KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Biotechnology

Sequence- and structure guided engineering of proteins and enzymes for biotechnology and health applications

KAREN SCHRIEVER

Stockholm, Sweden 2023

Sequence- and structure guided engineering of proteins and enzymes for biotechnology and health applications

KAREN SCHRIEVER

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday, March 24th 2023 at 1:00 p.m. in Air & Fire, Science for Life Laboratory, Tomtebodavägen 23A, Solna.

Doctoral Thesis in Biotechnology
KTH Royal Institute of Technology
Stockholm, Sweden 2023

ISBN 978-91-8040-509-6
TRITA-CBH-FOU-2023:8

Printed by: Universitetsservice US-AB, Sweden 2023

Abstract

Proteins are highly diverse and sophisticated biomolecules that represent a cornerstone of biological structure and function and have been exploited in man-made applications for thousands of years. Those proteins that facilitate chemical reactions at physiologically relevant time-scales are referred to as enzymes. Understanding the connections between proteins' functions and their structures, mechanisms and evolution allows to engineer them towards desired properties for various applications. The aim of the work presented in this thesis is to assess different protein engineering approaches and workflows in the context of health and biotechnology applications. Four proteins were studied and/or engineered towards different outcomes using either sequence-based information, structural information or a combination thereof. In **paper I** a sequence-based approach was applied to optimise vaccine candidates for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Specifically, ancestral sequence reconstruction was used to generate highly stable and soluble antigens that could be produced in high quantities in a low-throughput and structure-independent manner. These ancestral antigens interacted with antibodies from recovered patients and served as scaffolds to host a domain of the extant antigen to further enhance antibody engagement. **Paper II** and **III** applied enzyme engineering to terpene cyclases in a health and biocatalysis context, respectively. In **paper II** a structure-based approach was used to understand the fundamental principles underlying the catalytic mechanism of an enzyme in human steroid metabolism. Specifically, solvent access tunnels were identified and modified to probe the role of activation entropy in human oxidosqualene cyclase, which drastically modified the temperature dependence of catalysis. This finding may also have implications for engineering related plant enzymes for production of industrially relevant compounds in heterologous hosts. In **paper III** sequence- and structure based approaches were used together to engineer substrate specificity in a promiscuous bacterial terpene cyclase. Specifically, the structure of a stable reconstructed ancestor of spiroviolene synthase was determined in order to understand the molecular basis of substrate promiscuity and engineer highly selective variants that retained thermostability. The presented workflow is relevant for engineering these enzymes as biocatalysts for production of terpene-based high value compounds. In **paper IV** the metabolite regulation of a flux-controlling enzyme in the Calvin cycle was studied to eventually engineer it for enhanced growth of autotrophic production hosts. Specifically, interactions between a bifunctional cyanobacterial fructose-1,6-bisphosphatase and a panel of metabolites were identified using a proteomics approach and verified by *in vitro* experiments. A synergistic regulation involving the enzyme's redox state and glyceraldehyde 3-phosphate was discovered, which has implications for integrated metabolic and enzyme engineering approaches involving this biocatalyst. In summary, the results presented herein highlight the utility of integrating several different engineering approaches for proteins used in health and biotechnology applications.

Sammanfattning

Proteiner är mycket diversa och sofistikerade biomoleyler som representerar en hörnsten för biologisk struktur och funktion och har tagits till vara i tillämpade produkter sen flera tusen år tillbaka. De proteiner som underlättar att kemiska reaktioner händer under en fysiologiskt relevant tidsram kallas för enzymer. En förståelse av sammanhangen mellan proteiners funktion och deras strukturer, mekanismer och evolution möjliggör att utveckla åtråvärda egenskaper hos de olika tillämpningarna. Målet med det presenterade arbetet i denna avhandling är att granska olika inriktningar och arbetsflöden för att utveckla proteiner med tillämpningar i områdena hälsa och bioteknik. Fyra proteiner studerades och/eller utvecklades mot olika rön med hjälp av sekvensbaserad information, strukturbaserad information eller en kombination av dessa. I **Artikel I** tillämpades en sekvensbaserad inriktning för att optimera en vaccinkandidat mot severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, svår akut respiratorisk sjukdom coronavirus 2 på svenska). Konkret, så användes rekonstruktion av förfädersproteiner för att generera mycket stabila och lösliga antigener som kunde produceras i stora mängder. Metoden var inte beroende av att testa många proteiner eller strukturell information. Dessa förfädersantigener interagerade med antikroppar från tillfrisknade patienter och kunde användas som strukturell bas för att hysa en domän som tillhör ett nuvarande antigen med syftet att ytterligare förstärka antikroppsinteraktionerna. I **Artikel II** och **III** användes enzymteknik för att utveckla terpencyklaser med tillämpningar inom områdena hälsa respektive biokatalys. I **Artikel II** tillämpades en strukturbaserad inriktning för att förstå de fundamentala principerna som ligger till grund för en enzymmekanism inom den mänskliga steroid metabolismen. Konkret, så identifierades och modifierades accesstunnlar för vatten med syftet att studera aktiveringsentropins roll för humant oxidosqualencyklas, vilket ledde till en drastisk förändring i katalysens temperaturberoende. Denna insikt kan komma at ha betydelse för utvecklingen av relaterade växtenzymer med syfte att producera industriellt värdefulla kemiska föreninger i cellfabriker. I **Artikel III** användes sekvensbaserade och strukturbaserade metoder tillsammans för att utveckla substratspecificitet i ett bakteriellt terpencyklas som katalyserar flera reaktioner. Konkret, så löstes strukturen av ett stabilt, rekonstruerat förfädersenzym till spiroviolensyntas för att förstå den molekylära grunden till att enzymet kan katalysera flera reaktioner och för att utveckla mycket selektiva varianter med bibehållen termisk stabilitet. Det presenterade arbetsflödet är relevant för att utveckla dessa enzymer till industriella biokatalysatorer för att producera terpenbaserade kemiska högvärdesföreningar. I **Artikel IV** studerades hur ett enzym som kontrollerar flödet genom Calvincykeln regleras av metaboliter för att som slutmål utveckla Calvincykeln mot ökad produktion i autotrofiska produktionsvärdar. Konkret, så identifierades interaktionerna mellan ett bifunktionellt fruktos-1,6-bisfosfatas från cyanobakterier och en utvald grupp metaboliter med hjälp av en proteomikmetod och verifierades sedan med hjälp av *in vitro* experiment. En synergistisk reglering upptäcktes som involverar enzymets redoxtillstånd och metaboliten glyceraldehyd 3-fosfat och som har konsekvenser för hur detta enzym behöver modifieras för att kunna appliceras inom metabolismteknik. Sammanfattningsvis visar resultaten i denna avhandling nyttan av att integrera flera olika ingenjörsmässiga strategier för att skräddarsy proteiner med tilllämpningar i hälsa och bioteknik.

Thesis defence

This thesis will be defended on **March 24th, 2023** at **13:00** in conference rooms Air and Fire, Science for Life Laboratory, Tomtebodavägen 23A, Solna, Sweden, for the degree of Doctor of Philosophy (PhD) in Biotechnology.

Respondent:
Karen Schriever, M. Sc. in Biochemistry from Heidelberg University
Department of Fibre and Polymer Technology,
KTH Royal Institute of Technology, Stockholm, Sweden

Faculty opponent:
Prof. Eric A. Gaucher
Department of Biology, Georgia State University, Atlanta, Georgia, USA

Chairperson:
Assoc. Prof. Carsten Mim
Department of Biomedical Engineering and Health Systems,
KTH Royal Institute of Technology, Stockholm, Sweden

Evaluation committee:

Prof. Cathleen Zeymer
Department of Chemistry, Technical University of Munich, Munich, Germany

Prof. Sara Snogerup Linse
Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden

Professor Henrik Toft Simonsen
Department of Biology and Biochemistry, Université Jean Monnet, Saint-Étienne, France

Respondent's main supervisor:
Assoc. Prof. Per-Olof Syrén
Department of Fibre and Polymer Technology,
KTH Royal Institute of Technology, Stockholm, Sweden

Respondent's co-supervisor:
Prof. Paul Hudson
Department of Protein Science,
KTH Royal Institute of Technology, Stockholm, Sweden

Appended papers

This thesis is based on the following articles and manuscripts. Full versions are appended at the end of the thesis with permission of the copyright holders.

I D. Hueting*, K. Schriever*, F. Zuo, L. Du, H. Persson, C. Hofström, M. Ohlin, K. Walldén, L. Hammarström, H. Marcotte, Q. Pan Hammarström, J. Andréll, P.-O. Syrén (2022). Design, structure and plasma binding of ancestral β-CoV scaffold antigens.

*These authors contributed equally to this work
Manuscript in revision; pre-print available on Research Square; https://doi.org/10.21203/rs.3.rs-1909545/v1

II K. Schriever*, D. Hueting*, A. Biundo, C. Kürten, T. Braun, S. Govindarajan, C. Gustavsson, P.-O. Syrén (2023). Designed out-of-active-site mutations in human oxidosqualene cyclase modulate the activation entropy and enthalpy of the cyclization reaction.

*These authors contributed equally to this work
Manuscript

III K. Schriever, P. Saenz-Mendez, R. Srilakshmi Rudraraju, N. M. Hendrikse, E. P. Hudson, A. Biundo, R. Schnell, P.-O. Syrén (2021). Engineering of Ancestors as a Tool to Elucidate Structure, Mechanism, and Specificity of Extant Terpene Cyclase. *J. Am. Chem. Soc.*, *143, 10,* 3794–3807

IV E. Sporre*, J. Karlsen*, K. Schriever, J. Asplund Samuelsson, M. Janasch, L. Strandberg, D. Kotol, L. Zeckey, I. Piazza, P.-O. Syrén, F. Edfors, E. P. Hudson (2023). Metabolite interactions in the bacterial Calvin cycle and implications for flux regulation.

*These authors contributed equally to this work
Manuscript in review; pre-print available on bioRxiv; https://doi.org/10.1101/2022.03.15.483797

Respondent's contributions to appended papers

I Contributed to conceptualisation, planned and performed a major part of the experiments, performed data analysis and visualisation, wrote the manuscript together with co-first author and corresponding authors and with contributions from co-authors.

II Planned and performed a major part of the experiments, performed data analysis and visualisation, wrote the manuscript together with co-first author and corresponding author.

III Contributed to conceptualisation, planned and performed the majority of the experiments, performed data analysis and visualisation, wrote the manuscript together with corresponding author and with contributions from co-authors, created the cover art.

IV Planned and performed part of the experiments, performed minor part of data analysis and visualisation, wrote a minor part of the manuscript, assisted in scientific discussions and revising the manuscript.

Related work not included in the thesis

B. Guo, S. Reddy Vanga, X. Lopez-Lorenzo, P. Saenz-Mendez, S. Rönnblad Ericsson, Y. Fang, X. Ye, K. Schriever, E. Bäckström, A. Biundo, R. A. Zubarev, I. Furó, M. Hakkarainen P.-O. Syrén (2022). Conformational Selection in Biocatalytic Plastic Degradation by PETase. *ACS Catal.*, 12, 6, 3397–3409

Commonly used abbreviations

ASR	Ancestral sequence reconstruction
ACE2	Angiotensin converting enzyme 2
AnSA	Ancestral scaffold antigen
C. necator (*cn*)	*Cupriavidus necator*
cryo-EM	Cryogenic electron microscopy
DLS	Dynamic Light Scattering
E. coli (*ec*)	*Escherichia coli*
FPP	Farnesyl pyrophosphate
F/SBPase	Fructose-1,6-/sedoheptulose-1,7-bisphosphatase
GAP	Glyceraldehyde 3-phosphate
GC	Gas chromatography
GGPP	Geranylgeranyl pyrophosphate
hOSC	Human oxidosqualene cyclase
MD simulations	Molecular dynamics simulations
MSA	Multiple sequence alignment
nanoDSF	Nano differential scanning fluorimetry
NTD	N-terminal domain
PDI	Polydispersity Index
RBD	Receptor binding domain
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SHC	Squalene hopene cyclase
S protein	Spike protein
SvS	Spiroviolene synthase
Synechocystis (*syn*)	*Synechocystis sp. PCC6803*
TC	Terpene cyclase
WT	Wild type

Contents

1 - Protein Stability

1.1 Brief introduction to protein structure

Proteins are bio-macromolecules consisting of amino acid residues that fold into complex three-dimensional (3D) structures, which enables them to fulfil very specific functions. Proteins are ubiquitous in every living organism and perform various tasks among many others in the form of enzymes that catalyse chemical reactions (**chapter 2**), structural proteins that build up the texture of certain tissues, motor proteins that help cells move and regulatory proteins that steer the cell cycle or immune system components that can recognise and react to pathogens. This thesis is focused on proteins in health contexts (surface antigens in **paper I**, enzymes in **paper II**) as well as proteins in biotechnological contexts (enzymes in **papers III** and **IV**).

A protein's 3D structure is mainly governed by the interactions between the sidechains and backbones of the different amino acids that the protein is composed of, but can also be impacted by interactions with its immediate environment, such as other proteins or small molecules in solution. In most proteins, short segments fold into local structures that are stabilised by backbone interactions within the segment. These folded segments are referred to as elements of secondary structure, the most common being α-helices, β-sheets and β-turns. These secondary structure elements are typically connected by more flexible loops and interact in all three dimensions *via* backbone and sidechains of amino acids to create super-secondary structures (also called motifs). Whereas the global arrangement of secondary structure elements and super-secondary structures within one protein chain is referred to as the tertiary structure of the protein, the overall structure resulting from the interaction of several separate protein chains is referred to as the quaternary structure. Although protein functionality is highly complex and context-dependent, it is well established that the single or multiple tertiary or quaternary structures that a protein can assume have a defining impact on its functionality. Understanding the 3D structure of proteins is therefore of high interest to scientists, as illustrated by the various experimental and computational techniques that exist to predict or measure protein structures. Whereas cryogenic electron microscopy (cryo-EM) was used to solve and study the quaternary structure of a surface antigen in **paper I**, X-ray crystallography was used in **paper III** to delineate the structure-function relationship of a terpene cyclase. In **papers II** and **IV**, on the other hand, X-ray crystal structures were available from literature beforehand and were used to understand and modify native protein function.

1.2 The folding energy landscape and protein stability

Proteins are biologically produced as a linear chain of amino acids (corresponding to the primary structure) on the ribosomes. As a result, a protein needs to first undergo a folding process to attain its final shape.

Protein folding is a highly complex process and some elements of protein folding are still not fully understood. It has been known since for over half a century that small monomeric proteins are able to spontaneously refold after denaturation,[1] indicating that the amino acid sequence dictates a protein's fold due to interactions between sidechains and backbones of the different amino acids. It is also well established that the spontaneous collapse of hydrophobic regions to attain dense solvent-excluded packing (also known as the hydrophobic effect) is strongly entropy favoured due to the release of water molecules to the bulk solvent and is a major driver of protein folding early on in the folding trajectory.[2] However, several complex aspects of protein folding make it more challenging to understand and predict protein folding trajectories. Not all segments of a nascent protein are for instance simultaneously available for folding during the progressive addition of amino acids by the ribosome. Moreover, many proteins require specialised chaperones to fold into their correct shape. Some proteins are intrinsically disordered and only fold in a particular context, *e.g.* in the presence of a particular ligand or binding partner,[3] whereas other proteins that undergo conformational changes have several stably folded states. Adding to this complexity is the fact that many proteins fold (and denature or aggregate) in a stepwise manner progressing through meta-stable folded intermediates.[4,5]

Conceptually, the energetics of protein folding can be described by an energy function in which unfolded or partly folded states correspond to higher energy levels and folded states correspond to lower energy levels. The global minimum (or minima) of the resulting folding landscape represents the natively folded conformation(s) of the protein. Transitions between different folded states (local minima) occur through unfolded and energetically unfavourable states that occupy the maxima in the folding energy landscape. In this context, a protein's stability can be thought of as the energy difference between the natively folded, stable state and unfolded, unstable states. In practice, this energy difference corresponds to the amount of additional energy a system can absorb in form of heat, changes in salinity or pH, UV-irradiation or other structural stressors, before the energetic barrier for unfolding is overcome. The folding landscape for one protein may change at *e.g.* different temperatures or pH values, due to the changed energetics of desolvation/solvation and charges, such that a single natively folded conformation may not always dominate across all possible physiological conditions.

On a molecular level, protein stability is afforded by interactions that lower the energy of a protein conformation with respect to unfolded conformations. Besides the hydrophobic effect mentioned above, these interactions comprise *e.g.* electrostatic interactions between charged side chains (ca. <1 kcal mol^{-1} and 4-5 kcal mol^{-1} on the protein surface and in the core, respectively),[6,7] cation-π interactions between and positive charges and aromatic rings (ca. 0.5 kcal mol^{-1}),[8] hydrogen bonds between polar side chains and backbones (ca. 0.5 - 4 kcal mol^{-1}) and Van-der-Waals interactions (ca. 0.1 -1 kcal mol^{-1})[9]. The total favourable energy contribution from interactions across all amino acids of a protein balances the entropic penalty associated with folding the protein into a single or few narrowly constrained conformation(s). On the other hand, protein stability is also promoted by interactions that raise the energy of alternative or unfolded protein conformations with respect to the folded conformation, such as buried unbalanced charges or hydrogen bonds, entropically unfavourable trapped solvent molecules, repulsion of identical charges or steric clashes. Algorithms that were developed to design stable proteins based on structural considerations take either option into account and are referred to as positive and negative design approaches, respectively.

1.3 Marginal stability of mesophilic proteins

Most proteins found in mesophilic species, *i.e.* species that are adapted to living at intermediate temperatures, are found to be only marginally stable. On average, the protein stabilisation energy, the energy difference between folded and unfolded conformations, is rather moderate (ca. 5-12 kcal mol^{-1}),[10] which corresponds to the breaking of just a few hydrogen bonds. In fact, several human diseases are caused by protein mutations that destabilise particular proteins, such as p53 mutations in cancer or fibril formation of proteins in neurodegenerative diseases, resulting in differences in effective available protein concentration and the accumulation of toxic aggregates in cells. This highlights that even subtle mutations can suffice to undermine the marginal stability of mesophilic proteins with grave consequences.

In some cases, the stability of proteins may be limited by the fact that certain protein regions need to be flexible, *e.g.* to allow for conformational change during catalysis. If the protein would be more rigid, it would not be able to perform the required movement and would essentially become inactive, a compromise that is sometimes described as the "functionality-stability trade-off".[11] Nevertheless, the existence of many highly stable and functional proteins in extremophiles as well as the exceptional stability of some *de novo* designed proteins,[12,13] highlight that marginal protein stability is not necessarily a molecular requisite of functional protein dynamics in general.

Instead, the absence of positive selection towards higher stability in mesophiles is a more plausible explanation for marginal protein stability.[14] As a result, and due to genetic drift, original structural features conferring protein stability, such as hydrogen bonds on the surface of proteins, may be lost over time. Moreover, since protein stability is inherently related to protein folding, a protein's interaction with the host cell's protein folding and homeostasis machinery (such as chaperones or the unfolded protein response pathway) influences effective protein stability.[15] Protein stability is also impacted by the interactions of a protein with other proteins or small molecules in a crowded cell environment. All of these interactions can buffer destabilising mutations and may therefore allow them to accumulate over time as long as they do not hamper the core functionality of the respective protein. The cell's folding machinery and milieu therefore contribute to diminishing the selection pressure for high stability on mesophilic proteins. When producing proteins heterologously in another organism than the original species in a laboratory setting, such buffering effects *e.g.* interactions with specific chaperones may be absent or modified in the novel host, and can result in a higher degree of misfolding and thus low soluble protein yields.

1.4 Why enhance protein stability?

Many proteins are used in applied contexts, such as enzymes in biosynthetic or chemo-enzymatic catalytic processes or antibodies and protein-binding proteins in medical applications. In these contexts, it can be desirable to enhance protein stability, for example to enable the use of enzymes at higher operating temperatures or other harsh reaction conditions, as frequently required in industrial processes. High protein stability can also increase the half-life of biocatalysts in synthetic applications (thus increasing effective turnover) or of biopharmaceuticals inside the body and may extend shelf-life times. Another reason for enhancing a protein's stability is the fact that stable proteins can often be more easily produced in a heterologous expression system with high soluble yields compared to less stable proteins, thus reducing the relative cost of protein production. Importantly, enhanced protein stability also facilitates a protein's biophysical characterization, since stable proteins are typically easier to work with *in vitro* and better amenable to structural studies (**paper III**).[16] Finally, stable proteins are generally associated with greater evolvability.[17] Proteins often need to be adapted to specific industrial tasks that do not correspond to their natural activity, typically by introducing several mutations. It is estimated that ca. 80 % of all mutations that can theoretically be introduced into an average protein sequence result in misfolding and are thus detrimental to protein activity.[18] Very stable proteins that have a high energy barrier of unfolding, are more likely to retain their overall fold and may buffer the accumulation of several destabilising mutations. Consequently, protein stability is also advantageous for engineering and adapting proteins for a specific task in various industrial applications. In **chapter 3.4** different techniques for enhancing protein stability are outlined.

2 - Enzymes

2.1 Enzymatic rate enhancement

2.1.1 Thermodynamics of chemical reactions

The driving force of a chemical reaction is defined by its Gibb's free energy ΔG which is described by Equation 1, in which ΔH is the change in enthalpy and ΔS is the change in entropy that are associated with the transition of substrates to products of the reaction.

$$\Delta G = \Delta H - T \times \Delta S \qquad 1$$

The enthalpy can be thought of as the summed amount of energy resulting from intra- and inter-molecular interactions in the system, such as *e.g.* hydrogen bonds and electrostatic interactions, whereas the entropy can – in simplified terms - be thought of as the orders of degrees of freedom that a system can occupy. If $\Delta G < 0$, a reaction will occur spontaneously (exergonic reaction), whereas free energy needs to be supplied to the system in order for an endergonic reaction ($\Delta G > 0$) to proceed. An example of a ubiquitous exergonic reaction in Nature is the hydrolysis of a phosphate group from adenosine triphosphate (ATP) , which has a ΔG value of -7.3 kcal mol^{-1} at standard conditions (ΔG^0, 25 °C, 1.0 M of all reaction components, 1.0 atm pressure).[19]

While processes such as ATP hydrolysis are thermodynamically favoured, *i.e.* the driving force of the reaction would make them occur spontaneously, many such processes occur very slowly (the rate of spontaneous ATP hydrolysis in solution for instance is 3.2×10^{-5} s^{-1} at pH 8.4[20]). A classic example for this is cellular aerobic respiration of glucose to CO_2 and water by a set of metabolic processes (including glycolysis, the citric acid cycle and oxidative phosphorylation), which thermodynamically releases 690 kcal mol^{-1} of Gibb's free energy but would never occur spontaneously in water at relevant timescales.

2.1.2 Transition state theory

The molecular mechanism of a reaction is typically depicted along a reaction coordinate (Fig. 1). Many reactions are characterised by the occurrence of multiple reaction intermediates; short-lived molecules that are instable (high positive ΔG to the ground state) but that can be experimentally isolated under special experimental conditions. Each two mechanistic steps of a reaction are connected through a transition state in which bonds are partially broken and formed to equal extents. By definition, transition states represent the maximum of the Gibb's free energy profile between two mechanistic states along the reaction coordinate (Fig. 1). As such, they have estimated lifetimes in the order of single bond vibrations and can therefore be inferred but are highly challenging to observe or isolate. According to simplified Eyring transition state theory (Equation 2), the rate constant of a reaction k depends on the difference in Gibb's free energy between the ground state (system state before the reaction) and the Gibb's free energy of the transition state $\Delta G^{\ddagger}$, also referred to as the Gibb's free energy of activation (Fig. 1) as well as the reaction temperature T. All other factors in Equation 2 are constants including the Boltzmann constant k_B, Planck's constant h and the universal gas constant R.

$$k = \frac{k_B T}{h}\, e^{-\frac{\Delta G^{\ddagger}}{RT}} \qquad 2$$

Specifically, in reactions involving more than one mechanistic step the one with the highest energy transition state contributes most significantly to the overall rate of reaction and is therefore referred to as the rate-limiting step. Since the rate constant exponentially decays with $\Delta G^{\ddagger}$, a decrease of $\Delta G^{\ddagger}$ by 1.4 kcal mol^{-1} at 25 °C would translate to a 10 x rate enhancement.

Enzymes are protein catalysts that enhance the rate of chemical reactions by lowering the Gibb's free energy of the transition state or by altering the reaction mechanism to proceed through a lower energy transition state (Fig. 1). In fact, one of the hallmarks of enzyme catalysis is the stabilisation of transition states by amino acid side-chains of the enzyme as compared to aqueous solution. Enzymatic rate enhancements (quantified as the quotient of the rate constants of catalysed *vs.* un-catalysed reactions) can range from

10^6 (*e.g.* chorismate mutase[21]) to 10^{26} (*e.g.* bacterial sulfatases[22]) and thus enable chemical reactions to happen at physiologically relevant time scales in Nature.

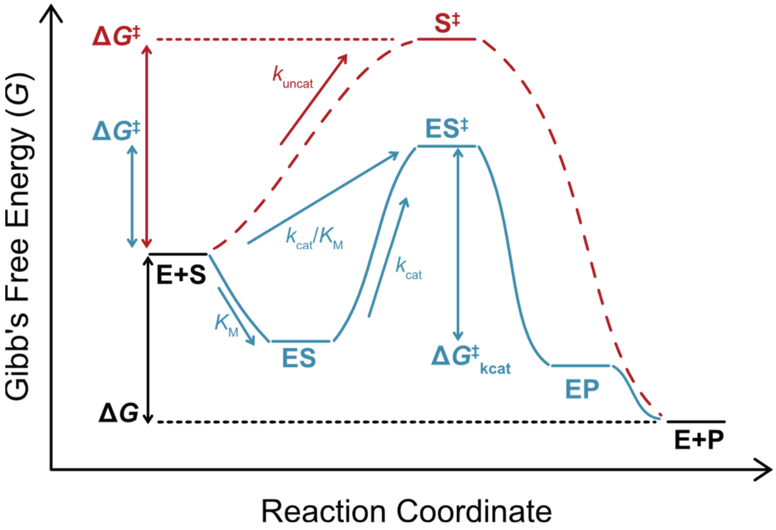


Figure 1. Energy profile of an enzymatic reaction (teal) compared to the un-catalysed reaction (red). The change in Gibb's free energy (ΔG) for the reaction from substrates to products is shown in black. Activation Gibb's free energy ($\Delta G^‡$) is shown for the un-catalysed reaction (red) and for the enzyme-catalysed reaction (teal). Enzymatic activation Gibb's free energy ($\Delta G^‡$) specifically refers to the difference between ground state and transition state ($\Delta G^‡_{\mathrm{kcat/Km}}$). The respective difference between the Michaelis complex and transition state ($\Delta G^‡_{\mathrm{kcat}}$) applies under saturating conditions (see **chapter 2.2.1**) and is indicated in teal. Rate constants associated with the formation of un-catalysed transition state (k_{uncat}) and catalysed transition state ($k_{\mathrm{cat}}/K_{\mathrm{M}}$) from the ground state as well as the formation of catalysed transition state from the Michaelis complex (*k*cat) are indicated. Additional transition states such as between ground state and Michaelis complex as well as between enzyme-product complex and free products are omitted for clarity.

2.2 Enzyme kinetics

2.2.1 Michaelis-Menten model

In order to compare activities of one enzyme with different substrates or between different enzyme variants it is important to describe enzyme-catalysed rates quantitatively in a universal manner, which is achieved by enzyme kinetics.

The overall rate of an enzyme-catalysed unimolecular reaction depends both on the rate of formation of an enzyme-substrate complex (ES, also called Michaelis complex), and on the rate of conversion to the product (Equation 3) in which E is enzyme, S is substrate, P is product and k_1, k_{-1}, k_2 *and* k_{-2} are rate constants associated with the respective steps.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E + P \qquad 3$$

Under the assumption that product is absent in the beginning of the reaction, the reverse reaction (*i.e.* starting from product and enzyme to form the enzyme-substrate complex) becomes negligible and the initial reaction rate v_0 can thus be expressed by the formation of product over time, according to Equation 4.

$$v_0 = \frac{d[P]}{dt} = k_2\,[ES] = k_{cat}\,[ES] \qquad 4$$

The rate constant k_2 represents the unimolecular rate constant for transformation of the ES complex to form the product and is referred to as k_{cat} (also called turnover number). Assuming (i) that the rate of association of E and S (Equation 5) and ES-dissociation (Equation 6) are equal (quasi-steady state approximation according to Briggs-Haldane[23]),

$$\frac{d[ES]}{dt} = k_1[E][S] \qquad 5$$

$$-\frac{d[ES]}{dt} = k_{-1}[ES] + k_{cat}\,[ES] \qquad 6$$

(ii) that free substrate [S] is in excess and therefore quasi-constant and (iii) that the total enzyme E_{tot} can be available as free enzyme E or enzyme-substrate complex ES (Equation 7)

$$[E]_{tot} = [E] + [ES] \qquad 7$$

the concentration of [ES] can be obtained by Equation 8, which can be reformed to Equation 9, in which K_M is the Michaelis-constant.

$$k_1([E]_{tot} - [ES]) \times [S] \; = \; k_{-1}[ES] + \; k_{cat}\,[ES] \qquad 8$$

$$[ES] = \frac{k_1[E]_{tot} \times [S]}{k_{-1} + k_{cat} + k_1[S]} = \frac{[E]_{tot} \times [S]}{\frac{k_{-1} + k_{cat}}{k_1} + [S]} = \frac{[E]_{tot} \times [S]}{K_M + [S]} \qquad 9$$

When substituting equation 9 into equation 4, the rate of an enzymatic reaction can thus be expressed according to the Michaelis-Menten equation (equation 10)[24]

$$v_0 = \frac{k_{cat} \times [E]_{tot} \times [S]}{K_M + [S]} = \frac{v_{max} \times [S]}{K_M + [S]} \qquad 10$$

The kinetic parameters k_{cat} (in units of s^{-1}) and K_{M} (in units of M) are enzyme-inherent constants that characterise the enzyme in a particular state (*e.g.* for a particular substrate at a defined buffer composition, pH and temperature). The measured initial rate of reaction (*i.e.* the linear accumulation of product over time) increases with the substrate concentration until it asymptotically approaches v_{max}. In this condition, the enzyme is saturated and no further rate increase will occur with increases in substrate concentration (Fig. 2A).

The Michaelis-Menten equation describes initial rates for enzyme-catalysed unimolecular reactions, which are common in terpene cyclases, as presented in **paper II** and **III**. While rate equations for bimolecular and higher order molecular reactions can be mathematically derived using similar assumptions as the Michaelis-Menten equation, many bimolecular reactions are practically treated as quasi-unimolecular by adding one of the substrates in such excess that its concentration can be considered constant. This is *e.g.* the case for hydrolysis reactions, as exemplified by fructose-1,6-bisphosphatase (FBPase) (**paper IV**) in which water molecules are available in excess over the other substrate fructose-1,6-bisphosphate.

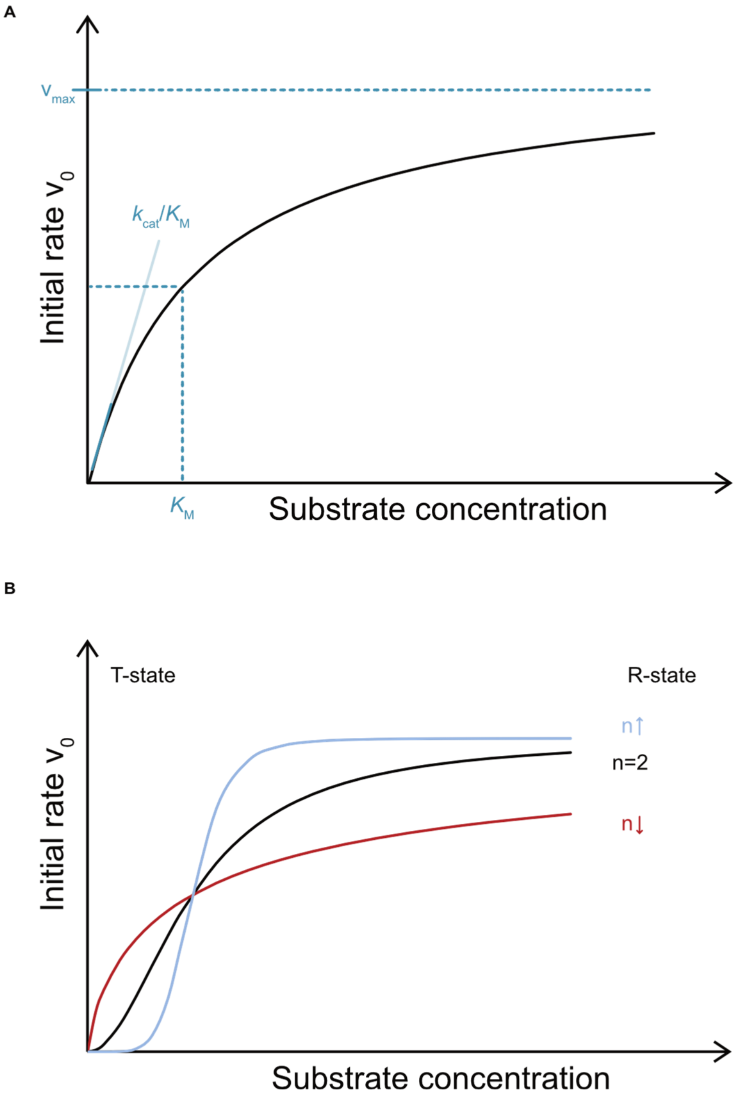


Figure 2. Representative kinetic plots. (A) A representative Michaelis-Menten kinetic plot is shown. At very low substrate concentrations the initial rate increases linearly with the second order rate constant k_{cat}/K_M (shown as teal line). **(B)** Representative cooperative kinetic plots are shown. The black line represents the kinetic behaviour of a cooperative enzyme with a Hill coefficient of two. The effects of increasing (n=6) and decreasing (n=0.75) Hill coefficient on the cooperative plot at otherwise identical kinetic parameters is shown in blue and red, respectively.

The Michaelis-constant is defined as (k_{-1} + k_{cat})/k_1 (equation 9). Assuming that the rate constant for the chemical conversion k_{cat} is much lower and thus negligible compared to rate constants for substrate binding (k_1) and dissociation (k_{-1}) (rapid equilibrium approximation), K_M approximates the dissociation constant K_D for the Michaelis complex (k_{-1}/k_1). For this reason, the K_M value is frequently interpreted as a type of dissociation constant that describes substrate affinity to the enzyme.

For the particular case in which [S] = K_M, the initial rate corresponds exactly to half the maximum rate (equation 11). The K_M value therefore describes the substrate concentration at which half the maximum velocity is attained (Fig. 2A).

$$v_0 = \frac{v_{max} \times [S]}{K_M + [S]} = \frac{v_{max} \times [S]}{[S] + [S]} = \frac{v_{max} \times [S]}{2\,[S]} = \frac{1}{2}\, v_{max} \qquad 11$$

At very high substrate concentrations ([S] >> K_M), equation 10 can be simplified to equation 12, which represents a zero-order rate equation.

$$v_0 = \frac{v_{max} \times [S]}{K_M + [S]} \approx \frac{v_{max} \times [S]}{[S]} = v_{max} \qquad 12$$

At very low substrate concentrations ([S] << K_M) equation 10 can be simplified to equation 13, which represents a second-order rate equation with the rate constant k_{cat}/K_M. In this far-from-saturation substrate concentration range, initial rates increase linearly with substrate concentration (Fig. 2A). This behaviour is the case for human oxidosqualene cyclase (hOSC), as discussed in **paper II**.

$$v_0 = \frac{v_{max} \times [S]}{K_M + [S]} \approx \frac{v_{max} \times [S]}{K_M} = \frac{k_{cat}}{K_M} \times [S] \times [E]_{tot} \qquad 13$$

In summary, enzyme rates can be characterised by the apparent second-order rate constant k_{cat}/K_M, which is also referred to as "catalytic efficiency" and mathematically describes how both the process of substrate binding and generation of the transition state influence the overall reaction rate.

2.2.2 Cooperativity and allostery

Despite its usefulness and wide-spread use, the Michaelis-Menten equation is not suitable to describe more complex kinetic behaviours, such as *e.g.* in case of substrate activation or inhibition.

Similarly, cooperative enzyme kinetics require the modification of the standard Michaelis-Menten equation to equation 14, in which n is the Hill coefficient and K' is an equilibrium constant that takes into account multiple sequential binding events.

$$v_0 = \frac{v_{max} \times [S]^n}{K_M^n + [S]^n} = \frac{v_{max} \times [S]^n}{K' + [S]^n} \qquad 14$$

The quaternary structures of cooperative enzymes consist of several identical catalytic subunits that do not bind the substrate with equal affinity. Instead, the binding of substrate to one enzyme subunit changes the quaternary structure in such a way, that substrate binding in the next subunit is either enhanced (positive cooperativity, $n > 1$) or decreased (negative cooperativity, $0 < n < 1$). Since cooperativity describes the effective change of protein structure at a site other than where the original substrate binds, it represents a particular form of allostery.

Michaelis-Menten-Hill plots for positively cooperative enzymes are characterised by a sigmoidal shape (Fig. 2B, black curve). This shape reflects a first gradual and then exponential transition of the empty enzyme from its "tense" state (T-state) - in which the equilibrium is shifted towards free enzyme - to its "relaxed" state (R-state). A high Hill coefficient can be biologically interpreted as an increased sensitivity to the available substrate concentration; while the enzyme's inherent activity is very low, it readily increases as soon as a certain substrate concentration is available. In this work, cooperative kinetics were observed for a dimeric terpene cyclase (**paper III**) and a tetrameric phosphatase (**paper IV**).

Cooperative enzymes are frequently regulated by allosteric effectors, which are small molecules such as metabolites that effect a conformational change of the enzyme distal to their binding site. Allosteric inhibitors can increase the Hill-coefficient of a cooperative enzymatic reaction by preferentially binding to and stabilizing the T-state of the enzyme. As a consequence, the resulting kinetic plot will appear even more sigmoidal (Fig. 2B, blue curve) and reaction rates at low substrate concentration will be decreased compared to the un-inhibited reaction. On the contrary, allosteric inducers may decrease the Hill-coefficient of a cooperative enzyme reaction by preferentially binding to and stabilizing the R-state of the enzyme. Consequently, the kinetic plot will appear less sigmoidal (Fig. 2B, red curve) and reaction rates at low substrate concentrations will be increased compared to the un-induced reaction. In **paper IV** allosteric effects of various metabolites on enzymes of the Calvin cycle were analysed in different autotrophic bacteria.

2.3 Catalytic mechanisms

As described in **chapter 2.1.2**, enzymes enhance reaction rates either by reducing the Gibb's free energy of the transition state or by providing an alternative reaction mechanism that involves a different stabilised transition state corresponding to a lower $\Delta G^{\ddagger}$. In most enzyme mechanisms this is the result of optimal binding of the transition state to the enzyme active site, as initially suggested by Linus Pauling in 1946.[25]

Another aspect of enzyme catalysis refers to specific substrate recognition. As early as 1894, a model of geometric complementarity between enzymes and substrates was taken forward by Emil Fischer, which he figuratively described as lock and key, respectively.[26] While this model laid the foundation for the description of enzyme substrate interactions, it did not yet specifically account for enzyme dynamics. By now, two well-established conceptual models describe how enzymes afford optimal complementarity of their active site with the substrate through dynamic rearrangements. According to the induced fit model,[27] an enzyme initially exists in a conformation that is not perfectly complementary to the ligand. The substrate binding event then induces subtle conformational rearrangements of the enzyme such that its binding in the changed active site becomes more energetically favourable. In the conformational selection model,[28,29] the enzyme samples multiple different conformations in the ground state, some of which are perfectly complementary to the ligand. The ligand then “selects” the optimal conformation and stabilises it. These two mechanisms are closely related - the first involving a conformational change after a binding event, the latter involving a conformational change before the binding event.[30,31]

Besides general principles of enzyme catalysis, such as transition state stabilisation, enclosure of reactants in close proximity and optimal spatial arrangement to each other, or control of active site solvation, different sub-groups of amino acids typically perform more specific catalytic tasks. According to a review of the Mechanism and Catalytic Site Atlas,[32] a manually curated database of enzyme mechanisms covering 734 entries with detailed information about mechanistic steps (as of February 2023),[33] catalytic amino acids can be assigned to assume reactant, spectator and interaction-roles.

2.3.1 Breaking and forming of covalent bonds

Catalytic residues within the “reactant role” category are involved in breaking and forming bonds, and can be further grouped as performing covalent catalysis, hydrogen species transfer catalysis or electron transfer catalysis. Covalent catalysis involves the transient formation of covalent bonds between a catalytic enzyme residue and the substrate - most commonly *via* nucleophilic attack of a side-chain (such as cysteine, lysine or serine) on the ligand. Less frequent mechanisms involve electrophilic attack or radical mechanisms. Catalysis by hydrogen species transfer comprises mechanisms by which amino acid side-chains donate or accept either protons, hydrogen radicals or hydrides to or from the ligand. Among these, the transfer of protons is by far the most common. Typical amino acids involved in general acid and base catalysis are histidine, glutamate and aspartate and somewhat less frequently lysine, cysteine, tyrosine and serine. Tyrosine also commonly relays protons from catalytic residues to the ligand and *vice versa*. This is believed to be the case in hOSC (**paper II**), in which the substrate is initially protonated by a catalytic aspartate and finally deprotonated by a tyrosine residue that relays the proton to the catalytic base, a histidine residue (**chapter 4.2.3**). Finally, electron transfer catalysis in which catalytic residues accept and donate either electron pairs or single electrons or mediate electron tunnelling exist but are not very common.

2.3.2 Facilitation of chemical reactions

Within the "spectator role" category, catalytic residues do not directly react with the ligand but support the main catalytic mechanism and can be sub-grouped into electrostatic interactors, activators and steric interactors. Electrostatic interactors are involved in electrostatic stabilisation of the substrate and transition state or electrostatic destabilisation of the ground state. Electrostatic stabilisation is the most common among these mechanisms. In fact, dielectric constants in enzyme active sites have been estimated to be as low as 3-4,[34] which is in the same range as concrete (water has a dielectric constant of ca. 80 at 20 °C[35]). Therefore, any electrostatic interactions that may be considered weak to moderate in water due to hydration effects would have a much stronger effect in an enzyme active site.[36] Electrostatic stabilisation is achieved *e.g.* by ionic interactions between a (partially) charged transition state and residues with acid or base properties Another mechanism that is highly relevant for terpene cyclase catalysis (**paper II** and **paper III**) is electrostatic stabilisation of cations by delocalised π-charges in aromatic sidechains (tryptophan, tyrosine, and phenylalanine). Activators in contrast, amplify the function of residues involved in the reactant role category, by either enhancing or decreasing the reactant side-chain's acidity, nucleophilicity or redox potential. An example of this is observed in hOSC (**paper II**) in which two cysteine residues in vicinity of the catalytic aspartate are believed to acidify and thus activate the latter to protonate the substrate. Finally, the spectator role category also comprises steric interactors that direct the steric outcome of the reaction by steric hindrance or facilitation. This mechanism is crucial in terpene cyclase active sites (**paper II** and **paper III**) which pre-fold long and flexible linear poly-isoprene substrates into a conformation that facilitates catalytic cyclization.

2.3.2 Cofactors

The third category involving residues with assigned "interaction roles" is more heterogeneous and describes catalytic residues that interact with other molecules in the active site. This involves mostly hydrogen bond or polar and packing interactions with cofactors (metal ions) or coenzymes (small molecules), which in turn can mediate the above-mentioned catalytic functions such as hydrogen species transfer.

Approximately half of catalytic reactions curated in the Mechanism and Catalytic Site Atlas occur with help of a cofactor or coenzyme.[37] In this thesis, a dimeric diterpene terpene cyclase discussed in **paper III** and FBPase originating from the *Synechocystis sp. PCC6803* Calvin cycle discussed in **paper IV** required a divalent metal cofactor for catalysis. In fact, the inhibitory effect of some metabolites such as citrate and ATP on *Synechocystis* FBPase are likely caused by chelation of its cofactor. By contrast, the monomeric triterpene cyclase discussed in **paper II** catalyses its cyclisation reaction without aid of a cofactor.

2.4 Enthalpy and entropy in enzyme catalysis

2.4.1 Implications of reaction temperature for enthalpy and entropy

Gibb's free energy of activation is related to activation enthalpy and activation entropy according to equation 15.

$$\Delta G^{\ddagger} = \Delta H^{\ddagger} - T \times \Delta S^{\ddagger} \qquad 15$$

If a reaction has a negative value for $\Delta S^{\ddagger}$, meaning that the degrees of freedom are reduced in the transition state compared to the ground state, the reaction needs to overcome an entropic barrier for the reaction to occur ($-T \text{ x } \Delta S^{\ddagger} > 0$). The effect of an entropic energy barrier is amplified at increased temperatures, therefore entropy-penalised reactions will be even less favourable at high temperatures. Conversely, entropy-favoured reactions ($-T \text{ x } \Delta S^{\ddagger} < 0$) would be highly favourable at high temperatures. Due to the fact that high entropy barriers are often accompanied by lower enthalpy barriers and *vice versa* (enthalpy-entropy compensation[38]), high-temperature reactions are therefore likely to be associated with favourable entropy and less favourable enthalpy of activation. In contrast, lower-temperature reactions typically have a more favourable activation enthalpy term and unfavourable activation entropy term, since the entropic cost carries less weight at lower temperatures.

Another way to mathematically describe the same relationship is by substituting equation 15 into the Eyring equation (equation 2).

$$k = \frac{k_B T}{h} e^{-\frac{\Delta H^{\ddagger}}{RT}} e^{\frac{\Delta S^{\ddagger}}{R}} \qquad 16$$

From the resulting equation 16 it becomes apparent that – assuming an unfavourable activation enthalpy ($\Delta H^{\ddagger} > 0$) - the exponent of the first exponential term ($-\Delta H^{\ddagger}/RT$) becomes negative and its absolute value therefore increases with increasing temperatures. For reactions with unfavourable activation enthalpy, an increasing temperature therefore contributes to exponential rate enhancement or – in other words – a decreasing temperature therefore contributes to exponential rate decay. In summary, reactions at low temperatures therefore benefit from less unfavourable activation enthalpy, whereas reactions at high temperature benefit from favourable activation entropy.

2.4.2 Predominance of enthalpic enzyme catalysis

The catalytic mechanisms discussed this far, such as forming and breaking bonds in reaction with the substrate and stabilisation of the transition state by electrostatic interactions and steric interactions, involve different forms of electronic interactions. They lower the Gibb's free energy of activation $\Delta G^{\ddagger}$ mainly by decreasing the internal energy of the transition state *i.e.* the activation enthalpy $\Delta H^{\ddagger}$ by several kcal mol^{-1}. Enthalpic catalysis (in particular *via* electrostatic effects in pre-organised active sites) is considered the main mode of enzyme operation.[39,40] This observation is further supported by the fact that most cellular metabolites are charged.[41] Similarly, Wolfenden and coworkers have demonstrated that enzymatic rate enhancement is mainly conferred by decreased activation enthalpies compared to the un-catalysed reaction.[42] They further reason that catalytic rates in early enthalpy-driven enzymes (*i.e.* reactions proceeding with a low enthalpy barrier) would have automatically increased as the Earth cooled down from primordial warmer temperatures in line with the relationship outlined above (equation 16).[43,44] Furthermore, enzymes originating from psychrophilic (cold-adapted) organisms have been observed to display lower (*i.e.* more favourable) enthalpy of activation than their mesophilic counter-parts at the expense of a lower (*i.e.* more unfavourable) entropy of activation.[45,46] Interestingly, an analysis of thermodynamic parameters obtained from structural simulations of psychrophilic, mesophilic and thermophilic citrate synthases indicated that structural changes associated with lowered activation enthalpy were located outside of the active site.[47] Similar observations were made when comparing mesophilic and psychrophilic trypsin.[48] In this case mutations and altered mobility in surface residues were found to contribute to lowering the activation enthalpy in the cold-adapted homologue, indicating involvement of long-range interactions.

2.4.3 Solvent entropy in enzyme reactions involving charge transfer

Despite the pre-dominant role of enthalpy in enzyme catalysis, the contribution of entropy cannot be entirely discounted. Several, albeit few, examples of enzyme reactions involving a low activation entropic penalty or even a favourable activation entropy have been described to date, including cytidine deaminase,[49] bacterial squalene hopene cyclase (SHC),[50] elongation factor Tu (EF-Tu) GTPase[51] and peptide bond formation on the ribosome.[52]

In the 1970s Jencks proposed an enzymatic principle involving ground state destabilisation, which he termed the Circe effect.[53] According to this hypothesis an enzyme binds its substrate tightly (in non-reactive regions) so that it compensates for positioning the substrate's reactive group into a destabilizing environment. As a consequence, the Gibb's free energy of the ES-complex is raised (Fig. 1), so that the difference from the ES-complex to the transition state $(\Delta G^{\ddagger}_{kcat})$ is reduced. The principle has often been used to interpret the role of entropy in enzyme reactions: the large enthalpy benefit of substrate binding mostly compensates the entropic cost of confining the substrate, so that there is no further entropy cost of reaching the transition state from the ES-complex $(\Delta S^{\ddagger}_{kcat})$, resulting in a low entropy barrier or even favourable activation entropy. Despite the wide-spread use of this explanation in textbooks, the importance of such enzymatic ground state destabilization for catalysis has been heavily debated.[54-56] For instance, Åqvist and colleagues used structural simulations obtained at different temperatures to uncover that cytidine deaminase did not operate by the Circe effect.[57,58] By comparing the enzyme-catalysed and non-catalysed reactions, the authors identified that the appearance of a zwitterionic intermediate in the modelled step-wise aqueous reaction was responsible for an entropic penalty ($T \times \Delta S^{\ddagger}$ of -11.9 kcal mol^{-1} at 298K) and refer to the entropic strain of solvation as cause for the observed penalty. The enzyme-catalysed step-wise reaction in contrast was shown to proceed with $T \times \Delta S^{\ddagger}$ close to zero (+0.7 kcal mol^{-1} at 298K). The authors emphasise that the loss of entropy penalty in the enzyme was not due to substrate-fixation (Circe-effect) but rather due to the fact that the enzyme mechanism proceeded *via* a different reaction mechanism that avoided the presence of the zwitterionic species.[46] Likewise, un-catalysed ester aminolysis is associated with a strong entropy penalty which can be explained from solvation of a zwitterionic transition state.[46] In ribosome-catalysed ester aminolysis (peptide bond formation) the entropic cost of this solvent reorganisation was reduced due to a pre-organised hydrogen bonding network.[46,59] Moreover, charge delocalisation in the transition state (and thus reduced entropic cost of solvation) was also identified as a likely contributor to a high positive $T \times \Delta S^{\ddagger}$ in EF-Tu GTPase (+7 kcal mol^{-1} at physiological temperatures).[46,51] In summary, Åqvist *et al.* conclude that the bypassing and mitigation of charges in transition states and intermediates and associated desolvation is responsible for low entropy barriers or even favourable entropy in the studied enzymes and that solvent entropy might therefore be particularly crucial for processes involving charge transfer and separation.[46] The importance of solvent contribution to favourable activation entropy was further indirectly observed for the bacterial thermophilic triterpene cyclase SHC.[50] **Paper II** investigates the role of solvent-entropy in the catalytic mechanism of hOSC – an essential enzyme in human steroid metabolism.

2.5 Enzyme evolution

2.5.1 Enzyme promiscuity

In the mid-1970s Yčas and Jensen proposed that modern-day specialised enzymes have developed from low-activity generalist ancestors.[60,61] By now, it is well established that also most extant specialist enzymes harbour additional low-level promiscuous activities, sometimes involving unrelated substrates or mechanisms.[62-65] In fact, it is estimated that an average enzyme may harbour 10 promiscuous activities.[65] The prevalence and extent of promiscuity is exemplified by a systematic study of the haloalkanoate dehalogenase superfamily in which 217 enzyme family members accepted a median of 15.5 substrates (observed maximum substrate acceptance was 143 of 167 tested substrates).[66] The structural origin of latent promiscuity has been identified and characterised in detail for many different enzymes.[67] Tawfik and coworkers pointed out that "floppy" enzyme active sites that dynamically sample multiple sub-states of altering conformation, protonation state *etc.* are likely to promote promiscuous activities that make use of less abundant sub-states,[68] explaining a general correlation between enzyme flexibility and promiscuity.[62] The presence of promiscuity may complicate the description of novel enzymes due to incomplete determination of their reaction spectrum. For instance, serum paraoxonase-1 was erroneously identified as organophosphate hydrolase until its major activity as lactonase was uncovered.[69] For this reason, inferring enzyme function solely based on homology, as is frequently performed in databases, entails a certain risk of mis-classification.[70]

2.5.2 Evolution from promiscuous side-function to major enzyme activity

Under a given selection pressure, an enzyme has the potential to evolve into a highly proficient catalyst for initially promiscuous side-activities *via* accumulation and selection of mutations, insertions and deletions or recombination with other genes.[71] The potential obstacle of satisfying two separate and possibly contradictory evolutionary pressures – one towards the original function and one towards the novel function – can be alleviated by gene duplication, a mechanism that accelerates enzyme evolution by maintaining the original primary enzyme activity in one duplicate, while allowing a novel promiscuous function to evolve in parallel in the duplicated gene.[72,73]

As discussed in **chapter 1.4**, a large share (ca. 70-80%) of possible mutations that a protein can undergo are likely to be deleterious for its structure and original function,[18,74] which would lead to the elimination of such variants at early stages of evolution. Therefore, many mutations require additional compensatory or facilitating adaptive mutations that, taken by themselves, may seem neutral or irrelevant to the enzyme function at first. Interdependency between different positions of the same protein is referred to as epistasis,[71] where positive epistasis describes the benefit of two mutations that occur in conjunction with each other being greater than the sum of their individual benefits. Since mutations outside of the active site are less prone to impair enzyme activity and thus be eliminated in the early course of evolution, adaptive mutations are more likely to be found in the periphery of the protein and are often involved in epistatic networks. Tawfik and colleagues discuss how second- and third-shell adaptive interactions shift the distribution of active site sub-states that are responsible for promiscuity,[75] rather than that they "create something from nothing".[68] The structural and functional implications of residues outside of the active site are often challenging to predict. This is illustrated by an example in which mutations of two non-active site residues in a TEM-1 β-lactamase individually modified the active site conformation resulting in higher enzyme activity. The effected shifts in the active site were however mutually exclusive, resulting in negative epistasis of these two mutations when combined.[76] Similarly, the mutation of surface residues in **paper II** resulted in an altered ligand binding pose and thermodynamic activation parameters. Finally, surface interactions with other proteins in a crowded cell milieu that buffer mutations along an evolutionary trajectory may further contribute to evolutionary rates being greater on protein surfaces than in core residues.[77-79] Alas, such effects may not be captured by expression of enzymes in standard expression hosts typically used in experimental evolutionary studies.

2.5.3 The mediocrity of evolved enzymes

The most efficient enzymes that are known to date are "diffusion-limited", which means that their substrate binding and conversion is so rapid that virtually every enzyme-substrate collision results in conversion to product and the diffusion of substrates or products into or out of the active site become rate-limiting. These enzymes have evolved to maximise k_{cat}/K_M values to the physically possible upper limit (10^9 - 10^{10} s^{-1} M^{-1}).[68] One example is constituted by carbonic anhydrase (k_{cat}/K_M of 10^{10} s^{-1} M^{-1} or higher), an enzyme that regulates the pH in blood.[80,81] However, examples of such "super-enzymes" are very rare and the analysis of kinetic parameters across multiple enzyme families has revealed that most enzymes did not evolve to maximise catalytic efficiency. In fact, the median k_{cat}/K_M value for all enzymes in the Braunschweig Enzyme database (BRENDA) was found to be around 10^5 s^{-1} M^{-1}, indicating that on average enzymes operate at catalytic efficiency of several orders of magnitude below the diffusion limit.[82] The average enzyme therefore undergoes ca. 1000 times more futile than successful collisions with its substrate(s).[68]

Mediocre enzyme efficiency likely stems from both evolutionary and physicochemical constraints.[82] Regarding the first constraint, Newton *et al.*, summarise that "evolution is not the pursuit of perfect enzymes".[70] Evolutionary selection occurs on the organismal and not on the molecular level and the correlation of organismal fitness with enzyme kinetic parameters is not always straightforward.[70,83] This complex correlation is exemplified by a study in which an initial small 7% change in k_{cat}/K_M for an essential synthetic enzyme reaction drastically impacted organism growth rate, whereas the effect of a subsequent 200% increase in k_{cat}/K_M was almost negligible.[70,84] Likewise, the modulation of kinetic parameters of enzymes operating at high activity often barely impacts organism survival.[41,85] One potential explanation for this may simply lie in the fact that organismal survival does not benefit from an improvement in an enzyme's catalytic efficiency beyond a certain activity threshold.[82] In the absence of further selection pressure, the evolutionary trajectory of k_{cat}/K_M values comes to a halt. In this context it is interesting to mention, that average k_{cat} and k_{cat}/K_M values of enzymes originating from central carbon and energy metabolism exceed those of enzymes from secondary metabolism by ca. 30 and 6-fold, respectively. Ron Milo and coworkers conclude that enzymes of central carbon metabolism are subject to higher evolutionary pressure to maximise $k_{\mathrm{cat}}/K_{\mathrm{M}}$ in order to sustain high metabolic flux through a pathway without requiring an organism to produce large amounts of protein. The same evolutionary pressure does not apply to secondary metabolic enzymes that sustain much lower flux rates.[82] This hypothesis also highlights the importance of scaling kinetic parameters by the amount of available free enzyme when assessing effective enzyme activity in the context of the cell.[70,83] The maximisation of kinetic parameters is further compromised by the co-existence of multiple and potentially contradictory and unknown evolutionary pressures on the same enzyme. Besides activity and catalytic efficiency, multiple parameters are honed during an enzyme's evolution, including *e.g.* protein homeostasis, stability, solubility and oligomerisation, cofactor availability and recycling, regulation by other proteins and metabolites or the association with other enzymes to maximise metabolite channelling.[86,87] The evolution of regulatory mechanisms, for instance, represents another likely reason for the low observed k_{cat} and $k_{\mathrm{cat}}/K_{\mathrm{M}}$ values in enzymes originating from secondary metabolism.[82]

Due to these additional constraints on enzymes evolving *in vivo* the relationship between organismal "fitness" and kinetic parameters such as k_{cat} and $k_{\mathrm{cat}}/K_{\mathrm{M}}$ values is not necessarily straightforward, complicating evolutionary interpretation of the latter. The strong dependency of k_{cat} and $k_{\mathrm{cat}}/K_{\mathrm{M}}$ values on the specific assay conditions, which may not necessarily represent realistic physiological conditions,[88] further results in incongruity of measured *in vitro* and *in vivo* kinetic parameters.[89] Finally, the assumptions underlying the Michaelis-Menten equation (see **chapter 2.2.1** for details) may not necessarily apply to many enzymes for which kinetic parameters are reported in databases.[68] In addition to the discussed evolutionary constraints, Bar-Even *et al.* suggest that K_{M} and consequently $k_{\mathrm{cat}}/K_{\mathrm{M}}$ values may be subject to some degree of physicochemical constraints.[82] They observed that reaction mechanisms involving a higher number of substrates were associated with lower K_{M} values for each substrate. Likewise, for small substrates up to 350 g mol^{-1}, K_{M} values generally decreased with increasing molecular mass and hydrophobicity of the substrate. This correlation was however not equally strong across different EC-classes and applied more stringently to reactions involving simpler catalytic mechanisms. Moreover, k_{cat} values did not show correlation with these substrate properties.

3 – Protein and Enzyme Engineering

3.1 Protein and enzyme engineering objectives

The high-level reasons for engineering proteins are as diverse as optimising specific enzymes for industrial catalysis, developing enzyme catalysts for novel reactions and pathways, expanding the substrate scope of an existing enzyme, enhancing the cost-benefit ratio of industrial protein synthesis, designing artificial regulatory mechanisms for pharmaceutical purposes and many more. Besides these application-oriented reasons, protein engineering may also be performed for academic purposes to delineate principles and trends in protein structure, function and evolution, to study the role of individual residues in a particular enzyme mechanism or to benchmark novel engineering techniques. Another important reason for engineering proteins is to expand the range of accessible methods with which to study a particular protein. Common engineering objectives generally comprise increased protein stability (see **chapter 1.4**), solubility, production yields, (stereo-)selectivity, activity with desired substrates or the adaptation of a protein to a novel environment (*e.g.* in a reactor or in presence of solvents or salts).

The vast majority of experimental protein engineering approaches introduce amino acid exchanges into a protein by directly modifying the gene encoding it, producing the variant proteins and then comparing a particular phenotypic property of the variants to the reference protein either *in vitro* or *in vivo* to assess the effect of the mutations. In contrast, some engineering techniques also focus on non-genetic modifications, such as stimulation of alternative splicing *in vivo,* chemical modifications of proteins or crosslinking with protein binders. This thesis focuses on protein engineering techniques that involve modifications of a protein's amino acid sequence, which is why non-genetic protein engineering, albeit an interesting field of research, is not further discussed.

3.2 Implications of enzyme evolutionary principles for enzyme engineering

Since most enzymes operate at intermediate turnover numbers and catalytic efficiencies (see **chapter 2.5.3**) and are only marginally stable (see **chapter 1.3**), many enzymes carry the potential for further optimisation by enzyme engineering. As Jayaraman *et al.* point out, the evolution of novel biochemical function often starts with an increased expression of an existing promiscuous enzyme.[71,90]Analogously, the improvement of protein stability and solubility towards enhanced protein expression represents an important starting point for further engineering of desired enzyme function. Existing promiscuous enzymes can often be modified to catalyse novel reactions *e.g.* with substrates that don't occur naturally. In some cases, enzymes have even been engineered to catalyse reactions that are not found in Nature, such as carbon-silicon bond formation.[91] Typically, the strongest modification of enzyme features (*e.g.* catalytic efficiency or protein stability) occurs early in evolutionary trajectories, whereas the improvement per mutation levels off with the accumulation of more adaptive mutations, a concept that is referred to as "diminishing returns".[71,92,93] This principle from natural evolution also applies in laboratory evolution of enzymes towards desired properties.

The issues that complicate the evolutionary interpretation of kinetic parameters (as discussed in **chapter 2.5.3**) or protein stability (as discussed in **chapter 1.3**) can also present challenges for experimental protein and enzyme engineering. For instance, the buffering effect of host-specific chaperones on mutations that would otherwise be detrimental to the structural integrity of a protein may be absent in typical laboratory expression hosts, such as *Escherichia coli*, *Saccharomyces cerevisiae* or Chinese Ovarian Hamster cell lines, resulting in misfolded and thus inactive variants. Another major factor complicating protein and enzyme engineering is epistasis (discussed in **chapter 2.5.2**). Both rational and agnostic protein engineering techniques (**chapter 3.3.1** and **chapter 3.3.2**) have limitations that make it challenging to accurately account for epistasis. The long-range effects of some out-of-active site mutations on the conformational ensemble of the active site are difficult to predict or even more so to design. This is additionally complicated by the fact that epistatic interactions may comprise several interdependent residues that form an intricate network. As Tawfik and colleagues point out, active sites of artificially designed enzymes are often highly dynamic and populate multiple sub-states, several of which are unproductive and the introduction of additional adaptive mutations is required to shift the conformational ensemble towards the most productive sub-state.[68,94,95] Epistasis further slows down evolutionary rates[71] (*e.g.* enabling mutations need to occur first to allow otherwise deleterious mutation to be fixated) and the likelihood that enabling and deleterious mutations occur in a sensible chronological order within directed evolution approaches may be limited. Another complicating factor discussed in **chapter 2.5.3** is the dependence of kinetic parameters on assay conditions. By quantifying the success of an experimental protein optimisation using empirical methods, methodological biases may unwittingly influence the assessment of engineering success ("you get what you screen for"). An enzyme may for instance be iteratively evolved towards high turnover numbers using a particular *in vitro* assay, inadvertently selecting for tolerance towards the assay buffer components.[88] When using the optimised enzyme in a different buffer or *in vivo*, its activity may not be as superior as expected.

3.3 Protein and enzyme engineering approaches

In general, enzyme engineering approaches can be classified along a gradient of techniques that range from entirely rational approaches on one end to entirely "random" approaches through a spectrum of semi-rational approaches in between. While the two approaches are discussed separately below, it is important to mention that many engineering endeavours combine elements of both approaches to make full use of their respective strengths.

3.3.1 Rational engineering

Fully rational engineering of protein properties, such as thermostability or catalytic activity, requires extensive knowledge about the protein and its structure-function relationship, *e.g.* the identity of residues in the active site or those that communicate an allosteric regulatory signal. It also requires general knowledge about the expected effects of particular types of mutations, such as the introduction of buried charges. While some protein structures may be challenging to solve by structural biology techniques, advances in cryogenic electron microscopy (cryo-EM) and the recent advent of the reasonably accurate machine-learning guided protein structure prediction algorithm AlphaFold[96] are promising developments that may mitigate this limitation in the future. Nevertheless, not only structural but also mechanistic knowledge, which may be more challenging and resource-intensive to obtain, is required for this type of engineering. In the conceptually simplest case of rational engineering, a single residue exchange is introduced into the protein and its effect is subsequently evaluated by experiment. An example of this is represented by the targeted exchange of a tryptophan residue for an acidic aspartate in a Precambrian β-lactamase, which successfully introduced a Kemp eliminase activity into the enzyme at a site distant to the β-lactamase active site.[97] Rational enzyme engineering approaches can incorporate features of both positive and negative design, although positive design is more common. Negative design in this context describes the strategy of purposefully destabilising non-productive catalytic sub-states of a conformational ensemble.[68] *De novo* enzyme design also falls under the umbrella of rational protein engineering. In this approach, biochemical functionality is not adapted from a pre-existing major or promiscuous enzyme activity. Instead, optimised transition states are designed from scratch in a minimalist active site ("theozyme") and then grafted into a protein scaffold that is suitable to host them.[98-101] As of now, many *de novo* designed enzymes don't reach catalytic efficiency of natural enzymes without further optimization,[94,102] highlighting that our understanding of catalysis and how it is affected by residues outside the active site are still amendable.[103] An alternative to this bottom-up *de novo* design approach is based on exploiting the catalytic proficiency of cofactors by incorporating them into *de novo* designed scaffolds, such as *e.g.* heme-based *de novo* enzymes or metalloenzymes.[104-106]

3.3.2 Directed evolution (Agnostic engineering)

On the other end of the spectrum, proteins can be adapted towards a desired phenotype in an approach that mimics Darwinian species evolution, *i.e.* several iterations of creating a diverse set of homologous sequences that slightly vary from the reference sequence, followed by selection for variants with desired phenotypes. In the most pointed manifestation of this directed evolution approach, the positions, amino acid identities as well as amount of introduced exchanges are chosen stochastically, which is experimentally achieved using error-prone PCR techniques. In principle, no knowledge about the reference protein is required other than its sequence, which is in stark contrast to the detailed structural and mechanistic knowledge required for rational protein engineering. The approach is therefore referred to as "agnostic" (meaning unaware or independent of the target enzyme's structure-activity relationship) in the context of this thesis. However, other limitations may curb the usage of directed evolution techniques. Considering that a vast majority of potential mutations are expected to be detrimental or neutral, the diversity of functional proteins generated in a single step, which is limited by the error rate of the polymerase and transformation efficiency of common laboratory protein production hosts, may not be sufficient to sample relevant beneficial mutations. Therefore, fully agnostic directed evolution studies typically require multiple rounds of diversification, screening and selection, which can be resource-intensive and impede reproducibility of evolutionary trajectories. Moreover, the phenotypic outcome of the mutations need to be amenable to high throughput screening, *e.g.* by coupling a particular enzyme trait to a host organism's survival (such as β-lactamases conferring antibiotic resistance) or by generation of a colorimetric or fluorescent read-out for quantification of protein function. Despite such limitations, there are countless studies that have successfully employed directed evolution to improve major, promiscuous or *de novo* designed functions of enzymes.[102,107,108] Moreover, multiple semi-rational approaches have been developed to focus directed evolution libraries on more narrow function-enriched sequence space. This can for instance be achieved by smart recombination of functional gene segments,[109,110] confining random mutagenesis to a relevant protein domain or by site-saturation mutagenesis of positions that are recognised to be relevant by structure or sequence analysis (such as iterative and combinatorial active site saturation testing).[111,112]

3.4 Engineering of protein stability

3.4.1 Structure-based approaches

Enhancing protein stability is a very common protein engineering objective and as for other objectives, rational, semi-rational and agnostic approaches for improving stability are available. In directed evolution approaches, thermostability can be selected for by exposing library variants to heat prior to high-throughput screening of protein function to identify variants that maintain their functional fold at high temperature.[113] One example is represented by the increase of a terpene cyclase melting temperature by 12 °C using two rounds of directed evolution with a fluorescent activity reporter probe.[114] Moreover, protease susceptibility of surface-displayed libraries has been proposed as a high throughput assay to probe protein stability,[115] which may be useful in the context of directed evolution.

As described in **chapter 1.2** multiple structural factors contribute to the relative stabilisation of a protein's ground state or destabilisation of unfolded states, resulting in the protein assuming its native fold in a particular environment. Therefore, the engineering of protein thermostability can often be successfully performed using structure-guided rational approaches, such as enhancing core packing with large hydrophobic residues, introduction of surface-charges, rational design of hydrogen bonding networks or the rigidification of mobile structural elements using prolines and disulphide bonds.[14,116,117] A timely example is the structure-based stabilisation of the major surface antigen of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which in its wild type form is highly unstable and challenging to work with in the laboratory (see **chapter 4.1.1**).[118,119] The energy scoring function used by Rosetta software[120] is fairly accurate at predicting protein stability based on structural denominators and some proteins that were designed *de novo* using this software suite exhibit remarkable stability,[121-123] with an early example being the designed Top7 fold.[12] However, redesign of an existing mesophilic protein's stability can be much more challenging, since naturally evolved epistatic effects may not be immediately apparent and can result in loss of protein folding or function when introducing purportedly stabilizing mutations.

3.4.2 Sequence-based approaches

In contrast to structure-based protein stabilisation, which is dependent on a high-resolution structure, sequence-based approaches rely on phylogenetic information rather than structural input. Consensus sequence design describes the concept of selecting the most common residue at any given position based on a multiple sequence alignment of homologous sequences and has frequently resulted in highly stable protein variants.[124,125] The rationale behind the method is the assumption that the conservation of a particular amino acid identity at a given position across multiple protein orthologues likely indicates that this residue is relevant for protein stability and function and has thus been positively selected for, since it would otherwise have been cleared by genetic drift. However, several studies have observed that proteins generated by the consensus sequence approach show compromised activity or do not optimally express or fold.[126,127] This may be partly due to the reason that epistatic networks in consensus sequences are disturbed, which can impact an enzyme's dynamics.[128,129] Moreover, one major methodological drawback with the consensus sequence design approach is the risk of oversampling particular phyla for which many sequences are available in databases, which biases the alignment.[128,130] Another issue identified with the generation of consensus sequences lies in the lack of reproducibility in deriving the consensus sequence, since subjective input regarding frequency cut-offs at positions with multiple residue options and the choice of remaining non-conserved positions can have a large impact on consensus sequence generation.[128]

It is important to mention that methods that combine structural considerations with phylogeny guidance, such as PROSS[131] or FireProt[132] have shown great success. The emergence of AlphaFold as openly accessible tool for reliably predicting protein structure,[96] is likely to make proteins for which no structure is resolved more amenable to such hybrid approaches.

3.4.3 Ancestral Sequence Reconstruction

Ancestral sequence reconstruction (ASR) is a method used to study proteins that represent shared ancestors of a group of assumed descendant extant sequences. The idea of reconstructing and experimentally studying ancestral proteins was originally proposed by Pauling and Zuckerkandl in 1963[133] and the first reconstructed gene was experimentally characterised in 1995.[134] While the technique was mostly employed to study the evolutionary history of certain genes in the beginning, it has recently seen a surge in interest as applied protein engineering technique.

ASR represents a sequence-based method that not only takes into account a multiple sequence alignment like in consensus sequence design, but also the phylogenetic relationship between the included sequences, thereby accounting for epistatic interactions and co-evolution of amino acid residues. In a typical ASR workflow, a phylogenetic tree is constructed based on available sequences as well as statistical and evolutionary models. Different approaches can be employed to construct the tree (discussed in more detail in **Chapter 5.1.1**), followed by sequence inference at nodes, which represent so-called ancestors. Since information harboured by the entire phylogenetic tree is used for reconstructing sequences at nodes, oversampling of closely related sequences in the multiple sequence alignment has less impact on the overall inferred sequences.

One of the most commonly observed features across multiple ASR studies and reviews is that proteins corresponding to reconstructed ancestral sequences (herein referred to as ancestral proteins) often are more thermostable than extant (meaning present-day) homologues, yet often show similar activity levels to extant proteins.[130,135-139] For instance, reconstructed ancestral Elongation Factor Tu showed a similar dynamic profile to an extant thermophilic homologue, indicating that epistatic networks required for protein function appeared to be captured in reconstructed ancestral proteins, which was not the case in the consensus sequence control.[129] Thomson *et al.* point out that thermostability as a phenotypic outcome in reconstructed ancestors was found to be remarkably robust across multiple studies that have analysed the impact of different method-related biases on the reconstruction.[130,140] The utility of reconstructed, stable ancestors that behave similarly to extant homologues is particularly relevant for protein pharmaceuticals, as shown for coagulation factor VIII.[141] A reconstructed mammalian ancestral factor VIII shared a large amount of predicted B-cell epitopes with the extant human protein, thus reducing the potential for unwanted immunogenic responses to replacement therapy. Simultaneously, cross-reactivity of the ancestral proteins with inhibitory antibodies in plasma of patients with haemophilia A was reduced compared to using human factor VIII.

The structural and evolutionary reasons for increased stability in many ancestral proteins are frequently discussed. Several studies have observed a correlation between particular structural traits in individual ancestral proteins and ancestral thermostability.[130] Examples include increased hydrophobicity in ancestral EF-Tu proteins,[142] salt bridges formed in ancestral and thermostable extant adenylate kinases[143] or the reduction of nonpolar accessible surface area and increased inter-subunit salt bridges and hydrogen bonds in ancestral nucleoside diphosphate kinase.[130,144] In contrast, other studies did not observe clear structural changes between ancestors and extant proteins that would allow to pinpoint the role of individual residues or types of interactions.[130,145] Apart from the structural aspects underlying ancestral thermostability, several hypotheses regarding its evolutionary origin exist. One common hypothesis is based on the observation that, during certain stretches of time, ancient Earth was considerably warmer than today,[140,146] necessitating a general thermophile-like behaviour in proteins, which then lost thermostability due to absence of evolutionary pressure throughout mesophilic speciation as Earth cooled down. This hypothesis is supported by the fact that many ancient Precambrian enzymes display increases in melting temperatures (20 – 30 °C) that are challenging to achieve with other approaches such as rational structural engineering.[136,138,145,147] However, increases in protein stability can sometimes also be observed for younger ancestors.[130] Another important factor that may contribute to ancestral protein stability lies in the fact that they may have been required to fold in the absence of highly specialised and proficient protein folding and homeostasis machineries.[128,147] As discussed in **Chapter 1.3**, such folding-assisting systems can buffer the accumulation of destabilising mutations and therefore likely contributed to the evolution of marginal stability in many mesophilic proteins. This hypothesis of unassisted ancestral folding seems to be supported by the observation that many ancestral proteins are observed to be more soluble and obtained more easily in higher yields in both prokaryotic and mammalian expression systems,[139] indicating that they misfold less frequently. For instance, a higher expression per transcript was observed for ancestral factor VIII production in mammalian cells,[141] suggesting a reduction in futile folding cycles leading to protein degradation. Finally, a correlation between protein stability and protein evolvability has been observed.[17,148] This is partly due to the fact that stable proteins are able to maintain their fold (and thus function), even if several destabilising mutations occur. This is exemplified by the observation that an extant Rubisco species did not tolerate mutations well when subjected to a directed evolution regime (76 % inactive clones), whereas a reconstructed ancestral Rubisco species only exhibited 26 % inactive clones, when subjected to the same treatment.[149] It may therefore also be plausible that emerging stable protein variants more frequently continue evolving towards new functions and thus become ancestral proteins by definition. These three hypotheses behind ancestral protein stability – Earth temperature, unassisted protein folding and evolvability of stable proteins – do not exclude one another.

The described properties of reconstructed ancestral proteins, such as stability, solubility and functionality, make ASR a suitable protein engineering technique both as a standalone approach as well as in conjunction with other agnostic or rational engineering approaches. The number of cases in which ASR has been used to engineer proteins has continuously increased in the past decade.[130,139,147] Some examples of enzymes engineered for improved thermostability comprise cytochrome P450s,[138] haloalkane dehalogenases,[137] PETase[150] or an amino acid binding protein for development of an L-arginine biosensor.[151] Similarly, ASR has been utilised to obtain more soluble proteins, which typically manifests itself in higher expression yields in heterologous production hosts, as shown for example for endo-β-glucanase,[152] or the ancestral factor VIII discussed above.[141] These properties also make ancestral enzymes highly suitable for structural studies and many crystal structures of ancestral proteins have been reported.[129,136,153,154] Several studies have used ancestral proteins as starting point for further engineering by directed evolution, such as DNA-shuffling of ancestral clouds of cytochrome P450s (a collection of several plausible ancestral sequences representing the same node),[138] directed evolution of Rubisco[149] and evolution of a reconstructed phytase as additive for animal feeds.[155] Other studies have used ancestral proteins as robust scaffolds for rational (re)design, such as the design of a *de novo* Kemp eliminase activity in ancestral β-lactamases,[97] redesign of binding sites for fluorescence-based biosensors[151] or tailoring substrate activity in L-amino acid oxidases[156]. In the work presented in this thesis, proteins reconstructed by ASR were used as robust scaffolds for further engineering, namely grafting of domains with important surface epitopes from extant viruses into ancestral scaffold antigens (**paper I**), as well as engineering selectivity in a stable ancestral terpene cyclase based on structural considerations (**paper III**).

3.5 Protein and enzyme engineering in biomedical applications

The tools and principles that have been employed in the field of biocatalytic research in principle also extend to biomedical applications. Among the properties of protein therapeutics that can be improved by engineering, stability is particularly relevant. Protein stability has implications for therapeutic dose effects *in vivo* (increased bioavailability and half-life, reduced susceptibility to proteases), drug administration (decreased requirement for frequent administrations, new routes of administration, easier handling during administration) as well as logistics and cost of manufacturing, distribution and storage (production yield and purity, shelf-life).[157] The *in vivo* benefits of enhanced protein stability also apply to proteins that are administered indirectly in the form of ribonucleic acids (RNA). This particular field has recently experienced a surge in interest due to the rapid development and approval of first-generation mRNA vaccines against SARS-CoV-2 during the Covid-19 pandemic.[158] In the broad sense, protein and enzyme engineering can be applied in biomedical contexts for improving (i) antibody therapeutics (ii) synthetic biology circuits (iii) non-antibody protein therapeutics and (iv) immunogens for next-generation vaccines. Iterative improvement of target affinity in antibodies (or other binders) for treatment of *e.g.* cancer or auto-immune diseases is frequently performed by directed evolution in iterative rounds of diversification and affinity enrichment.[159] Synthetic biology circuits describe a highly heterogeneous group and involve applications as diverse as programmable chimeric antigen receptor (CAR-)T cells for personalised cancer therapy, engineered transcription factors for gene therapy, optogenetic switches and many more, as reviewed by Bojar and Fussenegger.[160] Engineering of non-antibody protein therapeutics and immunogens will be introduced in more detail in the following two sections.

3.5.1 Non-antibody protein therapeutics

Many human diseases are caused by the total lack, reduced availability or compromised biological activity of particular endogenous proteins. Protein replacement therapy describes the administration of modified or non-modified proteins to supplement the compromised protein function in patients. The first example of successful protein replacement therapy was the administration of recombinant human insulin to diabetes patients, which was approved by the U.S. Food and Drug Administration (FDA) in 1982. Recombinant enzymes can also serve as protein drugs, which is particularly relevant in the treatment of metabolic disorders. Initially, such efforts focused on substituting deficient proteins with homologues from other species or recombinant human proteins. In recent decades, protein engineering has been applied to further improve the stability, targeted delivery and immunogenicity of protein and enzyme therapeutics, as reviewed by Dellas *et al.*[157] An example of agnostic protein engineering applied to a therapeutic enzyme is the directed evolution of phenylalanine ammonia lyase towards improved protease resistance for treatment of phenylketonuria.[161] Examples of rational engineering include the computational redesign of an endopeptidase towards high acid-tolerance and gliadin activity for treatment of coeliac disease and the rational engineering of a microbial uricase towards high activity and low immunogenicity (*via* engineering of biochemical modification sites) for the treatment of gout.[162,163] Ancestral sequence reconstruction has recently been applied to several protein drugs, including the stabilisation of Factor VIII in coagulation therapy,[141] improvement of immunogenicity in stable ancestral uricases for gout treatment,[164] generation of highly active ancestral iduronate-2-sulfatase for treatment of Hunter syndrome as well as stable Phe/Tyr-ammonia lyases for treatment of phenylketonuria.[165,166] Although protein replacement therapy is not directly discussed in this thesis, these examples underline the potential of ancestral sequence reconstruction as a tool to improve protein drugs and thus the technique's transferability from biocatalytic to medical applications.

3.5.2 Immunogen engineering

Immunogen engineering is a field that aims to improve protective immune responses of vaccines by avoiding B-cell immunodominance and eliciting broadly protective antibodies against quickly mutating viruses.[167] Typical examples of immunodominance are the preferential production of antibodies against the rapidly mutating head-domain, rather than the conserved stem-domain of antigens such as influenza hemagglutinin (HA) or against the post-fusion rather than the infectious pre-fusion conformation of fusion protein antigens such as respiratory syncytial virus (RSV) F protein. Immunogen engineering efforts require a high degree of knowledge about the antigen of interest and the immunogenicity of its different epitopes. In general, rational engineering approaches aim to direct the immune response towards epitopes that are known to elicit broad protection. Immunofocusing by negative design involves removing undesired epitopes by truncation (*e.g.* subunit vaccines against SARS-CoV-2 using only the receptor binding domain (RBD) of the spike protein[168] or stem-only mini-HAs[169]) or masking them with engineered glycosylation sites.[170,171] Positive design approaches are centred on increasing the specific response towards the desired epitope. Examples constitute the selective unmasking of otherwise glycan-shielded epitopes[172] or grafting the desired epitope (*e.g.* a conserved stem HA-epitope) into a homologous antigen (*e.g.* from a non-circulating influenza strain) or an unrelated structural scaffold.[173] The rationale behind this grafting approach is that such chimeric antigens will cause immunogenic competition between the novel epitopes (originating from the scaffold antigen) and the grafted epitope. Due to memory for the latter from previous infection or vaccination, the expansion of the respective memory B-cell repertoires supersedes maturation of naïve B-cells towards the novel epitopes, even if the grafted epitope was previously sub-dominant. Such a chimera approach for influenza HA has been shown to elicit broadly protective antibodies in mouse immunisation studies.[174] In fact, the careful co-ordination of prime and boost immunisations with similar but different immunogens generally appears to be a promising route for the immunofocusing on sub-dominant epitopes.[175-177] Caradonna and Schmidt point out that it is critical to obtain a more profound understanding of how the antigenic distance between the prime and boost-antigens (*i.e.* how many antibody interactions they share) affects the efficacy of heterologous prime-boost regimes.[167] Focusing the immune response on important pre-fusion epitopes by stabilisation of metastable surface glycoproteins in the pre-fusion conformation represents another common immunofocusing strategy. This approach was first applied to influenza HA in the 1990s by rationally designing prolines and artificial disulphide bridges into dynamically rearranging protein domains.[178,179] Since then, this approach has been applied to many other viral antigens,[180-183] including the prominent example of stabilising SARS-CoV-2 spike protein by two or six rationally designed proline residues (S-2P[184] and HexaPro,[118] respectively, see **chapter 4.1.1**). In **paper I**, the SARS-CoV-2 spike protein was stabilised in the closed pre-fusion conformation using ancestral sequence reconstruction in absence of the respective proline residues. In a next step, the RBD domain (Wuhan wild type sequence) was grafted into the ancestral scaffolds, replacing the ancestral RBD domains. Both the ancestral antigens as well as the RBD-grafted ancestral antigens bound antibodies from convalescent patient blood sera, the latter with neutralising capacity. This finding indicates that ancestral sequence reconstruction represents a promising avenue towards the recapitulation of conserved epitopes in immunogens, which has also been suggested based on the broad protection that engineered virus strains displaying ancestral H5N1 influenza antigens elicited in ferrets.[185] In this context, it is interesting to mention that influenza H5 HA antigens designed by an iterative consensus approach were also shown to confer broad protection against diverse influenza H5 isolates.[186-188] Since the ancestral spike proteins incorporate sequence information from several sarbecoviruses, they are likely compatible and adaptable to accommodate RBD-domains of newly emerging SARS-CoV-2 variants and sarbecoviruses. Importantly, ancestral spike antigens may also represent interesting vaccine candidates for boosting sub-dominant B-cell repertoires against conserved epitopes from prior immunisation (either by vaccination or natural infection).

4 – Proteins and Enzymes in this Thesis

4.1 Spike protein

In early 2020 the World Health Organisation declared the spread of Covid-19, the disease caused by the novel coronavirus SARS-CoV-2, a global pandemic. As of February 2023 more than 6.8 million Covid-19 related deaths have been reported worldwide,[189] the estimated number including unreported or undetected cases being between 16.7 and 27.3 million.[190] Several developments have curbed death tolls since 2022, among others the establishment of herd immunity (achieved from previous infection and roll-out of vaccines) as well as concomitant displacement of the original strains by highly contagious but less lethal sub-strains.

4.1.1 Structure and stabilisation of the SARS-CoV-2 spike protein

SARS-Cov-2 is an RNA-virus and carries its cargo in a membrane envelope that is decorated with several membrane proteins. The trimeric spike protein (S protein) which protrudes from the viral surface in high density is the most distinctive surface antigen and gives the virus its name of being crowned (adorned). This heavily glycosylated fusion protein mediates viral phagocytosis and is functionally equivalent to *e.g.* HA in influenza, or the F protein in paramyxoviruses. The S protein mediates viral entry into the cell by binding to the human angiotensin-converting enzyme 2 (ACE2), whereupon it rearranges from its metastable pre-fusion conformation to the stable post-fusion conformation.[191] The S protein is anchored to the viral membrane with an extended stalk-like domain (S2 subunit), which is covered by a globular head-like domain (S1) that faces the target cell surface. The S1 subunit consists of the RBD that engages the ACE2 receptor as well as the N-terminal domain that packs against the RBD of the neighbouring monomer. The N-terminal domain (NTD) shows the lowest sequence conservation within the family of β-coronaviruses.[192] The individual monomers of the RBD-trimer can transiently undergo a hinge-like motion from the tightly packed stable “down-conformation" to an exposed less stable “up-conformation”, in which state they are accessible to the receptor. The same movement has also been identified in structural studies of the closely related β-coronaviruses SARS-CoV and Middle East Respiratory Syndrome coronavirus (MERS-CoV).[193,194] Upon binding to the receptor, the S protein is cleaved into its subunits by a target cell protease at a furin cleavage site, resulting in shedding of the soluble S1 subunit. The virus membrane-bound S2 subunit subsequently folds back onto itself, driven by the multimerisation of two so-called heptad repeat (HR) domains. This folding motion pulls the viral membrane closer towards the host cell membrane, enabling fusion of the two.

The native S protein is intrinsically unstable and challenging to work with due to its low expression yields in mammalian expression systems. McLellan and coworkers (and others) introduced two proline mutations into the S2 fusion domain to prevent its dynamic rearrangement in analogy to the stabilization of other viral glycoproteins described in **chapter 3.5.2**.[184,195] The positions to mutate were directly inferred from sequence alignment to homologous β-coronavirus spike proteins, for which this stabilization strategy had previously been successful.[183,196] This modification is commonly known as the "S-2P" variant of the protein. The Pfizer and Moderna mRNA vaccines both encode the S-2P modification of the spike protein,[197] without which the antigen is highly instable. By further removing the furin cleavage site and fusing a viral trimerisation domain to the C-terminus of the exodomain, the authors succeeded in stabilizing the protein in the pre-fusion conformation and obtained a 3.5 Å cryo-EM structure.[184] This structure revealed that the RBD was predominantly in a one-RBD-up conformation. Despite stabilisation of this variant relative to the native S protein, expression yields and stability of the S-2P variant were still very low. The authors therefore used the obtained structure as basis for further rational engineering, introducing additional prolines, disulphide bonds, cavity-filling mutations and salt bridges.[118] In this rational design approach 100 structure-guided single residue variants of S-2P were generated, expressed and characterised and the most beneficial mutations were subsequently combined to yield the HexaPro variant, which carries four additional proline mutations to S-2P. This final variant exhibited 10-fold higher expression yields, 5 °C increase in melting temperature and tolerated storage at room temperature and multiple freeze-thaw cycles.[118] In a similar approach, Langedijk and coworkers stabilised the S protein in the pre-fusion conformation using rational mutations including several proline substitutions.[119]

In **paper II**, two reconstructed ancestral S proteins were found to reside exclusively in the all RBD-down pre-fusion conformation, as confirmed by 2.6 and 2.8 Å resolution cryo-EM structures.

4.1.2 SARS-CoV-2 vaccines based on the spike protein

The S protein forms the basis of most vaccines that are approved or currently under development.[197] Many anti-S antibodies elicited by previous infection and/or vaccination that bind to the RBD are neutralising, which means that they can block ACE2 binding upon (re-)infection with the virus. Early studies in the pandemic indicated that monoclonal antibodies directed against the RBD of SARS-CoV S protein did not bind to the RBD of SARS-CoV-2 S protein, despite them sharing a high degree of structural homology and sequence similarity (high conservation of ACE2-binding residues).[184] This finding indicated that small differences in the RBD can render the coronavirus S protein immune to established immune responses. Indeed, the RBD sequence is heavily mutated in SARS-CoV-2 sub-strains reported this far which has resulted in a certain degree of immune evasion.[198] Antibodies directed against the S2-domain or non-RBD epitopes on the S1-domain might additionally confer non-neutralising protection by mechanisms such as antibody-dependent cellular cytotoxicity, phagocytosis and complement activation.[199]

4.2 Terpene cyclases

4.2.1 Terpenes and terpenoids

Terpenes and terpenoids comprise a vast and highly diverse and abundant group of natural compounds that are mainly produced in secondary metabolism of plants, several bacteria and fungi and a few insects while animals proceed through intermediate steps of terpene synthesis. Terpenes consist of elongated fused isoprene (C-5) units (Fig. 3A) that are arranged into one or several rings of varying stereochemical complexity (Fig. 3B). Classification of terpenes is performed based on the amount of fused isoprene units into mono- (C-10), sesqui- (C-15), di-(C-20), sester- (C-25), tri-(C-30), sesqua- (C-35) and tetraterpenes (C-40) or polyterpenes (which are less common).

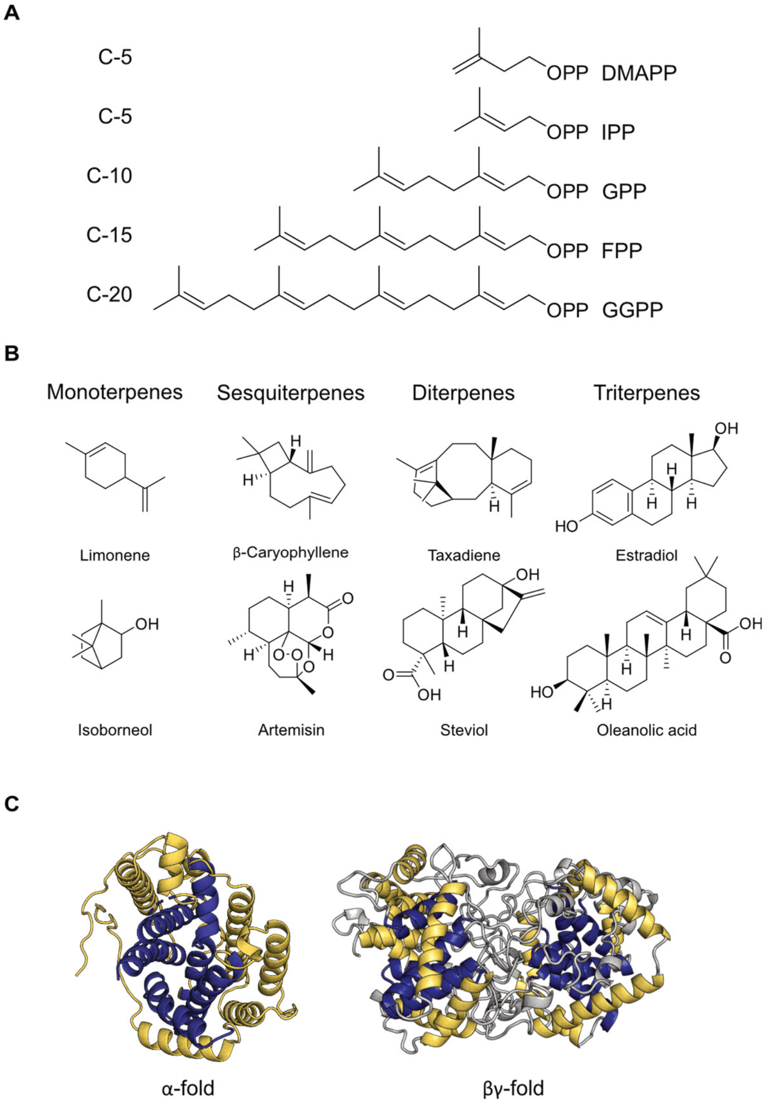


Figure 3. Examples of terpenes and terpene cyclases. (A) In terpene biosynthesis DMAPP and IPP are condensed to form GPP. Condensation with additional molecules of IPP results in elongated terpene substrate chain of varying lengths of which a few examples are shown. **(B)** Examples of cyclic terpenes and terpenoids of different chain lengths. **(C)** Representative structures of the α-fold, which is typical for class I terpene cyclases (SvS-A2 shown) and the βγ-fold, which is typical for class II terpene cyclases (hOSC shown). Inner and outer barrel helices are shown dark blue and gold, respectively.

Terpenes are frequently decorated with functional groups, in which case they are referred to as terpenoids (for simplicity, both terpenes and terpenoids are referred to as *terpenes* in this work). The variation of factors such as terpene length (C-5 to C-40), the number of rings (1-5), size of the different rings, stereochemistry and potential functional groups affords a high number of possible combinations, explaining the vastness of terpenes identified to date (>80.000 compounds)[200]. In fact, terpenes together with flavonoids make up the largest group of natural compounds. Owing to their structural diversity, terpenes have a broad array of different physiological functions and properties.

Terpenes fulfil ecological functions as diverse as plant hormones, insect deterrents, light-harvesting in photosynthesis, attraction of pollinators, membrane regulators, biosynthetic building blocks and have even been suggested to initiate natural cloud seeding.[201] Due to their hydrophobicity and size, many terpenes are volatile and thus responsible for the typical fragrance of coniferous trees and plant essential oils. Their taste and fragrance properties make them popular additives in the food (e.g. limonene, steviol-derivatives) and cosmetic (e.g. menthol) industry. Moreover, terpenes are responsible for many beneficial health effects and are the active ingredient in essential oils that have been used in traditional herbal medicine for centuries. For instance, many terpenes display antibacterial activity (*e.g.* limonene, eucalyptol, menthol, sabinene, carvone and oleanolic acid). Other pharmaceutical properties that attract industrial pharmaceutical interest comprise antiviral (isoborneol, borneol, betulinic acid), antiparasitic (artemisinin), anti-inflammatory (α-Phellandrene, Ginsenoside), analgesic (menthol) and antitumorigenic (perillyl alcohol, geraniol, paclitaxel) activities (some examples shown in Fig. 3B).[202,203] Last but not least, the industrial use of terpenes also expands to pest control, biofuels as well as environmentally friendly solvents and starting points for polymer materials.[204,205]

Terpenes are synthesised via one of two routes in Nature – the mevalonate pathway (which is common in all domains of life) and the methyl-erythritol-phosphate pathway (which is more common in prokaryotes and some plants). Both pathways converge in the production of the C-5 building blocks isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) which form the linear terpene substrate geranyl pyrophosphate (C-10) in a condensation reaction (Fig. 3A) that is catalysed by prenyltransferase enzymes. The fusion of additional C-5 units or fusion of two molecules of geranyl pyrophosphate give rise to linear terpene substrates of varying chain length (Fig. 3A). In the most complex step of terpene biosynthesis the linear substrates are then cyclised to the complex ring structures discussed above. The family of enzymes that perform the last step in terpene biosynthesis is the terpene cyclases family.

4.2.2 Class I and II terpene cyclases

Terpene cyclases (TCs) catalyse the cyclisation of the linear terpene substrates to the plethora of structures discussed above *via* the formation of an initial carbocation followed by several cascade-like carbocationic cyclisation steps, which results in the position of the charge being moved around the ligand. In many TCs, the cyclisation steps are accompanied by methyl- and hydride shifts, further contributing to the structural diversity created by these enzymes. Based on the mechanism by which the initial carbocation is generated, terpene cyclases can be classified into class I and class II terpene cyclases. The former group initiates ionisation by metal cofactor-mediated abstraction of the terminal pyrophosphate group, whereas the latter initiates ionisation by protonation of a terminal double bond or epoxide group mediated by a catalytic aspartic acidic residue.[200] Both classes of enzymes then transiently stabilise subsequent carbocation intermediates mainly through cation-π interactions with aromatic residues that line the active site. Moreover, it is well established that the active site serves as template to pre-fold the substrate into a conformation that facilitates cyclisation.[200]

Terpene cyclases of both classes consist of α-helical folds. A symmetric α-helical bundle of eight helices that harbour a central active site cavity is characteristic for class I TCs (Fig. 3C).[200,206] The so called α-fold may have arisen from gene duplication of a 4-helix bundle protein.[207] In contrast, class II TCs harbour the active site at the interface of two fused α-helical folds, referred to as β- and γ-fold, respectively (Fig. 3C). While these two domains are highly similar, they are not identical and also likely evolved from a gene duplication event.[200] The α-fold is typical of class I TCs, however some class I TC representatives have also been observed in combinations of α, β and γ-domains such as αβ or αβγ and some class II TCs have been described to have an αβγ arrangement. In several of these examples one or two domains represent inactive evolutionary remnants.[200,206,208] Independent of the domain architecture, TCs are observed to assume different oligomeric states, such as monomers (e.g. bacterial *ent*-copalyl diphosphate synthase,[209] class II di-TC with βγ-architecture), dimers (e.g. (+)-bornyl diphosphate synthase,[210] class I mono-TC with αβ-architecture) or hexameric (fusicoccadiene synthase,[211] mixed class I/class II di-TC with αα-architecture).

Class I TCs catalyse the cyclisation reaction of linear terpene substrates by metal cofactor assisted abstraction of the terminal pyrophosphate group. Most frequently, a tri-magnesium ion cluster fulfils this role, although in a few cases transition metal ions such as cobalt and manganese have been described to assume this function.[200] The metal ions are coordinated by two highly conserved metal binding motifs – the so called "aspartate rich DDxx(x)D" and the "NSE (or DTE)"-motifs, which are located on opposing walls on the top of the active site. In contrast, class II TCs initiate carbocation formation by metal-independent general acid catalysis at the interface of the β- and γ-domains, resulting in protonation of a double bond or epoxide group on the linear substrate. Many but not all class II TCs harbour the catalytic aspartate residue in a DxDD motif that is similar yet unrelated to the DDxx(x)D in class I TCs.[200] Besides these well-known signature motifs of TCs, several sub-groups of TCs carry additional sequence motifs that have been associated with different functions. These motifs include *e.g.* the QW-motif in triterpene cyclases, which has multiple occurrences in bacterial squalene hopene cyclase as well as in human oxidosqualene cyclase and is thought to confer additional enzyme stability due to side-chain stacking.[212] Another example is the "RY-dimer" located ca. 80-90 residues downstream of the NSE-motif in bacterial class I terpene cyclases, which was described to assist in binding of the substrate pyrophosphate group.[213,214] Apart from these conserved motifs, TCs generally show little sequence conservation and clusters of plant terpene cyclases that were high in sequence similarity did not correlate with chemical similarity of the formed products.[215]

Many terpene cyclases have been reported to exhibit very low k_{cat} (<< 1 s^{-1}) or k_{cat}/K_M values,[216] which may reflect the secondary metabolite nature of terpenes. Due to the complexity of their reaction mechanism, rational engineering is challenging. Many studies on terpene cyclase engineering focus on improving the flux through metabolic pathways leading up to terpene synthesis,[217-219] whereas fewer studies report direct engineering of thermostability and catalytic properties of terpene cyclases, *e.g.* by site-directed mutagenesis in the active site,[37,220] or directed evolution.[114]

4.2.3 Human oxidosqualene cyclase

Human oxidosqualene cyclase (hOSC) is a class II triterpene cyclase that catalyses the cyclisation of the linear triterpene (*S*)-2,3-oxidosqualene to the tetracyclic terpenoid lanosterol in a single enzymatic step encompassing eight intermediates and nine transition states.[221] The enzyme is also referred to as lanosterol synthase and homologues exist in most animals, yeasts, prokaryotes (squalene-hopene cyclase) and plants (cycloartenol synthase). All steroids are metabolically derived from lanosterol (and cycloartenol in plants), including cholesterol, sex hormones, corticosteroids and the vitamin D precursor, which are all crucial molecules in human health and disease (as well as phytosteroids and saponins in plants, many of which are of industrial interest).

Human oxidosqualene cyclase is a butterfly-shaped enzyme, in which the two wings are made up by similar, yet non-identical domains that harbour the active site at the domain interface (Fig. 3C, right panel).[222] The enzyme represents the archetypical βγ domain architecture of class II TCs. Each domain consists of an αα-barrel with an inner and outer ring of 5 or 6 α-helices, the axis of which extends from the figurative nick of the butterfly wing to the domain interface. The N-terminus of the enzyme consists of a loop and β-sheet that insert between the two domains at the active site. Domain 1 is anchored to the membrane of the endoplasmic reticulum *via* a helix that inserts into the membrane in a perpendicular fashion as well as by surrounding hydrophobic patches on the protein surface. The substrate is abstracted from the membrane through a hydrophobic access tunnel that extends directly to the active site.

The cyclisation cascade is initiated by protonation of the epoxide group on the substrate by a catalytic aspartate residue (D455) in domain 2 that is acidified by two adjacent cysteine residues.[222,223] The cyclisation cascade then proceeds through multiple carbocation intermediates, forming the four sterol rings in a consecutive manner. The second part of the reaction involves several hydride and methyl shifts, which also proceed through carbocation intermediates.[221] The active site is lined with multiple aromatic residues that are critical in stabilising these reactive carbocation intermediates but also serve as a template for a productive substrate binding pose,[221,222,224] which is particularly important given the substrate's large size (C_{30}) and amount of freely rotatable bonds.[200] Finally, the last cation intermediate is deprotonated to yield lanosterol, most likely *via* Y503 that relays the proton to the catalytic base H232, which is located in domain 1.

Since the hOSC reaction is the first dedicated as well as the most complex step in cholesterol biosynthesis, its reaction mechanism has been subject of intensive research ever since the 1960s.[225,226] **Paper II** investigates the role of solvent entropy in the hOSC reaction.

4.2.4 Spiroviolene synthase

Spiroviolene synthase (SvS) from *Streptomyces violens* is a bacterial class I diterpene cyclase that catalyses the cyclization of the linear diterpene geranylgeranyl pyrophosphate (GGPP) to the spirocyclic compound spiroviolene in a single enzymatic step encompassing six intermediates. The enzyme and its product were first described by Rabe *et al.*, who further characterised the intermediates that occur along the reaction trajectory using Nuclear Magnetic Resonance (NMR) studies.[227]

Three ancestors of SvS were reconstructed in 2018, of which the variant SvS-A2 (2 nodes upstream of the extant enzyme in the phylogenetic tree) showed the most interesting properties from an enzyme engineering point of view.[126] This ancestral TC could be obtained at higher yields and showed an increase in thermal stability (melting temperature increased by 13 °C).

In **paper III** the structure activity relationship of SvS-A2 was studied by X-ray crystallography, molecular mechanics simulations of all reaction intermediates and mutagenesis studies and used as a basis to construct a high fidelity homology model of the extant SvS enzyme.

4.3 Fructose-1,6-bisphosphatase

Fructose-1,6-bisphosphatase (FBPase) is a ubiquitous enzyme that catalyses the irreversible hydrolysis of fructose-1,6-bisphosphate (FBP) to fructose-6-phosphate (F6P). Mammalian FBPase was first described in 1943 and despite being initially considered an ancient and simple enzyme from carbohydrate metabolism it has now been recognised that the enzyme performs multiple complex tasks in animal cells.[228,229]

4.3.1 Classes of FBPases

FBPases can be grouped into five classes based on their phylogenetic distance.[230] Several prokaryotes, such as *E. coli* have been described to harbour FBPases from several classes (*e.g.* class I/II or class II/III).

Class I FBPases are the most widespread and occur in eukaryotes and bacterial prokaryotes. Examples of class I FBPase are *e.g.* pig kidney FBPase or the gene product of *fbp* in *E. coli*. In prokaryotes, animals and plants, class I FBPase is located in the cytosol and catalyses FBP hydrolysis in the context of gluconeogenesis reverse to the phosphofructokinase reaction of glycolysis.[231-233] Members of this enzyme family are characterised by the FBPase-fold and bind three divalent metal ions (such as Mg^{2+}, Mn^{2+} or Zn^{2+}) in their active site. Class I FBPases are highly sensitive to inhibition by AMP (IC_{50} values in the low micromolar range) and regulation of FBPase is important to avoid futile cycles of FBP hydrolysis and synthesis in gluconeogenesis and glycolysis, respectively. Plants harbour a second class I FBPase isozyme, additionally to the gluconeogenic cytosolic one, which localises to the chloroplasts.[234] Chloroplast class I FBPases are also metal-dependent but function in the context of the Calvin-Benson-Bassham cycle (Calvin cycle), the most prominent CO_2-assimilating pathway in autotrophic organisms. The chloroplast enzymes share high sequence similarity with cytosol FBPases but are insensitive to AMP inhibition. Additionally, an inserted loop above the active site introduces a conformational switch that is dependent on the redox state of two disulphides, rendering chloroplast FBPases sensitive to activation by thiol-reductants such as dithiothreitol (DTT) instead.[235,236] These two types of class I FBPases – AMP-sensitive and DTT-insensitive cytosolic FBPase and AMP-insensitive and DTT-sensitive chloroplast FBPase are the only FBPases that occur in eukaryotes, whereas prokaryotes can also carry representatives of other classes of FBPase.

Class II FBPases occur in different prokaryotes, such as the gene products of *glpX* and *yggF* in *E. coli*. Representatives of this class contain a FBPase_glpX domain that is similar to the core fold of lithium sensitive phosphatases.[237] They share little sequence identity to class I FBPases (*e.g. E. coli* class I and II FBPases share only 10% sequence identity), despite adopting a similar overall layered αβαβα structure.[233] The gene product of *E. coli glpX* was shown to be insensitive to AMP but sensitive to phosphate.[230] Some bacteria, such as *M. tuberculosis*, only possess class II FBPase to perform gluconeogenesis, whereas the biological role of class II FBPase in bacteria that have both class I and II FBPases is not exactly clear, even though they are also believed to contribute to gluconeogenesis.[233] For instance, the depletion of chromosomal class II FBPase in *E. coli* still sustained growth on gluconeogenic substrates, whereas depletion of class I FBPase did not.

FBPases of class III, IV and V occur in bacteria such as *B. subtilis*, archaea and thermophilic prokaryotes, respectively.[238]

4.3.2 Bifunctional F/SBPase

Several bifunctional FBPases have been described in proteobacteria and cyanobacteria.[237] These enzymes additionally catalyse the hydrolysis of sedoheptulose-1,7-bisphosphate (SBP) and are therefore referred to as F/SBPase. Both reactions occur in the Calvin cycle and are catalysed by two separate enzymes in plants.

The first confirmed bifunctional F/SBPase was described in *Cupriavidus necator,* a facultatively chemolithoautotrophic betaproteobacterium in 1995.[239] While the bifunctional enzyme was required for autotrophic growth, a separate FBPase gene that is not linked to the Calvin cycle operon is assumed to perform gluconeogenesis in this organism.[239,240] Bifunctional

F/SBPase has also been reported in photoautotrophic cyanobacteria such as *Synechocystis sp. PCC6803* (referred to as *Synechocystis* in this work) and *Synechococcus sp. PCC7942* (referred to as *Synechococcus* in this work).[241,242] The dual F/SBPase is essential for survival in cyanobacteria and is predicted to be important for Calvin cycle flux.[242,243] In line with this observation, several studies have observed enhanced photosynthetic activity in high light conditions in microalgae and plants, when overexpressing cyanobacterial F/SBPase.[244-246] No molecular evidence for active site discrimination between FBP and SBP has been observed this far.[247]

A phylogenetic analysis performed in 2012 suggested to group bifunctional F/SBPases into two separate classes in analogy to mono-functional FBPases (**chapter 4.3.1**).[237] Proteobacterial F/SBPases (such as *C. necator* F/SBPase) should be referred to as class I F/SBPases since they contain the typical FBPase domain of class I FBPases. Cyanobacterial F/SBPases (such as *synechocystis* FBPase) should be referred to as class II F/SBPases since they contain the FBPase_glpX domain of class II FBPases. **Paper IV** studied the metabolite regulation of F/SBPase from *C. necator* (class I) and *Synechocystis* (class II) and implications for the flux through the Calvin cycle.

4.3.3 Unique biochemical characteristics of cyanobacterial F/SBPase

Cyanobacterial F/SBPases occupy a unique position within the FBPase family and exhibit low sequence identity to other common FBPases. In a phylogenetic tree they cluster closest with FBPases from sulphate and sulphur-metabolizing bacteria.[237]

*Syn*F/SBPase (class II F/SBPase) assumes a homotetrameric AMP-stabilised T-state in which AMP binds between each two monomers at the central tetramer interface and the substrate binds at the surface of each monomer (see Feng *et al.* and **paper IV** for a figure).[248] The structure of the corresponding R-state is not available. The residues involved in substrate interaction are conserved in *E. coli glpX*, whereas the residues involved in AMP binding are not.

*Syn*F/SBPase is not closely related to class I chloroplast FBPases (<10 % sequence identity) but is assumed to fulfil a similar role, as it was observed that the enzyme was activated by DTT similar to the chloroplast enzymes.[248]

Although a model has been proposed,[248] the molecular mechanism of redox activation is not yet fully established.[247] Surprisingly, *syn*F/SBPase is also inhibited by low micromolar concentrations of AMP in analogy to cytosolic FBPases,[248] whereas chloroplast enzymes and also *E. coli glpX* - the closest non-cyanobacterial homologue - are insensitive to AMP. The *syn*F/SBPase AMP-binding site is distinct from that of cytosolic FBPases, highlighting that the two AMP regulatory mechanisms are likely the product of convergent evolution. *Syn*F/SBPase therefore combines redox regulation and AMP regulation, which are features that are representative of photosynthetic chloroplast FBPases and gluconeogenic cytosolic FBPases, respectively. Since only one type of FBPase was identified in *Synechocystis* thus far, this unique double regulation may indicate that the enzyme fulfils a dual role in both metabolic pathways with respective associated regulatory signatures.[242,248,249] In *Synechococcus* in contrast, both mono and bi-functional FBPases have been observed.[241] A double metabolic role of *syn*F/SBPase is likely associated with a high degree of additional regulatory mechanisms. In **Paper IV** the metabolite regulation of *syn*F/SBPase was studied in detail.

5 – Background on Methods in this Thesis

5.1 Computational techniques

5.1.1 Ancestral Sequence Reconstruction

As described in **chapter 3.4.3** the sequences of ancestral proteins can be computationally inferred to study protein evolution or as protein engineering tool. In the following, the ASR workflow is discussed, comprising selection of extant sequences, multiple sequence alignment (MSA), construction of a phylogenetic tree and inference of ancestral sequences.[130]

In a first step, a set of homologous amino acid or nucleotide sequences is retrieved from sequence databases, including a few sequences from a related outgroup. Note that in this context homology (*i.e.* an evolutionary relationship *via* shared ancestry) is typically assumed from sequence similarity. In **paper I** the basic local alignment search tool (BLAST) algorithm[250] was used to find homologous amino acid sequences of the SARS-CoV-2 S protein, including distantly related sequences of mink coronaviruses as outgroup. While it is generally preferred to include as many homologous sequences from as many taxa as possible,[251] the inclusion of too diverse sequences can increase the uncertainty of the following alignment. Likewise, the inclusion of too similar sequences, such as single residue mutants of a protein does not add much evolutionary information relevant for the entire protein family, while biasing the alignment, due to overweighting of the represented sequence. An example of this is represented by SARS-CoV-2 S protein sequences obtained from different clinical isolates of the same sub-strain in **paper I**. To improve the quality of the included set of sequences, it is advisable to generate an initial multiple sequence alignment (MSA, see next paragraph) with all sequences, followed by manual exclusion of too diverse and similar sequences as well as redundant entries and sequences that contain obvious artefacts from non-curated database entries (*e.g.* frame shift mutations or long insertions or deletions). Also very short sequences that align only to a section of the remaining sequence alignment should be excluded at this step. The removal of such sequences facilitates a more accurate alignment of the remaining sequences[252] and as Thomson *et al.* point out, "the sequence collection is perhaps the one most important factor influencing the quality of the inference".[130] Typically, several iterative cycles of MSA and sequence removal are performed to select the final set of sequences. While this step is necessary to improve the quality of alignment and accuracy of reconstruction, it introduces a user bias into the method and thus reduces reproducibility. Automated tools such as *e.g.* for alignment trimming may aid in counteracting such biases,[253] but cannot completely replace the manual sequence curation step.

Next, the final set of sequences is aligned using a software to perform MSA. The alignment is scored using amino acid substitution matrices, such as the BLOSUM matrix that judge the likelihood of specific amino acid substitutions.[254] The quality of the alignment greatly influences the accuracy of constructed phylogeny and consequently the final inference.[255,256] Alignment accuracy in turn is directly impacted by the quality of input sequences (where erroneous insertions and deletions have been shown to be particularly problematic) and to a lesser extent by the chosen alignment algorithm.[130] In **paper I** the MUSCLE (Multiple Sequence Comparison by Log-Expectation) algorithm[257] was used to construct the MSA.

The MSA is then used as basis for inferring the evolutionary relationship between all extant sequences. In **paper I** a maximum likelihood algorithm (see next paragraph) was used for construction of the phylogenetic tree in IQ-Tree.[258] In maximum likelihood approaches, tree generation is performed in multiple replicates (500-2000) and bootstrap values on each branch are used to assess the proportional replicate frequency and thus reliability of the indicated local topology. It is important to note that the phylogenetic tree in the described workflow is based on a single gene or protein (gene tree) and does not necessarily correspond to the phylogenetic tree that reflects how the respective organisms evolved as a whole (species tree).[130]

Finally, the sequences of proteins in all ancestral nodes are inferred. Three common statistical approaches are used for this purpose, among which maximum likelihood methods are the most widely used. Maximum parsimony approaches minimise the total number of amino acid exchanges required to give rise to the observed phylogeny.[259] This method is one of the oldest used for ASR. However, it is not frequently used to date, given the fact that the method does not account for uneven amino acid substitution likelihoods as well as evolutionary trajectories involving multiple residue exchanges at the same position.[130] Maximum likelihood approaches take into account the different likelihoods of specific amino acid substitutions, which are summarised in empirical substitution matrices (evolutionary models). Such evolutionary models also account for the fact that not all positions evolve at the same rate by sampling from a model-dependent evolutionary rate distribution function. A maximum likelihood algorithm maximises the probability of each amino acid identity in an ancestral node, so that it would give rise to all derived extant amino acid identities in the equivalent position, considering the given tree topology and evolutionary model.[92] Since there is no way to predict which evolutionary model applies to the given phylogeny, an initial fitting step is typically performed to asses which model produces the highest overall likelihood with the given dataset. Many maximum-likelihood ASR software packages such as MEGA[260] (used in **paper I**) incorporate the most common evolutionary models, including *e.g.* the Dayhoff,[261] Le and Gascuel[262] (LG) Whelan and Goldman[263] (WAG, used in **paper I**) or Jones-Taylor-Thornton[264] (JTT) models. Despite the availability of multiple evolutionary models, the difference between these models was found to affect the inference accuracy less strongly than the MSA.[265] In summary, maximum likelihood approaches result in a single sequence estimate for each ancestral node, in which the most likely amino acid identity from the probability distribution at each position is selected As such, the reconstructed sequence constitutes a representation of different possible ancestral states rather than the actual ancestral state. The method does not reflect uncertainties in the tree and evolutionary model but for the purpose of reconstructing ancestral proteins for engineering, the reported accuracy of maximum likelihood approaches may be considered sufficient.[266,267] Finally, Bayesian inference approaches are similar to maximum likelihood approaches but address such uncertainties by simultaneously sampling ancestral amino acid identities and differences in the provided tree and evolutionary model to provide posterior probability distributions at each position. While the results are considered accurate, this third statistical approach is computationally expensive.

It is important to keep in mind that one of the greatest factors confounding the generation of accurate phylogenetic trees (and hence ASR) is horizontal gene transfer (alongside gene duplication/loss and gene fusion/fission events[268]), which is common in bacterial evolution.[269] The occurrence of horizontal gene transfer interrupts the assumed directionality of evolution, according to which all extant proteins derive from a last common ancestor in separate branches that are independent of one another. One way to address such issues is to consolidate the information from the inferred gene tree at hand with a species tree that accounts for more complex evolutionary events, an approach for which several algorithms have been developed.[268]

5.1.2 Caver

Caver 3.0 is a software used to identify and describe tunnels (pathways from the surface to an internal cavity) and channels (pathways from the surface to another area on the surface) in dynamic protein structures.[270,271] Such pathways may serve as transport routes for ligands, ions or solvent to the protein centre and dynamically change with time.[271-275]

The software considers all atoms in a static protein structure as spheres of equal radius (equivalent to the smallest atom radius of the entire structure) in a Voronoi diagram - a 2D geometrical representation of the protein structure, in which centres of polygons correspond to the atom spheres and edges of polygons are set to equally divide the space between two adjacent centres. Tunnels or channels are geometrically determined as lowest-cost paths along vertices and edges in the diagram, connecting a user-defined starting coordinate to the bulk solvent. The cost in this context is defined by the probability of transportation along the identified pathway and is set to prefer short and wide tunnels. After the lowest-cost tunnel is identified, all pathways within a specified distance around it are discarded to avoid redundancy before the next-lowest-cost tunnel is identified. All paths along Voronoi edges that exceed a user-defined lower-limit radius are excluded from the analysis. The pathway is visually represented in 3D by a sequence of spheres along the pathway axis, in which each individual sphere extends to the protein's Van-der-Waals surface, thus showing the maximum size of a spherical probe that could move along the tunnel.

In a first step all pathways are identified in each of a provided ensemble of aligned protein structures (*i.e.* snapshots from Molecular Dynamics (MD) simulations). Next, the identified pathways are clustered across the entire ensemble and ranked by increasing cost of the cluster averaged over the total amount of snapshots ("priority"). Finally, characteristics of individual tunnels, such as their position-dependent radius and bottle-neck radius, length and cost are output per static structure and the frequency and priority of clusters across the ensemble is also indicated, allowing their dynamic analysis.

Modifying the transport through tunnels has been used to engineer *e.g.* substrate specificity and enzyme activity and can be achieved by mutating bottleneck residues to modify tunnel diameter or changing the side-chain properties of tunnel-lining residues.[276-279] In **paper II**, Caver 3.0 was used to identify and modify solvent access tunnels to the active site in a human triterpene cyclase.

5.2 Experimental techniques to assess protein stability and solubility

The proteins in this thesis were expressed recombinantly in *E. coli* BL21(DE3) (**papers II, III and IV**) as well as human embryonic kidney cells (Expi293F) (**paper I**). Stability and solubility of purified proteins were determined using the techniques described in this chapter.

5.2.1 (nano-) Differential Scanning fluorimetry

Differential scanning fluorimetry (DSF) quantifies the thermal unfolding of proteins *via* shifts in fluorescence that occur when the protein structure unwinds. The method allows determination of a melting temperature (T_M) that quantitatively reflects protein stability and can be used to compare protein variants or the same protein at different conditions. Conventional DSF, which was used in **paper I**, is based on the fluorescent signal produced when an organic dye (SyproOrange) binds to hydrophobic residues in the protein that become accessible as a consequence of unfolding.[280,281] The melting temperature (T_M), which is the temperature at which half of the protein population is denatured, is quantified as the midpoint of the fluorescence increase. NanoDSF (**paper I, III and IV**) quantifies the change in intrinsic fluorescence of buried tryptophan (and to a lesser extent tyrosine) residues (emission peak at 330 nm) that occurs as they become exposed to the solvent as a consequence of thermal unfolding (emission peak of surface-exposed tryptophan residues is 350 nm).[280,282] In nanoDSF the T_M is determined as the temperature at which the F350/F330 ratio (ratio of fluorescence at 350 and 300 nm, respectively) changes most rapidly, corresponding to the maximum of its first derivative. The nanoDSF device used in this thesis (Prometheus NT.48, NanoTemper Technologies) further measures thermal aggregation alongside unfolding by detecting the back-reflection (backscattering) of a light beam that passes the sample twice and is influenced by sample turbidity.[283] Such aggregation data are presented for different antigens in **paper I** and for a Calvin cycle enzyme in **paper IV**.

5.2.2 Thermal shift and aggregation assays

A thermal shift assay compares a protein's T_M in absence and presence of a ligand or cofactor, based on the notion that the ligand stabilises the protein.[281,284] Such changes may occur both as a result of thermodynamic stabilisation as well as conformational changes induced by such an interaction. In **paper IV** a thermal shift assay was used to verify binding of several metabolites to two purified proteins. Moreover, a thermal aggregation assay, which is based on the same principle but using backscattering data as a readout instead of fluorescence data (see **Chapter 5.2.1**), was also performed in the same paper.

5.2.3 Dynamic Light Scattering

Dynamic light scattering (DLS) is a technique used to determine the size of particles in suspension based on the measurement of scattered light at a particular angle.[285] More specifically, the technique measures fluctuations in the measured scattering intensity over time that occur as a result of Brownian particle motion. From these fluctuations the particle's hydrodynamic radius (radius of a sphere that would diffuse at the same velocity as the particle) can be calculated from the diffusion coefficient. Since the intensity of angularly scattered light is inversely proportional to the particle radius to the sixth power, this method is particularly sensitive to large particles even when present in small quantities, making the technique suitable for detection of aggregates in protein samples.[285,286] While the resolution of DLS is not sufficient to differentiate individual oligomeric states of a protein, the method allows to measure the heterogeneity of the sample, which is quantified as the polydispersity index (PDI). A low PDI (< 0.25) indicates monodispersity, *i.e.* a homogenous sample population, whereas higher PDI values (0.25-1) indicate the co-presence of particles of different sizes. In **paper I** DLS was used to assess solubility (*i.e.* the absence of aggregates) of different antigens both directly after purification as well as during storage at different temperatures.

5.3 Enzyme assays

The most straightforward way to determine the rate at which enzymes catalyse reactions is to quantify either the amount of consumed substrate, generated product or both over time.

5.3.1 Detection by Gas Chromatography

In **paper II** and **III**, gas chromatography coupled with flame ionisation detection (GC-FID) was used to quantify the concentration of substrates (**paper II**) as well as products (**paper II**, **paper III**) in enzymatic reactions to calculate the rate of their change over time. This method is based on the chromatographic separation of volatile analytes by temperature, which are burnt upon elution from the column, producing an electrical signal in the detector that is proportional to the sample amount.[287] Absolute sample quantities can be calculated by comparing elution peak areas to those of a standard curve recorded at identical experimental conditions. Alternatively, the sample concentration can be calculated by relating the sample peak area to that of an internal standard (Int Std) of known concentration (typically an alkane of medium chain length), which is spiked into the sample solvent or the reaction mixture. In that case, a relative response factor (RRF), which can be measured (**paper II**) or calculated[288] (**paper III**), is required to adjust for compound-dependent differences in ionisation response upon combustion, according to equation 17.

$$conc\ (Sample) = \frac{area(Sample)}{area(Int\ Std)} \frac{1}{RRF} \times conc(Int\ Std) \qquad 17$$

The identity of products in **paper II** and **III** were ascertained using a mass spectrometry (MS) detector, which is connected within the same GC instrument. Moreover, **paper III** used the relative retention time of the analytes compared to an alkane standard mix to calculate their linear retention indices,[289] which represents a system- and method-independent value that can be used to determine analyte identity in reference to literature values.

Prior to their detection by gas chromatography, compounds need to be extracted from aqueous reaction mixtures using organic solvents, such as ethyl acetate for more polar compounds (lanosterol, **paper II**) or hexane for hydrophobic compounds (spiroviolene, **paper III**).

The volatility of some compounds may need to be enhanced by derivatisation in order for them to be separated by GC.[290] Derivatisation methods describe the chemical modification of polar groups, such as *e.g.* esterification or amidification of carboxylic acids to decrease polarity. A very efficient way of enhancing compound volatility is the formation of trimethylsilyl ethers from alcohols using N-Methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA).[290,291] This silylation derivatisation was used to enhance the volatility of lanosterol in **paper II**.

5.3.2 Detection by Malachite Green assay

In **paper III** and **IV**, a Malachite Green assay was used to quantify product concentrations.[292] This widely-used phosphate detection assay is based on the generation of a complex between inorganic phosphate (such as released by phosphatase activity in **paper IV**), malachite green and ammonium molybdate. This complex, which is formed under acidic conditions, absorbs light at 620-650 nm. By quantifying the colorimetric signal with a spectrophotometer in relation to a linear phosphate standard curve, the amount of phosphate in the reaction can be determined. In **paper III** the inorganic pyrophosphate released by the enzyme was first hydrolysed to inorganic phosphate by adding an inorganic pyrophosphatase to the reaction buffer prior to colorimetric detection.[292]

5.3.3 Kinetic analysis

In **papers II-IV** kinetic parameters were obtained by quantifying initial rates (*i.e.* the linear accumulation of product over time) at different substrate concentrations. Data were fit to equations 10 and 14 using non-linear regression to determine kinetic parameters k_{cat}, K_{M}, $k_{\mathrm{cat}}/K_{\mathrm{M}}$ and the Hill coefficient. Comparison of quality of fit between the two equations (as measured by R^2 value) was used to judge whether cooperativity was present.

Since no saturation within the tested substrate range was observed for the enzyme in **paper II**, $k_{\mathrm{cat}}/K_{\mathrm{M}}$ values were determined from the linear slope of the plot, as described in equation 13.

5.3.4 Eyring transition state analysis

Equation 18 can be formed from equation 16 using apparent second order enzyme rate constants.

$$ln\left(\frac{\frac{k_{cat}}{K_M}}{\frac{k_BT}{h}}\right) = -\frac{\Delta H^\ddagger}{R} \times \frac{1}{T} + \frac{\Delta S^\ddagger}{R} \qquad 18$$

In order to determine activation enthalpy and entropy of hOSC (**paper II**) apparent second order rate constants ($k_{\mathrm{cat}}/K_{\mathrm{M}}$) were measured at different temperatures. By plotting their logarithm over the inverse absolute temperature according to equation 17, the activation enthalpy and entropy can be obtained from the slope and intercept of the resulting linear regression, respectively.

5.4 Structural Biology

The aim of structural biology techniques is to determine a high-resolution model of a protein's atomic coordinates in order to visualise its molecular structure.

5.4.1 X-ray crystallography

X-ray crystallography is based on the diffraction of X-rays by proteins that are arranged in a crystal lattice.[293] First, protein samples are prepared at high concentration and are then screened with a panel of different crystallisation buffers that differ in type and concentration of precipitant. Crystallisation describes the ordered precipitation of a protein into a lattice. The crystal is then incrementally rotated in an X-ray beam to record diffraction patterns from all angles, which is typically performed at specialised national synchrotron facilities. Diffraction patterns arise from the constructive interference of elastically scattered electrons resulting from the symmetry of the crystal. Using sophisticated image analysis software the diffraction patterns are combined to infer electron density maps. Typically phase information is derived from homologous proteins for which structures are available from before (molecular replacement) or from selenomethionine residues that are incorporated into the protein structure. Finally, a structural model of the protein is constructed by fitting the polypeptide chain into the electron density map.

X-ray crystallography can generally yield high resolution structures of proteins across a vast range of sizes. The major bottleneck with this technique lies in the unpredictability and potential tediousness of the protein crystallisation process, as well as the requirement for a high concentration protein sample.[294] X-ray crystallography was used to determine the structure of a bacterial diterpene cyclase in **paper III**.

5.4.2 Cryogenic electron microscopy

Cryogenic electron microscopy (cryo-EM) is a microscopy technique that uses a beam of high-speed electrons instead of a beam of focused rays of light to image samples, thus overcoming the diffraction limited resolution of conventional light microscopy. In single particle analysis,[295,296] biological samples are prepared in aqueous solution and rapidly frozen into a thin layer on a grid by plunge-freezing the sample into liquid ethane. The rapid freezing procedure promotes the formation of non-crystalline vitreous ("glass-like") ice and ideally freezes the protein sample in its native conformation. In the microscope, electrons pass through the frozen biological sample and are scattered by specimen atoms. In oversimplified terms, interference patterns resulting from different scattering centres in the sample are recorded as 2D projections on a screen. Since protein molecules are frozen in various orientations across the section of vitreous ice, many individual orientations are recorded in a single projection. By grouping snapshots of different protein orientations within and across many projections recorded over time, they can then be averaged to yield higher resolution 2D-images of the protein from different angles (referred to as 2D classes). Subsequently, a 3D map of the particle electron density is reconstructed from all 2D classes using sophisticated image analysis software. Finally, a structural model of the protein is constructed by fitting the polypeptide chain into the electron density map.

Cryo-EM has the advantage that structures are resolved in their native conformations, thus avoiding artefacts resulting from crystallisation, but is somewhat limited to proteins greater than 80-100 kDa in size. Cryo-EM was used to determine the structure of trimeric ancestral antigens in **paper I**. The ISOLDE software package was used to model atom coordinates into the electron density map using MD simulations with an AMBER force field in ChimeraX.[297,298]

6 – Present Investigation

6.1 Paper I – Design, structure and plasma binding of ancestral β-CoV scaffold antigens

6.1.1 Context within this thesis

The cryo-EM structure of SARS-CoV-2 S-2P, a pre-fusion stabilised SARS-CoV-2 spike protein variant (**chapter 4.1.1**) was published online in February 2020, only months after the first cases were officially reported and even before the WHO declared the SARS-CoV-2 outbreak a pandemic.[184] This structure (and others reported shortly after) have greatly facilitated understanding molecular details of how SARS-CoV-2 enters its target cells. One aspect that may have contributed to this remarkable speed is the fact that the two respective proline mutations could be inferred from homologous virus proteins, for which the equivalent structure-derived stabilisation had been previously reported (see **chapters 3.5.2** and **4.1.1**). Several studies have further improved S protein stability by rational structural protein engineering, such as *e.g.* introducing disulphide and salt bridges or improving packing of residues in the protein core.[118,119,299-301] Not discounting the value of these studies, which have been essential for development of vaccines and treatments, these approaches all share the limitation of requiring previous knowledge about the general biology and structure activity relationship of coronavirus spike proteins.

In this context, we wondered if the stabilisation effects generally observed in reconstructed ancestors (including several therapeutic proteins (**chapter 3.4.3** and **3.5.1**)) would also apply to the highly unstable SARS-CoV-2 S protein. The rationale behind this approach is two-fold. Firstly, this approach obviates the need for available structural and functional knowledge, as it is based exclusively on available sequence information. The required screening effort is also expected to be minimal, since reconstructed ancestral proteins are typically folded and functional (see Thomson *et al.*[130] for a schematic). An ASR-based antigen stabilisation approach would therefore be particularly useful in the case of an emerging virus for which the surface antigen cannot be stabilised based on previous structural knowledge. Secondly, ancestral proteins consolidate sequence information from several related species or even entire clades (depending on the age of the ancestor). An ancestral antigen is therefore likely to accurately capture conserved epitopes, which would make it a suitable vaccine candidate for immunofocusing (**chapter 3.5.2**). This property is particularly important in light of the SARS-CoV-2 S protein's rapid real-time evolution, resulting in highly mutated strains that evade prior immunity.

6.1.2 Investigation

In a first step, a phylogenetic tree was constructed using the amino acid sequences of SARS-CoV-2 S protein homologues, most of which were available at the onset of the pandemic. Next, ASR was performed using a maximum likelihood approach and the sequences of ancestors (referred to as ancestral scaffold antigens (AnSAs)), corresponding to nodes 3, 5, 6 and 9 upstream of the SARS-CoV-2 S protein were retrieved (see appended manuscript for the tree). The evolutionary history of the SARS-CoV-2 virus and its S protein is complex and recombination events involving *e.g.* the S protein RBD from strains outside of the sarbecovirus clade have been reported.[302] Since these more complex evolutionary events are not accounted for in the employed evolutionary model, the presented phylogenetic tree cannot be used to make conclusive observations about SARS-CoV-2 evolution. This methodological simplification was however considered tenable in the context of this study given the primary aim of using ASR as a straightforward sequence-informed protein engineering technique, rather than a fully accurate account of SARS-CoV-2 evolution. A reconstruction excluding those sequences from the alignment that were not available at the beginning of the pandemic produced almost identical ancestral sequences for AnSA-5 and -6. In line with β-coronavirus sequence diversity being greatest in the NTD (**chapter 4.1.1**), we observed that the ancestors differed most strongly from the WT protein in this domain (43-70% sequence identity) as well as in the RBD (49-77% sequence identity).

In agreement with multiple studies that have observed enhanced expression yields of ancestral proteins in laboratory host strains (see **chapter 3.4.3)**, AnSA-5 and -6 could be expressed in high yield and purity in Expi293F cells, which were in the same range as observed for HexaPro (data shown in appended manuscript). In contrast, AnSA-3 showed lower yields than

HexaPro and AnSA-9 could not be detected by SDS-PAGE, which is why further discussion will focus on AnSA-5 and -6. As mentioned in **chapter 4.1.1**, HexaPro was obtained from an extensive library screening approach, whereas the two highly-expressing proteins in this study were obtained after testing four variants. Moreover, DLS experiments indicated that AnSA-5 and AnSA-6 exhibited similar polydispersity indices and a somewhat higher ratio of soluble to aggregate peak than HexaPro (data shown in appended manuscript).

The structural integrity of AnSA-5 and -6 was assessed using cryo-EM, readily yielding 2.6 and 2.8 Å 3D reconstructions, respectively. Both proteins were found in the locked pre-fusion conformation with all three RBD domains in the down-conformation (shown for AnSA-5 in Fig. 4). A high local resolution extending to surface-exposed loop regions allowed a detailed structural analysis, especially in AnSA-5. When mapping the ancestral mutations onto the structure, it becomes apparent that these are primarily located on the surface in the S1 subunit (Fig. 5A). Moreover, many of these mutations are involved in hydrogen bonds close to domain interfaces within and across monomers (Fig. 4A, Fig. 5B). One example, shown in the bottom left inset of Fig. 4A, is the hydrogen bond between S284 in the NTD and the backbone of F310 close to the RBD in the same monomer. The respective segments are further hydrogen bonded to the SD2 domain *via* T604 and Q602, such that the three loops at the triple domain interface are stabilised. Another example, which is shown in the top right inset of Fig. 4A, is a hydrogen bond network between two residues in the RBD of one monomer (S386 and E490) and one residue in the NTD of the neighbouring monomer (K226). For a detailed discussion of additional examples the reader is referred to the appended manuscript. In contrast to the rational stabilisation approaches discussed above, which focus on physically locking dynamic regions *via* targeted individual mutations, it appears that AnSA-5 and -6 are mainly stabilised by inter-domain hydrogen bonding networks.

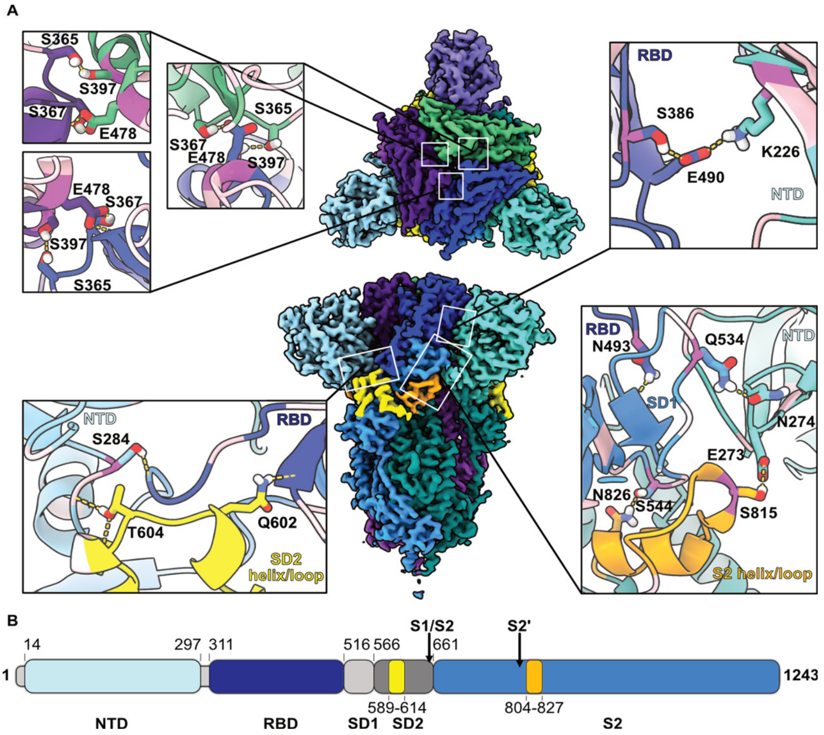


Figure 4. Cryo-EM structure of AnSA-5. (A) Top and side view of AnSA-5 electron density map. In each chain of the trimer the NTD is coloured in a lighter shade, the RBD in a darker shade and the remaining monomer in an intermediate shade (blue for chain A, green for chain B and purple for chain C). Particular motifs discussed in more detail in the appended manuscript are coloured yellow and orange. Interdomain hydrogen bond networks are shown as insets with Cα atoms of ancestral mutations highlighted in pink (dark pink if they enable a novel hydrogen bond *via* sidechains with different properties). A glycosylation on residue 274 is omitted for clarity. **(B)** Domain overview as bar graph (colours shown for chain A). The SD1 and SD2 domains are coloured light and dark grey (grey omitted in (A) for clarity). Figure from submitted manuscript,[303] prepared in collaboration with J. Andréll.

Next, we assessed the suitability of AnSA-5 and -6 as potential immunogens by first measuring different correlates of protein stability. DLS measurements performed after 30 days storage at 4 °C, room temperature and 37 °C indicated that AnSA-5 and -6 remained highly soluble at physiological temperatures, whereas HexaPro showed an increase in the relative amount of aggregates (Fig. 6). Importantly, both AnSA-5 and -6 interacted with blood plasma samples of convalescent Covid-19 patients (Fig. 7), albeit with lower affinity than observed for HexaPro, which may be due to the heavily mutated ancestral RBD domains. It is highly likely that the antibodies in patient plasma samples bind to conserved epitopes that are shared among HexaPro, AnSA-5/-6 and the S-protein of the respective SARS-CoV-2-strain that the patients encountered. While such antibodies may not be neutralising *per se*, they may still hold the potential for conferring broad protection by other mechanisms of antibody-mediated immunity. AnSA-5 and -6 therefore represent interesting immunogen candidates to elicit such antibody repertoires, although further experiments are required to confirm this hypothesis.

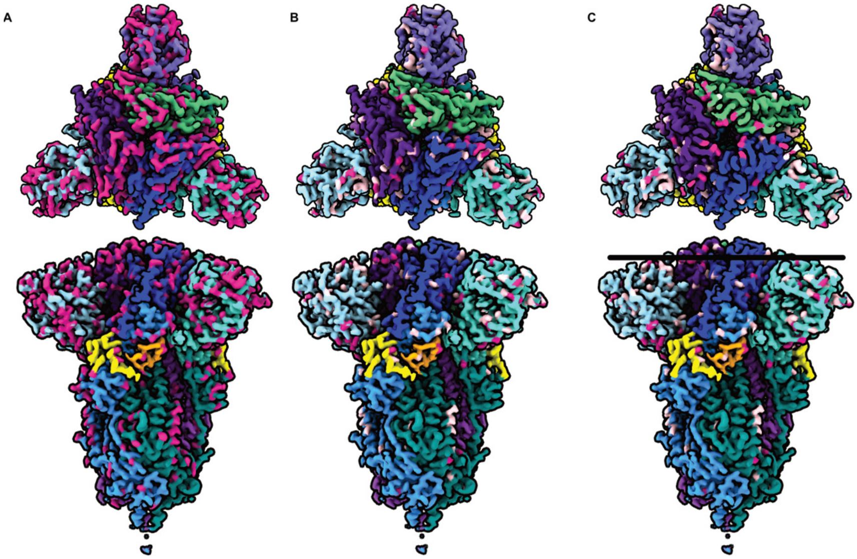


Figure 5. Distribution of ancestral mutations on AnSA-5. Top and side view of electron density map. **(A)** All mutated residues in AnSA-5 compared to the SARS-CoV-2 WT S protein are highlighted in dark pink. **(B)** Same structure as in (A) in which ancestral residues that are not involved in hydrogen bonds are omitted. Ancestral mutations that have non-similar or weakly similar side chain properties and form hydrogen bonds *via* their sidechains are shown in dark pink, other mutations (similar side chain properties or interactions *via* the backbone) are shown in light pink. **(C)** Same view as in (B) with the top-view marginally clipped along the z-axis (approximate clipping height indicated by black line). Figure modified from submitted manuscript.[303]

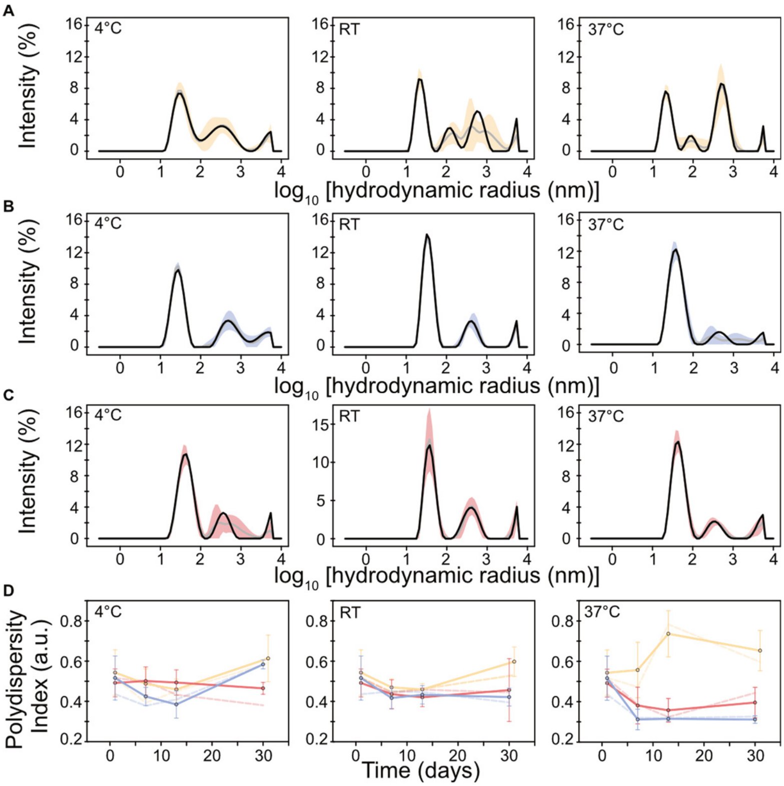
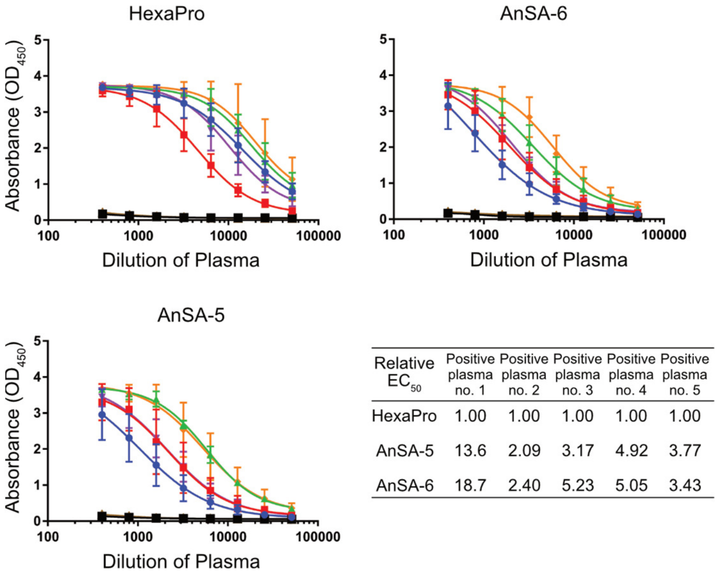


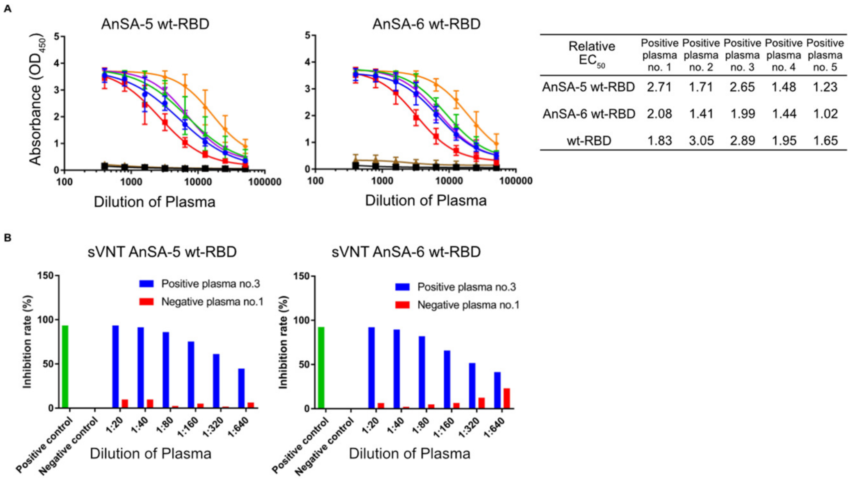
Figure 6. Solubility of HexaPro, AnSA-5 and AnSA-6 after 4 week storage at different temperatures. DLS particle size distributions are plotted by intensity for **(A)** HexaPro **(B)** AnSA-5 and **(C)** AnSA-6. Averages from five technical repeat measurements are shown and standard deviations are shown as coloured area plots. **(D)** Polydispersity indices at different time points during storage were derived from DLS plots. Figure from submitted manuscript.[303]

Finally, AnSA-5 and -6 were used as scaffolds to host the Wuhan WT RBD domain, thereby replacing the respective ancestral RBD domains, which reduced expression yields and solubility but restored ACE2 receptor-binding, as quantified by surface plasmon resonance. Nevertheless, expression yields were still higher than reported for S-2P. A likely reason for reduced expression yields compared to the original AnSAs may lie in the disruption of interdomain hydrogen bonds involving the ancestral RBDs, as discussed above. The reconstitution of the WT-RBD domain into the ancestral background increased affinity between the antigens and convalescent plasma samples (Fig. 8A). Moreover, it is highly likely that this increase is due to the added interactions with the WT RBD domain as demonstrated by the fact that a plasma sample successfully intercepted binding of these chimera spike proteins to the ACE2 receptor in a surrogate virus neutralisation assay (Fig. 8B). These results highlight that ancestral spike proteins can serve as scaffolds to host the RBD of a modern evolved virus. Since the AnSAs comprise sequence information from multiple sarbecoviruses, it is possible that they would be compatible with the RBD domains of emerging SARS-CoV-2 variants of concern but also novel sarbecoviruses.



Relative EC_{50}	Positive plasma no. 1	Positive plasma no. 2	Positive plasma no. 3	Positive plasma no. 4	Positive plasma no. 5
HexaPro	1.00	1.00	1.00	1.00	1.00
AnSA-5	13.6	2.09	3.17	4.92	3.77
AnSA-6	18.7	2.40	5.23	5.05	3.43

Figure 7. Interaction of HexaPro, AnSA-5 and AnSA-6 with plasma samples of convalescent Covid-19 patients. Binding (from triplicates) as quantified by enzyme-linked immunosorbent assay. COVID-19 convalescent plasma samples: 1 (blue circles), 2 (red squares), 3 (green triangles), 4 (purple triangles), 5 (orange diamonds). Negative (pre-pandemic) plasma samples: 1 (black squares), 2 (brown triangles). EC_{50} values were obtained from non-linear regression and are normalised to the EC_{50} value of HexaPro per plasma sample. Figure modified from submitted manuscript,[303] prepared by F. Zuo and colleagues.



Relative EC_{50}	Positive plasma no. 1	Positive plasma no. 2	Positive plasma no. 3	Positive plasma no. 4	Positive plasma no. 5
AnSA-5 wt-RBD	2.71	1.71	2.65	1.48	1.23
AnSA-6 wt-RBD	2.08	1.41	1.99	1.44	1.02
wt-RBD	1.83	3.05	2.89	1.95	1.65

Figure 8. Interaction of AnSA-5 and AnSA-6 harbouring the WT-RBD domain with plasma samples of convalescent Covid-19 patients. (A) Binding (from triplicates) as quantified by enzyme-linked immunosorbent assay. COVID-19 convalescent plasma samples: 1 (blue circles), 2 (red squares), 3 (green triangles), 4 (purple triangles), 5 (orange diamonds). Negative (pre-pandemic) plasma samples: 1 (black squares), 2 (brown triangles). EC_{50} values were obtained from non-linear regression and are normalised to the EC_{50} value of HexaPro per plasma sample. (B) Surrogate virus neutralisation assay of patient sera binding to AnSA-5/6 harbouring the WT-RBD domain. Figure modified from submitted manuscript,[303] prepared by F. Zuo and colleagues.

6.1.3 Concluding remarks

In summary, **paper I** highlights that ancestral sequence reconstruction could successfully be used to yield stable, soluble and highly expressing S protein variants in a straightforward manner. The amount of tested samples was considerably reduced compared to conventional rational antigen stabilisation approaches and no structural input was required to achieve stabilisations on par with reported literature standards. While these reconstructed S proteins do not represent actual ancestral evolutionary states, they may constitute potentially useful immunological tools.

Importantly, these ancestral scaffold antigens did not bind to the ACE2 receptor but interacted with plasma samples of convalescent patients. This observation strongly suggests that AnSAs display conserved epitopes on their surface that are recognised by broadly reactive antibodies and likely target the S2 subunit. The AnSAs presented in this work may therefore be useful as probes to isolate and identify such cross-reactive antibodies. Moreover, they may potentially elicit cross-reactive antibody responses when used as vaccine, although this hypothesis requires further experimental validation.

The exchange of the ancestral RBD domain for the corresponding WT domain underscores the importance of the reconstructed ancestors as scaffolds to be adaptable to emerging sarbecovirus and novel SARS-CoV-2 strains. Finally, the high fidelity display of such conserved epitopes may make AnSAs interesting candidates for boosting sub-dominant antibody repertoires from prior immunisations (resulting from vaccination, infection or both) in analogy to designed heterologous prime-boost vaccination schemes (see **Chapter 3.5.2**).

6.2 Paper II – Designed out-of-active site mutations in human oxidosqualene cyclase modulate the activation entropy and enthalpy of the cyclization reaction

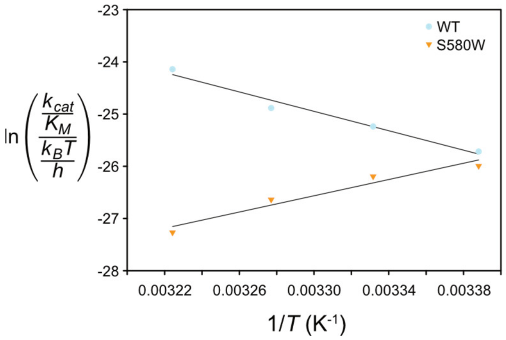
6.2.1 Context within this thesis

Human OSC is an essential enzyme in human steroid metabolism and has been intensively studied ever since its discovery in the 1960s (**chapter 4.2.3**). Since the enzyme catalyses the first dedicated step in cholesterol biosynthesis, it represents an interesting target for treatment of hypercholesterolemia and related cardiovascular diseases.[304] Statins (HMG-CoA-reductase inhibitors) in comparison target sterol biosynthesis several metabolic steps upstream and have been associated with causing some (albeit minor) side effects, such as myopathies and increased risk of diabetes mellitus.[305] As mentioned in **chapter 3.1**, one objective for engineering enzymes is to better understand principles of their catalytic mechanisms. By understanding molecular features that systematically increase or decrease enzyme activity, novel inhibitor design strategies that aim to mimic the effect of engineered mutations may potentially be taken forward. Since its structure is known since 2004 and many aspects of the reaction mechanism have been identified,[200,222] hOSC lends itself well to rational engineering approaches.

In 2015 and 2020, comprehensive QM-MM/MD studies by Wu and colleagues covering all reaction steps have further contributed to clarifying mechanistic aspects of this enzyme.[221,224] The latter study also pointed out the distribution of electrostatic potentials in the active site and enzyme intermediates as an important aspect in the reaction.[221] As discussed in **chapter 2.4.3**, some enzyme reactions involving charge transfer and separation have been associated with positive activation entropy contributions to catalysis.[46] In fact, a representative of squalene hopene cyclases (bacterial hOSC homologues that share ca. 25% sequence identity with hOSC) was found to exhibit a large positive activation entropy at its physiological temperature (55 °C).[50] However, this particular SHC originated from a thermophile and as outlined in **chapter 2.4.1**, it is plausible that thermophilic enzyme reactions would benefit more strongly from higher activation entropies than mesophilic enzyme reactions. In this work, the activation entropy contribution to the hOSC cyclisation reaction, which operates at a physiological temperature of around 37 °C, was studied and modified using a rational enzyme engineering approach.

6.2.2 Investigation

In a first step, the temperature dependence of apparent second order rate constants for the hOSC reaction were determined using a kinetic assay coupled with GC-FID detection. Next, Eyring transition state analysis (**chapters 2.4.1** and **5.3.4**) was used to determine the activation enthalpy and entropy of the reaction from the dataset. Indeed, a positive $T \times \Delta S^{\ddagger}$ term of + 7.0 kcal mol^{-1} (at 37 °C) could be determined (Fig. 9, cyan circles), indicating that the transition state of the rate-limiting step (most likely the generation of the first carbocation by protonation of the epoxide moiety[221]) is described by more degrees of freedom (*i.e.* lower degree of order) than the ground state. The value is lower than for the thermophilic SHC but in the same order of magnitude as reported for *E. coli* EF-Tu GTPase (at 25 °C).[306] The reaction was further associated with an activation enthalpy penalty of 22.0 kcal mol^{-1}, possibly describing the enthalpic instability of the carbocation generated in the first reaction step.



	$\Delta H^{\ddagger}$ (kcal/mol)	$T^{*}\Delta S^{\ddagger}$ [a] (kcal/mol)	$\Delta G^{\ddagger}$ [a] (kcal/mol)
WT	22.0	7.0	15.0
S580W	-15.5	-32.2	16.7

[a] at 310 K

Figure 9. Eyring transition state analysis of hOSC WT and S580W variant. Plots are based on analysis of k_{cat}/K_M values at different temperatures. Data points are obtained from triplicates (detailed analysis described in appended manuscript). Enthalpy, entropy and Gibb's free energy of activation for the variants were obtained from the plot and are indicated in the table. The S580W variant is discussed further below. Figure modified from manuscript.

As discussed above, an increase of solvent entropy was identified in the enzymatic reaction compared to the un-catalysed reaction in several enzymes that have been described to proceed at low activation entropy barriers (or even favourable activation entropies). Although computational simulation of the un-catalysed OSC reaction and calculation of its associated activation enthalpy and entropy are outside the scope of this study, it is reasonable to assume that the un-catalysed reaction would be associated with a considerable entropy penalty due to the fixation of many freely rotatable bonds in the substrate, excluding the substrate as source of positive activation entropy. In contrast, the solvent may be a likely source of favourable activation entropy in hOSC. In a next step, we therefore studied the hydration of the hOSC active site. An intermediate snapshot from MD simulations (based of the product-bound crystal structure as proxy for the pre-folded transition state) revealed that 15 water molecules were found within a radius of 6 Å from the ligand (Fig. 10). These water molecules clustered in proximity to the catalytic aspartate (D455) and to a lesser extent in proximity to the catalytic base (H232). Positive activation entropy may for instance be achieved by their expulsion from the active site or translocation compared to the ground state (which is composed of unbound substrate, enzyme and solvent).

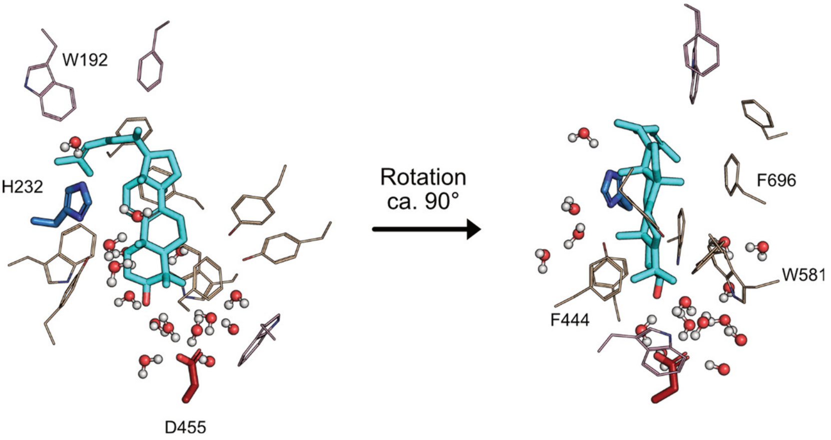


Figure 10. Active site hydration in hOSC. Close-up view of the hOSC active site (snapshot 100 of MD simulations is shown). The protein backbone is not shown for clarity. Lanosterol is shown as cyan sticks, aromatic side-chains lining the active site are shown as lines, and catalytic residues D455 and H232 are shown as red and blue sticks, respectively. All water molecules within a radius of 6 Å of lanosterol are shown as ball and sticks. Figure from manuscript.

An analysis of cavities in hOSC across the MD simulation trajectory using Caver 3.0 software[271] (see **chapter 5.1.2**) revealed multiple prominent tunnels approaching the active site from several directions (Fig. 11). The most relevant tunnels that could be identified were (T1) the suggested re-protonation tunnel, which extends from the active site to the surface through the inner α-helical barrel in domain 1 and which was discovered in conjunction with the presentation of the hOSC crystal structure in 2004[222]; (T2) a tunnel extending from the same opening as T1 perpendicular to T1 to the "top" of the enzyme surface; and (T3) which represents the highly hydrophobic substrate access tunnel. Other important tunnels are discussed in more detail in the appended manuscript. Many of the observed tunnels, including T1 and T2 are lined by hydrophobic residues in proximity to the active site (with exception of the catalytic D455 and adjacent cysteine residues) but become gradually more hydrophilic when approaching the enzyme surface. Therefore, water molecules could be expected to move along such tunnels rather quickly without being slowed down by solvation.[307]

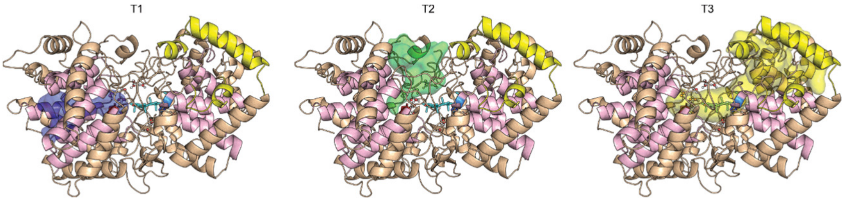


Figure 11. Overview of tunnels in hOSC. The inner α-helical barrels of both domains are coloured in light pink, whereas the outer α-helical barrels are coloured in wheat. The membrane insertion helix and surrounding hydrophobic surface patch are shown as yellow cartoon. The substrate is shown in cyan sticks and tunnels are shown as transparent surfaces. The hydrophobic substrate access tunnel T3 is shown in yellow since it is directly connected to the membrane insertion helix. Water molecules within a sphere of 6 Å from lanosterol are shown in sphere and stick representation. For figures of additional tunnels, the reader is referred to the appended manuscript. Figure modified from manuscript.

In a next step, we aimed to engineer these cavities by rational design and to assess the consequences on activation entropy. Specifically, residues at tunnel bottlenecks were identified and mutated to larger residues. One example is constituted by the exchange of S580 for a tryptophan, which is discussed in more detail below. For the discussion of other variants, the reader is referred to the appended manuscript. As intended, the S580W variant showed a difference in T2 passing by this residue (Fig. 12). Besides a reduction in tunnel diameter, the tunnel hydration was strongly affected. Only three water molecules were found within a radius of 3 Å from W581, a residue located at the tunnel opening towards the active site (Fig. 12B), whereas nine water molecules were found within the corresponding radius in the WT (Fig. 12A). This is most likely the result of hydrophobic interactions between the introduced W580 and surrounding hydrophobic residues such as Y528, Y530 and W581 (Fig. 12B). Likewise, active site hydration in proximity to the tunnel opening (close to D455) was reduced. Moreover, the ligand was observed to localise more closely to hydrophobic residues such as W581 and F696 in the active site compared to the WT. In line with these structural observations, a drastic change in the activation enthalpy and entropy compared to the WT was observed in kinetic experiments. The S580W cyclisation reaction was associated with a strong favourable activation enthalpy (-15.5 kcal mol^{-1}), whereas the activation entropy was highly unfavourable (-32.2 kcal mol^{-1} at 37 °C). The difference to the WT is substantial with $\Delta_{\mathrm{WT\text{-}S580W}}\Delta H^{‡}$ of

37.5 kcal mol^{-1} and $\Delta_{\mathrm{WT\text{-}S580W}}\Delta S^{\ddagger}$ of 39.2 kcal mol^{-1} (at 37 °C). Not only was the driving force of the reaction reversed in this variant compared to the WT, but also the slope of the Eyring plot (Fig. 9, orange triangles). In fact, enzyme activity steadily decreased from 22 °C to 37 °C, which is in stark contrast to the WT reaction. However, no major difference in the product peak was observed by GC (results shown in appended manuscript), indicating that the mutation only indirectly affected the reaction.

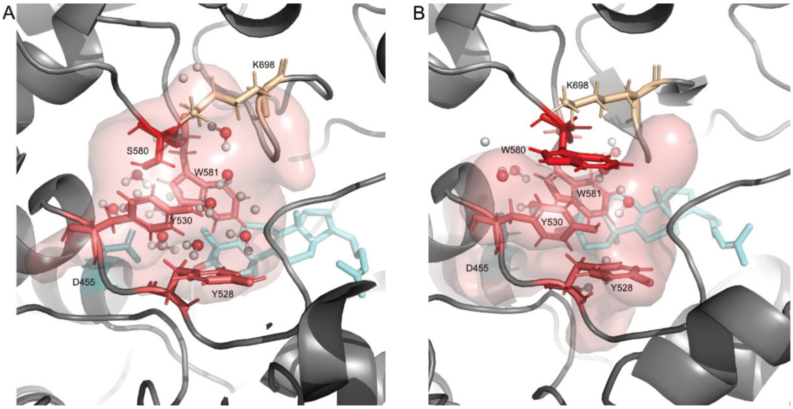


Figure 12. Close-up on T2 in (A) hOSC WT and (B) S580W variant. Images represent snapshots 100 from MD simulations. Lanosterol is shown in cyan sticks and the mutated residue 580 as red sticks. Hydrophobic residues lining the tunnel and interacting with W580 are shown as light red sticks. Figure from manuscript, prepared by D. Hueting.

Finally, we were interested in studying the interplay between different tunnel mutations as well as other types of mutations. In a first step, a small rational library of variants that combined each three different tunnel-constricting residues was constructed ("KTH"-library), which resulted in mostly inactive variants. However, the combination of tunnel residue mutations from this library with mutations from an unconstrained library ("DNA"-library) yielded several variants ("All"-variants) that showed very low preference for warm reaction temperatures over cold temperatures or reversed temperature dependence in analogy to the S580W variant (Fig. 13). Interestingly, no major changes in solubility or melting temperature were observed among the variants of all three libraries, indicating that changes in temperature preference are not the result of changed thermal stability. Moreover several of these variants (*e.g.* the All6-variant, which is discussed in the appended manuscript) displayed mutations on or close to the enzyme surface, suggesting that surface mutations may allow mutations in tunnels to occur, indicating a certain degree of epistasis. While further experiments would be required to elucidate the exact structural reason for this finding, initial observations from simulations indicate that the active site may change as result of the surface mutations.

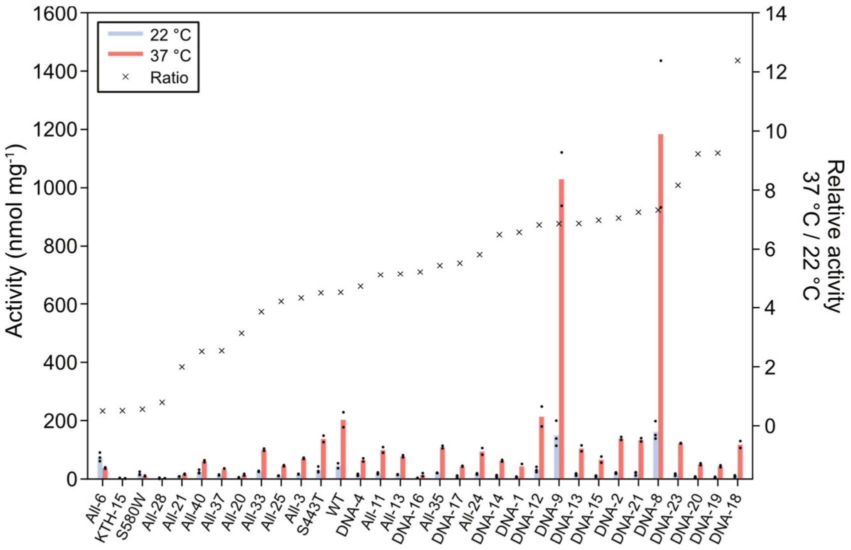


Figure 13. Temperature dependence of hOSC library variants. Lanosterol formation by different hOSC variants at 22 °C and 37 °C as measured by GC-FID. All variants from the three libraries that were active at 37 °C in an initial screen (KTH-, DNA-, All-libraries, see appended manuscript for detailed discussion) as well as the S443T and S580W variant were included. Data points are shown as black dots (triplicates for reactions at 22 °C, duplicates for reactions at 37 °C). The ratio of lanosterol formed at 37 °C over 22 °C is indicated as crosses (secondary y-axis). Figure from manuscript.

6.2.3 Concluding remarks

The results discussed in **paper II** emphasise the importance of activation entropy for the hOSC reaction at physiological temperatures. Moreover, the potential movement of active site water molecules that are supplied through a network of different tunnels is suggested as likely cause for the positive activation entropy. A single residue exchange in a newly identified solvent tunnel reduced activation entropy by almost 40 kcal mol^{-1} compared to the WT at the enzyme's physiological reaction temperature. Moreover, the screening of multiple library variants indicated that the activity of such tunnel variants at different temperatures may be further modulated by a combination of surface residues, which affect the active site in ways that are challenging to predict.

Within the greater context, the results in **paper II** open up the question whether small molecules could be designed to block identified tunnels, such as to mimic the effect of the S580W mutation. The rationale behind this approach would be that a tunnel obstruction resulting in reversed temperature dependence would render hOSC essentially inactive at human body temperature. Due to their Nature, tunnels should in principle be accessible for small molecules from the enzyme surface, although the respective tunnel characteristics would certainly impact effective access.

Positive activation entropies and their modulation by tunnel-obstructing mutations had previously been reported for thermophilic SHC and were herein confirmed for mesophilic hOSC.[50] One may therefore hypothesise that the same trends likely apply to other homologous class II triterpene cyclases, such as occur in yeast or plants and produce compounds of industrial interest. The consequences of tunnel-mutations on temperature dependence may potentially be interesting in the light of expressing such enzymes for triterpene production in transgenic plants or cell factories with a different temperature optimum.

6.3 Paper III – Engineering of Ancestors as a Tool to Elucidate Structure, Mechanism and Specificity of Extant Terpene Cyclase

6.3.1 Context within this thesis

While lanosterol, the terpene discussed in **paper II** is relevant as endogenous steroid precursor in animal physiology, the majority of terpenes are secondary metabolites produced in plants, fungi and microbes. Due to their diverse and unique properties many of these terpenes are industrially relevant as fuels, materials, pharmaceuticals and consumer products (**chapter 4.2.1**) such as *e.g.* limonene (C-10), artemisinin (C-15) or steviol (C-20). A few terpenes may be extracted from natural resources such as pinene from turpentine, a by-product of the pulp and paper industry,[308] or steviol glycosides from leaves of the candyleaf plant. But even though the summed production of all terpenes in Nature is immense, the respective quantities of individual terpenes can be very low, complicating their extraction (as exemplified by harmful early harvest routes of Paclitaxel from bark of the Pacific yew[309]). Therefore heterologous synthesis routes in fast-growing plants or metabolically engineered microorganisms, such as *e.g. E. coli* and *Saccharomyces cerevisiae*, are a topic of intensive research.[310-314]

Since many terpene cyclases have been described to exhibit low rates (k_{cat} values of less than one turnover per minute) as well as pronounced promiscuity, such heterologous production would benefit from using engineered terpene cyclases.[315,316] Engineering terpene cyclases is further desirable from a synthetic chemistry perspective, given their versatility in stereospecific carbon bond formations.[317] However, engineering efforts are impeded by the complexity as well as diversity of terpene cyclase reaction mechanisms. Since even slight modifications in the active site architecture of terpene cyclases can affect both pre-folding and cation formation,[318,319] structural knowledge about how these enzymes mediate specific carbocation transfers in their active sites greatly facilitates their rational engineering.[315,316,320]

Given the success of ASR in stabilising many proteins and enzymes (see *e.g.* **paper I**) and reports of multiple crystal structures of ancestral proteins,[136,153,154,321,322] we aimed to use this technique in order to obtain a terpene cyclase crystal structure as basis for further structure based rational engineering of its specificity.

6.3.2 Investigation

Spiroviolene synthase (see **chapter 4.2.4**) is a bacterial class I di-terpene cyclase that produces a spirocyclic compound in a single enzymatic step involving six intermediates. With a complex reaction mechanism[227] as well as low catalytic efficiency (109 s^{-1} M^{-1})[126] SvS represents a typical example of a terpene cyclase. Moreover, the enzyme was found to promiscuously accept a shorter sesquiterpene substrate (FPP, farnesyl pyrophosphate),[126] which it converts at a catalytic efficiency of 38 s^{-1} M^{-1}.

As discussed above, structural information is highly beneficial for engineering terpene cyclases. Due to low expression yields and the protein's tendency to precipitate when concentrating it, we could not obtain SvS at sufficiently high concentrations to proceed with crystallisation trials. In a first step we therefore determined the crystal structure of SvS-A2 instead, a reconstructed ancestor of SvS with an increased melting temperature compared to the WT[126] (**chapter 4.2.4**), to a resolution of 2.3 Å (Fig. 14A). The enzyme consists of two monomeric canonical α-fold domains (**chapter 4.2.2**) arranged in an antiparallel fashion, each harbouring the active site at the centre of the inner α-helical barrel. Both SvS-A2 as well as SvS-WT were confirmed to be dimeric in solution using multi angle light scattering experiments. Next, a structural model of SvS-A2 docked to pre-folded substrate was constructed using molecular mechanics simulations (Fig. 14B). The DDxx(x)D-motif (violet in Fig. 14B) and the NSE-motif (teal in Fig. 14B) are located across of each other on the upper walls of the active site and coordinate the tri-metal Mg^{2+} cluster between them. The metal ion cluster in turn binds to the pyrophosphate moiety of the substrate geranylgeranyl pyrophosphate (GGPP).

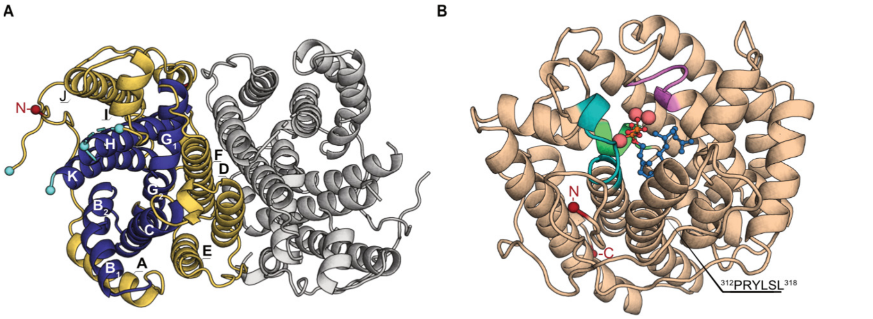


Figure 14. SvS-A2 crystal structure. (A) Dimeric crystal structure of SvS-A2 (top view). Each monomer represents the typical α-fold of this enzyme family. Helices in monomer A are labelled and the inner and outer α-helical barrels are coloured blue and yellow, respectively. Residues adjacent to unresolved loops are highlighted as cyan spheres. **(B)** Model of pre-folded GGPP (blue sticks and spheres) and the tri-Mg^{2+} cluster (pink spheres) docked to the active site of SvS-A2 monomer (missing loops modelled). The DDxx(x)D- and NSE metal binding motifs are highlighted in violet and teal, respectively. The effector motif is highlighted in green (see appended paper for a discussion of this motif). Figure modified from Schriever *et al.*[323]with permission by the journal.

SvS-A2 and SvS-WT share 77% sequence identity and catalyse the same major and promiscuous reactions, indicating that the active site is likely highly conserved between the two enzymes. Indeed, the majority of ancestral mutations in SvS-A2 could be mapped to the enzyme surface, whereas almost no ancestral mutations were observed in the inner α-helical barrel. It is therefore reasonable to assume that the structure of SvS-A2 would serve as a suitable template for obtaining a high quality homology model of SvS-WT. However, an initial homology model showed sub-optimal average quality scores, which were particularly critical in the regions involving the two metal binding motifs (Fig. 15A, "SvS-WT-Hom1"). An analysis of the ancestral mutation distribution indicated a surface patch of 5 consecutive ancestral mutations at the "bottom" of the enzyme, which is positioned ca. 18 Å away from the active site. In an attempt to further increase homology of the ancestral protein to the WT, these five residues as well as the only ancestral mutation in the otherwise conserved metal binding site were mutated back to the extant sequence in an attempt to further increase sequence identity.

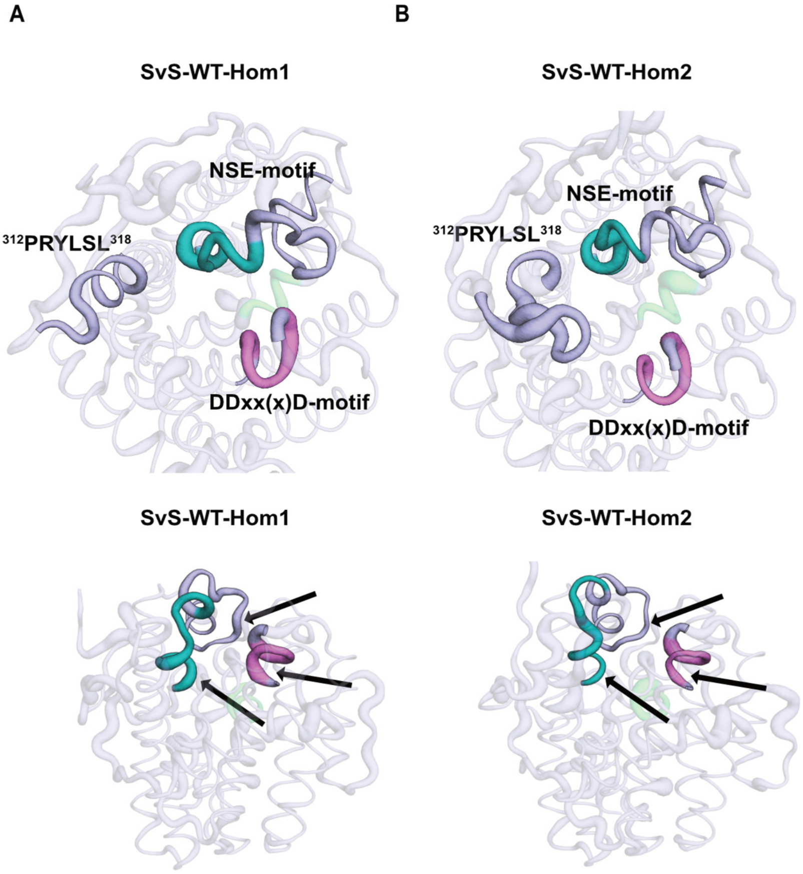


Figure 15. Comparison of different SvS-WT homology models. Top view (upper panel) and side view (lower panel) of SvS-WT homology models constructed using **(A)** the SvS-A2 crystal structure and **(B)** the crystal structure of the modified surface variant of SvS-A2 as templates. The thickness of the cartoon represents the local quality score of the model (Z-score), where a thicker cartoon indicates poorer model quality (lower Z-score). Arrows in the lower panel indicate areas close to the conserved metal binding motifs in which the local Z-score is improved in SvS-WT-Hom2. Figure modified from Schriever *et al.*[323]with permission by the journal.

This surface variant displayed similar thermostability to SvS-A2 and its crystal structure (2.4 Å resolution) was highly similar to that of the original SvS-A2 crystal structure. Additionally, 5 out of 17 residues that were unresolved in the initial SvS-A2 crystal structure could be resolved in a loop that closes the active site. When constructing a homology model of SvS-WT using this surface variant structure as template, both the overall quality scores as well as the local quality scores in the metal binding motifs improved (Fig. 15B, "SvS-WT-Hom2"). This second homology model was finally energy minimised using the AMBER force field.

Both SvS-A2 and SvS-WT catalyse a promiscuous sesquiterpene cyclisation (yellow and orange bars, Fig. 16A) alongside the major diterpene cyclisation (blue bars, Fig. 16A). Using GC-MS analysis the major product of the FPP cyclisation reaction was identified as hedycaryol (detected as its thermal rearrangement product elemol), whereas farnesol was detected as a side product (see appended paper for details). In competition assays as well as single-substrate assays, SvS-A2 was found to show higher relative diterpene over sesquiterpene conversion than SvS-WT (Fig. 16A). Likewise, computationally determined binding free energies indicated that both enzymes had a greater affinity for GGPP than FPP, but that the difference in respective binding free energies ($\Delta_{\mathrm{GGPP\text{-}FPP}}\Delta G_{\mathrm{bind}}$) was greater in SvS-A2 than SvS-WT. The qualitative agreement between preferences in enzymatic activity and calculated substrate binding suggest that higher selectivity in SvS-A2 compared to SvS-WT may partly stem from differences in substrate binding. Due to differences in kinetic behaviour between SvS-A2 and SvS-WT[126] and between the two different substrates (Fig. 16B,C) it can however not be ruled out that rate-limiting steps may differ between the two enzymes and therefore impact the respective selectivities.

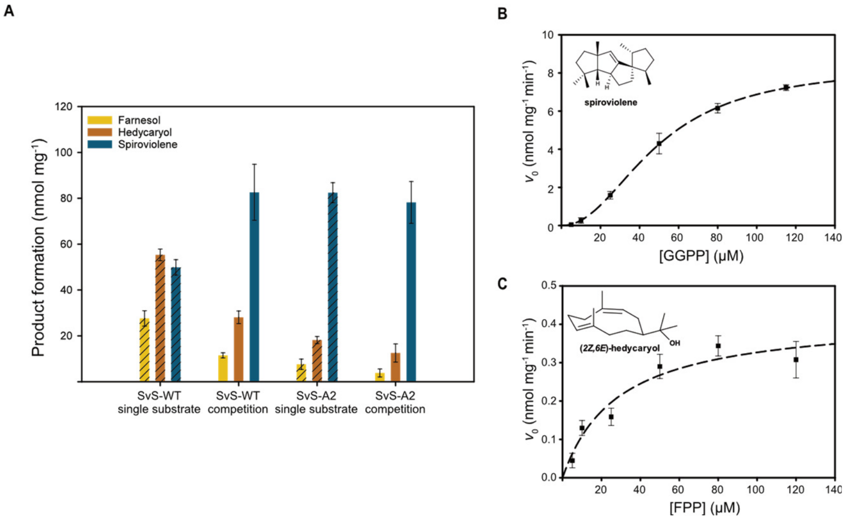


Figure 16. Promiscuous SvS activity. (A) Formation of product from FPP and GGPP in SvS-WT and SvS-A2 as measured by GC-FID in single substrate assays (striped bars) or competition assays (filled bars). **(B)** Initial rates for formation of spiroviolene and **(C)** hedycaryol in SvS-A2 were determined at different substrate concentrations. Lines represent data fit to the Hill equation (B) and Michaelis-Menten equation (C). Kinetic parameters determined from the fits are listed in the appended paper. Error bars are from triplicates. Figure modified from Schriever *et al.*[323] with permission by the journal.

The reaction intermediates from SvS and hedycaryol synthase had been previously identified by NMR spectroscopy.[227,324] In order to understand which residues are responsible for substrate specificity in SvS, these intermediates were analysed in the structural context of both the SvS-A2 and SvS-WT active sites, which are highly conserved. Snapshots of the different intermediates towards spiroviolene formation interacting with the active site in SvS-A2 (Fig. 17) and SvS-WT revealed several carbocation-π interactions, a common catalytic mechanism in this enzyme family (**chapter 4.2.2** and **paper II**). For example the first intermediate was observed to be stabilised by cation-π interactions involving *e.g.* W79 and F84 and possibly W82 and W156 (Fig. 17). Equivalent analyses of snapshots towards hedycaryol formation revealed cation stabilisation in a more confined area of the active site, involving the pyrophosphate anion that is eliminated from the substrate in the first reaction step as well as F84, which is conserved in related bacterial class I sesquiterpene cyclases. These results show how SvS (both the wild type and ancestor) can efficiently stabilise carbocations for both reactions, explaining the enzymes' promiscuous activities.

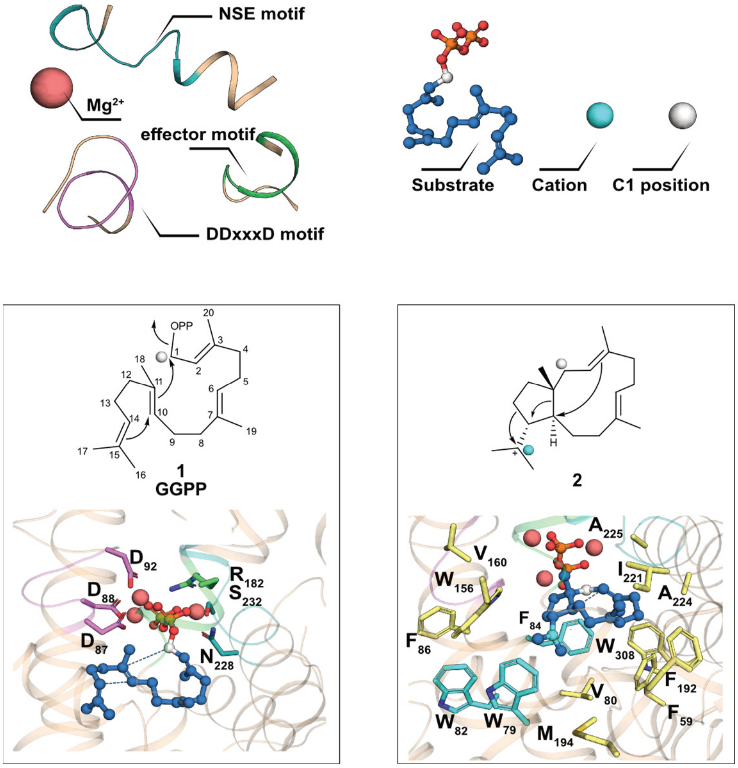


Figure 17. Snapshots of the reaction mechanism for spiroviolene formation from GGPP in the active site of SvS-A2. Only the first two snapshots (substrates and first carbocation intermediate) are shown for brevity. The reader is referred to the appended paper for a full version of this figure, comprising all reaction steps as well as the equivalent analysis for SvS-WT. The same interactions are observed in the active site of the SvS-WT homology model. The electron flow proposed by NMR experiments[227] is indicated by arrows in the 2D scheme. Residues that are involved in cation-π interactions are shown as cyan sticks. Figure modified from Schriever *et al.*[323] with permission by the journal.

We next aimed to enhance substrate specificity in the ancestral enzyme by rational engineering. A small library of 25 variants in or close to the active site was designed, mostly targeting residues adjacent to the carbocation-stabilising residues or different motifs in the enzyme to modify steric access to these. Using the Malachite Green assay, 3 variants were found to be promiscuous, 7 variants were inactive and 11 and 4 variants were found to be diterpene- or sesquiterpene-specific, respectively (Fig. 18A).

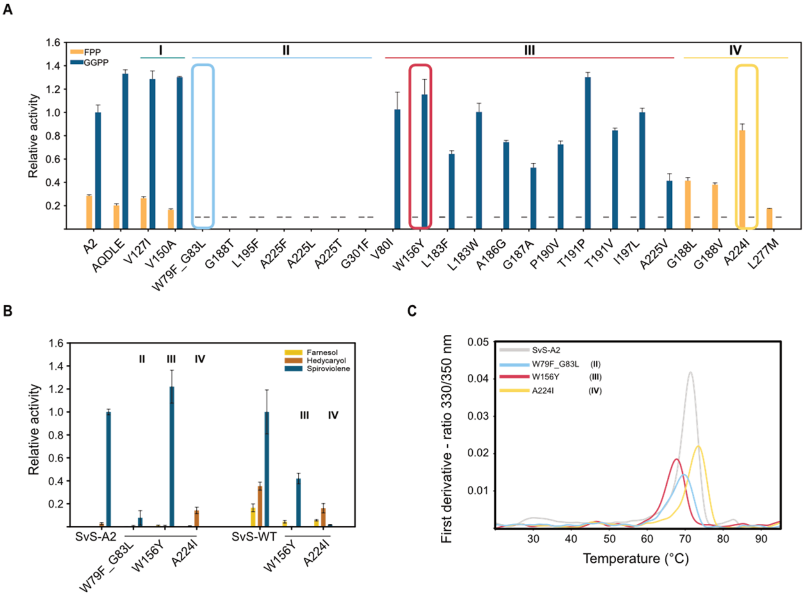


Figure 18. Engineered substrate specificity in SvS variants. (A) Activity of SvS-A2 variants with FPP or GGPP measured by Malachite Green assay using single substrates. Activity is shown relative to diterpene activity in SvS-A2. Dashes indicate values below the sensitivity threshold. **(B)** Product formation with FPP or GGPP in SvS-A2 variants and corresponding SvS-WT variants in a competition assay measured by GC-FID. Activities are given relative to SvS-A2 diterpene activity for SvS-A2 variants and relative to SvS-WT diterpene activity for SvS-WT variants. Error bars are from triplicates **(C)** Thermal unfolding of representative variants of SvS-A2 measured by nanoDSF. Figure modified from Schriever *et al.*[323] with permission by the journal.

The W156Y exchange (red box) for instance yielded a highly specific diterpene cyclase, likely due to the introduction of additional cation-π interactions with the first formed intermediate in the diterpene cyclisation trajectory. In contrast, the A224I exchange (yellow box) resulted in a highly specific sesquiterpene cyclase, which may be explained by its steric effect in proximity to the NSE motif, shielding off the larger diterpene substrate from the active site. Importantly, these two single residue exchanges could also be transferred to SvS-WT (Fig. 18B), resulting in the respective specificity trends. This observation highlights that results obtained from rational engineering of an ancestral protein can inform mutagenesis of its descendent protein to achieve a similar result in the WT background. Finally, thermostability of SvS-A2 was not strongly affected by these mutations (Fig. 18C). It is unlikely that the combination of residue exchanges conferring stability and specificity to SvS would have been immediately apparent from the extant enzyme itself. In summary these results showcase how an ancestral terpene cyclase serves as starting point for further structure guided rational engineering without compromising its thermostability.

6.3.3 Concluding remarks

In summary, the results presented in **paper III** represent an example of how an ancestral enzyme's stability rendered it more easily amenable to structural studies than its extant counterpart, as has been suggested in the literature.[138] The obtained structural information allowed the construction of a high quality homology model of extant SvS, for which no structural information was available, although further modification of the ancestral template resulted in improvement of the homology model's quality. Moreover, the obtained structure allowed the detailed analysis of interactions between the enzyme active site and different carbocation intermediates of the reaction mechanism towards both a diterpene and a sesquiterpene product, explaining which residues enable the enzyme's observed promiscuity. Finally, this structural knowledge was used to design enzyme variants with high specificity for either of the two substrates without compromising stability in the ancestral scaffold.

Paper III therefore demonstrates that reconstructed ancestors represent excellent starting points for further rational engineering of terpene cyclases. Due to their complex structure-activity-relationships and because many sequences but much fewer structures are available in databases, the approach of reconstructing a stable ancestor as basis for structure-guided engineering is particularly useful for this enzyme family. However, the workflow conceptually also applies to many other biocatalysts that catalyse highly complex reactions.

6.4 Paper IV – Metabolite interactions in the bacterial Calvin cycle and implications for flux regulation

6.4.1 Context within this thesis

One overarching aim of metabolic engineering is to obtain fast-growing organisms for industrial applications, such as CO_2-assimilating bacteria that can be tweaked to produce value compounds including biofuels and material precursors but typically grow slowly.[325] High flux through CO_2-assimilation pathways has direct implications for an autotrophic organism's growth rate and biomass accumulation. It is therefore of interest to understand and eventually manipulate the flux-bottlenecks of such pathways, *e.g.* by means of improving enzymes that catalyse rate-limiting reactions. Since the overexpression of F/SBPases holds promise for increasing growth of photosynthetic organisms (see **chapter 4.3.2**), this enzyme family represents a particularly interesting engineering target in this context.

As discussed in **chapter 2.5.3**, enzymes have not necessarily evolved to maximise rates since they are exposed to multiple evolutionary pressures simultaneously.[82] One such evolutionary constraint on kinetic parameters is the evolution of enzyme regulatory mechanisms, including among others the regulation by endogenous metabolites, which can endow enzymes with the ability to quickly react to sudden fluctuations in metabolism without modifying protein homeostasis.[326,327] If rational enzyme engineering is performed in the context of metabolic engineering, it is relevant to account for *in vivo* regulation in the enzyme design strategy. Understanding F/SBPase regulation in its native environment is thus prerequisite for integrated enzyme and metabolic engineering efforts involving this enzyme.

6.4.2 Investigation

The Calvin cycle is the most common of all carbon-assimilating metabolic pathways that have been described.[328,329] It consists of 13 enzymatic reactions (Fig. 19), which sum up to the following overall conversion (where GAP is glyceraldehyde 3-phosphate and P_i is inorganic phosphate).

$$3\ CO_2 + 6\ NADPH/H^+ + 9\ ATP + 5\ H_2O \rightarrow GAP + 6\ NADP^+ + 9\ ADP + 8\ P_i$$

In this paper, the regulation of bacterial Calvin cycle enzymes by selected metabolites was studied using interaction proteomics, *in vitro* studies and modelling of metabolic flux control coefficients. Specifically, similarities and differences between Calvin cycle regulation in each two model organisms that represent the domains of photoautotrophic bacteria (*Synechocystis* and *Synechococcus*) and chemolithoautotrophic bacteria (*C. necator* and *Hydrogenophaga pseudoflava*) were assessed. While all of these organisms use the Calvin cycle to assimilate CO_2, they differ in their energy source and microbial lifestyle which likely affects the regulatory metabolite pool. A particular focus was directed to the metabolite regulation of *syn*F/SBPase.

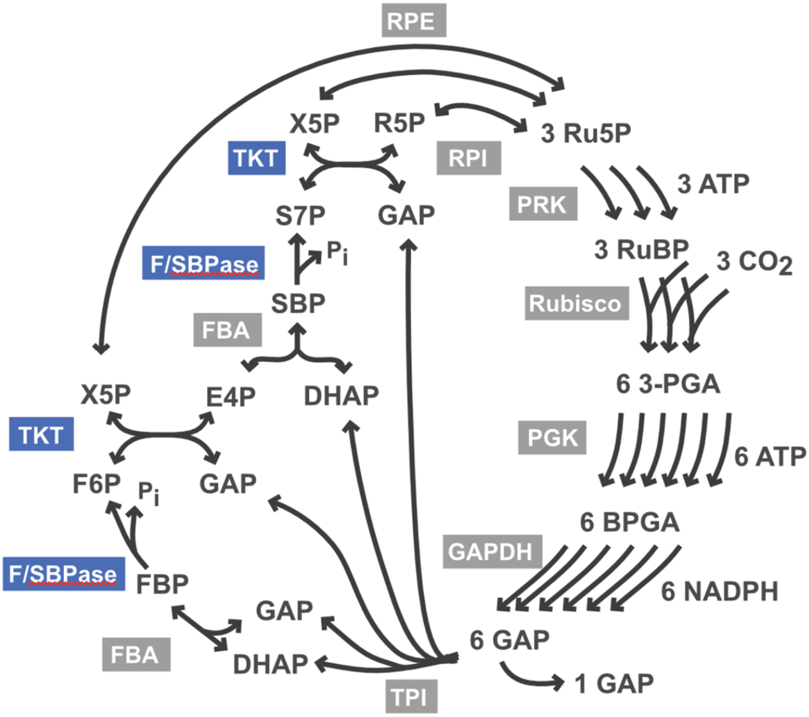


Figure 19. Schematic overview of the Calvin cycle. Figure from submitted manuscript,[330] prepared by E. P. Hudson.

In a first step, limited proteolysis small molecule mapping (LiP-SMap[331]) proteomics were used to uncover any potential interactions between a panel of metabolites (selected to represent energy, redox or metabolic status of the cell) and the proteome of the four bacterial strains. The method is based on the differential accessibility of proteinase K cleavage sites on proteins in absence and presence of an interacting metabolite, *e.g.* due to transient shielding of the digestion site by the metabolite or due to conformational changes occurring as a result of the interaction. Cell lysates were stripped of endogenous metabolites by gel filtration and subsequently incubated with the respective metabolites in presence of added Mg^{2+}. Although separate interaction proteomics experiments were performed using two different concentrations of each respective metabolite (typically 1 and 10 mM), a particular emphasis was put on those interactions observed at high metabolite concentrations, as they simulate spikes in metabolite concentrations occurring due to sudden metabolic shifts (*e.g.* changing light conditions for cyanobacteria). All interactions that were both significant and showed a change greater than two-fold were clustered according to the molecular function of the respective protein (based on KEGG orthology groups). A principal component analysis (PCA) performed per metabolite indicated whether the metabolite affected similar or different molecular functions in the four bacterial strains (data shown in appended manuscript), with a larger separation on the PCA plots indicating larger differences. GAP (one of the end products of the cycle) and Acetyl-CoA showed many interactions with proteins from all four species, yet displayed different interaction patterns between photo- and chemoautotrophs. Also, ATP, GTP and citrate showed multiple interactions with varying degrees of similarity between photoautotrophs and chemoautotrophs. These wide-spread interactions are likely not very specific and perhaps reflect the reactive aldehyde Nature of GAP, as well as Mg^{2+}- chelating properties of the other named metabolites,[332] since many proteins and enzymes require Mg^{2+} both for catalytic function and structural stability.

Next, any interactions that occurred specifically in the Calvin cycle or adjacent pathways were analysed per species (data shown in appended paper). Again, extensive interactions with GAP, Acetyl-CoA, ATP and to a somewhat lesser extent GTP and citrate were observed in many enzymes in all four species.

Most other metabolites in contrast, showed fewer and more species-specific interactions with Calvin cycle enzymes. F/SBPase in particular showed many interactions across the entire metabolite panel. Among these, many interactions occurred both somewhat enzyme-specific (meaning that the same metabolite didn't affect many other enzymes) as well as species-specific (meaning that the interaction didn't occur in all four bacterial strains).

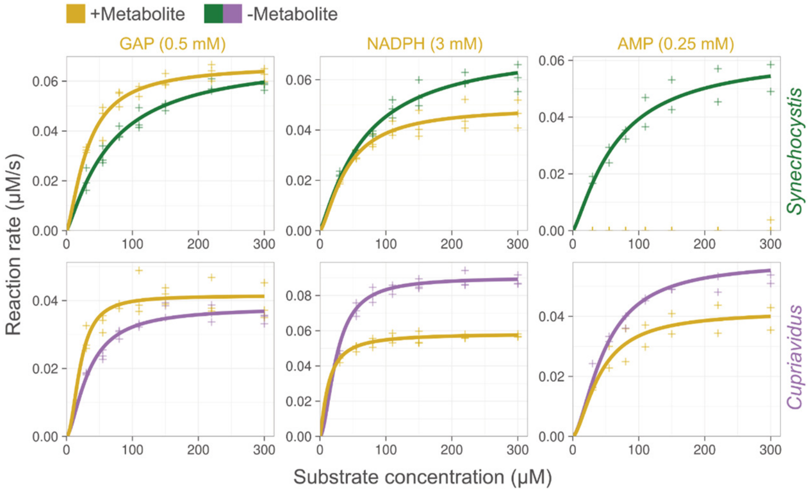


Figure 20. Kinetic assays of *syn*F/SBPase and *cn*F/SBPase in absence and presence of different metabolites. 0.42 ng uL^{-1} enzyme were incubated with different concentrations of FBP at 30 °C and initial rates were determined using a Malachite Green assay. Crosses represent data points, lines represent data fit to the Hill equation. Figure from submitted manuscript,[330] prepared by J. Karlsen.

Metabolite-enzyme interactions detected by interaction proteomics experiments don't automatically reflect a perturbation of enzyme function by the respective metabolite. Therefore, *in vitro* kinetic assays were performed in presence and absence of some of the identified interacting metabolites to validate above results for F/SBPase in one photoautotroph (*Synechocystis*) and one chemoautotroph (*C. necator*). GAP (at 0.5 mM) was identified as inducer and NADPH (at 3 mM) as inhibitor of F/SBPase from both organisms (Fig. 20). AMP was further included as a control, since it is well established that this metabolite strongly inhibits *syn*F/SBPase (see **chapter 4.3.3**) but its effects on *cn*F/SBPase were not previously known. We observed that *cn*F/SBPase was also inhibited by AMP, albeit to a much lower extent than *syn*F/SBPase (Fig. 20), which seems to support the hypothesis that *cn*F/SBPase does not perform gluconeogenic function in comparison to *syn*F/SBPase, as tight AMP regulation is associated with cytosolic gluconeogenic FBPases (see **chapter 4.3.1** for a detailed discussion).

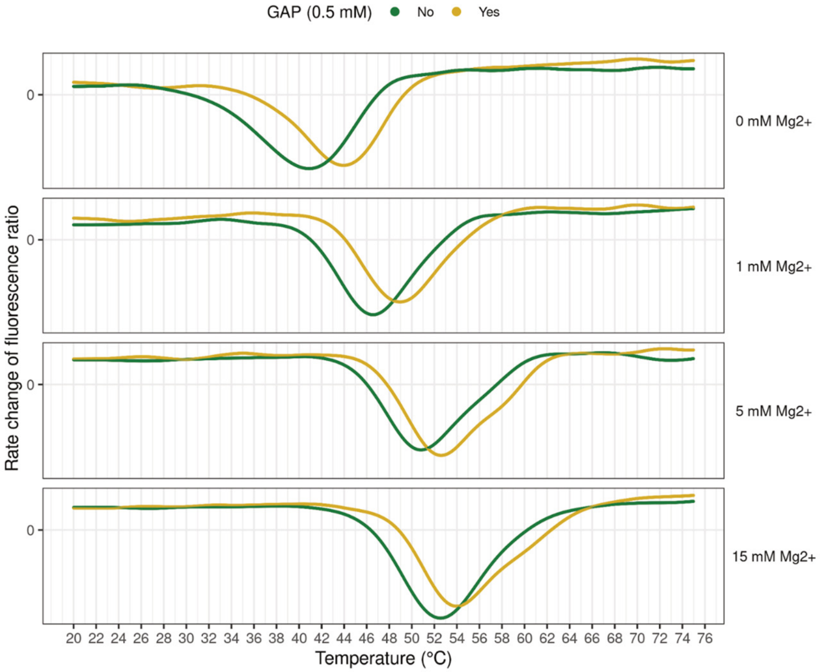


Figure 21. Thermal shift assay of *syn*F/SBPase. The enzyme was incubated with ± 0.5 mM GAP at 10 mM DTT and Mg^{2+} concentrations ranging from 0 to 15 mM, as indicated. Thermal unfolding was measured using nanoDSF. Figure from submitted manuscript,[330] prepared by J. Karlsen.

As discussed above, GAP is among the metabolites that produce multiple interactions in all species. In order to exclude that the observed GAP stimulation of *syn*F/SBPase is due to a Mg^{2+}-dependent effect (as discussed for *e.g.* ATP above), thermal shift assays of the enzyme were performed at different concentrations of Mg^{2+} in presence of DTT (Fig. 21). A significant shift in melting temperature was observed between *syn*F/SBPase when treated with GAP (yellow traces) and a buffer control (green traces), confirming that GAP binds to *syn*F/SBPase. The shift was not dependent on the Mg^{2+} concentration in the buffer, further confirming that the stimulating effect of GAP is not Mg^{2+}-related. Last but not least, these experiments revealed a gradual thermal shift of non-metabolite bound *syn*F/SBPase at increasing concentrations of Mg^{2+} (comparison of green traces across panels in Fig. 21), which likely represents a population distribution between two conformations. This observation is in line with analytical size exclusion chromatography experiments reported by Feng *et al.* that observed a new peak occurring upon addition of Mg^{2+} to the enzyme.[248] It is therefore likely that metabolites that compete with the enzyme for Mg^{2+}-binding (such as ATP and Acetyl-CoA) indirectly change the enzyme's conformation to an unproductive state, which is why they occur as Lip-SMap hits.

Since *syn*F/SBPase is strongly regulated by both AMP and redox state (see **chapter 4.3.3**), we were interested in studying whether the inductive GAP effect acted in synergy with redox regulation. Thermal aggregation data (Fig. 22A) revealed two distinctly scattering protein species in oxidised *syn*F/SBPase (main peak and peak shoulder in red trace), the second of which disappeared in presence of DTT (green trace). Moreover the apex of the peak moved between these two extremes at intermediate DTT concentrations, indicating a population distribution between the two states dependent on the redox state. When pre-incubating *syn*F/SBPase at 30 °C (the temperature used in kinetic assays) for different durations before measurement, an increase of the peak shoulder over time could be observed (Fig. 22B). The time-dependent increase likely reflects protein aggregation (alternatively a change in oligomeric state), which would be expected upon a gradual accumulation of unspecific disulphide bridges within and across protein monomers and molecules. This time-dependent increase of the peak was further amplified in presence of GAP (Fig. 22C), indicating that GAP may accelerate aggregation in oxidative conditions. In line with this observation, GAP was found to inhibit *syn*F/SBPase activity compared to metabolite-free enzyme in absence of DTT (data shown in appended paper). In summary, kinetic and thermal aggregation data indicate that redox regulation and GAP regulation of *syn*F/SBPase act in synergy with each other.

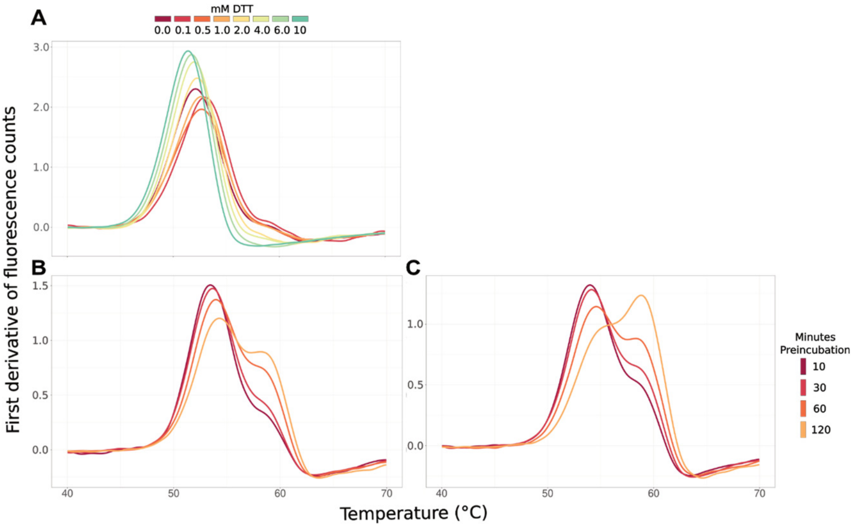


Figure 22. Thermal aggregation assays of *syn*F/SBPase. The enzyme was incubated **(A)** with DTT concentrations varying from 0 to 10 mM **(B)** for indicated pre-incubation times at 30 °C in absence of DTT and **(C)** for indicated pre-incubation times at 30 °C in absence of DTT and presence of 0.5 mM GAP. Thermal aggregation was measured using backscattering in nanoDSF. Figure modified from submitted manuscript,[330] originally prepared in collaboration with E. Sporre.

Since a reductive environment signals an abundance of light (*i.e.* energy availability), it seems plausible that *syn*F/SBPase, an enzyme that controls Calvin cycle flux to a certain degree (see appended paper for details), may be regulated by GAP (an end product of the Calvin cycle) by a feed-forward type regulation to enhance the rate of CO_2 assimilation. The experimental results suggest that if the light conditions change rapidly, *syn*F/SBPase would be oxidised and GAP may consequently synergistically inhibit the enzyme to rapidly slow down the cycle during an energy shortage. Due to the reactive aldehyde nature of GAP, it is possible that such interactions could be mediated by post-translational modification of nucleophilic amino acids.[333,334] Additional experiments are required to confirm this hypothesis and assess both the exact molecular mechanism of GAP-redox synergy as well as its magnitude *in vivo*. Moreover, the described metabolite regulations were only measured for the enzyme's FBPase reaction. Additional kinetic experiments would further be required to assess if observed metabolite effects would apply equally to the enzyme's SBPase activity. Nevertheless, findings from thermal unfolding and aggregation assays should also apply to the SBPase activity. For discussion of additional results the reader is referred to the appended manuscript.

6.4.3 Concluding remarks

The results discussed in **paper IV** indicate that the Calvin cycle is subject to multiple layers of metabolite regulation in both photoautotrohic and chemoautotrohic organisms. In this thesis a particular emphasis was put on the analysis of *syn*F/SBPase regulation by metabolites. NADPH and GAP were identified as novel inhibitor and inducer of the enzyme, respectively. GAP regulation was observed to occur synergistically with redox regulation, suggesting that the enzyme's *in vivo* activity is delicately fine-tuned by different metabolites. This finding underscores how multifaceted and complex enzyme regulation is expected to be in crowded cell environments. When trying to alleviate bottle-necks in Calvin cycle flux by enzyme engineering, it is crucial to take such interactions into account. The peptide-level resolution of changes measured by the LiP-SMap approach in this study may give a first idea of where metabolites may interact, although further structural studies are required to confirm such interactions and allow for their rational engineering.

7 Concluding Remarks and Outlook

The work presented in this thesis aims to apply principles from different disciplines of protein and enzyme engineering to contexts that are relevant in the greater frame of health and biotechnology applications, such as the generation of vaccines, understanding principles of human steroid metabolism for inhibitor design, improving specificity and activity in promiscuous biosynthetic enzymes and improving microbial hosts for CO_2-neutral production of value compounds. Several different protein engineering approaches, that each entail unique advantages and disadvantages, have been combined in the different projects of this thesis. One aspect that is challenging to address in many engineering approaches are long range interactions. Understanding how residues in the protein periphery affect active site conformation, electrostatics and dynamics as well as protein folding is challenging and predicting how the introduction of mutations impact such complicated networks even more so. The application of ancestral sequence reconstruction in an engineering context bridges the two domains of rational and agnostic engineering. The method starts from sequence input and in principle does not require knowledge about a protein's structure or function. However, it relies on rational phylogenetic input. In a way, ASR describes an agnostic engineering approach that focuses the experimentally sampled sequence-space onto proteins that are assumed to have been previously sampled by evolution and are therefore likely folded and somewhat functional. Epistatic networks are not expected to be disturbed as much since ASR takes co-evolution of interdependent protein residues into account.

Inspired by studies that used ASR for generation of stable and functional therapeutic proteins and the prospect of stabilisation independent of structural information, we capitalised on this method for generating sarbecovirus S protein antigens. Highly stable and soluble S proteins could be obtained in high yields without using structural knowledge and by testing only four variants experimentally. These variants occupied an all-RBD-down pre-fusion structure, as confirmed by high resolution cryo-EM and appeared to display conserved epitopes that could be recognised by antibodies elicited against SARS-CoV-2 S protein. Moreover, these proteins served as scaffolds to host the receptor binding domain of the SARS-CoV-2 S protein, highlighting their compatibility with protein domains from extant evolved viruses and their potential utility as immunogen scaffolds that can be adapted to emerging viruses and virus variants. In the future, it would be interesting to study *in vivo* stability of the antigens in animal studies and to see if they would be able to elicit broadly protective antibodies against different SARS-CoV-2 strains as well as other sarbecoviruses when used as vaccine candidates. Application of this approach to other virus families would further shed light on the universality of this approach. Since antigens constructed by ASR comprise sequences from several related virus strains, they are likely to display conserved epitopes that are shared among the group of extant viruses and moreover be compatible to host grafted extant epitopes from existing and emerging viruses of the same clade. Constructing a vaccine based on such an antigen may therefore potentially elicit cross-protective antibody responses against related viruses and be adaptable to specific strains if needed. If proven effective, the described approach would be a powerful tool to develop vaccines for infectious diseases against which there currently are no effective vaccines available (*e.g.* the Marburg virus) or as a targeted boosting strategy for viruses that necessitate frequent vaccine updates due to viral mutations (*e.g.* Influenza). Finally, future studies may investigate if ancestral antigens can be used for focusing sub-dominant B-cell repertoires from a heterologous primer (*e.g.* vaccination or infection) when used directly or when used as scaffold for grafted epitopes of interest.

One aspect of ancestral proteins that is frequently investigated is the change in thermal protein stability and related change in catalytic temperature optimum compared to extant proteins. In the next study, we focused on understanding the temperature dependence of oxidosqualene cyclase - an important enzyme in human cholesterol metabolism. Since cholesterol is crucial in regulating membrane homeostasis at fluctuating temperatures, it is conceivable that the temperature dependence of enzymes involved in its biosynthesis evolved to adapt to the organism's environmental temperature. For instance, a thermophilic bacterial homologue of OSC exhibits optimal catalytic activity at 55 °C *in vitro*, whereas OSC exhibits optimal catalytic activity around 37 °C. Evolutionary rates are expected to be greater on the periphery, especially in the case of enzymes that have highly complex pre-arranged active sites such as hOSC. We therefore studied in which way mutations adjacent to the active site as well as the protein surface of OSC impact the enzyme's catalytic temperature dependence. We found that OSC rate enhancement is mediated by favourable activation entropy, which was severely impacted by obstructing solvent access tunnels to the active site by mutagenesis. In fact, the effect of such an obstruction by a single amino acid exchange led to reversal of temperature dependence, such that cold temperatures were preferred. Moreover, additional mutations on the surface of the enzyme further modulated temperature dependence, highlighting the importance of long-range effects and epistasis in regulating enzyme temperature dependence. The observation may have implications for novel drug design approaches to target this enzyme, which is deregulated in some cardiovascular diseases. It would further be interesting to study whether the modification of activation entropies indeed represents an evolutionary adaptation to different environmental temperatures in this enzyme family by studying activation entropy and tunnel distributions along reconstructed evolutionary trajectories using ASR. Since OSC (and the related SHC) are not the only enzymes described to date that operate at positive activation entropies it would further be interesting to gauge if the impact of solvent access tunnels on activation entropy if specific to the OSC/SHC family or more widespread. Lastly, it would be interesting to apply the same engineering approach to related plant triterpene cyclases that produce compounds of industrial interest in order to adapt them for applications that operate at different temperatures than the enzyme's temperature optimum (*e.g.* expression in other plants, cell factories or *in vitro*).

Since highly stable and soluble proteins are considered to be more easily evolvable, we next explored the use of ASR as a means to generate a starting point for further engineering. Such approaches have been discussed in the literature with the concept of "re-booting" evolution from an optimised starting point towards a desired outcome. In this context, we were interested in combining ASR with rational structure-guided engineering. This is particularly useful for the enzyme family of terpene cyclases (the same enzyme family that hOSC belongs to), which are characterised by highly complex active site architectures. We obtained a crystal structure of a stable reconstructed ancestor of a bacterial class I terpene cyclase to understand how the enzyme and its ancestor both chaperone two different highly complex cyclisation reactions in their active sites. Based on the generated molecular insight about substrate promiscuity, two single-residue exchanges could be identified in the active site that defined the enzyme's preference for either of the substrates. Importantly, these single-residue switches could be transferred into the sequence of the extant enzyme, resulting in the same selectivity trend. In applications, in which proteins are required to deviate as little as possible from an extant starting point (*e.g.* due to immunogenicity concerns in protein therapeutics) an easy-to-handle reconstructed ancestor may serve as a suitable proxy to pinpoint mechanistic details of individual mutations, which can then be transferred back to the extant system. The results highlight that ASR is particularly powerful at generating starting points for further engineering, rather than simply constituting the end-point of an engineering effort. This has implications for the optimisation of biosynthetic enzymes with complex reaction mechanisms to desired tasks, *e.g.* the production of value compounds such as terpenes in transgenic organisms.

Autotrophic bacteria represent particularly attractive hosts for this purpose, given the prospect of CO_2-neutral production of terpenes and other compounds that can be used as biofuels, material precursors and pharmaceutical compounds, all of which are currently mostly sourced from crude oil. This however requires engineering both the metabolism as well as optimising growth rates of slowly growing autotrophic bacteria. With that in mind, we turned our attention to understanding the regulation of a flux-controlling enzyme in the Calvin cycle of *Synechocystis* in the last study of this thesis. Using a combination of interaction proteomics and *in vitro* studies, we found that *syn*F/SBPase in reductive conditions was induced by GAP, an end-product of the Calvin cycle, indicating a feed-forward type regulation. Moreover, we found that GAP inhibited the same enzyme in non-reductive conditions, likely *via* an aggregation-mediated mechanism. In the future, it would be interesting to elucidate the exact molecular mechanism of *syn*F/SBPase regulation by DTT and GAP and to gauge its physiological relevance in the WT as well as engineered strains of *Synechocystis*. One approach could be to unambiguously identify and mutate GAP interaction sites in the enzyme. Since this enzyme family represents a somewhat unique regulatory niche, it would also be interesting to study the evolutionary history of metabolite regulation in this bacterial sub-clade.

In summary, the work presented in this thesis highlights the utility of combining different strategies when engineering proteins and enzymes for different applications. **Paper I** and **III** show that ancestral sequence reconstruction represents a highly useful tool to obtain stable proteins, for which structures can be solved and which can serve as starting point for further optimisation. Ancestral mutations were mostly found dispersed on the protein surface in both studies, which represents a mutation type that would be challenging to introduce to the same extent by structure-guided rational engineering. The combination of ASR as an approach to modulate protein stability *via* the protein surface and subsequent structure-guided engineering to narrow down on specificity represents a promising workflow that should be applicable to multiple other proteins in health and biotechnology contexts. **Paper II** presents the idea of rationally modifying the enzyme periphery to modulate thermodynamic parameters of activation, which could potentially be used in conjunction with ASR in the future. **Paper IV** finally highlights that protein regulation is highly complex and needs to be taken into account for integrated enzyme and metabolic engineering approaches.

Acknowledgements

Science is always a team effort and I want to thank so many people that have contributed to my research and education in some way or another.

The funding of the work presented in this thesis by the **Swedish Research Council (VR)** is greatly acknowledged. I would also like to thank the **German Chemical Society** and **ÅForsk Foundation** for financially supporting my attendance at international conferences and conferences and the **International Union of Biochemistry and Molecular Biology** for granting financial support for a research stay abroad. I would also like to express my gratitude to **Per-Olof** for the continuous review of the entire dissertation as it evolved and to **Natalie** and **Paul** for their valuable input, as well as to **Mats Martinell** for reviewing the dissertation and his feedback.

First of all, thank you **Per-Olof** for your constant support and encouragement throughout the past years. Thank you for letting me explore new ideas and projects freely, while providing valuable scientific guidance, feedback and suggestions. And thank you for giving me the opportunity to work on all these fun projects in such a collaborative working environment.

Thank you **Paul** for being such an involved co-supervisor and the continuous collaboration throughout the past years. Thank you for "adopting" me as an extended group member into the Cyano-group, always offering me another perspective on my projects and countless spontaneous scientific discussions in the lab.

To all current and previous members of the **PEOZ group** – thank you for the amazing time together! **David**, you have been my closest collaborator throughout these years and I want to thank you for all the team-work, shared ups and downs of the PhD and our projects, laughter and late labnight Foodora dinners. And for putting on good music in the lab, of course. **Ximena**, thank you for always being so caring and helpful and the uncountable "shall-we-just-take-a-quick-coffee-break?"-breaks. I am glad that you joined our group and brought so much energy and laughter with you. And thank you for taking on the co-responsibility for our little GC-baby with so much optimism and positivity. Thank you **Antonino,** for guiding me on my first steps at SciLifeLab, the fun collaborations, for being the most social person I know and for initiating so many fun projects and events. Thank you **Natalie,** for being such a caring and supportive person, role-model and friend. I am glad we got to collaborate on the SvS project and had a great time together with you in Whistler. I hope we will continue our virtual fika-breaks and real-life meet-ups in the future! Thank you **Sissi**, for your scientific advice, always being eager to help but also for shared laughter and all the long conversations about science, academia and life in general. **Elisabeth**, thank you for your positive attitude and humour and your initiative to organise group get-togethers. Thank you, **Sara E.**, for all the fun time and laughter we spent together during the first year of my PhD and always making sure there was enough snacks around to survive a long lab-day. Thank you to **Sudarsan**, for always sharing his knowledge and to **Patricia** for the collaboration on the SvS-project. And thank you to **Caroline**, **Arne**, **Tamás**, **Nasim**, **Heba**, **Wissam** and **Ranjani** for the fun time spent together at some point or another in the past years. Thank you also to all students in the group and in particular to my students **Johan**, **Philip** and **Albert**.

Thank you to all current and previous members of the **Cyano-group**, **Jonas group** and other groups on the floor for making **gamma-5** the warmest and most inviting workplace I could have ever asked for. I really cherish the lunchtime conversations on the most random topics, squeezed together at the two tables in the tiny kitchen. Thank you **Jan**, for all the great times, hikes and trips, teaching us non-Swedes about ice hockey, the collaboration on the LipS-Map project and for taking the stress out of a busy day in the lab. And of course thank you for introducing me to the choir. A big thank you to **Markus**, for his advice, being so caring, such a good listener and helping me get started in the group and for initiating activities such as hikes and museum trips. **Ivana**, I can't tell you how grateful I am for all your advice throughout the years. Your general scientific as well as practical lab-knowledge and creative lab-hacks are beyond comparison and you have helped me countless times, when I was stuck. Thank you for being a great listener and for often sharing the less crowded lab-shifts during the lock-down. And of course, thank you for teaching me almost everything that I know about the GC and for being so patient with me while learning. **Emil**, thank you for the close collaboration on the LipS-Map and FUREE projects. I really enjoyed working with you and the fun times during our pipetting-workout kinetic assays. Thank you for always being helpful and supportive. Thank you **Olivia**, for sharing the first months of my PhD journey with me, for all the walks in Råsunda during the pandemic and for being such a lovely and creative person and friend. Thank you **Nick**, for being so caring, brightening everyone's day with your humour and for taking so good care of lab and candy-corner supplies and to **Rui** for always being so helpful and knowledgeable. Thank you **Matthias**, for always being in a good mood and for all interesting discussions about proteins, academia and life in general. I am glad we got to share the last meters of our PhD journeys while writing up our dissertations. **Deike**, thank you for the fun conversations, being so caring and a 'Fels in der Brandung' on gamma-5. **Kiyan**, thank you for your legendary Friday-lab Spotify list. Also thank you to all the other members of the two groups for the fun time and scientific discussions that we shared. You all contributed to the working environment and my PhD experience in some way or another.

Thank you to all collaborators and co-authors that I got a chance to work with throughout these years. I am especially grateful to **Juni**, for sharing her wisdom on protein science, mammalian cell culture, and cryo-EM. Thank you for the fun collaboration and all shared laughter and enthusiasm. And thank you for sharing my passion for (or rather obsession with) stationery and unusual but practical lab-consumables. I am also particularly grateful to **Robert Schnell** for sharing his advice on proteins and enzymes during the collaboration on the SvS project. Thank you to **Anders Olsson**, **Camilla Hofström** and other members of the **DDDP** for generously letting me use lab equipment and for your valuable advice on proteins and protein biophysics.

Thank you to all other PhD students, PostDocs and PIs at SciLifeLab that have made my time here so memorable. Thank you to **Alba, Dörte, Marianna, Marco, Aswathy**, **Vaishnovi, Marcel, Eva** and others for the teamwork in the PhD and PostDoc council. A special thank you to **Alba** and **Dörte** for organising the first SciLifeLab minisymposium together – I learnt a lot with and from you in the process and always enjoy catching up with you.

I am also very grateful to **Dana Reichmann** and her group for hosting me at HUJI in winter 2021, sharing their expertise, being so welcoming and making sure I had a great stay. In particular would like to thank **Tal** for sharing his knowledge on HDX-MS, recommending the best coffee-houses in Jerusalem and for being so patient and fun to work with. Also thank you to **Eitan Lerner** and **Paul** for introducing me to the world of single molecule fluorescence spectroscopy.

Thank you to the members of the **Coating Division** for always welcoming us SciLifeLabers over to the KTH campus. Thank you **Eva**, for being supportive and interested in my research and **Linda**, for always being so friendly and positive. A super big Grazie to **Rosella** for welcoming me in Stockholm and for all the wonderful trips we shared to the North of Sweden and exploring the Stockholm surroundings together, no matter the weather. And thank you for your tips on printing the thesis, I am so glad we shared this last step of the PhD journey together.

To my friends in science or outside – thank you for your support throughout the last years! I am very glad that I got to share the PhD experience with some of you and learn from you. **Marlena**, thank you for the daily WhatsApp pep talks during our shared writing-phase.

And most importantly, thank you to my family. Thank you to my wonderful **Herde** for the constant support, encouragement, inspiration and love. You have been with me steps of the way. **Ludvig**, thank you for simply everything. Your immeasurable patience and support in the last months has been my anchor. I am looking forward to our next adventures together.

Bibliography

1. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* **181**, 223-30 (1973).
2. Nicholls, A., Sharp, K.A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281-96 (1991).
3. Oldfield, C.J. & Dunker, A.K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**, 553-84 (2014).
4. Brockwell, D.J. & Radford, S.E. Intermediates: ubiquitous species on folding energy landscapes? *Curr Opin Struct Biol* **17**, 30-7 (2007).
5. Dobson, C.M. Protein folding and misfolding. *Nature* **426**, 884-90 (2003).
6. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A.R. Strength and co-operativity of contributions of surface salt bridges to protein stability. *J Mol Biol* **216**, 1031-44 (1990).
7. Albeck, S., Unger, R. & Schreiber, G. Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *J Mol Biol* **298**, 503-20 (2000).
8. Prajapati, R.S., Sirajuddin, M., Durani, V., Sreeramulu, S. & Varadarajan, R. Contribution of Cation−π Interactions to Protein Stability. *Biochemistry* **45**, 15000-15010 (2006).
9. Pollard, T.D., Earnshaw, W.C., Lippincott-Schwartz, J. & Johnson, G.T. *Cell Biology*, (Elsevier, Philadelphia, Pennsylvania, 2017).
10. Branden, C.I. & Tooze, J. *Introduction to protein structure*, (Garland Science, New York, 2012).
11. Teufl, M., Zajc, C.U. & Traxlmayr, M.W. Engineering Strategies to Overcome the Stability–Function Trade-Off in Proteins. *ACS Synthetic Biology* **11**, 1030-1039 (2022).
12. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-8 (2003).
13. Soares, T.A., Boschek, C.B., Apiyo, D., Baird, C. & Straatsma, T.P. Molecular basis of the structural stability of a Top7-based scaffold at extreme pH and temperature conditions. *J Mol Graph Model* **28**, 755-65 (2010).
14. Goldenzweig, A. & Fleishman, S.J. Principles of Protein Stability and Their Application in Computational Design. *Annu Rev Biochem* **87**, 105-129 (2018).
15. Bershtein, S., Mu, W., Serohijos, A.W., Zhou, J. & Shakhnovich, E.I. Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Mol Cell* **49**, 133-44 (2013).
16. Deller, M.C., Kong, L. & Rupp, B. Protein stability: a crystallographer's perspective. *Acta Crystallogr F Struct Biol Commun* **72**, 72-95 (2016).
17. Bloom, J.D., Labthavikul, S.T., Otey, C.R. & Arnold, F.H. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **103**, 5869-74 (2006).
18. Boucher, J.I., Bolon, D.N. & Tawfik, D.S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Sci* **25**, 1219-26 (2016).
19. Blanco, A. & Blanco, G. Chapter 14 - Carbohydrate Metabolism. in *Medical Biochemistry* (eds. Blanco, A. & Blanco, G.) 283-323 (Academic Press, 2017).
20. Trentham, D.R., Eccleston, J.F. & Bagshaw, C.R. Kinetic analysis of ATPase mechanisms. *Q Rev Biophys* **9**, 217-81 (1976).
21. Fersht, A. *Structure and Mechanism in Protein Science*, (World Scientific Publishing Co. Pte. Ltd. , Singapore).
22. Edwards, D.R., Lohman, D.C. & Wolfenden, R. Catalytic Proficiency: The Extreme Case of S–O Cleaving Sulfatases. *Journal of the American Chemical Society* **134**, 525-531 (2012).
23. Briggs, G.E. & Haldane, J.B. A Note on the Kinetics of Enzyme Action. *Biochem J* **19**, 338-9 (1925).
24. Michaelis, L., Menten, M.L., Johnson, K.A. & Goody, R.S. The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry* **50**, 8264-9 (2011).
25. Pauling, L. Molecular architecture and biological reactions. *Chemical and engineering news* **24**, 1375-1377 (1946).
26. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **27**, 2985-2993 (1894).
27. Koshland, D.E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci U S A* **44**, 98-104 (1958).

28. Gianni, S., Dogan, J. & Jemth, P. Distinguishing induced fit from conformational selection. *Biophys Chem* **189**, 33-9 (2014).

29. Changeux, J.P. & Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol Rep* **3**, 19 (2011).

30. Weikl, T.R. & Paul, F. Conformational selection in protein binding and function. *Protein Sci* **23**, 1508-18 (2014).

31. Guo, B. et al. Conformational Selection in Biocatalytic Plastic Degradation by PETase. *ACS Catalysis* **12**, 3397-3409 (2022).

32. Ribeiro, A.J.M., Tyzack, J.D., Borkakoti, N., Holliday, G.L. & Thornton, J.M. A global analysis of function and conservation of catalytic residues in enzymes. *J Biol Chem* **295**, 314-324 (2020).

33. Ribeiro, A.J.M. et al. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res* **46**, D618-d623 (2018).

34. Mertz, E.L. & Krishtalik, L.I. Low dielectric response in enzyme active site. *Proc Natl Acad Sci U S A* **97**, 2081-6 (2000).

35. Archer, D.G. & Wang, P. The dielectric constant of water and Debye-Hückel limiting law slopes. *Journal of physical and chemical reference data* **19**, 371-411 (1990).

36. Fried, S.D. & Boxer, S.G. Electric Fields and Enzyme Catalysis. *Annu Rev Biochem* **86**, 387-415 (2017).

37. Ronnebaum, T.A., Eaton, S.A., Brackhahn, E.A.E. & Christianson, D.W. Engineering the Prenyltransferase Domain of a Bifunctional Assembly-Line Terpene Synthase. *Biochemistry* **60**, 3162-3172 (2021).

38. Fox, J.M., Zhao, M., Fink, M.J., Kang, K. & Whitesides, G.M. The Molecular Origin of Enthalpy/Entropy Compensation in Biomolecular Recognition. *Annu Rev Biophys* **47**, 223-250 (2018).

39. Warshel, A. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J Biol Chem* **273**, 27035-8 (1998).

40. Warshel, A. et al. Electrostatic Basis for Enzyme Catalysis. *Chemical Reviews* **106**, 3210-3235 (2006).

41. Wagner, A. *Robustness and evolvability in living systems*, (Princeton university press, 2013).

42. Wolfenden, R. & Snider, M.J. The depth of chemical time and the power of enzymes as catalysts. *Acc Chem Res* **34**, 938-45 (2001).

43. Stockbridge, R.B., Lewis, C.A., Jr., Yuan, Y. & Wolfenden, R. Impact of temperature on the time required for the establishment of primordial biochemistry, and for the evolution of enzymes. *Proc Natl Acad Sci U S A* **107**, 22102-5 (2010).

44. Wolfenden, R. Primordial chemistry and enzyme evolution in a hot environment. *Cell Mol Life Sci* **71**, 2909-15 (2014).

45. Siddiqui, K.S. & Cavicchioli, R. Cold-adapted enzymes. *Annu Rev Biochem* **75**, 403-33 (2006).

46. Åqvist, J., Kazemi, M., Isaksen, G.V. & Brandsdal, B.O. Entropy and Enzyme Catalysis. *Accounts of Chemical Research* **50**, 199-207 (2017).

47. Bjelic, S., Brandsdal, B.O. & Åqvist, J. Cold Adaptation of Enzyme Reaction Rates. *Biochemistry* **47**, 10049-10057 (2008).

48. Isaksen, G.V., Åqvist, J. & Brandsdal, B.O. Protein surface softness is the origin of enzyme cold-adaptation of trypsin. *PLoS Comput Biol* **10**, e1003813 (2014).

49. Snider, M.J., Gaunitz, S., Ridgway, C., Short, S.A. & Wolfenden, R. Temperature effects on the catalytic efficiency, rate enhancement, and transition state affinity of cytidine deaminase, and the thermodynamic consequences for catalysis of removing a substrate "anchor". *Biochemistry* **39**, 9746-53 (2000).

50. Syrén, P.O., Hammer, S.C., Claasen, B. & Hauer, B. Entropy is key to the formation of pentacyclic terpenoids by enzyme-catalyzed polycyclization. *Angew Chem Int Ed Engl* **53**, 4845-9 (2014).

51. Åqvist, J. & Kamerlin, S.C. Exceptionally large entropy contributions enable the high rates of GTP hydrolysis on the ribosome. *Sci Rep* **5**, 15817 (2015).

52. Sievers, A., Beringer, M., Rodnina, M.V. & Wolfenden, R. The ribosome as an entropy trap. *Proc Natl Acad Sci U S A* **101**, 7897-901 (2004).

53. Jencks, W.P. Binding energy, specificity, and enzymic catalysis: the circe effect. *Adv Enzymol Relat Areas Mol Biol* **43**, 219-410 (1975).

54. Villa, J. et al. How important are entropic contributions to enzyme catalysis? *Proc Natl Acad Sci U S A* **97**, 11899-904 (2000).

55. Warshel, A., Florián, J., Strajbl, M. & Villà, J. Circe effect versus enzyme preorganization: what can be learned from the structure of the most proficient enzyme? *Chembiochem* **2**, 109-11 (2001).

56. Shurki, A., Štrajbl, M., Villà, J. & Warshel, A. How Much Do Enzymes Really Gain by Restraining Their Reacting Fragments? *Journal of the American Chemical Society* **124**, 4097-4107 (2002).

57. Kazemi, M., Himo, F. & Åqvist, J. Enzyme catalysis by entropy without Circe effect. *Proc Natl Acad Sci U S A* **113**, 2406-11 (2016).

58. Mulholland, A.J. Dispelling the effects of a sorceress in enzyme catalysis. *Proc Natl Acad Sci U S A* **113**, 2328-30 (2016).

59. Trobro, S. & Aqvist, J. Mechanism of peptide bond synthesis on the ribosome. *Proc Natl Acad Sci U S A* **102**, 12395-400 (2005).

60. Ycas, M. On earlier states of the biochemical system. *J Theor Biol* **44**, 145-60 (1974).

61. Jensen, R.A. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* **30**, 409-25 (1976).

62. Khersonsky, O. & Tawfik, D.S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* **79**, 471-505 (2010).

63. Pandya, C., Farelli, J.D., Dunaway-Mariano, D. & Allen, K.N. Enzyme promiscuity: engine of evolutionary innovation. *J Biol Chem* **289**, 30229-30236 (2014).

64. Copley, S.D. An evolutionary biochemist's perspective on promiscuity. *Trends Biochem Sci* **40**, 72-8 (2015).

65. Copley, S.D. Shining a light on enzyme promiscuity. *Curr Opin Struct Biol* **47**, 167-175 (2017).

66. Huang, H. et al. Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc Natl Acad Sci U S A* **112**, E1974-83 (2015).

67. Velez Rueda, A.J., Palopoli, N., Zacarías, M., Sommese, L.M. & Parisi, G. ProtMiscuity: a database of promiscuous proteins. *Database (Oxford)* **2019**(2019).

68. Bar-Even, A., Milo, R., Noor, E. & Tawfik, D.S. The Moderately Efficient Enzyme: Futile Encounters and Enzyme Floppiness. *Biochemistry* **54**, 4969-77 (2015).

69. Khersonsky, O. & Tawfik, D.S. Structure-reactivity studies of serum paraoxonase PON1 suggest that its native activity is lactonase. *Biochemistry* **44**, 6371-82 (2005).

70. Newton, M.S., Arcus, V.L., Gerth, M.L. & Patrick, W.M. Enzyme evolution: innovation is easy, optimization is complicated. *Curr Opin Struct Biol* **48**, 110-116 (2018).

71. Jayaraman, V., Toledo-Patiño, S., Noda-García, L. & Laurino, P. Mechanisms of protein evolution. *Protein Sci* **31**, e4362 (2022).

72. Roth, C. et al. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol* **308**, 58-73 (2007).

73. Näsvall, J., Sun, L., Roth, J.R. & Andersson, D.I. Real-time evolution of new genes by innovation, amplification, and divergence. *Science* **338**, 384-7 (2012).

74. Guo, H.H., Choe, J. & Loeb, L.A. Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences* **101**, 9205-9210 (2004).

75. Tawfik, D.S. Messy biology and the origins of evolutionary innovations. *Nat Chem Biol* **6**, 692-6 (2010).

76. Dellus-Gur, E. et al. Negative Epistasis and Evolvability in TEM-1 β-Lactamase--The Thin Line between an Enzyme's Conformational Freedom and Disorder. *J Mol Biol* **427**, 2396-409 (2015).

77. Tóth-Petróczy, A. & Tawfik, D.S. Slow protein evolutionary rates are dictated by surface-core association. *Proc Natl Acad Sci U S A* **108**, 11151-6 (2011).

78. Levy, E.D., De, S. & Teichmann, S.A. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci U S A* **109**, 20461-6 (2012).

79. Wyganowski, K.T., Kaltenbach, M. & Tokuriki, N. GroEL/ES buffering and compensatory mutations promote protein evolution by stabilizing folding intermediates. *J Mol Biol* **425**, 3403-14 (2013).

80. Koenig, S.H. & Brown, R.D., 3rd. H 2 CO 3 as substrate for carbonic anhydrase in the dehydration of HCO 3. *Proc Natl Acad Sci U S A* **69**, 2422-5 (1972).

81. Hasinoff, B.B. Kinetics of carbonic anhydrase catalysis in solvents of increased viscosity: a partially diffusion-controlled reaction. *Arch Biochem Biophys* **233**, 676-81 (1984).

82. Bar-Even, A. et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402-10 (2011).

83. Newton, M.S. et al. Structural and functional innovations in the real-time evolution of new (βα)(8) barrel enzymes. *Proc Natl Acad Sci U S A* **114**, 4727-4732 (2017).

84. Klesmith, J.R., Bacik, J.P., Michalczyk, R. & Whitehead, T.A. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth Biol* **4**, 1235-43 (2015).

85. Hartl, D.L., Dykhuizen, D.E. & Dean, A.M. Limits of adaptation: the evolution of selective neutrality. *Genetics* **111**, 655-74 (1985).

86. Agozzino, L. & Dill, K.A. Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proc Natl Acad Sci U S A* **115**, 9092-9097 (2018).

87. Levy, E.D. & Teichmann, S. Structural, evolutionary, and assembly principles of protein oligomerization. *Prog Mol Biol Transl Sci* **117**, 25-51 (2013).

88. Meini, M.R., Tomatis, P.E., Weinreich, D.M. & Vila, A.J. Quantitative Description of a Protein Fitness Landscape Based on Molecular Features. *Mol Biol Evol* **32**, 1774-87 (2015).

89. Wright, B.E., Butler, M.H. & Albe, K.R. Systems analysis of the tricarboxylic acid cycle in Dictyostelium discoideum. I. The basis for model construction. *J Biol Chem* **267**, 3101-5 (1992).

90. Hittinger, C.T. & Carroll, S.B. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**, 677-81 (2007).

91. Kan, S.B., Lewis, R.D., Chen, K. & Arnold, F.H. Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science* **354**, 1048-1051 (2016).

92. Tokuriki, N. et al. Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. *Nat Commun* **3**, 1257 (2012).

93. Newton, M.S., Arcus, V.L. & Patrick, W.M. Rapid bursts and slow declines: on the possible evolutionary trajectories of enzymes. *J R Soc Interface* **12**(2015).

94. 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 U S A* **109**, 10358-63 (2012).

95. Sykora, J. et al. Dynamics and hydration explain failed functional transformation in dehalogenase design. *Nat Chem Biol* **10**, 428-30 (2014).

96. Senior, A.W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020).

97. Risso, V.A. et al. De novo active sites for resurrected Precambrian enzymes. *Nat Commun* **8**, 16113 (2017).

98. Jiang, L. et al. De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-91 (2008).

99. Richter, F., Leaver-Fay, A., Khare, S.D., Bjelic, S. & Baker, D. De novo enzyme design using Rosetta3. *PLoS One* **6**, e19230 (2011).

100. Kries, H., Blomberg, R. & Hilvert, D. De novo enzymes by computational design. *Curr Opin Chem Biol* **17**, 221-8 (2013).

101. Kipnis, Y. et al. Design and optimization of enzymatic activity in a de novo β-barrel scaffold. *Protein Sci* **31**, e4405 (2022).

102. Obexer, R., Pott, M., Zeymer, C., Griffiths, A.D. & Hilvert, D. Efficient laboratory evolution of computationally designed enzymes with low starting activities using fluorescence-activated droplet sorting. *Protein Eng Des Sel* **30**, 531 (2017).

103. Coulther, T.A., Pott, M., Zeymer, C., Hilvert, D. & Ondrechen, M.J. Analysis of electrostatic coupling throughout the laboratory evolution of a designed retroaldolase. *Protein Sci* **30**, 1617-1627 (2021).

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

105. Mathieu, E. et al. Rational De Novo Design of a Cu Metalloenzyme for Superoxide Dismutation. *Chemistry* **26**, 249-258 (2020).

106. Klein, A.S. & Zeymer, C. Design and engineering of artificial metalloproteins: from de novo metal coordination to catalysis. *Protein Eng Des Sel* **34**(2021).

107. Brandenberg, O.F., Chen, K. & Arnold, F.H. Directed Evolution of a Cytochrome P450 Carbene Transferase for Selective Functionalization of Cyclic Compounds. *Journal of the American Chemical Society* **141**, 8989-8995 (2019).

108. Shi, L. et al. Complete Depolymerization of PET Waste by an Evolved PET Hydrolase from Directed Evolution. *Angew Chem Int Ed Engl* (2023).

109. Voigt, C.A., Martinez, C., Wang, Z.G., Mayo, S.L. & Arnold, F.H. Protein building blocks preserved by recombination. *Nat Struct Biol* **9**, 553-8 (2002).

110. Fox, R.J. et al. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol* **25**, 338-44 (2007).

111. Reetz, M.T., Wang, L.W. & Bocola, M. Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. *Angew Chem Int Ed Engl* **45**, 1236-41 (2006).

112. Acevedo-Rocha, C.G., Hoebenreich, S. & Reetz, M.T. Iterative saturation mutagenesis: a powerful approach to engineer proteins by systematically simulating Darwinian evolution. *Methods Mol Biol* **1179**, 103-28 (2014).

113. Soh, L.M.J. et al. Engineering a Thermostable Keto Acid Decarboxylase Using Directed Evolution and Computationally Directed Protein Design. *ACS Synthetic Biology* **6**, 610-618 (2017).

114. Lauchli, R. et al. High-throughput screening for terpene-synthase-cyclization activity and directed evolution of a terpene synthase. *Angew Chem Int Ed Engl* **52**, 5571-4 (2013).

115. Rocklin, G.J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168-175 (2017).

116. Kazlauskas, R. Engineering more stable proteins. *Chem Soc Rev* **47**, 9026-9045 (2018).

117. Yang, H., Liu, L., Li, J., Chen, J. & Du, G. Rational design to improve protein thermostability: recent advances and prospects. *ChemBioEng Reviews* **2**, 87-94 (2015).

118. Hsieh, C.L. et al. Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **369**, 1501-1505 (2020).

119. Juraszek, J. et al. Stabilizing the closed SARS-CoV-2 spike trimer. *Nature Communications* **12**, 244 (2021).

120. Leman, J.K. et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* **17**, 665-680 (2020).

121. Huang, P.S. et al. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* **12**, 29-34 (2016).

122. Polizzi, N.F. et al. De novo design of a hyperstable non-natural protein–ligand complex with sub-Å accuracy. *Nature Chemistry* **9**, 1157-1164 (2017).

123. Caldwell, S.J. et al. Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion. *Proc Natl Acad Sci U S A* **117**, 30362-30369 (2020).

124. Lehmann, M. et al. The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng* **15**, 403-11 (2002).

125. Kiss, C., Temirov, J., Chasteen, L., Waldo, G.S. & Bradbury, A.R. Directed evolution of an extremely stable fluorescent protein. *Protein Eng Des Sel* **22**, 313-23 (2009).

126. Hendrikse, N.M., Charpentier, G., Nordling, E. & Syrén, P.O. Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *Febs j* **285**, 4660-4673 (2018).

127. Risso, V.A., Gavira, J.A., Gaucher, E.A. & Sanchez-Ruiz, J.M. Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins* **82**, 887-96 (2014).

128. Trudeau, D.L., Kaltenbach, M. & Tawfik, D.S. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol Biol Evol* **33**, 2633-41 (2016).

129. Okafor, C.D. et al. Structural and Dynamics Comparison of Thermostability in Ancient, Modern, and Consensus Elongation Factor Tus. *Structure* **26**, 118-129.e3 (2018).

130. Thomson, R.E.S., Carrera-Pacheco, S.E. & Gillam, E.M.J. Engineering functional thermostable proteins using ancestral sequence reconstruction. *J Biol Chem* **298**, 102435 (2022).

131. Goldenzweig, A. et al. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol Cell* **63**, 337-346 (2016).

132. Musil, M. et al. FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res* **45**, W393-w399 (2017).

133. Pauling, L., Zuckerkandl, E., Henriksen, T. & Lövstad, R. Chemical paleogenetics. *Acta chem scand* **17**, S9-S16 (1963).

134. Jermann, T.M., Opitz, J.G., Stackhouse, J. & Benner, S.A. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* **374**, 57-9 (1995).

135. Perez-Jimenez, R. et al. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol* **18**, 592-6 (2011).

136. 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**, 2899-902 (2013).

137. Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R. & Damborsky, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *Chembiochem* **18**, 1448-1456 (2017).

138. Gumulya, Y. et al. Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nature Catalysis* **1**, 878-888 (2018).

139. Spence, M.A., Kaczmarski, J.A., Saunders, J.W. & Jackson, C.J. Ancestral sequence reconstruction for protein engineers. *Current Opinion in Structural Biology* **69**, 131-141 (2021).

140. Gaucher, E.A., Govindarajan, S. & Ganesh, O.K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704-707 (2008).

141. Zakas, P.M. et al. Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat Biotechnol* **35**, 35-37 (2017).

142. Gromiha, M.M., Pathak, M.C., Saraboji, K., Ortlund, E.A. & Gaucher, E.A. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins* **81**, 715-21 (2013).

143. Nguyen, V. et al. Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* **355**, 289-294 (2017).

144. Akanuma, S. et al. Experimental evidence for the thermophilicity of ancestral life. *Proc Natl Acad Sci U S A* **110**, 11067-72 (2013).

145. Ingles-Prieto, A. et al. Conservation of protein structure over four billion years. *Structure* **21**, 1690-7 (2013).

146. Robert, F. & Chaussidon, M. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature* **443**, 969-72 (2006).

147. Risso, V.A., Sanchez-Ruiz, J.M. & Ozkan, S.B. Biotechnological and protein-engineering implications of ancestral protein resurrection. *Current Opinion in Structural Biology* **51**, 106-115 (2018).

148. Tokuriki, N. & Tawfik, D.S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604 (2009).

149. Gomez-Fernandez, B.J. et al. Directed -in vitro- evolution of Precambrian and extant Rubiscos. *Sci Rep* **8**, 5532 (2018).

150. 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).

151. Whitfield, J.H. et al. Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci* **24**, 1412-22 (2015).

152. Barruetabeña, N. et al. Resurrection of efficient Precambrian endoglucanases for lignocellulosic biomass hydrolysis. *Communications Chemistry* **2**, 76 (2019).

153. Ortlund, E.A., Bridgham, J.T., Redinbo, M.R. & Thornton, J.W. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317**, 1544-8 (2007).

154. Nicoll, C.R. et al. Ancestral-sequence reconstruction unveils the structural basis of function in mammalian FMOs. *Nat Struct Mol Biol* **27**, 14-24 (2020).

155. Ladics, G.S. et al. Safety evaluation of a novel variant of consensus bacterial phytase. *Toxicology Reports* **7**, 844-851 (2020).

156. Nakano, S. et al. Ancestral L-amino acid oxidases for deracemization and stereoinversion of amino acids. *Communications Chemistry* **3**, 181 (2020).

157. Dellas, N., Liu, J., Botham, R.C. & Huisman, G.W. Adapting protein sequences for optimized therapeutic efficacy. *Curr Opin Chem Biol* **64**, 38-47 (2021).

158. Vavilis, T. et al. mRNA in the Context of Protein Replacement Therapy. *Pharmaceutics* **15**(2023).

159. Hoogenboom, H.R. Selecting and screening recombinant antibody libraries. *Nat Biotechnol* **23**, 1105-16 (2005).

160. Bojar, D. & Fussenegger, M. The Role of Protein Engineering in Biomedical Applications of Mammalian Synthetic Biology. *Small* **16**, e1903093 (2020).

161. Chng, C. et al. Engineered phenylalanine ammonia lyase polypeptides. US Patent Application, WO2018148633 (2018).

162. Wolf, C. et al. Engineering of Kuma030: A Gliadin Peptidase That Rapidly Degrades Immunogenic Gliadin Peptides in Gastric Conditions. *J Am Chem Soc* **137**, 13106-13 (2015).

163. Nyborg, A.C. et al. A Therapeutic Uricase with Reduced Immunogenicity Risk and Improved Development Properties. *PLoS One* **11**, e0167935 (2016).

164. Li, Z., Hoshino, Y., Tran, L. & Gaucher, E.A. Phylogenetic Articulation of Uric Acid Evolution in Mammals and How It Informs a Therapeutic Uricase. *Mol Biol Evol* **39**(2022).

165. Hendrikse, N.M. et al. Ancestral lysosomal enzymes with increased activity harbor therapeutic potential for treatment of Hunter syndrome. *iScience* **24**, 102154 (2021).

166. 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).

167. Caradonna, T.M. & Schmidt, A.G. Protein engineering strategies for rational immunogen design. *NPJ Vaccines* **6**, 154 (2021).

168. Yang, S. et al. Safety and immunogenicity of a recombinant tandem-repeat dimeric RBD-based protein subunit vaccine (ZF2001) against COVID-19 in adults: two randomised, double-blind, placebo-controlled, phase 1 and 2 trials. *Lancet Infect Dis* **21**, 1107-1119 (2021).

169. van der Lubbe, J.E.M. et al. Mini-HA Is Superior to Full Length Hemagglutinin Immunization in Inducing Stem-Specific Antibodies and Protection Against Group 1 Influenza Virus Challenges in Mice. *Front Immunol* **9**, 2350 (2018).

170. Duan, H. et al. Glycan Masking Focuses Immune Responses to the HIV-1 CD4-Binding Site and Enhances Elicitation of VRC01-Class Precursor Antibodies. *Immunity* **49**, 301-311.e5 (2018).

171. Boyoglu-Barnum, S. et al. Glycan repositioning of influenza hemagglutinin stem facilitates the elicitation of protective cross-group antibody responses. *Nature Communications* **11**, 791 (2020).

172. Liu, W.C., Jan, J.T., Huang, Y.J., Chen, T.H. & Wu, S.C. Unmasking Stem-Specific Neutralizing Epitopes by Abolishing N-Linked Glycosylation Sites of Influenza Virus Hemagglutinin Proteins for Vaccine Design. *J Virol* **90**, 8496-508 (2016).

173. Sesterhenn, F. et al. De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science* **368**(2020).

174. Bajic, G. et al. Structure-Guided Molecular Grafting of a Complex Broadly Neutralizing Viral Epitope. *ACS Infect Dis* **6**, 1182-1191 (2020).

175. Sesterhenn, F. et al. Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen. *PLoS Biol* **17**, e3000164 (2019).

176. Sprenger, K.G., Louveau, J.E., Murugan, P.M. & Chakraborty, A.K. Optimizing immunization protocols to elicit broadly neutralizing antibodies. *Proc Natl Acad Sci U S A* **117**, 20077-20087 (2020).

177. Kaku, C.I. et al. Broad anti-SARS-CoV-2 antibody immunity induced by heterologous ChAdOx1/mRNA-1273 vaccination. *Science* **375**, 1041-1047 (2022).

178. Godley, L. et al. Introduction of intersubunit disulfide bonds in the membrane-distal region of the influenza hemagglutinin abolishes membrane fusion activity. *Cell* **68**, 635-45 (1992).

179. Qiao, H. et al. Specific single or double proline substitutions in the "spring-loaded" coiled-coil region of the influenza hemagglutinin impair or abolish membrane fusion activity. *J Cell Biol* **141**, 1335-47 (1998).

180. McLellan, J.S. et al. Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. *Science* **342**, 592-8 (2013).

181. Krarup, A. et al. A highly stable prefusion RSV F vaccine derived from structural analysis of the fusion mechanism. *Nat Commun* **6**, 8143 (2015).

182. Rutten, L. et al. Structure-Based Design of Prefusion-Stabilized Filovirus Glycoprotein Trimers. *Cell Rep* **30**, 4540-4550.e3 (2020).

183. Pallesen, J. et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc Natl Acad Sci U S A* **114**, E7348-e7357 (2017).

184. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263 (2020).

185. Ducatez, M.F. et al. Feasibility of reconstructed ancestral H5N1 influenza viruses for cross-clade protective vaccine development. *Proceedings of the National Academy of Sciences* **108**, 349-354 (2011).

186. Giles, B.M. & Ross, T.M. A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. *Vaccine* **29**, 3043-54 (2011).

187. Giles, B.M., Bissel, S.J., Dealmeida, D.R., Wiley, C.A. & Ross, T.M. Antibody breadth and protective efficacy are increased by vaccination with computationally optimized hemagglutinin but not with polyvalent hemagglutinin-based H5N1 virus-like particle vaccines. *Clin Vaccine Immunol* **19**, 128-39 (2012).

188. Giles, B.M. et al. A computationally optimized hemagglutinin virus-like particle vaccine elicits broadly reactive antibodies that protect nonhuman primates from H5N1 infection. *J Infect Dis* **205**, 1562-70 (2012).

189. Mathieu, E. et al. Coronavirus Pandemic (COVID-19), https://ourworldindata.org/coronavirus, accessed on Feb 6, 2023. (2020-2023).

190. The Economist. The pandemic's true death toll, https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates, accessed on Feb 12, 2023. (2020-2023).

191. Huang, Y., Yang, C., Xu, X.F., Xu, W. & Liu, S.W. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* **41**, 1141-1149 (2020).

192. Zhu, C. et al. Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. *J Med Virol* **93**, 5729-5741 (2021).

193. Gui, M. et al. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Research* **27**, 119-129 (2017).

194. Yuan, Y. et al. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nature Communications* **8**, 15092 (2017).

195. Walls, A.C. et al. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

196. Kirchdoerfer, R.N. et al. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Scientific Reports* **8**, 15701 (2018).

197. Heinz, F.X. & Stiasny, K. Distinguishing features of current COVID-19 vaccines: knowns and unknowns of antigen presentation and modes of action. *NPJ Vaccines* **6**, 104 (2021).

198. Lupala, C.S., Ye, Y., Chen, H., Su, X.D. & Liu, H. Mutations on RBD of SARS-CoV-2 Omicron variant result in stronger binding to human ACE2 receptor. *Biochem Biophys Res Commun* **590**, 34-41 (2022).

199. Vanderven, H.A. & Kent, S.J. The protective potential of Fc-mediated antibody functions against influenza virus and other viral pathogens. *Immunol Cell Biol* **98**, 253-263 (2020).

200. Christianson, D.W. Structural and Chemical Biology of Terpenoid Cyclases. *Chem Rev* **117**, 11570-11648 (2017).

201. Spracklen, D.V., Bonn, B. & Carslaw, K.S. Boreal forests, aerosols and the impacts on clouds and climate. *Philos Trans A Math Phys Eng Sci* **366**, 4613-26 (2008).

202. Yang, W. et al. Advances in Pharmacological Activities of Terpenoids. *Natural Product Communications* **15**, 1934578X20903555 (2020).

203. Del Prado-Audelo, M.L. et al. Therapeutic Applications of Terpenes on Inflammatory Diseases. *Front Pharmacol* **12**, 704197 (2021).

204. Stamm, A. et al. Pinene-Based Oxidative Synthetic Toolbox for Scalable Polyester Synthesis. *JACS Au* **1**, 1949-1960 (2021).

205. Ninkuu, V. et al. Biochemistry of Terpenes and Recent Advances in Plant Protection. *Int J Mol Sci* **22**(2021).

206. Cao, R. et al. Diterpene cyclases and the nature of the isoprene fold. *Proteins* **78**, 2417-32 (2010).

207. Huang, H. et al. Structure of a membrane-embedded prenyltransferase homologous to UBIAD1. *PLoS Biol* **12**, e1001911 (2014).

208. Oldfield, E. & Lin, F.Y. Terpene biosynthesis: modularity rules. *Angew Chem Int Ed Engl* **51**, 1124-37 (2012).

209. Rudolf, J.D. et al. Structure of the ent-Copalyl Diphosphate Synthase PtmT2 from Streptomyces platensis CB00739, a Bacterial Type II Diterpene Synthase. *J Am Chem Soc* **138**, 10905-15 (2016).

210. Whittington, D.A. et al. Bornyl diphosphate synthase: structure and strategy for carbocation manipulation by a terpenoid cyclase. *Proc Natl Acad Sci U S A* **99**, 15375-80 (2002).

211. Chen, M., Chou, W.K., Toyomasu, T., Cane, D.E. & Christianson, D.W. Structure and Function of Fusicoccadiene Synthase, a Hexameric Bifunctional Diterpene Synthase. *ACS Chem Biol* **11**, 889-99 (2016).

212. Wendt, K.U., Poralla, K. & Schulz, G.E. Structure and function of a squalene cyclase. *Science* **277**, 1811-5 (1997).

213. Baer, P. et al. Induced-fit mechanism in class I terpene cyclases. *Angew Chem Int Ed Engl* **53**, 7652-6 (2014).

214. Dickschat, J.S. Bacterial terpene cyclases. *Nat Prod Rep* **33**, 87-110 (2016).

215. Durairaj, J. et al. An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157-165 (2019).

216. Huang, Z.-Y., Ye, R.-Y., Yu, H.-L., Li, A.-T. & Xu, J.-H. Mining methods and typical structural mechanisms of terpene cyclases. *Bioresources and Bioprocessing* **8**, 66 (2021).

217. Bian, G. et al. Metabolic Engineering-Based Rapid Characterization of a Sesquiterpene Cyclase and the Skeletons of Fusariumdiene and Fusagramineol from Fusarium graminearum. *Org Lett* **20**, 1626-1629 (2018).

218. Zhu, F. et al. In vitro reconstitution of mevalonate pathway and targeted engineering of farnesene overproduction in Escherichia coli. *Biotechnol Bioeng* **111**, 1396-405 (2014).

219. Dahl, R.H. et al. Engineering dynamic pathway regulation using stress-response promoters. *Nat Biotechnol* **31**, 1039-46 (2013).

220. Schotte, C., Lukat, P., Deuschmann, A., Blankenfeldt, W. & Cox, R.J. Understanding and Engineering the Stereoselectivity of Humulene Synthase. *Angew Chem Int Ed Engl* **60**, 20308-20312 (2021).

221. Diao, H. et al. Biosynthetic Mechanism of Lanosterol: A Completed Story. *ACS Catalysis* **10**, 2157-2168 (2020).

222. Thoma, R. et al. Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **432**, 118-22 (2004).

223. Corey, E.J. et al. Studies on the Substrate Binding Segments and Catalytic Action of Lanosterol Synthase. Affinity Labeling with Carbocations Derived from Mechanism-Based Analogs of 2,3-Oxidosqualene and Site-Directed Mutagenesis Probes. *Journal of the American Chemical Society* **119**, 1289-1296 (1997).

224. Chen, N., Wang, S., Smentek, L., Hess, B.A., Jr. & Wu, R. Biosynthetic Mechanism of Lanosterol: Cyclization. *Angew Chem Int Ed Engl* **54**, 8693-6 (2015).

225. van Tamelen, E.E., Willett, J.D., Clayton, R.B. & Lord, K.E. Enzymic Conversion of Squalene 2,3-Oxide to Lanosterol and Cholesterol. *Journal of the American Chemical Society* **88**, 4752-4754 (1966).

226. Willett, J.D., Sharpless, K.B., Lord, K.E., van Tamelen, E.E. & Clayton, R.B. Squalene-2,3oxide, an intermediate in the enzymatic conversion of squalene to lanosterol and cholesterol. *J Biol Chem* **242**, 4182-91 (1967).

227. Rabe, P. et al. Mechanistic Investigations of Two Bacterial Diterpene Cyclases: Spiroviolene Synthase and Tsukubadiene Synthase. *Angew Chem Int Ed Engl* **56**, 2776-2779 (2017).

228. Gomori, G. Calcification and Phosphatase. *Am J Pathol* **19**, 197-209 (1943).

229. Gizak, A., Duda, P., Wisniewski, J. & Rakus, D. Fructose-1,6-bisphosphatase: From a glucose metabolism enzyme to multifaceted regulator of a cell fate. *Adv Biol Regul* **72**, 41-50 (2019).

230. Brown, G. et al. Structural and biochemical characterization of the type II fructose-1,6-bisphosphatase GlpX from Escherichia coli. *J Biol Chem* **284**, 3784-92 (2009).

231. Williams, M.K. & Kantrowitz, E.R. Isolation and sequence analysis of the cDNA for pig kidney fructose 1,6-bisphosphatase. *Proceedings of the National Academy of Sciences* **89**, 3080-3082 (1992).

232. Daie, J. Cytosolic fructose-1,6-bisphosphatase: A key enzyme in the sucrose biosynthetic pathway. *Photosynth Res* **38**, 5-14 (1993).

233. Donahue, J.L., Bownas, J.L., Niehaus, W.G. & Larson, T.J. Purification and characterization of glpX-encoded fructose 1, 6-bisphosphatase, a new enzyme of the glycerol 3-phosphate regulon of Escherichia coli. *J Bacteriol* **182**, 5624-7 (2000).

234. Kelly, G.J., Zimmermann, G. & Latzko, E. Fructose-bisphosphatase from spinach leaf chloroplast and cytoplasm. in *Methods in Enzymology*, Vol. 90 (ed. Wood, W.A.) 371-378 (Academic Press, 1982).

235. Marcus, F., Moberly, L. & Latshaw, S.P. Comparative amino acid sequence of fructose-1,6-bisphosphatases: identification of a region unique to the light-regulated chloroplast enzyme. *Proc Natl Acad Sci U S A* **85**, 5379-83 (1988).

236. Chiadmi, M., Navaza, A., Miginiac-Maslow, M., Jacquot, J.P. & Cherfils, J. Redox signalling in the chloroplast: structure of oxidized pea fructose-1,6-bisphosphate phosphatase. *Embo j* **18**, 6809-15 (1999).

237. Jiang, Y.H., Wang, D.Y. & Wen, J.F. The independent prokaryotic origins of eukaryotic fructose-1, 6-bisphosphatase and sedoheptulose-1, 7-bisphosphatase and the implications of their origins for the evolution of eukaryotic Calvin cycle. *BMC Evol Biol* **12**, 208 (2012).

238. Fujita, Y. et al. Identification and expression of the Bacillus subtilis fructose-1, 6-bisphosphatase gene (fbp). *J Bacteriol* **180**, 4309-13 (1998).

239. Yoo, J.G. & Bowien, B. Analysis of the cbbF genes from Alcaligenes eutrophus that encode fructose-1,6-/sedoheptulose-1,7-bisphosphatase. *Curr Microbiol* **31**, 55-61 (1995).

240. Windhövel, U. & Bowien, B. On the operon structure of the cfx gene clusters in Alcaligenes eutrophus. *Arch Microbiol* **154**, 85-91 (1990).

241. Tamoi, M., Ishikawa, T., Takeda, T. & Shigeoka, S. Molecular characterization and resistance to hydrogen peroxide of two fructose-1,6-bisphosphatases from Synechococcus PCC 7942. *Arch Biochem Biophys* **334**, 27-36 (1996).

242. Tamoi, M., Murakami, A., Takeda, T. & Shigeoka, S. Acquisition of a new type of fructose-1,6-bisphosphatase with resistance to hydrogen peroxide in cyanobacteria: molecular characterization of the enzyme from Synechocystis PCC 6803. *Biochim Biophys Acta* **1383**, 232-44 (1998).

243. Janasch, M., Asplund-Samuelsson, J., Steuer, R. & Hudson, E.P. Kinetic modeling of the Calvin cycle identifies flux control and stable metabolomes in Synechocystis carbon fixation. *J Exp Bot* **70**, 973-983 (2019).

244. Miyagawa, Y., Tamoi, M. & Shigeoka, S. Overexpression of a cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase in tobacco enhances photosynthesis and growth. *Nat Biotechnol* **19**, 965-9 (2001).

245. Ogawa, T. et al. Enhancement of photosynthetic capacity in Euglena gracilis by expression of cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase leads to increases in biomass and wax ester production. *Biotechnol Biofuels* **8**, 80 (2015).

246. López-Calcagno, P.E. et al. Stimulating photosynthetic processes increases productivity and water-use efficiency in the field. *Nat Plants* **6**, 1054-1063 (2020).

247. Cotton, C.A., Kabasakal, B.V., Miah, N.A. & Murray, J.W. Structure of the dual-function fructose-1,6/sedoheptulose-1,7-bisphosphatase from Thermosynechococcus elongatus bound with sedoheptulose-7-phosphate. *Acta Crystallogr F Struct Biol Commun* **71**, 1341-5 (2015).

248. Feng, L. et al. Structural and biochemical characterization of fructose-1,6/sedoheptulose-1,7-bisphosphatase from the cyanobacterium Synechocystis strain 6803. *Febs j* **281**, 916-26 (2014).

249. Tamoi, M., Takeda, T. & Shigeoka, S. Functional Analysis of Fructose-1,6-Bisphosphatase Isozymes (fbp-I and fbp-II Gene Products) in Cyanobacteria. *Plant and Cell Physiology* **40**, 257-261 (1999).

250. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

251. Pollock, D.D., Zwickl, D.J., McGuire, J.A. & Hillis, D.M. Increased taxon sampling is advantageous for phylogenetic inference. *Syst Biol* **51**, 664-71 (2002).

252. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-77 (2007).

253. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-3 (2009).

254. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-9 (1992).

255. Vialle, R.A., Tamuri, A.U. & Goldman, N. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Mol Biol Evol* **35**, 1783-1797 (2018).

256. Aadland, K. & Kolaczkowski, B. Alignment-Integrated Reconstruction of Ancestral Sequences Improves Accuracy. *Genome Biol Evol* **12**, 1549-1565 (2020).

257. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).

258. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-74 (2015).

259. Fitch, W.M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology* **20**, 406-416 (1971).

260. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549 (2018).

261. Dayhoff, M. A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **5**, suppl. 3 (1987).

262. Le, S.Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**, 1307-20 (2008).

263. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691-9 (2001).

264. Jones, D.T., Taylor, W.R. & Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-82 (1992).

265. Abadi, S., Azouri, D., Pupko, T. & Mayrose, I. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications* **10**, 934 (2019).

266. 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).

267. Hanson-Smith, V., Kolaczkowski, B. & Thornton, J.W. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol* **27**, 1988-99 (2010).

268. Merkl, R. & Sterner, R. Ancestral protein reconstruction: techniques and applications. *Biol Chem* **397**, 1-21 (2016).

269. Ochman, H., Lawrence, J.G. & Groisman, E.A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).

270. Petřek, M. et al. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics* **7**, 316 (2006).

271. Chovancova, E. et al. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* **8**, e1002708 (2012).

272. Coulombe, R., Yue, K.Q., Ghisla, S. & Vrielink, A. Oxygen access to the active site of cholesterol oxidase through a narrow channel is gated by an Arg-Glu pair. *J Biol Chem* **276**, 30435-41 (2001).

273. Huang, X., Holden, H.M. & Raushel, F.M. Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annu Rev Biochem* **70**, 149-80 (2001).

274. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423**, 949-55 (2003).

275. Barney, B.M., Yurth, M.G., Dos Santos, P.C., Dean, D.R. & Seefeldt, L.C. A substrate channel in the nitrogenase MoFe protein. *J Biol Inorg Chem* **14**, 1015-22 (2009).

276. Huang, X. & Raushel, F.M. An engineered blockage within the ammonia tunnel of carbamoyl phosphate synthetase prevents the use of glutamine as a substrate but not ammonia. *Biochemistry* **39**, 3240-7 (2000).

277. Chaloupková, R. et al. Modification of activity and specificity of haloalkane dehalogenase from Sphingomonas paucimobilis UT26 by engineering of its entrance tunnel. *J Biol Chem* **278**, 52622-8 (2003).

278. Li, G. et al. Simultaneous engineering of an enzyme's entrance tunnel and active site: the case of monoamine oxidase MAO-N. *Chem Sci* **8**, 4093-4099 (2017).

279. Kim, S.M. et al. O2-tolerant CO dehydrogenase via tunnel redesign for the removal of CO from industrial flue gas. *Nature Catalysis* **5**, 807-817 (2022).

280. Gao, K., Oerlemans, R. & Groves, M.R. Theory and applications of differential scanning fluorimetry in early-stage drug discovery. *Biophysical Reviews* **12**, 85-104 (2020).

281. Lo, M.C. et al. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Anal Biochem* **332**, 153-9 (2004).

282. Alexander, C.G. et al. Novel microscale approaches for easy, rapid determination of protein stability in academic and commercial settings. *Biochim Biophys Acta* **1844**, 2241-50 (2014).

283. NanoTemper Technologies. How protein aggregation can change the course of your experiment, https://nanotempertech.com/blog/how-protein-aggregation-can-change-the-course-of-your-experiment/, accessed on Feb 14, 2023. (2019).

284. Niesen, F.H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat Protoc* **2**, 2212-21 (2007).

285. Stetefeld, J., McKenna, S.A. & Patel, T.R. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophys Rev* **8**, 409-427 (2016).

286. Lorber, B., Fischer, F., Bailly, M., Roy, H. & Kern, D. Protein analysis by dynamic light scattering: methods and techniques for students. *Biochem Mol Biol Educ* **40**, 372-82 (2012).

287. McWilliam, I.G. & Dewar, R.A. Flame Ionization Detector for Gas Chromatography. *Nature* **181**, 760-760 (1958).

288. de Saint Laumer, J.Y. et al. Prediction of response factors for gas chromatography with flame ionization detection: Algorithm improvement, extension to silylated compounds, and application to the quantification of metabolites. *J Sep Sci* **38**, 3209-3217 (2015).

289. Vandendool, H. & Kratz, P.D. A generalization of the retention index system including linear temperature programmed gas-liquid partiton chromatography. . *J Chromatogr* **11**, 463-71 (1963).

290. Moros, G., Chatziioannou, A.C., Gika, H.G., Raikos, N. & Theodoridis, G. Investigation of the derivatization conditions for GC-MS metabolomics of biological samples. *Bioanalysis* **9**, 53-65 (2017).

291. Bowden, J.A. et al. Enhanced analysis of steroids by gas chromatography/mass spectrometry using microwave-accelerated derivatization. *Anal Chem* **81**, 6725-34 (2009).

292. Vardakou, M., Salmon, M., Faraldos, J.A. & O'Maille, P.E. Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* **1**, 187-96 (2014).

293. Drenth, J. *Principles of protein X-ray crystallography*, (Springer Science & Business Media, 2007).

294. Chernov, A.A. Protein crystals and their growth. *J Struct Biol* **142**, 3-21 (2003).

295. Sigworth, F.J. Principles of cryo-EM single-particle image processing. *Microscopy (Oxf)* **65**, 57-67 (2016).

296. Carroni, M. & Saibil, H.R. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods* **95**, 78-85 (2016).

297. Croll, T.I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519-530 (2018).

298. Pettersen, E.F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82 (2021).

299. Henderson, R. et al. Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nature Structural & Molecular Biology* **27**, 925-933 (2020).

300. McCallum, M., Walls, A.C., Bowen, J.E., Corti, D. & Veesler, D. Structure-guided covalent stabilization of coronavirus spike glycoprotein trimers in the closed conformation. *Nature Structural & Molecular Biology* **27**, 942-949 (2020).

301. Xiong, X. et al. A thermostable, closed SARS-CoV-2 spike protein trimer. *Nature Structural & Molecular Biology* **27**, 934-941 (2020).

302. Boni, M.F. et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology* **5**, 1408-1417 (2020).

303. Hueting, D. et al. Design, structure and plasma binding of ancestral β-CoV scaffold antigens, https://doi.org/10.21203/rs.3.rs-1909545/v1. (2022).

304. Rabelo, V.W., Romeiro, N.C. & Abreu, P.A. Design strategies of oxidosqualene cyclase inhibitors: Targeting the sterol biosynthetic pathway. *J Steroid Biochem Mol Biol* **171**, 305-317 (2017).

305. Bellosta, S. & Corsini, A. Statin drug interactions and related adverse reactions: an update. *Expert Opin Drug Saf* **17**, 25-37 (2018).

306. Johansson, M., Bouakaz, E., Lovmar, M. & Ehrenberg, M. The kinetics of ribosomal peptidyl transfer revisited. *Mol Cell* **30**, 589-98 (2008).

307. Tunuguntla, R.H. et al. Enhanced water permeability and tunable ion selectivity in subnanometer carbon nanotube porins. *Science* **357**, 792-796 (2017).

308. Gscheidmeier, M. & Fleig, H. Turpentines. in *Ullmann's Encyclopedia of Industrial Chemistry*.

309. Gersmann, H. & Aldred, J. Medicinal tree used in chemotherapy drug faces extinction, https://www.theguardian.com/environment/2011/nov/10/iucn-red-list-tree-chemotherapy, accessed on Feb 11, 2023. (2011).

310. Sarria, S., Wong, B., Martín, H.G., Keasling, J.D. & Peralta-Yahya, P. Microbial Synthesis of Pinene. *ACS Synthetic Biology* **3**, 466-475 (2014).

311. Meadows, A.L. et al. Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* **537**, 694-697 (2016).

312. Jongedijk, E. et al. Biotechnological production of limonene in microorganisms. *Appl Microbiol Biotechnol* **100**, 2927-38 (2016).

313. Jiang, Z. et al. Agronomic and chemical performance of field-grown tobacco engineered for triterpene and methylated triterpene metabolism. *Plant Biotechnol J* **16**, 1110-1124 (2018).

314. Moser, S. & Pichler, H. Identifying and engineering the ideal microbial terpenoid production host. *Applied Microbiology and Biotechnology* **103**, 5501-5516 (2019).

315. Greenhagen, B.T., O'Maille, P.E., Noel, J.P. & Chappell, J. Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proc Natl Acad Sci U S A* **103**, 9826-31 (2006).

316. Kampranis, S.C. et al. Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: structural insights into the evolution of terpene synthase function. *Plant Cell* **19**, 1994-2005 (2007).

317. Oberhauser, C. et al. Exploiting the Synthetic Potential of Sesquiterpene Cyclases for Generating Unnatural Terpenoids. *Angew Chem Int Ed Engl* **57**, 11802-11806 (2018).

318. O'Maille, P.E. et al. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat Chem Biol* **4**, 617-23 (2008).

319. Jia, M., Zhou, K., Tufts, S., Schulte, S. & Peters, R.J. A Pair of Residues That Interactively Affect Diterpene Synthase Product Outcome. *ACS Chemical Biology* **12**, 862-867 (2017).

320. Yoshikuni, Y., Ferrin, T.E. & Keasling, J.D. Designed divergent evolution of enzyme function. *Nature* **440**, 1078-82 (2006).

321. Kratzer, J.T. et al. Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc Natl Acad Sci U S A* **111**, 3763-8 (2014).

322. Kaltenbach, M. et al. Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat Chem Biol* **14**, 548-555 (2018).

323. Schriever, K. et al. Engineering of Ancestors as a Tool to Elucidate Structure, Mechanism, and Specificity of Extant Terpene Cyclase. *J Am Chem Soc* **143**, 3794-3807 (2021).

324. Baer, P. et al. Hedycaryol synthase in complex with nerolidol reveals terpene cyclase mechanism. *Chembiochem* **15**, 213-6 (2014).

325. Krieg, T., Sydow, A., Faust, S., Huth, I. & Holtmann, D. CO(2) to Terpenes: Autotrophic and Electroautotrophic α-Humulene Production with Cupriavidus necator. *Angew Chem Int Ed Engl* **57**, 1879-1882 (2018).

326. Werner, A., Broeckling, C.D., Prasad, A. & Peebles, C.A.M. A comprehensive time-course metabolite profiling of the model cyanobacterium Synechocystis sp. PCC 6803 under diurnal light:dark cycles. *Plant J* **99**, 379-388 (2019).

327. Jaiswal, D. & Wangikar, P.P. Dynamic Inventory of Intermediate Metabolites of Cyanobacteria in a Diurnal Cycle. *iScience* **23**, 101704 (2020).

328. Berg, I.A. Ecological aspects of the distribution of different autotrophic CO2 fixation pathways. *Appl Environ Microbiol* **77**, 1925-36 (2011).

329. Asplund-Samuelsson, J. & Hudson, E.P. Wide range of metabolic adaptations to the acquisition of the Calvin cycle revealed by comparison of microbial genomes. *PLoS Comput Biol* **17**, e1008742 (2021).

330. Sporre, E. et al. Metabolite interactions in the bacterial Calvin cycle and implications for flux regulation, https://doi.org/10.1101/2022.03.15.483797 (2022).

331. Piazza, I. et al. A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* **172**, 358-372.e23 (2018).

332. Pecoraro, V.L., Hermes, J.D. & Cleland, W.W. Stability constants of Mg2+ and Cd2+ complexes of adenine nucleotides and thionucleotides and rate constants for formation and dissociation of MgATP and MgADP. *Biochemistry* **23**, 5262-71 (1984).

333. Moellering, R.E. & Cravatt, B.F. Functional lysine modification by an intrinsically reactive primary glycolytic metabolite. *Science* **341**, 549-53 (2013).

334. Coukos, J.S., Lee, C.W., Pillai, K.S., Liu, K.J. & Moellering, R.E. Widespread, Reversible Cysteine Modification by Methylglyoxal Regulates Metabolic Enzyme Function. *ACS Chemical Biology* **18**, 91-101 (2023).