that even ancestral nodes close to the last universal common ancestor (LUCA) can be reliably reconstructed and resurrected in some studies. Up to date, and during the last 30 years of ASR research, tens of different ancestral proteins have been reconstructed and resurrected in many different studies[133,134]. Resurrected ancestral proteins have been used as molecular tools to explore and study relevant evolutionary processes and hypothesis that may help to explain the evolution of proteins[135], but also as important scaffolds with particular properties of interest for protein engineering and biotechnological applications[136,137]. Implications of ASR in both evolutionary biochemistry and biotechnology are discussed in the following sections.

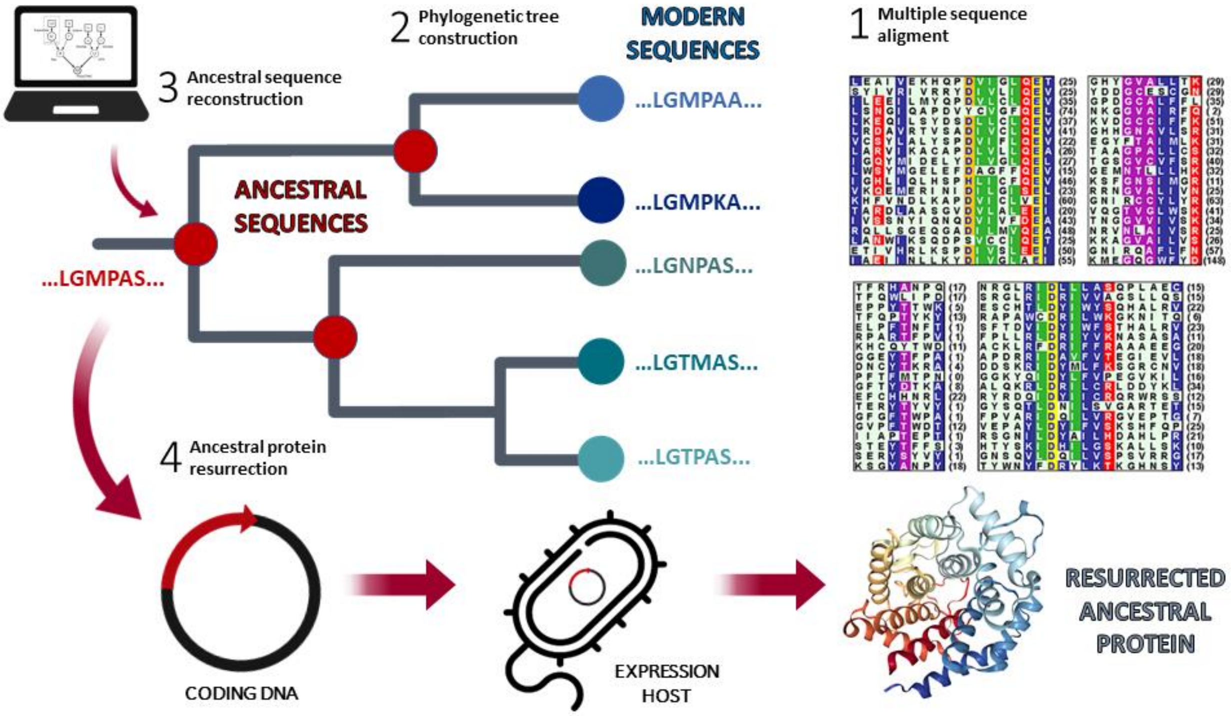

2 Phylogenetic tree construction
MODERN SEQUENCES
1 Multiple sequence aligment
3 Ancestral sequence reconstruction
ANCESTRAL SEQUENCES
...LGMPAS...
...LGMPAA...
...LGMPKA...
...LGNPAS...
...LGTMAS...
...LGTPAS...
4 Ancestral protein resurrection
CODING DNA
EXPRESSION HOST
RESURRECTED ANCESTRAL PROTEIN

**Figure 6**. The workflow for ancestral protein resurrection. The resurrection of ancestral proteins involves many different steps based on computational and experimental approaches. (1) Sequence comparison between extant protein homologs is performed with a multiple sequence alignment. (2) Evolutionary relationships between the modern sequences are established by building a phylogenetic tree based on the multiple sequence alignment. (3) The most probable sequences in the ancestral nodes of the phylogenetic tree are statistically inferred and reconstructed by applying suitable models of amino acid substitution. (4) The ancestral proteins are resurrected by synthesizing an artificial gene encoding the sequence of the ancestral protein that will be produced and purified from a suitable expression host.

## Evolutionary implications of resurrected ancestral proteins

By definition, ASR is a methodology that relies in the inference of the most probable sequences corresponding to ancestral nodes in a phylogenetic tree that represents the evolution of a protein. There is an inherent uncertainty in the prediction of the ancestral sequences[133,138–140], particularly in the case of residues with high divergence that are not involved in functionalities. However, validation can be achieved to some extent at the phenotypic level[141], since in many cases the biochemical and biophysical properties of the resurrected proteins lead to convincing evolutionary narratives and rationalize evolutionary information from different sources. This has allowed the use of resurrected ancestral proteins as molecular tools to address important problems and questions in evolution[135]. Traditionally, horizontal analysis of proteins has been applied to address problems in protein evolution, based on the sequence comparison of two related proteins to identify the specific residues responsible for different properties or functionalities. However, horizontal analyses usually fail to identify sequence differences that explain functional differences because of the intrinsic difficulty to identify a small set of substitutions that explains complex functionalities or properties, and the impossibility to directly capture the effects of epistasis in protein sequence-function evolution[142]. These limitations are easily addressed when evolution is vertically analyzed by reconstructing the sequences of ancestral proteins to understand how a protein family's sequence, structure and function changed over time and which sequence substitutions led to the shifts in the protein features[135]. Moreover, vertical analysis allows to investigate all features of sequence and structure irrespective of whether they were caused by functional optimization, historical contingency, or tinkering. In addition to the vertical approach for analyzing protein evolution, ASR allows to reconstruct proteins with estimated ages of even billions of years dating to ancestors close to LUCA and the origin of life. Characterizing such old ancestral proteins not only provides information about the biochemistry in the early steps of life but can also be used as proxies to infer the environmental conditions in which the most ancient organisms lived.

There are several ASR studies from which important evolutionary conclusions can be drawn. Some relevant examples will be briefly cited in this section to illustrate the wide range of evolutionary implications for ASR studies. Studies of sequence and structure evolution have revealed important insights about how sequence in a particular protein family evolutionary trajectory changes while maintaining the structural fold intact[143,144]. Also, the most immediate phenotypic observation that is usually made when studying resurrected ancestral proteins is that they display high thermostability in comparison to their modern counterparts[136]. Evolutionary studies in different protein families support this trend in proteins dating from the Precambrian era, which plausibly reflects the thermophilic nature of ancient life[145]. Some examples of thermostable ancestral proteins include reconstructed elongation factors[146], thioredoxins[147], β-lactamases[148], nucleoside diphosphate kinases[149] or adenylate kinases[150], all of them pointing to thermophilicity of the ancestral life. Additionally, ancestral proteins frequently display higher levels of conformational flexibility likely linked to promiscuity[148]. Studying the evolutionary trajectories of some ancestral flexible proteins revealed progressive rigidification with a direct impact on protein function and leading to more specialist descendants[151–154]. Some other ancestral features are also degraded during evolution leading to descendants that do not display the ancestral properties, such as faster folding rates[155] or unexpected ligand binding[156]. Enzymatic catalysis is the fundamental biochemical property of many proteins, and its evolution in terms of emergence, efficiency or specificity can also be addressed by characterizing ancestral proteins[148,157–161]. In particular, ASR has been successfully applied to study the emergence of catalysis in non-catalytic binding ancestors[159–161] or to study the evolution of a catalytic machinery[158]. Finally, conclusions about the biological and biochemical evolution of species can be derived from the experimental characterization of ancestral proteins. Some relevant examples regarding ancestral proteins from *Homo sapiens* include the characterization of ancestral uricases that revealed the molecular origins of gout[162] and the characterization of ancestral alcohol dehydrogenases that helped to uncover when alcohol was introduced in our ancestors' diet[163]. In general ASR has revealed many insights about protein molecular evolution and how it impacts on protein function and structure.

## Biotechnological and protein engineering implications of resurrected ancestral proteins

Resurrected ancestral proteins have proven to be valuable not only for addressing evolutionary issues but also for their potential in biotechnological and protein engineering applications. The interest of ancestral proteins for biotechnological applications and their use as scaffolds for protein engineering experiments resides in the fact that ancestral proteins are conceived as "different" from their modern counterparts. Ancestral proteins mainly differ in terms of sequence, particularly when very old phylogenetic nodes are targeted for reconstruction. In fact, reconstruction of Precambrian proteins shows large numbers of amino acid differences with their modern descendants[136]. However, in terms of biochemical and biophysical properties, resurrected ancestral proteins were likely adapted to function in intracellular and extracellular environments that completely differ from the conditions hosting modern proteins. Therefore, ancestral proteins are expected to display "unusual" or even "extreme" features that allowed them to work under the different environmental conditions of past life but are not conserved in the modern proteins. Experimental characterization of ancestral proteins has revealed the existence of unusual properties that, in many cases, match evolutionary narratives of the ancestral life. But additionally, these ancestral features of resurrected proteins are the basis of many different biotechnological applications. Therefore, ASR can be conceived as a methodology to uncover unexplored regions of the sequence space by generating new protein sequences with an evolutionary support that may display interesting properties in terms of protein engineering and biotechnological applications[133,136,137].

Many studies have focused on generating novel ancestral proteins with applications of industrial biotechnological interest. The examples include the generation of a wide range of different enzymatic systems with different applications in biocatalysis, synthetic biology, and chemical synthesis such as ancestral cellulases[164], lytic polysaccharide monooxygenases[165], laccases[166], cytochrome P450s[167], diterpene cyclases[168], haloalkane dehalogenases[169] , PETases[170], L-amino acid oxidases[171–173], ω-transaminases[174], lipases[175], ene-reductases[176], alcohol dehydrogenases[177], RNA ligases[178], carboxylic acid reductases[179] and thioredoxins[180]. Additionally, ancestral sequence reconstruction has also been applied for the generation of proteins displaying altered properties with therapeutic and biomedical applications. In some cases, displaying activities *in vivo* suitable for applications as drugs. This includes the generation of ancestral uricases[162], coagulation factors[181,182], iduronate-2-sulfatase[183], phenylalanine tyrosine ammonia-lyases[184], cytidine and adenine base editors[185], and CRISPR-associated endonucleases[186].

Besides the biotechnological and biomedical applications of resurrected ancestral proteins, the combination of unusual features usually displayed by ancestral proteins motivate their use as scaffolds for protein engineering studies aiming to generate new functionalities (i.e., novel catalytic activities) and to gain insight into their emergence. This idea constitutes the central focus of the doctoral thesis, which is deeply described and detailed in the “Hypothesis and objectives” section. The most relevant and important features that have been described in ancestral proteins within the scope of this thesis are described in the following sections, emphasizing their significance for protein engineering and biotechnological applications.

### *Enhanced expression levels*

Protein function in a cellular context involves interactions with other sub-cellular components, including other proteins. Modern proteins are adapted to interact with modern cellular environments as they have coevolved with their interaction partners. However, as resurrected ancestral proteins likely performed their functions in completely different cellular contexts, it is expected that ancestral proteins display altered patterns of interactions with other modern cellular contexts and components. The most relevant example of altered pattern is directly related with the protein folding process. Protein folding is a complex molecular process that happens *in vivo* and requires in many cases the assistance of chaperones to ensure the correct folding of the proteins. Molecular chaperones are, therefore, an outcome of molecular evolution that likely emerged at some point in order to assist the impaired folding of proteins. Ancestral proteins likely had to correctly fold *in vivo* without the assistance of chaperones, just following the most fundamental physicochemical principles of protein folding. Therefore, efficient ancestral mechanisms of correct folding may have been intrinsically encoded in the sequences of ancestral proteins. These ancestral mechanisms of unassisted efficient protein folding plausibly contribute, together with other factors, to the enhanced expression levels and solubility reported for many different ancestral protein systems that have been reconstructed and resurrected[138,164,166,168,172,185,187–192]. High expression levels are certainly convenient when preparing proteins with biotechnological applications, but also for protein engineering experiments in which high amounts of soluble protein facilitates experimentation and characterization of the engineered versions.

### *Promiscuity*

Enzymatic promiscuity can be defined as the capability of an enzyme to catalyze a biochemical reaction different from the reaction for which it has specialized through evolution[193]. On the other side, substrate promiscuity refers to the ability of an enzyme to catalyze a specific biochemical reaction against different substrates[20,194,195]. Promiscuity is a biochemical feature commonly observed in resurrected ancestral proteins. In many cases, ancestral proteins are able to catalyze chemical reactions with a wider range of substrates in comparison to their modern counterparts. A plausible evolutionary explanation for this feature is found in the widespread idea that ancient proteins were generalists with a broad substrate scope[24,196] that likely performed different chemical reactions in the context of an unevolved ancestral metabolism based in non-specialized ancestral enzymes. Duplication from these promiscuous multifunctional ancestral enzymes likely led to function partitioning and diversification into more specialist modern enzymes[140,196]. In modern proteins, promiscuity is rare but observed in enzymes that catalyze chemical reactions with a wide substrate promiscuity or even in enzymes that display low-level activities with no physiological relevance. This kind of promiscuity can be considered as a vestige of the proposed generalist nature of ancestral enzymes. In any case, ASR usually leads to the generation of ancestral multifunctional promiscuous enzymes that can be used as starting points in protein engineering campaigns to generate enzymes with new catalytic activities not found in their modern descendants. This promiscuity can be either exploited in terms of substrate promiscuity to enhance low levels of activity against interesting substrates, or to enhance minimal levels of enzymatic promiscuity to develop new catalysis in a promiscuous scaffold[197,198].

### *Increased conformational diversity/flexibility*

Promiscuity in enzymes is linked to a higher conformational diversity/flexibility in the protein scaffold[199,200]. Protein structures are not static scaffolds with a unique fixed structural conformation. Instead, protein structures should be understood as an ensemble of different conformations which are dynamically interconverting in the native state. Therefore, a higher conformational flexibility may facilitate the exploration of different catalytically active conformations of the enzyme that increase its substrate or enzymatic promiscuity by favoring the correct binding and catalysis of different substrates. The view of protein structures as an ensemble of conformations agrees with the evolutionary adaptability of proteins in which the same structural scaffold can adopt new functions[201]. Mutations accumulated during evolution shape the conformational dynamics of the proteins and shift the distribution of the ensemble towards conformations that promote the emergence of new functionalities[202,203] or the adaptation to new environments[204].

Resurrected ancestral proteins usually display a higher conformational flexibility in comparison to their modern descendants[148,151,156], directly linked to their promiscuity as a feature that allowed them to perform a higher number of biological functions in the context of a primitive metabolism. A plausible explanation for the ancestral flexibility relies on the Dayhoff's hypothesis. As proteins were sequentially formed through incrementation of their length through the proposed evolutionary mechanisms, more complex catalytically-active, but poorly structured globular structures started to form, lacking a well packed hydrophobic core and displaying a higher conformational flexibility linked to some degree of promiscuity[205]. Subsequent evolution then optimized packing and led to well-structured and more rigid modern proteins with lower conformational flexibility and higher functional specialization[206]. However, this should not distract us from the fact that conformational flexibility of the protein scaffold is an interesting biophysical feature that can be exploited in terms of protein engineering and biotechnological applications. State of the art computational methods to design *de novo* enzymes do not capture all elements of enzymatic function, and only address the role of conformational flexibility in the catalytic residues. Large-scale enzyme design efforts that take into account the role of the global conformational flexibility of the scaffold in the enzymatic activity are still not feasible, showing a lack of knowledge about the role of flexibility in protein catalysis. ASR provides a unique opportunity to address this issue through the resurrection of flexible ancestral proteins that can be used as starting point scaffolds for the design of new enzyme functionalities[136,207]. Conformational flexibility should facilitate the binding of substrates and the transition states needed to promote the enzymatic catalysis through sampling the ensemble of structural conformations, finding a minor conformation responsible for the emergence of a novel catalytic functionality that would be enriched by subsequent evolutionary optimizations[136,208]. An extensive successful application of flexible ancestral scaffolds as starting points for *de novo* enzyme design would be a turning point to achieve the routinely design of *de novo* enzymes, but it would also shed some light on the fundamental molecular features that determine the role of conformational flexibility in the emergence, evolution, and optimization of enzymatic catalysis.

### *Higher thermostability*

A remarkable number of ASR studies have reported the resurrection of ancestral proteins that display unusual thermostability enhancements, in particular when resurrecting old Precambrian ancestral nodes[146–150]. This feature most likely reflects a plausible high-temperature environment for ancient life, consistent with different evolutionary narratives that support a hot start of life and a thermophilic ancient life[209–211]. The increments in thermostability obtained through ASR are usually in the order of tens of degrees, which is larger than computational estimates of stability biases of ancestral reconstruction[212], supporting the idea that hyperstability is a genuine ancestral feature. From the biotechnological and protein engineering point of view, high thermostability is a convenient property because low stability compromises several applications for the enzymes. Actually, the increases of thermostability obtained with ASR usually outperform the enhancements achieved through rational design of stabilizing mutations or through directed evolution[213]. Most importantly, high thermostability increases the evolvability of the protein scaffold by increasing its mutational robustness facilitating to accept destabilizing mutations that may confer new functionalities to the protein[121].

Overall, ASR provides an excellent approach to explore the protein sequence space and obtain new proteins with unusual properties. In particular, new scaffolds that combine high thermostability and conformational flexibility display a huge potential, as these two features are main contributors to protein evolvability. High thermostability provides a more robust scaffold that tolerates destabilizing mutations that may generate new functionalities. Besides, high conformational flexibility allows for a more extensive exploration of conformations that may promote catalysis. Therefore, ancestral proteins that display both high thermostability and conformational flexibility could serve as superior evolvable scaffolds for the design of *de novo* enzymatic activities

# HYPOTHESIS AND OBJECTIVES

Due to their exceptional catalytic and functional properties, enzymes are highly attractive catalysts to perform specific chemical reactions in a sustainable and efficient manner, making them ideal for biotechnological applications. Consequently, enzymes stand as the key enabling technology in the field of biocatalysis. However, the main limitation of biocatalysis resides in the fact that enzymes do not capture most of the chemistry that can be practically performed. It has been estimated that around 250 different chemical reactions are catalyzed in the natural world by enzymes, while the number of different chemical reactions described in the scientific literature is more than 60,000[214]. Furthermore, the number of practically applicable chemical reactions is constantly growing and most of them are completely unnatural reactions that cannot be catalyzed by enzymes found in nature[215]. This limitation leaves biocatalysis behind in practical biotechnological applications despite of its huge potential. This problem has been approached with the development of experimental methodologies and computational algorithms which have provided biocatalysis with new tools for the design and engineering of artificial enzymes capable to perform unnatural chemical reactions[216]. More specifically, computational rational design of new active sites has allowed researchers to design and implement novel artificial enzymatic activities in inert protein scaffolds with moderate catalytic parameters. Computational optimization and experimental directed evolution have proven to be extremely useful tools to evolve and optimize the novel active sites in order to reach catalytic parameters on the same order as natural enzymes.  However, despite decades of effort and progress, the number of successfully designed and optimized *de novo* enzymes is very limited, showing that the overall problem of designing new actives sites in proteins "*a la carte*" for a given enzymatic reaction is still far from solved[217,218].

Regardless of the limitations in the practical applications of designed *de novo* enzymes, the engineering and evolution of novel active sites from scratch in inert protein scaffolds constitutes a valuable source of molecular evolutionary information. Structural crystallographic studies, kinetic analysis, and tracing of evolutionary trajectories during directed evolution of the designed enzymes, are extremely helpful approaches to uncover fundamental principles and molecular mechanisms likely followed by nature to generate new catalysis in proteins. These fundamentals and mechanisms are not just useful to understand the emergence and role of enzymatic catalysis in the origin of life, but they can also be mimicked experimentally in order to generate new enzymes with specific chemical catalysis. It is therefore conceivable to think that routinary generation of novel enzymes with artificial activities will be only plausible once we truly understand how nature performs the same tasks by following specific and defined evolutionary molecular mechanisms. However, at the same time, the efficient design, engineering, and optimization of *de novo* active sites in proteins constitutes a perfect approach to further increase our knowledge about the origins of catalysis. This immediately leads to

**the practical unsolved paradox that underlies the limitations in *de novo* enzyme design** - We need to deeply understand the natural mechanisms for emergence of enzymatic catalysis in order to replicate them and design new enzymes. But, at the same time, successful design of novel enzymes is the best way to increase our knowledge about the origins of enzymatic catalysis. This interdependence between the understanding of enzymatic catalysis and design of novel active sites underlies the limitations and pitfalls met during decades of extensive *de novo* enzyme design.

The vast majority of protein engineering efforts directed towards the generation of novel active sites as well as the studies aimed to understand the evolution of catalytic function have been performed using modern proteins as scaffolds. Despite decades of concerted effort in protein engineering and evolution experiments, *de novo* design of active sites has been achieved in modern scaffolds with a very limited success[217,218], comprising just a few examples of *de novo* enzymes that catalyze specific model reactions[38–42,47–51,91]. As a result, **the generation of *de novo* enzymes remains as a major unsolved problem in protein science**. Secondly, the natural evolution of a catalytic activity has only been observed in enzymes that evolve from previously promiscuous predecessors by recycling modern catalytic machineries[18–21]. However, **the emergence of completely novel enzymatic activities and active sites from scratch in non-catalytic scaffolds has not been observed in nature**, and therefore, the molecular mechanisms and forces that drive the emergence of catalysis in protein scaffolds remain elusive. These two observations motivate the **main hypothesis** of the doctoral thesis – We propose that use of **modern proteins as scaffolds** for engineering and evolutionary studies likely **hinders further progress** in both fields. The rationale behind this idea is that modern proteins display highly specialized and optimized scaffolds resulted from millions of years of evolution that perform specific functions under specific conditions. These highly evolved scaffolds are therefore not suitable starting points for the generation of new catalytic mechanisms, and display active sites with architectures that do not resemble to the original minimal active sites that provided the scaffolds with initial catalysis. **Instead, using resurrected ancestral proteins as scaffolds with a much lower degree of evolution and optimization, and displaying ancestral features that promote evolvability may be a more successful and fruitful strategy to approach this kind of studies**. In summary, the working hypothesis that constitutes the central axis of this doctoral thesis is the idea that resurrected **ancestral proteins may provide superior and more effective scaffolds for protein engineering and evolutionary studies**, in comparison to their modern counterparts. This general hypothesis is further detailed in the following paragraphs:

1. The efficient engineering of new catalytic activities and design of artificial active sites is likely hampered in modern protein scaffolds because of their intrinsic specialization. Modern proteins are the result of millions of years of evolution that shaped protein structure, dynamics, and the internal organization of functional groups. This led to specifically optimized protein scaffolds which have

evolved to function under specific conditions and (usually) for a particular molecular task. Therefore, engineering a new function in a highly specialized modern protein would likely be extremely difficult because the overall scaffold has already optimized its functional and structural elements through evolution for a specific function, which inherently leads to a decrease of the evolvability.

2. On the other hand, resurrected ancestral proteins may be better scaffolds for engineering as they likely display scaffolds with a lower degree of functional and structural specialization. Specifically, ancestral proteins usually display promiscuous enzymatic activities which are likely linked to a higher conformational flexibility and diversity of the scaffold. Additionally, ancestral proteins often display higher thermostabilities that facilitates the incorporation of destabilizing yet functional mutations that may generate a new functionality. Therefore, the lower degree of specialization in ancestral proteins reflected in their promiscuity and flexibility, along with higher thermostabilities, promote the evolvability in ancestral scaffolds and could facilitate the generation of novel active sites.
3. Studying molecular evolution of enzymatic catalysis may be an unproductive task in modern enzymes. Modern enzymes and their catalytic active sites have been evolving for millions of years. Multiple mutations have likely accumulated in or around the active site to modulate their catalytic properties. As a consequence, when studying and comparing the active sites in modern enzymes we can just analyze the final outcome of an intensive evolutionary process that may have completely masked, or even substituted, the original catalytic machineries and architectures that originally led to catalysis. This phenomenon has been observed during experimental directed evolution of *de novo* designed active sites[87,88]. Therefore, attempting to decipher the molecular origins of catalysis by studying modern active sites is most likely an unfruitful approach.
4. Resurrected ancestral proteins may serve as models to study catalytic active sites which most likely reflect more plausible approximations to the original catalytic machineries, specifically when resurrecting ancestral enzymes from phylogenetic nodes close to LUCA. Examination of the architecture and organization of ancestral active sites could be a more valid approach to determine the minimal requirements for enzymatic catalysis. Additionally, experimental evolution of ancestral enzymes that display conformational flexibility and analysis of the evolutionary trajectories is a viable approach to understand how evolution introduces changes in the catalytic machineries, but also in the protein scaffold by rigidifying the structures and enriching conformational states that promote catalysis. Therefore, ancestral proteins may serve as more suitable models to understand the minimal requirements for enzyme catalysis and the mechanisms that drive functional optimization during evolution.

To conclude, **the main goal of the doctoral thesis is to demonstrate the superiority of resurrected ancestral proteins as scaffolds for protein engineering approaches and their significance in evolutionary studies**. To achieve this, two resurrected ancestral systems, namely ancestral β-lactamases and ancestral glycosidases, were used as scaffolds in various experiments. The engineering and evolutionary approaches applied in these studies aimed not only to reveal the fundamental mechanisms of molecular evolution, but also to develop innovative methods for generating new protein scaffolds or evolving enzymatic activities. The experimental findings and conclusions of the doctoral thesis are presented in four research papers that approach the main hypothesis from different angles. The thesis results are divided into two chapters, each focused on a resurrected ancestral protein system, and include two research papers that address the specific research objectives. This clear structure ensures a comprehensive and organized presentation of the experimental results.

**SPECIFIC OBJECTIVE 1**: Evolution and optimization of an artificial *de novo* active site, previously engineered in a resurrected ancestral β-lactamase, that catalyzes the Kemp elimination reaction. This research aims to provide insights into the plausible molecular mechanisms by which a *de novo* minimal active site evolves to optimize a novel catalytic activity and achieve efficient levels of catalytic efficiency. Additionally, engineering efforts will be used to develop new strategies for optimizing artificial enzymatic reactions. The objective comprises of two specific goals:

- Optimization of the initial artificial *de novo* active site through an ultra-low throughput computational-guided screening. This screening aims to identify mutations in the active site that increase the catalytic efficiency. This is detailed in the first publication titled "Enhancing a *de novo* enzyme activity by computationally-focused ultra-low-throughput screening"
- Further evolution of the optimized artificial active site by generating new interactions in the active site. This will be achieved by extending the protein sequence and exploring alternative combinations of the sequence space. This is described in the second publication titled "Efficient base-catalyzed Kemp elimination in an engineered ancestral enzyme"

By focusing on these specific goals, the research will provide a structured approach to gain valuable insights into the evolution and optimization of artificial enzymatic reactions and identify novel strategies to enhance their catalytic efficiency. Ultimately, this will provide insights about the natural molecular mechanisms of enzymatic optimization.

**SPECIFIC OBJECTIVE 2**: Characterization of novel ancestral scaffolds displaying a typical TIM barrel fold in order to understand the emergence of *de novo* cofactor-based enzymatic functionality. The objective is composed of three specific goals aimed at identifying plausible molecular mechanisms for the emergence of *de novo* active sites in highly evolvable ancestral scaffolds.

- Identification of novel promising ancestral scaffolds that combine ancestral biochemical and biophysical features that promote evolvability. This will allow exploration of novel scaffolds for the development of multiple strategies for *de novo* active site generation. This will be presented in the third publication titled "Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase"
- Characterization of the unexpected binding of a cofactor in a resurrected ancestral TIM barrel and its evolutionary implications. This will help to understand the emergence of a novel active site and catalysis in a protein scaffold. This is also discussed in the third publication, as well as in the fourth and last publication included in this thesis titled "Protection of catalytic cofactors by polypeptides as a driver for the emergence of primordial enzymes"
- Use a cofactor binding ancestral protein as a model to understand the evolutionary driving forces that facilitated the association between cofactors and polypeptides in the origin of life. This will be presented in the fourth publication.

By fulfilling these specific objectives and goals, a better understanding of the emergence of enzymatic functions and active sites in ancestral proteins will be gained, by using a cofactor assisted catalysis in the resurrected scaffold as a model.

# RESULTS

## Chapter 1 – Engineering and optimization of a *de novo* enzymatic activity in a resurrected ancestral scaffold

### Introduction and background

In chapter 1 we present all the results obtained from the experiments performed in order to engineer and evolve a *de novo* active site generated in a resurrected ancestral β-lactamase used as a scaffold. Initial work on this project was previously published and demonstrated how resurrected ancestral proteins (in particular resurrected ancestral β-lactamases) may serve as better scaffolds for the engineering of *de novo* active sites and the generation of artificial catalysis[207]. The main discoveries of this paper will be discussed in the following paragraphs as an introduction before presenting the main findings of this thesis.

The main goal of this initial work was to demonstrate that resurrected ancestral β-lactamases are better scaffolds for the engineering and generation of an artificial *de novo* active site. In order to do so, single hydrophobic-to-ionizable mutations were introduced in ancestral and modern scaffolds by following a "chemical intuition" approach to generate a catalytic active site for the Kemp elimination reaction. This approach is backed up by the idea that partially buried groups with perturbed properties are often involved in catalysis[219,220]. Therefore, introduction of perturbed buried residues in suitable scaffolds may have plausibly provided a feasible route for the generation of novel active sites. Kemp elimination reaction, a benchmark reaction for rational enzyme design, was selected as a target enzymatic reaction to be catalyzed by the designed *de novo* active site (Figure 7A). To generate the novel active site a rational approach was taken by assuming that substituting a buried tryptophan residue, which shares a similar chemical structure with the Kemp elimination substrate (Figure 7B), with a negatively charged residue would create a cavity for substrate binding and introduce a catalytic buried base for the proton abstraction reaction.

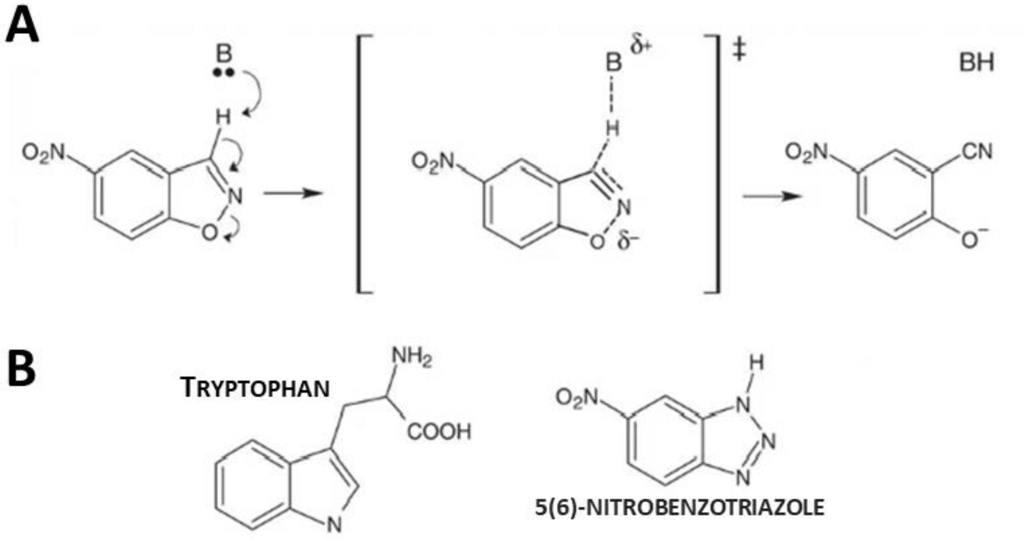


**Figure 7**. **(A)** Kemp elimination reaction. A catalytic base (represented as B) catalyzes the proton abstraction from the benzisoxazole ring of the substrate (5-nitrobenzisoxazole), resulting in ring-opening forming the cyanophenol product. **(B)** Structural comparison of a tryptophan residue and a transition state analogue of the Kemp elimination reaction 5(6)-nitrobenzotriazole. Structural similarity suggests that the rational substitution of a buried tryptophan with a negatively charged residue would generate a binding pocket for the Kemp substrate and the catalytic base for proton abstraction. Adapted from Risso VA *et al*, Nature Communications, 8:16113, 2017.

To test this idea, several resurrected ancestral and modern β-lactamases were assessed as starting points to evaluate the rational design. Additionally, conformational flexibility was used as a biophysical feature to guide the engineering approach, as it has been previously described to be a common feature of resurrected ancestral β-lactamases likely linked to functional promiscuity[199,200]. Therefore, conformational flexibility in the ancestral and modern protein scaffolds was assessed by nuclear magnetic resonance (NMR) relaxation determinations, revealing a higher flexibility in the ancestral Gram-negative common ancestor (GNCA) node in comparison to the modern TEM-1 β-lactamase. In particular, a large accumulation of residues with relaxation rates that include a conformational exchange contribution was identified in a particular region of the GNCA ancestral scaffold encompassing helix h1 (residues 26–41), helix h11 (residues 271–290) and the loops 225–229 and 252–257, which contains a buried tryptophan residue at the position 229 (Figure 8A). Therefore, residue W229 was selected as a suitable target for the generation of a *de novo* active site for the Kemp elimination reaction through a hydrophobic-to-ionizable residue replacement. Simple W229D replacements in most of the resurrected ancestral β-lactamases led to substantial levels of Kemp elimination, particularly in the alternative ancestral node GNCA-4. In contrast, W229D variants built in 10 different modern β-lactamases led to undetectable levels of Kemp elimination. The high levels of Kemp elimination in the GNCA-4 scaffold were linked to the higher conformational flexibility of the ancestral scaffold (Figure 8B). This is an intuitive idea if considering the shape similarity between a tryptophan residue and the Kemp elimination substrate. Replacement of a tryptophan with a basic residue will not generate an active site unless conformational rearrangements facilitated by an enhanced conformational flexibility allow alternative configurations to avoid clashes between the substrate and the catalytic residue. X-ray crystallography studies and molecular dynamics simulations indicated that the conformational flexibility of the ancestral node was critical to facilitate the correct binding and positioning of the substrate and transition state in the active site to facilitate catalysis (Figure 8C).

The design of a minimal *de novo* active site through a single hydrophobic-to-ionizable substitution W229D was further improved with a second single mutation F290W that facilitated the stabilization of the transition state through a face-to-edge interaction with the new tryptophan residue. In comparison with other artificial Kemp eliminases described in the literature, most of the minimalistic variants designed in the resurrected ancestral β-lactamases scaffolds displayed higher levels of catalytic efficiency than other minimalistic or rational iterative computational designs using modern proteins as starting points (Figure 9). Additionally, the best variant W229D/F290W designed in the ancestral node GNCA-4 displayed a catalytic efficiency less than two orders of magnitude lower than the best Kemp eliminase reported by the time of this work. However, it is important to note that the best design was the outcome of a much more difficult and complex rational design process based in a computational theozyme which was further improved after 17 rounds of directed evolution.

In conclusion, this work demonstrated that resurrected ancestral proteins displaying specific biochemical and biophysical features related with a higher evolvability might generally be better scaffolds for *de novo* active site generation than their modern counterparts. In particular, this study revealed how single mutations can lead to novel active sites in specific structural regions of a protein scaffold. More specifically, substitutions of buried hydrophobic with ionizable residues with perturbed physicochemical properties appears as a plausible mechanism for the emergence of novel catalysis. This supports the idea that minimalistic design of novel active sites is a valid approach to reveal evolutionary mechanisms of catalytic innovation. Additionally, the results endorse the notion that conformational flexibility is essential to assist the emergence and subsequent evolution of new active sites by facilitating and improving substrate and transition state binding, through the sampling of a diverse ensemble of potentially productive conformations.

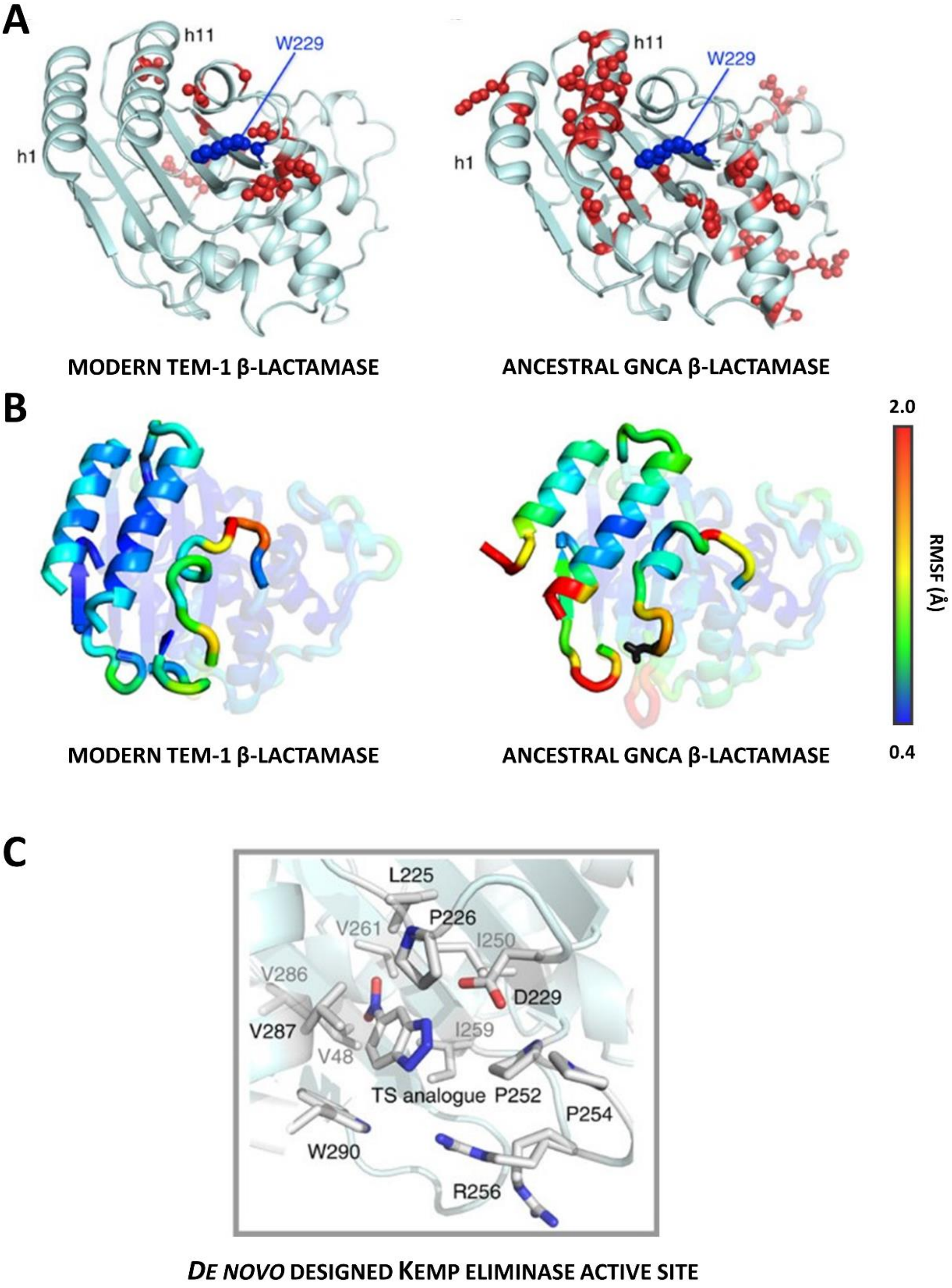


**Figure 8**. **(A)** Structures of the modern TEM-1 β-lactamase on the left (PDB 1BTL) and the ancestral β-lactamase GNCA on the right (PDB 4B88). Residues with relaxation rates that include a conformational exchange contribution are highlighted in red. Note that most of them are located in the region encompassed by the h1 and h11 α-helices. **(B)** Same structures colored by the calculated root mean square fluctuations (RMSF) of their Cα atoms, with a color scale on the right to show the values in Å. This figure represents the relative mobilities of the ancestral and modern scaffold and highlights the higher conformational flexibility in the ancestral GNCA β-lactamase that facilitates the *de novo* catalysis. **(C)** Blow-up of the de novo designed Kemp eliminase active site design W229D/F290W on the ancestral GNCA4 scaffold with the bound transition state (TS) analogue. Residues surrounding and interacting with the TS analogue are highlighted. Adapted from Risso VA *et al*, Nature Communications, 8:16113, 2017.

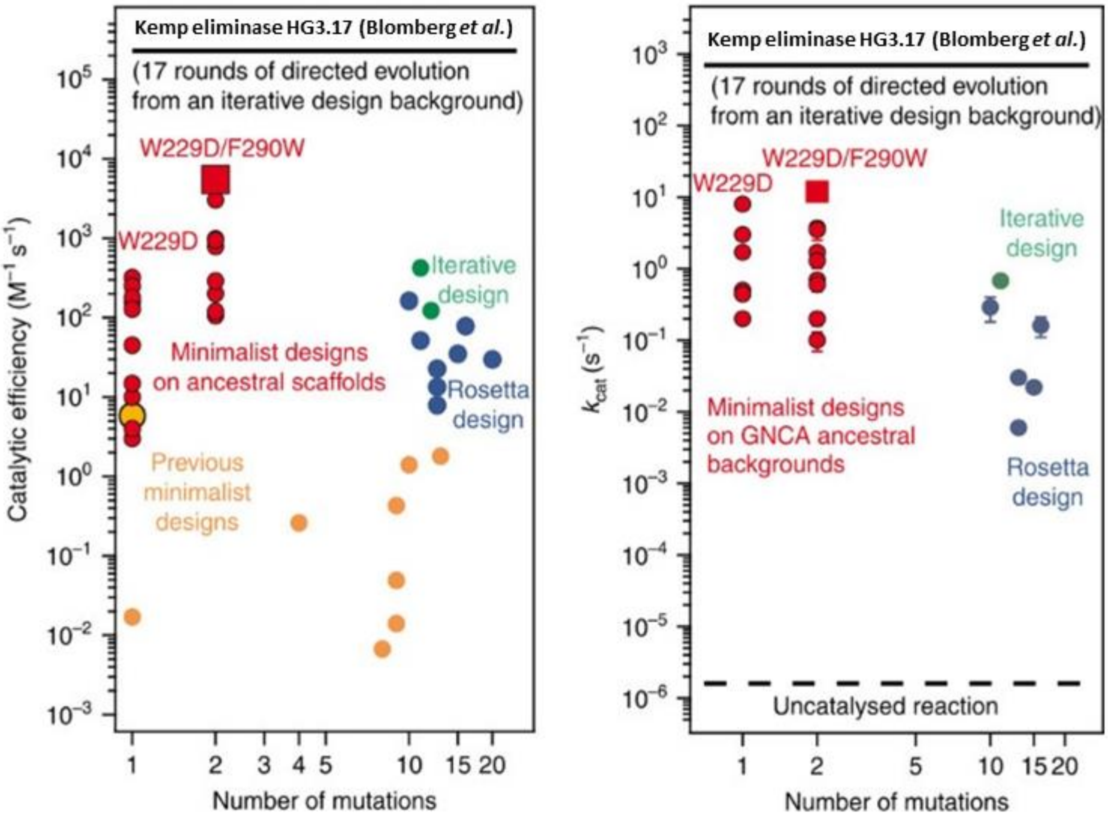


**Figure 9**. Comparison of the Kemp eliminase catalytic efficiencies on the left ($k_{cat}/K_M$) and turnover numbers on the right ($k_{cat}$) of artificial *de novo* Kemp eliminases designed up to 2017. In red, values for the single (W229D) and double (W229D/F290W) minimalistic designs on resurrected ancestral β-lactamase scaffolds. In orange, values for previous minimalistic designs on modern scaffolds. In blue, values for Kemp eliminases designs based on Rosetta calculations. In green, values for Kemp eliminases designs based on an iterative approach that involves subsequent optimization steps on the basis of 3D-structural information and molecular dynamics simulations. The black line of the top represents the values for the best Kemp eliminase designed by 2017 from Blomberg *et al*. The black dashed line on the bottom represents the uncatalyzed reaction for comparison with all the designs. Adapted from Risso VA *et al*, Nature Communications, 8:16113, 2017.

This work established for the first time the hypothesis that resurrected ancestral proteins might be superior starting points for the generation of novel active sites than modern proteins, and therefore better scaffolds to study and understand the natural principles of catalytic innovation. In this thesis, we have used the engineered Kemp eliminase GNCA-4 with the W229D/F290W design as a starting point to further evolve the Kemp eliminase active site, aiming to reveal fundamental mechanisms for the optimization of minimalistic *de novo* active sites. The next two sections of this chapter feature two published research papers that showcase the findings of his research[221,222]. The first paper titled “Enhancing a *de novo* enzyme activity by computationally-focused ultra-low-throughput screening” shows how the application of FuncLib, a computational methodology to guide the evolution of enzymes, leads to optimized variants of the artificial Kemp eliminase based in a resurrected ancestral β-lactamase with an improved geometric preorganization of the active site. The second paper titled “Efficient Base-

Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme” demonstrates how the artificial Kemp elimination reaction is further improved through the introduction of additional second-shell interactions via engineering of additional residues in the protein sequence and subsequent optimization via low-throughput screening directed evolution.

# PUBLICATION I

## Enhancing a *De Novo* Enzyme Activity by Computationally-Focused Ultra-Low-Throughput Screening

Chemical Science

**Showcasing research from Professors Kamerlin and Sanchez Ruiz's laboratories, Department of Chemistry-BMC, Uppsala University, Uppsala, Sweden, and Departamento de Química Física, Universidad de Granada, Granada, Spain.**

Enhancing a *de novo* enzyme activity by computationally-focused ultra-low-throughput screening

This work explores the application of a novel computational approach, FuncLib, to computationally-focus protein engineering effort on a *de novo* active site capable of Kemp elimination incorporated onto a Precambrian β-lactamase scaffold. Funclib uses a combination of phylogenetic analysis and Rosetta design to rank multi-point enzyme variant on the basis of predicted stability. Using FuncLib, it was possible to obtain a designed Kemp eliminase with a catalytic efficiency and turnover number ($\sim 2 \times 10^4$ M$^{-1}$ s$^{-1}$ and $\sim 10^2$ s$^{-1}$) in the range expected for average naturally occurring enzymes.
Artwork created by Joppe van der Spoel, Studio de Wilde Muis, with input from Irmeli Barkefors, Uppsala University.

## As featured in:

See Jose M. Sanchez-Ruiz, Shina C. L. Kamerlin *et al.*, *Chem. Sci.*, 2020, **11**, 6134.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

Registered charity number: 207890

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2020, **11**, 6134

All publication charges for this article have been paid for by the Royal Society of Chemistry

# Enhancing a *de novo* enzyme activity by computationally-focused ultra-low-throughput screening†

Valeria A. Risso,[a] Adrian Romero-Rivera,[b] Luis I. Gutierrez-Rus,[a] Mariano Ortega-Muñoz,[c] Francisco Santoyo-Gonzalez,[c] Jose A. Gavira,[d] Jose M. Sanchez-Ruiz*[a] and Shina C. L. Kamerlin*[b]

Directed evolution has revolutionized protein engineering. Still, enzyme optimization by random library screening remains sluggish, in large part due to futile probing of mutations that are catalytically neutral and/or impair stability and folding. FuncLib is a novel approach which uses phylogenetic analysis and Rosetta design to rank enzyme variants with multiple mutations, on the basis of predicted stability. Here, we use it to target the active site region of a minimalist-designed, *de novo* Kemp eliminase. The similarity between the Michaelis complex and transition state for the enzymatic reaction makes this system particularly challenging to optimize. Yet, experimental screening of a small number of active-site variants at the top of the predicted stability ranking leads to catalytic efficiencies and turnover numbers ($\sim 2 \times 10^4$ $M^{-1}$ $s^{-1}$ and $\sim 10^2$ $s^{-1}$) for this anthropogenic reaction that compare favorably to those of modern natural enzymes. This result illustrates the promise of FuncLib as a powerful tool with which to speed up directed evolution, even on scaffolds that were not originally evolved for those functions, by guiding screening to regions of the sequence space that encode stable and catalytically diverse enzymes. Empirical valence bond calculations reproduce the experimental activation energies for the optimized eliminases to within ~2 kcal mol$^{-1}$ and indicate that the enhanced activity is linked to better geometric preorganization of the active site. This raises the possibility of further enhancing the stability-guidance of FuncLib by computational predictions of catalytic activity, as a generalized approach for computational enzyme design.

Received 5th April 2020
Accepted 18th May 2020

DOI: 10.1039/d0sc01935f

rsc.li/chemical-science

## Introduction

Enzymes are green catalysts with unmatched catalytic proficiencies,[1] and with widespread applications in biotechnology as extracellular catalysts for a host of (bio)chemical processes, from organic synthesis to developing new pharmaceuticals, biofuels, or bioremediation agents, to name but a few examples (see *e.g.* ref. 2 and 3 for an overview). To be able to efficiently control the physicochemical properties of enzymes in a tailored fashion is therefore a problem with major economic implications, leading to extensive research effort in this direction.[4] However, natural enzymes have had millions of years to evolve to their modern catalytic efficiencies, and therefore mimicking this process whether *in vitro* or *in silico* is a non-trivial undertaking, in particular due to the immensity of the sequence space that needs exploring, and the very high frequency of catalytically detrimental mutations.[5,6] Directed evolution revolutionized experimental protein engineering efforts, by vastly expanding the sequence space accessible to protein engineers by several orders of magnitude, with low overhead.[7–9] Despite its many advantages, as a caveat, directed evolution is time-consuming, typically requiring many rounds of medium or high-throughput screening to achieve suitable levels of enzyme catalysis from a starting, low seed level.[10] Nevertheless, it has facilitated the development of a wide diversity of biotechnological applications of proteins.

Recent years have seen an explosion of interest also in computational enzyme design,[11–14] propelled in large part by early successes in *de novo* enzyme design through grafting

[a] *Departamento de Química Física, Facultad de Ciencias, Unidad de Excelencia de Química aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain. E-mail: sanchezr@ugr.es*

[b] *Science for Life Laboratory, Department of Chemistry-BMC, Uppsala University, BMC Box 576, S-751 23 Uppsala, Sweden. E-mail: lynn.kamerlin@kemi.uu.se*

[c] *Departamento de Química Orgánica, Facultad de Ciencias, Unidad de Excelencia de Química aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain*

[d] *Laboratorio de Estudios Cristalográficos, Instituto Andaluz de Ciencias de la Tierra, CSIC, Unidad de Excelencia de Química aplicada a Biomedicina y Medioambiente (UEQ), University of Granada, Avenida de las Palmeras 4, 18100 Armilla, Granada, Spain*

† Electronic supplementary information (ESI) available: Additional simulation details and table of the full list of variants predicted by FuncLib. See DOI: 10.1039/d0sc01935f

computationally designed active site models onto natural protein scaffolds (*e.g.* ref. 15–17, among others). We note, however, that while impressive, this approach typically generates enzymes with only modest catalytic activities, which again require many rounds of directed evolution before reaching catalytic efficiencies[10,18] that are comparable to naturally occurring enzymes.[19]

In light of the above, the use of computation to focus and speed up directed evolution is of considerable interest. Indeed, there have been substantial advances in this field, with many new screening approaches being put forward, based on sequence, structural or even dynamical information gained from simulations (see *e.g.* ref. 20–30). In addition, machine learning shows great promise as a screening tool in enzyme design studies.[31–34] Still, the best engineered enzymes, with catalytic efficiencies comparable to natural enzymes, are more often the results of intensive directed evolution efforts starting from low-activity rational designs.[10,18]

The sluggishness of the common directed evolution procedures has to do, at least in part, with the fact that most variants in a random library with a substantial mutational load will include mutations that are deleterious in terms of fundamental protein biophysical properties, such as stability and folding. FuncLib[28] is a novel automated method for designing multipoint mutations at enzyme active sites by combining phylogenetic analysis and Rosetta design calculations. FuncLib does not *per se* predict mutations that enhance catalysis, but rather suggests variants with multiple mutations that generate stabilizing interacting networks at the active site, thus focusing the search to safe regions of the sequence space. Furthermore, FuncLib can be used to target regions that are expected to be relevant for catalysis, thus avoiding the inefficiency associated with probing catalytically neutral mutations. We note here that while one might intuitively expect trade-offs between catalytic activity and stability, this is not necessarily the case *a priori*: that is, it has been experimentally demonstrated that it is possible to enhance stability through either engineering[36] or directed evolution,[37] without compromising activity. Here, we apply the FuncLib approach to the enhancement of the activity of a *de novo* enzyme activity previously generated by minimalist rational design.[35] Specifically, we recently demonstrated that a simple hydrophobic-to-ionizable residue substitution (Fig. 1) is sufficient to generate a *de novo* active site capable of highly proficient Kemp eliminase activity for the cleavage of 5-nitrobenzisoxazole in Precambrian β-lactamases obtained by ancestral inference,[35] with the best of our designs ($k_{\text{cat}}/K_{\text{M}} \sim 5 \times 10^3$ $\text{M}^{-1}$ $\text{s}^{-1}$ and $k_{\text{cat}} \sim 10$ $\text{s}^{-1}$ at alkaline pH) showing catalytic proficiencies only two orders of magnitude lower than the best designed Kemp eliminase obtained through iterative design followed by 17 rounds of directed evolution.[38]

There are a number of reasons Kemp elimination is particularly attractive as a model system for *de novo* enzyme design studies. (1) It provides a simple activated model for proton abstraction by carbon, (2) as a non-natural reaction it means that no natural enzyme has evolved to catalyse this reaction, reducing the risk of contamination from natural enzymes, and (3) for historical reasons, Kemp elimination has often been used as a benchmark for enzyme (and other catalyst) design studies,[15,35,38–47] providing extensive examples of designed constructs against which to compare our engineered β-lactamases. Certainly, Kemp elimination is a facile reaction that requires a simple catalytic machinery (essentially, a catalytic base to abstract the proton). However, and as a relevant point in the context of this work, it is difficult to generate high levels of Kemp eliminase activity because the transition state is so similar to the reactant state, both in terms of structure and in terms of charge distribution.[48] That is, structurally, the overall geometry of both the substrate and transition state are similar. Therefore, moving from the substrate to the transition state does not bring about a large change in the spatial arrangement of the interacting moieties that could be used as a basis for the preferential stabilization of the transition state. In addition, charge build-up on the ring oxygen at the transition state (Fig. 1) is highly delocalized into the aromatic ring of the substrate.[48–50] Therefore, it is highly challenging to use improved transition state stabilization or manipulate active site polarity as a means to achieve

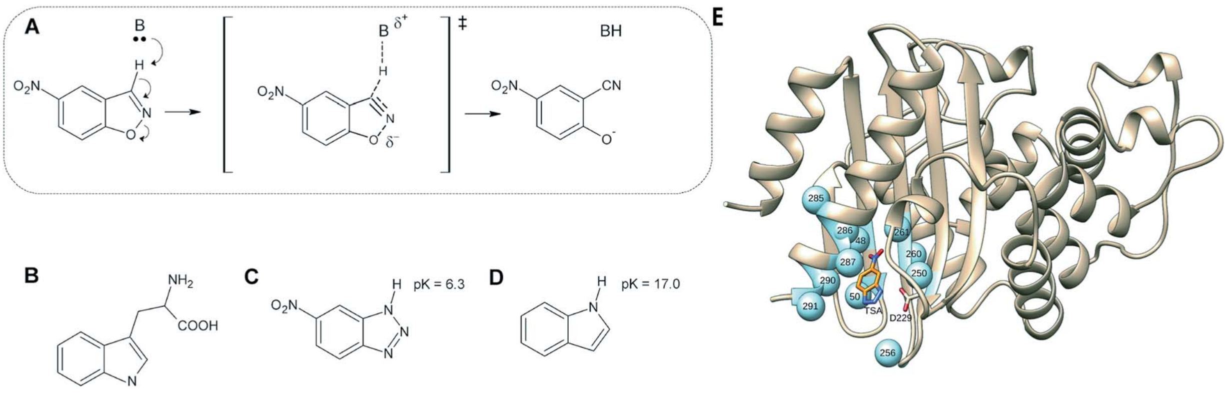


Fig. 1 (A) Kemp elimination of 5-nitrobenzisoxazole showing a proposed transition state structure. For comparison, shown here are also the structures of (B) tryptophan, (C) a transition-state analog and (D) indole. (E) 3D-structure of the background GNCA4-WT *de novo* enzyme (PDB ID: 5FQK,[35] referred to throughout as GNCA4-WT), showing both the position of the bound transition analogue, as well as the key residues we targeted using FuncLib (shown as spheres). Panels (A–D) were originally published in ref. 35. Reproduced here with permission from ref. 35. Copyright 2017, the authors. Published under a CC-BY license (http://creativecommons.org/licenses/by/4.0/).

substantial gains in catalysis. As an illustration of this, Hilvert and coworkers[50] have recently explored in detail the contribution of oxyanion hole stabilization to the highly proficient Kemp eliminase, HG3.17, and find the contribution of a key residue forming this oxyanion hole, Gln50, to be only modest, likely reflecting charge delocalization at the transition state.

Use of FuncLib allows us to consider the effect of mutations at 11 positions simultaneously, thus avoiding problems caused by epistasis which can lead to unpredictable (non-additive) effects on enzyme activity.[51–53] Remarkably, we find that screening of just 20 FuncLib predicted variants leads to substantial enhancement of our previous best Kemp eliminase. That is, experimental validation of the twenty best scoring FuncLib predictions through biochemical and structural analysis allows us to identify 4 variants with significantly enhanced catalytic efficiency and improved turnover number, the best of which reach catalysis levels ($k_{cat}/K_M$ of $\sim 2 \times 10^4$ M$^{-1}$ s$^{-1}$ and $k_{cat}$ of $\sim 10^2$ s$^{-1}$) for the cleavage of 5-nitrobenzisoxazole that compare favourably with that of naturally occurring enzymes.[19] In addition, we demonstrate that the empirical valence bond (EVB) approach[54] can reproduce the experimental free energy barriers for the optimized eliminases to within ~2 kcal mol$^{-1}$, raising the possibility of further enhancing the stability-guidance of FuncLib on the basis of EVB-based computational predictions of catalytic activity. Overall, we demonstrate a simple computational protocol with tremendous potential for biocatalysis.

## Materials and methods

### Initial screening using FuncLib

Initial design was performed using the FuncLib webserver (http://funclib.weizmann.ac.il/), as described in ref. 28. As our starting point, we selected all amino acids in close contact with the substrate for randomization by FuncLib, comprising of 11 starting positions (V48, D50, I250, R256, L260, V261, L285, V286, V287, W290 and H291, see Table S1†). The calculations were performed on Chain A of the crystal structure of the GNCA4-W229D/F290W variant (PDB ID: 5FQK,[35] henceforth referred to as GNCA4-WT), with the transition state analog, 6-nitrobenzotriazole, retained in the calculation, and the His tag removed. The multiple sequence alignment was performed using the default parameters, and the top twenty ranked designs based on their stability score were retained for further experimental and computational analysis.

### Empirical valence bond simulations

The empirical valence bond (EVB) approach[54] has been extensively used to successfully study enzyme catalysis in general,[55,56] and Kemp elimination in particular.[48,57–59] In this context, we recently used the EVB approach to study the evolution of multiple active site configurations[59] in the *de novo* designed Kemp eliminase, KE07.[15] In the present work, we follow the protocol presented in ref. 59. Our EVB simulations were performed using a simple two-state EVB model, describing the reactant and product states for the Kemp elimination reaction, with the side chain of D229 and the substrate included in the EVB region. All other residues were treated fully classically using the OPLS-AA force field.[60,61] All simulations were performed using the *Q* simulation package, version 5.10,[62] and a description of valence bond states and all EVB parameters used in the simulations are provided in the ESI of ref. 59.

EVB simulations were performed of the Kemp elimination reaction catalyzed by the GNCA4-WT β-lactamase, a series of additional single active site mutations of this variant used for calibration of the EVB simulations (G62S, A146G, A173V, L265Q, R256K, R256A), as well as the top-twenty ranked mutations predicted by the FuncLib web-server, based on both the structural predictions from FuncLib, and, where available, also crystal structures for comparison (for the three variants characterized in this work). Simulations of the GNCA4-WT variant were performed using the PDB ID: 5FQK,[35] and the best hits from the FuncLib webserver were simulated based on the PDB structures provided by FuncLib[28] with the substrate. The structures of all other variants were generated using SCWRL4.[63] In all cases, the substrate 5-nitrobenzisoxazole was manually placed in the active site in the position of the transition state analogue 5(6)-nitrobenzotriazole present in the crystal structure. Missing residues at the C- and N-termini of the protein were ignored for simplicity, and the first residue of the His-tag present in the initial crystal structure was retained for consistency (this was also the case for the FuncLib calculations).

The entire system was then solvated in a 23.5 Å spherical droplet of TIP3P water molecules,[64] centred on the CG atom of D229, and subject to surface-constrained all-atom solvent (SCAAS) boundary conditions.[65] The system was modelled using a multi-layer approach standard to such simulations in which all atoms within the inner 85% of the water droplet are allowed to move freely, the atoms in the external 15% of the droplet are restrained to their crystallographic positions using a 10 kcal mol$^{-1}$ Å$^{-2}$ harmonic positional restrained, and all atoms outside the droplet are fixed at their crystallographic positions using a 200 kcal mol$^{-1}$ Å$^{-2}$ harmonic position restraint. Only those ionizable residues that fall within the mobile region (inner 85%) of the simulation sphere were ionized during the simulations, all other ionizable residues outside the mobile region were kept in their charge neutral states to avoid instabilities introduced by having charges located outside the explicit simulation sphere. Protonation states of ionizable residues within the explicit simulation sphere, as well as histidine protonation patterns (both of which were validated by PROPKA 3.1 (ref. 66) and visual inspection), can be found in Table S2.†

All systems were subjected to an initial 3 ps minimization at 1 K using a 0.1 fs stepsize, in order to remove bad contacts in the system after solvation. During this simulation time, a 200 kcal mol$^{-1}$ Å$^{-2}$ harmonic restraint was placed on all protein and substrate atoms in the simulation to restrain them to their crystallographic positions. The step size was then increased to 1 fs for the remainder of the simulations (both equilibration and subsequent EVB simulations), and the temperature was gradually increased from 1 to 300 K while simultaneously dropping the harmonic restraints from 200 to 0.5 kcal mol$^{-1}$ Å$^{-2}$ on only the atoms in the EVB region (not taking into account the additional restraints on atoms outside the inner 85% of the water droplet). Once the system had reached 300 K, the system was subjected to a further 20 ns of equilibration. Each equilibration was performed ten times, with ten different sets of initial velocities, leading to 200 ns of equilibration time per system, and 5.4 μs of equilibration time over all systems considered in this work. The corresponding backbone root mean square deviations are shown in Fig. S1–S3.†

For each system, the endpoints of the ten equilibration runs were then used as starting structures for subsequent EVB simulations, with three additional equilibration runs of 500 ps in length being performed from each of these starting points, using new random velocities, in order to generate 30 discrete starting points for EVB simulations of each system. The EVB free energy perturbation/umbrella sampling (EVB-FEP/US) calculations were performed in 51 individual mapping frames of 100 ps simulation length each, leading to a total of 5.1 ns simulation time per individual EVB trajectory, 153 ns simulation time per system, and 4.590 μs of equilibration time over all systems considered in this work. The EVB parameters were calibrated using the uncatalyzed background reaction in aqueous solution as a baseline, as described in ref. 59. The same calibration as in our previous work[59] was used in the present study, and no new calibration was performed here with all EVB parameters used in this work presented in the ESI of ref. 59.

All simulations were performed using the Berendsen thermostat[67] with the leapfrog integrator, and with the solute and solvent coupled to individual heat baths. The bonds to hydrogen atoms were constrained using the SHAKE algorithm.[68] Cut-offs of 10 and 99 Å were used for the calculation of nonbonded interactions involving the protein and water molecules and the EVB region respectively (effectively no cut-off for the latter), and electrostatic interactions for all atoms falling beyond this cut-off were approximated using the local reaction field approach.[69] The non-bonded pairlist was updated every 30 fs. All simulation analysis was performed using the *Q*Calc module of *Q*,[62] and all structural analysis was performed using VMD version 1.9.3.[70] For full simulation details, see ref. 59.

### Protein expression, purification and library screening

The different β-lactamase variants studied in this work were purified using procedures previously described in detail in ref. 35 and 71. Briefly, genes for the His-tagged proteins were cloned into a pET24 vector with kanamycin resistance, were cloned into *E. coli* BL21(DE3) cells, and the proteins were purified by NTA affinity chromatography. Stock solutions for activity determinations and physicochemical characterization were prepared by exhaustive dialysis against the desired buffer.

Mutagenized libraries for screening studies were generated by error-prone PCR using the GeneMorph II Random Mutagenesis kit (Agilent) and transformed into *E. coli* Bl21 (DE3) and individual colonies were picked and grown in 96-well plates. The Kemp eliminase activity of ∼500 variants were assayed with 5-nitrobenzisoxazole (0.25 mM) in 96-well plates. This primary screening served to select variants that were subsequently prepared and tested on pure form. In most cases, this secondary screening implied the determination of profiles of activity *versus* substrate concentration.

### Stability determination

Thermal denaturation of the different β-lactamase variants studied in this work was studied using differential scanning calorimetry at a scan rate of 200 K per hour in HEPES 10 mM, 100 mM NaCl, pH 7 following protocols that have been previously described in detail.[71] A single transition was observed in thermograms of heat capacity *versus* temperature. Denaturation temperature values correspond to the maximum of the calorimetric transition.

### Activity determination

Determination of Kemp elimination activity were carried out at 25 °C HEPES 10 mM or 10 mM sodium phosphate (in all cases with 100 mM NaCl), depending on the pH range, as has been previously described in ref. 35. Experiments were routinely carried out in the presence of acetonitrile to increase the solubility of the substrate and expand its experimental concentration range, thus facilitating the detection of curvature in Michaelis plots and, therefore, the reliable determination of turnover numbers. 5% acetonitrile was used in most cases, although experiments with higher and lower acetonitrile contents were also performed (see the Results and discussion for details). It is to be noted that, even in those cases in which no acetonitrile is added on purpose, a small amount of the cosolvent is present because the stock solution of the substrate is prepared in acetonitrile. The approximate substrate ranges used depend on acetonitrile concentration, reflecting the substrate solubility (Table S3†).

Product formation in activity determinations was followed by measuring the absorbance at 380 nm and an extinction coefficient of 15 800 $M^{-1}$ $cm^{-1}$ was used to calculate rates. All measurements were corrected by a blank performed under the same conditions. This is particularly critical at basic pH values, where catalysis by the hydroxyl anions may lead to substantial blank values. Still, we made sure that the level of enzyme catalysis was significantly above the blanks, even at the more alkaline pHs studied.

Catalytic parameters were determined from the fit of the Michaelis–Menten equation to the experimental rate *vs.* substrate concentration profiles. As mentioned above, solubility limits the experimentally available substrate concentration range, making it essentially impossible to experimentally reach saturation. This is, in fact, a common occurrence in studies of Kemp eliminases, and should not prevent the determination of reasonable estimate of the turnover number, $k_{cat}$, provided that significant curvature is observed in the experimental Michaelis plots. For instance, the $k_{cat}$ value of the best Kemp eliminase reported to date[38] (∼700 ± 60 $s^{-1}$) was determined from the analysis of a Michaelis plot in which saturation was not achieved and only moderate curvature was observed, as is apparent in Fig. 2C of ref. 38, and similar curvatures are seen in most Michaelis plots reported here. Still, in order to ensure that the catalytic rate enhancements reported here are not artefactual,

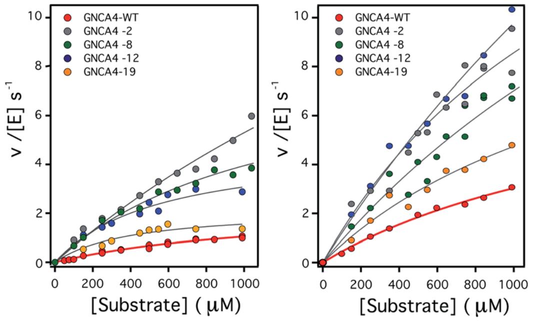


Fig. 2 Plots of Kemp eliminase activity *vs.* substrate concentration at (left) pH 7 and (right) pH 8.4. Activities were measured here for the background protein (GNCA4-WT) and for the 4 variants that display substantially enhanced catalysis at both pH values. Michaelis plots for all the top 20 variants from the FuncLib prediction can be found in Fig. S5.† The lines are the best fits of the Michaelis–Menten equation.

we have performed an extensive amount of experimental work under different conditions, including at different pH and acetonitrile concentrations, to allow for increased ranges of substrate concentration. The catalytic enhancements reported here are consistent over this variety of conditions.

### Crystallization, data collection and structure determination

In order to obtain single crystal structures of the three variants of the GNCA4 β-lactamases of interest to this work, we followed a similar protocol already described elsewhere.[35] The three proteins were subject to crystallization assays by the capillary counterdiffusion techniques[72] and by vapor-diffusion (VD) using the hanging drop set-up. We prepared a small screening around the known successful conditions previously used to crystallized GNCA4 and GNCA4-WT variants.[35] In brief, for counterdiffusion experiments, each protein was concentrated to 23–25 mg mL$^{-1}$, loaded in capillaries of 0.3 mm inner diameter and confronted to 5 M sodium formate in the pH range of 4.0 to 9.0. For VD 1 μL of protein solution was mixed with the reservoir, in a 1 : 1 ratio, and equilibrated against 500 μL of each precipitant cocktail (4 M sodium formate in the pH range of 4.0 to 9.0). The best-looking crystals of GNCA4-2 & GNCA4-12 were obtained at pH 4.0 using the counterdiffusion technique, while in the case of GNCA4-19, they grew at pH 7.0 in hanging drop.

Crystals were extracted from the capillary or fished directly from the drop, subject to cryo-protection by the equilibration with 15% (v/v) glycerol prepared in the mother liquid, with or without 1 mM of the transition-state analogue (5)6-nitrobenzotriazole (ST), flash-cooled in liquid nitrogen and stored until data collection. Crystals were diffracted at the XALOC beamline of the Spanish synchrotron light radiation source (ALBA, Barcelona). Data were indexed and integrated with XDS[73] and scaled with SCALA[74] of the CCP4 program suite.[75] Molecular replacement was performed in Phaser,[76] using the coordinates of GNCA4-WT (PDB ID: 5FQK[35]) as the search model. Refinement was initiated with the phenix.refine[77] module of the PHENIX suite,[78] followed by manual building and water inspection in Coot.[79] The final refinement of ligand coordinates, B-factors and occupancies was achieved following several cycles of refinement including Titration–Libration–Screw (TLS) parameterization. The final model coordinates were verified with Molprobity.[80] The resulting coordinates and the experimental structure factors have been deposited in the Protein Data Bank[81] (PDB IDs: 6TY6, 6TXD and 6TWW, for GNCA4-2, GNCA-12 and GNCA-19, respectively), and the corresponding crystallographic data statistics are provided in Table S4.†

## Results and discussion

### Attempting to increase *de novo* enzyme activity through random library screening

We previously used a minimalist approach (based on 1–2 mutations) to generate a completely new active site for Kemp elimination in ancestral β-lactamase scaffolds. We first attempted to enhance the activity level of our best *de novo* Kemp eliminase through using standard library screening procedures. A library of variants with random mutations and average mutational load of 3–5 mutations was prepared and 522 clones were tested, as we have described in the Materials and methods. The corresponding plot of activity relative to background *vs.* clone ranking is shown in Fig. S4.†

Of these clones, about 300 showed greatly diminished activity levels, suggesting that the encoded proteins may have failed to fold properly. We randomly chose 4 of these clones for protein preparation and, as expected, we found essentially no soluble protein. We also prepared the proteins for the top 10 clones shown in Fig. S4.† In the primary screening, these clones showed activity levels about twice or higher than that of the background variant. However, of these clones, only one was confirmed as a real positive in secondary screening carried out with the purified protein (Table 1). The corresponding variant

Table 1 Catalytic efficiencies and denaturation temperatures at pH 7 for the background GNCA4-WT variant, and the top 10 clones of the random library screening shown in Fig. S4[a]

| Clone | $k_{cat}/K_M$ (M$^{-1}$ s$^{-1}$) | $T_M$ (°C) |
|---|---|---|
| GNCA4-WT | 3047 ± 282 | 80 |
| 3C11 | 608 ± 68 | 77 |
| 4B4 | 1770 ± 126 | 81 |
| 8F11 | **5980 ± 117** | **80** |
| 6D5 | 2476 ± 420 | 81 |
| 7C1 | 600 ± 56 | 72 |
| 8E12 | 2222 ± 167 | 70 |
| 6A12 | 1036 ± 159 | 79 |
| 7D1 | 1880 ± 155 | 67 |
| 2H4 | 2280 ± 146 | ND |
| 5H8 | 2066 ± 67 | 64 |

[a] The values in this table reflect secondary screening performed after purification of the corresponding proteins. Denaturation temperatures ($T_M$) were derived from differential scanning calorimetry, and the catalytic parameters were obtained from fitting the Michaelis–Menten equation to the experimental rate *vs.* substrate concentration profiles. Note that only one of the variants (clone 8F11) shows mildly enhanced catalytic activity in this secondary screening. The $k_{cat}/K_M$ for the GNCA4-WT was originally presented in ref. 35. The $k_{cat}/K_M$ and $T_M$ of the most efficient clone (8F11) is highlighted in bold. All kinetic measurements were performed at 25 °C.

Table 2 A comparison of calculated and experimental activation free energies for the Kemp elimination of 5-nitrobenzisoxazole by the GNCA4-WT β-lactamase and a series of active site mutants[a]

| Variant | $k_{\text{cat}}$ | $K_{\text{m}}$ | $k_{\text{cat}}/K_{\text{m}}$ | $\Delta G^{\ddagger}_{\text{exp}}$ | $\Delta G^{\ddagger}_{\text{calc}}$ |
|---|---|---|---|---|---|
| GNCA4-WT (no His-tag) | 2.6 ± 0.44 | 1.5 ± 0.4 | 1705 ± 139 | 16.7 | 16.2 ± 0.1 |
| G62S | 3.64 ± 0.83 | 1.25 ± 0.45 | 2911 ± 401 | 16.7 | 16.3 ± 0.2 |
| A146G | 5.44 ± 0.77 | 2.34 ± 0.44 | 2328 ± 112 | 16.5 | 16.5 ± 0.2 |
| A173V | 3.78 ± 0.19 | 1.53 ± 0.12 | 2464 ± 62 | 16.7 | 16.9 ± 0.3 |
| L265Q | 4.4 ± 1.01 | 1.8 ± 0.58 | 2447 ± 242 | 16.6 | 16.7 ± 0.2 |
| R256K | 6.13 ± 1.76 | 3.2 ± 1.1 | 1542 ± 369 | 16.4 | 16.9 ± 0.2 |
| R256A | 4.80 ± 1.40 | 4.7 ± 1.6 | 875 ± 15 | 16.5 | 16.6 ± 0.3 |

[a] The GNCA4-WT β-lactamase, which is used as the baseline for our study, is referred to in this table as "wild-type" ("GNCA4-WT"). Note that this data for the "wild type" was measured without a His-tag in ref. 35, which accounts for the small difference with the data given in Table 1 (taken also from ref. 35). Kinetic measurements were performed as described in the Methodology section, and $k_{\text{cat}}$, $K_{\text{M}}$, and $k_{\text{cat}}/K_{\text{M}}$ values are provided in $\text{s}^{-1}$, mM, and $\text{M}^{-1}\ \text{s}^{-1}$, respectively. $\Delta G^{\ddagger}_{\text{exp}}$ and $\Delta G^{\ddagger}_{\text{calc}}$ denote the experimental and calculated activation free energies for these enzymes, in kcal $\text{mol}^{-1}$. $\Delta G^{\ddagger}_{\text{exp}}$ was derived from $k_{\text{cat}}$ using transition state theory, and $\Delta G^{\ddagger}_{\text{calc}}$ is shown as averages and standard error of the mean over thirty individual EVB trajectories per system. All the values in this table were measured at pH 7 with no acetonitrile (other than the small amount coming from the substrate stock solution). All kinetic measurements were performed at 25 °C.

included 6 mutations, with catalytic parameters that were only about two-fold higher than those of the background enzyme.

In order to determinate whether this rather moderate enhancement was due to cancelation between enhancing and deleterious effects of the different mutations, we determined the effect of the single mutations on Kemp eliminase activity. However, no strong cancellation was found (Table 2). Overall, these results highlight the low efficiency and limited enhancements that are typical of non-focused library screening. There is little doubt, of course, that a directed evolution experiment would eventually lead to substantial enhancements in activity, but this will likely require many rounds of library preparation and screening, and also the focus of this study is the extent to which computational approaches can be used to enhance enzyme activity *in lieu* of (otherwise more costly) directed evolution experiments.

### Generation and preliminary assessment of FuncLib predictions

As described in ref. 28, the purpose of FuncLib is to be used to design a small set of stable, efficient, and functionally diverse multipoint active-site mutants that are suitable for low-throughput experimental testing. Our starting point for the FuncLib design was the crystal structure of the most active Kemp eliminase, GNCA4-WT, characterized in our previous work[35] ($k_{\text{cat}}/K_{\text{M}}$ of 3047 ± 283 $\text{M}^{-1}\ \text{s}^{-1}$ at pH 7 for the protein with a His-tag) (PDB ID: 5FQK[35]). This structure was provided as a starting point to the FuncLib server, which is available at http://FuncLib.weizmann.ac.il. We selected 11 active site positions to diversify, comprising residues in close proximity to the substrate (Fig. 1). The resulting sequence space is shown in Table S1.† The diversification was performed using the default FuncLib parameters, and the transition state analog 5(6)-nitrobenzotriazole present in the crystal structure was retained as a proxy for the substrate 5-nitrobenzisoxazole. This yielded 3000 variants, ordered by the Rosetta scoring energy[82] (see the Table S5 and the ESI†).

One obvious feature in the FuncLib results is the frequent prediction among the highly scored variants of a phenylalanine residue at position 260 (*vs.* the Leu residue present in the background "WT" protein, denoted here as GNCA4-WT). This is interesting, because, although close to the *de novo* active site, position 260 belongs to a β-strand and its side chain is actually opposite the active site. Therefore, as a first step to explore the FuncLib predictions we assessed the effect of a single L260F mutation on Kemp elimination catalysis. We observe that this L260F mutation by itself is able to enhance both the catalytic efficiency and turnover number by about 2-fold (data not shown). While this is only a moderate increase in activity, it is already comparable to those for the single improved variant obtained from the screening of a non-focused, random mutation library (Table 1).

### Detailed experimental assessment of the FuncLib predictions

For a more detailed assessment, we prepared and determined both the stability and the Kemp eliminase activity of the 20 twenty top FuncLib predictions. The amino acid substitutions included in these variants are shown in Table S6.†

As mentioned before, FuncLib combines phylogenetic analysis and Rosetta calculations to suggest multiple mutations that generate stabilizing interacting networks at the active site. Indeed, the denaturation temperatures of the top 20 variants, as determined by differential scanning calorimetry demonstrate that all enzymes are stable, and two variants even appear to be somewhat more stable than the background (Table 3). This confirms that, despite the substantial number of mutations introduced, the FuncLib predictions avoid substantial protein destabilization. This should be compared with the top ten variants derived from the random library screening (Table 1) which, in some cases, display substantially diminished denaturation temperatures.

To assess the catalysis levels of the top 20 predicted FuncLib variants, we measured the kinetic activity of several of the predicted sequences at different substrate concentrations and at pH 7 and pH 8.4 (Fig. 2). The catalytic parameters for Kemp elimination catalyzed by the top 20 predicted variants span about two orders of magnitude. This wide range should not be surprising, because FuncLib is not intrinsically intended for predicting catalytically favorable mutations, but rather only to sharply focus the search to regions of the sequence space that encode stable proteins. Still, 4 out of the 20 variants tested display substantially enhanced Kemp eliminase activity with respect to the background variant, both at pH 7 and pH 8.4. The accurate determination of catalytic parameters (in particular the turnover number, $k_{cat}$) from the fitting of the Michaelis–Menten equation to the experimental profiles shown in Fig. 2 is impaired in many cases by the available substrate concentration range, which is in turn limited by substrate solubility. Therefore, we additionally determined rate *vs.* substrate concentration profiles in the presence of 5% acetonitrile, which increases substrate solubility by about 3-fold. This allows for an extended substrate concentration range, but at the slight expense of catalytic efficiency. Such studies in the presence of 5% acetonitrile were performed at pH 7 for all the 20 top variants of the FuncLib ranking (Table 3) and, as a function of pH for the 4 best variants. The corresponding profiles of catalytic efficiency and turnover *vs.* pH are compared with those for our background protein, GNCA4-WT in Fig. 3 and 4. These data confirm an enhancement of catalysis over background of up to about one order of magnitude, in particular in the $k_{cat}$ value.

Table 3 Catalytic parameters for the background and FuncLib variants of the GNCA4/W229F-F290W β-lactamase at pH 7 in the presence of 5% acetonitrile and denaturation temperatures at pH 7 for the same proteins[a]

| Variant | $k_{cat}$ ($s^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ ($M^{-1}$ $s^{-1}$) | $T_M$ (°C) |
|---|---|---|---|---|
| GNCA4-WT | 5.1 ± 0.8 | 3.7 ± 0.8 | 1360 ± 101 | 78.0 |
| GNCA4-1 | 0.22 ± 0.03 | 2.2 ± 0.6 | 102 ± 12 | 79.1 |
| GNCA4-2 | 28.9 ± 15 | 8.12 ± 5 | 3519 ± 401 | 78.4 |
| GNCA4-3 | 4.5 ± 1.6 | 3.3 ± 1.7 | 1348 ± 238 | 79.1 |
| GNCA4-4 | 0.12 ± 0.14 | 14 ± 18 | 8.7 ± 1.2 | 78.0 |
| GNCA4-5 | 2.8 ± 0.2 | 2.3 ± 0.3 | 1214 ± 11 | 77.5 |
| GNCA4-6 | 23 ± 18 | 24 ± 20 | 944 ± 54 | 78.8 |
| GNCA4-7 | 0.54 ± 0.06 | 2.8 ± 0.5 | 190 ± 12 | 77.7 |
| GNCA4-8 | 8.2 ± 1.2 | 2.8 ± 0.7 | 2856 ± 247 | 77.6 |
| GNCA4-9 | 0.17 ± 0.12 | 5.3 ± 5 | 31.7 ± 7.3 | 76.8 |
| GNCA4-10 | 0.7 ± 0.23 | 4.8 ± 2 | 190 ± 22 | 79.6 |
| GNCA4-11 | 2.7 ± 0.35 | 1.8 ± 0.4 | 1403 ± 153 | 76.4 |
| GNCA4-12 | 28 ± 12 | 6.8 ± 3.7 | 4127 ± 460 | 76.0 |
| GNCA4-13 | 0.4 ± 0.07 | 2.9 ± 0.7 | 132 ± 12 | 75.1 |
| GNCA4-14 | 1.06 ± 0.07 | 1.9 ± 0.2 | 560 ± 32.5 | 79.6 |
| GNCA4-15 | 3.1 ± 1.8 | 9.1 ± 6.3 | 339 ± 38 | 77.1 |
| GNCA4-16 | 1.8 ± 0.07 | 3.4 ± 1.9 | 532 ± 96 | 81.2 |
| GNCA4-17 | 0.06 ± 0.01 | 4.4 ± 1.5 | 15 ± 1.4 | 77.1 |
| GNCA4-18 | 4.3 ± 0.4 | 8.2 ± 0.8 | 524 ± 9.3 | 80.9 |
| GNCA4-19 | 7.1 ± 1.5 | 2.9 ± 0.9 | 2366 ± 271 | 77.9 |
| GNCA4-20 | 0.3 ± 0.02 | 1.2 ± 0.1 | 232 ± 16 | 83.9 |

[a] Catalytic parameters were determined at pH 7 in the presence of 5% acetonitrile and the His-tag, from fits of the Michaelis–Menten equation to the experimental profiles of rate *vs.* substrate concentration. The use of 5% acetonitrile extends the experimentally available substrate concentration range, but has a slightly detrimental effect on activity (see Fig. 4). This explains the difference between the value given in this table for the "wild type" protein and that given in Table 1. Michaelis plots for variants GNCA4-4 and GNCA4-6 were almost linear, even with the extended substrate concentration range allowed by the addition of 5% acetonitrile. This explains the large uncertainty associated to the determination of $k_{cat}$ and $K_m$ for these variants, specifically. Note that the number following "GNCA" in the variant column corresponds to the ranking of the FuncLib prediction, based on the Rosetta score, as provided in the ESI and in Table S5. The GNCA4-WT baseline variant is referred to here as the "wild-type" (GNCA4-WT). Denaturation parameters were determined at pH 7 by differential scanning calorimetry. For a list of mutations for each variant, see the ESI. All kinetic measurements were performed at 25 °C.

It is to be noted, nevertheless that while the addition of 5% acetonitrile has the crucial advantage of increasing the solubility of the substrate for the Kemp elimination reaction, thus expanding the experimental concentration range and allowing for more accurate determination of catalytic parameters, the presence of such a small amount of acetonitrile has a small detrimental effect on catalysis (a decrease of about 2-fold), likely in part through a general solvent effect. Therefore, in order to provide an assessment of the achieved levels of catalytic activity that are not perturbed by cosolvent effects, we performed experiments for the GNCA4-12 variant at pH 8 and several different concentrations of acetonitrile, and we extrapolated the kinetic parameters to zero solvent concentration, as shown in Fig. 4.

Increasing acetonitrile concentrations somewhat depresses the catalytic activity. Two factors may contribute to this. First, since acetonitrile increases substrate solubility, it is also stabilizing the free (non-bound) substrate and thus potentially increasing some of the relevant kinetic free energy barriers. In addition, the interaction of acetonitrile molecules with the protein may directly modify such barriers, through small alterations in the structure or dynamics. This second effect is specific, and may depend on the molecular features of this variant, thus leading to the different extrapolation behaviours. The conjunction of these two factors could perhaps be behind the somewhat complex dependency seen for the catalytic efficiency of the GNCA-12 variant (left panel, Fig. 4). In any case, these speculative interpretations do not affect the main point of Fig. 4, namely that the extrapolations to zero acetonitrile concentration are rather short (even for $k_{cat}$) and, therefore, there is little doubt about the reliability of the extrapolated values. The short extrapolation leads to a catalytic efficiency and a turnover number of about $2 \times 10^4$ $M^{-1}$ $s^{-1}$ and $10^2$ $s^{-1}$. These values are well within the ranges of catalytic parameters for modern natural enzymes and, in particular, the value $10^2$ $s^{-1}$ for $k_{cat}$ is about one order of magnitude higher than the median value of the $k_{cat}$ distribution for modern enzymes.[19]

Finally, we have used X-ray crystallography to determine the 3D-structures of the catalytically optimized GNCA4-2, GNCA4-12 and GNCA4-19 variants, the first of which has a transition state analogue bound at the *de novo* active site. These particular structures were chosen as they are all highly active variants, in terms of the measured rates within the available substrate concentration range (Fig. 2), with improved catalytic parameters over GNCA4-WT (Table 3 and Fig. 3). The protein backbones of these new structures are essentially superimposable with that of the background GNCA4-WT variant (Fig. 5A) and, therefore, the observed enhancement of catalysis is likely linked to small rearrangements in the *de novo* active site (Fig. 5B).

### Empirical valence bond calculations on the FuncLib predictions

The enhancements in catalytic activity reported above have been obtained by following a procedure that did not explicitly

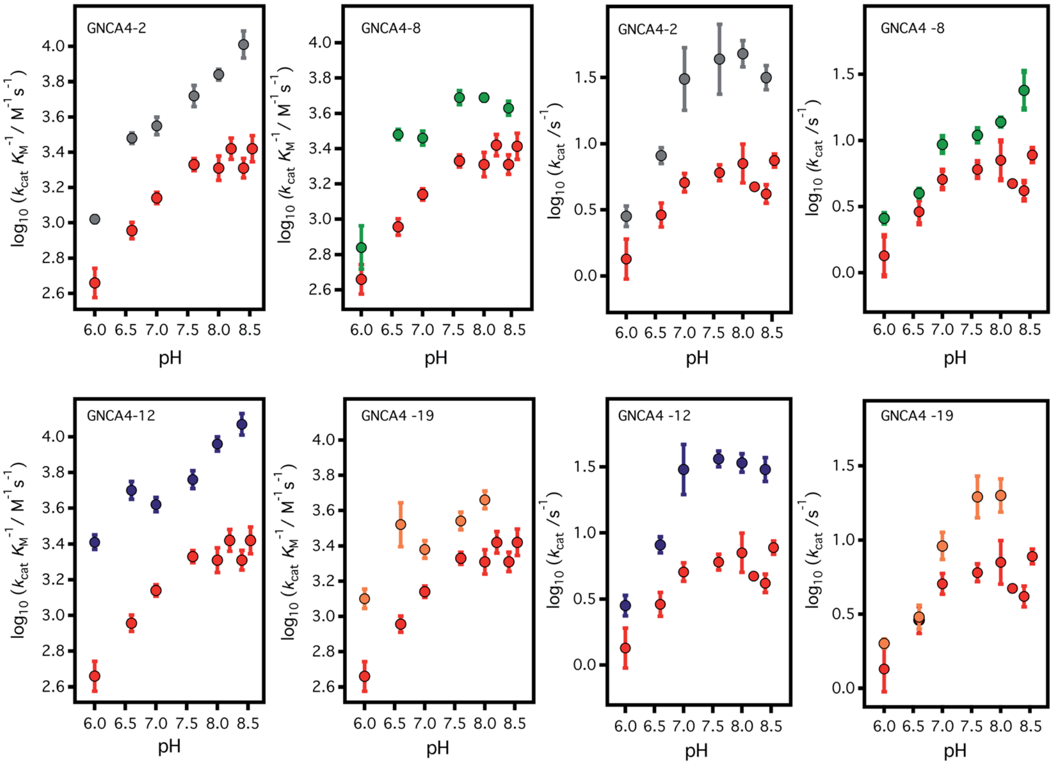


Fig. 3 Profiles of (left) catalytic efficiency and (right) turnover number for the 4 best FuncLib variants. In all cases, the profiles are compared with that of the background GNCA4-WT (red data points). All data were obtained in the presence of 5% acetonitrile to increase the substrate concentration range, and to allow for a more accurate determination of the catalytic parameters ($k_{cat}$ in particular). Acetonitrile, however, has a slightly detrimental effect on activity (Fig. 4) and, therefore, the values given here for the "wild type" protein are somewhat lower than those previously reported in ref. 35. Agreement is observed, however, upon extrapolation to 0% acetonitrile (Fig. 4).

take the structure or stabilization of the transition state into account. That is, we simply focused our screening to regions of the sequence space that are meaningful (positions near and at the active site) and also safe to mutate, in the sense that the predicted multiple-mutation variants are not stability-impaired and their folding is not compromised. We were then interested

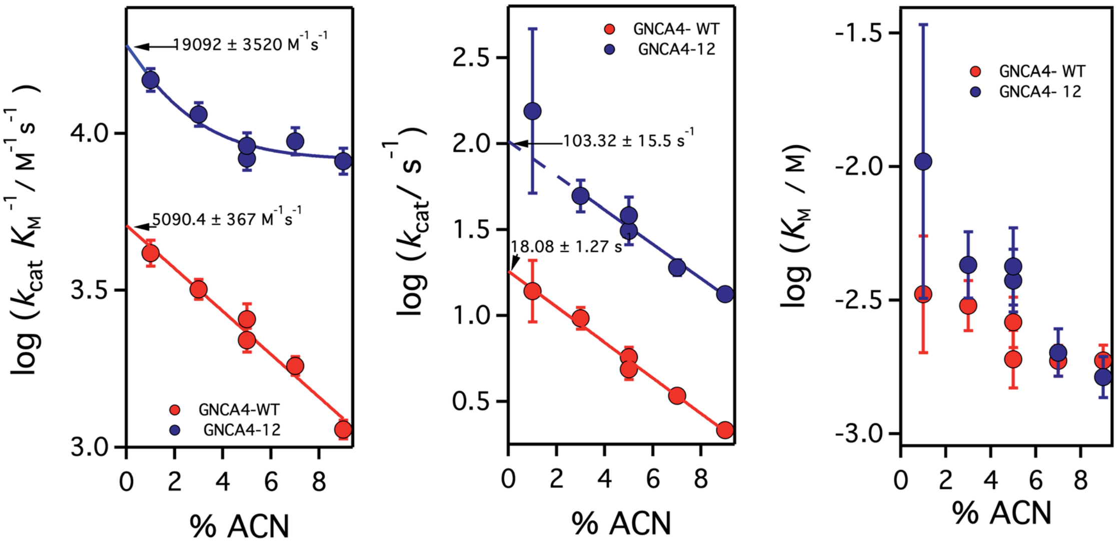


Fig. 4 Catalytic parameters for the activity of the background GNCA4-WT protein (red) and the GNCA4-12 variant from the FuncLib prediction, measured at pH 8 and at different acetonitrile (ACN) concentrations. The values were derived from the fitting of the Michaelis–Menten equation to profiles of rate *vs.* substrate concentration. Values of the catalytic parameters in the absence of acetonitrile are obtained through a short extrapolation, as shown. The values extrapolated for the "wild type" protein (red data point) are in good agreement with those reported in ref. 35 at basic pH.

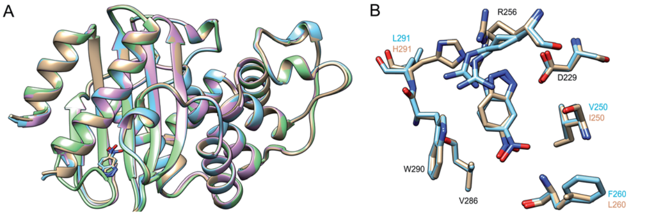


Fig. 5 (A) Superposition of the 3D crystal structures of the background GNCA4-WT (tan, PDB ID: 5FQK[35]) and the three FuncLib variants whose structure we have determined in this work, specifically the GNCA4-2 (light blue, PDB ID: 6TY6), GNCA4-12 (pink, PDB ID: 6TXD) and GNCA4-19 (green, PDB ID: 6TWW) variants. Highlighted here is also the position of the transition state analogue in the GNCA4-WT and GNCA4-2 variants. (B) A close-up of the *de novo* active site in these enzymes, superimposing the active sites of the background enzyme (tan) and the GNCA4-2 variant predicted from FuncLib (light blue, Table 3), with a transition state analogue bound in the active site. Note that we have changed the orientation of the active site compared to panel (A), to better highlight the changes in key active site side chains.

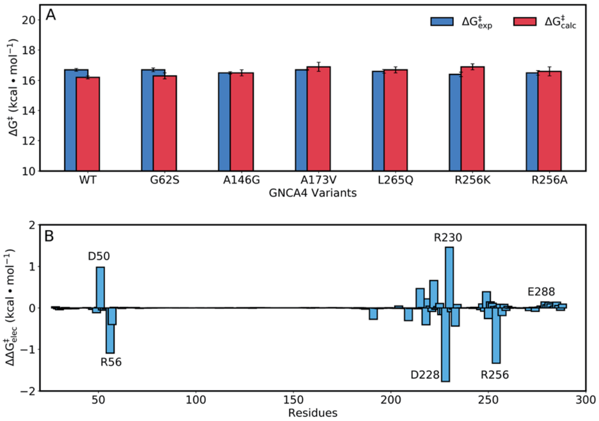


Fig. 6 (A) A comparison of calculated ($\Delta G^{\ddagger}_{\text{calc}}$) and experimental ($\Delta G^{\ddagger}_{\text{exp}}$) activation free energies for the Kemp elimination of 5-nitrobenzisoxazole by the GNCA4-WT β-lactamase, and a series of its active site mutants (see also Table 1). (B) The electrostatic contributions of individual residues to the calculated activation free energies ($\Delta\Delta G^{\ddagger}_{\text{elec}}$) for the Kemp elimination of 5-nitrobenzisoxazole by the GNCA4-WT β-lactamase (treated as the baseline 'wild-type' enzyme in this work). All values were obtained by applying the linear response approximation (LRA)[83,84] to the calculated EVB trajectories, as in our previous works,[85–87] and scaled assuming a dielectric constant of 4 for the highly hydrophobic environment of the *de novo* active site of this β-lactamase (Fig. 1). For the correlation between calculated and experimental values, see Fig. S6.†

in exploring the extent to which computational calculations on the catalytic step itself could be used to further focus and guide the screening. To this end, we have used the empirical valence bond (EVB) approach[54] to probe the catalytic activity of the FuncLib predictions, as this approach has been extensively used to successfully study enzyme catalysis in general,[56] and Kemp elimination in particular.[48,57–59] In particular, this allows us to build on our recent work,[59] in which used the EVB approach to study the evolution of multiple active site configurations in the *de novo* designed Kemp eliminase, KE07.[15] In the present work, we follow the protocol presented in ref. 59, as described in brief in the Materials and methods.

As our starting point, we benchmarked our empirical valence bond (EVB) model by performing simulations of our baseline enzyme, GNCA4-WT, as well as six active site mutants: G62S, A146G, A173V, R256A, R256K, L265Q, described in the section Attempting to increase *de novo* enzyme activity through random library screening. As can be seen from Table 2, the effect of these mutations on the catalytic activity is minimal, with a mere 3.3-fold difference in $k_{\text{cat}}/K_{\text{M}}$ ($\text{M}^{-1}$ $\text{s}^{-1}$) between the most and least active variants, an effect which is mainly caused by differences in $K_{\text{M}}$. The $k_{\text{cat}}$ values are very similar, resulting in activation free energies that are within 0.3 kcal mol$^{-1}$ of each other across the series. Following from this, our EVB simulations were able to reproduce the experimental activation free energy for both the GNCA β-lactamase W290D-F290W to within 0.5 kcal mol$^{-1}$ (Fig. 6A and Table 2).

Representative structures from our simulations of the GNCA4-WT β-lactamase are shown in Fig. 7, with average donor–acceptor distances from our simulations highlighted. The corresponding donor–acceptor distances and donor–hydrogen–acceptor angles for all variants shown in Table 2 can be found in Table S7.† Finally, the electrostatic contributions of individual residues to the calculated activation free energies can be found in Fig. 6B. These contributions were calculated by applying the linear response approximation (LRA)[83,84] to our calculated EVB trajectories, as in our previous work (*e.g.* ref. 85–87). From this data, it can be seen that the individual contributions of most residues to the calculated activation free energies is small (<2 kcal mol$^{-1}$), in line with the fact that the transition state is very similar in structure and in charge distribution to the Michaelis complex.

Having established that our EVB calculations can reliably reproduce the activation free energies of known enzyme variants, we then turned our attention to the top 20 ranked variants from diversification of 11 active site residues (Fig. 1, ESI†),

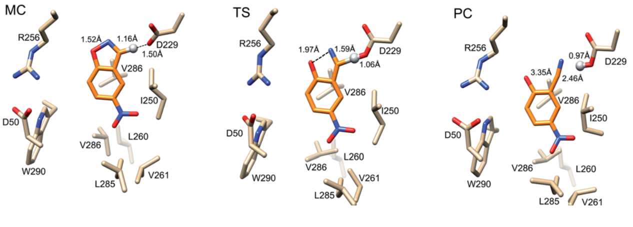


Fig. 7 Representative structures of the GNCA4-WT β-lactamase at the Michaelis complex (MC), transition state (TS), and product complex (PC) for the Kemp elimination reaction catalysed by this enzyme, extracted from EVB trajectories of this reaction. Structures were selected based on clustering analysis using the method of Daura *et al.*[88] as implemented in GROMACS 2016.4.[89,90] The clustering was performed at the MC, TS and PC independently, in order to obtain representative structures for each reacting state. Highlighted here are the donor–hydrogen, acceptor–hydrogen and oxygen–nitrogen distances that are changing during the reaction, and the proton being transferred is shown as a sphere for clarity. Distances are shown as average distances over the entire simulation trajectory (for the corresponding distances for other variants see Tables S7 and S8†).

obtained using FuncLib[28] as described in the Materials and methods. Note that the first variant in the ESI,† with serial number '0101010101010101010101', corresponds to the wild-type enzyme. For simplicity, these variants will be henceforth labelled 1 to 20, starting with the first mutated system, and following the FuncLib ranking.

Fig. 8 and Table S5† show an overview of the calculated activation free energies for the top 20 FuncLib variants. From this data, it can be seen that in the majority of variants, we obtain very little differences in activation free energy (similar to the prior results shown in Table 1), with at most 1 kcal mol$^{-1}$ improvement compared to GNCA4-WT. The only exception to this is a variant (GNCA4-4) with a high activation free energy of 20.3 kcal mol$^{-1}$. This is due to the introduction of an I250M substitution in this variant. Here, the longer side chain of methionine is located between the substrate and the catalytic D229 side chain, introducing steric hindrance in the active site that displaces the substrate from an optimal binding position and increases the D⋯A distance at the Michaelis complex substantially (see Table S8†). All other calculated values based on FuncLib predicted structures lie in the range of 15.3–17.4 kcal mol$^{-1}$, compared to a calculation activation free energy of 16.2 kcal mol$^{-1}$ for the wild-type enzyme (Table S5†). We note also that, in general, the 5 variants carrying the I250M substitution (GNCA4-4, GNCA-7, GNCA4-9, GNCA4-13 and GNCA4-17) show higher experimental activation free energies, in the range of 17.8–19.1 kcal mol$^{-1}$ (Table S6†), suggesting that this substitution is kinetically unfavourable.

Overall, there is (from a computational perspective) good agreement with the experimental values, with the calculated values falling to within 2 kcal mol$^{-1}$ of experiment, considering that unlike the calculations on the simpler single amino acid substitutions shown in Fig. 6, in the case of the FuncLib variants, we are now making predictions for the effect of multiple simultaneous variants using computationally predicted structures. Our data is also in agreement with other computational studies of Kemp elimination, that report activation free energies within 2 (or sometimes more) kcal mol$^{-1}$ from experiment.[48,49,91–95] We note that we have attempted to further refine our EVB calculations by exploring other approaches to generate the starting structures, such as predicting mutations using SCWRL[63] or inserting point mutations manually using the Dunbrack rotamer library,[96] as implemented in Chimera.[97] We tried comparing all three approaches using the GNCA4-2, GNCA4-10 and GNCA4-17 variants as model systems, as these variants show some of the greatest deviations from the experimental values (Table S5†). However, the resulting activation free energies were within 0.4 kcal mol$^{-1}$ of the values obtained using the FuncLib structures (Table S5†), with the exception of GNCA4-2/SCWRL which yielded an activation free energy of 16.0 ± 0.3 kcal mol$^{-1}$ in better agreement with the experimental value of 15.5 kcal mol$^{-1}$ (note however that in the case of GNCA4-2, there was a TSA bound in the active site to guide substrate placement). Therefore, we did not pursue these avenues further as we did not observe systematic improvement in our calculated activation free energies by using alternate approaches to generate the starting structures.

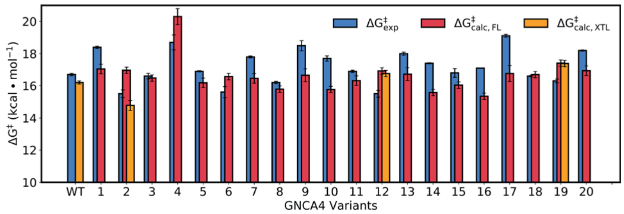


Fig. 8 Calculated activation free energies of the Kemp elimination of 5-nitrobenzisoxazole by the GNCA4-WT β-lactamase and the top 20 best scoring variants predicted by FuncLib[28] (labelled 1 through 20). Shown here are the experimental activation free energies ($\Delta G^‡_{exp}$) derived from $k_{cat}$ based on data presented in Table 3, as well as the corresponding calculated activation free energies based on either structures predicted from FuncLib ($\Delta G^‡_{calc,FL}$) or, where available, directly from crystal structures ($\Delta G^‡_{calc,XTL}$). All energies are presented in kcal mol$^{-1}$, and the calculated activation free energies are averages and standard error of the mean over 30 individual EVB trajectories per system, as described in the Materials and methods. The raw data for this figure can be found in Table S5,† and the correlations between experimental and calculated data can be found in Fig. S7.†

From a structural perspective, it can be seen from Table S8† that our EVB calculated transition states are very similar for the wild-type and all twenty simulated FuncLib variants, in terms of D–A distance and D–H⋯A angle. In addition, the electrostatic contributions of different residues are also relatively similar (Fig. S8†), which is unsurprising in light of the fact that, as discussed elsewhere,[48] the change in charge distribution between Michaelis complex and transition state is very small, making it hard to obtain any significant gains from electrostatic stabilization in this reaction. Where there are larger differences are in the structures of the reacting atoms at the Michaelis complex, where the D–A distance ranges from 2.64–4.25 Å, and the D–H⋯A angle ranges from 129.8–167.1°, with significant correlation between the calculated activation free energy and the D–H distance and D–H⋯A angle (Fig. 9A and B). That is, $R^2$ = 0.84, and −0.81 for the correlation between the calculated activation free energy and the D–H⋯A angle when taking into account only the wild-type enzyme and the FuncLib variants, and 0.82 and −0.78 for distances and angles, respectively, when including also the single residue substitutions considered in Table 2. In the case of the experimental data, we still have moderate correlation between the calculated and experimental activation free energies, still, $R^2$ = 0.56, and −0.57 for the correlation between the experimental activation free energies and to the calculated D–H distances and D–H⋯A angles.

We note that GNCA4-4 appears to be an outlier on this plot, with a D–A distance of >4.0 Å and a D–H⋯A angle of ∼130°. Removing this variant from the analysis (Fig. S9†) yields poor correlation between the geometric parameters and the calculated activation free energies. However, removing this variant still gives good correlations of $R^2$ = 0.68, and −0.64 for the correlation between the experimental activation free energies and to the calculated D–H distances and D–H⋯A angles, respectively. Finally, we obtain good correlations between log $k_{cat}/K_M$ and the calculated D–H distances and D–H⋯A angles, at $R^2$ = −0.71, and 0.71 respectively, when the GNCA4-4 variant is

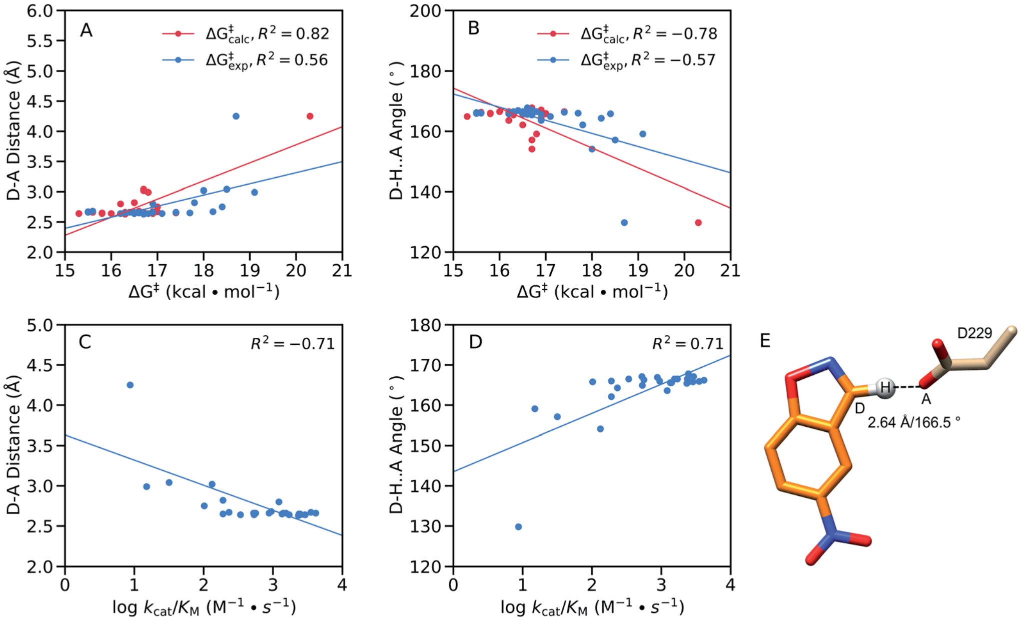


Fig. 9 Correlations between the calculated and experimental activation free energies and the (A and C) donor–acceptor (D–A) distances (Å) and (B and D) donor–hydrogen–acceptor (D–H⋯A) angles (°) in our EVB simulations, calculated based on the data presented in Tables 2, 3, S5, S7 and S8,† using linear regression analysis. Correlations between the geometric parameters and (A and B) calculated activation free energies or (C and D) log $k_{cat}/K_M$ are shown here for all variants considered in this work, both single-point mutations and FuncLib predictions. (E) Schematic overview of the orientation of the reacting fragments in the wild-type enzyme. The annotated distance and angle are the average values from our EVB simulations of the wild-type enzyme (Tables S7 and S8†).

included in the analysis (Fig. 9), and $R^2 = -0.76$ and 0.69 when the GNCA4-4 variant is omitted (Fig. S9†).

We note also that unlike in the case of these geometric parameters (Fig. 9), we do not observe significant correlations with other energetic features of the reaction such as the p$K_a$ of the catalytic base (predicted using PROPKA 3.1 (ref. 66)) or the reorganization energies. Therefore, it is likely that a significant component of the calculated changes in activity observed upon introduction of the amino acid substitutions predicted by FuncLib is better geometric preorganization of the active site for efficient proton abstraction from the substrate, as was also observed in the case of the crystal structure of the directed-evolution optimized Kemp eliminase HG3.17, compared to the computationally designed HG3.[38]

Finally, one additional feature that can be reducing the quality of our predictions is the fact that the FuncLib variants involve the introduction of up to nine mutations into each structure (of the eleven positions that were selected for randomization, see the ESI†), which is likely to compromise the quality of the FuncLib generated protein structures. To assess this, we also performed EVB simulations on the variants for which crystal structures were available: GNCA-2, GNCA-12 and GNCA-19. For these variants, the calculated values fall to within 1.3 kcal mol$^{-1}$ of the experimental values, and can deviate by up to 2.2 kcal mol$^{-1}$ from the values calculated from the FuncLib predicted structures. In the case of GNCA-2, which has the largest deviation between the calculated activation free energies using the crystal and FuncLib structures ($\Delta\Delta G^{\ddagger}_{calc}$ = 2.2 kcal mol$^{-1}$ with the crystal structure giving better agreement with experiment), we observed subtle structural differences the crystal and FuncLib structures (Fig. S10†). Specifically, we observe different rotamers of the R256 and L291 side chains, as well as also subtle displacements of both the side chain of the catalytic base D229 (which is further from the substrate in the FuncLib predicted structure). The shift in the position of the catalytic base D229 in particular likely plays a significant role in the higher calculated activation free energy for this variant when using the FuncLib predicted structure as a starting point.

Therefore, as can be seen from Table 2 and Fig. 6 and S7(C),† when only a few simultaneous substitutions are involved in generating the computationally predicted structure (as in our prior work[59,85–87,98,99]), or where a crystal structure of a variant with multiple amino acid substitutions is available, the EVB approach can reproduce experimental data with high fidelity in a wide range of systems. In addition, considering the potentially large structural perturbations involved, agreement within 2 kcal mol$^{-1}$ of experiment is still respectable, in line with or better than the agreement with experiment obtained in other computational studies of Kemp elimination,[48,49,91–95] and thus gives EVB great potential as a predictive tool for more complex reactions where the introduction of mutations have a larger energetic impact on the system, and thus better correlation with experiment would be expected as observed for example in ref. 85, 86 and 98–100. Based on this, we believe the EVB simulations can already act as a first step filter over the Rosetta scores predicted by FuncLib, as the latter in this case provided no correlation with experimental activities, despite being able to effectively predict variants with improved activity.

## Concluding remarks

Kemp elimination is a straightforward proton–abstraction reaction that can be performed by a simple molecular machinery consisting, at the bare minimum, of a catalytic base. Accordingly, *de novo* generation of enzyme active sites for Kemp elimination has proved amenable to rational design.[15,35,38,44,46,94,101] On the other hand, enhancing an already existing Kemp eliminase activity is challenging because of the similarity of the substrate and the transition state for the reaction,[48] which makes it difficult to find mutations that preferentially stabilize the transition state. Indeed, the best Kemp eliminases reported to date are the results of many rounds of directed evolution starting with rational designs.[38,102]

The starting point of the engineering efforts reported here is a Kemp eliminase we previously obtained through minimalist design on a β-lactamase background.[35] Our design took advantage of the conformational flexibility of an ancestral β-lactamase scaffold to produce both a suitable cavity and a catalytic base within it through a single mutation, while a second mutation enhanced relevant interactions at the *de novo* active site. This led to a $k_{\text{cat}}$ value of ~10 s$^{-1}$, which is about the turnover number for an average modern enzyme.[19] Such a comparatively high starting level of catalysis should further contribute to the (already difficult) task of enhancing Kemp eliminase activity and, indeed, as reported here, screening of 500 clones from a random library led to only one variant with a moderate catalysis improvement. It is remarkable against this backdrop, then, that screening of the 20 top variants from the FuncLib ranking produced 4 variants with improved catalysis (in terms of both $k_{\text{cat}}$ and $k_{\text{cat}}/K_{\text{M}}$), of which two showed order-of-magnitude enhancements, bringing $k_{\text{cat}}$ to the region of $10^2$ s$^{-1}$ (Fig. 4). This value compares well with the best Kemp eliminases reported to date, derived from extensive directed evolution efforts on complex rationally-designed backgrounds. It is in fact somewhat higher than values reported in ref. 102, and it is in the same range as the value (700 s$^{-1}$) reported in ref. 38, in both cases as the outcome of many rounds of directed evolution. Finally, the catalytic efficiency of our best Kemp eliminase ($k_{\text{cat}}/K_{\text{M}}$ of ~$2 \times 10^4$ M$^{-1}$ s$^{-1}$) is only about one order of magnitude below the values obtained from intensive directed evolution, namely $2.3 \times 10^5$ M$^{-1}$ s$^{-1}$ by Hilvert and coworkers (HG3.17),[38] and $5.7 \times 10^5$ M$^{-1}$ s$^{-1}$ reported by Tawfik and coworkers using a 5,7-dichloro Kemp substrate,[102] as well as $1.2 \times 10^5$ M$^{-1}$ s$^{-1}$ obtained by computational design using a minimalist approach using the HG3 eliminase as a starting scaffold while incorporating key mutations from the HG3 evolutionary trajectory towards HG3.17 into the design process towards the new Kemp eliminase, HG4.[103] This is significant because our crystal structures show that, unlike other Kemp eliminases such as HG3 (ref. 38) or KE07,[59] in the present case it was possible to obtain significant enhancements in catalytic activity without the need for major structural reorganization of the active site.

The striking efficiency of our success with FuncLib-based optimization can be put down to several factors. First, FuncLib is intended to predict stable enzyme variants, a prediction which is in fact confirmed by our thermal denaturation experiments on our Kemp eliminases (Table 3). Therefore, screening effort is not wasted in probing unstable variants that may not fold properly. Secondly, FuncLib can be used to target regions that are expected to be relevant for catalysis (the active site region in this work) and, therefore, screening efforts is not wasted in testing variants with mutations that do not impact catalysis ("neutral" variants). In fact, most of the tested 20 FuncLib predictions show Kemp elimination activities that differ substantially from that of the background used (Table 3 and Fig. 2–4). Thirdly, the fact that FuncLib directly predicts multi-point variants bypasses issues related to epistatic interactions between mutations.

Our results support, overall, that FuncLib predictions may provide an efficient computational methodology to speed up directed evolution by guiding screening to regions of the sequence space that are safe and catalytically-relevant. We have further shown here that the experimental free energy barriers for the optimized eliminases can be reproduced to within ~2 kcal mol$^{-1}$ by the empirical valence bond calculations. This is impressive in light of the very small changes in activity involved (from a thermodynamic perspective, Table 3) and thus the associated challenges of optimizing Kemp eliminase activity using electrostatics alone.[48,58] We note that other computational studies of Kemp elimination also report activation free energies with deviations within this range or up to several kcal mol$^{-1}$ from experiment.[48,49,91–95] In addition, whereas we and others have been able to obtain high fidelity with experimental values across a wide range of enzymes and enzyme variants even in the case of far more complex systems than the current Kemp eliminase.[98–100,104–109] This makes EVB useful as a predictive tool for systems where the changes in energy involved are not as subtle as in the case of Kemp elimination.

In addition, while the FuncLib algorithm focuses on optimizing stability and carries no information about the transition states involved, nevertheless, the best performing FuncLib variants do so due to improved geometric preorganization of the active site through optimizing of the D–H distance and D–H···A angle. This suggests that, in particular for more complex systems where mutations can introduce larger changes in activity, the FuncLib-based stability-guidance could be further refined and focused on the basis of the computational prediction of catalysis, at least in the initial stages of the directed evolution process, during which larger jumps in activity may be possible. This is significant, as FuncLib does not take any information about the substrate or transition state into account in the design process, and therefore while it targets the stability of the overall protein, it does not provide insight into how mutations will affect transition state stabilization.[28] Clearly, FuncLib can also be used as to generate stable scaffolds that can then be used as a basis for rational design efforts to insert specific physio-chemical properties (such as, for instance, engineering an oxyanion hole) into the active site of the enzyme of interest.

Taken together, the combination of experimental and computational work presented here both showcases the tremendous potential of FuncLib's evolutionary-based stability-screening protocol as a valuable tool in computational enzyme design, as well as the potential of ancestral enzymes as starting scaffolds for artificial enzyme engineering. Here, our crystal structures illustrate that significant gains in activity can be achieved without the need for corresponding significant active site rearrangement. Finally, it is important to note that FuncLib is based on sequence alignment, and thus it would be logical to assume that it would work best for enhancing the reactivity of an enzyme towards its native substrate(s). There remains, however, the question of whether it would also enable the design of function scaffolds that were not designed for those functions. By targeting a non-natural reaction in a *de novo* active site, we demonstrate that FuncLib is a broadly useful tool, that can also be used to design biological catalysts for anthropogenic substrates.

## Author contributions

VAR, ARR, JMSR, and SCLK conceptualized this work. JMSR and SCLK are responsible for data curation. All authors were responsible for investigation and formal analysis of data. JMSR and SCLK were responsible for funding acquisition. VAR, ARR, JMSR, MOM, FSG and JAG were responsible for design of methodology. SCLK and JMSR were responsible for project administration. SCLK, JMSR, JAG and FSG provided resources for this work. SCLK was responsible for software for the computational part of the manuscript. SCLK, JMSR, JAG and FSG were responsible for supervision of this work. VAR, ARR, JMSR and SCRLK were responsible for validation, visualization, writing and review and editing of the manuscript, with input from all authors.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

This work was supported by the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellowship to SCLK, grant 2018.0140), the Human Frontier Science Program (to JMSR and SCLK, grant RGP0041/2017), FEDER Funds/Spanish Ministry of Science, Innovation and Universities (grants BIO2015-66426-R and RTI2018-097142-B-100 to JMSR) and FEDER/Junta de Andalucia – Consejería de Economía y Conocimiento (to VAR, grant E.FQM.113.UGR18). Computer time was provided through a generous allocation by the Swedish National Infrastructure for computing (SNAC grants 2018/2-3 and 2019/2-1), and all calculations were performed on the Kebnekaise Cluster at HPC2N Umeå, and the Tetralith cluster at NSC Linköping. We thank Sarel Fleishman and Olga Khersonsky for introducing us to and assisting us in using FuncLib.

## References

1 R. Wolfenden and M. J. Snider, *Acc. Chem. Res.*, 2001, **34**, 938–945.

2 J.-M. Choi, S.-S. Han and H.-S. Kim, *Biotechnol. Adv.*, 2014, **33**, 1443–1454.

3 E. M. M. Abdelraheem, H. Busch, U. Hanefeld and F. Tonin, *React. Chem. Eng.*, 2019, **4**, 1878–1894.

4 K. Chen and F. H. Arnold, *Nat. Catal.*, 2020, **3**, 203–213.

5 L. Yuan, I. Kurek, J. English and R. Keenan, *Microbiol. Mol. Biol. Rev.*, 2005, **69**, 373–392.

6 N. Tokuriki, F. Stricher, L. Serrano and D. S. Tawfik, *PLoS Comput. Biol.*, 2008, **4**, e1000002.

7 F. H. Arnold, *Acc. Chem. Res.*, 1998, **31**, 125–131.

8 U. T. Bornscheuer, B. Hauer, K. E. Jaeger and U. Schwaneberg, *Angew. Chem., Int. Ed.*, 2019, **58**, 36–40.

9 G. Qu, A. Li, Z. Sun, C. G. Acevedo-Rocha and M. T. Reetz, *Angew. Chem., Int. Ed.*, 2020, **59**, 2–30.

10 C. Zeymer and D. Hilvert, *Annu. Rev. Biochem.*, 2018, **87**, 131–157.

11 P.-S. Huang, S. E. Boyken and D. Baker, *Nature*, 2016, **537**, 320–327.

12 M. C. C. J. C. Ebert and J. N. Pelletier, *Curr. Opin. Chem. Biol.*, 2017, **37**, 89–96.

13 A. Goldenzweig and S. J. Fleishman, *Annu. Rev. Biochem.*, 2018, **87**, 105–129.

14 V. V. Welborn and T. Head-Gordon, *Chem. Rev.*, 2019, **119**, 6613–6630.

15 D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Liang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik and D. Baker, *Nature*, 2008, **453**, 190–195.

16 L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas III, D. Hilvert, K. N. Houk, B. L. Stoddard and D. Baker, *Science*, 2008, **319**, 1387–1391.

17 J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St. Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael and D. Baker, *Science*, 2010, **329**, 309–313.

18 D. Hilvert, *Annu. Rev. Biochem.*, 2013, **82**, 447–470.

19 A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik and R. Milo, *Biochemistry*, 2011, **50**, 4402–4410.

20 A. Pavelka, E. Chovancova and J. Damborsky, *Nucleic Acids Res.*, 2009, **37**, W376–W383.

21 C.-Y. Chen, I. Georgiev, A. C. Anderson and B. R. Donald, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 3764–3769.

22 G. Kiss, D. Röthlisberger, D. Baker and K. N. Houk, *Protein Sci.*, 2010, **19**, 1760–1773.

23 S. Lindert, J. Meiler and J. A. McCammon, *J. Chem. Theory Comput.*, 2013, **9**, 3843–3847.

24 H. J. Wijma, R. J. Floor, S. Bjelic, S. J. Marrink, D. Baker and D. B. Janssen, *Angew. Chem., Int. Ed.*, 2015, **54**, 3726–3730.

25 A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik and S. J. Fleishman, *Mol. Cell*, 2016, **63**, 337–346.

26 M. C. Childers and V. Daggett, *Mol. Syst. Des. Eng.*, 2016, **2**, 9–33.

27 A. Romero-Rivera, M. Garcia-Borràs and S. Osuna, *ACS Catal.*, 2017, **7**, 8524–8532.

28 O. Khersonsky, R. Lipsh, Z. Avizemer, Y. Ashani, M. Goldsmith, H. Leader, O. Dym, S. Rogotner, D. L. Trudeau, J. Prilusky, P. Amengual-Rigo, V. Guallar, D. S. Tawfik and S. J. Fleishman, *Mol. Cell*, 2018, **72**, 178–186.e175.

29 A. Currin, J. Kwok, J. C. Sadler, E. L. Bell, N. Swainston, M. Ababi, P. Day, N. J. Turner and D. B. Kell, *ACS Synth. Biol.*, 2019, **8**, 1371–1378.

30 A. D. St-Jacques, M.-E. C. Eyahpaise and R. A. Chica, *ACS Catal.*, 2019, **9**, 5480–5485.

31 F. Cadet, N. Fontaine, L. Gangyue, J. Sanchis, M. N. F. Chong, R. Pandjaitan, I. Vetrivel, B. Offmann and M. T. Reetz, *Sci. Rep.*, 2018, **8**, 16757.

32 Z. Wu, S. B. J. Khan, R. D. Lewis, B. J. Wittmann and F. H. Arnold, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 8852–8858.

33 L. Guangyue, Y. Dong and M. T. Reetz, *Adv. Synth. Catal.*, 2019, **361**, 2377–2386.

34 S. Mazurenko, Z. Prokop and J. Damborsky, *ACS Catal.*, 2020, **10**, 1210–1223.

35 V. A. Risso, S. Martinez-Rodriguez, A. M. Candel, D. M. Krüger, D. Pantoja-Uceda, M. Ortega-Muñoz, F. Santoyo-Gonzlez, E. A. Gaucher, S. C. L. Kamerlin, M. Bruix, J. A. Gavira and J. M. Sanchez-Ruiz, *Nat. Commun.*, 2017, **8**, 16113.

36 B. van der Burg, G. Vriend, O. R. Veltman, G. Venema and V. G. Eisink, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 2056–2060.

37 K. Miyazaki, P. L. Wintrode, K. A. Grayling, D. N. Rubingh and F. H. Arnold, *J. Mol. Biol.*, 2000, **297**, 1015–1026.

38 R. Blomberg, H. Kries, D. M. Pinkas, P. R. E. Mittl, M. G. Grütter, H. K. Privett, S. L. Mayo and D. Hilvert, *Nature*, 2013, **503**, 418–421.

39 S. N. Thorn, R. G. Daniels, M.-T. M. Auditor and D. Hilvert, *Nature*, 1995, **373**, 228–230.

40 A. Genre-Grandpierre, C. Tellier, M.-J. Loirat, D. Blanchard, D. R. W. Hodgson, F. Hollfelder and A. J. Kirby, *Bioorg. Med. Chem. Lett.*, 1997, **7**, 2497–2502.

41 A. J. Kirby, F. Hollfelder and D. S. Tawfik, *Appl. Biochem. Biotechnol.*, 2000, **83**, 173–181.

42 E. W. Debler, S. Ito, F. P. Seebeck, A. Heine, D. Hilvert and I. A. Wilson, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 4984–4989.

43 F. P. Seebeck and D. Hilvert, *J. Am. Chem. Soc.*, 2005, **127**, 1307–1312.

44 M. Sparta and A. N. Alexandrova, *Mol. Simul.*, 2011, 557–571, 3Ams.

45 M. Merski and B. K. Shoichet, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 16179–16183.

46 W. Cullen, M. C. Misuraca, C. A. Hunter, N. H. Williams and M. D. Ward, *Nat. Chem.*, 2016, **8**, 231–236.

47 E. Sanxhez, S. Lu, C. Reed, J. Schmidt and M. Forconi, *J. Phys. Org. Chem.*, 2016, **29**, 185–189.

48 M. P. Frushicheva, J. Cao, Z. T. Chu and A. Warshel, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 16869–16874.

49 G. Jindal, B. Ramachandran, R. P. Bora and A. Warshel, *ACS Catal.*, 2017, **7**, 3301–3305.

50 H. Kries, J. S. Bloch, H. A. Bunzel, D. M. Pinkas and D. Hilvert, *ACS Catal.*, 2020, **10**, 4460–4464.

51 D. M. Weinreich, N. F. Delaney, M. A. DePristo and D. L. Hartl, *Science*, 2006, **312**, 111–114.

52 J. M. Sanchez-Ruiz, *Biochem. J.*, 2012, **445**, e1–e3.

53 T. N. Starr and J. W. Thornton, *Protein Sci.*, 2016, **25**, 1204–1218.

54 A. Warshel and R. M. Weiss, *J. Am. Chem. Soc.*, 1980, **102**, 6218–6226.

55 A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. Liu and M. H. M. Olsson, *Chem. Rev.*, 2006, **8**, 3210–3235.

56 A. Shurki, E. Derat, A. Barrozo and S. C. L. Kamerlin, *Chem. Soc. Rev.*, 2015, **44**, 1037–1052.

57 M. P. Frushicheva, J. Cao and A. Warshel, *Biochemistry*, 2011, **50**, 3849–3858.

58 A. Labas, E. Szabo, L. Mones and M. Fuxreiter, *Biochim. Biophys. Acta*, 2013, **1834**, 908–917.

59 N.-S. Hong, D. Petrović, R. Lee, G. Gryn'ova, M. Purg, J. Saunders, P. Bauer, P. D. Carr, C.-Y. Lin, P. D. Mabbitt, W. Zhang, T. Altamore, C. Easton, M. L. Coote, S. C. L. Kamerlin and C. J. Jackson, *Nat. Commun.*, 2018, **9**, 3900.

60 W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 11225–11236.

61 G. A. Kaminski, R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, *J. Phys. Chem. B*, 2001, **105**, 6474–6487.

62 J. Marelius, K. Kolmodin, J. Feierberg and J. Åqvist, *J. Mol. Graphics Modell.*, 1998, **16**, 213–225.

63 G. G. Krivov, M. V. Shapovalov and R. L. Dunbrack Jr, *Proteins*, 2009, **77**, 778–795.

64 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.

65 A. Warshel and G. King, *Chem. Phys. Lett.*, 1985, **121**, 124–129.

66 M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 525–537.

67 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

68 J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.

69 F. S. Lee and A. Warshel, *J. Chem. Phys.*, 1992, **97**, 3100.

70 W. Humphrew, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38.

71 V. A. Risso, J. A. Gavira, D. Meija-Carmona, E. A. Gaucher and J. M. Sanchez-Ruiz, *J. Am. Chem. Soc.*, 2013, **135**, 2899–2902.

72 F. Otalora, J. A. Gavira, J. D. Ng and J. M. Garcia-Ruiz, *Prog. Biophys. Mol. Biol.*, 2009, **101**, 26–37.

73 W. Kabsch, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 125–132.

74 P. Evans, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2006, **62**, 72–82.

75 Collaborative Computational Project Number 4, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1994, **50**, 760–763.

76 A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read, *J. Appl. Crystallogr.*, 2007, **40**, 658–674.

77 P. V. Afonine, M. Mustyakimov, R. W. Grosse-Kunstleve, N. W. Moriarty, P. Langan and P. D. Adams, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 1153–1163.

78 P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 213–221.

79 P. Emsley, B. Lohkamp, W. G. Scott and K. Cowtan, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 486–501.

80 V. B. Chen, W. B. Arendall III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 12–21.

81 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

82 A. V. Morozov and T. Kortemme, *Adv. Protein Chem.*, 2005, **72**, 1–38.

83 F. S. Lee, Z.-T. Chu, M. B. Bolger and A. Warshel, *Protein Eng., Des. Sel.*, 1992, **5**, 215–228.

84 I. Muegge, H. Tao and A. Warshel, *Protein Eng., Des. Sel.*, 1997, **10**, 1363–1372.

85 Y. S. Kulkarni, Q. Liao, D. Petrović, D. M. Krüger, B. Strodel, T. L. Amyes, J. P. Richard and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2017, **139**, 10514–10525.

86 Y. S. Kulkarni, Q. Liao, F. Byléhn, T. L. Amyes, J. P. Richard and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2018, **140**, 3854–3857.

87 Y. S. Kulkarni, T. L. Amyes, J. P. Richard and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2019, **141**, 16139–16150.

88 X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren and A. E. Mark, *Angew. Chem., Int. Ed.*, 1999, **38**, 236–240.

89 D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.

90 M. J. Abraham, D. van der Spoel, E. Lindahl and B. Hess, *GROMACS Development Team*, 2017.

91 A. N. Alexandrova, D. Röthlisberger, D. Baker and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2008, **130**, 15907–15915.

92 A. Labas, E. Szabo, L. Mones and M. Fuxreiter, *Biochim. Biophys. Acta, Proteins Proteomics*, 2013, **1834**, 908–917.

93 K. Świderek, I. Tuñón, V. Moliner and J. Bertran, *ACS Catal.*, 2015, **5**, 2587–2595.

94 A. Li, B. Wang, A. Ilie, K. D. Dubey, G. Bange, I. V. Korendovych, S. Shaik and M. T. Reetz, *Nat. Commun.*, 2017, **8**, 14876.

95 K. Świderek, I. Tuñón, V. Moliner and J. Bertran, *Chem.–Eur. J.*, 2017, **23**, 7582–7589.

96 M. V. Shapovalov and R. L. Dunbrack Jr, *Structure*, 2011, **19**, 844–858.

97 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Cough, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.

98 D. Blaha-Nelson, D. Krüger, K. Szeler, M. Ben-David and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2017, **139**, 1155–1167.

99 M. Purg, M. Elias and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2017, **139**, 17533–17546.

100 A. R. Calixto, C. Moreira, A. Pabis, C. Kötting, K. Gerwert, T. Rudack and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2019, **141**, 10684–10701.

101 I. V. Korendovych, D. W. Kulp, Y. Wu, H. Cheng, H. Roder and W. F. DeGrado, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 6823–6827.

102 O. Khersonsky, G. Kiss, D. Röthlisberger, O. Dym, S. Albeck, K. N. Houk, D. Baker and D. S. Tawfik, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 10358–10363.

103 A. Broom, R. V. Rakotoharisoa, M. C. Thompson, N. Zarifi, E. Nguyen, N. Mukhametzhanov, L. Liu, J. S. Fraser and R. A. Chica, 2020, bioRxiv, DOI: 10.1101/2020.03.19.999235.

104 M. Kazemi and J. Åqvist, *Nat. Commun.*, 2015, **6**, 7293.

105 G. V. Isaksen, J. Åqvist and B. O. Brandsdal, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 7822–7827.

106 G. Jindal, K. Slanska, V. Kolev, J. Damborsky, Z. Prokop and A. Warshel, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **116**, 389–394.

107 H. Yoon, L. N. Zhao and A. Warshel, *ACS Catal.*, 2019, **9**, 1329–1336.

108 Y. S. Kulkarni, T. L. Amyes, J. P. Richard and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2019, **141**, 16139–16150.

109 M. Ben-David, M. Soskine, A. Dubovetskyi, K.-P. Cherukuri, O. Dym, J. L. Sussman, Q. Liao, K. Szeler, S. C. L. Kamerlin and D. S. Tawfik, *Mol. Biol. Evol.*, 2020, **37**, 1133–1147.

# SUPPORTING INFORMATION

## Enhancing a *De Novo* Enzyme Activity by Computationally-Focused Ultra-Low-Throughput Screening

# Supporting Information for:

# Enhancing a *De Novo* Enzyme Activity by Computationally-Focused, Ultra-Low-Throughput Sequence Screening

Valeria A. Risso,[1] Adrian Romero-Rivera,[2] Luis I. Gutierrez-Rus,[1] Mariano Ortega-Muñoz,[3] Francisco Santoyo-Gonzalez,[3] Jose A. Gavira,[4] Jose M. Sanchez-Ruiz,[*,1] Shina C. L. Kamerlin[*,2]

1. Departamento de Química Física, Facultad de Ciencias, University of Granada, 18071-Granada, Spain
2. Science for Life Laboratory, Department of Chemistry-BMC, Uppsala University, BMC Box 576, S-751 23 Uppsala, Sweden
3. Departamento de Química Organica, Facultad de Ciencias, University of Granada, 18071-Granada, Spain
4. Laboratorio de Estudios Cristalograficos, Instituto Andaluz de Ciencias de la Tierra, CSIC-University of Granada, Avenida de las Palmeras 4, Granada 18100 Armilla, Spain

Corresponding author email addresses: sanchezr@ugr.es and lynn.kamerlin@kemi.uu.se

# Table of Contents

**Supplementary Figures……..……………………..………………………..………………………. S4**

**Figure S1**. Root mean square deviations during equilibration of simulations of the GNCA4-WT and of variants with single amino acid substitutions…………………………………………………….....S4

**Figure S2**. Root mean square deviations during equilibration of simulations of the twenty FuncLib variants..………….……………….. ………………………………………………………………..S5

**Figure S3**. Root mean square deviations during equilibration of simulations of the crystal structures of the three FuncLib variants…………………………………………….....…………..………. S6

**Figure S4**. Kemp eliminase activity of 522 clones from a random library on the *de novo* GNCA4-WT *β*-lactamase..……………………………………………..……………….……………………S6

**Figure S5.** Plots of Kemp eliminase activity *vs.* substrate concentration for FuncLib variants………S7

**Figure S6.** Correlation between calculated and experimental activation free energies for the Kemp elimination of 5-nitrobenzisoxazole by the GNCA4-WT and a series of active site mutants…………...S8

**Figure S7.** Correlation between calculated energetics and experimental activation free energies for the Kemp elimination of 5-nitrobenzisoxazole by the GNCA4-WT and the top twenty variants predicted from the FuncLib webserver………………………………………………………………...S9

**Figure S8**. Electrostatic contributions of individual residues to the calculated activation free energies for the Kemp elimination of 5-nitrobenzisoxazole by the top 20 best scoring GNCA4 variants predicted by FuncLib……..…..…............................................................................................................S10

**Figure S9**. Correlation between geometric parameters and calculated and experimental activation free energies and experimentally measured kinetics……………………………………………………S11

**Figure S10**. Comparison of the X-ray and FuncLib predicted structures of the GNCA4-2 variant…S12

**Supplementary Tables ……………………………………………………..…………………….... S13**

**Table S1**. Sequence space explored after diversification of 11 active site residues by the FuncLib webserver..………………………………………………………………………………………S13

**Table S2**. Ionized residues and protonation patterns of histidine residues during EVB simulations of the β-lactamase catalyzed Kemp elimination of 5-nitrobenzisoxazole............…………………......…..S14

**Table S3**. Percentage acetonitrile used at different substrate concentrations…………………………S14

**Table S4**. Data collection and refinement statistics of the 3D structural models..………………...…. S15

**Table S5**. Rosetta scores and calculated and experimental activation free energies for the GNCA4-WT *β*-lactamase, as well as the top twenty variants predicted from FuncLib..………………………..…….. S16

**Table S6.** Amino acid substitutions introduced in the twenty top-ranked FuncLib variants…………..S17

**Table S7**. Average donor-acceptor (D-A) distances and donor-hydrogen-acceptor (D-H…A) angles obtained from EVB simulations of different experimentally characterized variants of the GNCA4-WT *β*-lactamase..……………………..…………………………………………………………………….. S18

**Table S8**. Average donor-acceptor (D-A) distances and donor-hydrogen-acceptor (D-H…A) angles obtained from EVB simulations of the GNCA4-WT *β*-lactamase, as well as of the top twenty variants predicted from FuncLib. ……………………………………………………….……………………. S19

**Supplementary References ………………………………………………………..……………………S20**

## Supplementary Figures

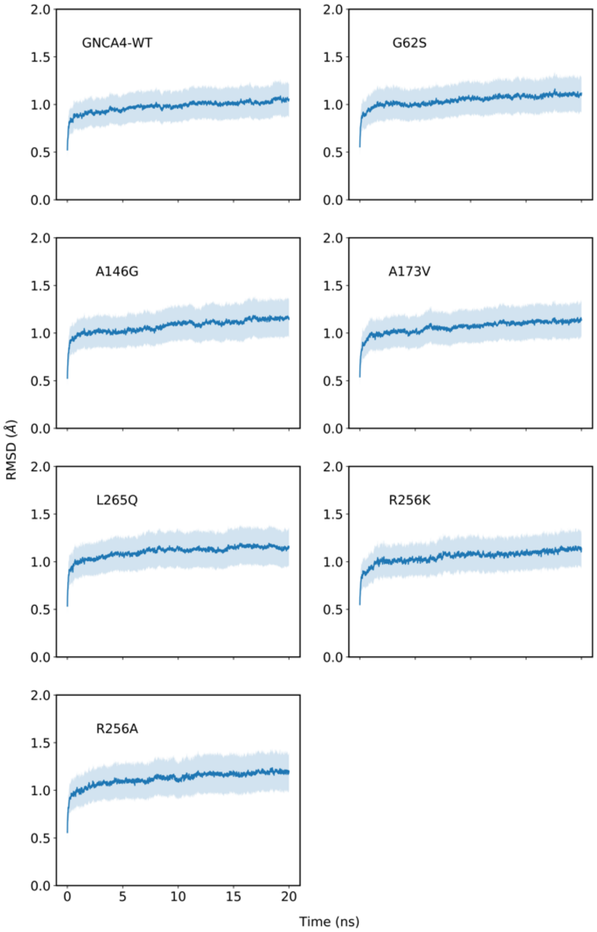


**Figure S1**. The root mean square deviations (RMSD, Å) of all backbone atoms of the GNCA4-WT and single amino acid substitutions used for the calibration of our EVB model, at the approximate EVB transition state (λ = 0.5) for the Kemp elimination reaction catalyzed by these enzymes. Data was collected every 10 ps from the initial equilibration runs, and is shown as averages and standard deviations over ten individual 20 ns MD simulations per system (*i.e.* 200 ns cumulative simulation time per system). The average RMSD per system is denoted by solid blue lines, and the standard deviations per point over all trajectories are illustrated by the shaded area on each plot.

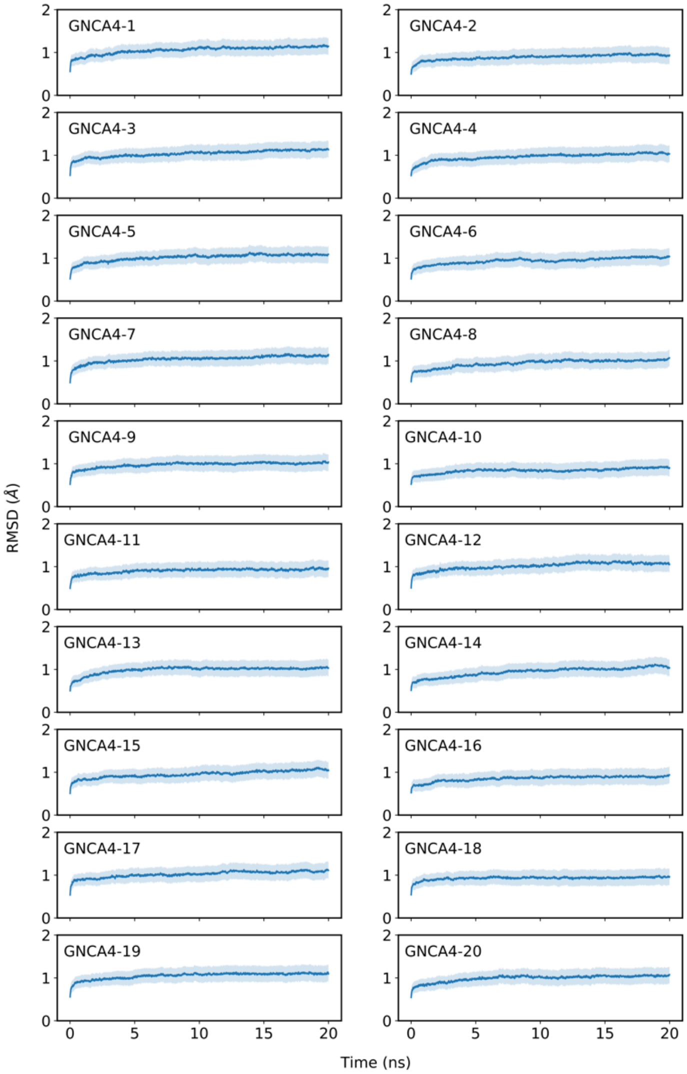


**Figure S2**. The root mean square deviations (RMSD, Å) of all backbone atoms of the twenty FuncLib variants studied in this work (computationally predicted structures), at the approximate EVB transition state ($\lambda = 0.5$) for the Kemp elimination reaction catalyzed by these enzymes. Data was collected every 10 ps from the initial equilibration runs, and is shown as averages and standard deviations over ten individual 20 ns MD simulations per system (*i.e.* 200 ns cumulative simulation time per system). The average RMSD per system is denoted by solid blue lines, and the standard deviations per point over all trajectories are illustrated by the shaded area on each plot.

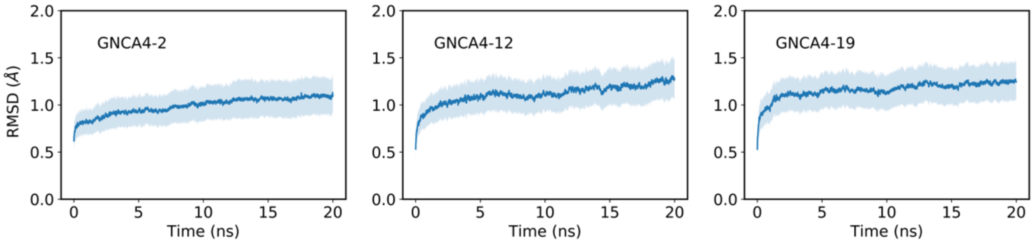


**Figure S3**. The root mean square deviations (RMSD, Å) of all backbone atoms of the three FuncLib variants studied in this work for which crystal structures are available, at the approximate EVB transition state (λ = 0.5) for the Kemp elimination reaction catalyzed by these enzymes. Data was collected every 10 ps from the initial equilibration runs, and is shown as averages and standard deviations over ten individual 20 ns MD simulations per system (*i.e.* 200 ns cumulative simulation time per system). The average RMSD per system is denoted by solid blue lines, and the standard deviations per point over all trajectories are illustrated by the shaded area on each plot.

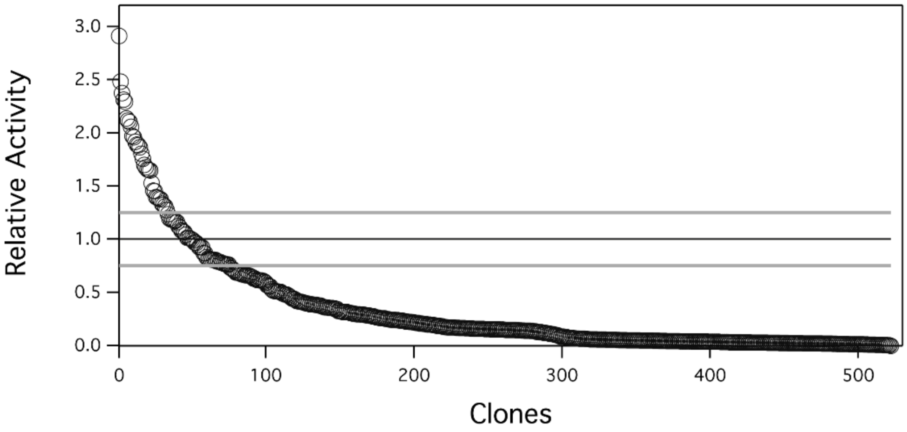


**Figure S4.** The Kemp eliminase activity of 522 clones from a random library prepared on the *de novo* GNCA4-WT *β*-lactamase (mutational load 3-5 mutations). The activity of these clones is shown relative to the activity of the background enzyme (shown as a black horizonal line). The grey horizontal lines represent the standard deviation interval for the background variant derived from measurements performed on 52 clones.

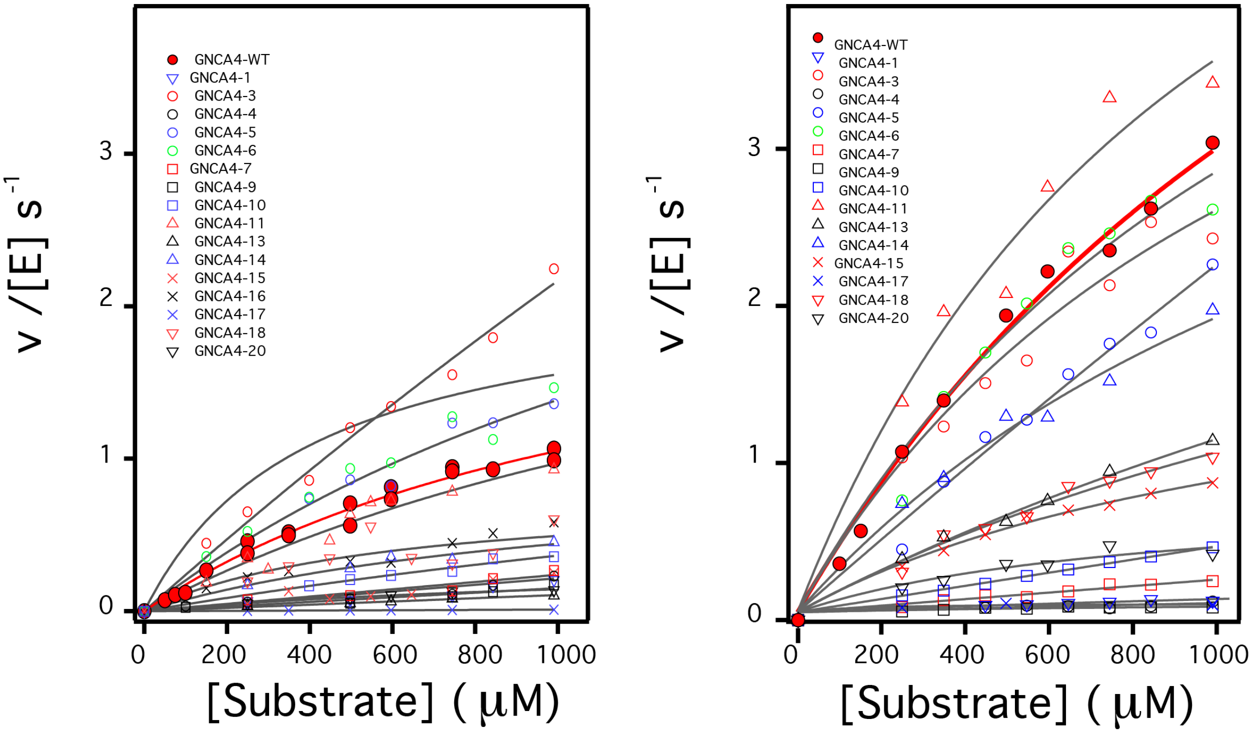


**Figure S5**. Plots of Kemp eliminase activity *vs*. substrate concentration at (**left**) pH 7 and (**right**) pH 8.4. Activities for the background protein (GNCA4-WT), as well as the 4 variants that display substantially enhanced catalysis at both pH values are found in **Figure 2**. Shown here are the activities of the GNCA4-WT and the remaining 16 variants from the top 20 variants from the FuncLib prediction (**Table S5**). The lines are the best fits of the Michaelis-Menten equation.

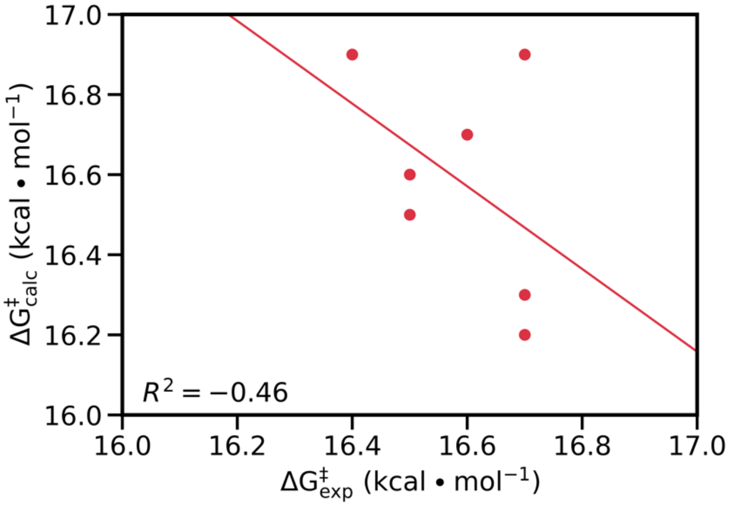


**Figure S6.** Correlation between calculated and experimental activation free energies for the Kemp elimination of 5-nitrobenzisoxazole by the GNCA4-WT and a series of active site mutants, calculated using linear regression analysis. The raw data for this figure is shown in **Table 2**. The correlation between the calculated and experimental activation free energies, calculated using linear regression analysis, is -0.46. Note that the differences in energies for each system are so small, that even very small thermodynamic differences can lead to weaker correlation with the experimental values.

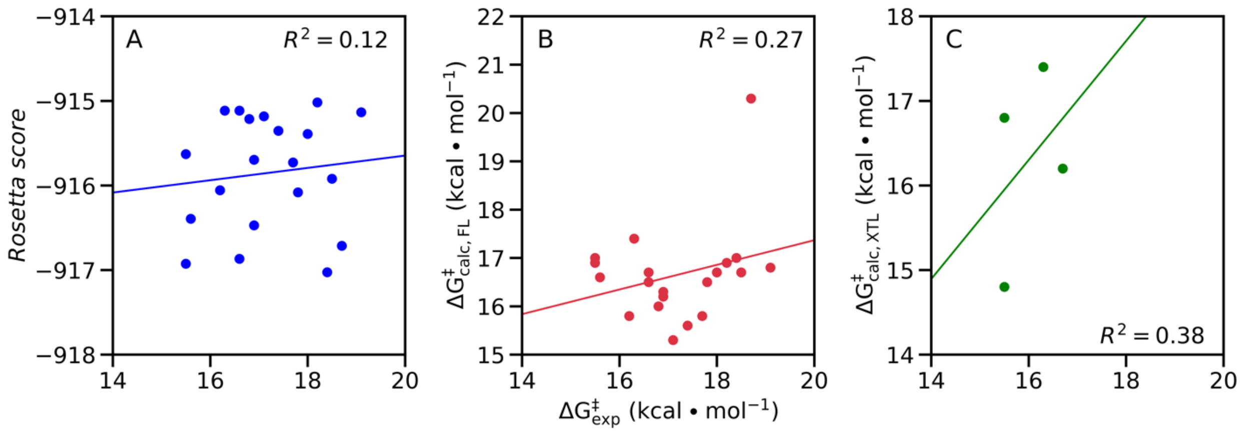


**Figure S7**. Correlation between (**A**) the Rosetta score from FuncLib and the experimental activation free energies ($\Delta G^{\ddagger}_{\mathrm{exp}}$), (**B**) the activation free energies calculated using the structure predictions from FuncLib ($\Delta G^{\ddagger}_{\mathrm{calc,FL}}$) and $\Delta G^{\ddagger}_{\mathrm{exp}}$, and (**C**) the activation free energies calculated directly from crystal structures, where available, and $\Delta G^{\ddagger}_{\mathrm{exp}}$. The raw data for this figure can be found in **Table S5**. Note that, for consistency, we did not include the GNCA4-WT in the correlation calculations for panels (**A**) and (**B**), as this is not a FuncLib predicted variant. As can be seen, in terms of the correlation between the calculated and experimental values, there is a weak correlation between calculated and experimental activation free energies ($R^2$ = 0.27, calculated using linear regression analysis, note that we have removed the GNCA4-WT from this correlation as this is not a FuncLib predicted structure). This is, however, due to the fact that the energy differences involved are, from a computational perspective, so small that even small deviations from the experimental value will lead to weak correlation with experiment. In terms of the comparison between the Rosetta score obtained from FuncLib (**Table S4**) and the experimental activation free energy, we obtain essentially no correlation with experiment ($R^2$ = 0.12, again omitting the GNCA4-WT for the same reason as above), which likely reflects the fact that the FuncLib ranking does not include any information about the substrate or transition state, and is based exclusively on the stability of the scaffold.[1] Similarly, for the variants where we have crystal structures available (GNCA4-WT, GNCA4-2, GNCA4-12 and GNCA4-19), we obtain similar correlation between calculated and experimental activation free energies ($R^2$ = 0.38), although this is a correlation over only 4 enzyme variants, and the energy difference between the calculated and experimental values is always within ~1 kcal·mol$^{-1}$ of the experimental value, indicating again that the weak correlation coefficients in this specific case are mainly due to the very small energy differences involved (which are within the resolution of EVB and other QM/MM methodologies as described in the main text), rather than a problem with the method.

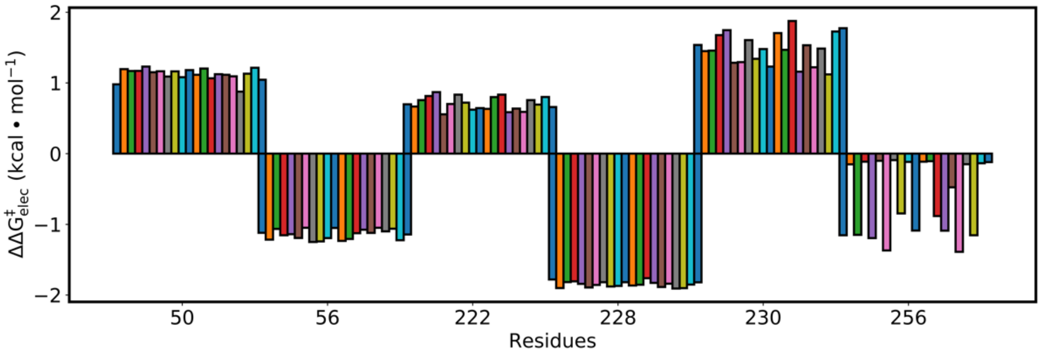


**Figure S8.** The electrostatic contributions of individual residues to the calculated activation free energies (ΔΔG‡elec) for the Kemp elimination of 5-nitrobenzisoxazole by the top 20 best scoring GNCA4 variants predicted by FuncLib.[1] All values were obtained by applying the linear response approximation (LRA)[2, 3] to the calculated EVB trajectories, as in our previous works,[4-6] and scaled assuming a dielectric constant of 4 for the highly hydrophobic environment of the *de novo* active site of this *β*-lactamase (**Figure 1**). Note that the deviations observed for residue 256 are due to the mutation of this residue (**Table S1**).

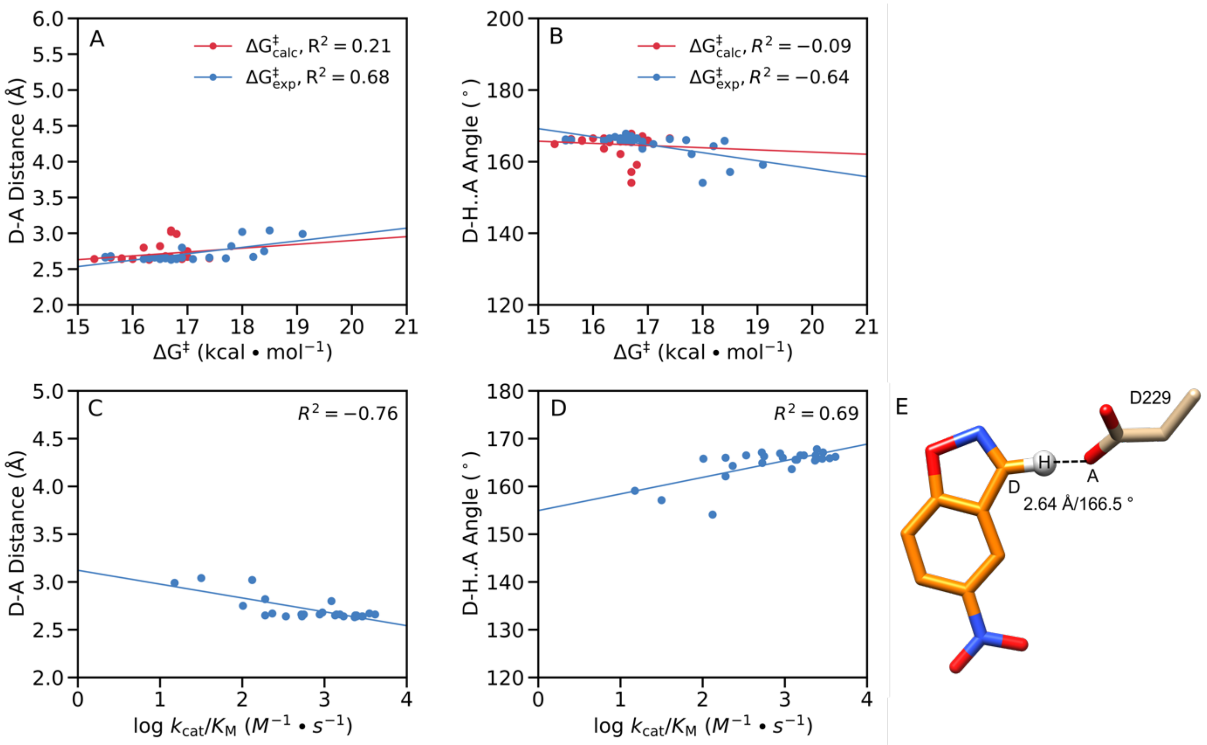


**Figure S9.** Correlations between the calculated and experimental activation free energies and the (**A**, **C**) donor-acceptor (D-A) distances (Å) and (**B**, **D**) donor-hydrogen-acceptor (D-H…A) angles (°) in our EVB simulations, calculated based on the data presented in **Tables 2, 3, S5, S7** and **S8**, using linear regression analysis. Correlations between the geometric parameters and (**A**, **B**) calculated activation free energies or (**C**, **D**) log $k_{\mathrm{cat}}/K_{\mathrm{M}}$ are shown here for all variants considered in this work, both single-point mutations and FuncLib predictions, with the exception of the GNCA4-4 variant, which is an outlier in the data as shown in **Figure 9**. (**E**) Schematic overview of the orientation of the reacting fragments in the wild-type enzyme. The annotated distance and angle are the average values from our EVB simulations of the wild-type enzyme (**Tables S7** and **S8**).

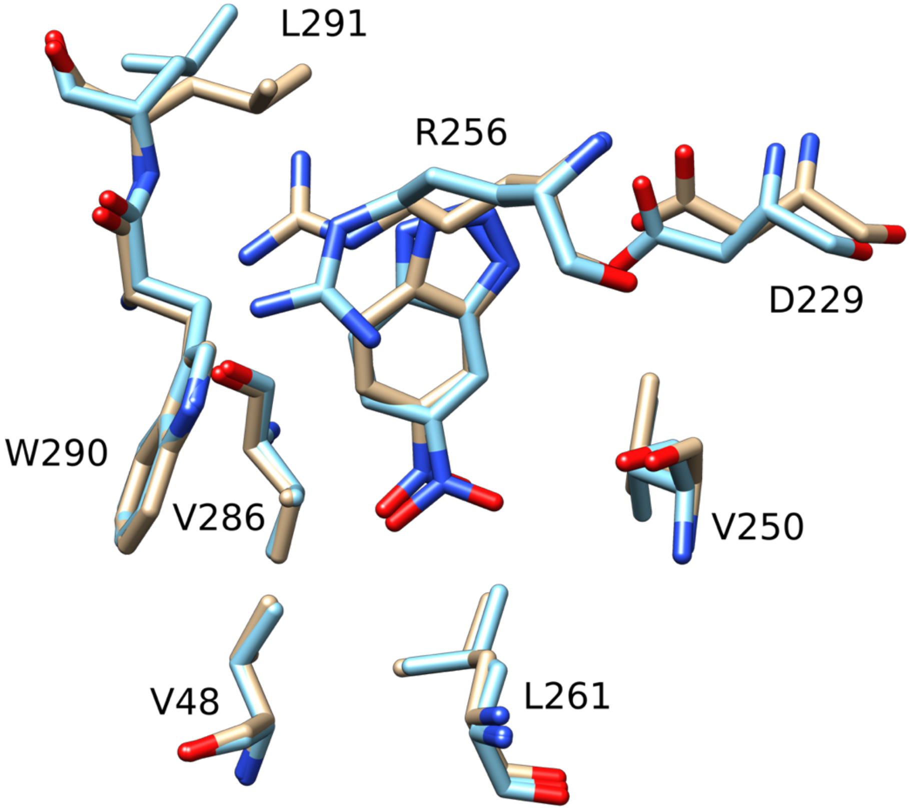


**Figure S10.** Overlay of the crystal structures of the GNCA4-2 variant obtained *via* (**blue**) X-ray crystallography (PDB ID: 6TY6) and (**tan**) FuncLib prediction.

# Supplementary Tables

**Table S1.** Sequence space explored after diversification of 11 active site residues by the FuncLib webserver.[1]

| **Original Residue** | **FuncLib Predictions** |
|---|---|
| V48 | VIL |
| D50 | D |
| I250 | ILMV |
| R256 | RHKQ |
| L260 | LFIMV |
| V261 | VILM |
| L285 | LAVW |
| V286 | VAILM |
| V287 | VAILMST |
| W290 | W |
| H291 | HEFIKLMNQRTV |

**Table S2.** List of ionized residues as well as the protonation patterns of histidine residues in EVB simulations of the *β*-lactamase catalyzed cleavage of 5-nitrobenzisoxazole *via* Kemp elimination.[a]

| Residue Type | Residue Number |
|---|---|
| Asp | 50, 209, 218, 228, 229, 233, 246, 273, 276 |
| Glu | 281 |
| Arg | 55, 56, 191, 204, 220, 222, 230, 256, 284 |
| Lys | 215, 219, 234 |
| His-δ | 122, 241 |
| His-ε | None |

[a] All residues not listed here were kept in their unionized forms during the simulations, as they fell outside the explicit simulation sphere (see the **Methodology** section of the main text). Protonation states and numbering based on residue numbering in PDB ID: 5FQK.[7]

**Table S3.** Percentage acetonitrile (%ACN) used during kinetic measurements at different substrate concentrations.

| %ACN | 5-Nitrobenzisoxazole Concentration Range (mM) |
|---|---|
| 1 | 0-1 |
| 3 | 0-2.2 |
| 5 | 0-2.4 |
| 7 | 0-2.9 |
| 9 | 0-2.9 |

**Table S4.** Data collection and refinement statistics of the 3D structural models.[a]

| | GNCA4-2 | GNCA4-12 | GNCA4-19 |
|---|---|---|---|
| PDB ID | 6TY6 | 6TXD | 6TWW |
| Wavelength (Å) | 0.97926 | 0.97926 | 0.97926 |
| Resolution range | 47.35 - 1.8 (1.864 - 1.8) | 71.74 - 2.0 (2.071 - 2.0) | 47.31 - 1.381 (1.431 - 1.381) |
| Space group | *I*222 | *I*222 | *P*6$_5$22 |
| Unit cell (Å, °) | 78.38 148.35 246.01<br>90 90 90 | 77.31 148.23 245.96<br>90 90 90 | 78.23 78.2 198.32<br>90 90 120 |
| Total reflections | 678420 (70746) | 425809 (42214) | 1134923 (70380) |
| Unique reflections | 130796 (13114) | 95172 (9368) | 74330 (7295) |
| Multiplicity | 5.2 (5.4) | 4.5 (4.5) | 15.3 (9.6) |
| Completeness (%) | 98.56 (99.39) | 98.99 (98.87) | 99.98 (100.00) |
| Mean I/sigma(I) | 10.77 (1.20) | 11.89 (2.63) | 27.24 (2.43) |
| Wilson B-factor (Å$^2$) | 30.31 | 28.62 | 15.91 |
| R-merge | 0.08955 (1.443) | 0.1031 (0.6784) | 0.04978 (0.6767) |
| CC1/2 | 0.997 (0.678) | 0.993 (0.81) | 1 (0.887) |
| Refinement | | | |
| R-work | 0.2005 (0.3722) | 0.2125 (0.2954) | 0.1505 (0.2004) |
| R-free | 0.2280 (0.3997) | 0.2373 (0.3297) | 0.1659 (0.2132) |
| CC(work) | 0.965 (0.831) | 0.953 (0.877) | 0.969 (0.930) |
| CC(free) | 0.952 (0.798) | 0.951 (0.838) | 0.975 (0.919) |
| Number of atoms | 7290 | 6729 | 2617 |
| protein | 6548 | 6248 | 2209 |
| ligands | 138 | 59 | 36 |
| solvent | 604 | 422 | 372 |
| Number of chains | 3 | 3 | 1 |
| RMS(bonds) (Å) | 0.019 | 0.003 | 0.013 |
| RMS(angles) (°) | 1.44 | 0.60 | 1.27 |
| Ramachandran favored (%) | 98.35 | 98.48 | 98.11 |
| Ramachandran outliers (%) | 0.00 | 0.00 | 0.00 |
| Rotamer outliers (%) | 2.38 | 1.25 | 0.87 |
| Average B-factor (Å$^2$) | 40.03 | 38.25 | 20.22 |
| macromolecules (Å$^2$) | 39.04 | 37.96 | 17.35 |
| ligands (Å$^2$) | 56.53 | 51.32 | 37.30 |
| solvent (Å$^2$) | 46.90 | 40.63 | 35.59 |
| Number of TLS groups | 18 | 21 | 5 |

[a] Statistics for the highest-resolution shell are shown in parentheses.

**Table S5**. Rosetta scores[8] and calculated and experimental activation free energies for the GNCA4-WT *β*-lactamase, as well as the top twenty variants predicted from FuncLib.[a]

| Variant | Rosetta Score | $\Delta G^{‡}_{exp}$ | $\Delta G^{‡}_{calc,XTL}$ | $\Delta G^{‡}_{calc,FL}$ | $\Delta\Delta G^{‡}_{exp\rightarrow calc,XTL}$ | $\Delta\Delta G^{‡}_{exp\rightarrow calc,FL}$ |
|---|---|---|---|---|---|---|
| GNCA4-WT | -906.616 | 16.7 | 16.2 ± 0.1 | - | -0.5 | - |
| GNCA4-1 | -917.026 | 18.4 | - | 17.0 ± 0.3 | - | -1.4 |
| GNCA4-2 | -916.926 | 15.5 | 14.8 ± 0.3 | 17.0 ± 0.2 | -0.7 | 1.5 |
| GNCA4-3 | -916.868 | 16.6 | - | 16.5 ± 0.2 | - | -0.1 |
| GNCA4-4 | -916.714 | 18.7 | - | 20.3 ± 0.5 | - | 1.6 |
| GNCA4-5 | -916.472 | 16.9 | - | 16.2 ± 0.3 | - | -0.7 |
| GNCA4-6 | -916.394 | 15.6 | - | 16.6 ± 0.2 | - | 1.0 |
| GNCA4-7 | -916.082 | 17.8 | - | 16.5 ± 0.3 | - | -1.3 |
| GNCA4-8 | -916.056 | 16.2 | - | 15.8 ± 0.2 | - | -0.4 |
| GNCA4-9 | -915.920 | 18.5 | - | 16.7 ± 0.4 | - | -1.8 |
| GNCA4-10 | -915.728 | 17.7 | - | 15.8 ± 0.2 | - | -1.9 |
| GNCA4-11 | -915.696 | 16.9 | - | 16.3 ± 0.3 | - | -0.6 |
| GNCA4-12 | -915.629 | 15.5 | 16.8 ± 0.2 | 16.9 ± 0.2 | 1.3 | 1.4 |
| GNCA4-13 | -915.391 | 18.0 | - | 16.7 ± 0.4 | - | -1.3 |
| GNCA4-14 | -915.354 | 17.4 | - | 15.6 ± 0.2 | - | -1.8 |
| GNCA4-15 | -915.214 | 16.8 | - | 16.0 ± 0.2 | - | -0.8 |
| GNCA4-16 | -915.183 | 17.1 | - | 15.3 ± 0.2 | - | -1.8 |
| GNCA4-17 | -915.135 | 19.1 | - | 16.8 ± 0.5 | - | -2.3 |
| GNCA4-18 | -915.116 | 16.6 | - | 16.7 ± 0.2 | - | 0.1 |
| GNCA4-19 | -915.115 | 16.3 | 17.4 ± 0.2 | 17.4 ± 0.2 | 1.1 | 1.1 |
| GNCA4-20 | -915.018 | 18.2 | - | 16.9 ± 0.3 | - | -1.3 |

[a] The GNCA4-WT *β*-lactamase, which is used as the baseline for our study, is referred to in this table as “wild-type” (“GNCA4-WT”). Experimental activation free energies ($\Delta G^{‡}_{exp}$) were derived from $k_{cat}$, where available, based on kinetic data presented in **Tables 2** and **3** (note that all calculations were performed without a His-tag, and therefore kinetic data from **Table 2** was used for the GNCA4-WT). Calculated activation free energies ($\Delta G^{‡}_{calc,XTL}$ if calculated based on an available crystal structure, and ($\Delta G^{‡}_{calc,FL}$ if calculated based on the structure predicted by FuncLib) are presented as average values and standard error of the mean over thirty independent EVB trajectories per system. The $\Delta\Delta G^{‡}$ values represent the difference between the experimental activation free energy and the calculated activation free energy based on crystal structures or structures obtained from FuncLib, respectively. All energies are presented in kcal·mol$^{-1}$. ‘-’ indicates ‘data not available’. For the full list of FuncLib[1] predictions, see the **Supplementary Data**.

**Table S6.** Amino acid substitutions introduced in the twenty top-ranked FuncLib[1] variants.

| **Variant** | **48** | **50** | **250** | **256** | **260** | **261** | **285** | **286** | **287** | **290** | **291** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GNCA4-WT | V | D | I | R | L | V | L | V | V | W | H |
| GNCA4-1 | V | D | I | H | F | I | L | V | L | W | I |
| GNCA4-2 | V | D | V | R | F | L | L | V | V | W | L |
| GNCA4-3 | V | D | I | H | F | I | L | V | V | W | H |
| GNCA4-4 | V | D | M | R | F | I | L | V | L | W | I |
| GNCA4-5 | V | D | I | H | F | L | L | V | V | W | I |
| GNCA4-6 | V | D | V | R | F | I | L | I | V | W | I |
| GNCA4-7 | V | D | M | H | F | L | L | V | V | W | K |
| GNCA4-8 | V | D | I | R | F | I | L | V | V | W | K |
| GNCA4-9 | V | D | M | H | F | I | L | V | A | W | H |
| GNCA4-10 | V | D | I | R | F | L | L | V | L | W | I |
| GNCA4-11 | V | D | V | H | F | L | L | V | T | W | H |
| GNCA4-12 | V | D | V | H | F | L | L | V | V | W | V |
| GNCA4-13 | V | D | M | R | F | L | L | V | V | W | H |
| GNCA4-14 | V | D | I | R | F | L | L | V | A | W | Q |
| GNCA4-15 | I | D | I | R | F | L | L | I | V | W | I |
| GNCA4-16 | V | D | V | R | V | L | L | V | L | W | N |
| GNCA4-17 | V | D | M | H | V | I | L | V | V | W | I |
| GNCA4-18 | V | D | V | R | I | I | L | I | V | W | L |
| GNCA4-19 | V | D | V | H | F | I | L | I | V | W | H |
| GNCA4-20 | V | D | I | H | I | I | L | V | T | W | N |

**Table S7**. Average donor-acceptor (D-A) distances and donor-hydrogen-acceptor (D-H…A) angles obtained from EVB simulations of different experimentally characterized variants of the GNCA4-WT *β*-lactamase.[a]

| **Variants** | **D-A** | | | **D-H..A** | | |
|---|---|---|---|---|---|---|
| | **MC** | **TS** | **PC** | **MC** | **TS** | **PC** |
| GNCA4-WT | 2.64±0.06 | 2.63±0.07 | 3.30±0.26 | 166.5±6.3 | 167.6±5.9 | 148.3±19.0 |
| A146G | 2.63±0.07 | 2.63±0.06 | 3.24±0.20 | 165.4±6.3 | 166.2±6.4 | 155.3±12.1 |
| A173V | 2.64±0.08 | 2.63±0.06 | 3.24±0.19 | 165.8±6.7 | 168.2±5.9 | 156.1±12.7 |
| G62S | 2.64±0.07 | 2.63±0.06 | 3.22±0.21 | 167.1±6.0 | 167.3±6.5 | 154.1±12.5 |
| L265Q | 2.65±0.08 | 2.63±0.06 | 3.25±0.19 | 167.8±6.9 | 169.1±6.1 | 155.0±14.9 |
| R256A | 2.66±0.07 | 2.64±0.06 | 3.29±0.20 | 166.9±6.0 | 167.8±6.4 | 155.3±10.6 |
| R256K | 2.66±0.08 | 2.65±0.07 | 3.29±0.22 | 166.5±6.6 | 166.3±5.7 | 151.3±19.2 |

[a] D-A distances are presented in Å and D-H..A angles are presented in °. Data is shown as average values and standard deviations over 30 independent trajectories. MC, TS and PC denote the Michaelis complex, transition state, and product complex, respectively.

**Table S8**. Average donor-acceptor (D-A) distances and donor-hydrogen-acceptor (D-H…A) angles obtained from EVB simulations of the GNCA4-WT *β*-lactamase, as well as of the top twenty variants predicted from FuncLib.[a]

| **Variants** | **D-A** | | | **D-H..A** | | |
|---|---|---|---|---|---|---|
| | **MC** | **TS** | **PC** | **MC** | **TS** | **PC** |
| GNCA4-WT | 2.64 ±0.06 | 2.63 ±0.07 | 3.30 ±0.26 | 166.5 ±6.3 | 167.6 ±5.9 | 148.3 ±19.0 |
| GNCA4-1 | 2.75 ±0.49 | 2.64 ±0.07 | 3.33 ±0.19 | 165.8 ±8.3 | 167.7 ±5.8 | 153.1 ±11.9 |
| GNCA4-2 | 2.67 ±0.07 | 2.64 ±0.07 | 3.32 ±0.18 | 165.9 ±6.9 | 169.0 ±5.6 | 154.2 ±12.7 |
| GNCA4-3 | 2.65 ±0.07 | 2.65 ±0.06 | 3.37 ±0.20 | 165.6 ±7.3 | 167.0 ±6.1 | 156.2 ±12.2 |
| GNCA4-4 | 4.25 ±0.96 | 2.65 ±0.07 | 3.37 ±0.20 | 129.8 ±21.5 | 167.1 ±7.1 | 151.0 ±15.3 |
| GNCA4-5 | 2.80 ±0.62 | 2.64 ±0.07 | 3.36 ±0.18 | 163.6 ±10.3 | 167.6 ±6.0 | 154.7 ±10.1 |
| GNCA4-6 | 2.68 ±0.10 | 2.64 ±0.06 | 3.23 ±0.16 | 166.0 ±6.5 | 167.4 ±6.0 | 154.9 ±10.2 |
| GNCA4-7 | 2.82 ±0.64 | 2.64 ±0.06 | 3.28 ±0.21 | 162.1 ±10.8 | 168.0 ±5.7 | 155.1 ±13.0 |
| GNCA4-8 | 2.64 ±0.07 | 2.64 ±0.07 | 3.33 ±0.20 | 165.8 ±6.3 | 167.7 ±5.6 | 154.7 ±12.7 |
| GNCA4-9 | 3.04 ±0.87 | 2.64 ±0.06 | 3.25 ±0.17 | 157.1 ±17.7 | 168.7 ±5.4 | 155.1 ±11.9 |
| GNCA4-10 | 2.65 ±0.08 | 2.64 ±0.07 | 3.32 ±0.23 | 166.0 ±5.8 | 168.0 ±5.3 | 154.0 ±11.6 |
| GNCA4-11 | 2.66 ±0.08 | 2.64 ±0.06 | 3.30 ±0.17 | 165.6 ±7.0 | 167.2 ±6.6 | 154.4 ±9.4 |
| GNCA4-12 | 2.66 ±0.08 | 2.64 ±0.07 | 3.25 ±0.16 | 166.2 ±6.7 | 168.5 ±5.5 | 153.9 ±10.0 |
| GNCA4-13 | 3.02 ±0.77 | 2.65 ±0.06 | 3.32 ±0.24 | 154.1 ±20.3 | 167.7 ±6.1 | 149.5 ±21.7 |
| GNCA4-14 | 2.66 ±0.08 | 2.65 ±0.06 | 3.27 ±0.19 | 166.3 ±5.7 | 167.4 ±6.1 | 152.6 ±10.9 |
| GNCA4-15 | 2.64 ±0.07 | 2.64 ±0.07 | 3.34 ±0.18 | 166.5 ±6.3 | 167.3 ±6.4 | 155.4 ±11.5 |
| GNCA4-16 | 2.64 ±0.07 | 2.64 ±0.06 | 3.23 ±0.15 | 164.9 ±6.4 | 168.4 ±5.8 | 152.2 ±9.0 |
| GNCA4-17 | 2.99 ±0.78 | 2.64 ±0.06 | 3.30 ±0.19 | 159.1 ±16.7 | 168.0 ±6.4 | 154.7 ±16.3 |
| GNCA4-18 | 2.66 ±0.08 | 2.64 ±0.07 | 3.26 ±0.19 | 167.1 ±5.9 | 167.8 ±5.6 | 152.0 ±14.4 |
| GNCA4-19 | 2.65 ±0.08 | 2.66 ±0.08 | 3.26 ±0.19 | 166.5 ±6.0 | 166.7 ±6.5 | 155.0 ±10.2 |
| GNCA4-20 | 2.67 ±0.07 | 2.65 ±0.07 | 3.24 ±0.17 | 164.3 ±8.0 | 166.7 ±6.3 | 154.0 ±10.0 |

[a] D-A distances are presented in Å and D-H..A angles are presented in °. Data is shown as average values and standard deviations over 30 independent trajectories. MC, TS and PC denote the Michaelis complex, transition state, and product complex, respectively.

# Supplementary References

1. O. Khersonsky, R. Lipsh, Z. Avizemer, Y. Ashani, M. Goldsmith, H. Leader, O. Dym, S. Rogotner, D. L. Trudeau, J. Prilusky, P. Amengual-Rigo, V. Guallar, D. S. Tawfik and S. J. Fleishman, *Mol. Cell*, 2018, **72**, 178-186.e175.
2. F. S. Lee and A. Warshel, *J. Chem. Phys.*, 1992, **97**, 3100.
3. I. Muegge, H. Tao and A. Warshel, *Protein Eng. Des. Sel.*, 1997, **10**, 1363-1372.
4. Y. S. Kulkarni, Q. Liao, D. Petrović, D. M. Krüger, B. Strodel, T. L. Amyes, J. P. Richard and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2017, **139**, 10514-10525.
5. Y. S. Kulkarni, Q. Liao, F. Byléhn, T. L. Amyes, J. P. Richard and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2018, **140**, 3854-3857.
6. Y. S. Kulkarni, T. L. Amyes , J. P. Richard and S. C. L. Kamerlin, *J. Am. Chem. Soc.*, 2019, **141**, 16139-16150.
7. V. A. Risso, S. Martinez-Rodriguez, A. M. Candel, D. M. Krüger, D. Pantoja-Uceda, M. Ortega-Muñoz, F. Santoyo-Gonzlez, E. A. Gaucher, S. C. L. Kamerlin, M. Bruix, J. A. Gavira and J. M. Sanchez-Ruiz, *Nat. Commun.*, 2017, **8**, 16113.
8. A. V. Morozov and T. Kortemme, *Adv. Protein Chem.*, 2005, **72**, 1-38.

# PUBLICATION II

## Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme

International Journal of
*Molecular Sciences*

International Journal of
*Molecular Sciences*

Article

# Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme

Luis I. Gutierrez-Rus, Miguel Alcalde, Valeria A. Risso and Jose M. Sanchez-Ruiz

https://doi.org/10.3390/ijms23168934

International Journal of
*Molecular Sciences*

MDPI

*Article*

# Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme

**Luis I. Gutierrez-Rus [1], Miguel Alcalde [2], Valeria A. Risso [1,*] and Jose M. Sanchez-Ruiz [1,*]**

[1] Departamento de Quimica Fisica, Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain

[2] Department of Biocatalysis, Institute of Catalysis and Petrochemistry, CSIC, Cantoblanco, 28049 Madrid, Spain

* Correspondence: vrisso@ugr.es (V.A.R.); sanchezr@ugr.es (J.M.S.-R.)

**Abstract:** The routine generation of enzymes with completely new active sites is a major unsolved problem in protein engineering. Advances in this field have thus far been modest, perhaps due, at least in part, to the widespread use of modern natural proteins as scaffolds for de novo engineering. Most modern proteins are highly evolved and specialized and, consequently, difficult to repurpose for completely new functionalities. Conceivably, resurrected ancestral proteins with the biophysical properties that promote evolvability, such as high stability and conformational diversity, could provide better scaffolds for de novo enzyme generation. Kemp elimination, a non-natural reaction that provides a simple model of proton abstraction from carbon, has been extensively used as a benchmark in de novo enzyme engineering. Here, we present an engineered ancestral β-lactamase with a new active site that is capable of efficiently catalyzing Kemp elimination. The engineering of our Kemp eliminase involved minimalist design based on a single function-generating mutation, inclusion of an extra polypeptide segment at a position close to the de novo active site, and sharply focused, low-throughput library screening. Nevertheless, its catalytic parameters ($k_{cat}/K_M$~$2{\cdot}10^5$ $M^{-1}$ $s^{-1}$, $k_{cat}$~635 $s^{-1}$) compare favorably with the average modern natural enzyme and match the best proton-abstraction de novo Kemp eliminases that are reported in the literature. The general implications of our results for de novo enzyme engineering are discussed.

**Keywords:** enzyme design; de novo enzymes; Kemp elimination; ancestral enzymes; β-lactamases; focused library screening; focused directed evolution

check for updates

**Citation:** Gutierrez-Rus, L.I.; Alcalde, M.; Risso, V.A.; Sanchez-Ruiz, J.M. Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme. *Int. J. Mol. Sci.* **2022**, *23*, 8934. https://doi.org/10.3390/ijms23168934

Academic Editors: Carlos Gutierrez-Merino, José Manuel González Ros and Jose Javier Lopez

Received: 28 July 2022
Accepted: 9 August 2022
Published: 11 August 2022

## 1. Introduction

In the early 1960s, Linus Pauling and Emile Zurkenkandl published two papers that played a crucial role in the emergence and subsequent development of the molecular evolution field. In the first paper [1], they introduced the molecular clock hypothesis and the possibility of using sequences of homolog proteins to estimate species divergence times. In the second paper [2], they proposed that plausible approximations to the sequences of ancestral proteins can be derived from suitable analyses of the known sequences of their modern counterparts. Ancestral sequence reconstruction has been amply used in the post-genomic era as a tool to address fundamental problems in molecular evolution [3–5]. Furthermore, it has been found that ancestral proteins that are "resurrected" in the lab (i.e., the proteins encoded by the reconstructed sequences) often display interesting and even extreme properties [6–16] that, plausibly, reflect ancestral adaptations to unusual intra- and extra-cellular environments. Resurrected ancestral proteins often display high stability, supporting the frequently hypothesized thermophilic nature of ancient life. In addition, they often show efficient heterologous expression in modern hosts, possibly reflecting their emergence prior to the availability of efficient cellular folding assistance. Plausibly, unlike many modern proteins, they do not rely on having adapted to (i.e., having co-evolved with) the folding assistance machinery of the host to fold efficiently. Finally, in a significant number of ancestral reconstruction studies, the resurrected enzymes were found to be able to catalyze several related reactions. Such promiscuity, to use the accepted term in the field, may be consistent with Jensen's proposal, many years ago [17], that primordial enzymes were generalists. Alternatively, perhaps the promiscuous resurrected ancestral enzymes corresponded to pre-duplication nodes in the evolution of new functionalities [10].

Beyond specific evolutionary narratives, it is clear that the unusual properties of resurrected proteins may be valuable in biotechnological application scenarios [5,7,12–14,18–22]. Promiscuity, for instance, is most likely linked to conformational diversity, i.e., to the fact that the protein exists in solution as an ensemble of more or less related conformations, with different functionalities linked to different subsets of conformations [23]. It is well known that conformational diversity promotes evolvability, i.e., the capability to evolve new functions [24–27]. The reason for this is that if a few rare conformations are competent for the targeted activity, their population in the ensemble can be enhanced through rational design or, more likely, through standard directed evolution. Enhanced stability, in addition to being a biotechnologically useful property by itself, is known to contribute to evolvability [28], as functionally useful but destabilizing mutations will be accepted in a high-stability protein scaffold, while they will likely compromise proper folding in a moderately stable protein.

Furthermore, and contrary to naïve expectations, enhanced conformational diversity and high stability are not necessarily mutually exclusive features [29]. In an ancestral reconstruction exercise targeting the antibiotic-resistance enzyme β-lactamase, published approximately ten years ago [7], we found that the resurrected proteins corresponding to ancient Precambrian phylogenetic nodes were not only highly stable, with denaturation temperatures about 30 °C above those of their modern mesophilic counterparts, but also promiscuous, being able to degrade a variety of lactam antibiotics, including third-generation antibiotics. For comparison, TEM-1 β-lactamase, a typical modern β-lactamase, is a penicillin specialist that displays very low levels of activity with other lactam antibiotics. Subsequent experimental and computational studies [30,31] confirmed that the substrate promiscuity of the ancestral β-lactamases was, indeed, linked to enhanced conformational diversity. Resurrected Precambrian β-lactamases are, therefore, highly-stable and conformationally diverse to some substantial extent.

Overall, it emerges that resurrected ancestral proteins in general, and our highly stable and conformationally diverse Precambrian β-lactamases in particular, could perhaps provide superior starting points for protein engineering. In order to systematically explore this possibility, we recently started a research program aimed at using our resurrected ancestral β-lactamases as scaffolds for de novo enzyme generation, which is a fundamental unsolved problem in protein engineering [32]. We targeted Kemp elimination (Figure 1), a simple model of a fundamental chemical process—proton abstraction from carbon—and an extensively used benchmark in de novo enzyme engineering. Specifically, we aimed at generating efficient Kemp eliminases with minimal design and screening efforts. In our first step [33], we succeeded in using a single-mutation, minimalist design to generate significant levels of Kemp elimination activity in ancestral β-lactamase scaffolds. In our second step [34], we improved these starting activities on the basis of ultra-low-throughput screening that was computationally focused on the new active site region.

Here, we report the third step along these lines. The interactions at the original de novo active site have probably been optimized to a substantial extent by our previous efforts. Therefore, we have devised a different approach, based on the extension of the protein via the introduction of an additional polypeptide segment near the new active site. The rationale behind this approach is that the extra segment should display conformational diversity; we may expect that some conformations and sequence patterns generate new interactions that promote catalysis. Overall, the extra polypeptide segment offers new interaction possibilities that can presumably be exploited for de novo activity enhancement. In fact, by screening a small library focused on the additional segment in this work, we reached a catalytic efficiency of $k_{cat}/K_M$~$2{\cdot}10^5$ $M^{-1}$ $s^{-1}$ and a turnover number of

$k_{cat}$~635 $s^{-1}$. These Michaelis–Menten catalytic parameters compare favorably with those for the average modern natural enzyme [35] and match the best proton-abstraction de novo Kemp eliminases reported in the literature [36]. The general implications of these results for de novo enzyme engineering are discussed in Section 3 of this paper.

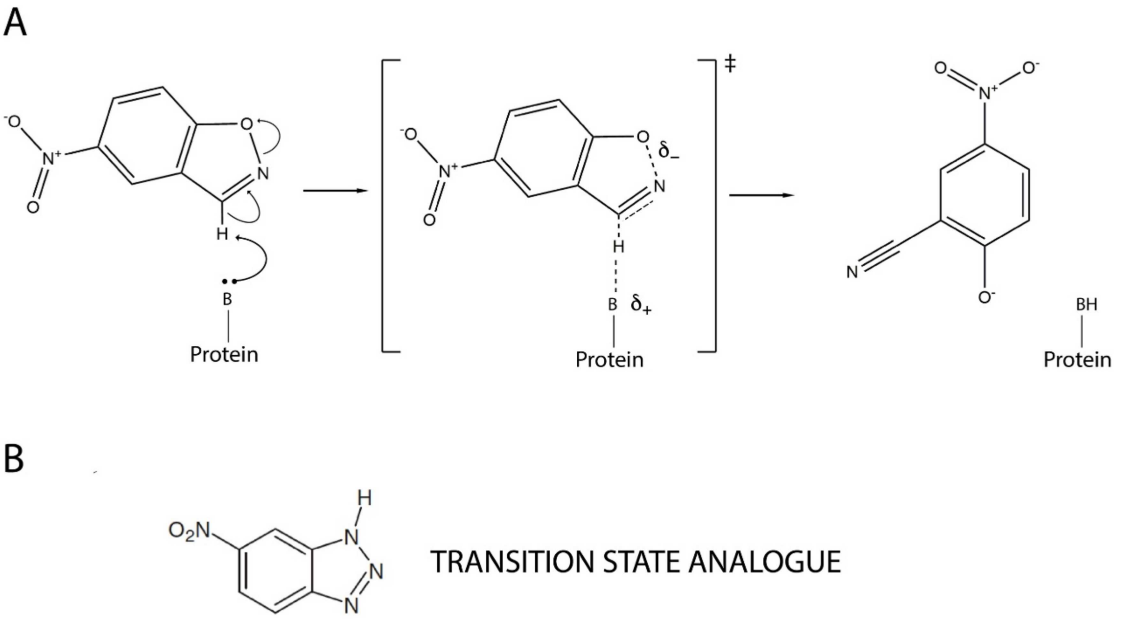


**Figure 1.** Mechanism of base-catalyzed Kemp elimination showing a proposed transition state structure (**A**). A transition state analogue, 5(6)-nitrobenzotriazole, is also shown (**B**).

## 2. Results

### *2.1. Kemp Eliminase Variants Used as Starting Point for This Work*

Recently [33], we generated a completely new active site for Kemp elimination—i.e., a site that was distinct from the antibiotic degradation active site—using a minimalist approach based on a single mutation. Specifically, a hydrophobic-to-ionizable mutation generated both a cavity for substrate binding and a catalytic base that was capable of performing proton abstraction. Both the high stability and the conformational diversity of the ancestral β-lactamase scaffolds that were used likely played roles in the success of the minimalist design [33,37]. The function-generating mutation, a tryptophan to aspartate amino acid replacement at position 229, is clearly disruptive and may lead to folding problems if implemented in a β-lactamase of moderate stability. Furthermore, the cavity produced by the mutation cannot exactly match the shape of the Kemp reactant; therefore, substrate binding must rely on local flexibility in the region of the new active site. Indeed, while the minimalist design was successful in a number of resurrected Precambrian β-lactamases, it failed to lead to significant de novo activity levels when implemented in 10 different modern β-lactamases.

The activity generated by W229D was found to be enhanced by a second F290W mutation [33], resulting in a catalytic efficiency of approximately $10^4$ $M^{-1}$ $s^{-1}$ with a turnover number of approximately 10 $s^{-1}$ for Kemp elimination catalysis. Figure 2A shows the 3D structure of the more active de novo Kemp eliminase that we initially obtained [33] through a scan of the function-generating W229D mutation on several resurrected ancestral β-lactamases. That structure includes a bound transition-state analogue (Figure 1B) that indicates the location of the engineered active site.

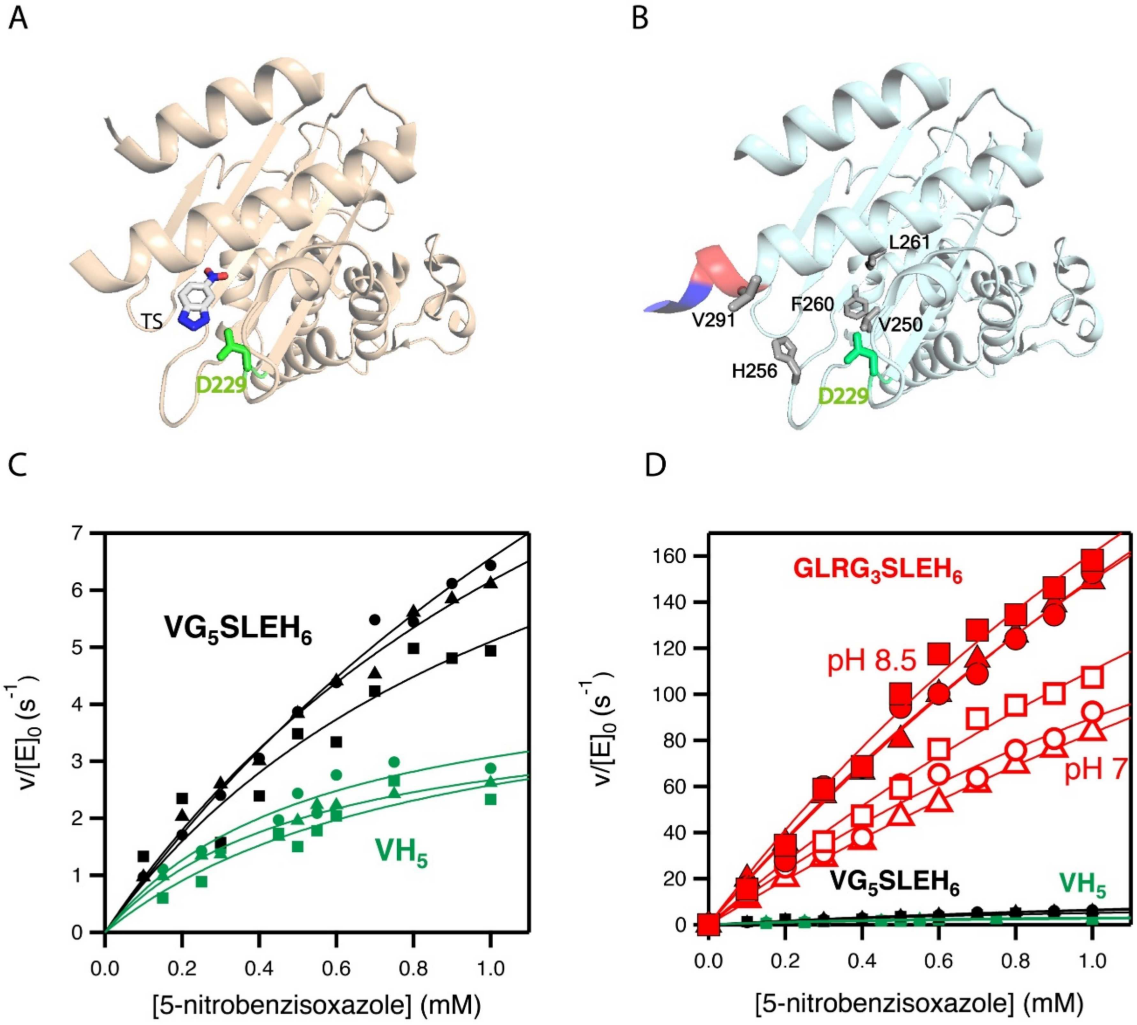


**Figure 2.** 3D structures and catalytic properties of the Kemp eliminase variants used as starting points for the enzyme engineering reported in this work. (**A**) Structure (PDB ID 5FQK) of the W229D/F290W variant of an ancestral β-lactamase with Kemp eliminase activity [33]. The W229D mutation generates the new function and the catalytic base (D229) introduced by the function-generation mutation is shown. The structure includes a transition state analogue (see Figure 1), which indicates the location of the new active site. (**B**) The Kemp elimination activity of the protein shown in A could be enhanced by several amino acid replacements at the de novo active site region, resulting from computationally focused ultra-low-throughput screening [34]. The activity-enhancing residues are highlighted in grey in the structure shown here (PDB ID 6TXD). The His-tag attached to the carboxyl terminus for purification purposes is also highlighted (red and blue colors). Note that one of the activity-enhancing mutations replaces the first histidine with a valine. Therefore, the protein shown has a VH$_5$ tail attached to the carboxyl terminus. The background variant used in this work has VG$_5$SLEH$_6$ attached to the carboxyl terminus, introducing a polypeptide segment between the caxboxyl terminus and the His-tag, but keeping the valine. (**C**) Michaelis–Menten profiles for the proteins with VH$_5$ and VG$_5$SLEH$_6$ attached to the carboxyl terminus. (**D**) The differences in catalysis observed in C are very small compared with the activity enhancement achieved in the screening efforts reported this work (variant with GLRG$_3$SLEH$_6$ attached to the carboxyl terminus). Note that three Michaelis–Menten profiles were independently determined for each of the variants shown in (**C**,**D**).

Subsequently [34], we used computationally focused screening to target the active site region, to further increase the Kemp elimination activity of our best W229D/F290W variant of an ancestral β-lactamase scaffold. The enhancement in catalytic efficiency that was obtained was moderate, but the turnover number was raised by approximately one order of magnitude. The mutations introduced at this stage are highlighted in the structure shown in Figure 2B. Note that in all our Kemp eliminases, the de novo active site is located near the carboxyl terminus and that His-tag is routinely attached to the carboxyl terminus residue to enable protein purification by affinity chromatography. Note also that one of the activity-enhancing mutations shown in Figure 2B actually replaces the first histidine residue of the purification tag with a valine.

### *2.2. Combinatorial Library Design and Screening*

In this work, we inserted a polypeptide segment between the carboxyl terminus residue and the His-tag (Figure 3A) of our previous best Kemp eliminase. This insertion retained the valine residue at the first position after the original carboxyl terminal residue, while the rest of the inserted segment included several glycine residues and a serine, following the known sequence design principles for soluble and flexible protein linkers [38]. Overall, the β-lactamase variant used here as starting point for directed evolution was identical to the best Kemp eliminase from our previous study [34], except for the presence of a (Gly)$_5$-Ser-Leu-Glu-(His)$_6$ segment between the extra valine at the carboxyl terminus and the purification His-tag. This insertion had a small effect on catalysis, as shown in Figure 2C by the Michaelis–Menten plots and catalytic parameters. In fact, as shown in Figure 2D, the effect of the insertion on catalysis was almost negligible when compared with the total activity enhancement that was achieved in this work.

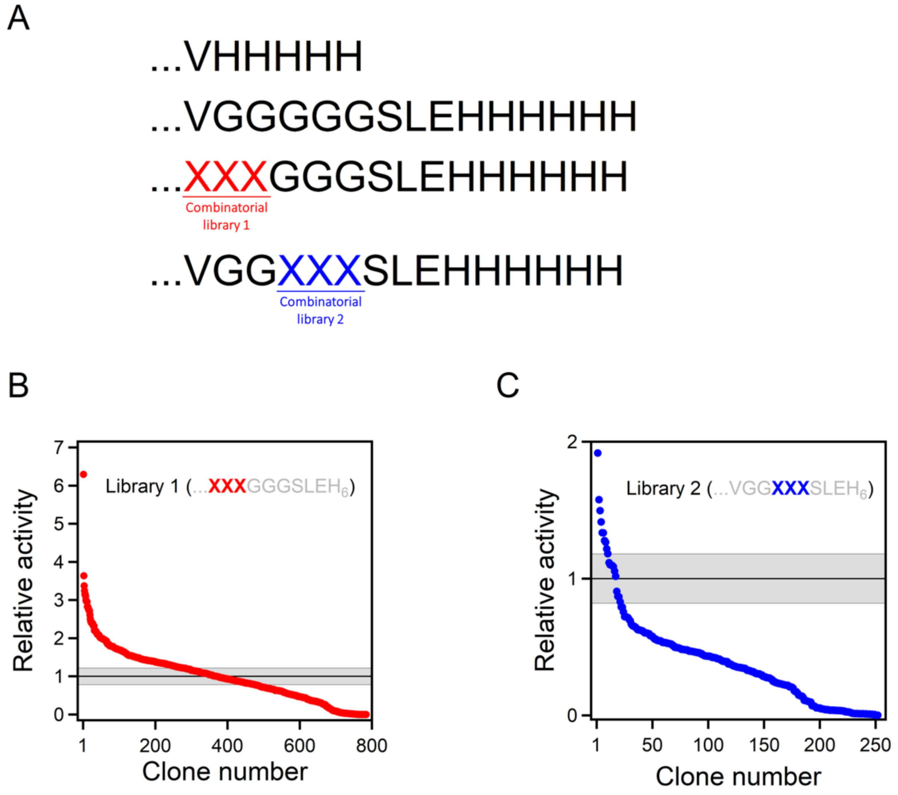


**Figure 3.** Combinatorial library screening for enhanced Kemp eliminase activity. (**A**) While the best Kemp eliminase from our previous work (34) had a VH$_5$ tail attached, the background variant used for library construction in this work had a VG$_5$SLEH$_6$ polypeptide attached to the carboxyl terminus. Two 8000-variant combinatorial libraries were prepared, including all possible combinations of the 20 amino acids at two sets of three positions (labelled XXX), as shown. (**B**,**C**) The result of the screening of the two libraries. Clones are ranked according to Kemp eliminase activity relative to the background variant. The grey strip represents the average activity of the background variant plus or minus the associated standard error.

The β-lactamase variant described above was used as background for a library that comprises all combinations of all possible amino acid residues at the three first positions after the carboxyl terminus (Figure 3A). The rationale behind this approach was that, as the carboxyl terminus is close to the de novo active site, some of the library variants may generate interactions that promote catalysis. The combinatorial library spanned $20^3$ = 8000 different amino acid sequences, and approximately 800 variants were screened for Kemp elimination activity, as described in Section 4 of this paper. In this primary screening, most variants displayed significant levels of Kemp elimination activity (Figure 3B), indicating that few of the mutations were disruptive and prevented proper folding. This is consistent with the fact that, in this case, library screening sampled the sequence space associated with a conformationally flexible polypeptide segment that was expected to remain largely exposed to the solvent in most cases. Nevertheless, a few of the variants were found to display substantially enhanced levels of catalysis for the Kemp elimination reaction (Figure 3B). The sequences of these variants were determined by Sanger sequencing; the corresponding combinations of residues at the randomized positions that enhance activity are provided in Table 1. Because catalysis relies on decreasing the activation free energy of the reaction, it appears reasonable to assume that these variants can generate interactions that stabilize the transition state (i.e., the chemical species at the top of the free energy barrier) for Kemp elimination at the de novo active site.

**Table 1.** Catalytic parameters for the cleavage of 5-nitrobenzisoxazole at pH 7 (HEPES 10 mM NaCl 100 mM) and 1% acetonitrile and 25 °C catalyzed by the engineered and evolved versions of Precambrian β-lactamases. For the best variant, data at pH 8.5 are included.

| Variant | Sequence [a] | $k_{cat}$ ($s^{-1}$) [b] | $K_M$ (mM) [b] | $k_{cat}/K_M$ ($s^{-1}$ $M^{-1}$) [b] | Tm (C°) [c] |
|---|---|---|---|---|---|
| V4 at pH 8.5 | … GLRGGG … | 635.0 ± 59.1 | 3.14 ± 0.49 | $(2.0 \pm 0.1) \times 10^5$ | - |
| V4 | … GLRGGG … | 407.5 ± 76.7 | 3.35 ± 0.91 | $(1.3 \pm 0.2) \times 10^5$ | 79.7 |
| V3 | … DIRGGG … | 202.9 ± 8.9 | 3.35 ± 0.21 | $(6.1 \pm 0.3) \times 10^4$ | 79.7 |
| V6 | … GLHGGG … | 74.7 ± 13.5 | 1.74 ± 0.46 | $(4.4 \pm 0.3) \times 10^4$ | 80.8 |
| V1 | … KLRGGG … | 54.1 ± 13.2 | 1.45 ± 0.37 | $(3.6 \pm 0.1) \times 10^4$ | 78.5 |
| V2 | … KSIGGG … | 54.7 ± 4.8 | 1.75 ± 0.32 | $(3.2 \pm 0.3) \times 10^4$ | 79.4 |
| V5 | … RGAGGG … | 55.4 ± 2.7 | 1.97 ± 0.1 | $(2.82 \pm 0.01) \times 10^4$ | 79.7 |
| V8 | … VGGRFI … | 53.5 ± 11.0 | 2.10 ± 0.61 | $(2.6 \pm 0.3) \times 10^4$ | 77.0 |
| V7 | … NNIGGG … | 55.4 ± 20.8 | 2.50 ± 1.43 | $(2.5 \pm 0.4) \times 10^4$ | 81.1 |
| V9 | … VGGAPL … | 23.7 ± 3.8 | 1.12 ± 0.16 | $(2.1 \pm 0.2) \times 10^4$ | 78.5 |
| V10 | … VGGGTP … | 15.9 ± 1.5 | 1.19 ± 0.32 | $(1.4 \pm 0.3) \times 10^4$ | 79.7 |
| BACKGROUND | … VGGGGG … | 16.3 ± 4.4 | 1.71 ± 0.48 | $(9.6 \pm 0.4) \times 10^3$ | 79.9 |

[a] The sequences of each variant at the randomly mutagenized regions in the libraries are shown. Variants are ranked according to the catalytic efficiency. All values correspond to pH 7, with the exception of the best variant (V4), for which data at pH 7 and pH 8.5 are provided. [b] The values shown for the Michaelis–Menten catalytic parameters are the average values from three independent replicates ($n$ = 3 independent determinations of the Michaelis–Menten profiles). The associated standard errors are also provided. See Supplementary Figure S1 and Supplementary Table S2 for the results of the analysis of the individual profiles. [c] Denaturation temperatures determined by differential scanning calorimetry (DSC) are shown. The error associated with $T_m$ determination by DSC is typically smaller than one degree.

### 2.3. Stability and Catalytic Parameters for the Improved Kemp Eliminases

The seven top variants of the primary screening described above were purified and their Michaelis–Menten profiles of rate versus substrate concentration were determined at pH 7 (Figure 4A and Supplementary Figure S1). This secondary screening confirmed the results of the primary screening, as all the selected variants showed substantially enhanced catalysis with respect to the library background (Figure 4A and Table 1). In particular, the variant with the sequence GLR at the three targeted positions showed an enhancement in catalytic parameters of approximately one order of magnitude over the library background. Note that most of the activity determinations reported in this work were performed at pH 7. However, it is known that the activity of Kemp eliminase enzymes based on the proton-abstraction mechanism may increase at basic pH, reflecting the deprotonation of the amino acid residue that gives rise to the catalytic base (the aspartate at position 229 in this case). Accordingly, we determined the profile of rate versus substrate concentration for the best GLR variant at pH 8.5. A significant increase in activity with respect to pH 7 was observed (Figure 2D), leading to a catalytic efficiency of $k_{cat}/K_M$~$2{\cdot}10^5$ $M^{-1}$ $s^{-1}$ and a turnover number of $k_{cat}$~635 $s^{-1}$ (Table 1).

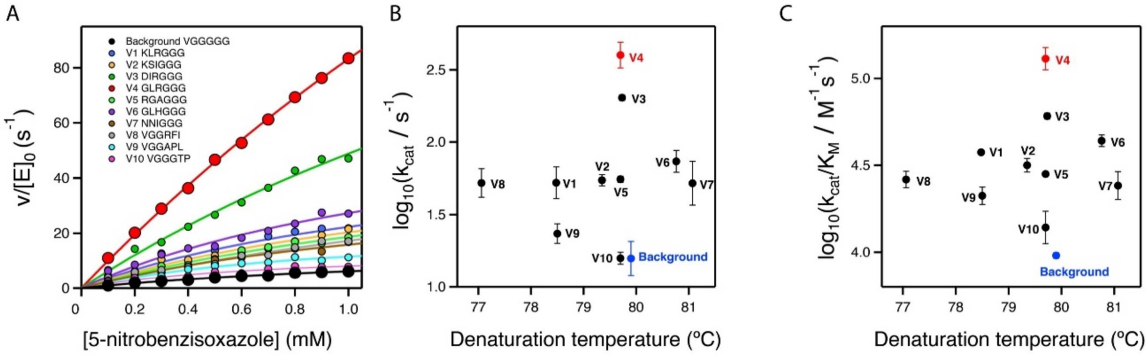


**Figure 4.** Secondary screening of the top variants from the primary library screening shown in Figure 3. (**A**) Michaelis–Menten profiles for the top variants at pH 7 (profiles for the best variant at pH 8.5 are shown in Figure 2B). The sequences at the relevant section of the included polypeptide are shown. Note that, for all variants, three independent Michaelis–Menten profiles were determined (see Supplementary Table S1 and Supplementary Figure S1); however, for the sake of clarity, only one representative profile for each variant is shown here. (**B**,**C**) Plots of Michaelis–Menten catalytic parameters for all the variants versus denaturation temperature, as determined by differential scanning calorimetry. These plots are scattergrams, indicating the absence of a significant stability/activity trade-off.

It is interesting that the achieved improvements in catalysis did not come with a significant cost in stability, as shown by the denaturation temperature values determined via differential scanning calorimetry (Figure 4B,C and Table 1). The lack of a substantial activity-stability trade-off can be attributed to the fact that our engineering approach did not manipulate already-existing interactions, but introduced new ones.

### 2.4. Structural Analysis of the Catalysis Enhancement

The two best Kemp eliminases from the library screening shared the presence of a bulky hydrophobic residue at the second position after the carboxyl terminus. It appears reasonable to assume that this hydrophobic residue establishes interactions that stabilize the transition state of the reaction and enhance catalysis. In order to explore this possibility, we applied AlphaFold2 [39,40] to the prediction of the 3D structure of our best Kemp eliminases (Figure 5). We carried out the prediction for sequences in which the function-generating mutation (tryptophan to aspartate at position 229 in the β-lactamase sequence) was omitted (Note, however, that the overall 3D structure predicted by AlphaFold2 remained essentially the same upon the W229D mutation: see Supplementary Figure S2). The reason for this is that W229 in β-lactamases is in roughly the same position as the bound transition state in the variants with the W229D mutation that showed Kemp elimination activity. Therefore, possible interactions that stabilize the bound transition state in the W229D variants may be suggested by the corresponding interactions with the tryptophan at position 229 in structures in which the W229D mutation is not included. Indeed, such interactions between the tryptophan and the hydrophobic residue at the second position are clearly suggested by the AlphaFold2 predictions (Figure 5). Finally, it is worth noting that the enhancement observed in the Kemp elimination activity was very unlikely to have been related to the lactam antibiotic activity of the β-lactamases we used as scaffolds for engineering, given that the two active sites (de novo for Kemp elimination and natural for antibiotic degradation) were well apart in both predicted and experimental structures (see Supplementary Figure S2).

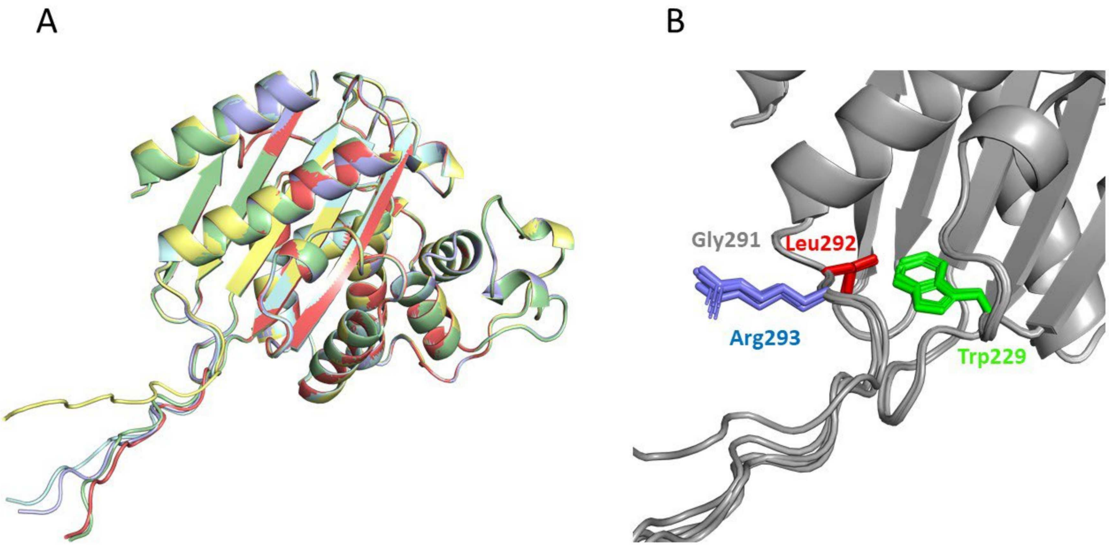


**Figure 5.** Prediction of the structure of the best Kemp eliminase obtained in this work by the program AlphaFold2 [39]. The prediction was carried out with a sequence that had a tryptophan at position 229, i.e., the function-generating W229D mutation was omitted. The reason for this was that the tryptophan at position 229 occupied a position close to that of the transition structure in the active Kemp eliminase (Figure 2A). Therefore, interactions with W229 in the predicted structure may conceivably correspond to interactions with the transition state that affects catalysis in the W229D variants. The top five predicted structures (**A**) are very close to the experimental structures shown in Figure 2, although the additional $GLRG_3SLEH_6$ polypeptide appears mostly extended and exposed to the solvent. However, a blow-up of the de novo active site region (**B**) reveals an interaction between the leucine of the polypeptide and the tryptophan at position 229 that could correspond to a kinetically relevant interaction with the transition state in the Kemp eliminase.

### *2.5. On the Possibility of Further Enhancements of Kemp Eliminase Activity*

The catalysis levels achieved in this work were certainly substantial. However, it seems reasonable to ask whether they could be further enhanced using the same general approach. As we screened less than 10% of the different amino acid sequences spanned by the combinatorial library, it is possible that better variants could be found by additional screening of the same library. More relevant is the fact that the library we used was sharply focused on three positions, which allowed us to screen a substantial fraction of the library in a reasonable time. Nevertheless, it is conceivable that mutations at farther positions in the inserted polypeptide can also increase the rate of Kemp elimination. To explore this possibility, we prepared a new 8000-variant combinatorial library focused at the fourth, fifth, and sixth positions after the carboxyl terminus (Figure 3C). Screening of approximately 250 variants from this library yielded results that were qualitatively similar to those obtained with the first library. Most of the variants displayed significant activity in the primary screening (Figure 3C), perhaps reflecting that the library samples the sequence space associated with a conformationally flexible polypeptide segment that likely remained largely exposed to the solvent in most cases. However, as was the case with the first library, a few of the variants were found to display substantially enhanced levels of catalysis for the Kemp elimination reaction (Figure 3C), a result that was confirmed by variant protein purification and by determination of the Michaelis–Menten profiles and catalytic parameters (Figure 4A and Table 1). Certainly, the catalysis enhancement obtained on the basis of a limited screening of this second library was smaller than that afforded by the GLR variant from the first library. However, it is clear that combined screening of the positions included in the libraries is likely to lead to additional increases in de novo catalysis (work in progress).

## 3. Discussion

The generation of de novo enzymes (i.e., enzymes with completely new active sites) is one of the major unsolved problems in protein engineering [32]. In addition to the obvious biotechnological implications, the results of de novo enzyme engineering studies may have immediate implications for our understanding of the origin of life. Most of the chemical reactions of life are extremely slow in the absence of enzyme catalysis. Several analyses support the view that diverse and specialized enzymes were already present in the last universal common ancestor [41,42]. It would seem reasonable to assume that efficient molecular mechanisms for the de novo emergence of enzymes and their subsequent optimization must exist. However, such efficient mechanisms are not apparent in the efforts of protein engineers to develop completely new enzyme functionalities. Only a limited number of de novo enzymes have been reported to date [32], and several of the most recent success stories in this field involve the recruitment of metals or metal-containing cofactors that already provide, by themselves, some starting level of catalysis. However, only approximately 30% of enzymes are metalloenzymes [43,44], and the mechanisms for the emergence of new enzymes that do not rely on metal recruitment remain poorly understood. Furthermore, most of the engineered de novo enzymes display very low activities; many rounds of laboratory-directed evolution, a highly time-consuming procedure, are often required to bring their catalysis to levels that are similar to those of modern natural enzymes [37,45–47].

Starting with the work of Tawfik, Baker and coworkers in 2008 [48], Kemp elimination has been extensively applied as a benchmark of de novo enzyme engineering. The reaction can occur through a base-catalyzed mechanism, as shown in Figure 1A, and it is in fact considered as a model for proton abstraction from carbon, a fundamental process in chemistry and biochemistry. It can also occur through a redox mechanism. Highly active de novo enzymes based on the recruitment of the heme cofactor for Kemp elimination catalysis have been recently reported [49,50]. Here, however, we are concerned with "traditional" Kemp elimination that is achieved through proton abstraction by a catalytic base (Figure 1A).

The best base-promoted Kemp eliminase to date (at least in terms of $k_{cat}$) was reported by Hilvert et al. in 2013 [36] and displays a catalytic efficiency of $k_{cat}/K_M = 2.3{\cdot}10^5\ M^{-1}{\cdot}s^{-1}$ and a turnover of $k_{cat} = 700 \pm 60\ s^{-1}$, values that compare favorably with those for the average modern natural enzyme ($k_{cat}/K_M$ about $10^5\ M^{-1}{\cdot}s^{-1}$ and a turnover of $k_{cat}$~$10\ s^{-1}$; [35]). Remarkably, the catalytic parameters for the best Kemp eliminase found in this work, $k_{cat}/K_M = (2.0 \pm 0.1){\cdot}10^5\ M^{-1}{\cdot}s^{-1}$ and $k_{cat} = 635 \pm 59\ s^{-1}$, match those for the best Kemp eliminase as previously reported in the literature. However, these two efficient de novo Kemp eliminases are the result of quite different design approaches and screening efforts. Hilvert's Kemp eliminase was the outcome of 17 rounds of directed evolution from a rationally designed de novo enzyme with a low but significant level of Kemp eliminase activity [36]. The starting background for this extensive screening effort was the result of a complex iterative procedure in which a failed (i.e., inactive) initial computational design was rescued on the basis of amino acid replacements that were suggested by analyses of molecular dynamics simulations and 3D crystallographic structures [51]. In contrast, the best Kemp eliminase reported here started with a minimalist design that targeted a conformationally flexible region in an ancestral β-lactamase scaffold [33]. Thus, a single mutation (W229D) generated a significant level of Kemp elimination activity that could be immediately enhanced by a second mutation (F290W) at the de novo active site. The Kemp elimination activity of this W229D/F2900W variant could be increased via computationally focused, ultra-low-throughput screening [34]. Specifically, screening of only 20 active-site variants at the top of the stability ranking predicted by the FuncLib approach [52] led to a substantial activity improvement. Finally, in this work we achieved further catalysis enhancement on the basis of the screening of only approximately 800 variants from a library that samples the sequence space of a short polypeptide segment engineered at the active site region.

Our success in arriving at an efficient base-promoted Kemp eliminase on the basis of a rather modest screening effort has immediate implications for the engineering of de novo enzymes. First, the catalysis-enhancing mutations identified in this work and in our previous work [33,34] do not increase the complexity of the catalytic machinery, which remains a simple proton abstraction by a catalytic base, as established by the function-generating mutation W229D. Rather, they appear to optimize interactions that stabilize the transition state for the reaction ([34,37] and this work). Overall, it emerges that large enhancements in de novo enzyme catalysis can be achieved through the fine-tuning of intramolecular interactions in the active site region. Second, this work demonstrated that a short polypeptide segment inserted near the new active site has the capability of generating such catalysis-enhancing interactions. This approach has several obvious advantages. The inserted polypeptide segment will, in principle, be flexible and exposed to the solvent to a substantial extent; thus, it will be unlikely to lead to disruptive interactions that compromise proper protein folding. In addition, being a short segment, the associated sequence space can be efficiently sampled on the basis of a moderate screening effort.

The proof of principle provided in this work takes advantage of the fact that the carboxyl terminus in β-lactamases is close to the location of the new active site; therefore, it can be easily used as the point of attachment of the new segment. This does not mean, however, that the extra-segment approach necessarily requires that the de novo active site is engineered near the carboxyl (or the amino) terminus of the protein. In fact, the new active site could be placed in any appropriate region of the protein (i.e., where design is successful), and then the carboxyl (or amino) terminus could be moved to a position close the new active site by engineering a protein variant with a suitable circular permutation [53].

## 4. Methods and Materials

### 4.1. Site-Saturation Mutagenic Libraries

Library preparation was performed using the QuikChange Lightning PCR method (Agilent # 210518). Three positions were simultaneously saturated with the mutagenic primers described in Supplementary Table S2 in order to generate a combinatorial library comprising 8000 different amino acid sequences. The recombinant plasmid pET24-GNCA4-12-5G_HT containing the gene of the background variant was used as template. The amplification reaction contained 5 μL of 10× QuikChange Lightning Buffer, 1 μL of dNTP mix, 1.5 μL of QuikSolution reagent, 1.25 μL of primers (10 μM each mix), template plasmid (50 ng), 1 μL of QuikChange Lightning Enzyme, and water, to a final volume of 50 μL. The conditions for the PCR were as follows: 1 cycle at 95 °C for 2 min, 18 cycles of denaturation at 95 °C for 20 s, annealing at 60 °C for 10 s, and extension at 68 °C for 3.5 min. The final extension step was carried out at 68 °C for 5 min. The PCR products were digested with DpnI. Two μL of the commercial DpnI solution were added to the PCR sample and the solution was incubated at 37 °C for 5 min. Afterwards, 2 μL of this solution were used to transform *E. coli* XL10-Gold Ultracompetent cells (45 μL) with a heat pulse. Subsequently, cells were suspended in 1 mL of SOC medium, incubated for 1 h at 37 °C, and plated on LB-agar containing 100 μg/mL kanamycin. To test the quality of the library, ten clones were randomly selected, their plasmids were extracted, and the gene of the β-lactamase variants was sequenced (Sanger).

### 4.2. Library Screening

*E. coli* BL21 (DE3) cells were transformed with the plasmid containing the mutant libraries or, as a control the variant used as background for library construction (Figure 3A), plated on LB-Kan agar and grown for 16 h at 37 °C. Individual colonies were picked and transferred into 44 mm deep well plates containing LB-Kan medium (0.2 mL) using a Pickolo colony picker with a Freedom EVO 200 robot from TECAN (Männedorf, Schweiz). Each plate contained an internal standard with the variant used as library background (column 7, rows A to H) and a negative control (column 1, row H). These master plates were incubated at 37 °C with shaking at 250 rpm. After 16 h, clones from this pre-culture were inoculated (using a cryo-replicator CR1000 from Enzyscreen, Haarlem, The Netherlands) into deep well plates with fresh LB-Kan (lysate plates). After 4 hours of incubation at 37 °C, 250 rpm, LB-Kan with IPTG was added. The plates were incubated at 25 °C with shaking at 250 rpm for 16 h. The plates were centrifuged at 3000 g, the medium was discarded, and the cell pellets were frozen at −80 °C. After ~2 h the frozen cell pellets were re-suspended in HEPES 100 mM, pH 7. After 60 min at 25 °C the lysates were centrifuged at 3000× *g* and the supernatant was used for the Kemp eliminase assay.

### 4.3. Kemp Eliminase Assay for Library Screening

With the help of a liquid handler station (Freedom EVO 200, TECAN, Männedorf, Schweiz), 100 µL of the supernatant from lysate plates were transferred to the reaction plates. The initial activities and residual activities values were determined by adding 100 µL of HEPES 100 mM buffer pH 7.0 containing 0.25 mM 5-nitrobenzisoxazole. Plates were stirred briefly and the absorption at 380 nm (extinction coefficient of 15,800 $M^{-1}$ $cm^{-1}$) was followed in kinetic mode in the plate reader (Tecan M200 Infinite Pro Microplate Reader, Männedorf, Schweiz). The values were normalized against the average value corresponding to the background variant used for library construction. The best variants according to this determination were subsequently prepared and tested on pure form, as described below.

### 4.4. Protein Expression and Purification

The various β-lactamase variants studied in this work were prepared and purified as described previously [33,34]. Briefly, genes cloned into a pET24-b vector with resistance to kanamycin were transformed into *E. coli* BL21 (DE3) cells. The proteins were prepared with a His-tag and purified by affinity chromatography. Stock solutions for activity determinations and physicochemical characterization were prepared by exhaustive dialysis against the desired buffer.

### 4.5. Determination of Profiles of Rate versus Substrate Concentration for Kemp Eliminases

Rates of Kemp elimination were determined by following product formation by measuring the absorbance at 380 nm. An extinction coefficient of 15,800 $M^{-1}$ $cm^{-1}$ was used to calculate rates from the initial linear changes in absorbance with time. Most measurements were performed at 25 °C in HEPES 10 mM NaCl 100 mM pH 7 and 1% acetonitrile, although determinations in the same buffer at pH 8.5 were also performed for the best Kemp eliminase found in this work. As the stock solution of the substrate is prepared in acetonitrile, a certain concentration of this cosolvent in the reaction mixture is unavoidable. Note, however, that the concentration of acetonitrile stated (1%) was the final concentration and it was the same for all the experiments. That is, variable amounts of acetonitrile were added to solutions with different substrate concentrations to ensure constancy of the final acetonitrile after adding different volumes of the substrate stock solution in acetonitrile. In this way, it was guaranteed that no artefactual distortion of the curvature of the Michaelis plots arose because of variable acetonitrile concentrations. The absorbance increase at 380 nm was linear during the measurement and was recorded during a 15s interval, ensuring constant initial velocity conditions. All activity measurements were corrected by a blank performed under the same conditions. However, the determined rates were always clearly above the blanks. Profiles of rate versus substrate concentration were used to calculate the values of the catalytic efficiency ($k_{cat}/K_M$), the turnover number ($k_{cat}$), and the Michaelis constant ($K_M$) by fitting the Michaelis–Menten equation to the experimental data, as we described previously [33,34]. It is to be noted that the experimental substrate concentration range was limited by substrate solubility and that, for an acetonitrile concentration in the reaction mixture of 1%, substrate concentrations above 1 mM are not possible. Increasing the amount of acetonitrile in the reaction mixture certainly increases substrate solubility and the maximum accessible substrate concentration, but it also increases the value of the Michaelis constant. With an experimental substrate concentration range of 0–1 mM and a Michaelis constant on the order of a few mM, only a small curvature can be observed in the experimental Michaelis plots, a fact that could be thought to compromise the reliable determination of $K_m$ and $k_{cat}$. It is important to note, however, that, for each of the Kemp eliminases characterized in detail in this work, we obtained and analyzed three different profiles of rate versus substrate concentration, starting with at least two different protein preparations. Similar curvatures and congruent values of $K_m$ and $k_{cat}$ were observed from the analysis of the three independent Michaelis profiles that were determined for each variant (Figures 2C,D, S1 and Supplementary Table S1).

### *4.6. Protein Stability Determinations*

Thermal stabilities of all the β-lactamase variants studied in this work were assessed by differential scanning calorimetry in HEPES 10 mM NaCl 100 mM pH 7 using a VP (Valerian Plotnikov) Capillary DSC (Microcal, Malvern), following protocols that were well established in our laboratory [7]. A typical calorimetric run involved several buffer–buffer baselines to ensure proper equilibration of the calorimeter, followed by runs with several protein variants with intervening buffer–buffer baselines. A single calorimetric transition was observed in all cases. We used the denaturation temperature, defined as the temperature corresponding to the maximum of the calorimetric transition, as a metric of protein stability.

### *4.7. Protein Structure Prediction*

The protein structure prediction of variants obtained by directed evolution was performed using the ColabFold notebook implemented in Google Colab [40], based on the AlphaFold2 algorithm [39]. Sequences were used as inputs, and structure prediction was performed with default parameters. PyMOL (PyMOL Molecular Graphics System, Version 2.4.1 Schrödinger, LLC, created by Warren Lyford Delano, Delano Scientific LLC, San Francisco, CA, USA) was used to visualize the predicted models and to inspect the roles of the different mutations.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/ijms23168934/s1.

**Author Contributions:** L.I.G.-R. and V.A.R. performed the library screening experiments, purified the selected β-lactamase variants, and determined their biophysical features. M.A. provided essential input regarding library design and the interpretation of library screening results. L.I.G.-R. carried out the structural analyses of Kemp eliminases using AlphaFold2. V.A.R. and J.M.S.-R. designed the research. J.M.S.-R. wrote the first draft of the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Human Frontier Science Program, grant number RGP0041/2017 (J.M.S.-R.), the Spanish Ministry of Science and Innovation/FEDER Funds, grant number RTI-2018-097142-B100 (J.M.S.-R.), and FEDER/Junta de Andalucia-Consejería de Economía y Conocimiento, grant number E.FQM.113.UGR18 (V.A.R.). Acquisition of the Freedom EVO 200 robot from TECAN used for library screening was made possible by grant EQC2019-006403-P to J.M.S.-R. from the Spanish Ministry of Science and Innovation/FEDER Funds.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All relevant data are included in the manuscript and in the Supplementary Materials.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the manuscript.

## References

1. Zuckerkandl, E.; Pauling, L. Molecular disease, evolution, and genic heterogeneity. In *Horizons in Biochemistry*; Kasha, A., Pullman, B., Eds.; Academic Press: New York, NY, USA, 1962; pp. 189–225.
2. Pauling, L.; Zuckerkandl, E. Chemical paleogenetics. Molecular "restoration studies" of extinct forms of life. *Acta Chem. Scan.* **1963**, *17*, S9–S16.
3. Benner, S.A.; Sassi, S.O.; Gaucher, E.A. Molecular Paleoscience: Systems biology from the past. In *Advances in Enzymology and Related Areas of Molecular Biology*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007; Volume 75, pp. 1–132.
4. Hochberg, G.K.A.; Thornton, J.W. Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biochem.* **2017**, *46*, 247–269. [CrossRef]
5. Gumulya, Y.; Gillam, E.M. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: The 'retro' approach to protein engineering. *Biochem. J.* **2017**, *474*, 1–19. [CrossRef]
6. Gaucher, E.A.; Govindarajan, S.; Ganesh, O.K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **2008**, *451*, 704–707. [CrossRef] [PubMed]
7. Risso, V.A.; Gavira, J.A.; Mejia-Carmona, D.F.; Gaucher, E.A.; Sanchez-Ruiz, J.M. Hypersability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.* **2013**, *135*, 2899–2902. [CrossRef]
8. Akanuma, S.; Nakajima, Y.; Kimura, M.; Nemoto, N.; Mase, T.; Miyazono, K.; Takonura, M.; Yamagishi, A. Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 11067–11072. [CrossRef]
9. Devamani, T.; Rauwerdink, A.M.; Lunzer, M.; Jones, B.J.; Mooney, J.L.; Tan, M.A.O.; Zhang, Z.-J.; Xu, J.-H.; Dean, A.M.; Kazalauslas, R.J. Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J. Am. Chem. Soc.* **2016**, *138*, 1046–1056. [CrossRef]
10. Siddiq, M.A.; Hochberg, G.K.; Thornton, J.W. Evolution of protein specificity: Insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **2017**, *47*, 113–122. [CrossRef]
11. Nguyen, V.; Wilson, C.; Hoemberger, M.; Stiller, J.B.; Agafonov, R.V.; Kutter, S.; English, J.; Theobald, D.L.; Kern, D. Evolutionary drivers of thermoadpatation in enzyme catalysis. *Science* **2017**, *355*, 289–294. [CrossRef]
12. Risso, V.A.; Sanchez-Ruiz, J.M.; Ozkan, S.B. Biotechnological and protein engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol.* **2018**, *51*, 106–115. [CrossRef]
13. Trudeau, D.L.; Tawfik, D.S. Protein engineers turned evolutionists—The quest for the optimal starting point. *Curr. Opin. Struct. Biol.* **2019**, *60*, 46–52. [CrossRef] [PubMed]
14. Spence, M.A.; Kaczmarski, J.A.; Saunders, J.W.; Jackson, C.J. Ancestral sequence resonctruction for protein engineers. *Curr. Opin. Struct. Biol.* **2021**, *69*, 131–141. [CrossRef] [PubMed]
15. Gamiz-Arco, G.; Gutierrez-Rus, L.; Risso, V.A.; Ibarra-Molero, B.; Hoshino, Y.; Petrovic, D.; Justicia, J.; Cuerva, J.M.; Romero-Rivera, A.; Seelig, B.; et al. Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase. *Nat. Commun.* **2021**, *12*, 380. [CrossRef]
16. Gamiz-Arco, G.; Risso, V.A.; Gaucher, E.A.; Gavira, J.A.; Naganathan, A.N.; Ibarra-Molero, B.; Sanchez-Ruiz, J.M. Combining ancestral reconstruction with folding-landscape simulations to engineer heterologous protein expression. *J. Mol. Biol.* **2021**, *433*, 167321. [CrossRef] [PubMed]
17. Jensen, R.A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **1976**, *30*, 409–425. [CrossRef] [PubMed]
18. Cox, V.E.; Gaucher, E.A. Engineering proteins by reconstructing evolutionary adaptive paths. *Methods Mol. Biol.* **2014**, *1179*, 353–363.
19. Zakas, P.M.; Brown, H.C.; Knight, K.; Meeks, S.L.; Spencer, H.T.; Gaucher, E.A.; Doering, C.B. Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat. Biotechnol.* **2017**, *35*, 35–37. [CrossRef]
20. Alcalde, M. When directed evolution met ancestral enzyme resurrection. *Microb. Biotechnol.* **2017**, *10*, 22–24. [CrossRef]
21. Delgado, A.; Arco, R.; Ibarra-Molero, B.; Sanchez-Ruiz, J.M. Using resurrected ancestral protein to engineer virus resistance. *Cell Rep.* **2017**, *19*, 1247–1256. [CrossRef]
22. Gomez-Fernandez, B.J.; Risso, V.A.; Rueda, A.; Sanchez-Ruiz, J.M.; Alcalde, M. Ancestral resurrection and directed evolution of fungal mesozoic laccases. *Appl. Environ. Microbiol.* **2020**, *86*, e00778-20. [CrossRef]
23. Khersonsky, O.; Tawfik, D.S. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **2010**, *79*, 471–505. [PubMed]
24. James, L.C.; Tawfik, D.S. Conformational diversity and protein evolution—A 60-year old hypothesis revisited. *Trends Biochem. Sci.* **2003**, *28*, 361–368. [CrossRef]
25. Copley, S.D. An evolutionary biochemist's perspective on promiscuity. *Trends Biochem. Sci.* **2015**, *40*, 72–78. [CrossRef]
26. Pabis, A.; Risso, V.A.; Sanchez-Ruiz, J.M.; Kamerlin, S.C.L. Cooperativity and flexibility in enzyme evolution. *Curr. Opin. Struct. Biol.* **2018**, *48*, 83–92. [CrossRef] [PubMed]
27. Petrovic, D.; Risso, V.A.; Kamerlin, S.C.L.; Sanchez-Ruiz, J.M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **2018**, *15*, 20180330. [CrossRef]
28. Bloom, J.D.; Labthavikul, S.T.; Otey, C.R.; Arnold, F.H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5869–5874. [CrossRef]
29. Jaenicke, R. Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 2962–2964. [CrossRef]

30. Zou, T.; Risso, V.A.; Gavira, J.A.; Sanchez-Ruiz, J.M.; Ozkan, B. Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. *Mol. Biol. Evol.* **2015**, *119*, 1323–1333. [CrossRef]
31. Modi, T.; Risso, V.A.; Martinez-Rodriguez, S.; Gavira, J.A.; Mebrat, M.D.; Van Horn, W.D.; Sanchez-Ruiz, J.M.; Ozkan, B. Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity. *Nat. Commun.* **2021**, *12*, 1852. [CrossRef]
32. Lovelock, S.L.; Crawshaw, R.; Basler, S.; Levy, C.; Baker, D.; Hilvert, D.; Green, A.P. The road to fully programmable protein catalysis. *Nature* **2022**, *606*, 49–58. [CrossRef]
33. Risso, V.A.; Martinez-Rodriguez, S.; Candel, A.M.; Krüger, D.M.; Pantoja-Uceda, D.; Ortega-Muñoz, M.; Santoyo-Gonzalez, F.; Gaucher, E.A.; Kamerlin, S.C.L.; Bruix, M.; et al. De novo active sites for resurrected Precambrian enzymes. *Nat. Commun.* **2017**, *8*, 16113. [CrossRef] [PubMed]
34. Risso, V.A.; Romero-Rivera, A.; Gutierrez-Rus, L.; Ortega-Muñoz, M.; Santoyo-Gonzalez, F.; Gavira, J.A.; Sanchez-Ruiz, J.M.; Kamerlin, S.C.L. Enhancing a de novo enzyme activity by computationally-focused ultra-low-throughput screening. *Chem. Sci.* **2020**, *11*, 6134–6148. [CrossRef] [PubMed]
35. Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D.S.; Milo, R. The moderately efficient enzyme: Evolutionary trends and physicochemical trends shaping enzyme parameters. *Biochemsitry* **2011**, *50*, 4402–4410. [CrossRef] [PubMed]
36. Blomberg, R.; Kries, H.; Pinkas, D.M.; Mittl, P.R.E.; Grütter, M.G.; Provett, H.K.; Mayo, S.L.; Hilvert, D. Precision is essential for efficient catalysis in an evolved enzyme. *Nature* **2013**, *503*, 418–421. [CrossRef] [PubMed]
37. Gardner, J.M.; Biler, M.; Risso, V.A.; Sanchez-Ruiz, J.M.; Kamerlin, S.C.L. Manipulating conformational dynamics to repurpose ancient proteins for modern catalytic functions. *ACS Cat.* **2020**, *10*, 4863–4870. [CrossRef]
38. Chen, X.; Zaro, J.; Shen, W.-C. Fusion protein linkers: Property, design and functionality. *Adv. Drug Deliv. Rev.* **2013**, *65*, 1357–1369. [CrossRef]
39. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
40. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making protein folding accessible to all. *Nat. Methods* **2022**, *19*, 679–682. [CrossRef]
41. Ouzonis, C.A.; Kunin, V.; Darzentas, N.; Goldovsky, L. A minimal estimate for the gene content of the last universal ancestor—Exobiology from a terrestrial perspective. *Res. Microbiol.* **2006**, *157*, 57–68. [CrossRef]
42. Weiss, M.C.; Sousa, F.L.; Mmjavac, N.; Neukirken, S.; Roettger, M.; Nelson-Sathi, S.; Martin, W.F. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **2016**, *1*, 16116. [CrossRef]
43. Holm, R.H.; Kennepohl, P.; Solomon, E.I. Structural and Functional Aspects of Metal Sites in Biology. *Chem. Rev.* **1996**, *96*, 2239–2314. [CrossRef] [PubMed]
44. Waldron, K.J.; Robinson, N.J. How do bacterial cells ensure that metalloproteins get the correct metal? *Nat. Rev. Microbiol.* **2009**, *7*, 25–35. [CrossRef] [PubMed]
45. Preiswerk, N.; Beck, T.; Schulz, J.D. Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 8013–8018. [CrossRef]
46. Obexer, R.; Godina, A.; Garrabou, X.; Mittl, P.R.E.; Baker, D.; Griffiths, A.D.; Hilvert, D. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **2017**, *9*, 50–56. [CrossRef]
47. Crawshaw, R.; Crossley, A.E.; Johannissen, L.; Burke, A.J.; Hay, S.; Levy, C.; Baker, D.; Lovelock, S.L.; Green, A.P. Engineering an efficient and enantioselective enzyme for the Morita-Baylis-Hillman reaction. *Nat. Chem.* **2022**, *14*, 313–320. [CrossRef] [PubMed]
48. Röthlisberger, D.; Khersonsky, O.; Wollacott, A.M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J.L.; Althoff, E.A.; Xangehllini, A.; Dym, O.; et al. Kemp elimination catalysis by computational enzyme design. *Nature* **2008**, *453*, 190–195. [CrossRef]
49. Li, A.; Wang, B.; Ilie, A.; Dubey, K.D.; Bange, G.; Korendovych, I.V.; Shaik, S.; Reetz, M.T. A redox-mediated Kemp eliminase. *Nat. Commun.* **2017**, *8*, 14876. [CrossRef]
50. Korendovych, I.; Bhattacharya, S.; Margheritis, E.; Takahashi, K.; Kulesha, A.; D'Souza, A.; Kim, I.; Tame, J.; Yoon, J.; Volkov, A.; et al. NMR-guided directed evolution. **2021**, *preprint from research square.* [CrossRef]
51. Privett, H.K.; Kiss, G.; Lee, T.M.; Blomberg, R.; Chica, R.A.; Thomas, L.M.; Hilvert, D.; Houk, K.N.; Mayo, S.L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 3790–3795. [CrossRef]
52. Khersonsky, O.; Lipsh, R.; Avizemer, Z.; Ashani, Y.; Goldsmith, M.; Leader, H.; Dym, O.; Rogotner, S.; Trudeaum, D.L.; Prikusky, J.; et al. Automated design of efficient and functionally diverse enzyme repertoires. *Mol. Cell* **2018**, *72*, 178–186.e5. [CrossRef]
53. Yu, Y.; Lutz, S. Circular permutation: A different way to engineer enzyme structure and function. *Trends Botechnol.* **2011**, *29*, 18–25. [CrossRef]

# SUPPORTING INFORMATION

## Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme

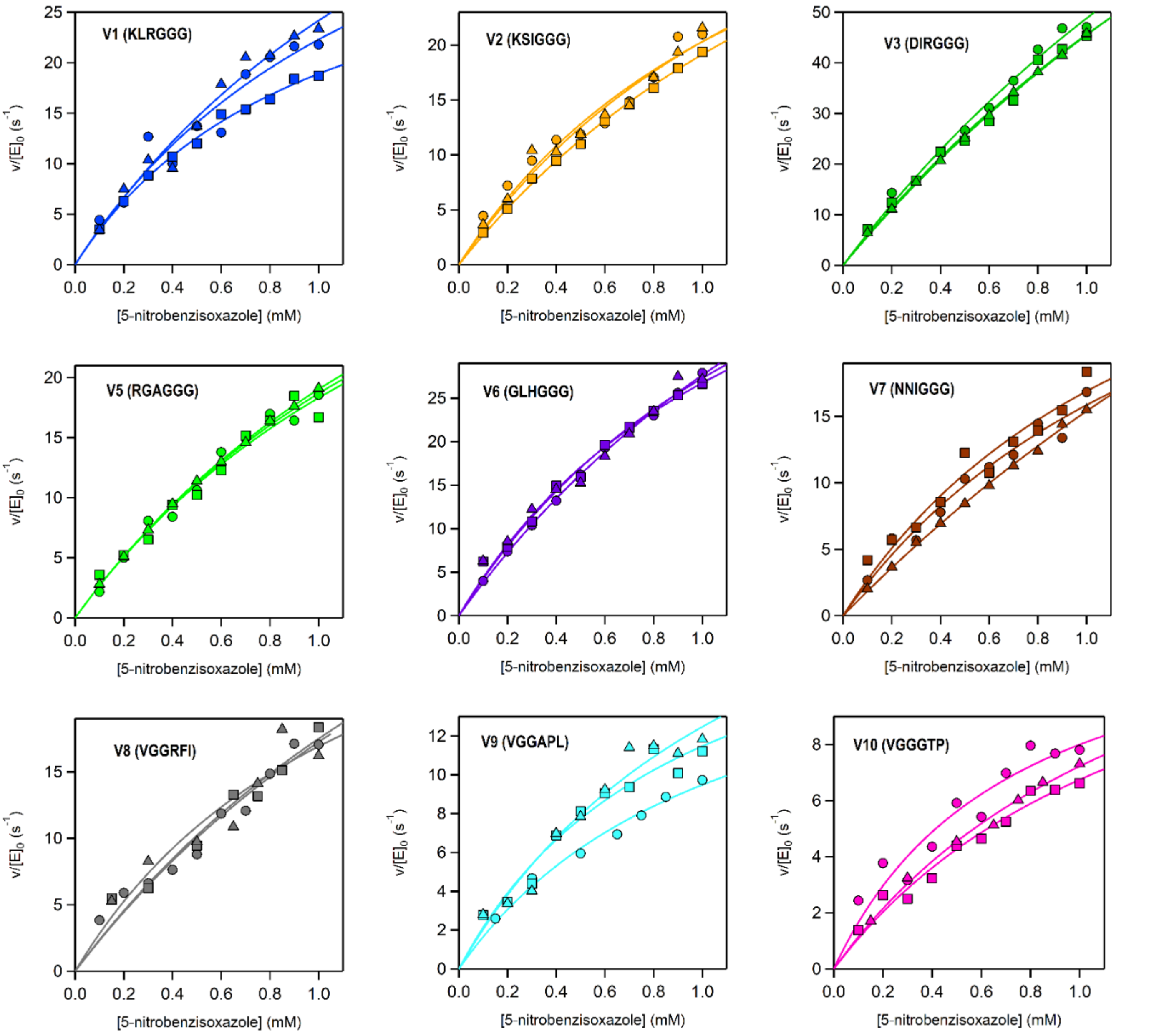


**Figure S1**. Michaelis-Menten profiles of the three independent determinations for the top variants from the primary library screening at pH 7. Note that the Michaelis-Menten profiles of the best variant (V4) are not included here (see Figure 2 in the main text). The sequences at the relevant section of the included polypeptide are shown for each variant.

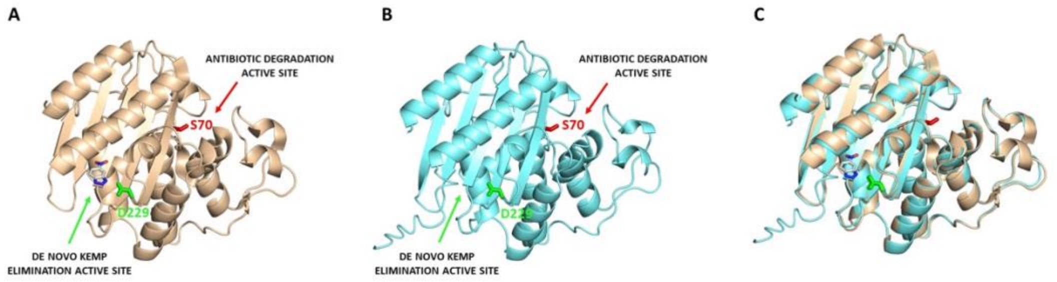


**Figure S2. (A)** Structure (PDB ID 5FQK) of the W229D/F290W variant of an ancestral $\beta$-lactamase with Kemp eliminase activity [33] showing location of the *de novo* Kemp elimination active site (identified by the bound transition-state analogue and the catalytic base D229) and the antibiotic ($\beta$-lactam) degradation active site (identified by the catalytic S70).**(B)** Same as in (A) but for the structure predicted by AlphaFold2 for the best Kemp eliminase found in this work. **(C)** Superposition of the experimental and predicted structures shown in (A) and (B).

**Table S1**. Catalytic parameters for the cleavage of 5-nitrobenzisoxazole at pH 7 (HEPES 10 mM NaCl 100 mM) and 1% acetonitrile and 25 ºC catalyzed by the engineered and evolved versions of Precambrian β-lactamases. Values of catalytic parameters derived from the fitting of the Michaelis-Menten equation are given for each of the three independent replicates. Errors are represented as the standard deviation derived from the fitting. The sequences for each variant are shown.

| Variant | Sequence | $k_{cat}$ ($s^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ ($s^{-1}$ $M^{-1}$) |
|---|---|---|---|---|
| V4<br>(pH 8.5) | GLRGGG | 703 ± 104 | 6.67 ± 0.65 | (1.9 ± 0.04) × $10^5$ |
| | | 606.4 ± 181 | 3.05 ± 1.14 | (2.0 ± 0.1) × $10^5$ |
| | | 596.9 ± 164 | 2.71 ± 0.85 | (2.2 ± 0.2) × $10^5$ |
| V4<br>(pH 7.0) | GLRGGG | 299.0 ± 99.8 | 2.33 ± 1.02 | (1.3 ± 0.1) × $10^5$ |
| | | 460.9 ± 53.3 | 4.54 ± 0.61 | (1.02 ± 0.02) × $10^5$ |
| | | 462.6 ± 143.0 | 3.19 ± 1.22 | (1.5 ± 0.1) × $10^5$ |
| V3 | DIRGGG | 206.8 ± 55.3 | 3.24 ± 1.07 | (6.4 ± 0.4) × $10^4$ |
| | | 190.4 ± 54.4 | 3.18 ± 1.12 | (6.0 ± 0.4) × $10^4$ |
| | | 211.4 ± 17.7 | 3.65 ± 0.37 | (5.8 ± 0.1) × $10^4$ |
| V6 | GLHGGG | 93.3 ± 6.5 | 2.38 ± 0.22 | (3.9 ± 0.1) × $10^4$ |
| | | 61.8 ± 7.4 | 1.31 ± 0.24 | (4.7 ± 0.3) × $10^4$ |
| | | 69.1 ± 14.9 | 1.53 ± 0.48 | (4.5 ± 0.4) × $10^4$ |
| V1 | KLRGGG | 53.8 ± 18.2 | 1.41 ± 0.72 | (3.8 ± 0.7) × $10^4$ |
| | | 38.0 ± 2.6 | 1.01 ± 0.12 | (3.8 ± 0.2) × $10^4$ |
| | | 70.4 ± 20.4 | 1.91 ± 0.76 | (3.7 ± 0.4) × $10^4$ |

| | | | | |
|---|---|---|---|---|
| V2 | KSIGGG | 49.1 ± 12.8 | 1.42 ± 0.55 | $(3.5 \pm 0.5) \times 10^4$ |
| | | 60.9 ± 5.2 | 2.18 ± 0.25 | $(2.8 \pm 0.1) \times 10^4$ |
| | | 54.1 ± 14.3 | 1.66 ± 0.63 | $(3.3 \pm 0.4) \times 10^4$ |
| V5 | RGAGGG | 55.0 ± 12.5 | 1.95 ± 0.61 | $(2.8 \pm 0.2) \times 10^4$ |
| | | 52.3 ± 15.2 | 1.86 ± 0.75 | $(2.8 \pm 0.3) \times 10^4$ |
| | | 59.0 ± 1.9 | 2.10 ± 0.09 | $(2.81 \pm 0.03) \times 10^4$ |
| V8 | VGGRFI | 61.8 ± 27.0 | 2.58 ± 1.45 | $(2.4 \pm 0.3) \times 10^4$ |
| | | 60.8 ± 34.0 | 2.47 ± 1.81 | $(2.4 \pm 0.4) \times 10^4$ |
| | | 38.0 ± 18.2 | 1.24 ± 0.95 | $(3.1 \pm 0.9) \times 10^4$ |
| V7 | NNIGGG | 41.0 ± 10.6 | 1.59 ± 0.60 | $(2.6 \pm 0.3) \times 10^4$ |
| | | 40.4 ± 11.2 | 1.39 ± 0.58 | $(2.9 \pm 0.4) \times 10^4$ |
| | | 84.8 ± 14.1 | 4.51 ± 0.88 | $(1.9 \pm 0.1) \times 10^4$ |
| V9 | VGGAPL | 20.0 ± 3.0 | 1.11 ± 0.27 | $(1.8 \pm 0.2) \times 10^4$ |
| | | 22.2 ± 3.9 | 0.94 ± 0.28 | $(2.4 \pm 0.3) \times 10^4$ |
| | | 28.9 ± 7.6 | 1.31 ± 0.53 | $(2.2 \pm 0.3) \times 10^4$ |
| V10 | VGGGTP | 17.6 ± 2.4 | 1.43 ± 0.30 | $(1.2 \pm 0.1) \times 10^4$ |
| | | 13.9 ± 2.8 | 0.74 ± 0.28 | $(1.9 \pm 0.3) \times 10^4$ |
| | | 16.1 ± 3.5 | 1.39 ± 0.45 | $(1.2 \pm 0.1) \times 10^4$ |
| BACKGROUND | VGGGGG | 21.9 ± 4.0 | 2.35 ± 0.56 | $(9.4 \pm 0.6) \times 10^3$ |
| | | 11.3 ± 3.8 | 1.21 ± 0.65 | $(9.3 \pm 1.9) \times 10^3$ |
| | | 15.7 ± 2.3 | 1.56 ± 0.33 | $(1.0 \pm 0.1) \times 10^4$ |

**Table S2**. Sequences for the mutagenic primers used to saturate three simultaneous positions of the included polypeptide for the generation of the combinatorial libraries.

| COMBINATORIAL LIBRARY | PRIMER NAME | PRIMER SEQUENCE |
|---|---|---|
| Library 1 (VGGGGG) | VHH_F | gtggtggtggtggtggtgctcgagtgaNNNNNNNNNaccacccacccacgccgccacaaccagacgc |
|  | VHH_R | gcgtctggttgtggcggcgtgggtgggtggtNNNNNNNNNtcactcgagcaccaccaccaccaccac |
| Library 2 (VGGGGG) | HHH_F | gtggtggtggtggtggtgctcgagtgagccaccaccNNNNNNNNNccacgccgccacaaccagacgc |
|  | HHH_R | gcgtctggttgtggcggcgtggNNNNNNNNNggtggtggctcactcgagcaccaccaccaccaccac |

# Chapter 2 – Characterization of resurrected ancestral TIM-barrels as scaffolds to engineer and understand the emergence of *de novo* catalysis

## Introduction and background

In chapter 2 we present the results obtained from experiments performed in order to characterize resurrected ancestral TIM-barrel proteins from an engineering and evolutionary standpoint. This project was conceived upon the results obtained from the experiments performed with resurrected ancestral β-lactamases, which showed that resurrected ancestral proteins may serve as superior scaffolds for protein engineering and evolutionary studies. In this case, we decided to reconstruct and resurrect ancestral proteins displaying a different fold than β-lactamases, but with the same goals in terms of engineering *de novo* enzymatic activities and understanding the origins of enzymatic catalysis. In particular, we decided to resurrect ancestral proteins that display a triose phosphate isomerase (TIM) barrel fold, a structurally repetitive protein architecture that displays biochemical and biophysical features of high interest in terms of engineering and evolutionary implications[223–225]. The TIM-barrel fold is a structurally repetitive protein architecture that resembles a "barrel" (Figure 10A) which consists of eight β-sheets connected to eight α-helices by flexible βα-loops and αβ-loops (Figure 10B). The β-sheets are arranged to form a barrel structure that sits in the core of the protein structure. Around the β-sheet barrel, the α-helices are organized surrounding the core of the structure and exposed to the solvent. The βα-loops that connect the β-sheets and α-helices are located on the "top" of the protein, displaying high conformational flexibility and accommodating the catalytic machinery of the active site (Figure 10A). The TIM barrel fold is ubiquitous in every modern proteome that, with a few exceptions, all its representative proteins act as enzymes[225]. The TIM-barrel fold represents about the 10% of all enzymes, harboring at least five of the six major enzymatic categories[226]. However, common ancestry between different TIM-barrel families cannot be demonstrated. Therefore, the TIM-barrel architecture can be considered as a "superfold"[226,227], as different proteins sharing this fold do not necessarily display an evolutionary relationship.

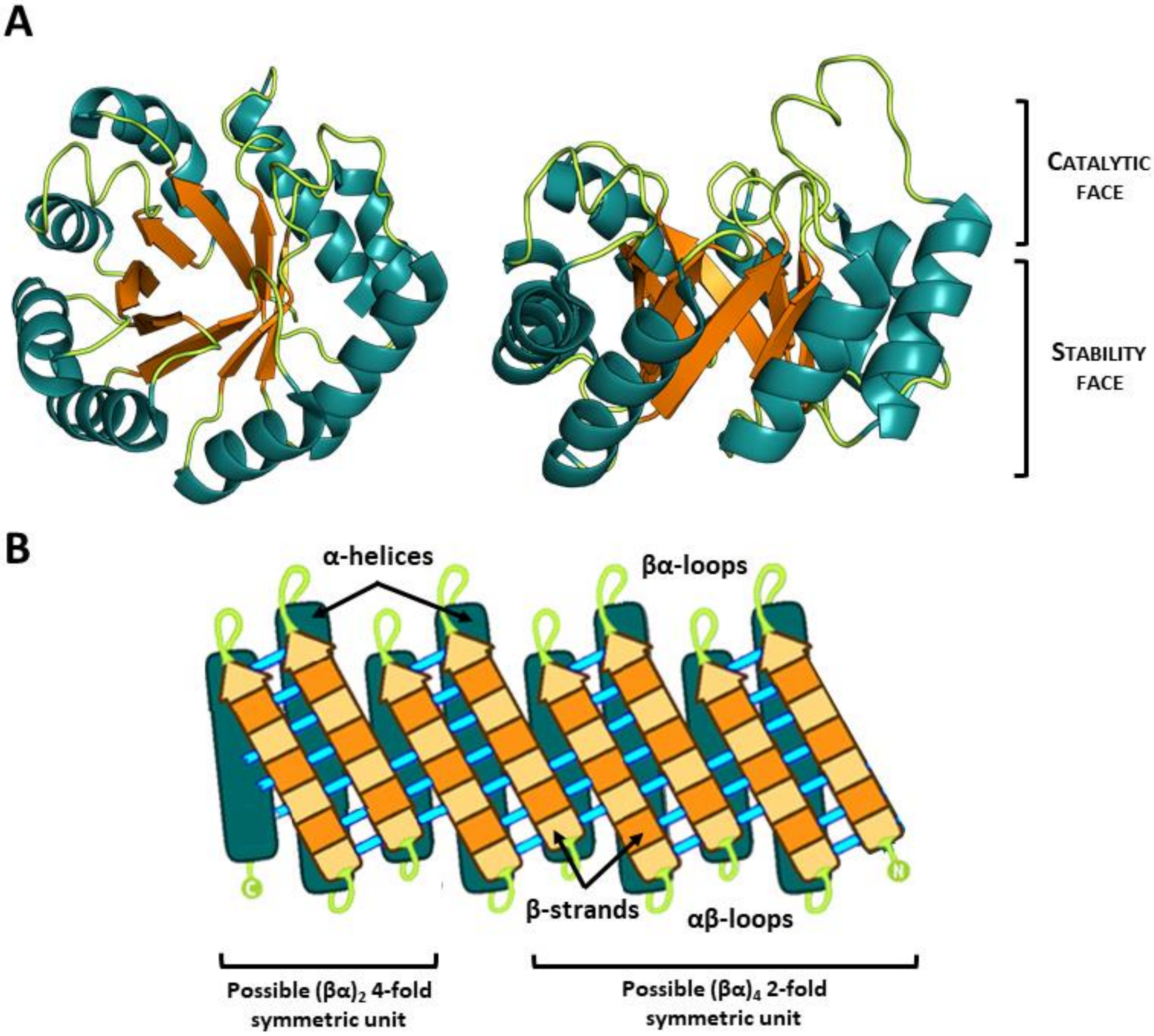

A
CATALYTIC FACE
STABILITY FACE
B
α-helices
βα-loops
β-strands
αβ-loops
C
N
Possible $(\beta\alpha)_2$ 4-fold symmetric unit
Possible $(\beta\alpha)_4$ 2-fold symmetric unit

**Figure 10**. (A) Backbone model of a typical TIM-barrel fold displaying the canonical barrel structure with eight β-sheets connected to eight α-helices by flexible βα-loops and αβ-loops (PDB: 1kv8). Helices are colored in teal, beta sheets are colored in orange, and loops are colored green. Sheets are colored in two shades of orange. Lighter shades indicate residues pointing inward, towards the barrel pore. Darker shades indicate residues pointing outward, towards the barrel core. Cyan lines represent the hydrogen bonding network between sheets in the core barrel. The catalytic face encompasses the "top" part of the protein with the βα-loops, that display high flexibility and harbors the catalytic machineries of the TIM-barrel enzymes. The stability face encompasses the rest of the protein architecture, the core barrel of β-strands and the surrounding solvent exposed α-helices. (B) Schematic representation of the canonical TIM-barrel secondary structure topology. The number of βα units in the TIM barrel architecture is always 8. TIM barrels naturally adopt a two or four-fold symmetry. Adapted from Nagarajan D, *et al*. BMC Biochemistry 16, 18 (2015).

The wide abundance of TIM-barrel enzymes reveals that nature has somehow “chosen” the TIM-barrel architecture as a default scaffold for the generation of many different catalytic machineries to catalyze unrelated biochemical reactions. This is possible because of the inherent high evolvability of the TIM-barrel fold, i.e., the ability to evolve many different functions in the same structural scaffold. The main contributor of the high evolvability in the TIM-barrel fold is the structural modularity of its three-dimensional structure, in which we can differentiate between a “stability face” that encompasses the β-sheet barrel and the surrounding α-helices, and a “catalytic face” that encompasses the top connecting βα-loops (Figure 10A). This modularity leads to a “division of labor”, in which the β-sheets and α-helices constitute a stable and rigid “core” of the protein that confers stability and robustness to the overall fold, while the βα-loops constitute a flexible region exposed to the solvent that contains the catalytic residues involved in the many different biochemical reactions. Evolutionary speaking, the structural modularity and division of labor may help to explain from a molecular standpoint how the TIM-barrel fold can explore a wide sequence space in the flexible catalytic loops leading to many different catalytic machineries and conformational states that facilitate the diversity of enzymatic functions, while maintaining the overall stability of the scaffold[224]. Overall, TIM-barrels contain the three major contributors of functional evolvability: structural modularity[228], conformational flexibility/diversity[199,200,229,230], and mutational robustness conferred by stability[121,231].

The repetitiveness of the TIM-barrel structure supports the idea that this fold could have evolved from previously existing smaller (β/α) proteins by following common mechanisms of protein innovation such as recombination or duplication[232–236]. In particular, evidence of a structural evolutionary link between the TIM-barrel fold and the ancient flavodoxin-like fold was demonstrated, contributing to the idea that protein superfolds share common ancestry in the context of the Dayhoff’s hypothesis[237,238], with the TIM-barrel fold as a canonical example. It is, therefore, conceivable to think that TIM-barrel proteins emerged from smaller proteins during the early evolution of the protein repertoire, playing a key role in the emergence of proteomes and life. It is also relevant to mention that part of the wide diversity of enzymatic functions carried out by TIM-barrel enzymes can be explained, in part, by the incorporation of a broad range of different metal and organic cofactors in the protein structure, including some putatively ancient cofactors[225,226,239]. The association between the highly evolvable TIM-barrel fold and the catalytic diversity provided by the cofactors is a trend that does not appear to be shared by other proteins in general. This supports the idea that TIM-barrel structures emerged during the early protein evolution steps and provided an ideal scaffold to promote the transition between prebiotic chemistry to proto biochemistry.

Therefore, as the TIM-barrel fold displays biochemical and biophysical features that contribute to a high mutational robustness and evolvability, it may represent an ideal protein scaffold for the engineering of *de novo* enzymatic activities. As the residues involved in enzymatic catalysis of TIM-barrels are located in the flexible loops of the catalytic face, different engineering strategies (i.e., rational design or directed evolution) can be easily applied in TIM-barrel proteins targeting the loops in order to generate a new catalytic active site, but without compromising the overall structure stability and folding. Additionally, the simple and repetitive structure of the TIM-barrel, its functional diversity in terms of enzymatic activities, and its ubiquity in modern proteomes and metabolic pathways directly points to the hypothesis that the TIM-barrel is most likely an extremely ancient architecture[240,241] that played a central role in the early evolution of protein-mediated catalysis and metabolic reactions[226,242,243]. These transitions were most likely facilitated by the incorporation of abiotic and organic catalysts in the TIM-barrel structures that later evolved into cofactor dependent TIM-barrel enzymes. As a result, the TIM-barrel also represents an ideal scaffold to study the natural evolutionary processes that led to the emergence of novel enzymatic reactions and catalytic machineries in protein scaffolds.

Given this background, we decided to combine the exceptional natural properties of TIM-barrels related with their robustness and functional diversity together with the common properties observed in resurrected ancestral proteins in terms of higher evolvability. This appears as an excellent strategy to generate novel protein scaffolds based in resurrected ancestral TIM-barrels as ideal scaffolds for (I) the generation and engineering of novel enzymatic activities and catalytic machineries and (II) for the study of evolutionary mechanisms and driving forces that underlie the emergence of enzymatic catalysis. This thesis involves the resurrections and characterization of several ancestral TIM-barrel proteins from different glycoside hydrolase families. The aim was to identify a suitable candidate for further engineering and evolutionary studies. One specific ancestral protein named N72, corresponding to a putative ancestor of bacterial and eukaryotic family-1 glycosidases, displayed functional features that make it an excellent scaffold for the purposes and objectives of the thesis. The findings of his research are presented in two separate research papers, both of which are included in the following sections of this chapter[156,244]. The first paper titled "Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase" presents the exceptional properties of the resurrected ancestral TIM-barrel glycosidase. In particular, a high thermostability is combined with a large conformational flexibility of the protein scaffold in comparison with its modern counterparts. However, the most relevant discovery is the fact that the ancestral TIM-barrel is able to bind a molecule of the redox cofactor heme which has potential implications in engineering and evolution. The second paper titled "Protection of catalytic cofactors by polypeptides as a driver for the emergence of primordial enzymes" presents that the heme binding ancestral TIM-barrel displays considerable levels of peroxidase activity exclusively catalyzed by the heme cofactor in absence of a peroxidase catalytic machinery. This study reveals how cofactor protection against different inactivation processes promotes the intrinsic catalytic power of the cofactor and leads to a higher number of turnovers during catalysis. This observation has important evolutionary implications, as it reveals how cofactor protection could serve as an evolutionary driving force that facilitated primordial polypeptide-cofactor associations during the emergence of catalysis in the origin of life.

# PUBLICATION III

## Heme-Binding Enables Allosteric Modulation in an Ancient TIM-Barrel Glycosidase

nature
communications

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20630-1 OPEN

# Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase

Gloria Gamiz-Arco[1,8], Luis I. Gutierrez-Rus[1,8], Valeria A. Risso[1], Beatriz Ibarra-Molero[1], Yosuke Hoshino[2], Dušan Petrović[3,7], Jose Justicia[4], Juan Manuel Cuerva[4], Adrian Romero-Rivera[3], Burckhard Seelig[5], Jose A. Gavira[6], Shina C. L. Kamerlin[3✉], Eric A. Gaucher[2✉] & Jose M. Sanchez-Ruiz[1✉]

Glycosidases are phylogenetically widely distributed enzymes that are crucial for the cleavage of glycosidic bonds. Here, we present the exceptional properties of a putative ancestor of bacterial and eukaryotic family-1 glycosidases. The ancestral protein shares the TIM-barrel fold with its modern descendants but displays large regions with greatly enhanced conformational flexibility. Yet, the barrel core remains comparatively rigid and the ancestral glycosidase activity is stable, with an optimum temperature within the experimental range for thermophilic family-1 glycosidases. None of the ~5500 reported crystallographic structures of ~1400 modern glycosidases show a bound porphyrin. Remarkably, the ancestral glycosidase binds heme tightly and stoichiometrically at a well-defined buried site. Heme binding rigidifies this TIM-barrel and allosterically enhances catalysis. Our work demonstrates the capability of ancestral protein reconstructions to reveal valuable but unexpected biomolecular features when sampling distant sequence space. The potential of the ancestral glycosidase as a scaffold for custom catalysis and biosensor engineering is discussed.

[1] Departamento de Quimica Fisica. Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain. [2] Department of Biology, Georgia State University, Atlanta, GA 30303, USA. [3] Science for Life Laboratory, Department of Chemistry-BMC, Uppsala University, BMC Box 576, S-751 23 Uppsala, Sweden. [4] Departamento de Quimica Organica. Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain. [5] Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, Minnesota, United States of America, & BioTechnology Institute, University of Minnesota, St. Paul, MN, USA. [6] Laboratorio de Estudios Cristalograficos, Instituto Andaluz de Ciencias de la Tierra, CSIC, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, Avenida de las Palmeras 4, Granada 18100 Armilla, Spain. [7] Present address: Hit Discovery, Discovery Sciences, Biopharmaceutical R&D, AstraZeneca 431 50 Gothenburg, Sweden. [8] These authors contributed equally: Gloria Gamiz-Arco, Luis I. Gutierrez-Rus. ✉email: lynn.kamerlin@kemi.uu.se; egaucher@gsu.edu; sanchezr@ugr.es

Pauling and Zuckerkandl proposed in 1963 that the sequences of modern protein homologs could be used to reconstruct the sequences of their ancestors[1]. While this was mostly only a theoretical possibility in the mid-twentieth century, ancestral sequence reconstruction has become a standard procedure in the twenty-first century, due to advances in bioinformatics and phylogenetics, together with the availability of increasingly large sequence databases. Indeed, in the last ~20 years, proteins encoded by reconstructed ancestral sequences ("resurrected" ancestral proteins, in the common jargon of the field) have been extensively used as tools to address important problems in molecular evolution[2,3]. In addition, a new and important implication of sequence reconstruction is currently emerging linked to the realization that resurrected ancestral proteins may display properties that are desirable in scaffolds for enzyme engineering[4–6]. For instance, high stability and substrate/catalytic promiscuity have been described in a number of ancestral resurrection studies[5,7]. These two features are known contributors to protein evolvability[8,9], which points to the potential of resurrected ancestral proteins as scaffolds for the engineering of new functionalities[4,10].

More generally, reconstruction studies that target ancient phylogenetic nodes typically predict extensive sequence differences with respect to their modern proteins. Consequently, proteins encoded by the reconstructed sequences may potentially display altered or unusual properties. Regardless of the possible evolutionary implications, it is of interest, therefore, to investigate which properties of putative ancestral proteins may differ from those of their modern counterparts and to explore whether and how these ancestral properties may lead to new possibilities in biotechnological applications. Here, we apply ancestral sequence reconstruction to a family of well known and extensively characterized enzymes. Furthermore, these enzymes display 3D-structures based on the highly common and widely studied TIM-barrel fold, a fold which is both ubiquitous and highly evolvable[11–13]. Yet, we find upon ancestral resurrection a diversity of unusual and unexpected biomolecular properties that suggest new engineering possibilities that go beyond the typical applications of protein family being characterized.

Glycosidases catalyze the hydrolysis of glycosidic bonds in a wide diversity of molecules[14]. The process typically follows a Koshland mechanism based on two catalytic carboxylic acid residues and, with very few exceptions, does not involve cofactors. Glycosidic bonds are very stable and have an extremely low rate of spontaneous hydrolysis[15]. Glycosidases accelerate their hydrolysis up to ~17 orders of magnitude, being some of the most proficient enzymes functionally characterized[16]. Glycosidases are phylogenetically widely distributed enzymes. It has been estimated, for instance, that about 3% of the human genome encodes glycosidases[17]. They have been extensively studied, partly because of their many biotechnological applications[14]. Detailed information about glycosidases is collected in the public CAZy database (Carbohydrate-Active enZYmes Database; http://www.cazy.org)[18] and the connected CAZypedia resource (http://www.cazypedia.org/)[19]. At the time of our study, glycosidases are classified into 167 families on the basis of sequence similarity. Since perturbations of protein structure during evolution typically occur more slowly than sequences change[20], it is not surprising that the overall protein fold is conserved within each family. Forty eight of the currently described glycosidase families display a fold consistent with the TIM barrel architecture. Often, common ancestry between different TIM-barrel families cannot be unambiguously demonstrated[12]. Therefore, the TIM-barrel may be considered as a "superfold" in the sense of Orengo et al.[21], and simply sharing this fold does not necessarily imply evolutionary relatedness.

Here, we study family 1 glycosidases, which are of the classical TIM-barrel fold. Family 1 glycosidases (GH1) commonly function as β-glucosidases and β-galactosidases, although other activities are also found in the family[22]. GH1 enzymes are present in the three domains of life and have been traced back to LUCA[23]. We focus on a putative ancestor of modern bacterial and eukaryotic enzymes and find a number of unusual properties that clearly differentiate the ancestor from the properties of its modern descendants. The ancestral glycosidase thus displays much-enhanced conformational flexibility in large regions of its structure. This flexibility, however, does not compromise stability as shown by the ancestral optimum activity temperature which is within the typical range for family 1 glycosidases from thermophilic organisms. Unexpectedly, the ancestral glycosidase binds heme tightly at a well-defined site in the structure with concomitant allosteric increase in enzyme activity. Neither metalloporphyrin binding nor allosteric modulation appears to have been reported for any modern glycosidases, despite the fact that these enzymes have been extensively characterized. Overall, this work demonstrates the potential of ancestral reconstruction as a tool to explore sequence space to generate combinations of properties that are unusual or unexpected compared to the repertoire from modern proteins.

## Results

**Ancestral sequence reconstruction**. Ancestral sequence reconstruction (ASR) was performed based on a phylogenetic analysis of family 1 glycosidase (GH1) protein sequences (see the Methods for details). GH1 protein homologs are widely distributed in all three domains of life and representative sequences were collected from each domain, including characterized GH1 sequences obtained from CAZy as well as homologous sequences contained in GenBank. The phylogeny of GH1 homologs consists of four major clades (Fig. 1a and Fig. S1). One clade is composed mainly of archaea and bacteria from the recently proposed Candidate Phyla Radiation (CPR)[24], while the other three clades include bacteria and eukaryotes. The archaeal/CPR clade largely contains uncharacterized proteins and was thus excluded from further analysis. In the bacterial/eukaryotic clades, eukaryotic homologs form a monophyletic clade within bacterial homologs. For our current study, the common ancestors of bacterial and eukaryotic homologs are selected for ASR analysis (N72, N73, and N125) because many homologs have been characterized and there is substrate diversity between the enzymes in the different clades.

**Selection of an ancestral glycosidase for experimental characterization**. We prepared, synthesized, and purified the proteins encoded by the most probabilistic sequences at nodes N72, N73, and N125. While the three proteins were active and stable, we found that those corresponding to N73 and N125 had a tendency to aggregate over time. We therefore selected the resurrected protein from node 72 for exhaustive biochemical and biophysical characterization. For the sake of simplicity, we will subsequently refer to this protein as the ancestral glycosidase in the current study.

It is important to note that the sequence of the ancestral glycosidase differs considerably from the sequences of modern proteins. The set of modern sequences used as a basis for ancestral reconstruction span a range of sequence identity (26–59%) with the ancestral glycosidase (Table S1). Also, using the ancestral sequence as the query of a BLAST (Basic Local Alignment Search Tool) search in several databases (non-redundant protein sequences, UniProtKB/Swiss-Prot, Protein Data Bank, Metagenomic proteins) yields a closest hit with only a 62% sequence identity to the ancestral glycosidase. These sequence differences translate into unexpected biomolecular properties.

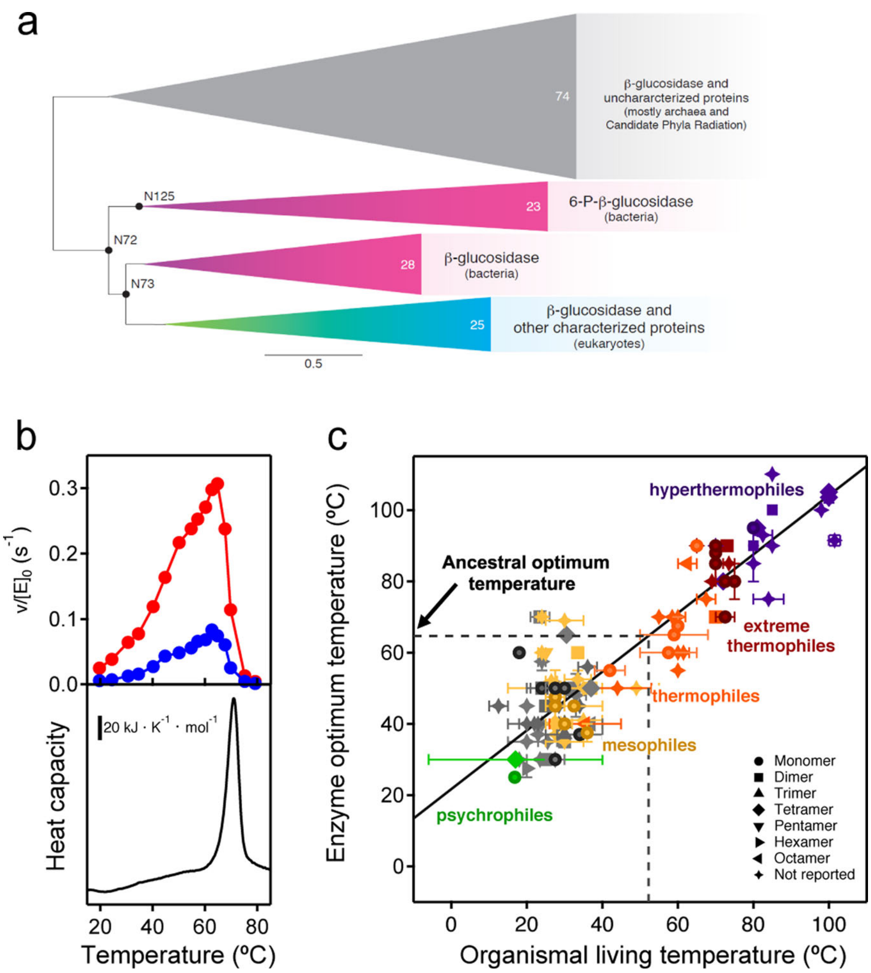


**Fig. 1 Ancestral sequence reconstruction of family 1 glycosidases (GH1) and assessment of ancestral stability. a** Bayesian phylogenetic tree of GH1 protein sequences using 150 representative sequences. Triangles correspond to four major well-supported clades (see supplemental Fig. S1 for nodal support) with common functions indicated. Numbers inside the triangles correspond to the number of sequences in each clade. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. Reconstructed ancestral sequences were inferred at the labeled nodes and the protein at node 72 was exhaustively characterized. **b** Determination of the optimum temperature for the ancestral glycosidase (upper panel) using two different substrates 4-nitrophenyl-β-D-glucopyranoside (red) and 4-nitrophenyl-β-D-galactopyranoside (blue). $v/[E]_0$ stands for the rate over the total enzyme concentration. The lower panel shows a differential scanning calorimetry profile for the ancestral glycosidase. Clearly, the activity drop observed at high temperature (upper panel) corresponds to the denaturation of the protein, as seen in the lower panel. **c** Plot of enzyme optimum temperature versus living temperature of the host organism for modern family 1 glycosidases. Data (Supplementary Dataset 1) are derived from literature searches, as described in Methods. Horizontal and vertical bars are not error bars, but represent ranges of organismal living temperatures and enzyme optimum temperatures when provided in the literature. Color code denotes the organisms that published literature describes as hyperthermophiles, extreme thermophiles, thermophiles, mesophiles, psychrophiles; gray color is used for organisms that have not been thus classified (plants that live at moderate temperatures in most cases). The line is a linear-squares fit ($T_{OPT} = 21.68 + 0.824T_{LIVING}$). Correlation coefficient is 0.89 and $p \sim 8.8 \times 10^{-45}$ (probability that the correlation results from chance). An environmental temperature of about 52 °C can be estimated from the optimum temperature of the ancestral glycosidase.

**Stability**. As it is customary in the glycosidase field, we assessed the stability of the ancestral glycosidase using profiles of activity versus temperature determined by incubation assays[25]. These profiles typically reveal a well-defined optimum activity temperature (Fig. 1b) as a result of the concurrence of two effects. At low temperatures, the expected Arrhenius-like increase of activity with temperature is observed. At high temperatures, protein denaturation occurs and causes a sharp decrease in activity. For the ancestral glycosidase, this interpretation is supported by differential scanning calorimetry data (lower panel in Fig. 1b) which show a denaturation transition that spans the temperature range in which the activity drops sharply.

The profiles of activity versus temperature (Fig. 1b) show a sharp maximum at 65 °C for the optimum activity temperature of the ancestral glycosidase. In order to ascertain the implications of this value for the evaluation of the ancestral stability, we have searched the literature on family 1 glycosidases for reported optimum temperature values (see Methods for details and Supplementary Dataset 1 for the results of the search). The values for ~130 modern enzymes show a good correlation with the environmental temperature of their respective host organisms (Fig. 1c). Therefore, the enzyme optimum temperature is an appropriate reflection of stability in an environmental context for this protein family. The optimum temperature value for ancestral glycosidase is within the experimental range of optimum activity temperatures for family 1 glycosidases from thermophilic organisms and it is consistent with an ancestral environmental temperature of about 52 °C (Fig. 1c).

**Conformational flexibility**. Remarkably, despite its "thermophilic" stability, large regions in the structure of the ancestral glycosidase are flexible and/or unstructured, as demonstrated by both experiment and computation (Fig. 2).

Proteolysis is known to provide a suitable probe of conformational diversity and the protein energy landscape[26], since most cleavable sites are not exposed in folded compact protein states. The ancestral glycosidase is highly susceptible to proteolysis and degradation is already apparent after only a few minutes incubation at a low concentration of thermolysin (0.01 mg/mL,

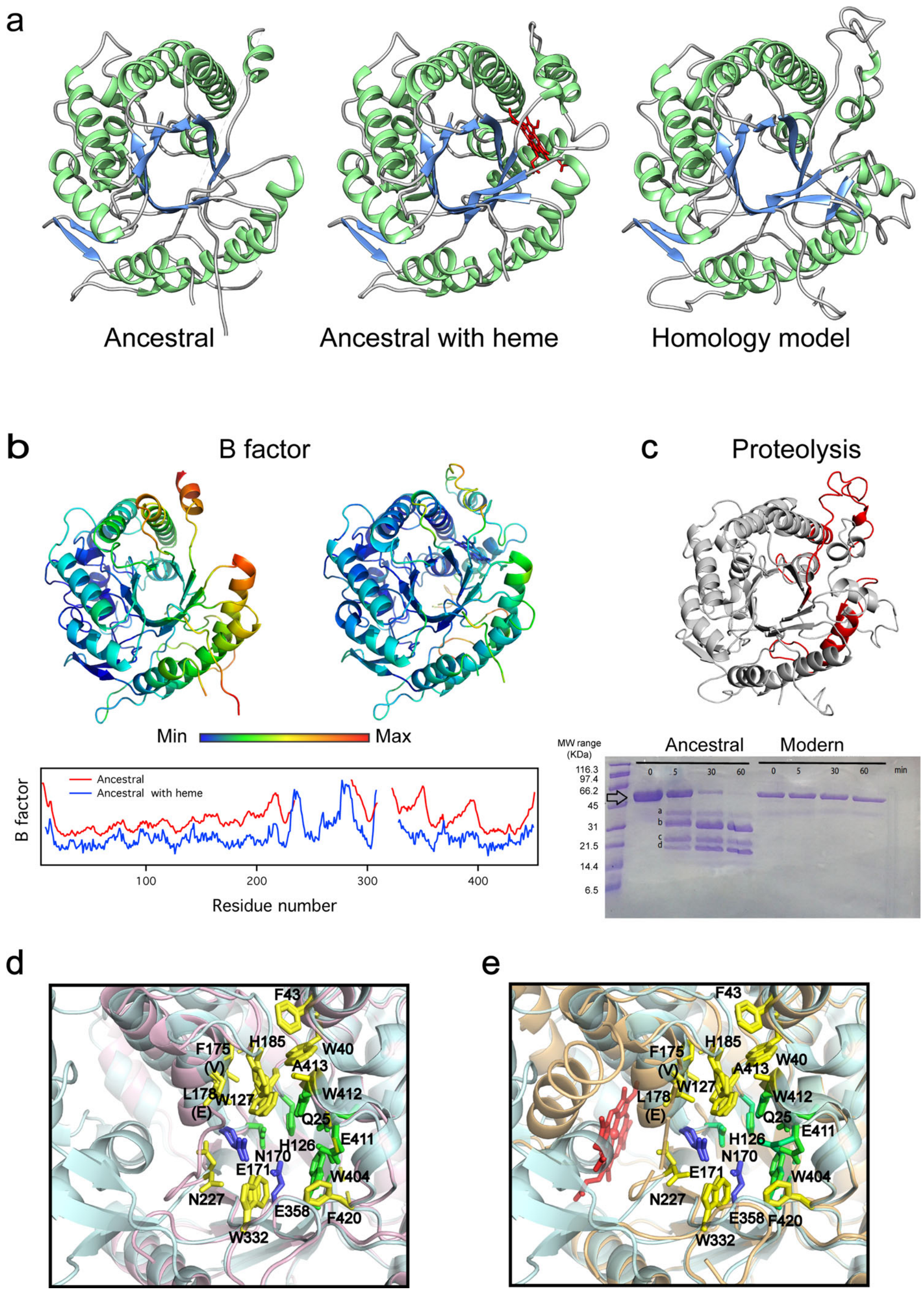


**Fig. 2 3D Structure of the ancestral glycosidase as determined by X-ray crystallography. a** Comparison between the ancestral structure determined in the absence (left) and presence (middle) of bound heme (red) and a homology model constructed as described in Supporting Information. The visual comparison reveals the missing sections in the electronic density of the ancestral protein, mostly in the protein without heme bound. **b** 3D structure of the ancestral protein without and with heme bound color-labeled according to normalized B-factor value and profiles of normalized B-factor versus residue number for the ancestral protein without (red) and with (blue) bound heme. Values are not shown for the sections that are missing in the experimental structures. **c** Proteolysis experiments with the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii*. The major fragments are labeled *a*, *b*, *c* and *d*. Molecular weights (MW) are shown for the markers used. Five independent experiments were performed with similar results. Mass spectrometry of the fragments predicts cleavage points within the red labeled sections in the shown structure. **d** Superposition of the structure of the ancestral glycosidase with that of the modern glycosidase from *Halothermothrix orenii* showing the critical active-site residues. **e** Superposition of the structures of the ancestral glycosidase without and with heme bound showing the critical active-site residues. In both (**d**) and (**e**), the highlighted active-site residues include the catalytic carboxylic acid residues (blue) and the residues involved in binding of the glycone (yellow) and aglycone (green) parts of the substrate.

Fig. 2c, and Fig. S2). Conversely, the modern glycosidase from the thermophilic *Halothermothrix orenii* remains essentially unaffected after several hours with the same concentration of the protease (Fig. 2c) or even with a ten times larger protease concentration. These two glycosidases, modern thermophilic and putative ancestral, are monomeric, as determined from gel filtration chromatography and analytical ultracentrifugation (Figs. S3 and S4) and display similar values for the optimum activity temperature (70 °C and 65 °C, respectively: Fig. 1b and S5). Therefore, their disparate susceptibilities to proteolysis can hardly be linked to differences in overall stability, but rather to enhanced conformational flexibility in the ancestral enzyme that exposes cleavable sites.

Furthermore, there is a large region missing in the electronic density map of the ancestral protein from X-ray crystallography (Fig. 2a), while the rest of the model agrees with a homology model based on modern glycosidase structures (see Supplementary Methods for details). At the achieved resolution of 2.5 Å, it should be possible to trace the course of a polypeptide chain in space, provided that such course is well defined. Therefore, the missing regions very likely correspond to regions of high flexibility. In addition, flexibility is also suggested by the B-factor values in regions that are present in the ancestral structure (Fig. 2b).

Lastly, molecular dynamics (MD) simulations (Fig. 3) also indicate enhanced flexibility in specific regions as shown by cumulative 15 μs simulations of the substrate-free forms of the ancestral glycosidase (both with and without heme: see below) as well as the modern glycosidase from *Halothermothrix orenii* (PDB ID: 4PTV)[27] [https://www.rcsb.org/structure/4PTV]. Both ancestral and modern proteins have the same sequence length, and similar protein folds with a root mean square deviation (RMSD) difference of only 0.7 Å between the structures. However, our molecular dynamics simulations indicate a clear difference in flexibility in the region spanning residues 227–334, which is highly disordered in the ancestral glycosidase but ordered and rigid in the modern glycosidase, with root mean square fluctuation (RMSF) values of <2 Å (Fig. 3). We also analyzed the interactions formed between residues 227–334 and the rest of the protein by counting the total intramolecular hydrogen bonds formed along the MD simulations. We observe that on average, the modern glycosidase forms 115 ± 11 hydrogen bonding interactions during our simulations, whereas the ancestral glycosidase forms either 101 ± 12/101 ± 13 hydrogen bonding interactions (in the presence of heme and absence of heme, respectively). This suggests that the higher number of intramolecular hydrogen bonds formed between residues 227–334 and the protein can contribute to the reduced conformational flexibility observed in the case of the modern glycosidase, compared to the ancestral glycosidase.

It is important to note that there is a clear structural congruence between the results of the experimental and computational studies described above. That is, the missing regions in the X-ray structure (Fig. 2a) match the high-flexibility regions in the molecular dynamics simulations (Fig. 3) and include the proteolysis cleavage sites determined by mass spectrometry (Fig. 2c). Overall, regions encompassing two alpha helices and several loops appear to be highly flexible or even unstructured in the ancestral glycosidase. The barrel core, however, remains structured and shows comparatively low conformational flexibility, which may explain the high thermal stability of the protein (see Discussion).

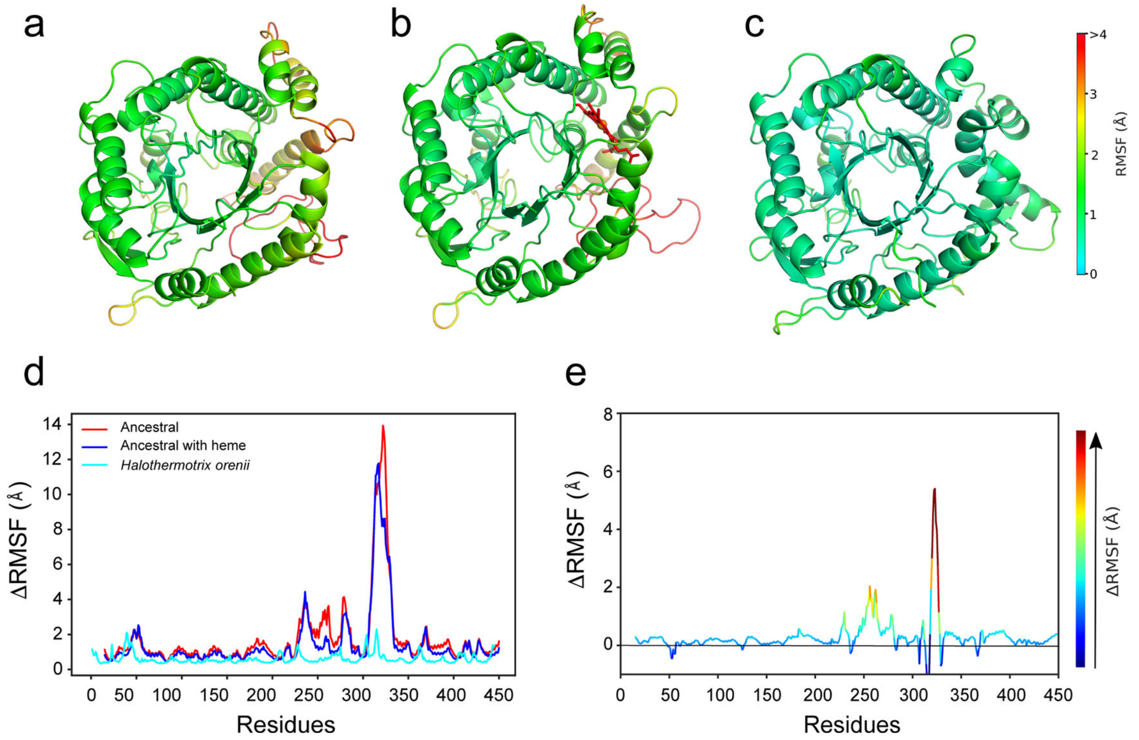


**Fig. 3 Molecular dynamics simulations.** Representative snapshots from molecular dynamics simulations of ancestral and modern glycosidases, showing the ancestral glycosidase both (**a**) without and (**b**) in complex with heme, as well as (**c**) the corresponding modern protein from *Halothermothrix orenii*. Structures were extracted from our simulations based on the average structure obtained in the most populated cluster using the hierarchical agglomerative algorithm implemented in CPPtraj[64]. All protein structures are colored by calculated root mean square fluctuations (RMSF) over the course of simulations of each system (see the color bar). Shown are also (**d**) absolute and (**e**) relative RMSF (Å) for each system, in the latter case showing the RMSF of the ancestral glycosidase without heme relative to the heme bound structure. Note the difference in the color bars between panels (**a**–**c**), which describes absolute RMSF per system, and panel (**e**), which describes relative RMSF. The numerical scale of the color bar on panel (**e**) corresponds to the *y*-axis of this panel.

**Catalysis**. We determined the Michaelis–Menten parameters for the ancestral enzyme with the substrates typically used to test the standard β-glucosidase and β-galactosidase activities of family 1 glycosidases (4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside) (Fig. 4 and Tables S2 and S3). We also compared the results with the catalytic parameters for four modern family 1 glycosidases, specifically those from *Halothermothrix orenii*, *Marinomonas sp.* (strain MWYL1), *Saccharophagus degradans* (strain 2-40 T), and *Thermotoga maritima* (Figs. S6–S9 and Tables S2 and S3). Modern glycosidases are highly proficient enzymes accelerating the rate of glycoside bond hydrolysis up to about 17 orders of magnitude[16]. The ancestral enzyme appears to be less efficient and shows a turnover number about two orders of magnitude below the values for the modern glycosidases studied here (Fig. 4). It is important to note, nevertheless, that the turnover number for the ancestral enzyme is still ~13 orders of magnitude higher than the first-order rate constant for the uncatalyzed hydrolysis of β-glucopyranosides, as determined by Wolfenden through Arrhenius extrapolation from high-temperature rates[15]. The catalytic carboxylic acid residues as well as the residues known to be responsible for the interaction with the glycone moiety of the substrate[28] are present in the ancestral enzyme and appear in the static X-ray structure in a configuration similar to that observed in the modern proteins

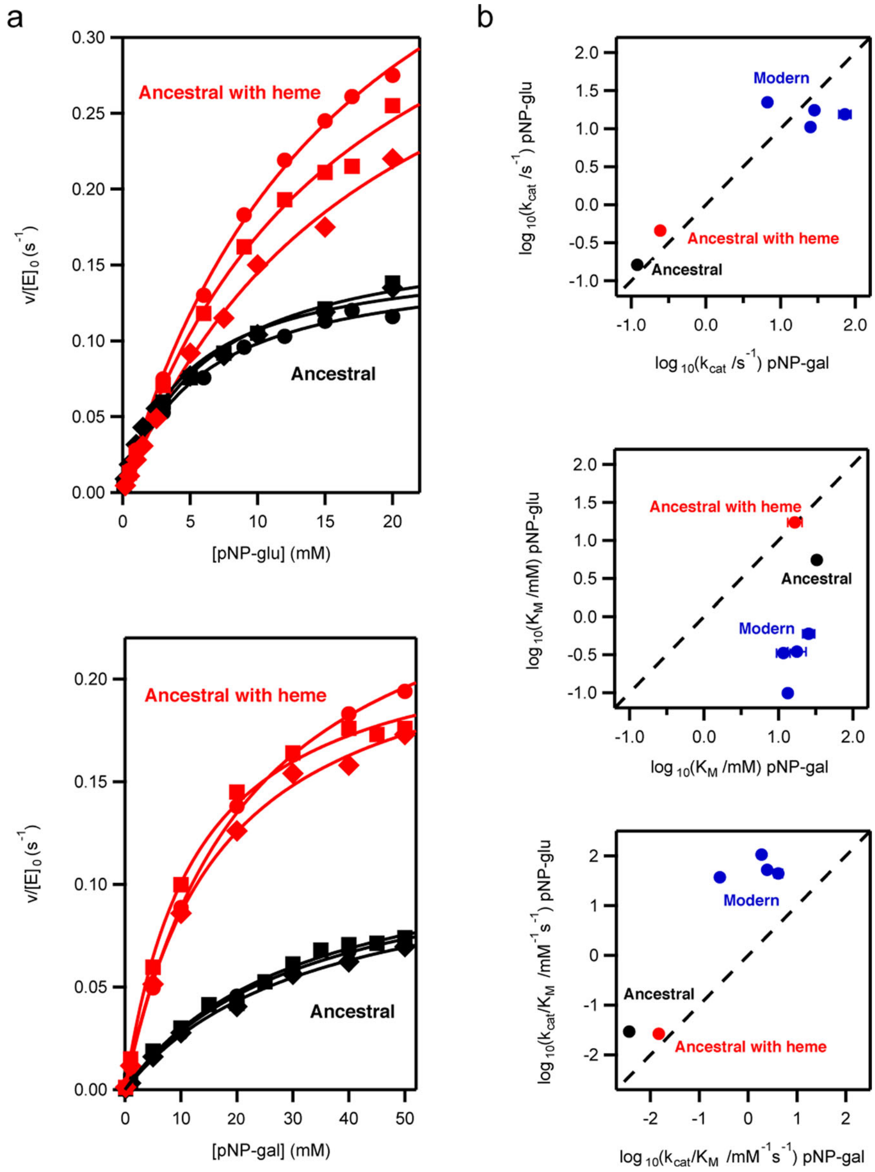


**Fig. 4 Ancestral versus modern catalysis by family 1 glycosidases. a** Michaelis plots of rate versus substrate concentration at pH 7 and 25 °C for hydrolysis of 4-nitrophenyl-β-D-glucopyranoside (upper panel) and 4-nitrophenyl-β-D-galactopyranoside (lower panel) catalyzed by the ancestral glycosidase with and without heme bound. v/[E]$_0$ stands for the rate over the total enzyme concentration. The lines are the best fits of the Michaelis–Menten equation. The different symbols (diamond, square, circle) refer to the triplicate experiments (involving two different protein preparations) performed for each protein/substrate combination. Michaelis plots for the four modern proteins studied in this work can be found in Figs. S6–S9. The values for the catalytic parameters derived from these fits are collected in Tables S2 and S3. **b** Logarithm of the Michaelis-Menten catalytic parameters for a glucopyranoside substrate versus a galactopyranoside substrate. pNP-glu and pNP-gal stand, respectively, for 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside. $k_{cat}$, $K_{M,}$ and $k_{cat}/K_M$ stand for the turnover number, the Michaelis constant and the catalytic efficiency. The values shown are averages of the values derived from the triplicates and the associated errors are the corresponding standard deviations. Note that, in most cases, the associated errors are smaller than the size of the data points.

(Fig. 2d). There are a few differences in the identity of the residues responsible for the binding of the aglycone moiety of the substrate[28], but these differences occur in positions that are variable in modern family 1 glycosidases (Fig. S10 and Table S4). Overall, the comparatively low activity of the ancestral protein is likely linked to its conformational flexibility. That is, the protein in solution is sampling a diversity of conformations of which only a few are active towards the common substrates. From an evolutionary point of view, the comparatively low ancestral activity may reflect an early stage in the evolution of family 1 glycosidases before selection favored greater turnover (see "Discussion").

Also, it is interesting to note that, although both β-glucosidase and β-galactosidase activities are typically described for family 1 glycosidases, these enzymes are commonly specialized as β-glucosidases[22]. This specialization does not occur, however, at the level of the turnover number, which is typically similar for both kinds of substrates. Instead, specialization occurs at the level of the substrate affinity, as reflected in lower values of the Michaelis constant ($K_M$) for β-glucopyranoside substrates as compared to β-galactopyranoside substrates[22]. This pattern is indeed observed in the modern enzymes we have studied (Fig. 4), which are described in the literature as β-glucosidases. On the other hand, this kind of specialization is not observed in the ancestral glycosidase, which shows similar $K_M$'s for the β-glucopyranoside and the β-galactopyranoside substrates. This lack of specialization may again reflect an early stage in the evolution of family 1 glycosidases, an interpretation which would seem generally consistent with the fact that resurrected ancestral proteins often display promiscuity[5,7,9,29]. On the other hand, it can be argued that the ancestral glycosidase was specialized for a different kind of substrate. To explore this possibility, we determined catalytic rates for a wide range of glycosidase substrates. These studies are briefly described below:

(1) Using the same methodology employed with *4*-nitrophenyl-β-D-glucopyranoside and *4*-nitrophenyl-β-D-galactopyranoside (Fig. 1), we determined profiles of catalytic rate versus temperature for the ancestral glycosidase and the modern glycosidases from *Halothermothrix orenii* and *Saccharophagus degradans* using as substrates *4*-nitrophenyl-β-D-fucopyranoside, *4*-nitrophenyl-β-D-lactopyranoside, 4-nitrophenyl-β-D-xylopyranoside and 4-nitrophenyl-β-D-mannopyranoside. In all cases (Fig. S11), we found the levels of catalysis of the ancestral protein to be reduced in comparison with the modern proteins. We also found that the levels of catalysis for the β-glucopyranoside and β-fucopyranoside substrates were similar, but this pattern is also observed with the modern proteins. (2) We carried out single activity determinations at 25 °C for the ancestral glycosidase with a wider range of substrates, including derivatives of disaccharides (maltose, cellobiose) and several substrates with an α anomeric carbon (Table S5). However, we did not find any substrate with a catalysis level substantially higher than that of those previously determined for 4-nitrophenyl-β-D-glucopyranoside and *4*-nitrophenyl-β-D-galactopyranoside and, in many cases (in particular with the α substrates), no substantial activity was detected. (3) Since some of the proteins that descended from the N72 node are 6-phosphate-β-glucosidases (Fig. 1A and S1), we tested the activity of our ancestral glycosidase against *4*-nitrophenyl-β-D-glucopyranoside-6-phosphate (Fig. S12). We found the catalytic efficiency to be ~40 fold smaller than that determined with the corresponding non-phosphorylated substrate. (4) Glycosidases are typically described[14] as being very promiscuous for the aglycone moiety of the substrate (the part of the substrate that is replaced with *p*-nitrophenyl in the substrates commonly used to assay glycosidase activity) while they are more specialized for the glycone moiety of the substrate. However, the flexibility in certain regions of the ancestral structure could perhaps favor the hydrolysis of substrates with larger aglycone moieties. To explore this hypothesis, we tested four synthetic substrates with aglycone moieties larger than the usual *p*-nitrophenyl group (Fig. S13). We revealed that ancestral levels of catalysis are substantially reduced with respect to those obtained for the modern glycosidase from *Halothermothrix orenii*, used here as comparison.

**Heme binding and allosteric modulation**. Overall, it appears reasonable that our resurrected ancestral enzyme reflects an early stage in the evolution of family 1 glycosidases, perhaps following a fragment fusion event (see Discussion), at which catalysis was not yet optimized and substrate specialization had not yet evolved. The presence of a large unstructured and/or flexible regions in the ancestral structure could perhaps reflect the absence of a small molecule that binds within that region. While these proposals are speculative, the experimental results described in detail below, show that the ancestral glycosidase does bind heme tightly and stoichiometrically at a site in the flexible regions. This was a completely unexpected observation given the large number of modern glycosidases that have been characterized in the absence of any porphyrin rings.

We curiously noticed that most preparations of the ancestral glycosidase showed a light-reddish color after elution from an affinity column (Fig. 5). UV–Vis spectra revealed the pattern of bands expected for a heme group[30], including the Soret band at about 400 nm and, in some cases, even the weaker α and β bands (i.e., the Q bands) in the 500–600 nm region (Fig. 6). From the intensity of the Soret band, a very low heme:protein ratio of about 0.02 was estimated for standard enzyme preparations, indicating that all the experiments described above were performed with essentially heme-free protein. However, the amount of bound heme in protein preparations was substantially enhanced by including hemin in the culture medium (heme with iron in the +3 oxidation state) or 5-aminolevulinic acid, the metabolic precursor of heme (Fig. 5). Heme:protein ratios of about 0.10 and 0.18, respectively, were then obtained (Fig. 6a). These results suggest that the ancestral glycosidase does have the capability to bind heme, but also that, as is commonly the case with modern heme-binding proteins[31], the limited amount of heme available in the expression host, combined with the high protein over-expression levels used, leads to low heme:protein ratios. The capability of the ancestral enzyme to bind heme was first shown by the in vitro experiments described next, and then confirmed by mass spectrometry and X-ray crystallography as subsequently described.

Heme has a tendency to associate in aqueous solution at neutral pH, a process that is reflected in a time-dependent decrease in the intensity of the Soret band, which becomes "flatter" upon the formation of dimers and higher associations[32]. However, the process is reversed upon addition of the essentially heme-free ancestral glycosidase (Fig. 6b), indicating that the protein binds heme and shifts the association equilibria towards the monomeric state. Remarkably, heme binding is also reflected in a several-fold increase in enzymatic activity which occurs on the seconds time scale when the heme and enzyme concentration are at a ~micromolar concentration (Fig. 6c). Determination of activity after suitable incubation times for different heme:protein ratios in solution yielded a plot with an abrupt change of slope at a stoichiometric ratio of about 1:1 (Fig. 6d). These experiments were carried out with ~micromolar heme and protein concentrations, indicating therefore a tight, sub-micromolar binding. This interaction was confirmed by microscale thermophoresis experiments that yielded an estimate of 547±110 nM for the heme dissociation constant (see "Methods" and Fig. S14 for details). Indeed, in agreement with this tight binding, increases in activity upon heme addition to a protein solution were observed (Fig. 6c) even with concentrations of ~50 nanomolar.

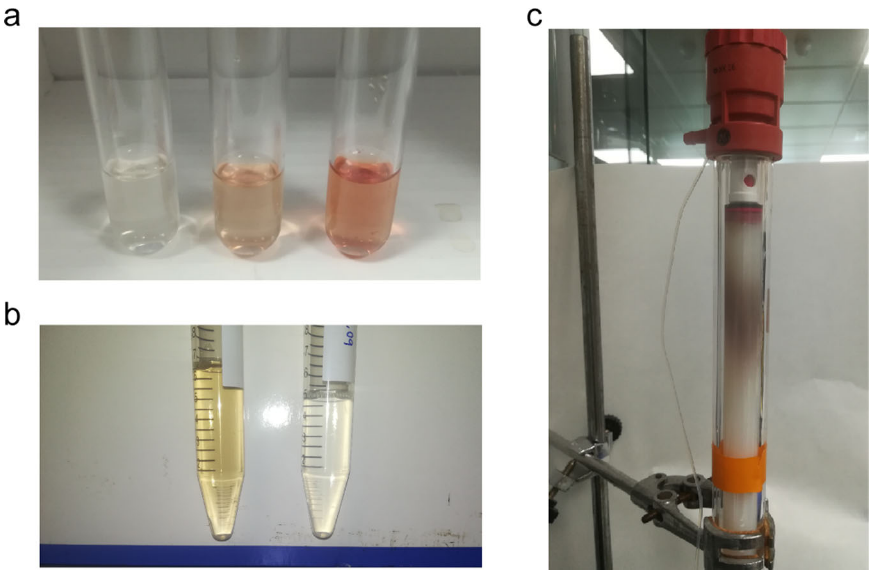


**Fig. 5 Heme binding to the ancestral glycosidase is visually apparent. a** The ancestral protein was prepared by Ni-NTA affinity chromatography. The pictures show the samples eluted from the columns for three different preparations that differed by the addition of 20 μM hemin (middle) and 0.4 mM 5-aminolevulinic acid (right) to the culture medium. Neither hemin nor 5-aminolevulinic acid had been added to the culture medium in the preparation on the left. Protein concentrations in these samples were ~10 mg/mL. **b** An ancestral glycosidase sample with a low amount of bound heme (right) was incubated with an excess of heme. A PD10 column and FPLC (fast protein liquid chromatrography) were then used to remove the unbound heme. The resulting protein preparation is shown on the left. Protein concentration is ~0.5 mg/mL. **c** The position of the ancestral protein with bound heme in a FPLC column is revealed by a reddish-brown band.

The 1:1 stoichiometry of heme/protein was confirmed by experiments in which the protein was incubated with an excess of heme and free heme was removed through exclusion chromatography (2 passages through PD10 columns). The protein was then quantified by the bicinchoninic acid method[33] with the Pierce™ BCA Protein Assay Kit while the amount of heme was determined using the pyridine hemochrome spectrum[34] after transfer to concentrated sodium hydroxide (see Methods for details). This resulted in a heme/protein stoichiometry of 1.03±0.03 from five independent assays.

The experiments described above allowed us to set up a procedure for the preparation of the ancestral protein saturated with heme and to use this preparation for activity determinations and crystallization experiments. The procedure (see Methods for details) involved in vitro reconstitution using hemin but did not include any chemical system capable of performing a reduction. It is therefore safe to assume that our heme-bound ancestral glycosidase contains iron in the +3 oxidation state. Activity determinations with the heme-saturated ancestral enzyme corroborated that heme binding increases activity by ~3 fold (see Michaelis plots in Fig. 4). Both mass spectrometry (Fig. S15) and X-ray crystallography confirmed the presence of one heme per protein molecule (Fig. 7, S16 and S17), which is located at the same site, with the same orientation and involved largely in the same molecular interactions in the three protein molecules (A, B, C) observed in the crystallographic unit cell. Besides interactions with several hydrophobic residues, the bound heme interacts (Fig. 7a) with Tyr264 of α-helix 8 (as the axial ligand), Tyr350 of α-helix 13, Arg345 of β-strand B and, directly via a water molecule, with Lys 261 of β-strand B, although this latter interaction is only observed in chain A. The bound heme shows B-factor values similar to those of the surrounding residues (Fig. 7b), it is well-packed and 95% buried (Fig. 7c). Indeed, the accessible surface area of the bound heme is only 43 Å$^2$ compared to the ~800 Å$^2$ accessible surface area for a free heme[35]. The interactions of the bound heme in the ancestral glycosidase are overall similar to those described for modern b-type heme proteins[35–37]. As observed in modern heme-binding proteins, the ancestral heme-binding pocket is enriched in hydrophobic and aromatic residues and propionate anchoring is achieved through interactions with arginine, tyrosine and lysine residues. Certainly, tyrosine, the axial ligand in the ancestral glycosidase, is not the most common axial ligand in modern heme proteins, but it is found in several cases, including catalases (see, Protein Data Bank (PDB) ID 1QWL for the 3D structure of the catalase from *Helicobacter pylori*). Interestingly, the amino acid residues that interact with the heme in the ancestral glycosidase are somewhat conserved, and are indeed the consensus residue from the set of modern glycosidases used as the starting point for ancestral reconstruction (Table S6). The fraction of each consensus residues in the modern protein is, however, less than unity and the sequences of modern glycosidases in the set differ from the ancestral sequence at many of the positions involved in heme interactions in the ancestral protein (Table S7).

Heme binding clearly rigidifies the ancestral protein, as shown by fewer missing regions in the electronic density map, in contrast to the structure of the heme-free protein (see Fig. 2a–c). This is also confirmed by molecular dynamics (MD) simulations of the ancestral glycosidase both with and without heme bound (Fig. 3 and S18). Figure S18 shows the backbone RMSD (root mean square deviation) over ten individual 500 ns MD simulations per system, and, from this data, it can be seen that while the RMSD is fairly stable in the case of the modern protein, the ancestral glycosidases (both with and without heme bound) are initially quite far from their equilibrium structures, due to the high flexibility of the missing regions of the protein which require substantial equilibration. In addition, we note that while the overall average RMSD for the ancestral protein with heme bound is slightly lower than for the ancestral protein without heme (Fig. S18), the standard deviation is higher. This is due to the greater flexibility of the reconstructed missing loop (see the “Methods” section), which allows it to sample a larger span of conformations depending on whether the loop is interacting with the bound heme or not (we observe both scenarios in our simulations of the heme-bound ancestral glycosidase).

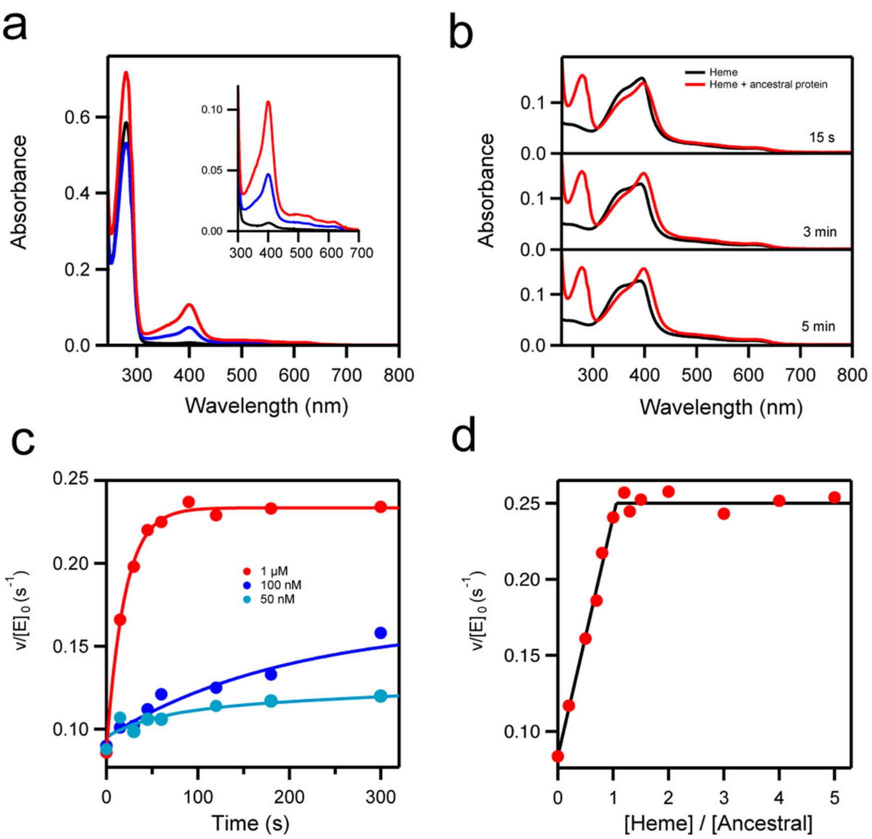


**Fig. 6 Heme binding to the ancestral glycosidase. a** UV–VIS spectra for preparations of the ancestral glycosidase showing the protein absorption band at about 280 nm and the absorption bands due to the heme (the Soret band at about 400 nm and the Q bands at higher wavelengths). Black color is used for the protein obtained using the original purification procedure without the addition of hemin or hemin precursor. Blue and red are used to refer, respectively, to preparations in which hemin and 5-aminolevulinic acid (the metabolic precursor of heme) were added to the culture medium. **b** Binding of heme to the ancestral glycosidase in vitro as followed by changes in VIS spectrum. Spectra of a heme solution 1 μM in the absence (black) or presence (red) of a similar concentration of ancestral protein. The "flat" Soret band of free heme is linked to its self-association in solution, while the bound heme is monomeric and produces a sharper Soret band. **c** Kinetics of binding of heme to the ancestral glycosidase as followed by the increase in enzyme activity (rate of hydrolysis of *4*-nitrophenyl-β-glucopyranoside; see Methods for details). In the three experiments shown a heme to protein molar ratio of 1.2 was used. The protein concentration in each experiment is shown. Note that activity increase is detected even with concentrations of 50 nM, indicating that binding is strong. The lines are meant to guide the eye and thus have no quantitative purpose. (**d**) Plot of enzyme activity versus [heme]/[protein] ratio in solution for a protein concentration of 1 μM. Activity was determined after a 5 min incubation and the plot supports a 1:1 binding stoichiometry. $v/[E]_0$ stands for the rate over the total enzyme concentration.

In contrast, in the absence of the heme, the loop is always in a flexible open conformation leading to a higher overall RMSD but a lower standard deviation as a narrower range of conformations are sampled in our simulations. As neither the loop nor the heme has access to the active site (Fig. S17), these differences are unlikely to have a direct effect on catalysis.

The higher flexibility of the ancestral protein without heme bound can also be seen from comparing RMSF (root mean square fluctuation) values across the protein. That is, the MD simulations performed without the heme bound show that most of the protein has higher flexibility (Fig. 3e, with ΔRMSF values greater than 0 in most of the sequence), particularly in the regions where the B-factors also indicate high flexibility (Fig. 2b). This is noteworthy, as the only difference in starting structure between the two sets of simulations is the presence or absence of the heme; the starting structures are otherwise identical. The MD simulations show that removing the heme from the heme-bound structure has a clear effect on the flexibility of the whole enzyme, increasing it relative to the heme bound structure (Fig. 3e), again also indicated by the B-factors (Fig. 2b). There are two regions where this difference is particularly pronounced. The first spans residues 25–265, which is located where the heme Fe(III) atom forms an interaction with the Tyr264 side chain as an axial ligand. Removing the heme removes this interaction, thus inducing greater flexibility in this region. The second region with increased flexibility spans 319–327, where again we observe that removing the heme increases the flexibility of this region.

Lastly, we note that the heme is located near the enzyme active site (at about 8 Å from the catalytic glutamate at position 171) but does not have direct access to this site as revealed in the structure (Fig. S17). Therefore, the increase in activity observed upon heme binding is an allosteric effect likely linked to dynamics (see "Discussion") since heme binding does not substantially alter the position/conformation of the catalytic carboxylic acids nor the residues involved in substrate binding according to the static X-ray structures (Fig. 2e). In fact, examining backbone RMSF values of key catalytic residues (Fig. S19) indicates that the flexibility of several of these residues is reduced upon moving from the ancestral glycosides without heme, to adding the heme, to the modern glycosidase, in a clear decreasing trend. We note that the observed effects are subtle and sub-Å; however, there are several experimental studies that suggest that sub-Å changes in dynamics can be catalytically important[38–40].

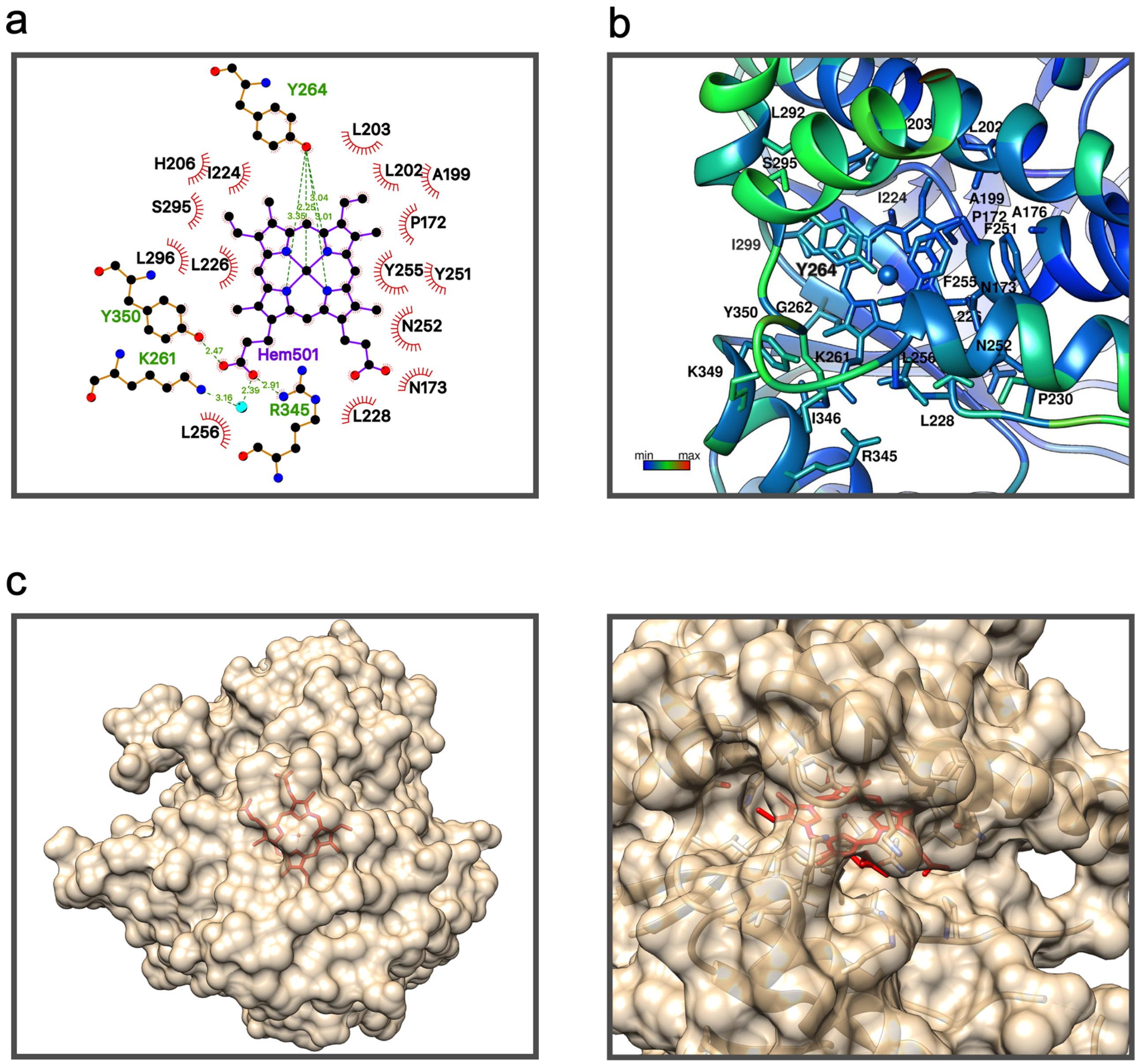


**Fig. 7 Local molecular environment of the bound heme in the ancestral protein. a** Schematic representation of the heme molecule and the neighbor residues in the 3D structure. **b** Heme group and residues directly interacting with the heme colored by B value. **c** Van der Waals surface of the ancestral protein shown in translucent brown, so that it becomes visually apparent that the heme (shown in red) is mostly buried.

## Discussion

The TIM-barrel is the most common enzyme fold, accounting for ~10% of known enzyme structures and providing a scaffold for an enormous diversity of biomolecular functions[11–13]. It is composed of eight parallel (β/α) units linked by hydrogen bonds forming a cylindrical core ("the barrel") with secondary structure elements connected by loops. The high capability of the fold to accommodate a wide diversity of different natural functions is likely linked to its modular architecture, with the barrel (and the αβ loops) providing stability and allowing a substantial degree of flexibility, variability, and, therefore, evolvability for the βα loops. That is, the barrel provides a stable platform that can accommodate loops of different sequences and conformations at the so-called catalytic face.

Remarkably, the differences in conformational flexibility between different parts of the molecule appear to be even more pronounced in our ancestral TIM-barrel glycosidase. Stability is still guaranteed by a rigid barrel core, but flexibility is greatly enhanced and extends to large parts of the structure, as shown by a combination of computational and experimental results. Conformational flexibility implies that the protein in solution is sampling a diversity of conformations. On the one hand, this may prevent the enzyme from reaching the highest levels of catalysis for a given natural reaction since the protein ensemble may not be shifted towards the most active conformations. Indeed, while modern glycosidases approach catalysis levels up to 17 orders of magnitude above the rate of spontaneous glycoside bond hydrolysis[16], the ancestral glycosidase displays turnover numbers about two orders of magnitude below the modern glycosidases studied here (Fig. 4 and Tables S2 and S3). On the other hand, flexibility is key to the emergence of new functions and contributes to evolvability, since minor conformations that catalyze alternative reactions may be enriched by subsequent evolution[41–44]. Therefore, the ancestral TIM-barrel described here holds promise as a scaffold for the generation of de novo catalysts, an important and largely unsolved problem in enzyme engineering. We have recently shown[45] that completely new enzyme functions can be generated through a single mutation that generates both a cavity and a catalytic residue, provided that conformational flexibility around the mutation site allows for substrate and transition-state binding[10,43,44]. The combination of a rigid core that provides stability with high flexibility in specific regions makes the ancestral protein studied here an excellent scaffold to develop this minimalist approach to de novo catalysis (work in progress).

Catalytic features of the ancestral glycosidase, such as diminished activity levels and lack of specialization for glucopyranoside substrates, would seem consistent with an early stage in the evolution of family 1 glycosidases. It has been proposed that TIM-barrel proteins originated through fusions of smaller fragments[46]. The high conformational flexibility in some regions of the ancestral glycosidase structure would then also seem consistent with an early evolutionary stage, since fragment fusion is not expected to immediately lead to efficient packing and conformational rigidity in all parts of the generated structure. On the other hand, the capability of the reconstructed ancestral glycosidase to bind heme tightly and stoichiometrically at a well-defined site is rather surprising. None of the ~5500 X-ray structures for the ~1400 glycosidases currently reported in CAZy shows a porphyrin ring. It is certainly possible that heme binding to the ancestral glycosidase is simply an accidental byproduct of the high conformational flexibility at certain regions of the structure, although the tightness of the binding and the specificity of the molecular interactions involved argue against this possibility. In any case, this is an issue that can be investigated by studying modern glycosidases. If heme binding is a functional ancestral feature (a product of selection), we may expect that at least some modern glycosidases show some inefficient, vestigial capability to bind heme, in keeping with the general principle that features that become less functional undergo evolutionary degradation[47,48]. No mention of heme binding to modern family 1 glycosidases can be found in the CAZypedia resource[22], but, of course, there is no reason why researchers in the glycosidase community should have tested heme-binding capabilities. As such, we have done so in this work for the four modern enzymes we already characterized in terms of catalysis (Fig. 4), i.e., the modern family 1 glycosidases from *Halothermothrix orenii*, *Thermotoga maritima*, *Marinomonas sp*. (strain MWL1), and *Saccharophagus degradans* (strain 2-40). When 5-aminolevulinic acid, the metabolic precursor of heme, was added to the culture medium, the four modern proteins were isolated with an appreciable amount of bound heme, although their heme-binding capability is clearly much reduced when compared with the ancestral glycosidase (Fig. 8 and Fig. S20). We furthermore carried out the same type of experiment with proteins corresponding to reconstructions at five nodes in the line of descent that leads from the ancestral glycosidase at node 72 to the modern glycosidase from *Halothermothrix orenii* (Fig. 8). We also found appreciable, but typically much lower amounts of bound heme, as compared with the "older" ancestral protein at node 72 (Fig. 8 and Fig. S21). Finally, we purified the heme-free forms of the proteins at these five ancestral nodes and studied their heme-binding capability in vitro. These proteins have glycosidase activity levels intermediate between those of the ancestral glycosidase at node 72 and the modern glycosidase from *Halothermothrix orenii* (Fig. S22) and unambiguously display in vitro heme-binding capability at micromolar concentrations, as shown by the presence of the Soret band in UV–VIS spectra (Fig. S23) and further confirmed by mass spectrometry (Figs. S24–S26). Evolutionary degradation of ancestral heme binding, however, is clearly revealed by analyses of elution profiles from gel filtration chromatography in terms of protein concentration, heme concentration, and glycosidase activity (Fig. 8). Thus, while heme binding to the most ancient studied node (our ancestral glycosidase at node 72) produces active monomers to a large extent, a trend towards a decreased amount of heme-bound monomers and appearance of higher association states upon heme binding is observed in the evolutionary line leading to the modern glycosidase from *Halothermothrix orenii*. One interesting possibility is that heme binding to the monomers of the less ancient proteins brings about conformational changes that trigger protein association.

It emerges that heme binding to the ancestral glycosidase at node 72 is not an oddity or an artifact of reconstruction. In contrast, it appears probable that heme binding to ancient family 1 glycosidases did specifically occur, and that it also underwent degradation at an early evolutionary stage to lead to a rudimentary capability with substantial variability, as it is commonly observed with vestigial features that are not subject to selection[47]. Overall, this suggests a complex evolutionary history for this family of enzymes involving perhaps a fortuitous (i.e., contingent) early fusion event with a heme-containing domain. In this scenario, heme had a functional role in the isolated heme-containing domain, which was no longer required when the domain was fused with the larger glycosidase scaffold, thus enabling the subsequent degradation of the heme-binding capability. In order to find some evidence for this hypothesis, we used the Dali sever[49] to search in the Protein Data Bank for structural alignments of the alpha helices involved in heme binding from our ancestral glycosidase. However, we did not find any convincing match, as the best obtained structural alignments had RMSD values of ~4 Å and Z scores of 2 or higher (see Fig. S27 for further details). It is possible that the structure of the ancestral heme-containing domain was distorted upon fusion and subsequent evolution, and is therefore difficult to identify in searches of modern protein structures. Another possibility is that there was never a fusion event with a heme-containing domain and that heme was already present even at the most ancient stages in the origins of family 1 glycosidases. This would be consistent with an interpretation of cofactors as molecular fossils that facilitated the primitive emergence of proteins by selecting them from a random pool of polypeptides[50].

It is important to note at this stage that relevant protein-engineering implications are independent of any evolutionary interpretations. In this context, heme-binding to the ancestral glycosidases, regardless of its evolutionary origin and implications, opens up new engineering possibilities. This is so because directed laboratory evolution can be used to enhance or modify any functionality, provided that a certain level of functionality is used to start the process. The capability to "seed" levels of new functionalities may become a critical bottleneck in protein-engineering projects. Our work uncovers a heme-binding capability and a possibility of allosteric regulation that were previously unknown in glycosidase enzymes. Potential practical implications are briefly discussed below.

Metalloporphyrins are essential parts of many natural enzymes involved in redox and rearrangement catalysis and can be engineered for the catalysis of non-natural reactions[51]. Remarkably, however, the combination of the highly evolvable TIM-barrel scaffold and the catalytically versatile metalloporphyrins is exceedingly rare among known modern proteins. A porphyrin ring is found in only 13 out of the 7637 PDB entries that are assigned the TIM-barrel fold according to the CATH classification[52]. These 13 entries correspond to just two proteins. One of them, uroporphyrinogen decarboxylase, is an enzyme involved in heme biosynthesis, while in the other identified case, flavocytochrome B2, the bound heme is far from the active site. By contrast, heme appears at about 8 Å from the catalytic Glu171 in our ancestral glycosidase. While its connection with the

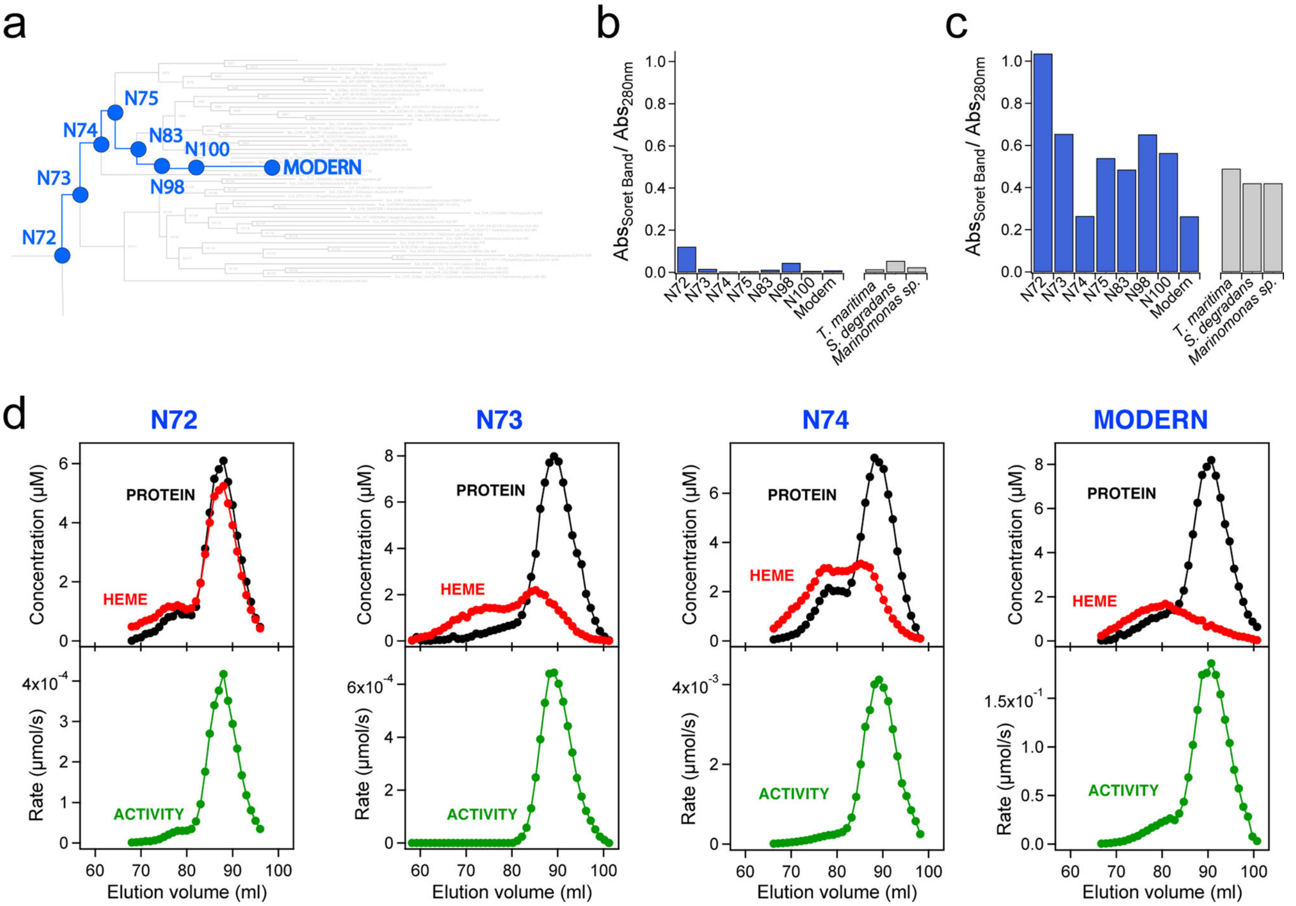


**Fig. 8 Evolutionary degradation of ancestral heme binding. a** Section of the phylogenetic tree used for the Bayesian analysis of family 1 glycosidases. The nodes in the evolutionary trajectory from node 72 to *Halothermothrix orenii* (labeled "MODERN") are highlighted and labeled. See Fig. S1 for a complete and detailed version of the phylogenetic tree. **b**, **c** Ratio of absorbance (Abs) at the maximum of the heme Soret band to the absorbance at the maximum of the protein aromatic absorption band for ancestral (see panel A) and modern family 1 glycosidases. Data in (**b**) correspond to protein preparations in which 0.4 mM 5-aminolevulinic acid (the metabolic precursor of heme) was added to the culture medium and the protein was purified by Ni-NTA affinity chromatography and further passage through a PD10 column (see Figs. S20 and S21 for further detail). For the data in (**c**), heme-free protein samples at ~5 μM were incubated for 1 h at pH 7 with a 5-fold excess of heme and free heme was removed through size exclusion chromatography (2 passages through PD10 columns) before recording the UV–VIS spectra (Fig. S23). **d** Profiles of protein concentration (bicinchoninic acid method[33]), heme concentration (pyridine hemochrome spectrum[3]) and glycosidase activity (with 0.25 mM 4-nitrophenyl-β-D-glucopyranoside) upon elution from gel filtration chromatography (HiLoad 16/600 Superdex 200 pg GE Healthcare). For these experiments, heme was gradually added to ~30 μM samples of ancestral (N72, N73, and N74) and modern (*Halothermothrix orenii*) glycosidases up to a ~5-fold excess and, after several hours, free heme was removed using PD10 columns. The elution volume for the main protein concentration peak is consistent with the monomeric association state (Fig. S3).

natural active site is blocked by several side chains in the determined structure (Fig. S17), it would appear feasible to use protein engineering to establish a conduit. As a simple illustration of this possibility, mutating Pro172, Asn173, Ile224, Leu226, Asn227, and Pro 272 to alanine in silico (Fig. S17 and Table S8) increase the accessible surface area of heme from barely 43 Å$^2$ to about 300 Å$^2$ and exposes the side of the heme facing the active site. The possibility thus arises that the engineering of metalloporphyrin catalysis, through rational design and/or laboratory evolution, would benefit from the evolutionary possibilities afforded by the flexible βα loops at a TIM-barrel catalytic face.

More immediate engineering possibilities arise from the allosteric modulation of catalysis in the ancestral protein, a phenomenon that, to our knowledge, has never been reported for modern glycosidases. Heme binding rigidifies the ancestral glycosidase and causes a several-fold activity enhancement. Heme is not expected to catalyze glycoside hydrolysis and, in any case, the bound heme does not have access to the active site in the experimentally determined ancestral structure. The activity enhancement upon heme binding is therefore an allosteric effect likely linked to dynamics, as it is the case with other allosteric effects reported in the literature[53]. Regardless of whether this feature is truly ancestral or just a byproduct of the enhanced conformational flexibility of the putative ancestral glycosidase, it is clear that it can provide a basis for biosensor engineering. For instance, computational design and laboratory directed evolution could be used to repurpose the heme-binding site for the binding of a targeted substance of interest and to achieve a large concomitant change in glycosidase activity. The development of this application should be facilitated by the availability of a wide diversity of synthetic chemical probes for the sensitive detection of glycosidase activity[17]. In total, we anticipate that unusual combinations of protein features will generate new possibilities in protein biotechnology and engineering.

## Methods

**Ancestral sequence reconstruction**. Characterized GH1 protein sequences were retrieved from the Carbohydrate-active enzyme database (CAZy)[18], including β-glucosidase (accession numbers: ACI19973.1 and AAL80566.1), 6-P-β-glucosidase (AIY91871.1), β-mannosidase (AAL81332.1), and myrosinase (AAK32833.1). These characterized protein sequences were utilized as seeds to identify additional homologous sequences. GH1 homologs were retrieved for all three domains of life from GenBank (http://www.ncbi.nlm.nih.gov/) using BLASTp, with the cutoff threshold of $<1 \times 10^{-5}$. Sequences with the minimum length of 300 amino acids were included in the dataset. Taxonomically redundant sequences were excluded. A total number of 150 sequences were collected for further analysis.

Sequences were aligned using T-Coffee. Initial non-bootstrapped phylogenetic trees were constructed using RAxML (ver. 8.2.11) to identify major clusters and to eliminate spurious sequences[54]. The RAxML analysis was performed with the hill-climbing mode using the gamma substitution model. These initial trees were used as starting point for more thorough Bayesian analysis. MrBayes (ver. 3.2.6) was conducted using the WAG amino acid replacement model with a gamma distribution and invariable sites model for at least 1,000,000 generations, with sampling at intervals of 100 generations, and two runs with four chains per run in order to monitor convergence[55]. Twenty-five percent of sampled points were discarded as burn-in. The tree topology was broadly identical between the RAxML and MrBayes analyses.

Ancestral sequences were reconstructed using FastML (ver. 3.1) with the WAG amino acid replacement model with a gamma distribution for variable replacement rates across sites[56].

**Database searches**. We searched the CAZy database in order to ascertain the presence of porphyrin rings in reported glycosidase structures. We systematically went through all 167 glycoside hydrolase families of the database in March 2020. For each family, we checked the structure section and we individually examined all the links provided to the protein data bank. Overall, we examined 5565 PDB files corresponding to 1435 different glycosidase enzymes. We did not find a single example of a reported structure with a bound porphyrin ring.

We also used the CAZy database as a starting point of an extensive literature search for optimum temperature values of family 1 glycosidases. We examined the section of characterized enzymes for family 1 glycosidases, the references included in the corresponding GenBank links, as well as the publications that cite those references in a Google Scholar search. Several hundred published articles were examined for experimental activity versus temperature profiles and reported values of the optimum temperature. We found such data for 126 different family 1 glycosidases. In many cases, the oligomerization state of the enzymes was also provided in the original references. The environmental temperatures (optimum growth temperatures) of the corresponding host organisms could be found in most cases in the "Bergey's Manual of Systematic Bacteriology" although, in some cases, literature searches were performed to find the optimum temperatures. Most organisms were classified as hyperthermophilic, extreme thermophilic, thermophilic, mesophilic or psychrophilic in Bergey's manual or the relevant literature references. We have used this classification to color code Fig. 1c, since it leads to clear and intuitive data clusters. All the information related to the values of the optimum temperature for activity and the organismal living temperature is collected in Supplementary Dataset 1.

In order to find examples of proteins with the TIM-barrel fold and a bound porphyrin ring in the reported structures, we checked (March-2020) all entries in the protein data bank that are classified as TIM-barrels in the CATH database. We examined a total of 7637 PDB files and found a porphyrin ring in only 13 of them. These 13 structures correspond to two proteins: flavocytochrome B2, a multi-domain protein in which the porphyrin ring is located in the non-TIM-barrel domain, and uroporphyrinogen decarboxylase, which is an enzyme involved in heme biosynthesis.

**Protein expression and purification**. The different proteins studied in this work were purified following standard procedures. Briefly, genes for the His-tagged proteins in a pET24b(+) vector with kanamycin resistance were cloned into *E. coli* BL21 (DE3) cells, and the proteins were purified by Ni-NTA affinity chromatography in HEPES buffer. The His tag was placed at the C-terminus, i.e., at a position that is well removed from the catalytic face of the barrel, the regions of enhanced conformational flexibility in the ancestral protein and the heme-binding site. Since the ancestral protein is susceptible to proteolysis, we included protease inhibitors in all steps of the purification (cOmplete® EDTA-free Protease Inhibitor Cocktail from Roche, ref. 11873580001). Protein solutions were prepared by exhaustive dialysis against the desired buffer (typically 50 mM HEPES pH 7) or by passage through PD10 columns. Protein purity was assessed by gel electrophoresis (Fig. S28). Proteins were properly folded, as judged by circular dichroism spectra (Fig. S29) [Far-UV CD spectra from 250 to 210 nm were recorded for extant and ancestral glycosidases, at 25 °C, using a Jasco J-715 spectropolarimeter equipped with a PTC-348WI. Buffer conditions were 50 mM HEPES, pH 7.0, protein concentration was within 0.2–0.6 mg/mL range and a 1 mm pathlength cuvette was used. An average of 30 scans was performed in each case. Blank subtraction was always carried out prior to mean residue ellipticity calculation, $[\Theta]_{MRW}$].

**Analytical ultracentrifugation**. Samples of the ancestral glycosidase in HEPES 50 mM, NaCl 150 mM, pH 7.0 were used. The assays were performed at 48,000 rpm (185,463 xg) in an XL-I analytical ultracentrifuge (Beckman-Coulter Inc.) equipped with both UV–VIS absorbance and Raleigh interference detection systems, using an An-50Ti rotor. Sedimentation profiles were recorded simultaneously by Raleigh interference and absorbance at 280 nm. Differential sedimentation coefficient distributions were calculated by least-squares boundary modeling of sedimentation velocity data using the continuous distribution c(s) Lamm equation model as implemented by SEDFIT 16.1c[57]. These experimental values were corrected to standard conditions using the program SEDNTERP[58] (version 20120111 Beta) to obtain the corresponding standard values (s20,w).

Sedimentation equilibrium assays (SE) for GH1-N72 were carried out at speeds ranging from 8,000 to 11,000 rpm (5,152 xg to 9,740 xg) and at 280 nm, using the same experimental conditions and instrument as in the SV experiments. A last high-speed run (48,000 rpm, 185,463 xg) was done to deplete protein from the meniscus region to obtain the corresponding baseline offsets. Weight-average buoyant MW of GH1-N72 were obtained by fitting a single-species model to the experimental data using the HeteroAnalysis 1.1.60 program[59] once corrected for temperature and solvent composition with the program SEDNTERP[58] (version 20120111 Beta).

**Preparation of the ancestral protein with bound heme and determination of the heme to protein ratio**. Stock solutions of hemin (heme with iron in the +3 oxidation state) were prepared daily in 1.4 M sodium hydroxide. Prior to use, the stock solution was diluted (typically 1:100) into HEPES buffer 50 mM, pH 7 and this solution was immediately used.

The ancestral protein with bound heme was prepared by incubating the protein with a 5-fold excess of heme for about one hour, followed by passage through a PD10 column and a Superdex-200 column to eliminate the non-bound heme. The heme to protein ratio in the resulting protein samples could be roughly estimated from the absorbance of the Soret band and the protein band at 280 nm in UV–VIS spectra. This procedure is not exact because the Soret band may depend on the interactions of the bound heme and, also, heme can show some absorption at 280 nm. For more accurate characterization protein concentration was determined by the bicinchoninic acid method[33] with the Pierce™ BCA Protein Assay Kit (ThermoFisher Scientific). A method based on pyridine hemochrome spectra[34] was used to determine the amount of heme. Briefly, 25 µl of 0.1 M potassium ferricyanide were added to a mixture of 2 mL or pyridine, 2 mL 0.1 M NaOH and 2 mL water. This solution was mixed with the protein solution in a 1:1 volume ratio and an excess of sodium dithionate was added. Lastly, the amount of heme was calculated from the absorbance of the pyridine hemochrome at 556 nm after correction for the absorbance of a blank. Using this approach, a heme to protein stoichiometry of 1.03±0.03 was determined from 5 independent measurements.

**UPLC mass spectrometry**. Ultra performance liquid chromatography (UPLC) was performed using a Waters Acquity H Class UPLC connected to a mass spectrometry Waters Synapt G2 Triwave® system. A 2.1 × 100 mm Protein BEH C4 column of 300 Å pore size and 2.1 µm particle size at a flow of 0.2 mL/min was used for chromatography. The mobile phase was a mixture of 0.1% formic acid–water (A) with 0.1% formic acid–acetonitrile (B) and the elution gradient were as follows: 0–10.33 min, 98–55% A; 10.33–20 min, 55–30% A; 20–21.57 min, 30% A; 21.57–23,33 min, 30–2% A; 23.33–30 min, stay 2% A. Mass spectrometry conditions were as follows: the ionization source of ESI was operated inion mode of positive (ESI + ) and 2.2 kV of capillary voltage. Temperature of desolvation was 400 °C, and ion source was 100 °C. Desolvent and cone gas (nitrogen) flow velocity were 600 L/h.

**Microscale thermophoresis quantification of heme-protein interaction**. The motion of molecules in microscopic temperature gradients (microscale thermophoresis) is sensitive to changes in properties induced by a binding event and can be used to quantify a diversity of intermolecular interactions[60]. For these experiments, we used a His-tagged protein labeled with a fluorescent probe using the His-tag labeling kit from NanoTemper technology. A 200 nanomolar protein solution was titrated at 25 °C with heme concentrations ranging from 55 nM to 2 µM. We did not use higher heme concentration to minimize the possibility of heme association, a process that would decrease the concentration of the monomeric heme that is competent for binding. The experiments were performed with Monolith NT.115 pico from NanoTemper technology. The data were acquired with MO. Control software, version 1.6 (NanoTemper Technologies GmbH). The binding curve and affinity were modeled and analyzed in the MO.Control software, version 1.6. Three replicate experiments were performed to yield an average value for the heme dissociation constant of 547 ± 110 nanomolar. Relevant plots and validation reports for the three experiments are shown in Fig. S14.

**Activity determinations**. Glucosidase and galactosidase activities were tested following the absorbance of p-nitrophenol at 405 nm upon the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside[27]. Rates were calculated from the initial absorbance vs. time slope and the known extinction coefficient of p-nitrophenol at pH 7. Experiments at different substrate concentrations were carried out to arrive at Michaelis plots for the ancestral and several modern glycosidases studied in this work. For the rate determination at a wide range of substrate concentrations we used a protocol designed to minimize any changes in buffer composition that could distort the profiles. Thus, for a rate measurement at a given substrate concentration, an enzyme solution in HEPES buffer at pH 7 was mixed with an equal volume of substrate dissolved in pure water. To minimize pH changes, the initial enzyme solution was prepared in 200 mM buffer to yield a final buffer concentration of 100 mM. We confirmed that the pH changes upon mixing were negligible. See legends to Figs. S6–S9 for details on data analysis.

As it is common in the literature, values of the optimum activity temperatures were determined from the profiles of activity versus temperature derived from measurements performed after several-minute incubations at each temperature[25]. Briefly, the protein was incubated at the desired temperature with 1 mM substrate in HEPES buffer 50 mM pH 7 and, after 10 min, the reaction was stopped by adding sodium carbonate to a concentration of 0.5 M. The amount of substrate hydrolyzed was determined from the absorbance of p-nitrophenol at 405 nm. We confirmed that the 10-min incubation only hydrolyzed a fraction of the substrate present and, therefore, that the amount of substrate hydrolyzed after a 10-min. incubation is a suitable metric of enzyme activity. Profiles of activity versus temperature were determined using both 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside. The profiles for the ancestral glycosidase are shown in Fig. 1b and those for the modern glycosidase from *Halothermothrix orenii* are given in Fig. S5. In all cases, the profiles show a sharp maximum from which an unambiguous determination of the optimum temperature is possible. Note also that there is good agreement between the optimum temperature values derived using the two different substrates. Differential scanning calorimetry experiments were performed as we have previously described in detail[9].

Glycosidase substrates were obtained from commercial sources, except 4-nitrophenyl-β-D-glucopyranoside-6-phosphate, which was prepared by us on the basis of a published procedure[61]. The chemical identity of the prepared compound was confirmed by mass spectrometry and nuclear magnetic resonance.

**Proteolysis experiments**. For proteolysis experiments, the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii* at a concentration of 1 mg/mL were incubated at 25 °C with thermolysin for different times in HEPES buffer 50 mM pH 7 containing 10 mM calcium chloride. Stock solutions of thermolysin were prepared fresh in the same solvent at a concentration of 1 mg/mL and were diluted 1:10 when added to the protein solution. The reaction was stopped by the addition of EDTA to a final concentration of 12.5 mM and aliquots were loaded into 15% (w/v) SDS-PAGE gels for electrophoresis. In some experiments, fragments separated by electrophoresis were extracted, desalted and subjected to LC-MS/MS analysis for mass determination. Fragment masses were determined by MALDI and their sequences were investigated using peptide mapping fingerprinting and MALDI-TOF/TOF (Fig. S2). This allowed us to locate approximately the cleavage sites, as shown in Fig. 2c.

**Crystallization and structure determination**. The ancestral glycosidase, dissolved in 150 mM NaCl, 50 mM HEPES pH 7.0, was concentrated to 35 mg/mL and to 70 mg/mL for the vapor-diffusion (VD) and counter-diffusion crystallization experiments, respectively. We checked by SDS electrophoresis that the concentrated protein used for crystallization was not proteolyzed. Hanging-drops VD experiments were prepared by mixing 1 µL of protein solution with the reservoir, in a 1:1 ratio, and equilibrated against 500 µL of each precipitant cocktail HR-I (Crystal Screen 1, Hampton Research). Capillary counter-diffusion experiments were set up in capillaries of 0.3 mm inner diameter using the CSK-24, AS-49 and PEG448-49 screening kits[62]. A similar procedure was followed for the crystallization of the ancestral glycosidase-heme complex, using two fixed concentrations at 75 and 30 mg/ml for the counter-diffusion and VD experiments. Experiments were performed at 293 K.

Crystals of the ancestral glycosidase were obtained only in condition #41 of HR-I, whilst the GH1N72-Heme complex crystallized in conditions #6 and #9 of HR-I and PPP8 of the mix of PEG counter-diffusion screen. Crystals were extracted either from the capillary or fished directly from the drop and subsequently cryoprotected by equilibration with 15 % (v/v) glycerol prepared in the mother liquid, flash-cooled in liquid nitrogen and stored until data collection. Crystals were diffracted at the XALOC beamline of the Spanish synchrotron light radiation source (ALBA, Barcelona). Indexed data were scaled and reduced using the CCP4 program suite[63].

Initial data sets were obtained for the ancestral glycosidase crystals diffracting the X-ray to 2.5 Å. The clean (without water, ligands, etc.) 3D model of the β-glucosidase from *Thermotoga maritima* (PDB ID. 2J78) [https://www.rcsb.org/structure/2J78] was used as search model for molecular replacement[63]. Two monomers were found in the asymmetric unit as expected from the Matthews coefficient for the P2(1) space group. Refinement, including Titration-Libration-Screw (TLS) parametrization, water pick, and model validation was carried out with PHENIX suite[64]. Unidentifiable amino acids in the highly disordered region have been assigned as poly-UKN chains C and D corresponding to the 18 and 14 (poly-Alanine) of chains A and B, respectively.

Crystals of the ancestral glycosidase-heme complex belong to the same space group than the ancestral glycosidase but were not isomorphous. The determined unit cell was bigger accommodating three polypeptide chains in the asymmetric as determined from the Matthews coefficient. A similar protocol was followed to place the three monomers in the unit cell by molecular replacement and to refine the structure. After a first refinement round the presence of one protoporphyrin ring in each polypeptide chain was determined. It was also clear that disordered regions of the ancestral glycosidase model were visible in the heme complex model.

The summary of data collection, refinement statistics, and quality indicators are collected in Table S9. The coordinates and the experimental structure factors have been deposited in the Protein Data Bank with ID 6Z1M [https://www.rcsb.org/structure/6Z1M] and 6Z1H [https://www.rcsb.org/structure/6Z1H] for the ancestral glycosidase with and without bound heme, respectively.

Figures displaying 3D-structures have been prepared using PyMOL (The Pymol Molecular Graphics System, Schrödinger, LLC). The 2D-interaction diagram of Fig. 7a was prepared using LigPlot+ (https://www.ebi.ac.uk/thornton-srv/software/LIGPLOT/).

**Molecular dynamics simulations**. Molecular simulations were performed on both ancestral glycosidases (this work, PDB ID: 6Z1M [https://www.rcsb.org/structure/6Z1M], 6Z1H [https://www.rcsb.org/structure/6Z1H]) and the modern glycosidase from *Halothermothrix orenii* (PDB ID: 4PTV [https://www.rcsb.org/structure/4PTV]). The structure of the heme-free ancestral glycosidase was obtained by manually deleting the heme coordinates from the corresponding heme-bound crystal structure. The missing regions of the ancestral glycosidases were reconstructed using MODELLER. Histidine protonation states were selected based on empirical $pK_a$ estimates performed using PROPKA 3.1 and visual inspection. All other residues were placed in their standard protonation states at physiological pH. The heme group was described using a bonded model, creating a bond between the Tyr264 side chain and the Fe(III) atom of the heme (Fig. S30). We used MCPB.py as implemented in AMBER19[65] to obtain the necessary parameters for creating the bonding pattern between the Fe(III) atom and the 4 nitrogen atoms of the heme and the tyrosine side chain oxygen. The resulting structure was then optimized, and frequency calculations were performed at the ωB97X-D/6-31 G* level of theory followed by the Seminario method[66] to obtain the force constants from the Hessian of the frequency calculation. This functional is a long-range (LC) corrected hybrid functional (see Chai and Head-Gordon[67] and references cited therein), which describes short-range interactions using an exchange functional, and long-range interactions using 100% Hartree-Fock exchange. ωB97X-D further improves on the concept of LC functionals, by systematic optimization of the functional, including the inclusion of an extra parameter that allows for an adjustable fraction of short-range exchange. In addition, this functional incorporates an empirical dispersion correction, following the DFT-D scheme[68]. We chose this functional for our parameterization as it yielded optimized structures of the heme complex that were similar to that observed in the heme complex, which is important in order to be able to maintain the heme in the binding pocket in a planar (non-distorted) conformation in our bonded model. This was further corroborated in our MD simulations, where the heme maintained a planar conformation without any unphysical distortions of dihedral angles. We note that only the heme and the Tyr264 side chain were considered as QM atoms in our model, as this is the region that was necessary to parameterize in our bonded model, following Li and Merz[69].

Partial charges were obtained for the heme and for the Tyr264 side chain at the HF/6-31 G* level of theory, using the restrained electrostatic potential (RESP) approach, following the MCPB.py protocol, and performing the calculations using Gaussian 09 Rev. E.01. Periodic boundary conditions (PBC) were used, and all systems were solvated in a truncated octahedral box filled with TIP3P water molecules[70], with 10 Å from the solute to the box edges in all directions. The truncated octahedron can fill space without leaving any gaps, and since our protein has a globular shape, a truncated octahedral box is the most suitable box shape to reduce the number of water molecules necessary to fill the box, which saves substantial computational time. with a distance of 10 Å from the solute to the surface of the box. The box was then filled with TIP3P water molecules. $Na^+$ and $Cl^-$ counter ions were added to the system to neutralize each enzyme. The protein was described using the AMBER ff14SB force field[71], and the heme was described using the General AMBER Force Field (GAFF)[72].

Following system preparation, the LEaP module of AMBER19 was used to generate the topology and coordinate files for the MD simulations, which were performed using the CUDA version of the PMEMD module of the AMBER19 simulation package. The solvated system was first subjected to a 5000 step steepest descent minimization, followed by a 5000 step conjugate gradient minimization with positional restraints on all heavy atoms of the solute, using a 5 kcal mol$^{-1}$ Å$^{-2}$ harmonic potential. The minimized system was then heated up to 300 K using the Berendsen thermostat[73], with a time constant of 1 ps for the coupling, and 5 kcal mol$^{-1}$ Å$^{-2}$ positional restraints (again a harmonic potential) applied during the heating process. The positional restraints were then gradually decreased to 1 kcal mol$^{-1}$ Å$^{-2}$ over five 500 ps steps of NPT equilibration, using the Berendsen thermostat and barostat to keep the system at 300 K and 1 atm. For the production runs, each system was subjected to either 500 ns of sampling in an NPT ensemble at constant temperature (300 K) and constant pressure (1 atm), controlled by the Langevin thermostat, with a collision frequency of 2.0 ps$^{-1}$, and the Berendsen barostat with a coupling constant of 1.0 ps. A 2 fs time step was used for all simulations, and snapshots were saved from the simulation every 5 ps. The SHAKE algorithm[74] was applied to constrain all bonds involving hydrogen atoms. A 10 Å cutoff was applied to all nonbonded interactions, with the electrostatic interactions being treated with the particle mesh Ewald (PME) approach[75]. 10 independent simulations were performed for each starting structure during 500 ns (for RMSD convergence, see Fig. S18). All the subsequent analyses were performed with the CPPTRAJ toolkit from Ambertools19[65]. Parameters used to describe the heme, input files, snapshots from our simulations, and simulation trajectories (with water molecules and ions removed to save file size) are available for download from Zenodo (https://zenodo.org) at https://doi.org/10.5281/zenodo.3857791.

**Reporting summary**. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Data included in the figures and supporting the findings of this study are available from the corresponding authors upon reasonable request. Atomic coordinates and the experimental structure factors have been deposited in the Protein Data Bank (https://www.rcsb.org) with ID 6Z1M and 6Z1H for the ancestral glycosidase with and without bound heme, respectively. Parameters used to describe the heme, input files, snapshots from our molecular dynamics simulations, and simulation trajectories (with water molecules and ions removed to save file size) are available for download from Zenodo (https://zenodo.org) at https://doi.org/10.5281/zenodo.3857791. Source data are provided with this paper.

Received: 3 June 2020; Accepted: 11 December 2020;
Published online: 15 January 2021

## References

1. Pauling, L. & Zuckerkandl, E. Chemical paleogenetics. Molecular “restoration studies” of extinct forms of life. *Acta Chem. Scan.* **17S**, 9–16 (1963).
2. Hochberg, G. K. A. & Thornton, J. W. Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).
3. Gumulya, Y. & Gillam, E. M. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the ‘retro’ approach to protein engineering. *Biochem. J.* **474**, 1–19 (2017).
4. Cole, M. F. & Gaucher, E. A. Exploiting models of molecular evolution to efficiently direct protein engineering. *J. Mol. Evol.* **72**, 193–203 (2011).
5. Risso, V. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Biotechnological and protein engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol.* **51**, 106–115 (2018).
6. Trudeau, D. L. & Tawfik, D. S. Protein engineers turned evolutionists—the quest for the optimal starting point. *Curr. Opin. Biotechnol.* **60**, 46–52 (2019).
7. Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).
8. Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167 (2009).
9. Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902 (2013).
10. Gardner, J. M., Biler, M., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. L. Manipulating conformational dynamics to repurpose ancient proteins for modern catalytic functions. *ACS Catal.* **10**, 4863–4870 (2020).
11. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).
12. Nagano, N., Orengo, C. A. & Thornton, J. M. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765 (2002).
13. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM barrel architecture facilitated the early evolution of protein-mediated metabolism. *J. Mol. Evol.* **82**, 17–26 (2016).
14. Grunwald, P. *Biocatalysis: Biochemical Fundamentals and Applications* 2nd edn. (World Scientific, New York, 2017).
15. Wolfenden, R., Lu, X. & Young, G. Spontaneous hydrolysis of glycosides. *J. Am. Chem. Soc.* **120**, 6814–6815 (1998).
16. Zechel, D. L. & Withers, S. G. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33**, 11–18 (2000).
17. Burke, H. M., Gunnlaugsson, T. & Scanian, E. M. Recent advances in the development of synthetic chemical probes for glycosidase enzymes. *Chem. Commun.* **51**, 10576–10588 (2015).
18. Lombard, V. et al. The Carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
19. CAZypedia Consortium. Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes. *Glycobiology* **28**, 3–8 (2018).
20. Ingles-Prieto, A. et al. Conservation of protein over four billion years. *Structure* **21**, 1–8 (2013).
21. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372**, 631–634 (1994).
22. Withers, S. Glycoside hydrolase family 1. CAZypedia, available at http://www.cazypedia.org/. Accessed 19 April 2020.
23. Weiss, C. W. et al. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
24. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
25. Wood, T. M. & Bhat, K. M. Methods for measuring cellulase activities. *Methods Enzymol.* **160**, 87–112 (1988).
26. Park, C., Zhou, S., Gilmore, J. & Marqusee, S. Energetics-based protein profiling on a proteomic scale: identification of proteins resistant to proteolysis. *J. Mol. Biol.* **368**, 1426–1437 (2007).
27. Hassan, N. et al. Biochemical and structural characterization of a thermostable β-glucosidase from *Halothermothrix orenii* for galacto-oligosaccharide synthesis. *Appl. Microbiol. Biotechnol.* **99**, 1731–1744 (2015).
28. Marana, S. R. Molecular basis of substrate specificity in family 1 glycoside hydrolases. *IUBMB Life* **58**, 63–73 (2006).
29. Devamani, T. et al. Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J. Am. Chem. Soc.* **138**, 1046–1056 (2016).
30. Vanderkooi, G. & Stotz, E. Reductive alteration of heme α hemochromes. *J. Biol. Chem.* **240**, 3418–3424 (1965).
31. Fiege, K., Querebillo, C. J., Hildebrandt, P. & Frankenberg-Dinkel, N. Improved method for the incorporation of heme cofactors into recombinant proteins using Escherichia coli Nissle 1917. *Biochemistry* **57**, 2747–2755 (2018).
32. Inada, Y. & Shibata, K. The Soret band of monomeric hematin and its changes on polymerization. *Biochem. Biophys. Res. Commun.* **9**, 323–327 (1962).
33. Smith, P. K. et al. Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150**, 76–85 (1985).
34. Berry, E. A. & Trumpower, B. L. Simultaneous determination of hemes a, b, and c from pyridine hemochrome spectra. *Anal. Biochem.* **161**, 1–15 (1987).
35. Smith, L. J., Kahramam, A. & Thornton, J. M. Heme proteins—diversity in structural characteristics, function, and folding. *Proteins* **78**, 2349–2368 (2010).
36. Schneider, S., Marles-Wright, J., Sharp, K. H. & Paoli, M. Diversity and conservation of interactions for binding heme in b-type heme proteins. *Nat. Prod. Rep.* **24**, 621–630 (2007).
37. Li, T., Bonkovsky, H. L. & Guo, J. Structural analysis of heme proteins: implications for design and prediction. *BMC Struct. Biol.* **11**, 13 (2011).
38. Wiita, A. P. et al. Probeing the chemistry of thioredoxin catalysis with force. *Nature* **450**, 124–127 (2007).
39. Sigala, P. A. et al. Testing geometrical discrimination within an enzyme active site: constrained hydrogen bonding in the ketosteroid isomerase oxyanion hole. *J. Am. Chem. Soc.* **130**, 13696–13798 (2008).
40. Lüdtke, S. et al. Sub-ångström-resolution crystallography reveals physical distortions that enhance reactivity of a covalent enzymatic intermediate. *Nat. Chem.* **5**, 762–767 (2013).
41. James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).
42. Bershtein, S. & Tawfik, D. S. Advances in laboratory evolution of enzymes. *Curr. Opin. Chem. Biol.* **12**, 151–158 (2008).
43. Petrović, D., Risso, V. A., Kamerlin, S. C. L. & Sanchez-Ruiz, J. M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **15**, 20180330 (2018).
44. Pabis, A., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. L. Cooperativity and flexibility in enzyme evolution. *Curr. Opin. Struct. Biol.* **48**, 83–92 (2018).
45. Risso, V. A. et al. *De novo* active sites for resurrected Precambrian enzymes. *Nat. Commun.* **8**, 16113 (2017).
46. Höcker, B., Beismann-Driemeyer, S., Heltwer, S., Lustig, A. & Sterner, R. Dissection of a $(\beta\alpha)_8$-barrel enzyme into two folded halves. *Nat. Struct. Biol.* **8**, 32–36 (2001).
47. Gamiz-Arco, G. et al. Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding. *Biochem. J.* **476**, 3631–3647 (2019).
48. Randall, R. N., Radford, C. E., Roof, K. A., Natarajan, D. K. & Gaucher, E. A. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.* **7**, 12847 (2016).
49. Holm, L. DALI and the persistence of protein shape. *Protein Sci.* **29**, 128–140 (2020).
50. Chu, X.-Y. & Zhang, H.-Y. Cofactors as molecular fossils to trace the origin and evolution of proteins. *ChemBioChem* https://doi.org/10.1002/cbic.202000027 (2020).

51. Chen, K. & Arnold, F. H. Engineering new catalytic activities in enzymes. *Nat. Catal.* **3**, 203–213 (2020).
52. Dawson, N. L. et al. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acid Res.* **45**, D289–D295 (2017).
53. Naganathan, A. N. Modulation of allosteric coupling by mutations: from protein dynamics and packing to altered native ensembles and function. *Curr. Opin. Struct. Biol.* **54**, 1–9 (2019).
54. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
55. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
56. Ashkenazy, H. et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* **40**, W580–W584 (2012).
57. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606–1619 (2000).
58. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. in *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (eds Harding, S. E., Rowe, A. J. & Horton, J. C.) 90–125 (Royal Society of Chemistry, Cambridge, 1992).
59. Cole, J. L. Analysis of heterogeneous interactions. *Methods Enzymol.* **384**, 212–232 (2004).
60. Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P. & Duhr, S. Molecular interactions studies using microscale thermophoresis. *Assay. Drug Dev. Technol.* **9**, 342–353 (2011).
61. Acebrón, I. et al. Structural basis of the substrate specificity and instability in solution of a glycosidase from *Lactobacillus plantarum*. *BBA—Proteins Proteom.* **1865**, 1227–1236 (2017).
62. González-Ramírez, L. A. et al. Efficient screening methodology for protein crystallization based on the counter-diffusion technique. *Cryst. Growth Des.* **17**, 6780–6786 (2017).
63. Collaborative, C. P. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **D 50**, 760–763 (1994).
64. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr.* **D 66**, 213–221 (2010).
65. Case, D. A. et al. *AMBER 2019.* (University of California, San Francisco, 2019).
66. Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **60**, 1271–1277 (1996).
67. Chai, J.-D. & Head-Godon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
68. Sato, T., Tsuneda, T. & Hirao, K. Long-range corrected density functional study on weakly bound systems: Balanced descriptions of various types of molecular interactions. *J. Chem. Phys.* **126**, 234114 (2007).
69. Li, P. & Merz, K. M. MCPB.py: a python based metal center parameter builder. *J. Chem. Inf. Model* **56**, 599–604 (2016).
70. Jorgensen, W. L., Chandrasenkar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
71. Maier, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters From ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
72. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
73. Beredsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
74. Ryckaert, J. P., Cicotti, G. & Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
75. Darden, T., York, D. & Pedersein, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

## Acknowledgements

This work was supported by Human Frontier Science Program Grant RGP0041 (J.M.S.-R., E.A.G., B.S., and S.C.L.K.), NIH grant R01AR069137 (E.A.G.), Department of Defense grant MURI W911NF-16-1-0372 (E.A.G.), the Swedish Research Council (2019-03499) (S.C.L.K.), the Knut and Alice Wallenberg Foundation (2018.0140 and 2019.0431) (S.C.L.K.), Spanish Ministry of Economy and Competitiveness/FEDER Funds Grants BIO2015-66426-R (J.M.S.-R.) RTI2018-097142-B-100 (J.M.S.-R.) and BIO2016-74875-P (J.A.G.). The simulations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX partially funded by the Swedish Research Council through grant agreement no. 2016-07213. We acknowledge the Spanish Synchrotron Radiation Facility (ALBA, Barcelona) for the provision of synchrotron radiation facilities and the staff at XALOC beamline for their invaluable support. We are also grateful to Victoria Longobardo Polanco (Proteomic Unit, Institute of Parasitology and Biomedicine "López-Neyra") for help with mass spectrometry experiments and data analyses and to Juan Román Luque Ortega (Molecular Interactions Facility, Centro de Investigaciones Biológicas Margarita Salas) for help with ultracentrifugation experiments and data analyses.

## Author contributions

B.S., S.C.L.K., E.A.G., and J.M.S.-R. designed the research. G.G.-A. and L.I.G.-R. prepared the protein variants and designed, performed and analyzed experiments addressed at determining their catalytic and biophysical features, under the supervision of V.A.-R. and B.I.-M., who also provided essential input regarding the interpretation of these properties. V.A.R. was in charge of mass spectrometry, ultracentrifugation, and thermophoresis experiments. Y.H. carried out ancestral sequence reconstruction under the supervision of E.A.G., who also provided essential input for the interpretation of the results in an evolutionary context. D.P. performed homology modeling under the supervision of S.C.L.K. Organic synthesis was performed by J.J. and J.M.C. who provided essential input regarding the properties of the synthesized compound. A.R.-R. carried out MD simulations under the supervision of S.C.L.K., and they provided the general interpretation and implications of the simulations. L.I.G.-R. and V.A.R. carried out protein crystallization. J.A.G. determined the X-ray structures and provided essential input regarding their interpretation and implications. J.M.S.-R. wrote the first draft of the manuscript to which B.S., S.C.L.K., and E.A.G. added crucial paragraphs and sections. All authors discussed the manuscript, suggested modifications and improvements, and contributed to the final version.

## Funding

Open Access funding provided by Uppsala University.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41467-020-20630-1.

**Correspondence** and requests for materials should be addressed to S.C.L.K., E.A.G. or J.M.S.-R.

**Peer review information** *Nature Communications* thanks Vickery Arcus, John Mitchell, Matilda Newton, and other, anonymous, reviewers for their contributions to the peer review of this work. Peer review reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# SUPPORTING INFORMATION

## Heme-Binding Enables Allosteric Modulation in an Ancient TIM-Barrel Glycosidase

# SUPPLEMENTARY INFORMATION FOR

## Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase

*Gloria Gamiz-Arco[1,7], Luis I. Gutierrez-Rus[1,7], Valeria A. Risso[1], Beatriz Ibarra-Molero[1], Yosuke Hoshino[2], Dušan Petrović[3,8], Jose Justicia[4], Juan Manuel Cuerva[4], Adrian Romero-Rivera[3], Burckhard Seelig[5], Jose A. Gavira[6], Shina C.L. Kamerlin[3]*, Eric A. Gaucher[2]*, Jose M. Sanchez-Ruiz[1]**

[1]Departamento de Quimica Fisica. Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain.

[2]Department of Biology, Georgia State University, Atlanta, GA 30306 U.S.A.

[3]Science for Life Laboratory, Department of Chemistry-BMC, Uppsala University, BMC Box 576, S-751 23 Uppsala, Sweden.

[4]Departamento de Quimica Organica. Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain.

[5]Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, Minnesota, United States of America, & BioTechnology Institute, University of Minnesota, St. Paul, Minnesota, United States of America.

[6]Laboratorio de Estudios Cristalograficos, Instituto Andaluz de Ciencias de la Tierra, CSIC, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, Avenida de las Palmeras 4, Granada 18100 Armilla, Spain.

[7]These authors contributed equally to this work

[8]Current address: Hit Discovery, Discovery Sciences, Biopharmaceutical R&D, AstraZeneca, 431 50 Gothenburg, Sweden.

*email: lynn.kamerlin@kemi.uu.se or egaucher@gsu.edu or sanchezr@ugr.es

## Supplementary Methods

### Homology model preparation

For the reconstructed ancestral sequence for node 72, 3D structures were initially prepared using homology modelling. In searching for the suitable templates, several criteria were imposed, i.e., high sequence identity[1,2] (>= 50%) and query sequence coverage (>= 90%), as determined by BLAST and HHBlits[1,2], as well as the oligomerization state of a monomer and belonging to the GH1 family in the CAZY database. The following PDB structures fulfilled the criteria, and were used to create homology models with SWISS-MODEL[3-5]: family 1 β-glucosidase from *Thermotoga maritima* (1W3J), β-glucosidase A from *Clostridium cellulovorans* (3AHX), engineered β-glucosidase from soil metagenome (3CMJ), GH1 β-glucosidase Td2F2 (3WH5), and β-glucosidase 1A from *Thermotoga neapolitana* (5IDI). The quality of the five homology models was assessed using the Swiss-Model provided scores and the structural parameters estimated with MolProbity[6] and the models from 1W3J, 3CMJ, and 5IDI templates were selected for the further work. While the protein core (the $(\beta\alpha)_8$ barrel) was generally modelled very well, major differences could be observed between the homology models in the region of the catalytic loops.

## Supplementary Tables

**Table S1.** Sequence identity of the ancestral glycosidase with the modern glycosidases in the set used as starting point for ancestral sequence reconstruction. The sequence identity values with the ancestral protein span the 0.26-0.59 range.

| | |
|---|---|
| Ancestral | 1,000 |
| Euk-XP_002676353.1.Naegleria.gruberi.NEG-M.Oth | 0,483 |
| Euk-OLQ05614.1.Symbiodinium.microadriaticum.SAR | 0,463 |
| Euk-OCF22674.1.Kwoniella.bestiolae.CBS.10118.Fg | 0,439 |
| Euk-KXS17974.1.Gonapodya.prolifera.JEL478.Fg.435 | 0,366 |
| Euk-KOO32978.1.Chrysochromulina.CCMP291.Oth.458 | 0,431 |
| Euk-KNC76787.1.Sphaeroforma.arctica.JP610.Met.475 | 0,389 |
| Euk-GAX29444.1.Fistulifera.solaris.SAR.446 | 0,519 |
| Euk-GAX16517.1.Fistulifera.solaris.SAR.450 | 0,500 |
| Euk-ETP15771.1.Phytophthora.parasitica.CJ01A1.SAR | 0,364 |
| Euk-ETP11111.1.Phytophthora.parasitica.CJ01A1.SAR | 0,516 |
| Euk-ETP02644.1.Phytophthora.parasitica.CJ01A1.SAR | 0,366 |
| Euk-EOD40853.1.Emiliania.huxleyi.CCMP1516.Oth.401 | 0,249 |
| Euk-EOD13795.1.Emiliania.huxleyi.CCMP1516.Oth.457 | 0,432 |
| Euk-EIE19776.1.Coccomyxa.subellipsoidea.C-169.Arpl | 0,271 |
| Euk-CHR-CAC34952.1.Piromyces.E2.Fg.463 | 0,393 |
| Euk-CHR-CAC08178.1.Homo.sapiens.Met.452 | 0,419 |
| Euk-CHR-BAO04178.1.Delphinium.grandiflorum.Arpl | 0,418 |
| Euk-CHR-BAE87009.1.Phanerochaete.chrysosporium.Fg | 0,464 |
| Euk-CHR-BAE63197.1.Aspergillus.oryzae.RIB40.Fg.460 | 0,440 |
| Euk-CHR-AGS32242.1.Coptotermes.gestroi.Met.460 | 0,424 |
| Euk-CHR-AAP13852.1.Bombyx.mori.Met.462 | 0,398 |
| Euk-CHR-AAL92115.1.Glycine.max.Arpl.456 | 0,377 |
| Euk-CHR-AAL37719.1.Solanum.lycopersicum.Arpl.461 | 0,435 |
| Euk-CHR-AAL25999.1.Brevicoryne.brassicae.Met.453 | 0,419 |
| Euk-CHR-AAK62412.1.Arabidopsis.thaliana.Arpl.465 | 0,404 |
| Euk-CHR-AAK32833.1.Arabidopsis.thaliana.Arpl.462 | 0,392 |
| Euk-CBJ30694.1.Ectocarpus.siliculosus.SAR.450 | 0,489 |
| Euk-AAG46031.1.Leishmania.infantum.Exc.456 | 0,371 |
| CPR-OYX53756.1.Sacchari.32-50-13.Oth.420 | 0,303 |
| CPR-OHA99561.1.Zambryski.R-O2_12_FULL_43_12b.Pc | 0,345 |
| CPR-OHA32072.1.Taylor.R-O2_01_FULL_45_15b.Pc | 0,340 |
| CPR-OHA13338.1.Taga.R-O2_01_FULL_39_11.Pc.405 | 0,317 |
| CPR-OHA09246.1.Sung.R-O2_01_FULL_59_16.Pc.399 | 0,311 |
| CPR-OGZ46118.1.Ryan.R-O2_01_FULL_48_27.Pc.416 | 0,330 |
| CPR-OGZ07021.1.Lloyd.R-O2_02_FULL_50_13.Pc.385 | 0,328 |

| | |
|---|---|
| CPR-OGZ06701.1.Lloyd.R-O2_02_FULL_50_11.Pc.387 | 0,332 |
| CPR-OGZ02593.1.Lipton.R-O2_01_FULL_53_13.Pc.396 | 0,315 |
| CPR-OGY99744.1.Lipton.R-O2_01_FULL_52_25.Pc.404 | 0,320 |
| CPR-OGY99381.1.Lipton.R-O2_01_FULL_52_25.Pc.410 | 0,305 |
| CPR-OGY71081.1.Jackson.R-O2_01_FULL_44_13.Pc.395 | 0,300 |
| CPR-OGY40781.1.Brenner.RIFOXYD1_FULL_41_16.Pc.389 | 0,312 |
| CPR-OGM99997.1.Yanofsky.R-O2_01_FULL_41_53.Pc.416 | 0,322 |
| CPR-OGM98994.1.Yanofsky.R-O2_01_FULL_41_26.Pc.386 | 0,314 |
| CPR-OGM90535.1.Wolfe.R-O2_01_FULL_38_11.Pc.412 | 0,336 |
| CPR-OGM00912.1.Uhr.RIFOXYC2_FULL_47_19.Pc.406 | 0,334 |
| CPR-OGL39085.1.Sacchari.R-O2_02_FULL_46_7.Oth | 0,311 |
| CPR-OGL30033.1.Sacchari.R-O2_02_FULL_47_12.Oth | 0,308 |
| CPR-OGK55604.1.Roizman.R-O2_01_FULL_45_11.Mc | 0,325 |
| CPR-OGJ70814.1.Peri.R-O2_12_FULL_53_10.Oth.390 | 0,318 |
| CPR-OGG92310.1.Kuenen.R-O2_12_FULL_42_13.Pc.406 | 0,316 |
| CPR-OGG87996.1.Kaiser.RIFOXYD1_FULL_42_15.Pc.382 | 0,312 |
| CPR-OGG55752.1.Kaiser.R-O2_01_FULL_55_37.Pc.387 | 0,317 |
| CPR-OGG06799.1.Gottesman.R-O2_01_FULL_42_12.Mc.399 | 0,326 |
| CPR-OGF90942.1.Giovannoni.R-O2_02_FULL_45_14.Pc | 0,351 |
| CPR-OGE98567.1.Doudna.R-O2_12_FULL_42_9.Pc.394 | 0,348 |
| CPR-OGE83632.1.Doudna.R-O2_02_FULL_43_13b.Pc.393 | 0,322 |
| CPR-OGE24950.1.Davies.R-O2_02_FULL_39_12.Mc.398 | 0,316 |
| CPR-OGE15085.1.Davies.GWA1_38_6.Mc.403 | 0,336 |
| CPR-KUK76511.1.WS6.34_10.Oth.396 | 0,338 |
| CPR-KKS47395.1.Giovannoni.GW2011_GWF2_42_19.Pc.397 | 0,351 |
| CPR-KKS25793.1.Jorgensen.GW2011_GWF2_41_8.Pc.408 | 0,332 |
| CPR-KKQ84990.1.Woese.GW2011_GWB1_38_8.Mc.399 | 0,347 |
| CPR-KKQ36955.1.Woese.GW2011_GWA1_37_7.Mc.396 | 0,353 |
| CPR-KKQ29214.1.Nomura.GW2011_GWA1_37_20.Pc.379 | 0,365 |
| CPR-KKQ12362.1.Moran.GW2011_GWF1_36_78.Pc.403 | 0,333 |
| Bac-WP_081838630.1.Thermogem.carboxidivorans.Clf | 0,511 |
| Bac-WP_053225736.1.Solirubrobacter.soli.Ac.440 | 0,483 |
| Bac-WP_033274371.1.Actinospica.acidiphila.Ac.393 | 0,313 |
| Bac-WP_028863797.1.Psychromonas.aquimarina.gP.448 | 0,415 |
| Bac-WP_026735568.1.Fischerella.PCC.9605.Cy.448 | 0,467 |
| Bac-WP_026389483.1.Acholeplasma.multilocale.Tn.438 | 0,336 |
| Bac-WP_016873414.1.Chlorogloeopsis.fritschii.Cy | 0,470 |
| Bac-WP_010471747.1.Acaryochloris.CCMEE.5410.Cy.441 | 0,488 |
| Bac-SEK57728.1.Rhodococcus.maanshanensis.Ac.387 | 0,300 |
| Bac-OUQ12786.1.Massiliomicrobiota.An142.Fm.454 | 0,449 |
| Bac-OUQ09514.1.Massiliomicrobiota.An142.Fm.466 | 0,340 |
| Bac-OIO58018.1.CG1_02_48_14.Mn.415 | 0,349 |

| | |
|---|---|
| Bac-OGS21660.1.RIFOXYA2_FULL_39_19.El.449 | 0,500 |
| Bac-OGF51021.1.RIFOXYA2_FULL_40_8.Frs.449 | 0,457 |
| Bac-OAA30738.1.Kosmotoga.arenicorallina.S304.Tg | 0,300 |
| Bac-KPM53585.1.Frankia.R43.Ac.388 | 0,322 |
| Bac-KIX85561.1.JCVI.TM6SC1.Dp.402 | 0,294 |
| Bac-GAT31492.1.Terrimicrobium.sacchariphilum.V.449 | 0,503 |
| Bac-EGF89970.1.Asticcacaulis.biprosthecum.C19.aP | 0,293 |
| Bac-EFH82146.1.Ktedonobacter.racemifer.Clf | 0,503 |
| Bac-CRX38655.1.Estrella.lausannensis.Chl.421 | 0,269 |
| Bac-CRH90323.1.Chlamydia.trachomatis.Chl.449 | 0,401 |
| Bac-CHR-CAC47470.1.Sinorhizobium.meliloti.1021.aP | 0,483 |
| Bac-CHR-BAC96154.1.Vibrio.vulnificus.YJ016.gP.445 | 0,481 |
| Bac-CHR-BAB37196.1.Escherichia.coli.Sakai.gP.464 | 0,366 |
| Bac-CHR-BAB36995.1.Escherichia.coli.Sakai.gP.462 | 0,346 |
| Bac-CHR-AIY95585.1.Bacillus.subtilis.Fm | 0,353 |
| Bac-CHR-AIY95329.1.Bacillus.subtilis.Fm | 0,365 |
| Bac-CHR-AIY91871.1.Bacillus.subtilis.Fm | 0,437 |
| Bac-CHR-AIY91623.1.Bacillus.subtilis.Fm | 0,473 |
| Bac-CHR-AGA60135.1.Microbacterium.Gsoil167.Ac.398 | 0,288 |
| Bac-CHR-AEI42200.1.Paenibacillus.mucilaginosus.Fm | 0,296 |
| Bac-CHR-ADD27066.1.Meiothermus.ruber.DSM.1279.DT | 0,524 |
| Bac-CHR-ADD01617.1.Thermoanaerobacter.italicus.Fm | 0,453 |
| Bac-CHR-ACV58907.1.Alicyclobaci.acidocaldarius.Fm | 0,551 |
| Bac-CHR-ACO44852.1.Deinococcus.deserti.VCD115.DT | 0,497 |
| Bac-CHR-ACM06095.1.Thermomicrobium.roseum.Clf | 0,521 |
| Bac-CHR-ACL70277.1.Halothermothrix.orenii.H.168.Fm | 0,557 |
| Bac-CHR-ACI19973.1.Dictyoglomus.thermophilum.Dg | 0,581 |
| Bac-CHR-ABU56651.1.Roseiflexus.castenholzii.Clf | 0,541 |
| Bac-CHR-ABR73190.1.Marinomonas.MWYL1.gP.444 | 0,449 |
| Bac-CHR-ABJ60960.1.Lactobacillus.gasseri.Fm | 0,340 |
| Bac-CHR-ABJ59900.1.Lactobacillus.gasseri.Fm | 0,494 |
| Bac-CHR-ABJ59596.1.Lactobacillus.gasseri.Fm | 0,362 |
| Bac-CHR-ABD82858.1.Saccharophagus.degradans.gP | 0,497 |
| Bac-CHR-ABD80656.1.Saccharophagus.degradans.gP | 0,443 |
| Bac-CHR-AAN58797.1.Streptococcus.mutans.UA159.Fm | 0,346 |
| Bac-CHR-AAK78365.1.Clostridium.acetobutylicum.Fm | 0,509 |
| Bac-CHR-AAG59862.1.Sphingomonas.paucimobilis.aP | 0,261 |
| Bac-CHR-AAB49339.1.Fusobacterium.mortiferum.Fs.442 | 0,525 |
| Bac-CHR-AAA24815.1.Dickeya.chrysanthemi.gP.456 | 0,365 |
| Bac-CCB88561.1.Simkania.negevensis.Z.Chl.422 | 0,270 |
| Bac-CAB95278.1.Streptomyces.coelicolor.A3_2.Ac.444 | 0,526 |
| Bac-BAY28878.1.Nostoc.carneum.NIES-2107.Cy.452 | 0,471 |

| | |
|---|---|
| Bac-BAM04503.1.Phycisphaera.mikurensis.Pl | 0,464 |
| Bac-BAL98072.1.Caldilinea.aerophila.DSM.14535.Clf | 0,522 |
| Bac-APF18099.1.Caldithrix.abyssi.DSM.13497.Cld.414 | 0,313 |
| Bac-AFG36462.1.Spirochaeta.africana.DSM.8902.Sp | 0,331 |
| Bac-AEP25088.1.Thermotoga.maritima.MSB8.Tg.443 | 0,593 |
| Bac-ADD01635.1.Thermoanaerobacter.italicus.Ab9.Fm | 0,578 |
| Bac-ADB52696.1.Conexibacter.woesei.DSM.14684.Ac | 0,496 |
| Bac-ACZ10042.1.Sebaldella.termitidis.ATCC.33386.Fs | 0,390 |
| Bac-ACZ09962.1.Sebaldella.termitidis.ATCC.33386.Fs | 0,399 |
| Bac-ACL69240.1.Halothermothrix.orenii.H.168.Fm.417 | 0,339 |
| Bac-ACI21065.1.Thermodesulfovib.yellowstonii.Nsr | 0,336 |
| Bac-ABJ83756.1.Solibacter.usitatus.Ellin6076.Ad | 0,305 |
| Bac-ABG04991.1.Rubrobacter.xylanophilus.Ac | 0,524 |
| Bac-AAK79373.1.Clostridium.acetobutylicum.Fm | 0,359 |
| Arc-SMD30897.1.Picrophilus.oshimae.Ery-Thp | 0,256 |
| Arc-Seed-BAA29440.1.Pyrococcus.horikoshii.Ery-Thc | 0,377 |
| Arc-OYT35865.1.Archaeoglobales.ex4484_92.Ery-Ach | 0,334 |
| Arc-OWP55065.1.Cuniculiplasma.C_DKE.Ery-Thp.470 | 0,263 |
| Arc-KYH41278.1.Bathyarchaeota.B26-2.Ery-Unc.477 | 0,278 |
| Arc-KJE49247.1.Acidiplasma.MBA-1.Ery-Thp.464 | 0,237 |
| Arc-CHR-ADL19795.1.Acidilo.saccharovorans.TK-Thp | 0,311 |
| Arc-CHR-ABW01492.1.Caldivir.maquilingensis.TK-Thp | 0,295 |
| Arc-CHR-AAL81332.1.Pyrococcus.furiosus.Ery-Thc | 0,294 |
| Arc-CHR-AAL80566.1.Pyrococcus.furiosus.Ery-Thc | 0,392 |
| Arc-CHR-AAL80197.1.Pyrococcus.furiosus.Ery-Thc | 0,275 |
| Arc-CHR-AAK43121.1.Sulfolobus.solfataricus.TK-Thp | 0,290 |
| Arc-CHR-AAD43138.1.Thermosphaera.aggregans.TK-Thp | 0,272 |
| Arc-BAB59827.1.Thermoplasma.volcanium.GSS1.Ery-Thp | 0,265 |
| Arc-AJB42198.1.Thermofilum.carboxyditrophus.TK-Thp | 0,285 |
| Arc-AJB41496.1.Thermofilum.carboxyditrophus.TK-Thp | 0,277 |
| Arc-AIF16120.1.marine.thaumarchaeote.TK-Thm | 0,345 |

**Table S2**. Catalytic parameters for the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside at pH 7 (HEPES buffer 100 mM) and 25 ºC catalyzed by modern and ancestral family 1 glycosidases. For each enzyme/substrate combination, the independent experimental replicates (involving at least two enzyme preparations) were performed. Values of catalytic parameters derived from the fitting of the Michaelis-Menten equation (See Figures S6-S9) are given for each replicate. Errors are standard deviations derived from the fits. The average values of the three replicates are given in Table S3.

| | 4-nitrophenyl-β-D-glucopyranoside | | | 4-nitrophenyl-β-D-galactopyranoside | | |
|---|---|---|---|---|---|---|
| | $k_{cat}$ ($s^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ ($s^{-1}$ $mM^{-1}$) | $k_{cat}$ ($s^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ ($s^{-1}$ $mM^{-1}$) |
| Ancestral glycosidase | 0.1739 ± 0.0074 | 6.190 ± 0.673 | 0.0281 ± 0.0020 | 0.1159 ± 0.0066 | 33.61 ± 3.75 | 0.00345 ± 0.00020 |
| | 0.1543 ± 0.0054 | 5.701 ± 0.580 | 0.0271 ± 0.0019 | 0.1233 ± 0.0087 | 31.74 ± 4.48 | 0.00388 ± 0.00029 |
| | 0.1578 ± 0.0071 | 4.76 ± 0.58 | 0.0331 ± 0.0028 | 0.1215 ± 0.0045 | 32.76 ± 2.39 | 0.00371 ± 0.00014 |
| Ancestral with bound heme | 0.521 ± 0.022 | 17.18 ± 1.31 | 0.0303 ± 0.0011 | 0.2801 ± 0.0056 | 21.57 ± 1.03 | 0.01298 ± 0.00037 |
| | 0.443 ± 0.034 | 16.18 ± 2.25 | 0.0274 ± 0.0018 | 0.2269 ± 0.0065 | 12.55 ± 1.15 | 0.0181 ± 0.0012 |
| | 0.414 ± 0.032 | 18.65 ± 2.40 | 0.0222 ± 0.0012 | 0.2321 ± 0.0074 | 16.89 ± 1.42 | 0.01374 ± 0.00075 |
| *Halothermothrix orenii* | 17.21 ± 0.25 | 0.3248 ± 0.022 | 52.99 ± 3.11 | 94.8 ± 4.3 | 26.14 ± 2.59 | 3.63 ± 0.20 |
| | 16.41 ± 0.42 | 0.342 ± 0.039 | 48.03 ± 4.71 | 62.54 ± 2.67 | 14.45 ± 1.85 | 4.33 ± 0.38 |
| | 13.285 ± 0.587 | 0.399 ± 0.071 | 33.33 ± 4.83 | 64.59 ± 1.25 | 14.40 ± 0.78 | 4.49 ± 0.16 |
| | 18.44 ± 0.27 | 0.367 ± 0.025 | 50.18 ± 2.84 | 25.99 ± 0.75 | 9.70 ± 0.94 | 2.68 ± 0.19 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *Thermotoga maritima* | 17.92 ± 0.16 | 0.348± 0.015 | 51.55 ± 1.82 | 30.77 ± 1.66 | 15.44 ± 2.41 | 1.99 ± 0.21 |
| | 17.77 ± 0.24 | 0.366 ± 0.022 | 48.48 ± 2.45 | 28.41 ± 0.73 | 10.36 ± 0.87 | 2.74 ± 0.17 |
| *Marinomonas sp.* (strain MWYL1) | 10.54 ± 0.22 | 0.095 ± 0.015 | 110.15 ± 16.20 | 28.51 ± 0.80 | 14.68 ± 1.14 | 1.94 ± 0.10 |
| | 10.38 ± 0.20 | 0.089 ± 0.013 | 116.43 ± 15.10 | 22.26 ± 0.70 | 13.20 ± 1.21 | 1.69 ± 0.11 |
| | 10.62 ± 0.41 | 0.112± 0.030 | 94.47 ± 22.80 | 24.96 ± 0.57 | 12.19 ± 0.89 | 2.05 ± 0.11 |
| *Saccharophagus degradans* (strain 2-40$^T$) | 25.19 ± 0.34 | 0.710 ± 0.053 | 35.46 ± 2.34 | 5.87 ± 0.29 | 22.20 ± 2.53 | 0.264 ± 0.018 |
| | 20.55 ± 0.30 | 0.545 ± 0.042 | 37.71 ± 2.58 | 7.21 ± 0.62 | 32.87 ± 5.46 | 0.219 ± 0.019 |
| | 21.09 ± 0.52 | 0.549 ± 0.070 | 38.43 ± 4.33 | 6.02 ± 0.20 | 22.20 ± 1.68 | 0.271 ± 0.012 |

**Table S3.** Catalytic parameters for the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside at pH 7 (HEPES buffer 100 mM) and 25 ºC catalyzed by modern and ancestral family 1 glycosidases. The values shown here are the average values, together with the corresponding standard deviations, of three independent replicates (see Table S2). n = 3 independent determinations of the Michaelis-Menten profiles.

| | 4-nitrophenyl β-D-glucopyranoside | | | 4-nitrophenyl β-D-galactopyranoside | | |
|---|---|---|---|---|---|---|
| | $k_{cat}$ ($s^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ ($s^{-1}$ $mM^{-1}$) | $k_{cat}$ ($s^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ ($s^{-1}$ $mM^{-1}$) |
| Ancestral | 0.1620 ± 0.0085 | 5.55 ± 0.59 | 0.0294 ± 0.0026 | 0.1202 ± 0.0032 | 32.70 ± 0.76 | 0.00368 ± 0.00018 |
| Ancestral with bound heme | 0.459 ± 0.045 | 17.3 ± 1.0 | 0.0266 ± 0.0036 | 0.246 ±0.024 | 17.0 ± 3.7 | 0.0149 ± 0.0023 |
| *Halothermothrix orenii* | 15.6 ± 1.7 | 0.355 ± 0.032 | 44.8 ± 8.4 | 74 ± 15 | 18.3 ± 5.5 | 4.15 ± 0.37 |
| *Thermotoga maritima* | 18.04 ± 0.29 | 0.3603 ± 0.0087 | 50.1 ± 1.3 | 28.4 ± 2.0 | 11.8 ± 2.6 | 2.47 ± 0.34 |
| *Marinomonas sp.* (strain MWYL1) | 10.51 ± 0.10 | 0.0980 ± 0.0097 | 107.0 ± 9.2 | 25.2 ± 2.6 | 13.4 ± 1.0 | 1.89 ± 0.15 |
| *Saccharophagus degradans* (strain 2-40) | 22.3 ± 2.1 | 0.601 ± 0.077 | 37.2 ± 1.3 | 6.37 ± 0.60 | 25.8 ± 5.0 | 0.251 ± 0.023 |

**Table S4.** Amino acid residues at critical active site positions in modern and ancestral glycosidases. The catalytic carboxylic acids, as well as the positions involved in the binding of the glycone and aglycone moieties of the substrate[7] are shown. Residues at those positions in the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii* are given. The last column provides the residue statistics for the set of modern glycosidases used as starting point for ancestral sequence reconstruction. Glycosidases are known to be somewhat specific for the glycone moiety of the substrate and much less specific for the aglycone moiety, which is reflected in a lower residue conservation at the protein residues involved in aglycone binding. See Figure S10 for a graphical illustration.

| | Ancestral glycosidase | Modern glycosidase from *Halothermothrix orenii* | Sequence statistics in the set of modern glycosidases used as a starting point for ancestral sequence reconstruction. The residue present in the ancestral protein is highlighted in bold |
|---|---|---|---|
| Catalytic residues | E171 | E166 | **E 98%** Q 0.7% R 0.7% S 0.7% |
| | E358 | E354 | **E 100%** |
| Residues involved in binding the glycone part of the substrate | Q25 | Q20 | **Q 98.7%** H 0.7% P 0.7% |
| | H126 | H121 | **H 97.3%** W 1.3% D 0.7% N 0.7% |
| | N170 | N165 | **N 99.3%** C 0.7% |
| | W404 | W401 | **W 98.7%** M 0.7% R 0.7% |
| | E411 | E408 | **E 84.7%** S 14.7% R 0.7% |
| | W412 | W409 | **W 86.7%** A 5.3% F 2.7% L 1.3% T 1.3% I 0.7% M 0.7% S 0.7% V 0.7% |
| Residues involved in binding the aglycone part of the substrate | W40 | W35 | **W 85.3%** S 2.7% V 2.7% A 2.0% I 2.0% F 1.3% L 1.3% M 1.3% C 0.7% T 0.7% |
| | F43 | F38 | **F 36%** W 36% Y 6.7% L 5.3% A 4.7% M 2.7% E 1.3% G 1.3% K 1.3% Q 1.3% V 1.3% I 0.7% S 0.7% T 0.7% |
| | W127 | W122 | **W 50.7%** F 39.3% Y 7.3% G 0.7% H 0.7% L 0.7% S 0.7% |
| | F175 | V170 | Y 21.3% V 18.0% **F 17.3%** S 12.0% P 6.0% I 4.7% A 4.0% M 3.3% L 2.7% Q 2.7% T 2.7% Q 2.7% T 2.7% W 2.7% N 2% C 0.7% |
| | L178 | E173 | **L 30.7%** M 10.0% G 8.0% N 8.0% A 6.7% Q 6.7% H 4.7% F 4.0% E 3.3% C 2.7% K 2.7% S 2.7% Y 2.0% D 1.3% P 1.3% R 1.3% T 1.3% V 1.3% W 0.7% |
| | H185 | H180 | **H 26.7%** W 21.3% F 20.7% L 5.3% Y 4.0% K 3.3% M 2.7% S 2.7% G 2.0% I 2.0% Q 2.0% R 2.0% A 1.3% D 1.3% V 1.3% E 0.7% N 0.7% |
| | N227 | N222 | **N 48.7%** A 16.0% H 8.0% S 6.0% D 4.7% L 3.3% I 2.0% Q 2.0% T 2.0% V 2.0% F 1.3% Y 1.3% C 0.7% E 0.7% G 0.7% R 0.7% |

| | | | |
|---|---|---|---|
| | W332 | W327 | **W 86%** Y 6.7% E 1.3% L 1.3% F 0.7% G 0.7% H 0.7% I 0.7% R 0.7% S 0.7% V 0.7% |
| | A413 | A410 | **A 40.7%** S 8.0% E 6.7% L 6.7% T 6.7% D 5.3% G 5.3% H 4.0% I 3.3% N 3.3% R 2.0% V 2.0% F 1.3% P 1.3% Q 1.3% C 0.7% K 0.7% M 0.7% |
| | F420 | F417 | **F 80.0%** Y 18.0% L 1.3% R 0.7% |

**Table S5.** Estimated rates of the hydrolysis of several substrates catalysed by the ancestral glycosidase at node 72. All substrates were assayed at concentration of 0.1 mM, 25 ºC, HEPES buffer 50 mM pH 7 with a concentration of enzyme of 4-5 μM. Rates were calculated from the time dependence of the absorbance due to the released 4-nitrophenolate after correction for the blank. A value of ~0 is reported when no significant rate enhancement over the blank was detected.

| **Substrate** | **Rate $s^{-1}$** |
|---|---|
| *4*-nitrophenyl-β-D-glucopyranoside | 1.1E-02 |
| *4*-nitrophenyl-β-D-galactopyranoside | 2.4E-03 |
| *4*-nitrophenyl-β-D-fucopyranoside | 3.4E-02 |
| *4*-nitrophenyl-β-D-mannopyranoside | 1.7E-04 |
| *4*-nitrophenyl-β-D-xylopyranoside | 7.9E-05 |
| *4*-nitrophenyl-β-D-glucuronide | 1.0E-04 |
| *4*-nitrophenyl-β-D-ribofuranoside | 3.1E-04 |
| *4*-nitrophenyl-β-D-thioglucopyranoside | ~0 |
| *4*-nitrophenyl-β-D-cellobioside | 1.3E-04 |
| *4*-nitrophenyl-β-D-lactopyranoside | 5.2E-05 |
| *4*-nitrophenyl-β-D-maltoside | 4.7E-05 |
| *4*-nitrophenyl-β-D-galactopyranoside 6P | ~0 |
| *4*-nitrophenyl-N-acetil-β-D-glucosamine | ~0 |
| *4*-nitrophenyl-N-acetil-β-D galactosamine | ~0 |
| *4*-nitrophenyl-α-D-maltoside | ~0 |
| *4*-nitrophenyl-α-L-fucopyranoside | ~0 |
| *4*-nitrophenyl-α-D-maltohexaoside | ~0 |
| *4*-nitrophenyl-α-D-galactopyranoside | ~0 |
| *4*-nitrophenyl-α-D-glucopyranoside | ~0 |
| *4*-nitrophenyl-α-D-xylopyranoside | ~0 |
| *4*-nitrophenyl-α-D-mannopyranoside | ~0 |
| *4*-nitrophenyl-α-D-rhamnopyranoside | ~0 |

**Table S6.** Statistics of amino acid occurrence in modern family 1 glycosidases at the positions involved in interactions with the heme in the ancestral glycosidase. The set of sequences used as starting point for ancestral reconstruction has been used for this calculation. The number of occurrences for the predicted ancestral residues are highlighted in bold. Note that, in all cases, the ancestral residue is the most common residue (i.e., the consensus residue) in the modern set. Still, in all positions other amino acid residues are also observed.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **N173** | 5 | 2 | **42** | 4 | - | 2 | 8 | 8 | - | 7 | 7 | 1 | 13 | 6 | - | 4 | 7 | 29 | 4 | 1 |
| **A199** | **37** | 7 | 6 | - | - | 5 | 2 | 14 | - | 8 | 15 | 4 | 10 | 7 | - | 10 | 7 | - | 14 | 4 |
| **L202** | 6 | - | - | - | - | 13 | 5 | - | - | 22 | **62** | 1 | 16 | 4 | - | 2 | 7 | - | - | 12 |
| **L203** | 22 | 1 | 4 | - | - | 1 | - | - | 1 | 32 | **62** | - | 7 | 10 | - | 2 | - | - | 1 | 7 |
| **H206** | - | - | 2 | - | - | 2 | - | - | **132** | 1 | - | - | - | - | - | 13 | - | - | - | - |
| **I224** | 4 | - | 3 | - | 9 | - | - | 1 | - | **72** | 18 | - | 6 | 6 | 5 | 4 | 2 | 5 | - | 15 |
| **L226** | 1 | - | 11 | 2 | - | 1 | 3 | - | 19 | 16 | **46** | 21 | 3 | 7 | 6 | 2 | - | - | 10 | 2 |
| **L228** | 9 | - | 17 | 1 | 6 | 1 | - | 13 | 1 | 13 | **27** | 1 | 11 | 15 | 2 | 8 | 4 | 1 | 7 | 13 |
| **F251** | 3 | 10 | 5 | 4 | - | 2 | 3 | - | 2 | 12 | 21 | 1 | 2 | **55** | - | 1 | 3 | 7 | 10 | 7 |
| **N252** | 6 | 10 | **11** | 3 | - | 11 | 8 | 5 | 11 | 3 | 5 | 4 | 2 | 17 | - | 11 | 9 | 13 | 5 | 4 |
| **F255** | 2 | - | - | - | 2 | - | - | - | 1 | 10 | 3 | - | 1 | **91** | - | 1 | - | 8 | 14 | 5 |
| **L256** | 13 | 1 | - | 2 | - | - | 1 | 7 | - | 16 | **57** | - | 8 | 11 | 4 | 5 | 4 | 2 | 4 | 4 |
| **K261** | 2 | 20 | 6 | 2 | 2 | - | 3 | 3 | - | 1 | 5 | **23** | - | 6 | - | 4 | 11 | 1 | 5 | - |
| **Y264** | - | - | - | - | - | - | - | - | - | 4 | 7 | 1 | 1 | 9 | - | 1 | - | - | **75** | 1 |
| **L296** | 7 | - | 8 | 6 | 5 | 12 | - | 2 | 8 | 16 | **32** | - | 2 | 6 | 4 | 12 | 6 | - | 8 | 16 |
| **S295** | 3 | 7 | 1 | 6 | 3 | 7 | 2 | 7 | 6 | 2 | 3 | 20 | - | 1 | 26 | **22** | 20 | - | 14 | - |
| **R345** | 6 | **26** | 6 | 16 | - | 5 | 20 | 3 | 7 | 2 | 4 | 11 | 3 | 5 | - | 1 | 4 | 15 | 14 | 2 |
| **Y350** | 8 | 4 | 7 | 3 | 2 | 3 | - | 11 | 3 | - | 8 | 3 | 1 | 10 | 4 | 4 | 6 | - | **63** | 1 |

**Table S7**. Statistics of number of amino acid differences between modern family 1 glycosidases and the ancestral glycosidases at the positions involved in heme binding in the latter. The set of sequences used as starting point for ancestral reconstruction has been used for this calculation (see Table S6). Note that all the modern sequences differ from the ancestral sequence in a significant number of positions.

| number of identical amino acids | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sequences with identical amino acids | - | 2 | 17 | 22 | 26 | 14 | 25 | 18 | 13 | 4 | 6 | 3 | - | - | - | - | - | - |

**Table S8.** Atomic surface area values ($\mathrm{\AA}^2$) for heme bound to the ancestral glycosidase and the bound heme upon mutating to alanine in silico residues that block its access to the active site (Pro172, Asn173, Ile224, Leu226, Asn227 and Pro 272). As reference, the values for free heme are given in the last column.

| Atom | Ancestral | Ancestral with mutations to alanine | Free heme |
|---|---|---|---|
| NB | 0 | 4.9 | 4.9 |
| ND | 0 | 1.9 | 4.8 |
| C1 | 0 | 4.2 | 6.6 |
| C1 | 0 | 3.7 | 6.3 |
| C1 | 0 | 3.8 | 6.5 |
| C1 | 0 | 3.2 | 6.7 |
| C2 | 0 | 2.8 | 3 |
| C2 | 0.1 | 2.6 | 5.5 |
| C2 | 0 | 2.6 | 6 |
| C2 | 0 | 1.7 | 4.7 |
| C3 | 0 | 2.7 | 4.8 |
| C3 | 0 | 0 | 4.5 |
| C3 | 0 | 2.6 | 5.5 |
| C3 | 0 | 0.2 | 3.2 |
| C4 | 0 | 3.3 | 6.5 |
| C4 | 0 | 2.9 | 6.4 |
| C4 | 0 | 5.9 | 6.6 |
| C4 | 0 | 2.5 | 6 |
| CA | 0 | 20.6 | 20.6 |
| CA | 0 | 2.6 | 24.5 |
| CA | 5.6 | 12.8 | 25.2 |
| CA | 0.1 | 0.1 | 16.8 |
| CB | 0.1 | 4.8 | 20.4 |
| CB | 0.2 | 0.2 | 58.6 |
| CB | 10.7 | 20.9 | 58.5 |
| CB | 0 | 17 | 20.6 |
| CG | 0 | 2.1 | 7.8 |
| CG | 0 | 5.5 | 8.2 |
| CH | 0 | 3.9 | 8.4 |
| CH | 0 | 11.3 | 15.1 |
| CH | 0 | 5.7 | 12 |
| CH | 0.4 | 6.2 | 13.6 |
| CM | 0 | 37.3 | 54.6 |
| CM | 0 | 2.6 | 54 |
| CM | 0 | 4 | 52.4 |
| CM | 2.5 | 3.1 | 52.8 |
| NA | 0 | 4.2 | 5.1 |
| NC | 0 | 4.4 | 5.7 |
| O1 | 10.8 | 39.9 | 52.1 |
| O1 | 4.8 | 21 | 43.8 |
| O2 | 4.6 | 15.7 | 37.3 |
| O2 | 3.1 | 16.9 | 50 |
| FE | 0 | 5.2 | 5.2 |

**Table S9.** Data collection and refinement statistics (values in parentheses are for highest-resolution shell).

| Protein | ancestral | ancestral-heme |
|---|---|---|
| PDB ID | 6Z1H | 6Z1M |
| Space group | P 21 | P 21 |
| Unit cell | | |
| a, b, c (Å) | 52.26 80.67 97.81 | 58.93 89.49 141.12 |
| β (°) | 100.06 | 94.211 |
| ASU | 2 | 3 |
| Resolution (Å) * | 43.38 - 2.5 (2.59 - 2.5) | 52.87 - 2.45 (2.54 - 2.45) |
| $R_{merge}$ (%)* | 8.90 (105.30) | 80.96 (80.85) |
| $I/\sigma_I$ * | 10.9 (1.7) | 10.15 (1.44) |
| Completeness (%)* | 99.92 (99.93) | 99.02 (99.57) |
| Unique reflections* | 27829 (2775) | 53447 (5335) |
| Multiplicity | 7.4 (7.5) | 3.2 (3.2) |
| Wilson B-factor | 59.93 | 50.34 |
| CC(1/2) * | 0.999 (0.707) | 0.997 (0.718) |
| **Refinement** | | |
| $R_{work}/R_{free}$ (%) | 19.59 / 24.10 | 17.55 / 22.09 |
| No. atoms | 6601 | 11026 |
| Protein | 6546 | 10676 |
| Ligands | 21 | 194 |
| Solvent | 34 | 156 |
| B-factor (Å$^2$) | 76.63 | 62.81 |
| R.m.s deviations | | |
| Bond lengths (Å) | 0.003 | 0.005 |
| Bond angles ($^0$) | 0.65 | 0.73 |
| Ramachandran (%) | | |
| Favored | 96.09 | 96.14 |
| Outliers | 0.13 | 0 |

*Statistics for the highest-resolution shell are shown in parentheses.

## Supplementary FIgures

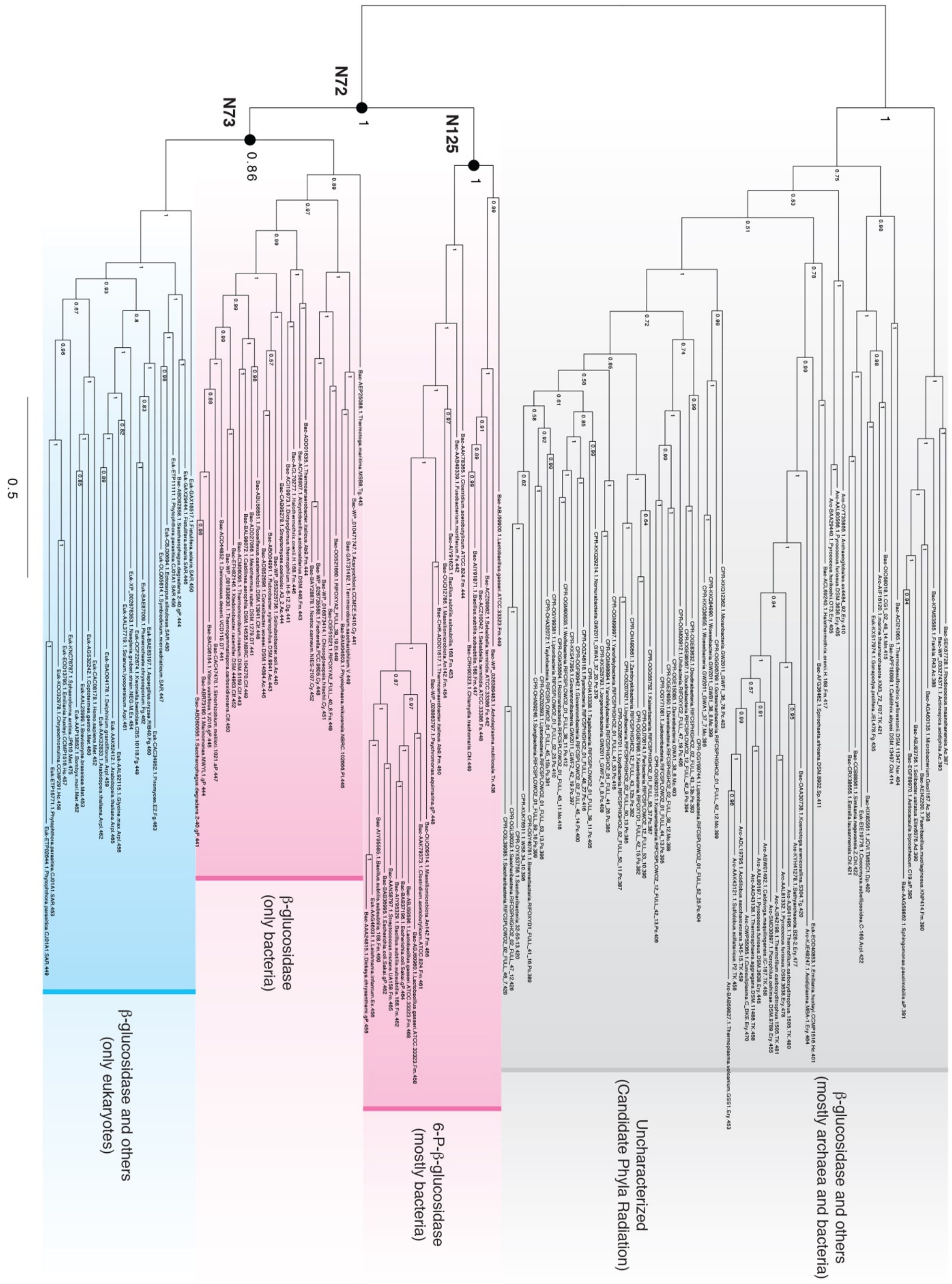


**Figure S1.** Bayesian analysis of family 1 glycosidases (GH1) protein sequences with sequence annotations. The annotation includes the accession number, the taxonomical information (domain name, phylum name and species name) and the sequence length. Three black dots indicate three ASR nodes (N72, N73 and N125). Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. Abbreviations: Ac =

Actinobacteria; aP = $\alpha$-Proteobacteria; Arc = Archaea; Arpl = Archaeplastida; Asg = Asgard group; Bac = Bacteria; Bc = Bacteroidetes; bP = $\beta$-Proteobacteria; Chl = Chlamydiae; Cld = Calditrichaeota (Caldithrix); Clf = Cloroflexi; CPR = Candidate Phyla Radiation; Cy = Cyanobacteria; Euk = Eukaryotes; Ex = Excavates; Fg = Fungi; Fm = Firmicutes; Frs = Fraserbacteria; Dg = Dictyoglomi; dP = $\delta$-Proteobacteria; Dp = Dependentiae (TM6); DPN = DPANN group; DT = Deinococcus-Thermus; Ery = Euryarchaeota; Fs = Fusobacteria; gP = $\gamma$-Proteobacteria; Hc = Hacrobia; Mc = Microgenomates; Met = Metazoa; Mn = Marinimicrobia; Nsr = Nitrospirae; Pc = Parcubacteria; Pl = Planctomycetes; V = Verrucomicrobia; SAR = SAR group; Sp = Spirochaetes; Tg = Thermotogae; TK = TACK group; Tn = Tenericutes.

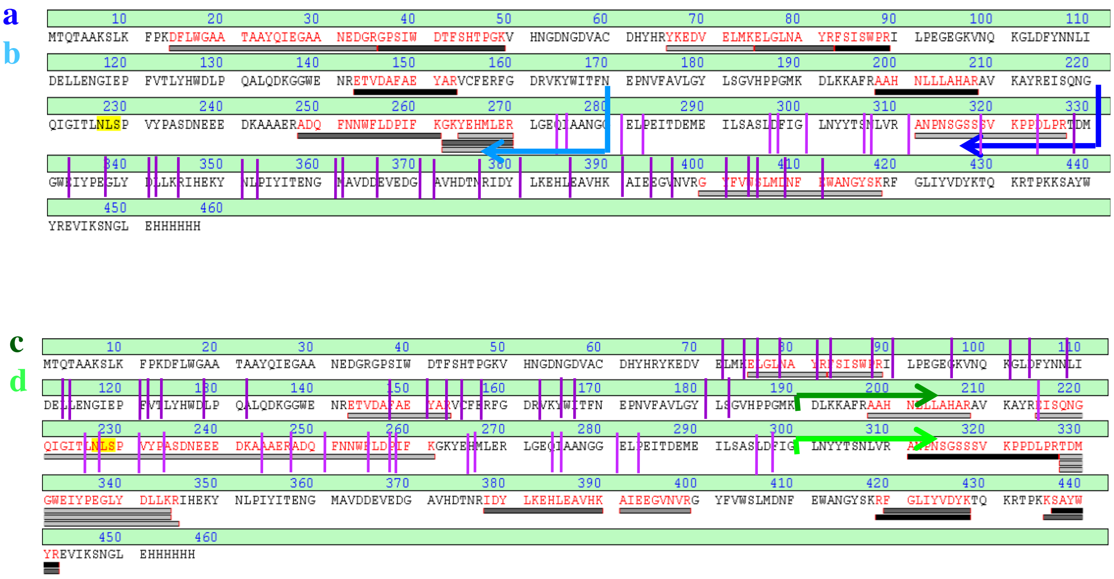

a
b
10 20 30 40 50 60 70 80 90 100 110
MTQTAAKSLK FPKDFLWGAA TAAYQIEGAA NEDGRGPSIW DTFSHTPGKV HNGDNGDVAC DHYHRYKEDV ELMKELGLNA YRFSISWPRI LPEGEGKVNQ KGLDFYNNLI
120 130 140 150 160 170 180 190 200 210 220
DELLENGIEP FVTLYHWDLP QALQDKGGWE NRETVDAFAE YARVCFERFG DRVKYWITFN EPNVFAVLGY LSGVHPPGMK DLKKAFRAAH NLLLAHARAV KAYREISQNG
230 240 250 260 270 280 290 300 310 320 330
QIGITLNLSP VYPASDNEEE DKAAAERADQ FNNWFLDPIF KGKYEHMLER LGEQIAANGG ELPEITDEME ILSASLDFIG LNYYTSNLVR ANPNSGSSSV KPPDLPRTDM
340 350 360 370 380 390 400 410 420 430 440
GWEIYPEGLY DLLKRIHEKY NLPIYITENG MAVDDEVEDG AVHDTNRIDY LKEHLEAVHK AIEEGVNVRG YFVWSLMDNF EWANGYSKRF GLIYVDYKTQ KRTPKKSAYW
450 460
YREVIKSNGL EHHHHHH
c
d
10 20 30 40 50 60 70 80 90 100 110
MTQTAAKSLK FPKDFLWGAA TAAYQIEGAA NEDGRGPSIW DTFSHTPGKV HNGDNGDVAC DHYHRYKEDV ELMKELGLNA YRFSISWPRI LPEGEGKVNQ KGLDFYNNLI
120 130 140 150 160 170 180 190 200 210 220
DELLENGIEP FVTLYHWDLP QALQDKGGWE NRETVDAFAE YARVCFERFG DRVKYWITFN EPNVFAVLGY LSGVHPPGMK DLKKAFRAAH NLLLAHARAV KAYREISQNG
230 240 250 260 270 280 290 300 310 320 330
QIGITLNLSP VYPASDNEEE DKAAAERADQ FNNWFLDPIF KGKYEHMLER LGEQIAANGG ELPEITDEME ILSASLDFIG LNYYTSNLVR ANPNSGSSSV KPPDLPRTDM
340 350 360 370 380 390 400 410 420 430 440
GWEIYPEGLY DLLKRIHEKY NLPIYITENG MAVDDEVEDG AVHDTNRIDY LKEHLEAVHK AIEEGVNVRG YFVWSLMDNF EWANGYSKRF GLIYVDYKTQ KRTPKKSAYW
450 460
YREVIKSNGL EHHHHHH

**Figure S2.** Estimation of the location of thermolysin cleavage sites from mass spectrometry and peptide-mapping fingerprinting. Potential thermolysin restriction sites are shown by vertical purple lines. Four thermolysin fragments were studied (a, b, c, and d: see left for color code). Fragment masses were determined by MALDI and their sequences were investigated using peptide mapping finger-printing and MALDI-TOF/TOF. Sequences for several sub-fragments (shown) could be thus determined and the length of the original fragments could be assessed. Fragments a and b extend approximately from the amino terminus to the dark and light blue arrows in the upper panel. Fragments c and d extend approximately from the dark and light green arrows in the lower panel to the carboxyl terminus. Comparison with the restriction sites allows a determination of the plausible thermolysin cleavage sites, as shown in Figure 2C of the main text.

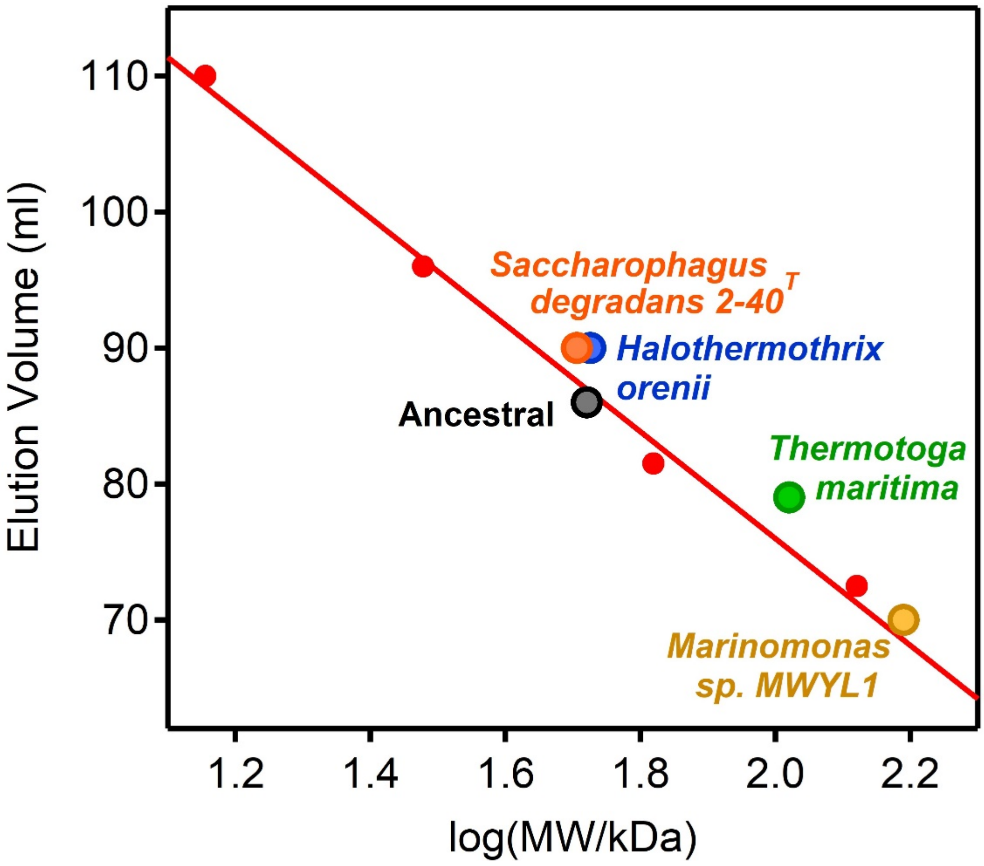


**Figure S3.** Assessment of association state of modern and ancestral glycosidases through gel filtration chromatography (HiLoad 16/600 Superdex 200 pg GE Healthcare). The molecular mass (MW) was estimated by the calibration curve of elution volume vs. log (MW). The protein markers used were: bovine serum albumin (monomer: 66 kDa, dimer: 132 kDa), deoxyribonuclease I from bovine pancreas (30.1 kDa) and lysozyme (14.3 kDa). The ancestral glycosidase and the modern glycosidases from *Saccharophagus degradans and Halothermothrix orenii* are monomers. The modern glycosidases from *Thermotoga maritima* is a dimer and the modern glycosidase from *Marinomonas sp. MWTL1* is a trimer.

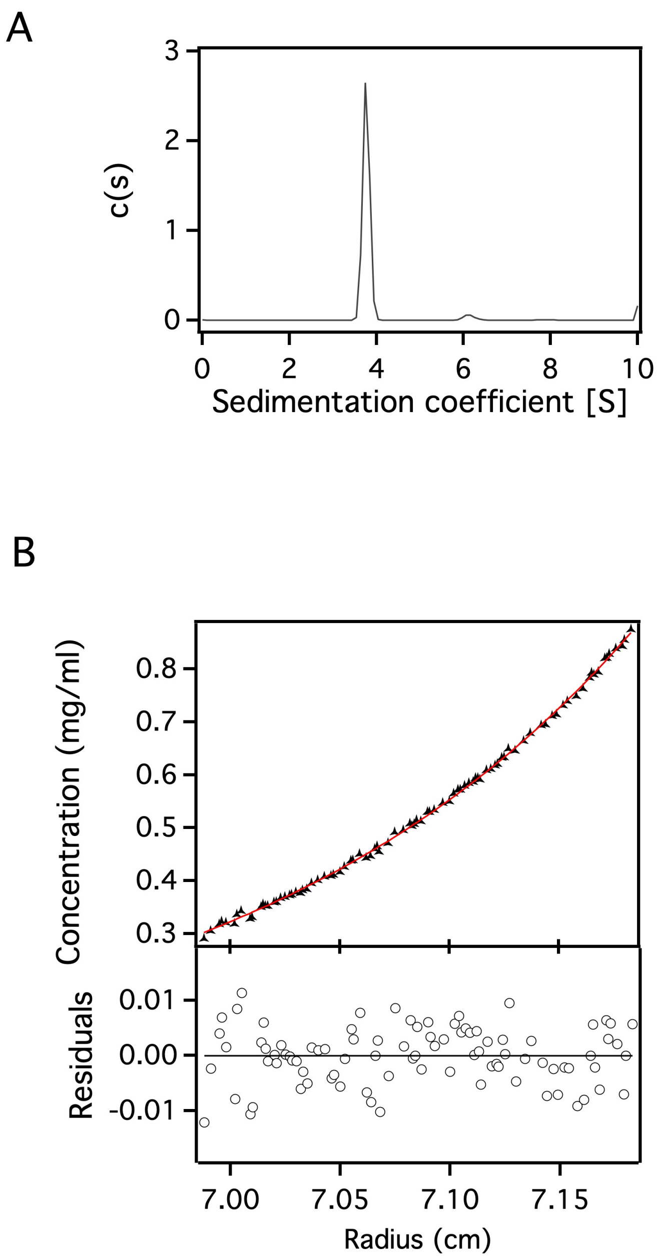


**Figure S4.** Assessing the association state of the ancestral glycosidase through analytical ultracentrifugation. A) Sedimentation velocity assay showing the sedimentation coefficient distribution c(s) corresponding to 0.24 mg/ml of purified protein. Peak at 3.7S is compatible with a globular monomer with the theoretical mass derived from the sequence. B) Sedimentation equilibrium assay. Upper panel: concentration gradient of experimental data (triangles) are presented together with best-fit analysis assuming protein monomer (red line). Lower panel: Difference between experimental data and estimated values for a protein monomer model (residuals in mg/mL). A molecular mass of 52900±192 Da is obtained, which is sufficiently close to the theoretical monomer mass calculated from the sequence (52542.79 Da) to rule out the dimeric and higher association states.

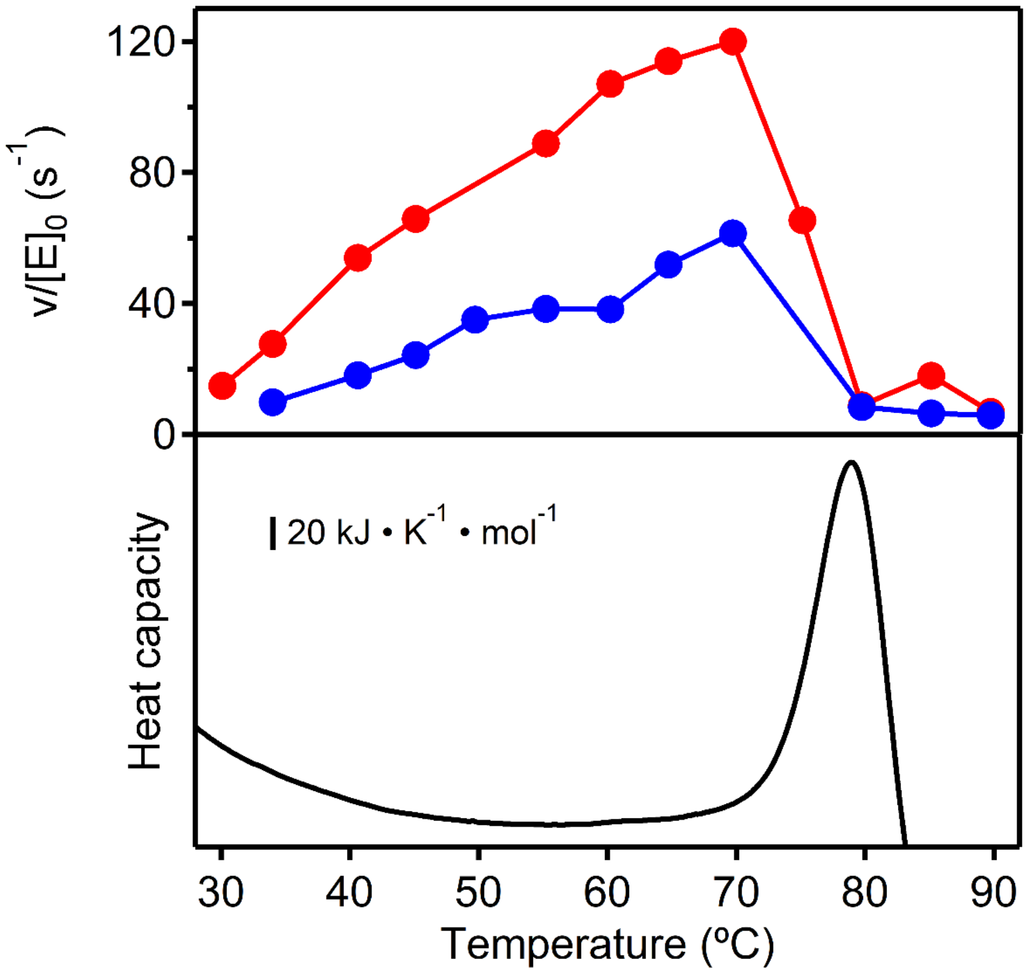


**Figure S5.** Determination of the optimum temperature for the modern glycosidase from *Halothermothrix orenii* using two different substrates 4-nitrophenyl-β-D-glucopyranoside (red) and 4-nitrophenyl-β-D-galactopyranoside (blue). The lower panel shows a differential scanning calorimetry profile for the enzyme under the same buffer conditions. Clearly, the activity drop observed at high temperature (upper panel) corresponds to the denaturation of the protein, as seen in the lower panel.

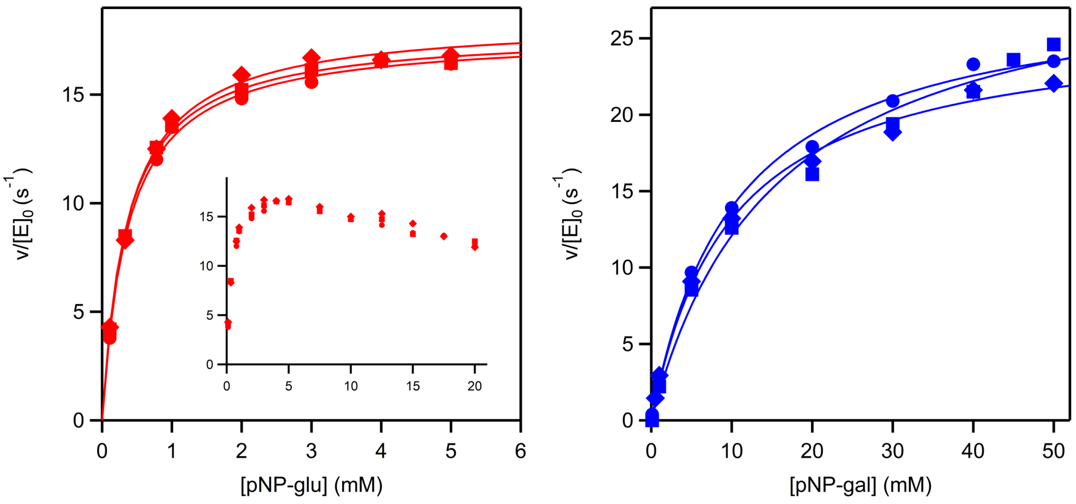


**Figure S6**. Michaelis plots of rate versus substrate concentration at pH 7 and 25 ºC for the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside catalyzed by the modern glycosidase from *Thermotoga maritima.* The different data points correspond to the 3 experimental replicates performed for each substrate, involving two different protein preparations in each case. The lines are the best fits of the Michaelis-Menten equation (see Tables S3 and S4 for the values derived from the fits). It is well known that glycosidase catalysis often shows kinetic complexities at high substrate concentrations, due to phenomena such as transglycosylation, inhibition by substrate or allosteric activation[8]. As a result of these complexities, Michaelis-Menten saturation kinetics are sometimes not observed. Here Michaelis-Menten saturation kinetics is not observed for 4-nitrophenyl-β-D-glucopyranoside in a wide concentration range (see inset in panel at the left). Therefore, only the data up to 6 mM have been used for the determination of the catalytic parameters from the fitting of the Michaelis-Menten equation.

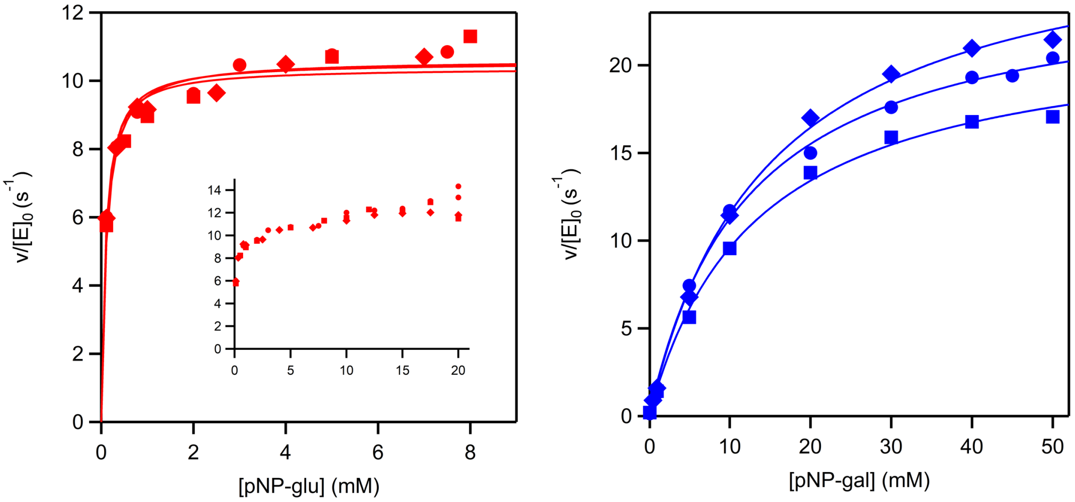


**Figure S7**. Michaelis plots of rate versus substrate concentration at pH 7 and 25 ºC for the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside catalyzed by the modern glycosidase from *Marinomonas sp.* (strain MWYL1). The different data points correspond to the 3 experimental replicates performed for each substrate, involving two different protein preparations in each case. The lines are the best fits of the Michaelis-Menten equation (see Tables S2 and S3 for the values derived from the fits). It is well known that glycosidase catalysis often shows kinetic complexities at high substrate concentrations, due to phenomena such as transglycosylation, inhibition by substrate or allosteric activation[8]. Here Michaelis-Menten saturation kinetics is not for 4-nitrophenyl-β-D-glucopyranoside in a wide concentration range (see inset in panel at the left). Therefore, only the data up to 8 mM have been used for the determination of the catalytic parameters from the fitting of the Michaelis-Menten equation.

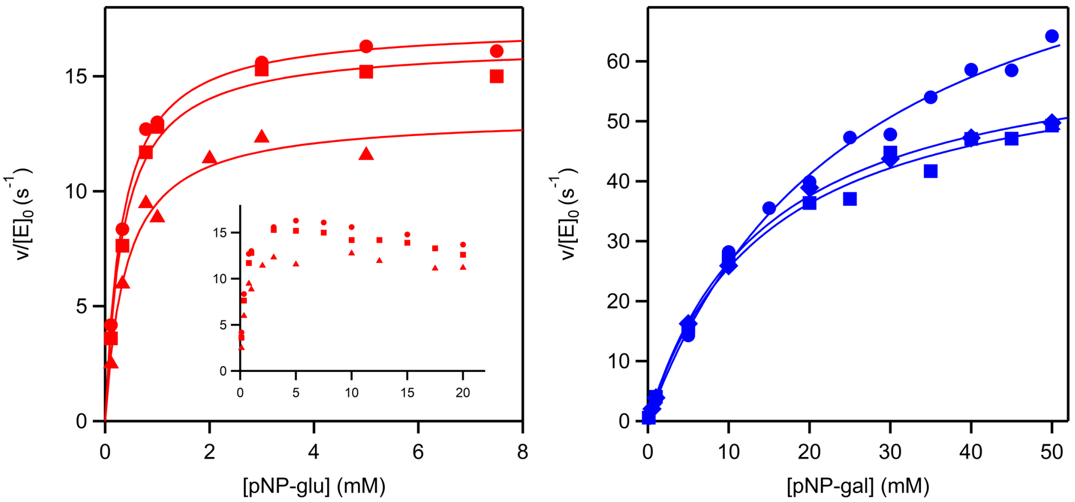


**Figure S8**. Michaelis plots of rate versus substrate concentration at pH 7 and 25 ºC for the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside catalyzed by the modern glycosidase from *Halothermothrix orenii.* The different data points correspond to the 3 experimental replicates performed for each substrate, involving two different protein preparations in each case. The lines are the best fits of the Michaelis-Menten equation (see Tables S2 and S3 for the values derived from the fits). It is well known that glycosidase catalysis often shows kinetic complexities at high substrate concentrations, due to phenomena such as transglycosylation, inhibition by substrate or allosteric activation[8]. Here Michaelis-Menten saturation kinetics is not for 4-nitrophenyl-β-D-glucopyranoside in a wide concentration range (see inset in panel at the left). Therefore, only the data up to 8 mM have been used for the determination of the catalytic parameters from the fitting of the Michaelis-Menten equation.

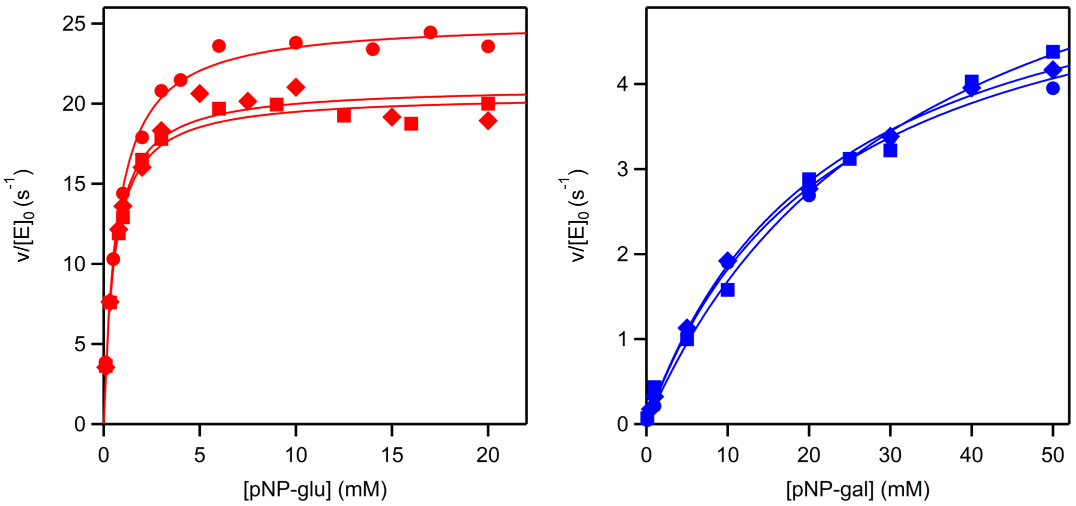


**Figure S9**. Michaelis plots of rate versus substrate concentration at pH 7 and 25 ºC for the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside catalyzed by the modern glycosidase from *Saccharophagus degradans.* The different data points correspond to the 3 experimental replicates performed for each substrate, involving two different protein preparations in each case. The lines are the best fits of the Michaelis-Menten equation (see Tables S2 and S3 for the values derived from the fits).

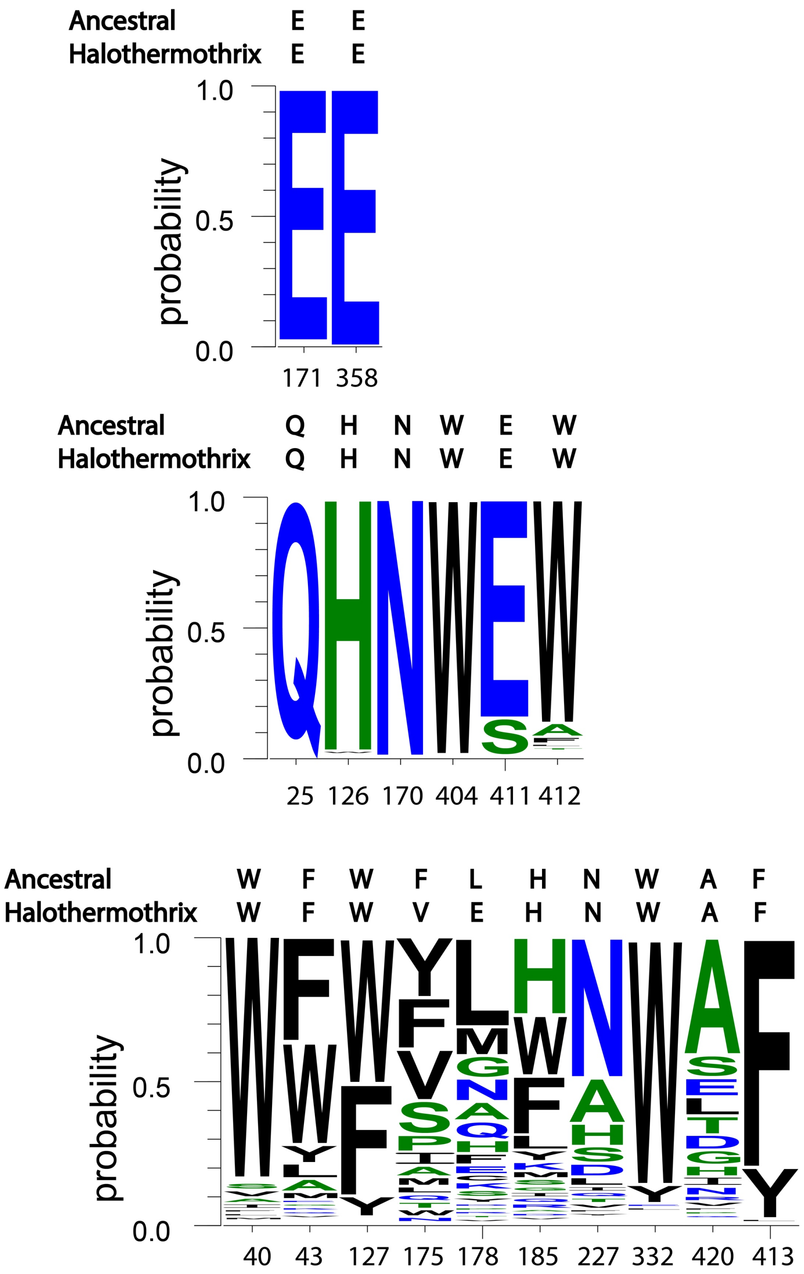


**Figure S10.** Statistics of residue occupancy at critical active site positions in the set of modern glycosidases used as starting point for ancestral sequence reconstruction (see also Table S4). The graphics shown refer to the catalytic carboxylic acids (upper), the positions involved in the binding of the glycone moiety of the substrate (middle) and the positions involved in the binding of the aglycone moiety of the substrate (lower). The sequences of the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii* are also given. Glycosidases are known to be somewhat specific for the glycone moiety of the substrate and much less specific for the aglycone moiety, which is reflected in lower residue conservation at the protein residues involved in aglycone binding.

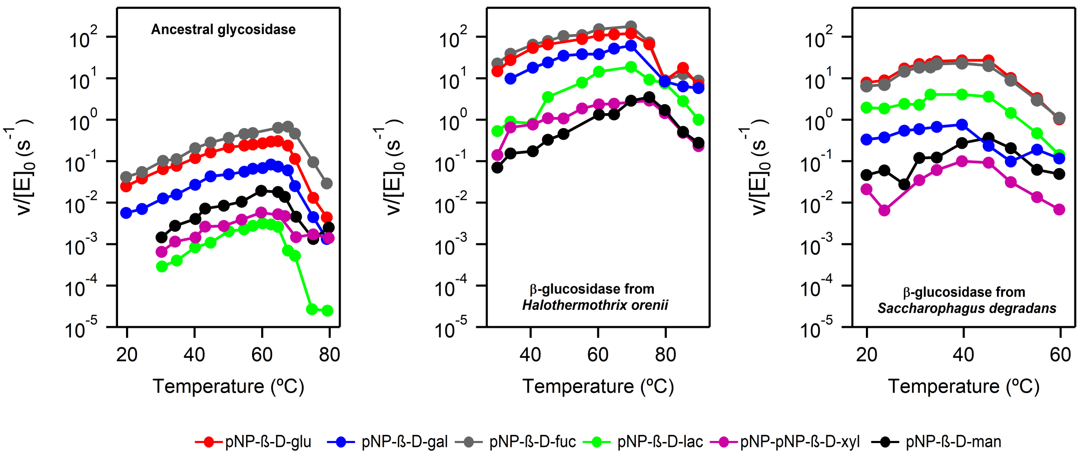


**Figure S11.** Profiles of activity versus temperature for the ancestral glycosidase (left) and the two modern glycosidases from *Halothermothrix orenii* and *Saccharophagus degradans* using the following substrates: *4*-nitrophenyl-β-D-glucopyranoside, *4*-nitrophenyl-β-D-galactopyranoside, *4*-nitrophenyl-β-D-fucopyranoside, *4*-nitrophenyl-β-D-lactopyranoside, *4*-nitrophenyl-β-D-xylopyranoside and *4*-nitrophenyl-β-D-mannopyranoside. Activity values were derived from determination of *p*-nitrophenolate after 10 minutes incubation of 1 mM substrate with the enzyme, as described in Methods.

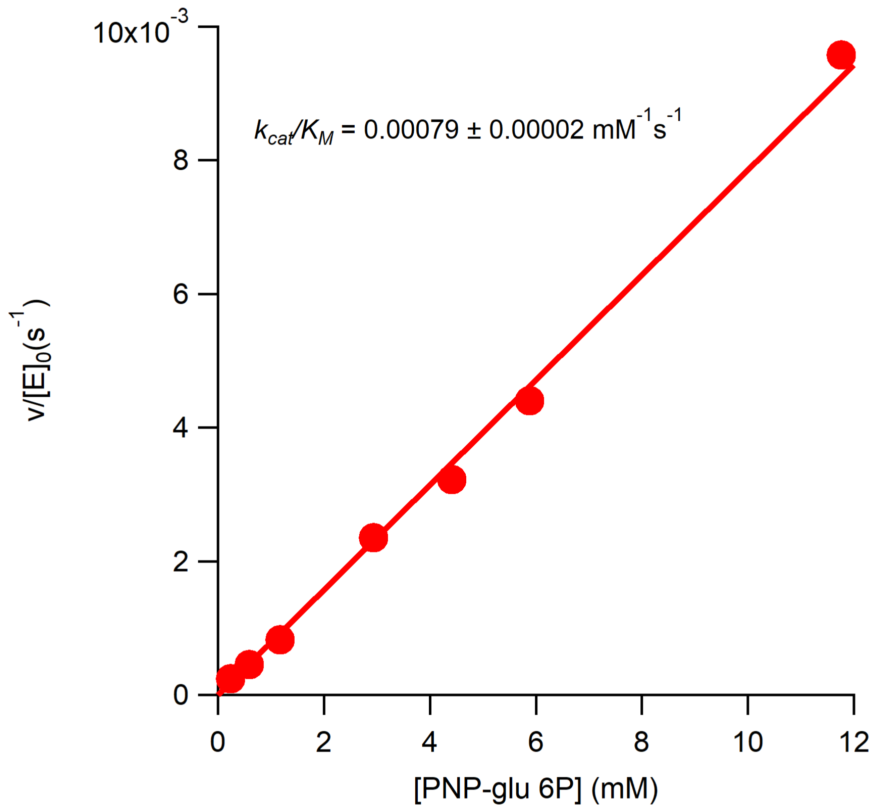


**Figure S12**. Michaelis plot of rate versus substrate concentration for the hydrolysis of *4*-nitrophenyl-β-D-glucopyranoside-6-phosphate catalysed by the ancestral glycosidase. No curvature is observed in the plot and, therefore, only the value for the catalytic efficiency can be derived from the experimental data. This value is about ~40 times lower than the catalytic efficiency with the corresponding non-phosphorylated substrate.

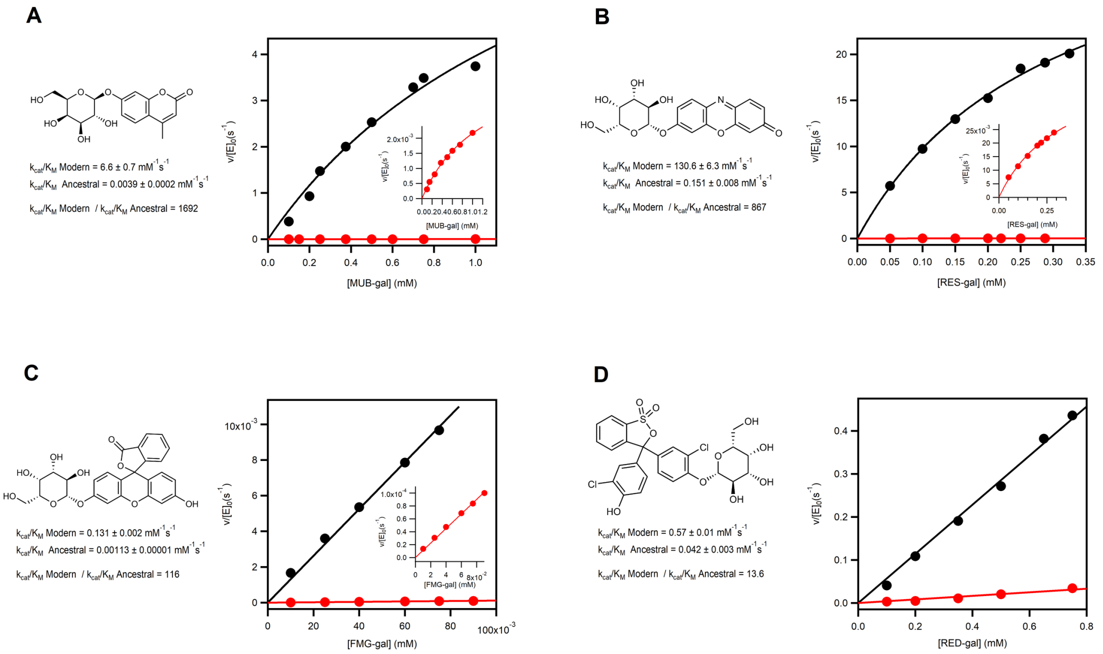
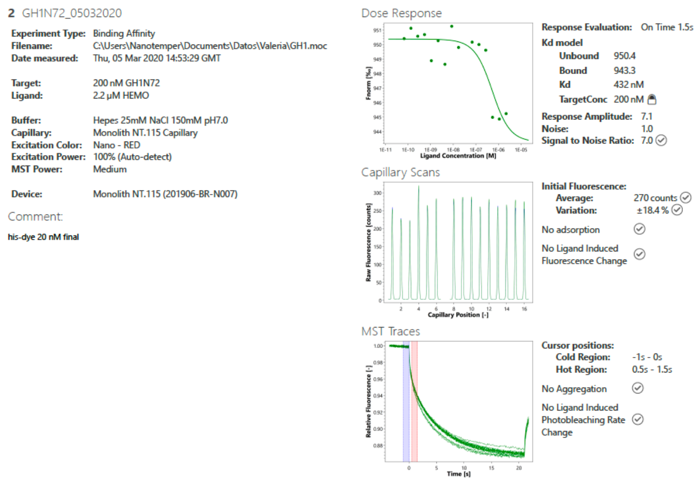


**Figure S13**. Michaelis plots of rate *versus* substrate concentration for the hydrolysis of the indicated substrates catalyzed by the ancestral glycosidase (red and insets) and the modern glycosidase from *Halothermothrix orenii* (black). All the substrates used are β-D-galactopyranosides with a large aglycone moiety. Catalytic efficiencies derived from the fits of the Michaelis-Menten equation are shown.



**2** GH1N72_05032020

**Experiment Type:** Binding Affinity
**Filename:** C:\Users\Nanotemper\Documents\Datos\Valeria\GH1.moc
**Date measured:** Thu, 05 Mar 2020 14:53:29 GMT

**Target:** 200 nM GH1N72
**Ligand:** 2.2 μM HEMO

**Buffer:** Hepes 25mM NaCl 150mM pH7.0
**Capillary:** Monolith NT.115 Capillary
**Excitation Color:** Nano - RED
**Excitation Power:** 100% (Auto-detect)
**MST Power:** Medium

**Device:** Monolith NT.115 (201906-BR-N007)

Comment:

his-dye 20 nM final

Dose Response

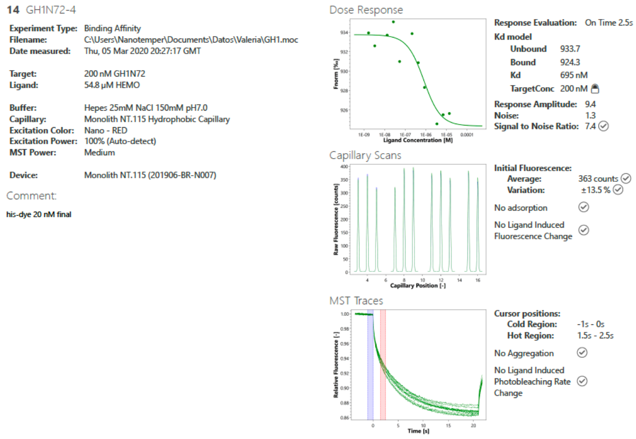
**Response Evaluation:** On Time 1.5s
**Kd model**
**Unbound** 950.4
**Bound** 943.3
**Kd** 432 nM
**TargetConc** 200 nM
**Response Amplitude:** 7.1
**Noise:** 1.0
**Signal to Noise Ratio:** 7.0 ✓

Capillary Scans

**Initial Fluorescence:**
**Average:** 270 counts ✓
**Variation:** ±18.4 % ✓
No adsorption ✓
No Ligand Induced Fluorescence Change ✓

MST Traces

**Cursor positions:**
**Cold Region:** -1s - 0s
**Hot Region:** 0.5s - 1.5s
No Aggregation ✓
No Ligand Induced Photobleaching Rate Change ✓



**14** GH1N72-4

**Experiment Type:** Binding Affinity
**Filename:** C:\Users\Nanotemper\Documents\Datos\Valeria\GH1.moc
**Date measured:** Thu, 05 Mar 2020 20:27:17 GMT

**Target:** 200 nM GH1N72
**Ligand:** 54.8 μM HEMO

**Buffer:** Hepes 25mM NaCl 150mM pH7.0
**Capillary:** Monolith NT.115 Hydrophobic Capillary
**Excitation Color:** Nano - RED
**Excitation Power:** 100% (Auto-detect)
**MST Power:** Medium

**Device:** Monolith NT.115 (201906-BR-N007)

Comment:

his-dye 20 nM final

Dose Response

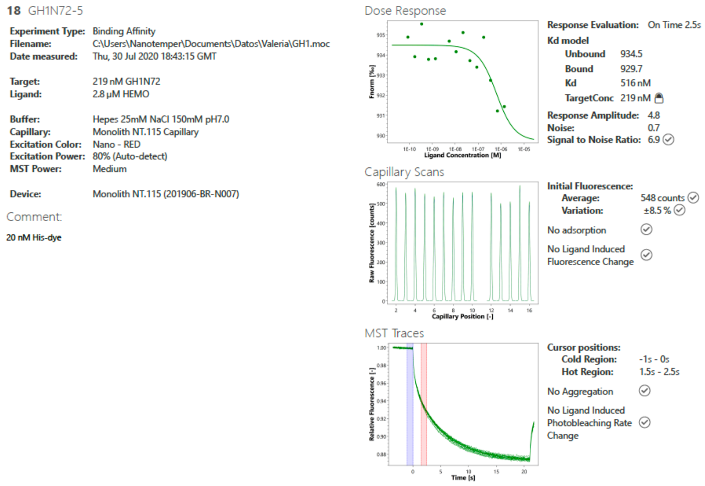
**Response Evaluation:** On Time 2.5s
**Kd model**
**Unbound** 933.7
**Bound** 924.3
**Kd** 695 nM
**TargetConc** 200 nM
**Response Amplitude:** 9.4
**Noise:** 1.3
**Signal to Noise Ratio:** 7.4 ✓

Capillary Scans

**Initial Fluorescence:**
**Average:** 363 counts ✓
**Variation:** ±13.5 % ✓
No adsorption ✓
No Ligand Induced Fluorescence Change ✓

MST Traces

**Cursor positions:**
**Cold Region:** -1s - 0s
**Hot Region:** 1.5s - 2.5s
No Aggregation ✓
No Ligand Induced Photobleaching Rate Change ✓



**18** GH1N72-5

**Experiment Type:** Binding Affinity
**Filename:** C:\Users\Nanotemper\Documents\Datos\Valeria\GH1.moc
**Date measured:** Thu, 30 Jul 2020 18:43:15 GMT

**Target:** 219 nM GH1N72
**Ligand:** 2.8 μM HEMO

**Buffer:** Hepes 25mM NaCl 150mM pH7.0
**Capillary:** Monolith NT.115 Capillary
**Excitation Color:** Nano - RED
**Excitation Power:** 80% (Auto-detect)
**MST Power:** Medium

**Device:** Monolith NT.115 (201906-BR-N007)

Comment:

20 nM His-dye

Dose Response

**Response Evaluation:** On Time 2.5s
**Kd model**
**Unbound** 934.5
**Bound** 929.7
**Kd** 516 nM
**TargetConc** 219 nM
**Response Amplitude:** 4.8
**Noise:** 0.7
**Signal to Noise Ratio:** 6.9 ✓

Capillary Scans

**Initial Fluorescence:**
**Average:** 548 counts ✓
**Variation:** ±8.5 % ✓
No adsorption ✓
No Ligand Induced Fluorescence Change ✓

MST Traces

**Cursor positions:**
**Cold Region:** -1s - 0s
**Hot Region:** 1.5s - 2.5s
No Aggregation ✓
No Ligand Induced Photobleaching Rate Change ✓

**Figure S14.** Relevant experimental data plots and validation reports for the quantification of heme binding to the ancestral glycosidase using microscale thermophoresis. Note that information for three replicate experiments is provided. Since bound heme is monomeric, while heme in solution at neutral pH has a tendency to associate, it is possible that the reported dissociation constants are overestimates (*i.e.,* binding could be even tighter than suggested by these values).

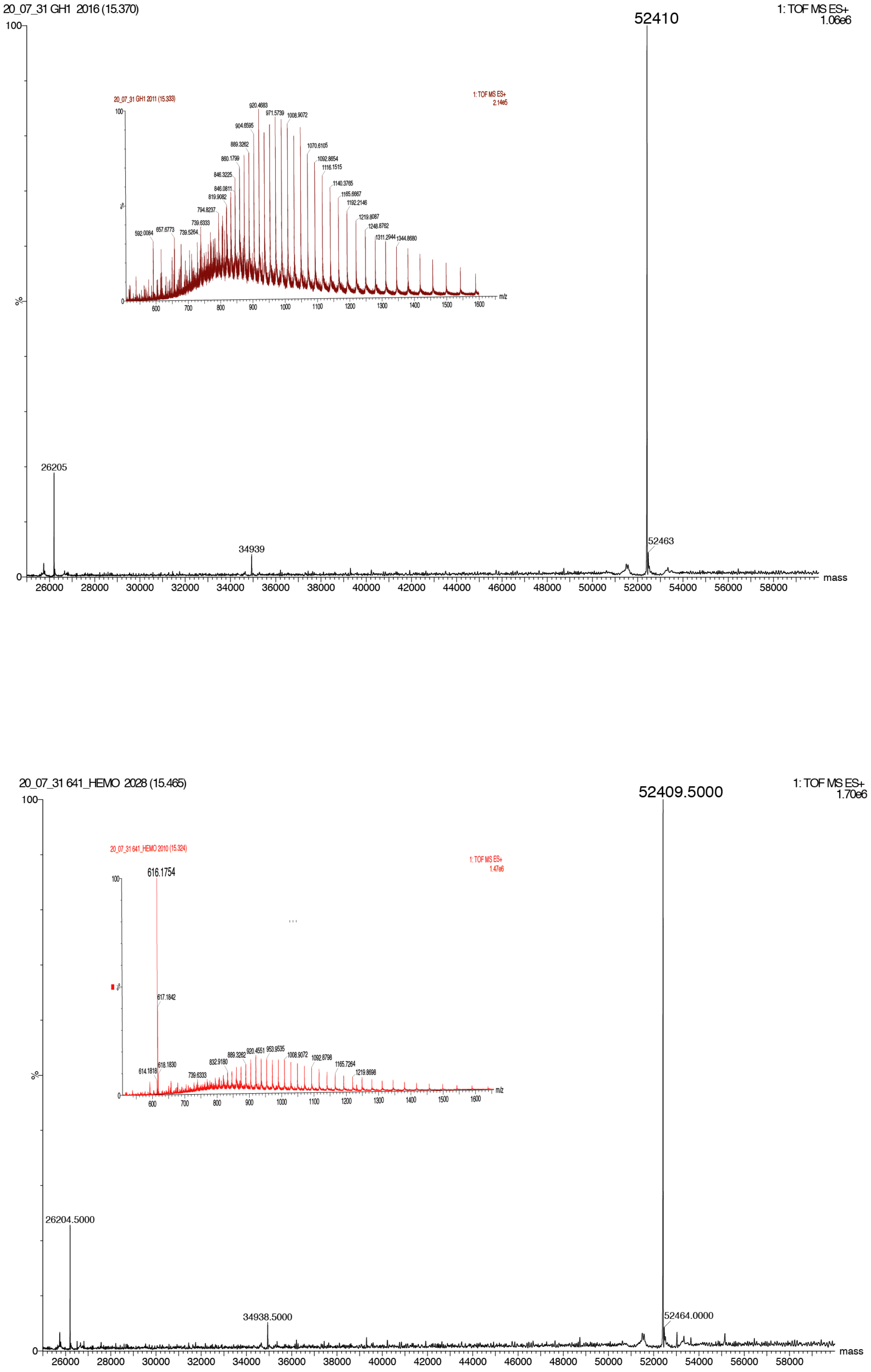


**Figure S15.** Mass spectra after UPLC elution of samples of the ancestral glycosidase originally without (upper panel) and with (lower panel) bound heme. The protein peak is apparent at a mass near the theoretical value calculated from the amino acid sequence (52542.79 Da). The insets correspond to the low mass range where heme is

expected to appear. A peak of mass essentially close to that expected for heme (651.94 Da) is observed only in the lower panel. The intensity of the heme peak is qualitatively similar to that of the protein peak, as expected from the 1:1 binding stoichiometry.

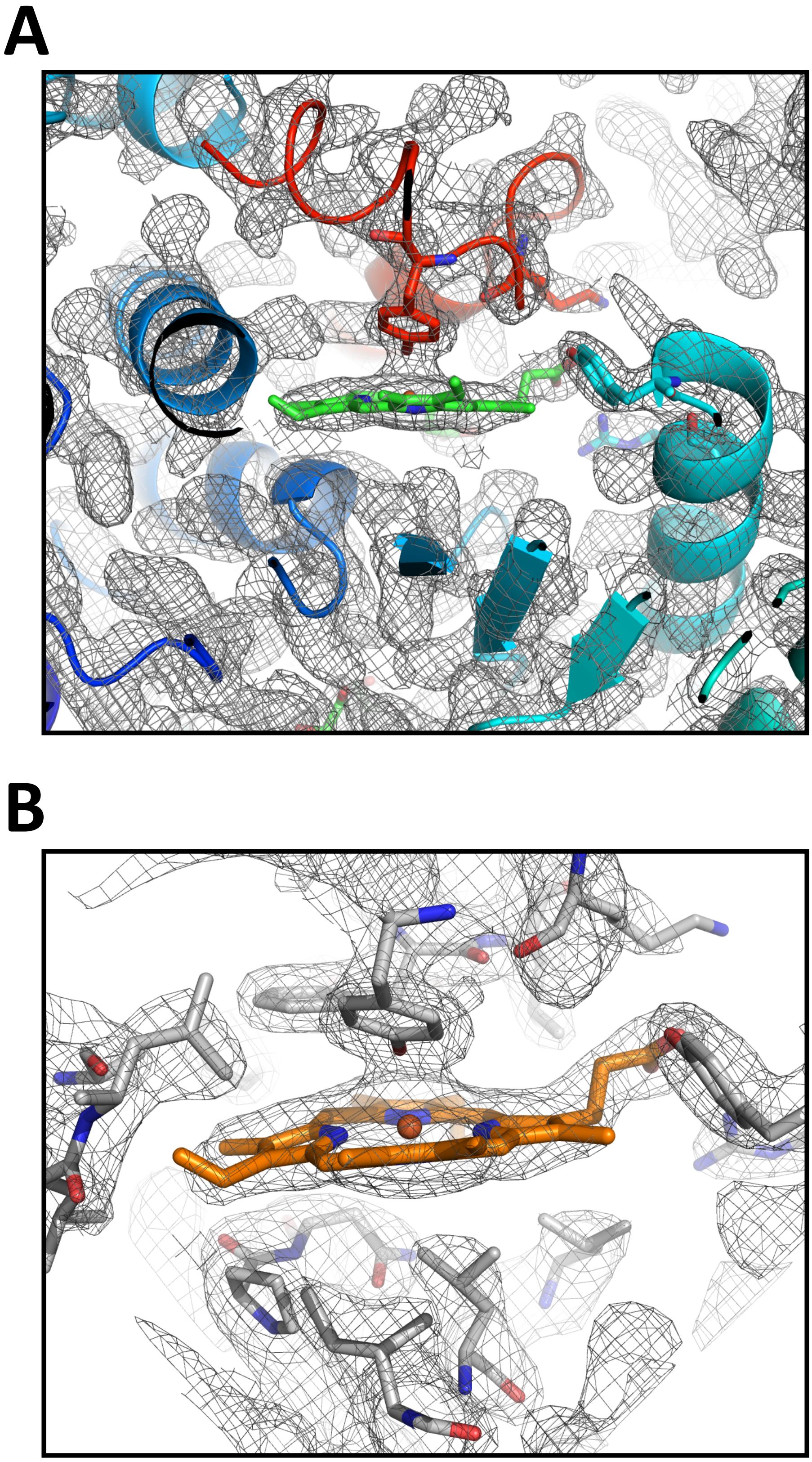


**Figure S16.** (A) Structure of the ancestral glycosidase showing the bound heme group into the well-defined |2Fo-Fc| electron density map contoured at 1 σ. The four amino acids involved in hydrogen bonds (See Figure 7A) are shown as sticks. B) Blow-up showing all residue participating in binding the heme group.

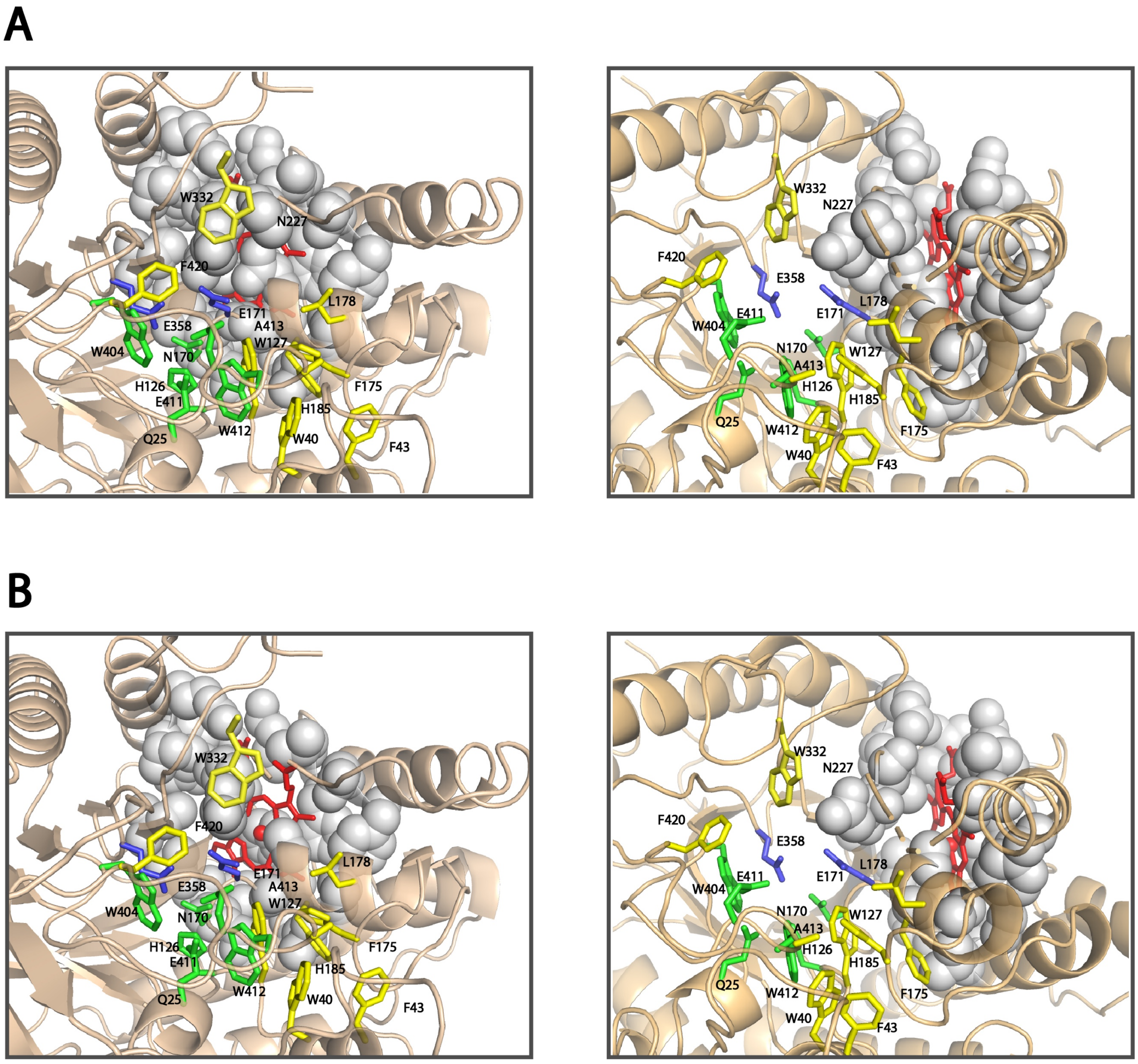


**Figure S17**. (A) View of the active site of the ancestral protein showing the catalytic carboxylic acid residues (blue) and the residues involved in binding of the glycone (yellow) and aglycone (green) parts of the substrate molecule. The residues that block the connection of the heme group with the active site are shown with van der Waals spheres and colored in grey. (B) Same as in C, but the residues blocking the connection of the heme with the active site have been computationally mutated to alanine, in such a way that now the iron of the heme group can be seen at the bottom of the active site in the chosen view.

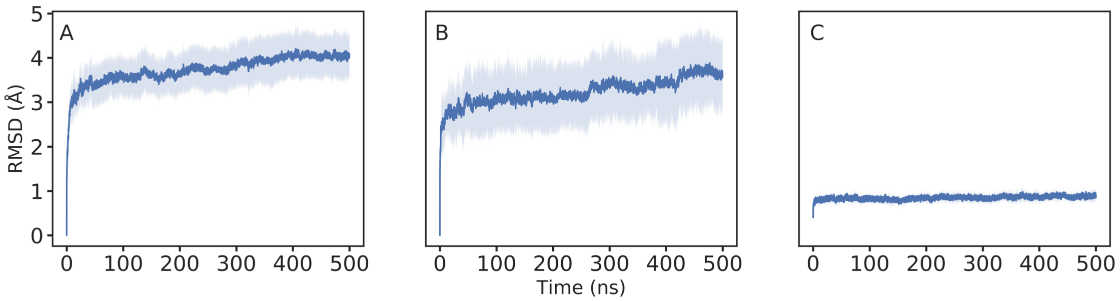


**Figure S18.** The root mean square deviations (RMSD, Å) of all backbone atoms of the (A) ancestral glycosidase without heme bound, (B) ancestral glycosidase with heme bound and (C) modern glycosidase from *Halothermothrix orenii* over ten individual 500 ns MD simulations per system (i.e. 5 μs cumulative simulation time per system). The center for the error band (solid blue line) shows the average RMSD obtained per system at a given time frame for each replica, while the standard deviations are given as the shaded areas on each plot. n = 10 independent simulations.

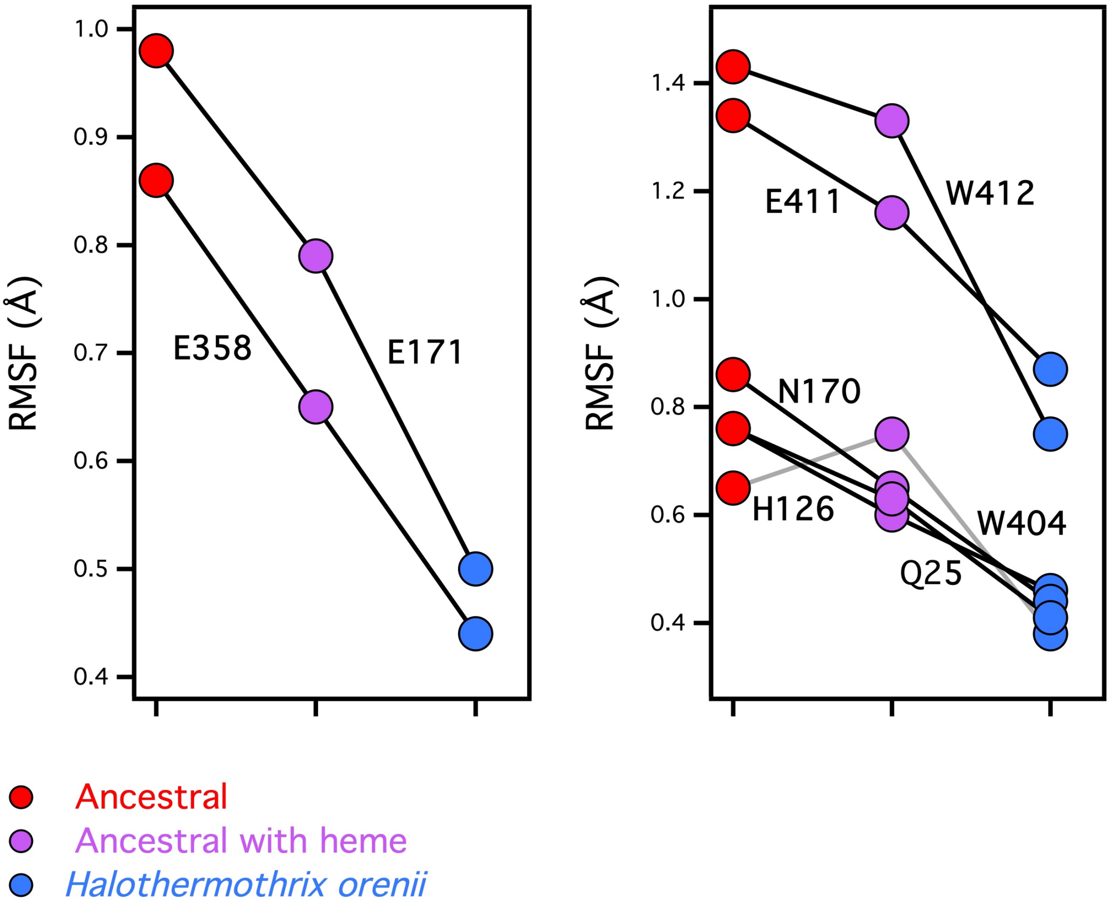


**Figure S19.** Root mean square deviations of all backbone atoms of the catalytic carboxylic acid residues (left) and residues involved in the binding of the glycone moiety of the substrate (right). Values are shown for the ancestral glycosidase with and without heme as well as for the modern glycosides from *Halothermothrix orenii*. This figure is a complement to Figure 3 in the main text, further details about the analysis are provided in the caption to Figure 3. See also Figures 2D and S10 for more information about the residues selected for this figure.

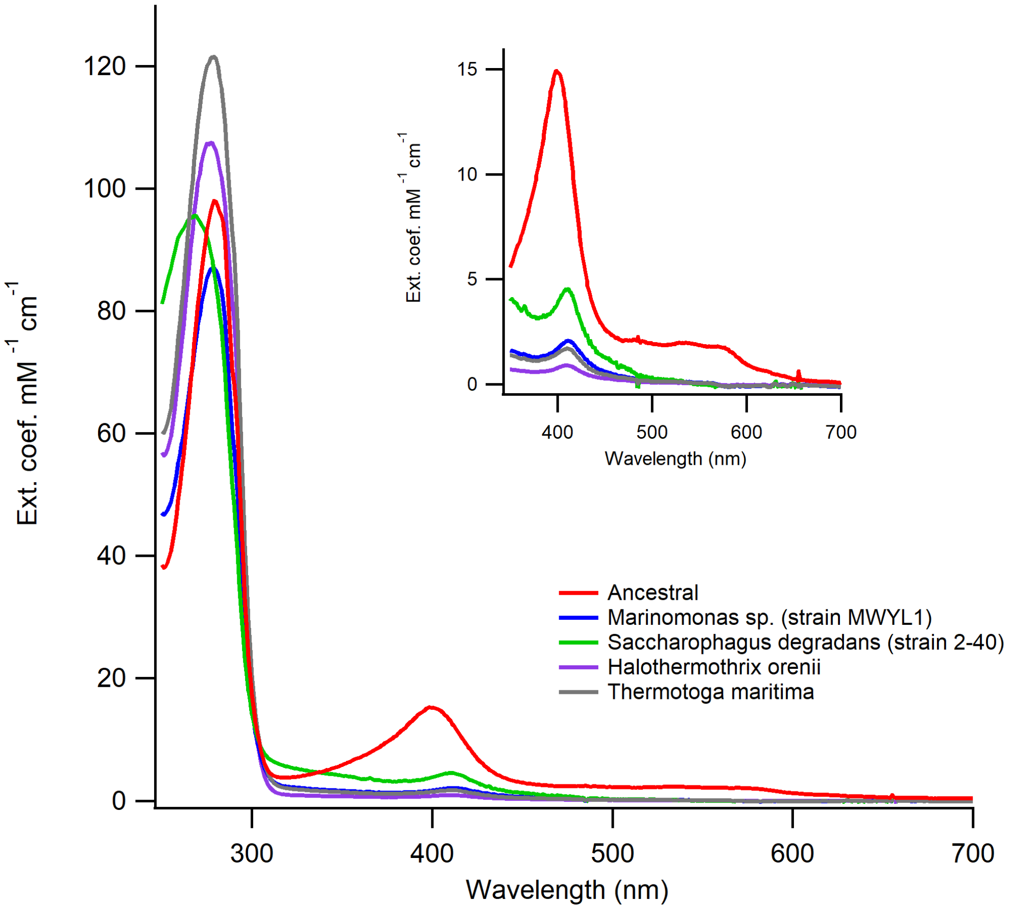


**Figure S20.** UV-VIS spectra of protein preparations of modern glycosidases showing the protein absorption band at 280 nm and the Soret heme band at about 400 nm. The inset is a blow-up of the Soret band region. For comparison, data for the ancestral glycosidase (corresponding to the reconstruction at node 72) are also included. In all preparations, 0.4 mM 5-aminolevulinic acid (the metabolic precursor of heme) was added to the culture medium and the protein was purified by Ni-NTA affinity chromatography and further passage through a PD10 column.

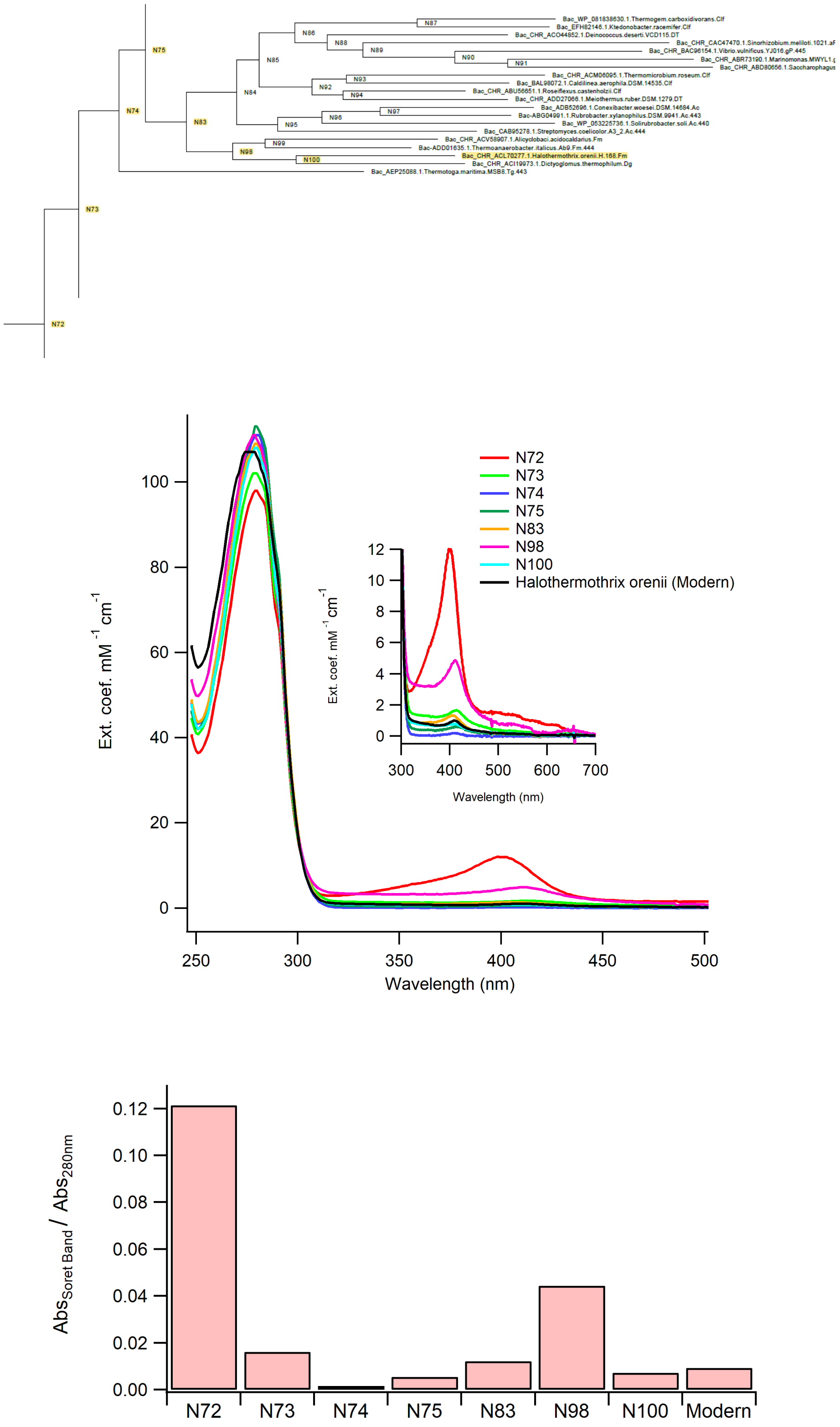


**Figure S21**. UV-VIS spectra of protein preparations corresponding to reconstructions of nodes in the line of descent leading from the ancestral glycosidase at node 72 to the modern glycosidase from *Halothermothrix orenii*. For comparison, data for the modern glycosidase are included. In all preparations, 0.4 mM 5-aminolevulinic acid (the metabolic precursor of heme) was added to the culture medium and the protein was purified by Ni-NTA affinity chromatography and further passage through a PD10 column. Upper panel: section of the phylogenetic tree used as a basis for sequence reconstruction (Figure S1) highlighting the nodes studied here. Middle panel: UV-VIS spectra showing the protein absorption band at 280 nm and the Soret heme band at about 400 nm. The inset is a blow-up of the Soret band region. Lower panel: ratio of absorbance at the maximum of the heme Soret band to the absorbance at the maximum of the protein aromatic absorption band.

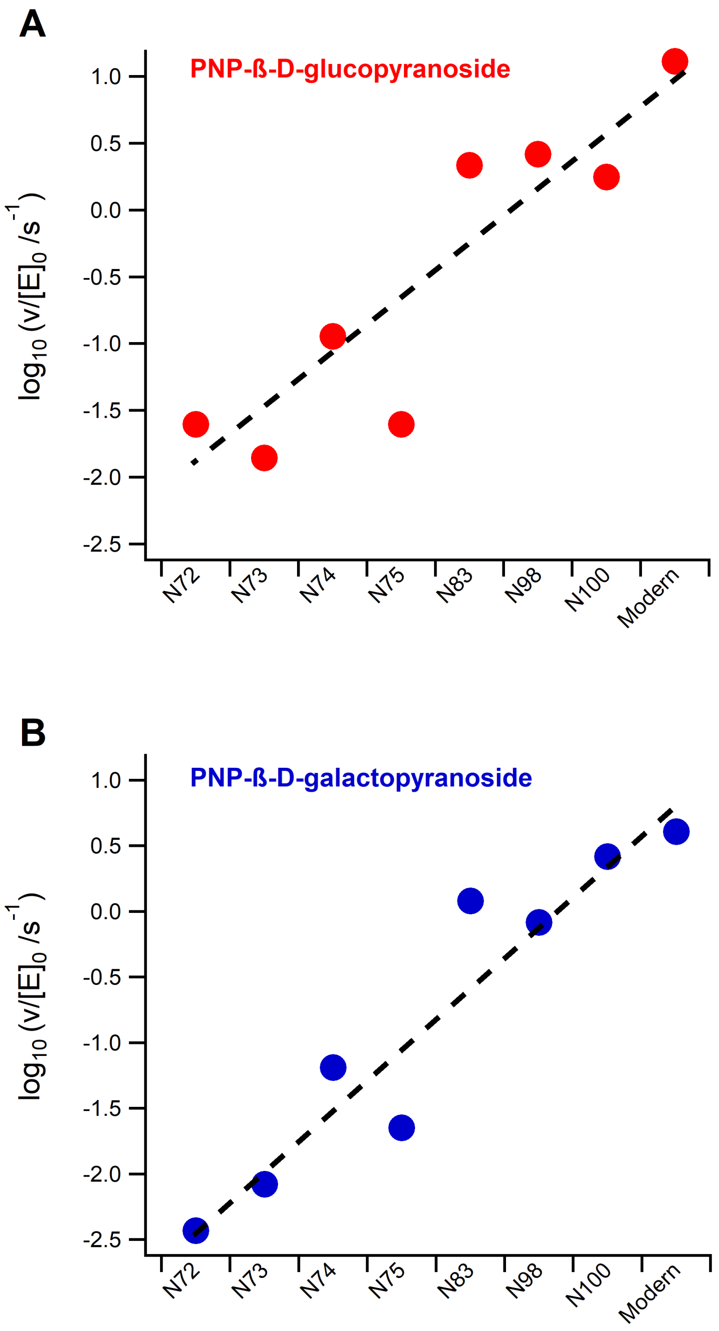


**Figure S22.** Glycosidase activity of protein preparations corresponding to reconstructions of nodes in the line of descent leading from the ancestral glycosidase at node 72 to the modern glycosidase from *Halothermothrix orenii* (labelled as modern). Assays were performed at pH 7 and 25 ºC with 1 mM concentration of *4*-nitrophenyl-β-D-glucopyranoside or *4*-nitrophenyl-β-D-galactopyranoside. The dashed lines are only meant to guide the eye.

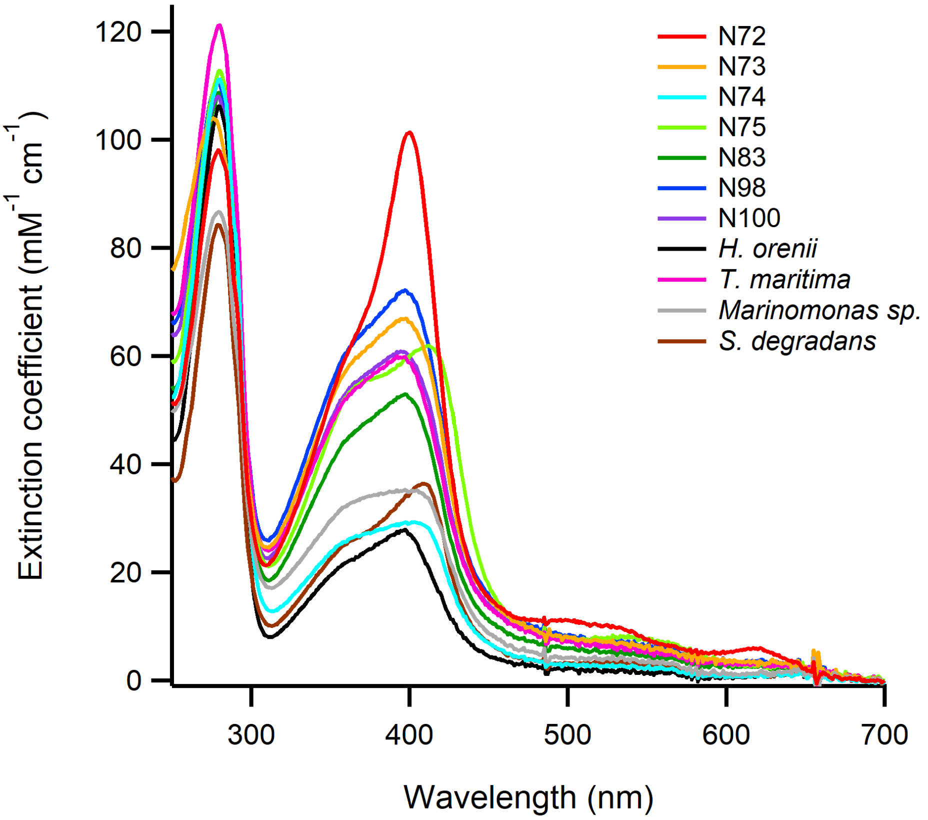


**Figure S23.** UV-VIS spectra of protein preparations corresponding to reconstructions of nodes in the line of descent leading from the ancestral glycosidase at node 72 to the modern glycosidase from *Halothermothrix orenii* (see Figure 8 in the main text and Figure S21). For comparison, data for three additional modern glycosidases are also included. In all preparations, heme-free protein samples at ~5 μM concentration were incubated with for 1 hour at pH 7 with a 5-fold excess of heme and free heme was removed by exclusion chromatography (2 passages through PD10 columns) before recording the UV-VIS spectra.

## N73

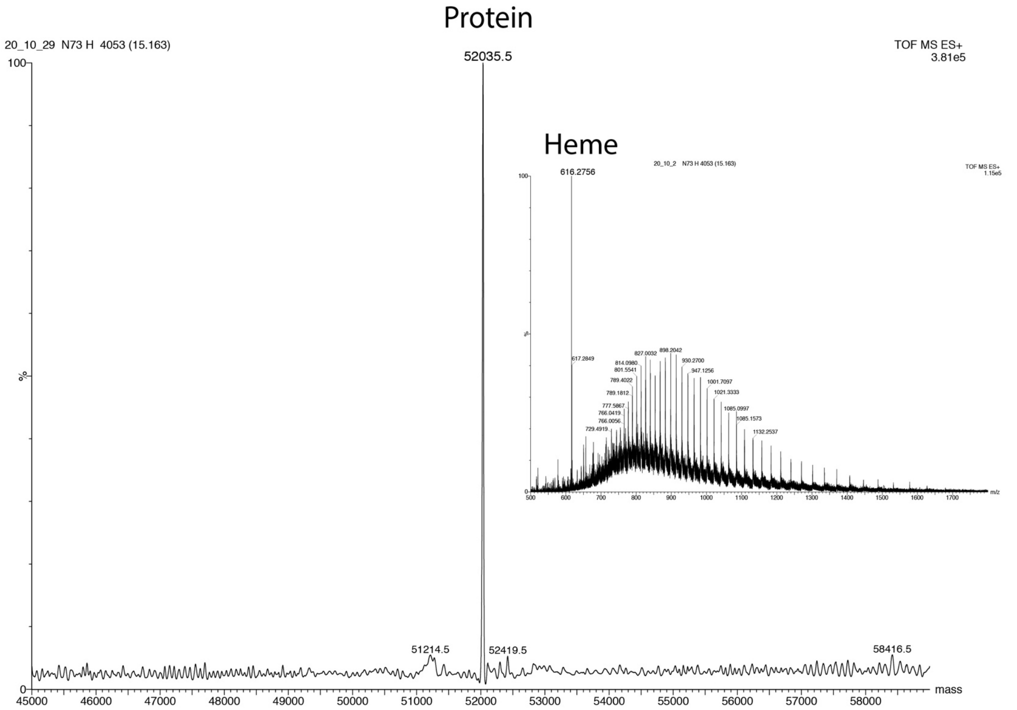


**Figure S24.** Mass spectra after UPLC elution of samples of the ancestral glycosidase corresponding to node 73 (Figure 8A and upper panel in Figure S21) with bound heme (see legends to Figures 8 and S21 for details). The protein peak is apparent at a mass near the theoretical value calculated from the amino acid sequence. The insets correspond to the low mass range where heme is expected to appear. A peak of mass close to that expected for heme is observed. This peak was not present in preparations of the protein without bound heme.

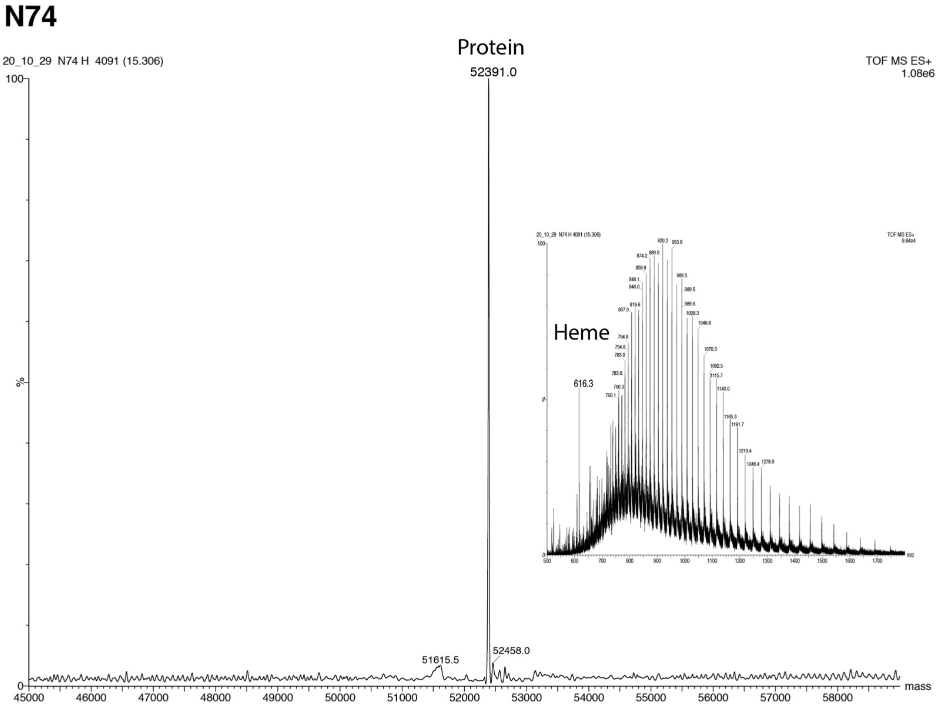


**Figure S25.** Mass spectra after UPLC elution of samples of the ancestral glycosidase corresponding to node 74 (Figure 8A and upper panel in Figure S21) with bound heme (see legends to Figures 8 and S21 for details). The protein peak is apparent at a mass near the theoretical value calculated from the amino acid sequence. The insets correspond to the low mass range where heme is expected to appear. A peak of mass close to that expected for heme is observed. This peak was not present in preparations of the protein without bound heme.

**N75**

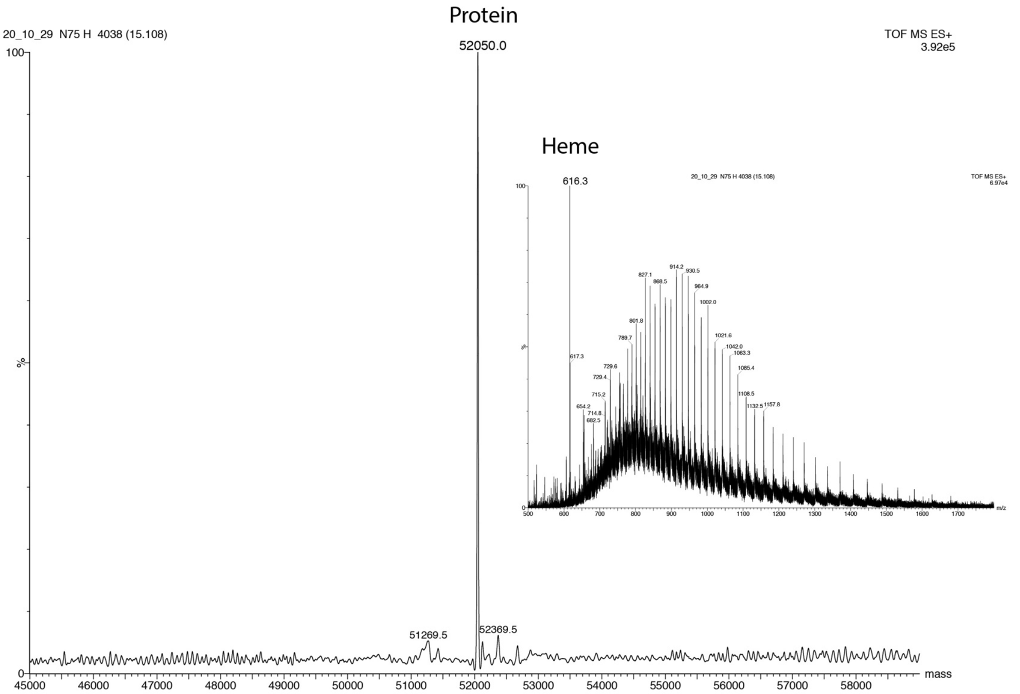


**Figure S26.** Mass spectra after UPLC elution of samples of the ancestral glycosidase corresponding to node 75 (Figure 8A and upper panel in Figure S21) with bound heme (see legends to Figures 8 and S21 for details). The protein peak is apparent at a mass near the theoretical value calculated from the amino acid sequence. The insets correspond to the low mass range where heme is expected to appear. A peak of mass close to that expected for heme is observed. This peak was not present in preparations of the protein without bound heme.

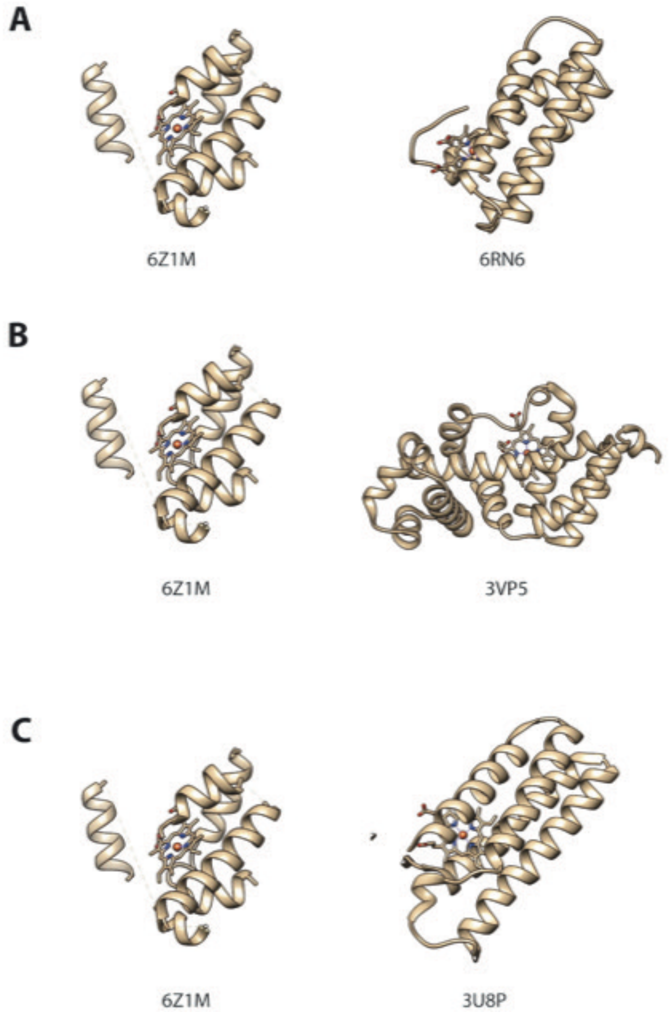


**Figure S27**. Comparison of the $\alpha$-helix structure around the bound heme in the ancestral glycosidase (left) with the results of a DALI search of the Protein Data Bank. The $\alpha$-helices involved in heme binding in our ancestral glycosidase were used as query for a structural alignment search. This resulted in 222 hits with RMSD values ranging from 1.6 to 11.4 Å. Only 3 of those hits had bound heme (shown here as A; B and C). The corresponding structures in the heme binding region are shown at the right. The corresponding structural alignments had RMSD ~4 Å and Z scores of 2 or higher.

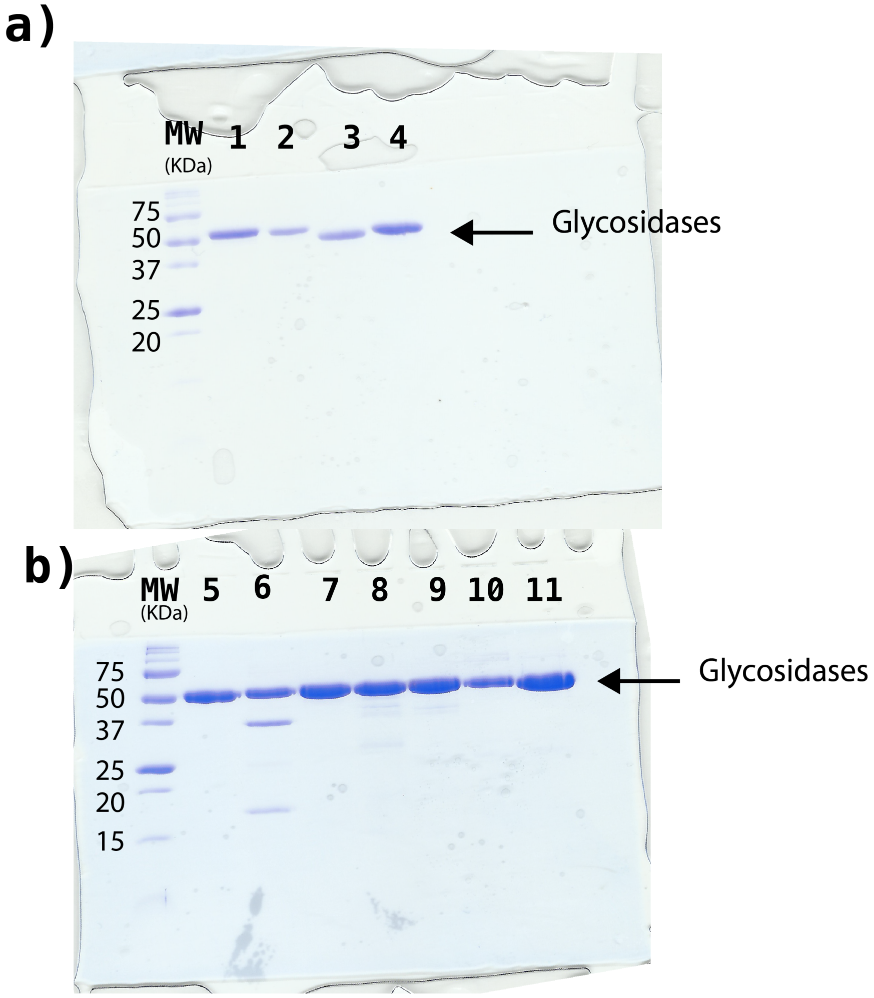


**Figure S28.** SDS gel electrophoresis of preparations of modern and ancestral proteins studied in this work. Code is as follows: MK: molecular weight markers. 1: *Halothermothrix orenii.* 2: *Thermotoga maritima.* 3: *Saccharophagus degradans (strain 2-40T).* 4: *Marinomonas sp. (strain MWYL1).* 5: N72. 6: N73. 7: N74. 8: N75. 9: N83. 10: N98. 11: N100 (Figure 8 in the main text and upper panel in Figure S21). Densitometry quantification indicated the the purity of the protein samples was in the range 93%-98%. Three independent experiments were performed with similar results.

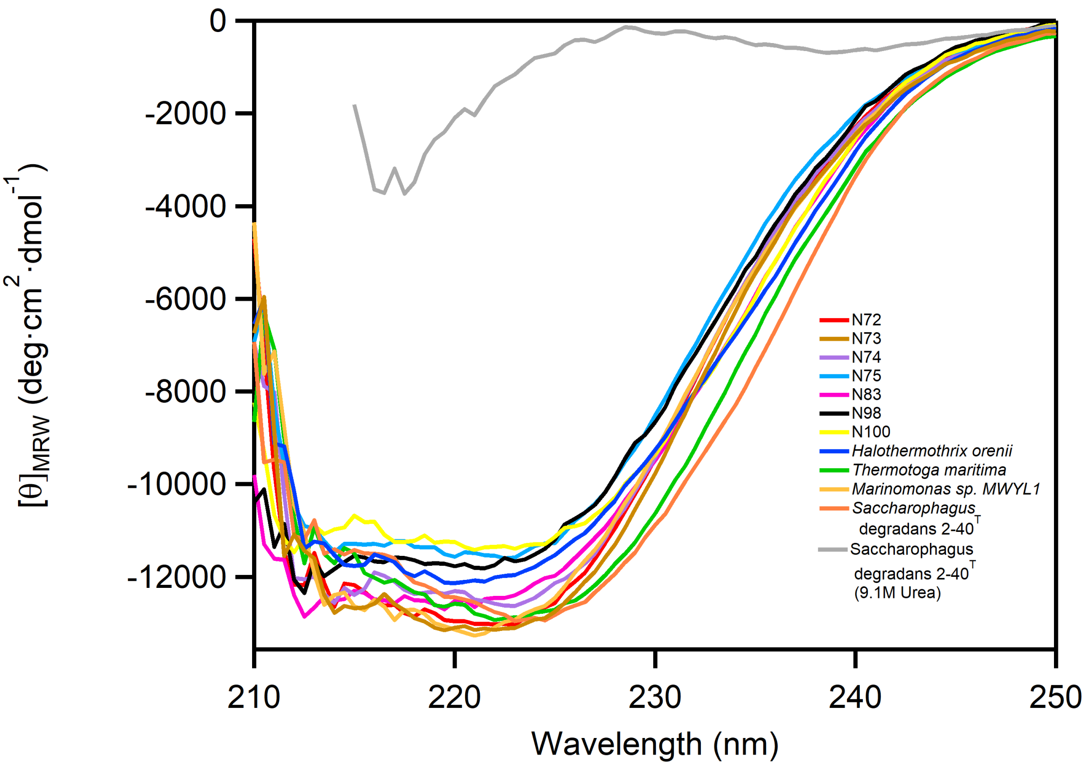


**Figure S29.** Circular dichroism spectra of modern and ancestral proteins studied in this work (see Figure 8 in the main text and upper panel of Figure S21 for the identification of the ancestral nodes). Conditions were 50 mM HEPES, pH 7.0, protein concentration within the 0.2-0.6 mg/mL range and a 1 mm pathlength cuvette. An average of 30 scans was performed in each case. Blank subtraction was always carried out prior to mean residue ellipticity calculation. For comparison, the spectra of *Saccharophagus degradans* glycosidase unfolded in 9.1 M urea is also reported.

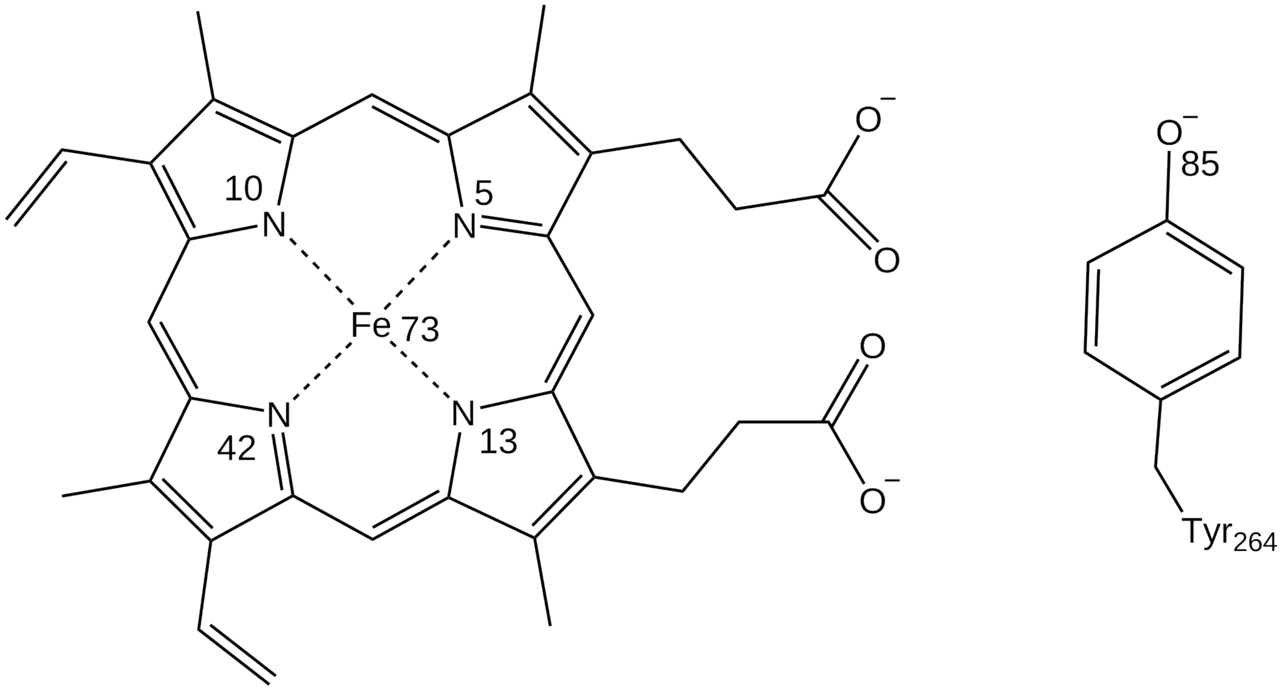


**Figure S30.** Atom numbering of the atoms considered in the bonded model between the Fe and the heme and Tyr264.

## References

1. Altschul, S.F., Madden, T.L., Schaäffer, A.A., Zhang. J., Zhang, Z., Miller. W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **25**, 3389-3402 (1997).
2. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9,** 173-175 (2011).
3. 3.Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195-201.
4. Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27,** 343-350 (2011).
5. Biasini, M., Bienert. S., Waterhouse, A., Arnold, K., Studer, G., Schimdt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L. & Schwede, T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acid Res. **42,** W252-W258 (2014).
6. Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., Jain, S., Lewis, S.M., Arendall, W.B. 3rd, Snoeyink, J., Adams. P.D., Lovell, S.C., Richardson, J.S. & Richardson, D.C. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27,** 293-315 (2018).
7. Marana, S.R. Molecular basis of substrate specificity in family 1 glycoside hydrolases. *IUBMB Life* **58**, 63-73 (2006).
8. Kuusk, S. & Väljamäe, P. When substrate inhibits and inhibitor activates: implications of β-glucosidases. *Biotechnol. Biofuels* **10,** 7 (2017).

# PUBLICATION IV

## (Preprint)

## Protection of Catalytic Cofactors by Polypeptides as a Driver for the Emergence of Primordial Enzymes

bioRχiv

# Protection of catalytic cofactors by polypeptides as a driver for the emergence of primordial enzymes

Luis I. Gutierrez-Rus[1], Gloria Gamiz-Arco[1], J.A. Gavira[2], Eric A. Gaucher[3], Valeria A. Risso[1], Jose M. Sanchez-Ruiz[1]*

[1]Departamento de Quimica Fisica, Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071-Granada, Spain.

[2]Laboratorio de Estudios Cristalograficos, Instituto Andaluz de Ciencias de la Tierra, CSIC, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, Avenida de las Palmeras 4, Granada 18100 Armilla, Spain.

[3]Department of Biology, Georgia State University, Atlanta, GA 30303

*Corresponding author: sanchezr@ugr.es

# Abstract

Enzymes catalyze the chemical reactions of life. For nearly half of known enzymes, catalysis requires the binding of small molecules known as cofactors. Polypeptide-cofactor complexes likely formed at a primordial stage and became starting points for the evolution of many efficient enzymes. Yet, evolution has no foresight so the driver for the primordial complex formation is unknown. Here, we use a resurrected ancestral TIM-barrel protein to identify one potential driver. Heme binding at a flexible region of the ancestral structure yields a peroxidation catalyst with enhanced efficiency when compared to free heme. This enhancement, however, does not arise from protein-mediated promotion of catalysis. Rather, it reflects protection of bound heme from common degradation processes and a resulting longer life time and higher effective concentration for the catalyst. Protection of catalytic cofactors by polypeptides emerges as a general mechanism to enhance catalysis and may have plausibly benefited primordial polypeptide-cofactor associations.

# Introduction

Life involves a vast network of chemical reactions, most of which would not occur at a sufficient rate in the absence of effective enzyme catalysis. With the obvious exception of ribozymes, enzymes are based on protein scaffolds. Most, if not all, of modern enzyme activities have evolved from prior enzyme activities (Ohno 1970; Khersonsky and Tawfik 2010). This evolutionary process is reasonably well-understood and protein engineers can mimic it in the lab using directed evolution (Campbell et al 2016; Zeymer and Hilvert 2018). On the other hand, it is a logical necessity that the first enzymes did not arise from previously existing "older" enzymes. That is, the first primordial enzymes must have necessarily arisen *de novo* in previously non-catalytic protein scaffolds. Reconstructions of the gene content of the last universal common ancestor (LUCA) support that a diversity of enzymes already catalyzed many different reactions at a very early evolutionary stage (Weiss et al 2016). It seems reasonable, therefore, that there are efficient mechanisms for the *de novo* emergence of completely new enzymes. Such mechanisms, however, remain elusive to protein engineers, who have found the generation of efficient *de novo* enzymes in the lab to be extremely challenging (Blomberg et al 2013; Risso et al 2017; Donnelly et al 2018; Lovelock et al 2022; Yeh et al 2023). Our limited grasp of the evolutionary mechanisms of *de novo* enzyme generation implies a serious gap in our understanding of the origin of life.

A substantial fraction (around 50%) of enzymes rely on cofactors for catalysis (Fischer et al 2010). This statement is true for many modern enzymes, but it also very likely holds for the most ancient enzymes, as indicated by reconstructions of the gene content of LUCA (Weiss et al 2016). Many cofactors, in particular metals and metal-containing organic cofactors, display by themselves significant levels of catalysis with a diversity of chemical reactions (Andreini et al 2008; Fischer et al 2010). Both polypeptides and cofactors were likely available already at a prebiotic stage (Frenkel-Pinter et al 2020; Goldman and Kacar 2021). Therefore, recruitment of catalytic cofactors by polypeptides would appear to provide a simple and straightforward mechanism for the *de novo* emergence of many of the primordial enzymes (although certainly not of all of them). On closer examination, however, this mechanism has a serious shortcoming: the driving force for the recruitment is not apparent at all. Certainly, once a polypeptide-cofactor complex has been formed, new possibilities arise in a Darwinian evolution scenario, including the enhancement of catalysis through mutations in the protein moiety. Yet, we know evolution has no foresight (Jacob 1977), so any driver of primordial polypeptide-cofactor association must have provided an immediate selective advantage, such as an instant enhancement in catalysis. However, the opposite would seem reasonable on general grounds. Association of a cofactor with a “naïve” polypeptide, i.e., a polypeptide that has not evolved to enhance catalysis by the cofactor, would simply decrease cofactor exposure to the environment and therefore its accessibility to substrates, thus likely impairing catalysis.

Here, we report experimental studies on the emergence of cofactor-based catalysis in an ancestral TIM-barrel protein. Our experiments reveal a general mechanism of immediate catalysis enhancement that may have plausibly provided a driving force for the polypeptide-cofactor association during a primordial period.

# Results

The TIM-barrel is the most widespread enzyme fold, providing a structural scaffold for a large diversity of modern enzyme functionalities (Wierenga 2001; Nagano et al 2002). In fact, the TIM-barrel fold has been proposed to have played an essential role in early metabolism (Goldman et al 2016). We recently reported an ancestral reconstruction exercise targeting family 1 glycosidases, which are of the TIM-barrel fold (Gamiz-Arco et al 2021). We found a putative common ancestor of eukaryotic and bacterial family 1 glycosidases that displayed unusual properties, including tight and stoichiometric binding of heme at a conformationally flexible region distinct from the glycosidase active site (figs. 1A, 1B and 1C). Remarkably, heme binding is extremely rare among modern TIM-barrel proteins (Gamiz-Arco et al 2021).

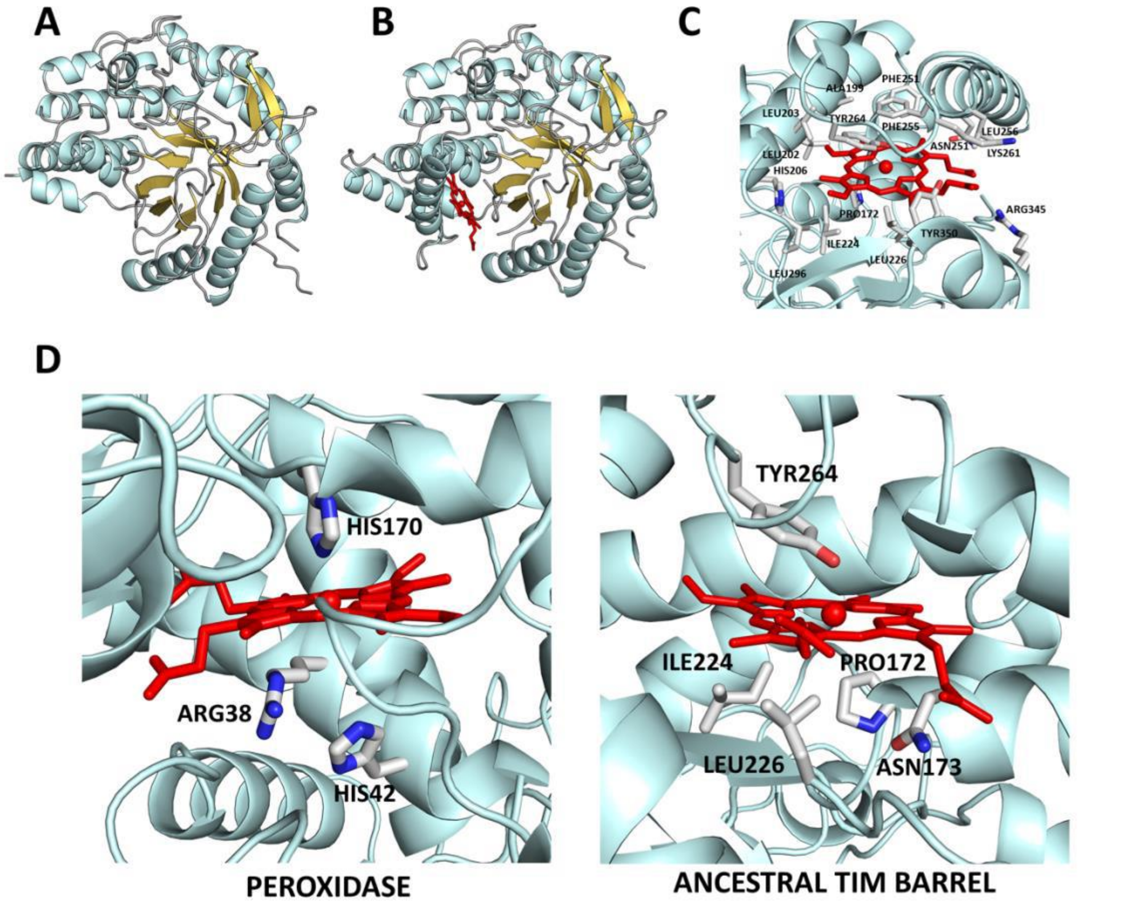


Fig. 1. 3D-structure of the ancestral TIM-barrel glycosidase studied in this work as determined by X-ray crystallography (Gamiz-Arco et al 2021). **(A)** Structure of the protein without heme bound (PDB ID 6Z1H). Note that the region where heme binds displays substantial conformational flexibility and that, as a result, there are missing sections in the electronic density maps (for further details, see Gamiz-Arco et al 2021). **(B)** Structure of the protein with heme bound (PDB ID 6Z1M). Heme binding rigidifies the protein and allows a larger part of the structure to be determined from the density maps, as it is apparent by comparing the structure shown in B with that shown in A (for further details, see Gamiz-Arco et al. 2021). **(C)** Zoom-in of the heme-binding region providing amino acid detail. **(D)** Comparison of the molecular environments of the heme in horseradish peroxidase (PDB ID 1HCH) and the ancestral TIM-barrel (PDB ID 6Z1M). Critical residues for peroxidase catalysis (the proximal histidine and the distal histidine and arginine) are highlighted in the peroxidase structure. In the ancestral TIM-barrel structure, the proximal residue is not a histidine and the residues at the opposite side of the heme ring (labeled) do not include histidines or arginines.

Free heme is known to display a low but measurable peroxidase activity (Brown et al 1969). Therefore, heme binding may be expected to confer some peroxidase activity to the ancestral TIM-barrel. We have used the common peroxidase substrate o-dianisidine (fig. 2A) to compare the peroxidase activity of heme with that of heme bound to the ancestral scaffold. Peroxidation of o-dianisidine leads to absorption in the visible region of the spectrum, which that can be used to follow the chemical reaction (Jenkins et al 2021). As it is customary in enzyme kinetics studies, we started with determinations of the initial rates of the reaction, which can be easily calculated from the initial increases of the absorbance due to the peroxidation product (fig. 2B). We performed a large number of experiments in solutions at different pH values. In most of the studied pH range, the peroxidase activity of the free heme is substantially higher than that of the heme-bound ancestral scaffold (fig. 2B). This result is not surprising, given that burial of the heme in the ancestral protein structure (see figure 7C in Gamiz-Arco et al 2021) should hinder the substrate access to the cofactor, thus impairing catalysis. Furthermore, the heme binding region of the ancestral structure has not evolved to promote cofactor-based catalysis. Enhancement of cofactor-based catalysis in modern peroxidase enzymes is linked to a specific molecular machinery that includes the proximal histidine and distal histidine and arginine residues (Poulos and Kraut 1980; Ortiz de Montellano 2010). Such assistance molecular machinery is lacking from the ancestral TIM-barrel (fig. 1D).

Subsequently, we determined the total yield of peroxidation product for several pH values at times ranging from 15 minutes to 4 hours. Strikingly (fig. 2C) we found a much higher conversion of substrate to product with the heme-bound ancestor when compared to free heme, in particular for reaction times on the order of hours. A reasonable explanation for these results is that free heme undergoes time-dependent processes that impair its peroxidase activity, while heme bound to the protein is protected and retains its peroxidase activity during a much longer time. To test this hypothesis, we performed experiments at longer durations and with repeated additions of substrate (fig. 2D). We found that, unlike the heme bound to the ancestral protein, free heme gradually loses its capability to peroxidize the freshly added substrate.

To further explore the mechanisms of heme protection upon binding to the ancestral protein, we performed kinetic experiments in which heme, either free or protein-bound, was incubated for given times in the reaction buffer prior to substrate addition. Profiles of product concentration versus reaction time for these experiments are shown in fig. 3A. Interestingly, we found the final substrate yield of reaction catalyzed by free heme to decrease with incubation time. This decrease is most likely linked to heme self-association during incubation time and the consequent reduction in the concentration of the more active heme monomer. In support of this interpretation, the alteration in the Soret band known to be linked to heme self-association (Inada and Shibata 1962) occurs in the same time scale as does the decrease in peroxidase activity observed upon incubation (figs. 3B and 3C). Heme bound to the ancestral protein cannot associate with other heme molecules and it is therefore protected from the loss of peroxidase activity caused by self-association.

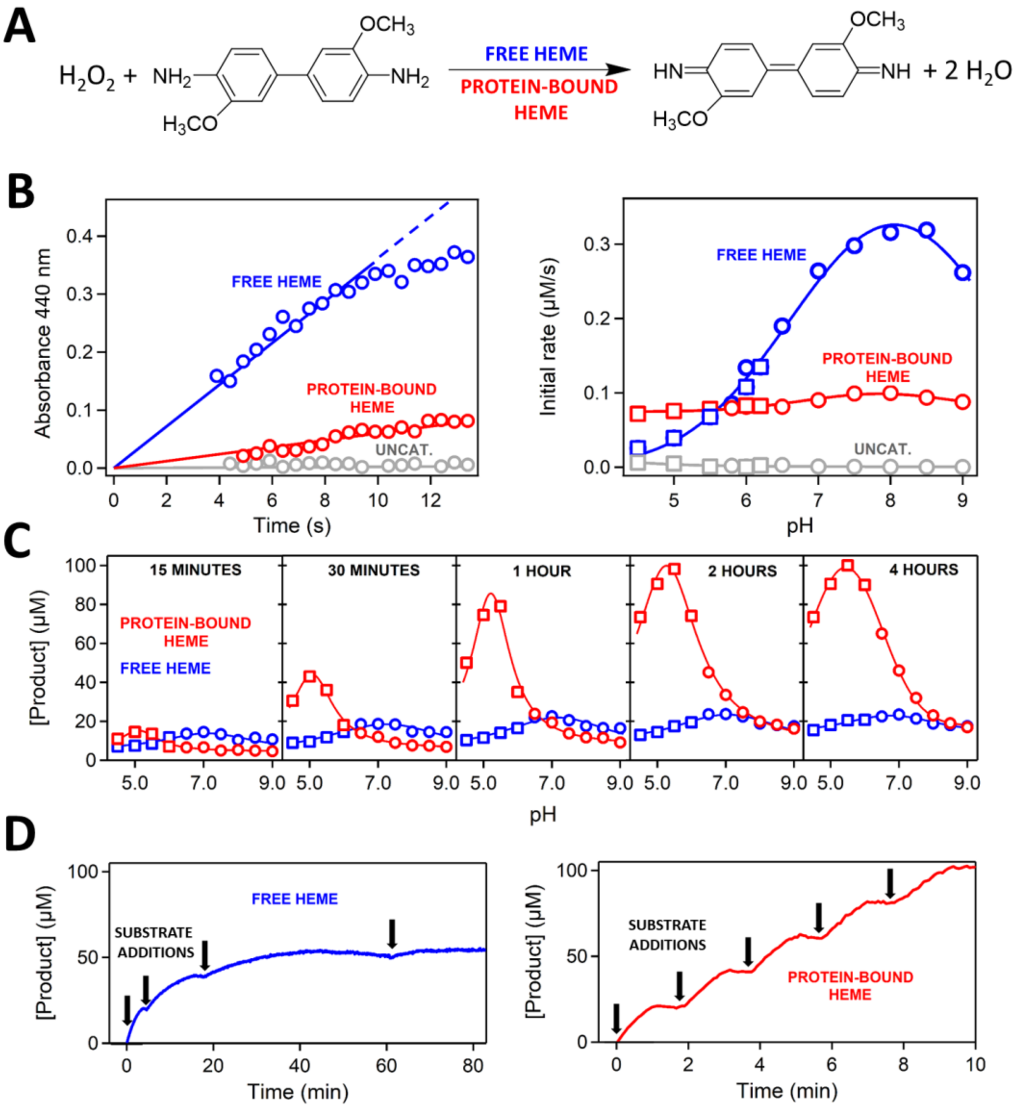


Fig. 2. Peroxidase activity of free heme versus heme bound to the ancestral TIM-barrel. (A) Reaction used to test peroxidase activity. The kinetics of o-dianisidine peroxidation can be determined by following the increase of absorbance at 440 nm upon peroxidation. (B) Left: representative examples of the profiles of absorbance at 440 nm versus time used to calculate the initial rates of the reaction. The example shown corresponds to pH 7, heme concentration of 0.4 µM (either free or protein bound) and initial concentrations of hydrogen peroxide and o-dianisidine of 10 mM and 0.1 mM, respectively. The data show that the initial reaction rate is higher with free heme as catalyst. Right: plot of initial rate versus pH for free heme, protein-bound heme and a control experiment in the absence of heme (labeled “uncat”). In most of the studied pH range, initial rates are higher with free heme as catalyst. Heme and initial substrate concentrations are the same as in B. Symbols refer to the buffer used: squares, 200 mM acetate, 150 mM NaCl; circles, 200 mM phosphate, 150 mM NaCl. (C) Profiles of amount of peroxidation product formed versus pH over long reaction times (shown inside the plots). As the reaction time increases, bound heme progressively becomes a more efficient peroxidation catalyst than free heme. Buffers, heme concentration and initial substrate concentrations are the same as in B. (D) Peroxidation kinetics over long reaction times with free heme (left) and protein-bound heme (right) as catalysts with repeated additions of 0.1 mM o-dianisidine substrate (labeled with arrows). Degradation of free heme catalysis is apparent by the progressive absence of peroxidation of freshly added o-dianisidine. By contrast, no such degradation is observed with protein-bound heme. The data correspond to pH 7, heme concentration of 2 µM (either free or protein bound) and initial concentration of hydrogen peroxide of 10 mM.

It is also clear from the strong curvature and the leveling-off of the kinetic profiles at long reaction times (fig. 3A and fig. 2D) that the degradation of the peroxidase activity of free heme is substantially accelerated after substrate addition. This points to an additional mechanism of inactivation acting during the catalytic cycle, likely involving chemical alterations of the heme caused by reactive oxygen species, as it has been described for heme in peroxidases (Valderrama et al 2002) and for free heme in solution (Brown et al 1968; Brown and Jones 1968). Remarkably, the loss of peroxidase activity during the catalytic cycle appears to be considerably limited in the restricted and defined molecular environment provided by the ancestral TIM-barrel for the bound heme (figs. 3A and 2D).

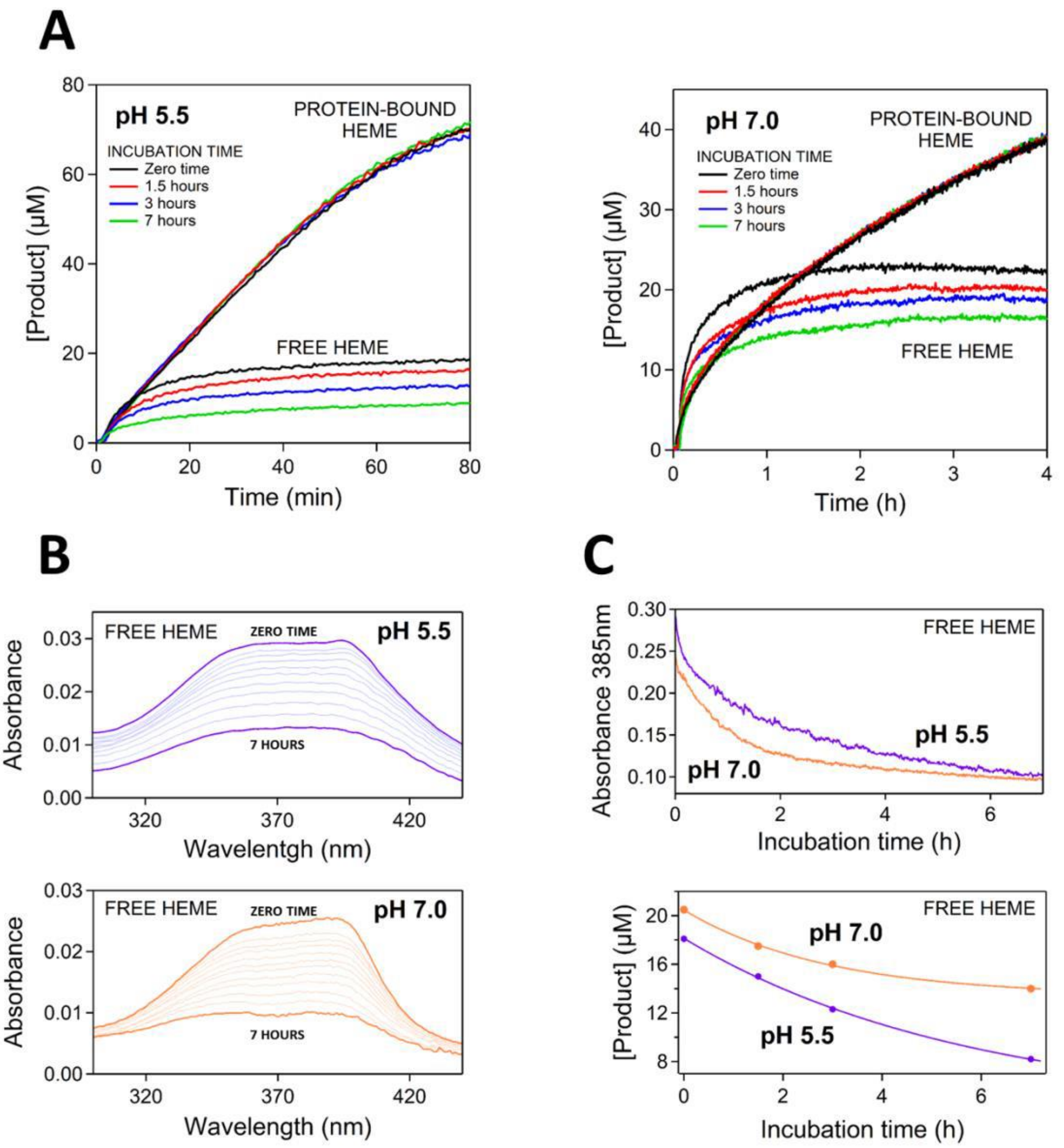


Fig 3. Degradation of catalysis by free heme during and before the peroxidation reaction. (A) Peroxidation kinetics over long reaction times with free heme and protein-bound heme as catalysts for pH 5.5 and pH 7. In all experiments, the heme concentration (either free protein-bound) was 0.4 µM and the initial concentrations of o-dianisidine and hydrogen peroxide were 0.1 mM and 10 mM, respectively. The experiments differ in the incubation duration, i.e., the time free heme or protein-bound heme were kept in the solution before addition of the substrates. The curvature of the kinetic profiles for free heme reveals severe catalysis degradation during the peroxidation reaction. Yet, substantial degradation also occurs during the incubation time as shown by decreasing final product yields with increasing incubation time. (B) Self-association of heme as revealed by flattening of the Soret band of free heme in solution. Spectra were collected approximately every 30 minutes from the dilution of the heme stock in the buffer (zero time) to seven hours. See legend to fig. 2 and Methods for buffer composition. (C) Plots of absorbance at 385 nm versus time (from the data given in B) and plots of final product yield versus incubation time (from the data given in A for free heme) Self-association occurs in the same time scale as (and it is the likely cause of) the degradation of catalytic potential of free heme during the incubation time.

# Discussion

Heme binding at a conformationally flexible region of an ancestral TIM-barrel leads to an enhancement in the efficiency of heme as a peroxidation catalyst, as shown by the increase in peroxidation reaction yield over a time scale of hours. Determination of initial reaction rates, however, indicates that binding does not improve the intrinsic peroxidase activity of heme. In fact, the opposite is true within most of the studied pH range: initial reaction rates are somewhat lower with the protein-heme complex as compared with free heme, likely reflecting limited substrate access to the substantially-buried bound heme. Therefore, the enhancement in the efficiency of peroxidation catalysis upon heme binding to the protein scaffold cannot be attributed to promotion of cofactor catalysis by the protein moiety, as is the case with modern peroxidases (fig. 1D). In fact, our experimental data (figs. 2 and 3) clearly support that the enhancement is due to the retardation of processes that impair the peroxidase activity of heme, resulting in a much longer life time and a higher effective concentration for the catalyst when it is bound to the protein.

The ancestral TIM-barrel we have used as protein scaffold in our experiments is a putative ancestor of bacterial and eukaryotic family-1 glycosidases (Gamiz-Arco et al 2021) and may perhaps probe an early stage in glycosidase evolution. However, in the context of the evolutionary history of proteins, it belongs to a period long after the generation of the first enzymes. Yet, the consequences for catalysis of heme binding to a conformationally flexible region of the ancestral TIM-barrel point to and mimic a plausible scenario for the primordial emergence of enzymes. We briefly elaborate this scenario below.

Reasonable chemical mechanisms for the prebiotic formation of polypeptides have been proposed (Frenkel-Pinter et al 2020). Indeed, it has been hypothesized that polypeptides already existed in the primordial RNA world, where they served as enhancers of ribozyme activity (Romero et al, 2016; Wolf and Koonin 2017). Regarding protein cofactors, their extreme evolutionary conservation strongly suggests that they are very ancient (Chu and Zhang 2020). The likely origin of inorganic cofactors in the geochemical environment in which life begun and that of organic cofactors (coenzymes) as parts of primordial ribozymes have been noted (Goldman and Kacar 2021). Overall, there can be little doubt that polypeptides and cofactors co-existed at some primordial stage. Then, their association may have immediately promoted catalysis by increasing the life time and the effective concentration of catalytic cofactors. It follows that, as soon as some means of transmission of genetic information, even if rudimentary and error-prone, was available, natural selection for polypeptides with increasing capability to protect catalytic cofactors became possible. This selection could have also favored cofactors that interact efficiently with polypeptides and become readily protected. Lastly, the stable and catalytically competent polypeptide-cofactors complexes thus formed provided starting points for the Darwinian evolution of a protein molecular machinery that assisted and further enhanced cofactor-based catalysis.

The protection mechanism for catalysis enhancement has been demonstrated in this work on the basis of the peroxidase activity of heme, but it should apply to many other cofactors. The emergence of enzymes based on catalytic iron-sulfur clusters may be a particularly relevant example. Iron and sulfur were abundant in the geochemical environment that likely hosted primordial life (Beinert 2000) and enzymes based on iron-sulfur clusters abound in reconstructions of the gene content of LUCA (Weiss et al 2016). Free iron-sulfur clusters have been found to mimic the basic features of protein-bound clusters, but in non-aqueous media and in the absence oxygen (Beinert 2000). Even if destruction by reaction with oxygen was not an issue in an anaerobic primordial environment (Weiss et al 2016), iron-sulfur clusters appear as obvious candidates for primordial catalysis enhancement through protection by polypeptides (Kim et al 2018).

Ferredoxins are iron-sulfur proteins that perform electron transfer in a diversity of biochemical transformations. About 60 years ago, Margaret Dayhoff noted the simplicity of ferredoxins, which consist of an inorganic active site and a very short polypeptide which she proposed to have emerged by duplication of smaller peptides (Eck and Dayhoff 1966). Modern ferredoxins may, therefore, be relics of primordial protection events. An intriguing, albeit speculative, possibility is that heme binding to our ancestral TIM-barrel glycosidase is also a relic of primordial protection, although, in this case, the ancestral functional feature has undergone evolutionary degradation and it is only observed in vestigial form in modern glycosidases (Gamiz-Arco et al 2021).

# Methods

The ancestral TIM-barrel glycosidase was prepared as previously described (Gamiz-Arco et al., 2021). Briefly, the gene for the His-tagged protein in a pET24 vector was cloned into E. coli BL21 (DE3) cells and the protein was purified using Ni-NTA chromatography. Protein prepared in this way typically has a very small amount of bound heme. Ancestral protein saturated with heme was prepared by incubation with 5-fold excess of heme followed by size-exclusion chromatography to eliminate non-bound heme, as we have previously described in detail (Gamiz-Arco et al 2021). The heme to protein ratio was found to be close to unity on the basis of absorbance determinations for the protein band at 280 nm and the heme Soret band.

As previously described (Gamiz-Arco et al., 2021), heme solutions were prepared by high-dilution (typically 1:1000) in the desired buffer of a stock solution in concentrated sodium hydroxide and used immediately. Stock solutions of heme were prepared daily. Heme concentrations in stock solution were determined from the absorbance of the heme Soret band at 385 nm using a known value of the extinction coefficient (Deniau et al 2003). Stock solutions of o-dianisidine were prepared by weight. Stock solutions of hydrogen peroxide were prepared by dilution of commercially available stock solution and their concentrations were determined from the absorbance at 240 nm using a known extinction coefficient (Jiang et al 1990). The peroxidation reaction was initiated by adding microliter volumes of the reactants to a 2-mL solution containing free heme or protein-bound heme. The reaction was followed by measuring the absorbance of the peroxidation product at 440 nm. A known value of the extinction coefficient (Jenkins et al 2021) was used calculate substrate concentration from absorbance values. Peroxidation experiments were performed in wide pH range using the following buffers: 200 mM acetate, 150 mM NaCl for the pH range 4.5-6.2 and 200 mM phosphate, 150 mM NaCl for the pH range 5.8-9.0.

# References

Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM. 2008. Metal ions in biological catalysis: from enzyme databases to general principles. J Biol Inorg Chem 13:1205-1218.

Beinert H. 2000. Iron-sulfur proteins: ancient structures, still full of surprises. J Biol Inorg Chem 5:2-15.

Blomberg R, Kries H, Pinkas DM, Mittl PRE, Grütter MG, Privett HK, Mayo SL, Hilvert D. 2013. Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature 503:418-421.

Brown SB, Jones P, Suggett. 1968. Reactions between haemin and hydrogen peroxide. Part1.-Ageing and non-destructuve oxidation of haemin. Transactions of the Faraday Society 64:986-993.

Brown SB, Jones P. 1968. Reactions between haemin and hydrogen peroxide. Part 2.-Destructive oxidation of heamin. Transactions of the Faraday Society 64:994-998.

Brown SB, Dean TC, Jones P, Kremer ML. Catalytic activity of haemin. 1970. Transactions of the Faraday Society 66:1485-1490.

Campbell E, Kaltenbach M, Correy GJ, Carr OD, Porebski BT, Livingstone EK, Afriat-Jurnow L, Buckle AM, Weik M, Hollfelder F, Tokuriki N, Jackson CJ. 2016. The role of protein dynamics in the evolution of new enzyme function. Nat Chem Biol 12:944-950.

Chu XY, Zhang HY. 2020. Cofactors as molecular fossils to trace the origin and evolution of proteins. ChemBioChem 21:3161-3168.

Deniau C, Gilli R, Izadi-Pruneyre N, Létoffé, Delepierre M, Wandersman C, Briand C, Lecroisey A. 2003. Thermodynamics of heme binding to the HasASM hemophore: effect of mutations and three key residues for heme update. Biochemistry 42:10627-10633.

Donnelly AE, Murphy GS, Digianantonio KM, Hecht MH. 2018. A *de novo* enzyme caalyzes a life-sustaining reaction in Escherichia coli. Nat Chem Biol 14:253-255.

Eck RV, Dayhoff MO. 1966. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science 152:363-366.

Fischer JD, Holliday GL, Rahman SA, Thornton JM. 2010. The structures and physicochemical properties of organic cofactors in biocatalysis. J Mol Biol 403:803-824.

Frenkel-Pinter M, Samanta M, Ashkenasy G, Leman LJ. 2020. Prebiotic peptides: molecular hubs in the origin of life. Chem Rev 120:4707-4765.

Gamiz-Arco G, Gutierrez-Ruz LI, Risso VA, Ibarra-Molero B, Hoshino Y, Petrovic D, Justicia J, Cuerva JM, Romero-Rivera A, Seelig B, Gavira JA, Kamerlin SCL, Gaucher EA, Sanchez-Ruiz JM. 2021. Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase. Nat Comunn 2:380.

Goldman AD, Beatty JT, Landweber LF. 2016. The TIM barrel architecture facilitated the early evolution of protein-mediated metabolism. J Mol Evol 82:17-26.

Goldman AD, Kacar B. 2021. Cofactors are remnants of life's origin and early evolution. J Mol Evol 89:127-133.

Inada Y, Shibata K. 1962. Soret band of monomeric hematin and its changes on polymerization. Biochem Biophys Res Commun 9:323-327.

Jacob F. 1977. Evolution and tinkering. Science 196:1161-1166.

Jenkins JMX, Noble CEM, Grayson KJ, Mulholland AJ, Anderson R. 2021. Substrate promiscuity of a *de novo* designed peroxidase. J Inorg Chem 217:111370.

Jiang ZY, Woollard ACS, Wolff SP. 1990. Hydrogen peroxide production during experimental protein glycation. FEBS Lett 268:69-71.

Khersonsky O, Tawfik DS. 2010. Enzyme promiscuity: a mechanistic and evolutioanary perspective. Annu Rev Biochem 79:471-505.

Kim JD, Pike DH, Tyryshkin AM, Swampa GVT, Raanan H, Montelione GT, Nanda V, Falkowski PG. 2018. Minimal heterochiral *de novo* designed 4Fe-4S binding peptide capable of robust electron transfer. J Am Chem Soc 140:11210-11213.

Lovelock SL, Crawshaw R, Basler S, Levy C, Baker D, Hilvert D, Green AP. 2022. The road to fully programmable protein catalysis. Nature 606:49-58.

Nagano N, Orengo CA, Thornton JM. 2002. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 321:741-765.

Ohno S. 1970. Evolution by gene duplication. Berlin, Germany, Springer.

Ortiz de Montellano PR. 2010. Catalytic mechanisms of heme peroxidases. In E. Torres, M Ayala (eds.) "Biochatalysis based on heme peroxidases" p 79-107. New York, USA, Springer.

Poulos TL, Kraut J. 1980. The stereochemistry of peroxidase catalysts. J Biol Chem 255:8199-8205.

Risso VA, Martinez-Rodriguez S, Candel AM, Krüger DM, Pantoja-Uceda D, Ortega-Muñoz M, Santoyo-Gonzalez F, Gaucher EA, Kamerlin SCL, Bruix M, Gavira JA, Sanchez-Ruiz JM. *De novo* active sites for resurrected Precambrian enzymes. Nat Commun 8:16113.

Romero MLR, Rabin A, Tawfik DS. 2016. Functional proteins from short peptides: Dayhoff's hypothesis turns 50. Angew Chem Int Ed 55:15966-15971.

Valderrama B, Ayala M, Vazquez-Duhalt R. 2002. Suicide inactivation of peroxidases and the challenge of engineering more robust enzymes. Chem Biol 9:555-565.

Weiss MC, Soussa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. Nat Microbiol 1:16116.

Wierenga RK. 2001. The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett 492:193-198.

Wolf YI, Koonin EV. 2007. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. Biol Direct 2:14.

Yeh AHW, Norn C, Kipnis Y, Tischer D, Pellock SJ, Evans, D, Ma P, Lee GR, Zhang JZ, Anishchenko I, Coventry B, Cao L, Dauparas J, Halabiya S, DeWitt M,

Carter L, Houk KN, Baker D. 2023. *De novo* design of luciferases using deep learning. Nature 614:774-780.

Zeymer C, Hilvert D. 2018. Directed evolution of protein catalysts. Annu Rev Biochem 87:131-157.

## Acknowledgments

This work was supported by Human Frontier Science Program grant RGP0041/2017 (J.M.S.R. and E.A.G.), National Science Foundation grant 2032315 (E.A.G.), Department of Defense grant MURI W911NF-16-1-0372 (E.A.G.), National Institutes of Health grant R01AR069137 (E.A.G.), Spanish Ministry of Science and Innovation / FEDER Funds grant PID2021-124534OB-100 (J.M.S.R.) and grant PID2020-116261GB-I00 (J.A.G.).

## Author Contributions

L.I.G.R. carried out protein preparation, and designed and performed the experimental determination peroxidase activity under the supervision of V.A.R.; G.G.A. provided essential input regarding the catalytic properties of the ancestral glycosidase; J.A.G. provided essential input regarding the structural interpretation of the protection mechanisms; E.A.G provided essential input regarding the interpretation of the data in an evolutionary context; J.M.S.R. designed the research and wrote the first draft of the manuscript; all authors discussed the manuscript, suggested modifications and improvements and contributed to the final version.

## Data Availability

All relevant experimental data are provided in the main-text figures. Tables with the data are available upon reasonable request.

Conflict of interest statement: The authors declare no competing interests.

# DISCUSSION

## Ancestral proteins as scaffolds for the engineering and evolution of artificial *de novo* active sites

Optimization of artificial active sites through directed evolution is the most straightforward approach to improve enzymatic activity in *de novo* enzyme design, regardless of whether *de novo* active sites are generated through computational rational or minimalistic design. In a typical directed evolution experiment, different strategies are usually followed by combining focused or non-focused random mutagenesis, along with shuffling mutant variants in order to generate a wider sequence diversity during the evolutionary process. The iterative application of directed evolution and the accumulation of beneficial mutations usually leads to variants with enhanced catalytic parameters, reaching high levels of enzymatic catalytic activity.  This is best illustrated by using as an example the best designed Kemp eliminase enzyme described in the literature. This artificial enzyme was designed completely *de novo* by using a TIM-barrel modern scaffold as a starting point. In this example, the authors generated an initial Kemp eliminase active site through an iterative and complex approach that combined computational design, X-ray crystallography and molecular dynamics simulations. The theozyme design was carried out in the binding pocket of the xylanase from *Thermoascus aurantiacus*. The resulting enzyme named HG3 differed from the wild type scaffold in 18 mutations located in the interior of the TIM-barrel and displayed a relatively low but considerable enzymatic activity with a $k_{cat}$ of 0.68 $s^{-1}$ and a $k_{cat}/K_M$ of 425 $M^{-1}$ $s^{-1}$. Further intensive optimization of the initial enzyme after 17 rounds of directed evolution and the introduction of 17 additional mutations yielded a highly active variant called HG3.17 displaying a $k_{cat}$ of 700 $s^{-1}$ and a $k_{cat}/K_M$ of 230000 $M^{-1}$ $s^{-1}$. The analysis of the evolutionary trajectories and outcomes during the optimization process revealed the importance of preorganization in the active site to stabilize the transition state and the correct positioning of catalytic groups as key components in generating a highly efficient *de novo* active site[39,86]. Additionally, during the iterative directed evolution steps, the role of conformational dynamics in catalysis was demonstrated by showing how enzymatic efficiency was increased through a gradual conformational selection towards conformational productive states over all unproductive conformations[125]. This work marked a milestone in *de novo* enzyme design by revealing fundamental mechanisms for the emergence, evolution and optimization of catalysis in an inactive protein scaffold. However, it also demonstrated that a considerable amount of experimental workforce and resources are still required to achieve functional and highly active *de novo* artificial enzymes. As stated in the hypothesis and objectives section of this thesis, we propose that the difficulty in designing *de novo* enzymes resides in the use of highly specialized and evolved modern proteins as scaffolds, which likely hampers the efficient and routinely generation of new artificial enzymes.

In the work presented in this thesis, we have demonstrated that comparable levels of *de novo* enzymatic catalysis can be achieved in a resurrected ancestral protein scaffold by following an easier minimalistic design approach and a low-throughput screening of mutant variants directed towards improvement of Kemp elimination reaction. This highlights the superior properties of resurrected ancestral proteins as scaffolds for the generation, evolution and optimization of *de novo* enzymatic activities in comparison with modern protein scaffolds. In the first place, an initial Kemp eliminase activity was designed into a resurrected ancestral β-lactamase by generating a minimalistic active site consisting of two single mutations, W229D and F290W[207]. The resulting enzyme named GNCA4 W229D/F290W displayed parameters for the Kemp elimination reaction with a $k_{cat}$ of 12 $s^{-1}$ and a $k_{cat}/K_M$ of 5500 $M^{-1}$ $s^{-1}$, considerable higher than those initially obtained in the computationally designed HG3 ($k_{cat}$ of 0.68 $s^{-1}$ and $k_{cat}/K_M$ of 425 $M^{-1}$ $s^{-1}$). The initial activity was achieved by introducing a minimalistic active site into a structural region of the protein scaffold with high conformational flexibility. This allowed the exploration of catalytically active conformations that are only found in ancestral scaffolds and not in modern ones[207]. The most direct implication of this study was that ancestral proteins likely outperform modern scaffolds for *de novo* enzyme design. This was demonstrated by the fact that only two mutations, guided by a "rational chemical intuition" approach, were needed to create an efficient artificial active site that only worked in the resurrected protein scaffolds. In contrast, following a complex iterative computational and experimental approach in a modern scaffold resulted in a total of 18 mutations to generate a less efficient active site. This thesis presents the results obtained on the later stages of optimization of the artificial enzymatic activity in the ancestral scaffold.

The results revealed that by using simpler experimental approaches, a comparable level of enzymatic catalysis can be achieved more easily in a resurrected ancestral scaffold, similar to that obtained in HG3.17 in which a much more complex and laborious process is required. In a first optimization step, we have improved the active site of our *de novo* Kemp eliminase based in a resurrected ancestral β-lactamase by following a computationally guided ultra-low throughput screening. The computational selection of mutations was performed by using FuncLib[245], an automated method for the design of multipoint mutations at enzyme active sites that uses phylogenetic information and Rosetta design calculations in order to identify plausible mutations that generate or maintain stabilizing interacting networks at the active site. The main goal of using FuncLib is to bypass high-throughput screening by testing a low number of variants predicted to preserve the overall stability, and therefore the activity, of the background enzyme. We applied FuncLib predictions targeting 11 residues to diversify located at the active site of GNCA4 W229D/F290W and screened just the top 20 variants from the prediction. As a result, the 20 tested variants remained stable with similar values of melting temperatures as determined by differential scanning calorimetry (DSC) and displayed substantial levels of Kemp elimination catalysis, in some cases even with enhanced catalytic parameters. This low throughput screening allowed to identify one specific variant named GNCA4-12 with enhanced catalysis, showing a $k_{cat}$ of ~100 $s^{-1}$ and a $k_{cat}/K_M$ of ~20000 $M^{-1}$ $s^{-1}$, which compares well with the best designed Kemp eliminase HG3.17, and with no concomitant loss of thermostability[221]. X-ray crystallography studies of the best variant GNCA4-12 with the transition state bound to the active site revealed that, despite the relatively large number of simultaneous mutations introduced in the artificial *de novo* active site (5 mutations), large conformational changes were not detected in comparison to the background scaffold. This indicated that the enhancement in catalysis was likely linked to small rearrangements in the *de novo* active site. Empirical valence bond (EVB) calculations and molecular dynamics (MD) simulations were performed and validated with the FuncLib predictions, showing a good agreement between the experimental and calculated values for activation free energy ($\Delta G^{\ddagger}$). But most importantly, evaluation of the calculated EVB trajectories and analysis of the individual contributions of each active site residue to the activation free energies revealed that the most significant component associated with the increased activity is a better geometric preorganization of the active site for an efficient proton abstraction from the substrate[221].

The results obtained from the first optimization step of the *de novo* artificial Kemp eliminase revealed important lessons with evolutionary and engineering implications. First, this study revealed that large enhancements of the catalytic activity in a *de novo* designed active site can be easily achieved through the introduction of a small number of mutations in a resurrected ancestral protein scaffold. In our ancestral lactamase-based Kemp eliminase, the introduction of just 5 computationally guided mutations in the active site, after screening an extremely low number of variants, led to an increase of about one order of magnitude in both $k_{cat}$ and a $k_{cat}/K_M$ values. This demonstrates that *de novo* catalysis in ancestral protein scaffolds can be easily optimized to reach higher levels of enzymatic catalysis, at least as easily as in modern scaffolds. Additionally, the results showed that the large increase in the enzymatic activity is achieved by improving the preorganization of the active site towards a better stabilization of the transition state without the need for corresponding significant active site rearrangements. These results agree with the observed optimization trajectories in HG3.17 identified through directed evolution and reaffirm the idea that catalytic enhancement through optimization of transition state stabilization stands as an immediate evolutionary molecular mechanism for the early optimization of novel active sites in enzymes. Secondly, from the protein engineering applications point of view, the results obtained by using FuncLib as a computational tool to guide the evolution of the artificial enzyme holds promise for further applications directed towards enhancement of engineered or natural enzymes. It is important to note that the enhancements in catalytic activity reported in this thesis have been obtained by following a procedure not intended for predicting catalytically favorable mutations. Instead, FuncLib is designed to sharply focus the search of the targeted residues to regions of the protein sequence space that encode stable proteins, without considering the protein structure or elements involved in the transition state stabilization. Therefore, application of FuncLib for protein engineering implies a more optimized screening effort, which is not wasted in exploring variants that may not fold properly or in structural regions which are not expected to be relevant for catalysis. Additionally, EVB calculations and MD simulations showed that FuncLib variants with enhanced catalytic activity could be potentially predicted by calculating the free energy values of the different mutants. EVB calculations could be therefore applied as a second filter to the top FuncLib predictions to identify plausible catalytically enhanced variants. Overall, our results support the idea that the application of FuncLib combined with EVB calculations may provide an efficient computational pipeline to speed up enzyme evolution by guiding the screening towards regions of the protein sequence space catalytically relevant and safe in terms of stability. This holds promise in systems where more complex catalytic machineries or mechanisms are involved in the enzymatic reactions and where larger changes in activity can be achieved, especially in the early stages of the directed evolution process. Taken together, the presented experimental and computational work collectively highlights the significant potential of FuncLib's evolutionary-based stability-screening protocol as a valuable tool in computational *de novo* enzyme design. Additionally, it demonstrates the potential of using ancestral enzymes as starting scaffolds for the design and engineering of artificial enzymes.

In the second optimization step, the Kemp elimination enzymatic activity was further enhanced in the resurrected ancestral scaffold by using directed evolution. However, instead of using an error prone PCR to introduce random mutations throughout the full protein sequence, mutagenesis was focused in a particular region of an extended engineered segment. As the artificial active site is located close to the C-terminal of the ancestral β-lactamase, we introduced an extra polypeptide segment in the C-terminal which contained 6 glycine residues along with some other residues. Then, the first three residues of the engineered extension (VGG) were targeted for focused directed evolution in order to explore the plausibility of generating new favorable interactions close to the active site, aiming to increase the enzymatic activity. A low throughput screening of about 800 clones was performed in order to identify highly active variants. As a result, one specific variant named GNCA12-V4 was identified to display substantially higher catalytic activity[222]. Purification and characterization of the engineered version of the Kemp eliminase revealed that GNCA12-V4 displayed enhancements of about one order of magnitude in its catalytic parameters reaching values of $k_{cat}$ of ~640 $s^{-1}$ and a $k_{cat}/K_M$ of ~200000 $M^{-1}$ $s$-$^1$, again without a concomitant loss of thermostability and reaching the same catalytic activity in terms of catalytic constant and catalytic efficiency as the best engineered Kemp eliminase described to date HG3.17.

A closer inspection of the engineered GNCA12-V4 revealed that the background sequence VGG was replaced for GLR. The introduction of a bulky hydrophobic (leucine) and a positively charged polar (arginine) residues close to the active site directly suggested that new interactions may be involved in the large enhancement of the enzymatic activity, most likely by favoring transition state stabilization. AlphaFold2 predictions of the 3D structure of GNCA12-V4 were performed in order to infer a plausible explanation for the catalytic enhancement. Essentially, the hydrophobic leucine residue generated during the directed evolution process is predicted to be positioned directly pointing to spot where the substrate and transition state are located in the active site, suggesting a stabilizing interaction with the transition state. Additionally, the positively charged arginine residue is predicted to be interacting through a hydrogen bond with an aspartic residue located right before the engineered aspartic residue that acts as the catalytic base in the proton abstraction step of the Kemp elimination reaction. This interaction could be essential for the correct positioning of the catalytic aspartic acid. But also, it could be facilitating the stabilizing interaction between the newly generated leucine and the transition state of the reaction. In any case, it appears that the engineered extended polypeptide is positioned in the active site acting as a "closing lid" that promotes a better positioning of catalytic residues and improves the stabilization of the active site. The emergence of a lid structure that enhances a *de novo* artificial activity has already been observed during the directed evolution of an artificial Diels-Alderase[90].

Again, the findings obtained from the second optimization step of the *de novo* artificial Kemp eliminase uncovered important insights with implications in protein evolution and engineering. This study reveals how the introduction of new extra residues in a protein sequence by following natural evolutionary processes may lead to the emergence of new interactions in the active site that improve the catalytic activity. This highlights the role of evolutionary mechanisms in protein evolution such as duplications or insertions, with the potential to introduce new residues and interactions that may lead to "saltation" events that bring about significant changes in a single mutational step. These mechanisms may be particularly relevant during the early evolution and optimization of novel enzymatic activities[246,247]. On the other side, our success in arriving at an efficient Kemp eliminase has immediate implications for the engineering of natural and engineered enzymes. We demonstrate that the insertion of short polypeptide segments in the enzyme sequence and close to the active site is a valid protein engineering approach with the potential to generate new transition state stabilizing interactions that enhance the enzymatic catalysis. This approach has as a main advantage the fact that as the segment is, in principle, flexible and exposed to the solvent, it is not expected to generate disruptive interactions with the protein that may compromise proper folding. Also, as the segment length is relatively short, the associated sequence space can be extensively explored through directed evolution with a moderately low screening effort. Ultimately, this engineering strategy provides a proof of principle for an underexplored protein engineering approach with the potential to enhance enzymatic activity even orders of magnitude.

In conclusion, our engineered *de novo* artificial Kemp eliminase highlights the benefits of using ancestral proteins as scaffolds for evolutionary and engineering studies. Our highly efficient Kemp eliminase was generated from a resurrected ancestral β-lactamase scaffold and displays the same catalytic parameters as the best base-catalyzed Kemp eliminase described in the literature. However, the experimental effort invested in this project has been substantially lower. The initial catalytic activity in our ancestral scaffold was designed based on a minimalistic catalytic machinery consisting of two mutations, using a chemical intuition approach. On the other side, the generation of the active site in the modern scaffold required a substantially higher amount of experimental workforce and resources, leading to a more complex and less efficient active site composed by 18 mutations. We then achieved rapid enhancement of our ancestral-scaffold based Kemp eliminase by screening 20 computationally predicted and 800 directed evolution generated variants. In contrast, the modern-scaffold based Kemp eliminase required 17 rounds of directed evolution, involving different diversity-generating strategies and screening more than 15000 clones, to achieve the same catalytic parameters. Through this process, our ancestral protein served as an effective scaffold to demonstrate new protein engineering methodologies and uncover the role of different evolutionary mechanisms in the emergence and evolution of new enzymatic activities. These studies have made remarkable progress in the field of resurrected ancestral proteins with significant implications for protein engineering and evolution. They support the main hypothesis of this thesis, which proposes that using ancestral proteins as scaffolds with a lower degree of structural and functional specialization may be a more effective strategy for comprehending the emergence, evolution, and design of new enzymatic functionalities.

## Ancestral proteins as scaffolds to reveal unexpected combinations of properties in the unexplored sequence space

Ancestral sequence reconstruction (ASR) and resurrection of the encoded proteins is commonly conceived as a methodology to navigate unexplored sequence space with an evolutionary-guided focus, that generally narrows the search to stable and functional proteins. One of the main outcomes of ASR is the generation of new proteins that may display extreme or unusual properties of interest based on their ancient features. Ancestral proteins can be then used as scaffolds for protein engineering experiments that seek to exploit their extreme ancient properties for different biotechnological applications[136,137]. But also, these proteins can be used as models to study and understand evolutionary processes associated with their ancestral features[135].

In this thesis, we report the reconstruction and resurrection of an ancestral TIM-barrel glycosidase, corresponding to a putative ancestor of bacterial and eukaryotic family 1 glycosidases, that displays unexpected properties with evolutionary and engineering implications[156]. Firstly, our ancestral TIM-barrel glycosidase displays a much higher conformational flexibility in comparison with its modern counterparts. In particular, flexibility is greatly enhanced in a large region of the TIM-barrel structure as demonstrated by experimental and computational results. This flexible domain appears as a disordered missing region in our X-ray crystallography studies. However, the disordered part of the structure does not compromise the foldability or stability of the overall protein maintained by a rigid barrel core structure. Flexibility also does not impair the natural glycosidase activity. Yet, our ancestral glycosidase displays lower levels of enzymatic catalysis with no preference for a particular glycosidase substrate. The lower and unspecialized catalytic activity is likely linked to the enhanced conformational diversity of the TIM-barrel, that leads the protein to explore catalytically unfavorable conformations and display a similar affinity for different types of sugars. From an evolutionary perspective, this combination of high thermostability, enhanced conformational flexibility and lower levels of unspecific catalysis are typical features observed in resurrected ancestral proteins that may reflect an early stage in the evolution of family 1 glycosidases[133,136,137]. As TIM-barrel proteins have been demonstrated to have a modular evolutionary origin mostly led by duplication and fusion events of smaller fragments[232–236], our ancestral TIM-barrel may represent an early ancestor that followed a fragment fusion event. Therefore, the highly flexible and disordered region would be consistent with the hypothesis of a recently fused domain that still displays an inefficient packing with the rest of the structure and a lack of conformational rigidity. Moreover, the increased conformational flexibility of the ancestral TIM-barrel stands as a promising feature with a considerable potential from the protein engineering point of view. Flexibility is a key contributor of protein evolvability and the emergence of new functionalities, as flexible proteins may explore minor conformations that facilitate low levels of new catalysis that might be subsequently improved by evolution[199–203]. Therefore, our ancestral TIM-barrel holds promise as a perfect scaffold for the generation of *de novo* catalysis. It displays a typical TIM-barrel fold with a modular structure and a region composed by flexible loops that provide a stable platform to accommodate the engineering of different sequences and conformations designed towards specific enzymatic reactions. But most importantly, in addition to the intrinsic interesting properties of TIM-barrels, our protein displays typical ancestral features that also contribute to a higher evolvability. More specifically, the combination of a rigid core that provides high thermostability with an enhanced conformational flexibility in a specific region makes our ancestral TIM-barrel an excellent scaffold for the engineering of *de novo* artificial active sites. In conclusion, the combination of ancestral thermostability and flexibility together with the intrinsic evolvability-promoting features of TIM-barrels leads to new possibilities in biotechnological applications that may not be accessible by exploring the modern descendants of our ancestral protein.

However, the most exciting and unexpected property that we described in our ancestral TIM-barrel was the ability to bind a heme molecule[156,244]. Heme binding was found as an unexpected, serendipitous discovery during the characterization of our ancestral protein, as heme binding has not been reported in glycosidases or TIM-barrel proteins in the literature. Heme is bound tightly and stoichiometrically in a specific buried site located in the highly flexible and disordered region of the TIM-barrel. Upon heme binding, the ancestral TIM-barrel undergoes a structural rearrangement process that leads to the rigidification of the flexible region and an allosteric modulation of the enzymatic activity that increase the catalytic power of the enzyme. This allosteric modulation is likely linked to the rigidification in the binding region, that may lead to a general rigidification of the TIM-barrel and narrows the conformational ensemble of the enzyme towards more active conformations. From the evolutionary point of view, whether heme binding is a genuine property of our ancestral protein is not apparent at all. Heme binding could be just a fortuitus event derived from the reconstruction process facilitated by the enhanced conformational flexibility of the ancestral TIM-barrel. However, a closer inspection into the heme binding site reveals that heme is bound tight to the protein, displays interactions also found in natural heme binding proteins, and leads to a functional improvement in the enzymatic catalysis of the ancestral glycosidase. From these results it appears probable that heme binding did occur in our ancestral TIM-barrel and underwent a progressive evolutionary degradation leading to the rudimentary and variable ability to bind heme that we detected in modern family 1 glycosidases. If that was the case, this observation would fit in the general principle that features that become less functional are evolutionary degraded[141,155] and would suggest a complex evolutionary history for this family of enzymes. One possibility is that the ancestral TIM-barrel was involved in a fusion event with a heme containing domain. In this scenario, heme could have been advantageous for the ancestral glycosidase during the first evolutionary steps, as it could be providing the TIM-barrel with more rigidity and a higher enzymatic activity. However, subsequent optimization of the protein dynamics through intermolecular interactions that improved packing and the enzymatic activity. Hence, heme was not needed in that specific region after fusing with the TIM-barrel, leading to a subsequent evolutionary degradation.

It is important to note that, regardless of the evolutionary origins and implications, heme binding opens up a wide range of biotechnological and protein engineering implications. The combination of a TIM-barrel with the heme cofactor bound to it has not been previously described in the literature. Additionally, we have demonstrated that our heme binding ancestral TIM-barrel displays low but considerable levels of peroxidase activity, which indicates that the intrinsic redox catalytic power of the heme cofactor is conserved upon binding to the ancestral protein[244]. Heme is an essential cofactor of many natural enzymes, and it is involved in a wide diversity of redox and rearrangement enzymatic reactions. Moreover, heme enzymes can be engineered to catalyze unnatural chemical reactions powered-up by the catalytic power of heme[216]. Therefore, our heme binding TIM-barrel shows a huge potential to be used as a starting point for the engineering and evolution of new enzymatic functionalities, based on the catalytic power of heme and supported by the stability and flexibility of the protein scaffold. This is particularly relevant for the field of *de novo* artificial enzyme design, where "seeding" minimum levels of new functionalities stands as a critical bottleneck for this kind of protein engineering and design projects[216–218]. Overall, our results demonstrate the potential of ancestral protein resurrection as a methodology to navigate unexplored protein sequence space and generate new proteins that capture and display unusual combinations of properties compared to the repertoire of modern proteins, with important implications for engineering and evolutionary studies.

# Ancestral proteins as scaffolds to understand the emergence of enzymatic catalysis

Despite the significant progress achieve during the last decades about enzyme structure, function, and evolution, we still are unable to explain how enzymatic catalysis firstly emerged in proteins during the early steps of protein evolution. Natural emergence of completely new catalytic machineries in proteins has not yet been observed in nature, as most new enzymatic functionalities arise from the evolution and modification of previously existing active sites in the highly evolved modern protein repertoire[22]. Additionally, modern enzymes have been evolving through millions of years of natural selection that have completely remodeled and even replaced the original active sites. Therefore, deciphering the minimal requitements of enzymatic catalysis by studying modern active sites appears as an unsuccessful approach. In this context, we propose ancestral protein resurrection as a more convenient methodology to approach the understanding on the origins of enzymatic catalysis. Resurrection and characterization of ancestral proteins corresponding to very ancestral nodes close to the origin of life might lead to the reconstruction of primordial active sites that capture the minimal requirements for a specific enzymatic reaction.

Dayhoff's hypothesis postulated that proteins firstly emerged from shorter yet functional peptides that were recruited from a random pool of abiotically generated peptides by natural selection[1,2]. These peptides were likely selected based on their chemical or biological primordial functionalities, as well as other physiochemical features such as foldability, stability or solubility[248]. One popular view about these primordial peptides is their role during the origin of life of facilitating the transition from prebiotic chemistry to proto metabolisms by interacting with pre-existing ribozymes, organic cofactors and inorganic metals involved in the catalysis of a wide diversity of chemical transformations[249–251]. Subsequent evolution of the cofactor-binding peptides likely led to more complex enzymes with catalytic machineries that depended on organic and inorganic cofactors. The outcome of this process is reflected in the modern proteomes, where a substantial fraction of enzymes (around 50%) contain organic or inorganic cofactors in the architecture of their active sites required for catalysis[69]. However, the driving forces that led this evolutionary association between cofactors and peptides are still unknown, which makes a critical gap in our understanding about the emergence of cofactor-dependent enzymatic catalysis in the origin of life.

In this thesis, we have demonstrated the role of an ancestral heme binding TIM-barrel scaffold as a model to understand the origins of cofactor-based enzymatic catalysis. First, we determined that the heme binding site in our ancestral TIM-barrel does not contain any catalytic machinery. Still, our heme bound ancestral TIM-barrel displays peroxidase activity, but with a lower level in comparison with free heme in solution in terms of initial activity[244]. This is compatible with our observations, as the burial of heme in the protein structure and the absence of catalytic residues that promote catalysis should hinder the substrate access to the catalytic cofactor and impair catalysis. However, experiments at longer times of reaction revealed that, unlike the heme bound to the protein, free heme undergoes a gradual loss of catalytic power that leads to complete inactivation. This suggests that the ancestral TIM-barrel provides mechanisms of protection to heme upon binding, which prevents the cofactor from inactivation and promotes catalysis during the enzymatic reaction.

The main contributor of heme inactivation is aggregation in aqueous solution, driven by the extremely low solubility of the porphyrin ring, and which leads to aggregated forms of heme that impair the catalytic activity[252,253]. However, in our experiments the inactivation effect is substantially accelerated after substrate addition, which points to additional mechanisms of inactivation acting during the catalytic cycle, likely involving chemical alterations of heme caused by reactive oxygen species[254–256]. Overall, our experiments suggest that heme binding in the protein scaffold leads to a protection and an enhancement in the catalytic efficiency of the cofactor led by the retardation of processes that impair its catalytic power. When bound to the TIM-barrel, heme is located in a buried hydrophobic environment that prevents the cofactor from aggregation and, therefore, inactivation. This protection mechanisms against aggregation could be extrapolated to other cofactors with low solubility and could help to explain the role of protection against aggregation as a driving force in the origin of cofactor dependent enzymes. This is, the early association between soluble peptides and insoluble organic and inorganic cofactors would have facilitated the catalysis of such cofactors by increasing its effective catalytic concentration in solution and preventing their inactivation by aggregation. Similarly, protection against inactivation during the catalytic cycle leading to a longer catalytic lifetime can also be applied to other cofactors. This is particularly relevant to iron-sulfur (FeS) clusters, which are found in the active sites of many enzymes inferred to be already present in the primordial proteome of LUCA[12]. FeS clusters are involved in many redox biochemical reactions and are heavily inactivated by oxygen and other reactive oxygen species[257,258]. Additionally, recruiting in buried binding sites could also prevent cofactor from interacting and reacting with other molecules of the environment that may lead to degradation or inactivation of the cofactors[259,260]. Therefore, association between cofactors and peptides in the early steps of protein evolution could have facilitated longer lifetimes for the catalytic activities of the cofactors, led by a protection against unspecific inactivating reactions.

In conclusion, the implications for catalysis of heme binding to a buried and flexible region in our ancestral TIM-barrel point to and reflect a plausible evolutionary route that led to the emergence of primordial cofactor dependent enzymes that will be described below. It is well accepted that polypeptides were already present in the primordial RNA word, generated through prebiotic chemical mechanisms and playing a key role as enhancers of ribozyme activity[248,251,261,262]. At the same time, organic and inorganic cofactors were already present in the geochemical environments where life begun, coexisting with primordial peptides and involved in prebiotic chemistry as primordial ribozymes[261–265]. Our results suggest that cofactor protection against aggregation and other inactivation events had a role as an evolutionary driving force that facilitated the association between cofactors and peptides. This association was beneficial for the ancient cofactor molecules, as protection was essential to promote the catalytic power and increase the effective concentration of cofactors, assisting their incorporation as essential parts of the primordial enzymatic repertoire. But also, the formation of cofactor associations could have also been a driver for the recruitment of functional peptides in the primordial protein evolution. In particular, cofactors could have played a role as nucleation points for primordial cofactors that facilitated the correct special configuration and formation of very ancient folds and even increased the solubility of these primordial peptides[249]. Therefore, cofactor interaction could be viewed as a driving force that facilitated the selection and recruitment of functional, folded, and soluble proteins from a random pool of polypeptides. Consequently, the formation of peptide-cofactor complexes during the early stages of protein evolution determined the original architectures of cofactor-dependent active sites. These associations were led by different evolutionary driving forces and resulted in the formation of protein structures that provided protection mechanisms for the promotion of cofactor catalysis but lacked catalytic machineries for efficient enzymatic reactions. However, the stable and catalytically active peptide-cofactor complexes provided efficient starting points for subsequent Darwinian evolution that eventually led to the formation of efficient catalytic machineries that assisted and further improved the cofactor-based catalysis.

It appears evident that having access to a model scaffold that reflects an early unevolved stage of protein-cofactor association is just possible through ancestral protein resurrection. Studying modern cofactor dependent enzymes will lead to active sites with a high degree of structural evolution and highly optimized catalytic machineries that do not reflect early original architectures. Overall, our findings support the role of resurrected ancestral proteins as effective model scaffolds to understand the emergence and evolution of enzymatic catalysis. Ancestral scaffolds provide access to evolutionary information contained in the original enzymatic architectures that have been masked through millions of years of evolution. Therefore, ancestral proteins can play a critical role in understanding the emergence and evolution of enzymatic catalysis during the origin of life.

# CONCLUSIONS

The different experimental results obtained from the studies presented in this thesis lead to the following relevant conclusions:

1. An ultra-low throughput screening of variants predicted by the protein engineering software FuncLib and validated with empirical valence bond calculations can optimize the minimalistic active site for the Kemp elimination reaction designed in the ancestral β-lactamase GNCA4.
2. The optimization of the *de novo* active site in the GNCA4 scaffold leads to the efficient variant GNCA4-12, which exhibits improved transition state stabilization and increased enzymatic activity due to better preorganization of the active site.
3. The results obtained after the evolution of GNCA4 confirm the role of transition state stabilization as an immediate evolutionary mechanism in the optimization of *de novo* enzymatic catalysis.
4. The combination of FuncLib predictions and empirical valence bond calculations is a promising computational pipeline for the easy engineering and evolution of *de novo* artificial active sites.
5. The engineering of an extra segment fragment into the GNCA4-12 variant sequence and directed evolution of a targeted sequence within lead to the improved variant GNCA12-V4, which displays outstanding catalytic parameters for the Kemp elimination reaction.
6. The results obtained after the evolution of the GNCA4-12 variant highlight the role of evolutionary mechanisms in the generation of new interactions in the active site for the evolution and optimization of *de novo* artificial active sites.
7. The engineering of extra fragments in protein sequences stands as an underexplored but promising strategy for the rapid and easy engineering and evolution of active sites.
8. Implementation of two different optimization strategies led to an efficient Kemp eliminase artificial enzyme, based on a resurrected ancestral β-lactamase, with outstanding catalytic parameters. This enzyme compares favorably with the best Kemp eliminase described in the literature, which was designed on a much more complex engineering process performed in a modern protein scaffold.
9. The use of a resurrected ancestral β-lactamase as a scaffold for the generation and further evolution of a novel enzymatic functionality demonstrates the superiority of ancestral proteins as more efficient scaffolds for the generation of *de novo* artificial active sites, in comparison with their modern counterparts.
10. Resurrection and characterization of N72, an ancestral β-glycosidase from the glycoside hydrolase family 1, led to the generation of an ancestral scaffold with an excellent combination of biochemical and biophysical properties.
11. Ancestral protein N72 displayed a catalytically active TIM-barrel fold that combined high thermostability provided by a rigid core and high conformational flexibility in a specific region of the protein structure.

12. Unexpectedly, our ancestral TIM-barrel protein N72 displayed the ability to bind heme tightly and stoichiometrically in a well-defined buried site located in the flexible region of the protein structure.
13. Upon heme binding, the TIM-barrel is rigidified through structural rearrangements that lead to an allosteric modulation of its natural enzymatic activity.
14. The heme-bound ancestral TIM-barrel displays low but substantial levels of peroxidase activity provided by the catalytic power of the heme cofactor. This minimal activity opens the door for further engineering studies aimed to generate novel enzymatic activities based on the redox power of heme.
15. The combination enhanced thermostability, increased conformational flexibility, and the ability to bind heme in our ancestral TIM-barrel leads to a protein scaffold with a perfect molecular context for the generation of *de novo* active sites.
16. The results obtained after characterization of our ancestral TIM-barrel N72 highlights the potential of ancestral sequence reconstruction as an outstanding approach to access unexplored protein sequence space and generate new protein variants with unusual combinations of properties.
17. Heme binding to the ancestral TIM-barrel promotes the protection of the heme cofactor from catalytic inactivation driven by aggregation and degradation. This protection leads to an improved catalytic activity in terms of a higher number of turnovers that is not promoted by any catalytic machinery in the protein.
18. The ancestral TIM-barrel N72 serves as a model scaffold to understand the origins of cofactor-dependent enzymatic catalysis, reflecting the role of cofactor protection as a driving force in the evolutionary association between peptides and cofactors in the early stages of protein evolution.
19. Overall, the results obtained with our ancestral β-lactamase and resurrected ancestral TIM-barrel highlight the role of ancestral proteins as model scaffolds to understand the emergence and evolution of enzymatic catalysis during the origin of life.

# Conclusiones

Los diferentes resultados experimentales obtenidos de los estudios presentados en esta tesis dan lugar a las siguientes conclusiones relevantes:

1. Un cribado de bajo número de variantes predichas por el software de ingeniería de proteínas FuncLib y validadas mediante cálculos de valencia empírica de enlace pueden optimizar el sitio activo minimalista para la reacción de eliminación de Kemp diseñado en la β-lactamasa ancestral GNCA4.
2. La optimización del sitio activo *de novo* en la proteína GNCA4 da lugar a una variante eficiente denominada GNCA4-12, que muestra una estabilización del estado de transición mejorada y una mayor actividad enzimática gracias a la mejor preorganización del sitio activo.
3. Los resultados obtenidos después de la evolución de GNCA4 confirman el papel de la estabilización del estado de transición como un mecanismo evolutivo inmediato en la optimización de catálisis enzimática *de novo*.
4. La combinación de predicciones procedentes de FuncLib y cálculos de valencia empírica de enlace aparece como una vía computacional prometedora para la ingeniería y evolución fáciles de sitios activos artificiales *de novo*.
5. La introducción de un fragmento de segmento extra en la secuencia de la variante GNCA4-12 y la evolución dirigida de parte de este segmento da lugar a la variante mejorada GNCA12-V4, que presenta parámetros catalíticos extraordinarios para la reacción de eliminación de Kemp.
6. Los resultados obtenidos tras la evolución de la variante GNCA4-12 destacan el papel de mecanismos evolutivos en la generación de nuevas interacciones en el sitio activo para la evolución y optimización de sitios activos artificiales *de novo*.
7. La introducción de fragmentos extra en la secuencia de proteínas aparece como una estrategia prometedora e inexplorada para la rápida y fácil evolución y optimización de sitios activos.
8. La implementación de dos estrategias de optimización diferentes ha dado lugar a la generación de una enzima artificial que funciona como una eliminasa de Kemp eficiente, basada en una β-lactamasa ancestral resucitada, y que presenta parámetros catalíticos extraordinarios. Esta enzima se compara favorablemente con la mejor Kemp eliminasa artificial descrita en la literatura, la cual fue diseñada mediante un proceso mucho más complejo realizado en una proteína moderna.
9. El uso de una β-lactamasa ancestral resucitada como andamiaje para la generación y posterior evolución de una nueva funcionalidad enzimática demuestra la superioridad de las proteínas ancestrales como puntos de inicio superiores para la generación de sitios activos artificiales *de novo*, en comparación con las proteínas modernas.
10. La resurrección y caracterización de la proteína N72, una β-glicosidasa ancestral procedente de la familia 1 de las glicosil hidrolasas, dio lugar a la generación de

un andamiaje ancestral con una excelente combinación de propiedades bioquímicas y biofísicas.

11. La proteína ancestral N72 muestra un plegamiento en forma de barril TIM catalíticamente activo que combina una alta termo estabilidad gracias al núcleo rígido y estable del barril, y una alta flexibilidad conformacional en una región específica de la estructura.
12. De forma inesperada, nuestra proteína ancestral N72 mostró la capacidad de unir hemo fuertemente y de forma estequiométrica en un sitio de unión enterrado y bien definido localizado en la región estructural flexible de la estructura.
13. Mediante la unión de hemo, el barril TIM aumenta su rigidez mediante reajustes estructurales que dan lugar a una modulación alostérica de su actividad enzimática natural.
14. El barril TIM ancestral con hemo unido muestra niveles bajos pero sustanciales de actividad enzimática peroxidasa gracias al poder catalítico e intrínseco redox del cofactor hemo. Esta actividad enzimática nueva mínima abre la puerta a futuros experimentos centrados en generar nuevas actividades enzimáticas basadas en el poder redox del hemo.
15. La combinación alta termo estabilidad, elevada flexibilidad conformacional y la capacidad de unir hemo en nuestro barril TIM ancestral da lugar a un andamiaje proteico con un contexto molecular perfecto para la generación de sitios activos *de novo*.
16. Los resultados obtenidos tras la caracterización de nuestro barril TIM ancestral N72 destacan el potencial de la reconstrucción ancestral de secuencias como una excelente metodología para acceder al espacio de secuencia sin explorar y generar nuevas variantes de proteínas con combinaciones inusuales de propiedades.
17. La unión de hemo al barril TIM ancestral promueve la protección del hemo frente a la inactivación catalítica provocada por la agregación y degradación del cofactor. Esta protección da lugar a una actividad catalítica mejorada en términos de un mayor número de reacciones catalizadas por cofactor que no está promovido por la participación de una maquinaria catalítica.
18. El barril TIM ancestral N72 sirve como un andamiaje proteico modelo para entender el origen de la catálisis dependiente de cofactor, reflejando el papel de la protección de los cofactores como una fuerza conductora en la asociación evolutiva entre cofactores y polipéptidos en las etapas más tempranas de la evolución de proteínas.
19. En general, los resultados obtenidos a través de nuestra β-lactamasa ancestral resucitada y nuestro barril TIM ancestral demuestran el papel de las proteínas ancestrales como andamiajes modelo para entender la emergencia y evolución de la catálisis enzimática durante el origen de la vida.

# Bibliography

1. Lupas, A. N., Ponting, C. P. & Russell, R. B. On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *J. Struct. Biol.* **134**, 191–203 (2001).
2. Eck, R. V. & Dayhoff, M. O. Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences. *Science* **152**, 363–366 (1966).
3. Friedberg, I. & Godzik, A. Connecting the Protein Structure Universe by Using Sparse Recurring Fragments. *Structure* **13**, 1213–1224 (2005).
4. Alva, V., Remmert, M., Biegert, A., Lupas, A. N. & Söding, J. A galaxy of folds. *Protein Sci.* **19**, 124-120 (2009).
5. Gonceàrenco, A. & Berezovsky, I. N. Computational reconstruction of primordial prototypes of elementary functional loops in modern proteins. *Bioinformatics* **27**, 2368–2375 (2011).
6. Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. Global view of the protein universe. *Proc. Natl. Acad. Sci.* **111**, 11691–11696 (2014).
7. Alva, V., Söding, J. & Lupas, A. N. A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **4**, e09410 (2015).
8. Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proc. Natl. Acad. Sci.* **114**, 11703–11708 (2017).
9. Kolodny, R., Nepomnyachiy, S., Tawfik, D. S. & Ben-Tal, N. Bridging Themes: Short Protein Segments Found in Different Architectures. *Mol. Biol. Evol.* **38**, 2191–2208 (2021).
10. Qiu, K., Ben-Tal, N. & Kolodny, R. Similar protein segments shared between domains of different evolutionary lineages. *Protein Sci.* **31**, (2022).
11. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
12. Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).

13. Romero Romero, M. L., Rabin, A. & Tawfik, D. S. Functional Proteins from Short Peptides: Dayhoff's Hypothesis Turns 50. *Angew. Chem. Int. Ed.* **55**, 15966–15971 (2016).

14. Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V. & Martin, W. F. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLOS Genet.* **14**, e1007518 (2018).

15. de Vries, H. Species and varieties: Their origin by mutation. Lectures delivered at the University of California. Second Edition, Corrected and Revised. By H. De Vries; edited by D. T. MacDougal. Chicago: Open Court Publishing Co.; London: Kegan Paul and Co., Ltd. (1906).

16. Goncearenco, A. & Berezovsky, I. N. Protein function from its emergence to diversity in contemporary proteins. *Phys. Biol.* **12**, 045002 (2015).

17. Söding, J. & Lupas, A. N. More than the sum of their parts: On the evolution of proteins from peptides: Review articles. *BioEssays* **25**, 837–846 (2003).

18. Pandya, C., Farelli, J. D., Dunaway-Mariano, D. & Allen, K. N. Enzyme Promiscuity: Engine of Evolutionary Innovation. *J. Biol. Chem.* **289**, 30229–30236 (2014).

19. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nat. Rev. Genet.* **11**, 572–582 (2010).

20. Copley, S. D. An evolutionary biochemist's perspective on promiscuity. *Trends Biochem. Sci.* **40**, 72–78 (2015).

21. Khersonsky, O. & Tawfik, D. S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).

22. Furnham, N., Dawson, N. L., Rahman, S. A., Thornton, J. M. & Orengo, C. A. Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies. *J. Mol. Biol.* **428**, 253–267 (2016).

23. Ouzounis, C. A., Kunin, V., Darzentas, N. & Goldovsky, L. A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. *Res. Microbiol.* **157**, 57–68 (2006).

24. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).

25. Wolfenden, R. & Snider, M. J. The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Acc. Chem. Res.* **34**, 938–945 (2001).

26. Feynman, R. *Richard Feynman's blackboard at time of his death*. California Institute of Technology, Image Archive ID: ct1:483 (1988).

27. Pauling, L. Molecular Architecture and Biological Reactions. *Chem. Eng. News* **24**, 1375–1377 (1946).

28. Warshel, A. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *J. Biol. Chem.* **273**, 27035–27038 (1998).

29. Warshel, A. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci.* **75**, 5250–5254 (1978).

30. Zhang, X. & Houk, K. N. Why Enzymes Are Proficient Catalysts: Beyond the Pauling Paradigm. *Acc. Chem. Res.* **38**, 379–385 (2005).

31. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **52**, 5700–5725 (2013).

32. Kries, H., Blomberg, R. & Hilvert, D. De novo enzymes by computational design. *Curr. Opin. Chem. Biol.* **17**, 221–228 (2013).

33. Tantillo, D. J., Jiangang, C. & Houk, K. N. Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **2**, 743–750 (1998).

34. Zanghellini, A. *et al.* New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).

35. Malisi, C., Kohlbacher, O. & Höcker, B. Automated scaffold selection for enzyme design. *Proteins Struct. Funct. Bioinforma.* **77**, 74–83 (2009).

36. Dahiyat, B. I. & Mayo, S. L. Protein design automation. *Protein Sci.* **5**, 895–903 (1996).

37. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci.* **103**, 16710–16715 (2006).

38. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).

39. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109**, 3790–3795 (2012).

40. Jiang, L. *et al.* De Novo Computational Design of Retro-Aldol Enzymes. *Science* **319**, 1387–1391 (2008).

41. Richter, F. *et al.* Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).

42. Siegel, J. B. *et al.* Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).

43. Bjelic, S. *et al.* Computational Design of Enone-Binding Proteins with Catalytic Activity for the Morita–Baylis–Hillman Reaction. *ACS Chem. Biol.* **8**, 749–757 (2013).

44. Bar-Even, A. *et al.* The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50**, 4402–4410 (2011).

45. Marshall, L. R., Zozulia, O., Lengyel-Zhand, Z. & Korendovych, I. V. Minimalist de Novo Design of Protein Catalysts. *ACS Catal.* **9**, 9265–9275 (2019).

46. Makhlynets, O. V. & Korendovych, I. V. Minimalist Design of Allosterically Regulated Protein Catalysts. in *Methods in Enzymology* vol. 580 191–202 (Elsevier, 2016).

47. Korendovych, I. V. *et al.* Design of a switchable eliminase. *Proc. Natl. Acad. Sci.* **108**, 6823–6827 (2011).

48. Raymond, E. A. *et al.* Design of an allosterically regulated retroaldolase: Design of an Allosterically Regulated Retroaldolase. *Protein Sci.* **24**, 561–570 (2015).

49. Moroz, Y. S. *et al.* New Tricks for Old Proteins: Single Mutations in a Nonenzymatic Protein Give Rise to Various Enzymatic Activities. *J. Am. Chem. Soc.* **137**, 14905–14911 (2015).

50. Merski, M. & Shoichet, B. K. Engineering a model protein cavity to catalyze the Kemp elimination. *Proc. Natl. Acad. Sci.* **109**, 16179–16183 (2012).

51. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci.* **98**, 14274–14279 (2001).

52. Harris, T. K. & Turner, G. J. Structural Basis of Perturbed pKa Values of Catalytic Groups in Enzyme Active Sites. *IUBMB Life Int. Union Biochem. Mol. Biol. Life* **53**, 85–98 (2002).

53. Warshel, A., Sharma, P. K., Kato, M. & Parson, W. W. Modeling electrostatic effects in proteins. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1764**, 1647–1676 (2006).

54. Garcia-Seisdedos, H., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. How many ionizable groups can sit on a protein hydrophobic core? *Proteins Struct Funct Bioinforma.* **80**, 1–7 (2012).

55. Holm, R. H., Kennepohl, P. & Solomon, E. I. Structural and Functional Aspects of Metal Sites in Biology. *Chem. Rev.* **96**, 2239–2314 (1996).

56. Punekar, N. S. *Enzymes: Catalysis, Kinetics and Mechanisms*. (Springer Singapore, 2018).

57. Lu, Y., Yeung, N., Sieracki, N. & Marshall, N. M. Design of functional metalloproteins. *Nature* **460**, 855–862 (2009).

58. Benson, D. E., Wisz, M. S. & Hellinga, H. W. Rational design of nascent metalloenzymes. *Proc. Natl. Acad. Sci.* **97**, 6292–6297 (2000).

59. Bhagi-Damodaran, A. *et al.* Why copper is preferred over iron for oxygen activation and reduction in haem-copper oxidases. *Nat. Chem.* **9**, 257–263 (2017).

60. Yeung, N. *et al.* Rational design of a structural and functional nitric oxide reductase. *Nature* **462**, 1079–1082 (2009).

61. Mirts, E. N., Petrik, I. D., Hosseinzadeh, P., Nilges, M. J. & Lu, Y. A designed heme-[4Fe-4S] metalloenzyme catalyzes sulfite reduction like the native enzyme. *Science* **361**, 1098–1101 (2018).

62. Jae Jeong, W. & Ju Song, W. Design and directed evolution of noncanonical β-stereoselective metalloglycosidases. *Nat. Commun.* **13**, 6844 (2022).

63. Faiella, M. *et al.* An artificial di-iron oxo-protein with phenol oxidase activity. *Nat. Chem. Biol.* **5**, 882–884 (2009).

64. Zastrow, M. L., Peacock, A. F. A., Stuckey, J. A. & Pecoraro, V. L. Hydrolytic catalysis and structural stabilization in a designed metalloprotein. *Nat. Chem.* **4**, 118–123 (2012).

65. Reig, A. J. *et al.* Alteration of the oxygen-dependent reactivity of de novo Due Ferri proteins. *Nat. Chem.* **4**, 900–906 (2012).

66. Cangelosi, D. V. M., Deb, D. A., Penner-Hahn, J. E. & Pecoraro, V. L. A de novo designed metalloenzyme for the hydration of CO2. (2015).

67. Mathieu, E. *et al.* Rational De Novo Design of a Cu Metalloenzyme for Superoxide Dismutation. *Chem. Eur. J.* **26**, 249–258 (2020).

68. Zastrow, M. L. & Pecoraro, V. L. Influence of Active Site Location on Catalytic Activity in *de Novo* -Designed Zinc Metalloenzymes. *J. Am. Chem. Soc.* **135**, 5895–5903 (2013).

69. Fischer, J. D., Holliday, G. L., Rahman, S. A. & Thornton, J. M. The Structures and Physicochemical Properties of Organic Cofactors in Biocatalysis. *J. Mol. Biol.* **403**, 803–824 (2010).

70. Lewis, J. C. & Ellis-Guardiola, K. Preparation of Artificial Metalloenzymes. in *Artificial Metalloenzymes and MetalloDNAzymes in Catalysis* 1–40 (Wiley-VCH Verlag GmbH & Co. KGaA, 2018).

71. Schwizer, F. *et al.* Artificial Metalloenzymes: Reaction Scope and Optimization Strategies. *Chem. Rev.* **118**, 142–231 (2018).

72. Petrik, I. D., Liu, J. & Lu, Y. Metalloenzyme design and engineering through strategic modifications of native protein scaffolds. *Curr. Opin. Chem. Biol.* **19**, 67–75 (2014).

73. Sharp, R. E., Moser, C. C., Rabanal, F. & Dutton, P. L. Design, synthesis, and characterization of a photoactivatable flavocytochrome molecular maquette. *Proc. Natl. Acad. Sci.* **95**, 10465–10470 (1998).

74. Nanda, V. *et al.* De Novo Design of a Redox-Active Minimal Rubredoxin Mimic. *J. Am. Chem. Soc.* **127**, 5804–5805 (2005).

75. Eggink, L. L. & Hoober, J. K. Chlorophyll Binding to Peptide Maquettes Containing a Retention Motif. *J. Biol. Chem.* **275**, 9087–9090 (2000).

76. Butterfield, S. M. & Waters, M. L. A Designed β-Hairpin Peptide for Molecular Recognition of ATP in Water. *J. Am. Chem. Soc.* **125**, 9580–9581 (2003).

77. Razeghifard, M. R. & Wydrzynski, T. Binding of Zn-Chlorin to a Synthetic Four-Helix Bundle Peptide through Histidine Ligation. *Biochemistry* **42**, 1024–1030 (2003).

78. Cochran, F. V. *et al.* Computational De Novo Design and Characterization of a Four-Helix Bundle Protein that Selectively Binds a Nonbiological Cofactor. *J. Am. Chem. Soc.* **127**, 1346–1347 (2005).

79. Watkins, D. W. *et al.* Construction and in vivo assembly of a catalytically proficient and hyperthermostable de novo enzyme. *Nat. Commun.* **8**, 358 (2017).

80. Grayson, K. J. & Anderson, J. R. The ascent of man(made oxidoreductases). *Curr. Opin. Struct. Biol.* **51**, 149–155 (2018).

81. Stenner, R., Steventon, J. W., Seddon, A. & Anderson, J. L. R. A de novo peroxidase is also a promiscuous yet stereoselective carbene transferase. *Proc. Natl. Acad. Sci.* **117**, 1419–1428 (2020).

82. Stenner, R. & Anderson, J. L. R. Chemoselective N−H insertion catalyzed by a de novo carbene transferase. *Biotechnol. Appl. Biochem.* **67**, 527–535 (2020).

83. Jenkins, J. M. X., Noble, C. E. M., Grayson, K. J., Mulholland, A. J. & Anderson, J. L. R. Substrate promiscuity of a de novo designed peroxidase. *J. Inorg. Biochem.* **217**, 111370 (2021).

84. Farid, T. A. *et al.* Elementary tetrahelical protein design for diverse oxidoreductase functions. *Nat. Chem. Biol.* **9**, 826–833 (2013).

85. Khersonsky, O. *et al.* Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci.* **109**, 10358–10363 (2012).

86. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).

87. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).

88. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).

89. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes: Robust Design and Optimization of Retroaldol Enzymes. *Protein Sci.* **21**, 717–726 (2012).

90. Preiswerk, N. *et al.* Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc. Natl. Acad. Sci.* **111**, 8013–8018 (2014).

91. Crawshaw, R. *et al.* Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**, 313–320 (2022).

92. Basler, S. *et al.* Efficient Lewis acid catalysis of an abiological reaction in a de novo protein scaffold. *Nat. Chem.* **13**, 231–235 (2021).

93. Studer, S. *et al.* Evolution of a highly active and enantiospecific metalloenzyme from short peptides. *Science* **362**, 1285–1288 (2018).

94. Chen, K. & Arnold, F. H. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci.* **90**, 5618–5622 (1993).

95. Stemmer, W. P. C. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389–391 (1994).

96. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).

97. Lutz, S. Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* **21**, 734–743 (2010).

98. Becker, S. Ultra-high-throughput screening based on cell-surface display and fluorescence-activated cell sorting for the identification of novel biocatalysts. *Curr. Opin. Biotechnol.* **15**, 323–329 (2004).

99. Agresti, J. J. *et al.* Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc. Natl. Acad. Sci.* **107**, 4004–4009 (2010).

100. Debon, A. *et al.* Ultrahigh-throughput screening enables efficient single-round oxidase remodelling. *Nat. Catal.* **2**, 740–747 (2019).

101. Molina, R. S. *et al.* In vivo hypermutation and continuous evolution. *Nat. Rev. Methods Primer* **2**, 36 (2022).

102. Nevin Gerek, Z., Kumar, S. & Banu Ozkan, S. Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol. Appl.* **6**, 423–433 (2013).

103. Campitelli, P., Modi, T., Kumar, S. & Ozkan, S. B. The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. *Annu. Rev. Biophys.* **49**, 267–288 (2020).

104. Romero-Rivera, A., Garcia-Borràs, M. & Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **7**, 8524–8532 (2017).

105. Wang, J. *et al.* Mapping allosteric communications within individual proteins. *Nat. Commun.* **11**, 3862 (2020).

106. Osuna, S., Jiménez-Osés, G., Noey, E. L. & Houk, K. N. Molecular Dynamics Explorations of Active Site Structure in Designed and Evolved Enzymes. *Acc. Chem. Res.* **48**, 1080–1089 (2015).

107. Childers, M. C. & Daggett, V. Insights from molecular dynamics simulations for computational protein design. *Mol. Syst. Des. Eng.* **2**, 9–33 (2017).

108. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).

109. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci.* **116**, 8852–8858 (2019).

110. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **10**, 1210–1223 (2020).

111. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **69**, 11–18 (2021).

112. Qiu, Y., Hu, J. & Wei, G.-W. Cluster learning-assisted directed evolution. *Nat. Comput. Sci.* **1**, 809–818 (2021).

113. Zeymer, C. & Hilvert, D. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem.* **87**, 131–157 (2018).

114. Quin, M. B. & Schmidt-Dannert, C. Engineering of Biocatalysts: from Evolution to Creation. *ACS Catal.* **1**, 1017–1021 (2011).

115. Dalby, P. A. Strategy and success for the directed evolution of enzymes. *Curr. Opin. Struct. Biol.* **21**, 473–480 (2011).

116. Bloom, J. D. & Arnold, F. H. In the light of directed evolution: Pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci.* **106**, 9995–10000 (2009).

117. Kipnis, Y. & Baker, D. Comparison of designed and randomly generated catalysts for simple chemical reactions: Designed vs. Randomly Generated Protein Catalysts. *Protein Sci.* **21**, 1388–1395 (2012).

118. Fasan, R., Meharenna, Y. T., Snow, C. D., Poulos, T. L. & Arnold, F. H. Evolutionary History of a Specialized P450 Propane Monooxygenase. *J. Mol. Biol.* **383**, 1069–1080 (2008).

119. Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci.* **102**, 606–611 (2005).

120. Besenmatter, W., Kast, P. & Hilvert, D. Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. *Proteins Struct. Funct. Bioinforma.* **66**, 500–506 (2006).

121. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci.* **103**, 5869–5874 (2006).

122. Hilvert, D. Design of Protein Catalysts. *Annu. Rev. Biochem.* **82**, 447–470 (2013).

123. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).

124. Bunzel, H. A. *et al.* Evolution of dynamical networks enhances catalysis in a designer enzyme. *Nat. Chem.* **13**, 1017–1022 (2021).

125. Otten, R. *et al.* How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science* **370**, 1442–1446 (2020).

126. Hong, N.-S. *et al.* The evolution of multiple active site configurations in a designed enzyme. *Nat. Commun.* **9**, 3900 (2018).

127. Warinner, C., Korzow Richter, K. & Collins, M. J. Paleoproteomics. *Chem. Rev.* **122**, 13401–13446 (2022).

128. Pauling, L. & Zuckerkandl, E. Chemical paleogenetics - Molecular restoration studies of extinct forms of life. *Acta Chem. Scand.* **17**, 9–16 (1963).

129. Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P. & Benner, S. A. The ribonuclease from an extinct bovid ruminant. *FEBS Lett.* **262**, 104–106 (1990).

130. Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. **345**, (1990).

131. Liberles, D. (2007). Ancestral Sequence Reconstruction (New York, NY:
Oxford University Press).

132. Atkinson, Q. D. The descent of words. *Proc. Natl. Acad. Sci.* **110**, 4159–4160 (2013).

133. Gumulya, Y. & Gillam, E. M. J. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. *Biochem. J.* **474**, 1–19 (2017).

134. Carletti, M. S. *et al.* Revenant: a database of resurrected proteins. *Database* **2020**, baaa031 (2020).

135. Hochberg, G. K. A. & Thornton, J. W. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).

136. Risso, V. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Biotechnological and protein-engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol.* **51**, 106–115 (2018).

137. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141 (2021).

138. Trudeau, D. L., Kaltenbach, M. & Tawfik, D. S. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol. Biol. Evol.* **33**, 2633–2641 (2016).

139. Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T. & Poon, A. F. Y. Ancestral Reconstruction. *PLOS Comput. Biol.* **12**, e1004763 (2016).

140. Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).

141. Randall, R. N., Radford, C. E., Roof, K. A., Natarajan, D. K. & Gaucher, E. A. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.* **7**, 12847 (2016).

142. Gerlt, J. A. & Babbitt, P. C. Enzyme (re)design: lessons from natural evolution and computation. *Curr. Opin. Chem. Biol.* **13**, 10–18 (2009).

143. Risso, V. A. *et al.* Mutational Studies on Resurrected Ancestral Proteins Reveal Conservation of Site-Specific Amino Acid Preferences throughout Evolutionary History. *Mol. Biol. Evol.* **32**, 440–455 (2015).

144. Ingles-Prieto, A. *et al.* Conservation of Protein Structure over Four Billion Years. *Structure* **21**, 1690–1697 (2013).

145. Romero-Romero, M. L. *et al.* Selection for Protein Kinetic Stability Connects Denaturation Temperatures to Organismal Temperatures and Provides Clues to Archaean Life. *PLoS One* **11**, e0156657 (2016).

146. Gaucher, E. A., Govindarajan, S. & Ganesh, O. K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704–707 (2008).

147. Perez-Jimenez, R. *et al.* Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* **18**, 592–596 (2011).

148. Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β-Lactamases. *J. Am. Chem. Soc.* **135**, 2899−2902 (2013).

149. Akanuma, S. *et al.* Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci.* **110**, 11067–11072 (2013).

150. Nguyen, V. *et al.* Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* **355**, 289–294 (2017).

151. Modi, T., Huihui, J., Ghosh, K. & Ozkan, S. B. Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble. Phil. Trans. R. Soc. **373**: 20170184 (2018).

152. Kim, H. *et al.* A Hinge Migration Mechanism Unlocks the Evolution of Green-to-Red Photoconversion in GFP-like Proteins. *Structure* **23**, 34–43 (2015).

153. Modi, T. *et al.* Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity. *Nat. Commun.* **12**, 1852 (2021).

154. Zou, T., Risso, V. A., Gavira, J. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Mol. Biol. Evol.* **32**, 132–143 (2015).

155. Gamiz-Arco, G. *et al.* Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding. *Biochem. J.* **476**, 3631–3647 (2019).

156. Gamiz-Arco, G. *et al.* Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase. *Nat. Commun.* **12**, 380 (2021).

157. Risso, V. A., Gavira, J. A., Gaucher, E. A. & Sanchez-Ruiz, J. M. Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins: Consensus vs. Ancestral Proteins. *Proteins Struct. Funct. Bioinforma.* **82**, 887–896 (2014).

158. Rauwerdink, A. *et al.* Evolution of a Catalytic Mechanism. *Mol. Biol. Evol.* **33**, 971–979 (2016).

159. Castro-Fernandez, V. *et al.* Reconstructed ancestral enzymes reveal that negative selection drove the evolution of substrate specificity in ADP-dependent kinases. *J. Biol. Chem.* **292**, 15598–15610 (2017).

160. Clifton, B. E. *et al.* Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nat. Chem. Biol.* **14**, 542–547 (2018).

161. Kaltenbach, M. *et al.* Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat. Chem. Biol.* **14**, 548–555 (2018).

162. Kratzer, J. T. *et al.* Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc. Natl. Acad. Sci.* **111**, 3763–3768 (2014).

163. Carrigan, M. A. *et al.* Hominids adapted to metabolize ethanol long before human-directed fermentation. *Proc. Natl. Acad. Sci.* **112**, 458–463 (2015).

164. Barruetabeña, N. *et al.* Resurrection of efficient Precambrian endoglucanases for lignocellulosic biomass hydrolysis. *Commun. Chem.* **2**, 76 (2019).

165. Barandiaran, L. *et al.* Enzymatic upgrading of nanochitin using an ancient lytic polysaccharide monooxygenase. *Commun. Mater.* **3**, 55 (2022).

166. Gomez-Fernandez, B. J., Risso, V. A., Rueda, A., Sanchez-Ruiz, J. M. & Alcalde, M. Ancestral Resurrection and Directed Evolution of Fungal Mesozoic Laccases. *Appl. Environ. Microbiol.* **86**, e00778-20 (2020).

167. Gumulya, Y. *et al.* Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nat. Catal.* **1**, 878–888 (2018).

168. Hendrikse, N. M., Charpentier, G., Nordling, E. & Syrén, P. Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *FEBS J.* **285**, 4660–4673 (2018).

169. Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R. & Damborsky, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem* **18**, 1448–1456 (2017).

170. Joho, Y. *et al.* Ancestral Sequence Reconstruction Identifies Structural Changes Underlying the Evolution of *Ideonella sakaiensis* PETase and Variants with Improved Stability and Activity. *Biochemistry* **62**, 437–450 (2023).

171. Nakano, S. *et al.* Ancestral L-amino acid oxidases for deracemization and stereoinversion of amino acids. *Commun. Chem.* **3**, 181 (2020).

172. Nakano, S., Minamino, Y., Hasebe, F. & Ito, S. Deracemization and Stereoinversion to Aromatic D-Amino Acid Derivatives with Ancestral L-Amino Acid Oxidase. *ACS Catal.* **9**, 10152–10158 (2019).

173. Ishida, C. *et al.* Reconstruction of Hyper-Thermostable Ancestral L-Amino Acid Oxidase to Perform Deracemization to D-Amino Acids. *ChemCatChem* **13**, 5228–5235 (2021).

174. Wilding, M. *et al.* Reverse engineering: transaminase biocatalyst development using ancestral sequence reconstruction. *Green Chem.* **19**, 5375–5380 (2017).

175. Ma, D. *et al.* Ancestral sequence reconstruction and spatial structure analysis guided alteration of longer-chain substrate catalysis for Thermomicrobium roseum lipase. *Enzyme Microb. Technol.* **156**, 109989 (2022).

176. Livada, J., Vargas, A. M., Martinez, C. A. & Lewis, R. D. Ancestral Sequence Reconstruction Enhances Gene Mining Efforts for Industrial Ene Reductases by Expanding Enzyme Panels with Thermostable Catalysts. *ACS Catal.* **13**, 2576–2585 (2023).

177. Chen, X. *et al.* Directed reconstruction of a novel ancestral alcohol dehydrogenase featuring shifted pH-profile, enhanced thermostability and expanded substrate spectrum. *Bioresour. Technol.* **363**, 127886 (2022).

178. Kajimoto, S. *et al.* Enzymatic Conjugation of Modified RNA Fragments by Ancestral RNA Ligase AncT4_2. *Appl. Environ. Microbiol.* **88**, e01679-22 (2022).

179. Thomas, A., Cutlan, R., Finnigan, W., van der Giezen, M. & Harmer, N. Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun. Biol.* **2**, 429 (2019).

180. Delgado, A., Arco, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. Using Resurrected Ancestral Proviral Proteins to Engineer Virus Resistance. *Cell Rep.* **19**, 1247–1256 (2017).

181. Zakas, P. M. *et al.* Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat. Biotechnol.* **35**, 35–37 (2017).

182. Zakas, P. M. *et al.* Molecular coevolution of coagulation factor VIII and von Willebrand factor. *Blood Adv.* **5**, 812–822 (2021).

183. Hendrikse, N. M. *et al.* Ancestral lysosomal enzymes with increased activity harbor therapeutic potential for treatment of Hunter syndrome. *iScience* **24**, 102154 (2021).

184. Hendrikse, N. M. *et al.* Exploring the therapeutic potential of modern and ancestral phenylalanine/tyrosine ammonia-lyases as supplementary treatment of hereditary tyrosinemia. *Sci. Rep.* **10**, 1315 (2020).

185. Koblan, L. W. *et al.* Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).

186. Alonso-Lerma, B. *et al.* Evolution of CRISPR-associated endonucleases as inferred from resurrected proteins. *Nat. Microbiol.* **8**, 77–90 (2023).

187. Manteca, A. *et al.* Mechanochemical evolution of the giant muscle protein titin as inferred from resurrected proteins. *Nat. Struct. Mol. Biol.* **24**, 652–657 (2017).

188. Gonzalez, D. *et al.* Ancestral mutations as a tool for solubilizing proteins: The case of a hydrophobic phosphate-binding protein. *FEBS Open Bio* **4**, 121–127 (2014).

189. Whitfield, J. H. *et al.* Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction: cpFLIPR: Improved Biosensor for L - Arginine. *Protein Sci.* **24**, 1412–1422 (2015).

190. Gomez-Fernandez, B. J. *et al.* Directed -in vitro- evolution of Precambrian and extant Rubiscos. *Sci. Rep.* **8**, 5532 (2018).

191. Li, D. *et al.* Consensus Mutagenesis and Ancestral Reconstruction Provide Insight into the Substrate Specificity and Evolution of the Front-End Δ6-Desaturase Family. *Biochemistry* **59**, 1398−1409 (2020).

192. Sun, Y., Calderini, E. & Kourist, R. A Reconstructed Common Ancestor of the Fatty Acid Photo-decarboxylase Clade Shows Photo-decarboxylation Activity and Increased Thermostability. *ChemBioChem* **22**, 1833–1840 (2021).

193. Gupta, R. D. Recent advances in enzyme promiscuity. *Sustain. Chem. Process.* **4**, 2 (2016).

194. Copley, S. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* **7**, 265–272 (2003).

195. Khersonsky, O., Roodveldt, C. & Tawfik, D. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **10**, 498–508 (2006).

196. O'Brien, P. J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, R91–R105 (1999).

197. Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167 (2009).

198. Bornscheuer, U. T. & Kazlauskas, R. J. Catalytic Promiscuity in Biocatalysis: Using Old Enzymes to Form New Bonds and Follow New Pathways. *Angew. Chem. Int. Ed.* **43**, 6032–6040 (2004).

199. James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).

200. Pabis, A., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. Cooperativity and flexibility in enzyme evolution. *Curr. Opin. Struct. Biol.* **48**, 83–92 (2018).

201. Haliloglu, T. & Bahar, I. Adaptability of protein structures to enable functional interactions and evolutionary implications. *Curr. Opin. Struct. Biol.* **35**, 17–23 (2015).

202. Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944–950 (2016).

203. Bhabha, G. *et al.* Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nat. Struct. Mol. Biol.* **20**, 1243–1249 (2013).

204. Isaksen, G. V., Åqvist, J. & Brandsdal, B. O. Enzyme surface rigidity tunes the temperature dependence of catalytic rates. *Proc. Natl. Acad. Sci.* **113**, 7822–7827 (2016).

205. Pohorille, A., Wilson, M. A. & Shannon, G. Flexible Proteins at the Origin of Life. *Life* **7**, 23 (2017).

206. Vamvaca, K., Vögeli, B., Kast, P., Pervushin, K. & Hilvert, D. An enzymatic molten globule: Efficient coupling of folding and catalysis. *Proc. Natl. Acad. Sci.* **101**, 12860–12864 (2004).

207. Risso, V. A. *et al.* De novo active sites for resurrected Precambrian enzymes. *Nat. Commun.* **8**, 16113 (2017).

208. Gardner, J. M., Biler, M., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. L. Manipulating Conformational Dynamics To Repurpose Ancient Proteins for Modern Catalytic Functions. *ACS Catal.* **10**, 4863–4870 (2020).

209. Martin, W., Baross, J., Kelley, D. & Russell, M. J. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814 (2008).

210. Robert, F. & Chaussidon, M. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature* **443**, 969–972 (2006).

211. Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409**, 1083–1091 (2001).

212. Williams, P. D., Pollock, D. D., Blackburne, B. P. & Goldstein, R. A. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. *PLoS Comput. Biol.* **2**, e69 (2006).

213. Wijma, H. J., Floor, R. J. & Janssen, D. B. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.* **23**, 588–594 (2013).

214. Lin, G.-M., Warden-Rothman, R. & Voigt, C. A. Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.* **14**, 82–107 (2019).

215. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).

216. Chen, K. & Arnold, F. H. Engineering new catalytic activities in enzymes. *Nat. Catal.* **3**, 203–213 (2020).

217. Korendovych, I. V. & DeGrado, W. F. Catalytic efficiency of designed catalytic proteins. *Curr. Opin. Struct. Biol.* **27**, 113–121 (2014).

218. Lovelock, S. L. *et al.* The road to fully programmable protein catalysis. *Nature* **606**, 49–58 (2022).

219. Isom, D. G., Cannon, B. R., Castañeda, C. A., Robinson, A. & García-Moreno E., B. High tolerance for ionizable residues in the hydrophobic interior of proteins. *Proc. Natl. Acad. Sci.* **105**, 17784–17788 (2008).

220. Pey, A. L., Rodriguez-Larrea, D., Gavira, J. A., Garcia-Moreno, B. & Sanchez-Ruiz, J. M. Modulation of Buried Ionizable Groups in Proteins with Engineered Surface Charge. *J. Am. Chem. Soc.* **132**, 1218–1219 (2010).

221. Risso, V. A. *et al.* Enhancing a de novo enzyme activity by computationally-focused ultra-low-throughput screening. *Chem. Sci.* **11**, 6134–6148 (2020).

222. Gutierrez-Rus, L. I., Alcalde, M., Risso, V. A. & Sanchez-Ruiz, J. M. Efficient Base-Catalyzed Kemp Elimination in an Engineered Ancestral Enzyme. *Int. J. Mol. Sci.* **23**, 8934 (2022).

223. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* (2001).

224. Höcker, B., Jürgens, C., Wilmanns, M. & Sterner, R. Stability, catalytic versatility and evolution of the (βα)8-barrel fold. *Curr. Opin. Biotechnol.* **12**, 376–381 (2001).

225. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (βα)8-Barrel Enzyme Fold. *Chem. Rev.* **105**, 4038−4055 (2005).

226. Nagano, N., Orengo, C. A. & Thornton, J. M. One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions. *J. Mol. Biol.* **321**, 741–765 (2002).

227. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. **372**, (1994).

228. Trifonov, E. N. & Frenkel, Z. M. Evolution of protein modularity. *Curr. Opin. Struct. Biol.* **19**, 335–340 (2009).

229. Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **324**, 203–207 (2009).

230. Petrović, D., Risso, V. A., Kamerlin, S. C. L. & Sanchez-Ruiz, J. M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **15**, 20180330 (2018).

231. Wagner, A. Robustness, evolvability, and neutrality. *FEBS Lett.* **579**, 1772–1778 (2005).

232. Strauslt, D. & Gilbert, W. Genetic Engineering in the Precambrian: Structure of the Chicken Triosephosphate Isomerase Gene. *Mol. Cell. Biol.* **5**, 3497–3506 (1985).

233. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. Structural Evidence for Evolution of the β/α Barrel Scaffold by Gene Duplication and Fusion. *Science* **289**, 1546–1550 (2000).

234. Henn-Sax, M., Höcker, B., Wilmanns, M. & Sterner, R. Divergent Evolution of (βα)8-Barrel Enzymes. *Biol. Chem.* **382**, 1315–1320 (2001).

235. Richter, M. *et al.* Computational and Experimental Evidence for the Evolution of a (βα)8-Barrel Protein from an Ancestral Quarter-Barrel Stabilised by Disulfide Bonds. *J. Mol. Biol.* **398**, 763–773 (2010).

236. Höcker, B., Beismann-Driemeye, S., Hettwer, S., Lustig, A. & Sterner, R. Dissection of a (βα)8-barrel enzyme into two folded halves. *Nat. Struct. Mol. Biol.* **8**, 32–36 (2001).

237. Farías-Rico, J. A., Schmidt, S. & Höcker, B. Evolutionary relationship of two ancient protein superfolds. *Nat. Chem. Biol.* **10**, 710–715 (2014).

238. Höcker, B., Schmidt, S. & Sterner, R. A common evolutionary origin of two elementary enzyme folds. *FEBS Lett.* **510**, 133–135 (2002).

239. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J. Mol. Evol.* **82**, 17–26 (2016).

240. Gilbert, W. & Glynias, M. On the ancient nature of introns. *Gene* **135**, 137–144 (1993).

241. Wang, M., Yafremava, L. S., Caetano-Anollés, D., Mittenthal, J. E. & Caetano-Anollés, G. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.* **17**, 1572–1585 (2007).

242. Anantharaman, V., Aravind, L. & Koonin, E. V. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7**, 12–20 (2003).

243. Yamada, T. & Bork, P. Evolution of biomolecular networks — lessons from metabolic and protein interactions. *Nat. Rev. Mol. Cell Biol.* **10**, 791–803 (2009).

244. Gutierrez-Rus, L. I. *et al.* Protection of catalytic cofactors by polypeptides as a driver for the emergence of primordial enzymes. *BioRxiv* (2023) doi:10.1101/2023.03.14.532612.

245. Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **72**, 178-186.e5 (2018).

246. Miton, C. M. & Tokuriki, N. Insertions and Deletions (Indels): A Missing Piece of the Protein Engineering Jigsaw. *Biochemistry* **62**, 148–157 (2023).

247. Jayaraman, V., Toledo-Patiño, S., Noda-García, L. & Laurino, P. Mechanisms of protein evolution. *Protein Sci.* **31**, e4362 (2022).

248. Fried, S. D., Fujishima, K., Makarov, M., Cherepashuk, I. & Hlouchova, K. Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids. *J. R. Soc. Interface* **19**, 20210641 (2022).

249. Chu, X. & Zhang, H. Cofactors as Molecular Fossils To Trace the Origin and Evolution of Proteins. *ChemBioChem* **21**, 3161–3168 (2020).

250. Goldman, A. D. & Kacar, B. Cofactors are Remnants of Life's Origin and Early Evolution. *J. Mol. Evol.* **89**, 127–133 (2021).

251. Freire, M. Á. Short non-coded peptides interacting with cofactors facilitated the integration of early chemical networks. *Biosystems* **211**, 104547 (2022).

252. Brown, S. B., Dean, T. C., Jones, P. & Kremer, M. L. Catalytic activity of haemin. Influence of dissolution conditions on activity. *Trans. Faraday Soc.* **66**, 1485 (1970).

253. Brown, S. B., Dean, T. C. & Jones, P. Catalatic activity of iron(III)-centred catalysts. Role of dimerization in the catalytic action of ferrihaems. *Biochem. J.* **117**, 741–744 (1970).

254. Brown, S. B., Jones, P. & Suggett, A. Reactions between haemin and hydrogen peroxide. Part 1.—Ageing and non-destructive oxidation of haemin. *Trans Faraday Soc* **64**, 986–993 (1968).

255. Brown, S. B. & Jones, P. Reactions between haemin and hydrogen peroxide. Part 2.—Destructive oxidation of haemin. *Trans Faraday Soc* **64**, 994–998 (1968).

256. Valderrama, B., Ayala, M. & Vazquez-Duhalt, R. Suicide Inactivation of Peroxidases and the Challenge of Engineering More Robust Enzymes. *Chem. Biol.* **9**, 555–565 (2002).

257. Beinert, H. Iron-sulfur proteins: ancient structures, still full of surprises. *JBIC J. Biol. Inorg. Chem.* **5**, 2–15 (2000).

258. Imlay, J. A. Iron-sulphur clusters and the problem with oxygen. *Mol. Microbiol.* **59**, 1073–1082 (2006).

259. Bonfio, C. The curious case of peptide-coordinated iron–sulfur clusters: prebiotic and biomimetic insights. *Dalton Trans.* **50**, 801–807 (2021).

260. Kim, J. D. *et al.* Minimal Heterochiral de Novo Designed 4Fe–4S Binding Peptide Capable of Robust Electron Transfer. *J. Am. Chem. Soc.* **140**, 11210–11213 (2018).

261. Kirschning, A. The coenzyme/protein pair and the molecular evolution of life. *Nat. Prod. Rep.* **38**, 993–1010 (2021).

262. Kirschning, A. Coenzymes and Their Role in the Evolution of Life. *Angew. Chem. Int. Ed.* **60**, 6242–6269 (2021).

263. Xavier, J. C., Hordijk, W., Kauffman, S., Steel, M. & Martin, W. F. Autocatalytic chemical networks at the origin of metabolism. *Proc. R. Soc. B Biol. Sci.* **287**, 20192377 (2020).

264. Keller, M. A., Turchyn, A. V. & Ralser, M. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. *Mol. Syst. Biol.* **10**, 725 (2014).

265. Russell, M. J. & Martin, W. The rocky roots of the acetyl-CoA pathway. *Trends Biochem. Sci.* **29**, 358–363 (2004).