
Australian National University

# Harnessing evolution-guided methods for protein engineering and applications in enzyme-mediated radical polymerisation

**Hannah Bott**

January 2025

A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

# Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of the author's knowledge, it contains no material previously published or written by another person, except where due reference is made in the text.

Hannah Bott

January 2025

# Acknowledgements

Firstly, I would like to thank my chair supervisor Colin Jackson for giving me this amazing opportunity and in particular for helping me get everything over the line. A huge thank you also to my Bristol supervisor Adam Perriman for giving me the opportunity to explore a new and interesting field in his lab. I am grateful also to Ross Anderson for letting me use his lab space and for his advice at a critical time. I would like to thank my panel members Gottfried Otting and Lara Malins for their time. To Lara Malins, thank you for being an inspiring role model for women in science.

This PhD would not have been possible without all the training and advice from many people. At the Australian National University, thank you to Lynn Tan, Joe Kaczmarski, Brendon Lee, James Anthony, Vanessa Vongsouthi, Adam Damry, Rebecca Frkic and Aidan Grosas for giving me advice and helping me learn how carry out various experiments. For the LacI project, thank you to Rosie Georgelin and Matthew Spence for the discussions and for the amazing work on the other parts of the project. Thank you to Edan Habel, Aidan Grosas, James Anthony and Mahakaran Sandhu who were willing to talk about their research at odd times of the night. I am also very grateful for all the other support, such as the talks with Surya Sundaramoorthy and the coffee runs with Cassidy Whitefield. A special thank you to Jake Saunders for the walks and talks and Skye Young for the talks too.

At the University of Bristol, I am hugely grateful to Mark Shannon for all his help and advice with hydrogels and ELMs and for the interesting science discussion. A big thank you as well to Charles de Kergeriou for his assistance with compression and tensile testing and advice on materials characterisation. Thank you to George Klemperer for showing me how to make alg/PNIPAm and Jiongyi Yan for all the discussions. Thank you to Fabrizio Scarpa for advice on materials. For learning how to transform yeast, a huge thank you to Ian Prosser and Paul Curnow. I am grateful to Chris Arthur for his help with mass spec and Dominic Alibhai for his confocal microscopy training. Thank you to Raquel Cruz Samperio, Holly Ford and Valeria Sandoval Torres for helping me find my way around new labs and answering all the inane questions that accompanies this. I am also very grateful to Ximena Vasto Anzaldo for all the encouragement.

For both universities, I am very grateful for the laboratory technicians, cleaners and HDR support staff who made this work possible.

Finally, the support of my family and friends has been invaluable. A special thank you to my mother, Elizabeth Walker, for always encouraging my education and supporting me to pursue this path. A special thank you also to my partner, Michael Hill, for always being there to support me through the highs and lows and for always believing in me. Words cannot describe how much your support has meant to me and I will forever be indebted to you.

This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

# Abstract

Proteins can be powerful tools for a myriad of applications; however, their uses are limited unless they can be engineered to purpose. Evolutionary methods of engineering proteins include ancestral sequence reconstruction (ASR) and directed evolution. ASR can be an effective method to study the sequence-function relationship over time. Additionally, ASR can predict thermostable ancestors that may have improved expression as well as improved temperature and solvent stability for future applications. Directed evolution can be useful for engineering a protein with a desired property as it samples a large sequence-function space. Whilst both ASR and directed evolution have been commonly used to engineer proteins for applications in the pharmaceutical, bioremediation and green chemistry fields, amongst others, their potential in understanding transcription factors and for enzyme-mediated radical polymerisation (EMRP) has been relatively underexplored. In this thesis, the utility of evolutionary methods in these fields was investigated using various biochemical techniques and material characterisation methods.

In Chapter 2, ligand binding of ancestrally reconstructed proteins of the lac repressor (LacI) family was characterised in detail to reveal an enthalpic-entropic trade-off in ligand binding along the LacI trajectory and increasing specialisation towards β-substituted D-galactosides as ligands. It was found that functional changes in ligand binding between two ancestors result from complex and likely epistatic interactions between residues. Hence, this research could inform future engineering efforts of LacI family transcription factors. Also, as the genotype-phenotype landscape of the LacI family of transcription factors has proved difficult to characterise in the past, this study of the LacI family highlights the ability of ASR to characterise the sequence-function landscape in a ‘smart’ way, thereby enabling deeper insights for future protein design.

ASR was further utilised in Chapter 3 to identify an ancestral horseradish peroxidase (HRP) with improved recombinant expression. HRP is a commonly used enzyme, however, its use is restricted because it is difficult to express recombinantly and plant-derived HRP has batch-to-batch variation. Due to these limitations, ASR was utilised to engineer HRP for improved recombinant expression. It was found that ASR successfully identified ancestral peroxidases that could be expressed recombinantly in E. coli in a soluble, active form. Selected ancestral peroxidases were demonstrated to have relatively high thermostability whilst still possessing

the desired peroxidase activity. It was also shown that one of the ancestral peroxidases could replace HRP in the EMRP of poly(*N*-isopropylacrylamide) (PNIPAm).

Chapter 4 further explores the use of evolutionary engineering methods in EMRP. Previously, HRP has commonly been used in EMRP of materials like alginate/PNIPAm. However, due to the limitations of HRP, it was investigated whether a peroxygenase engineered for high recombinant yields using directed evolution could be utilised instead. It was found that this peroxygenase, unspecific peroxygenase PaDa-I (UPO), could replace HRP in the formation of alg/PNIPAm to form a hydrogel with similar printability, thermosensitive properties and relatively similar mechanical properties to the HRP-formed alg/PNIPAm. Thus, the potential of UPO for use in EMRP reactions was established for the first time. It was then demonstrated that UPO-formed alg/PNIPAm containing K. phaffii or S. cerevisiae could act as a producer of eukaryotic proteins, thereby forming a 3D-printable, thermosensitive, engineered living material bioreactor.

Overall, this thesis highlights the utility of evolutionary methods in gaining further insights into protein genotype-phenotype landscapes and in engineering proteins for enzyme-mediated polymerisation applications.

# Table of Contents

# List of Abbreviations

| | |
|---|---|
| ABTS | 2,2'-Azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) |
| AICc | Akaike information criterion |
| alg | alginate |
| aLRT | approximate likelihood ratio test |
| Anc | ancestor |
| *Aru* | *Armoracia rusticana* (horseradish) |
| ASR | ancestral sequence reconstruction |
| AU | approximately unbiased (test) |
| BIC | Bayesian information criterion |
| CD | circular dichroism |
| DBD | DNA binding domain |
| di$H_2O$ | deionized water |
| DNA | deoxyribonucleic acid |
| DSF | differential scanning fluorimetry |
| EB | empirical Bayes |
| EC | enzyme class |
| EGFP | enhanced green fluorescent protein |
| ELM | engineered living material |
| EMRP | enzyme-mediated radical polymerisation |
| FRET | Förster resonance energy transfer |
| GalR | *gal* repressor |
| GFP | green fluorescent protein |
| GOx | glucose oxidase |
| GRAS | generally regarded as safe |
| HB | hierarchical Bayes |
| HDX-MS | hydrogen deuterium exchange mass spectrometry |
| HRP | horseradish peroxidase |
| indel | insertion and/or deletion |
| IPN | interpenetrating network |

| | |
|---|---|
| IPTG | isopropyl β-D-thiogalactopyranoside |
| ITC | isothermal titration calorimetry |
| $k_{cat}$ | turnover number |
| $K_D$ | equilibrium dissociation constant |
| $K_M$ | Michaelis constant |
| LacI | *lac* repressor |
| LBA | long branch attraction |
| LBD | ligand binding domain |
| LCST | lower critical solution temperature |
| MALDI-TOF | matrix-assisted laser desorption/ionization time of flight |
| MAP | maximum a posteriori |
| MBA | *N*,*N*′-methylenebisacrylamide |
| MCMC | Markov chain Monte Carlo |
| ML | maximum likelihood |
| MP | maximum parsimony |
| MRCA | most recent common ancestor |
| MS | mass spectrometry |
| MSA | multiple sequence alignment |
| $Mut^+$ | methanol utilisation plus |
| $Mut^S$ | slow methanol utilisation |
| MβDG | methyl β-D-galactopyranoside |
| NIPAm | N-isopropylacrylamide |
| PNIPAm | poly(*N*-isopropylacrylamide) |
| PP | posterior probability |
| RMSD | root mean square deviation |
| RNA | ribonucleic acid |
| ROS | reactive oxygen species |
| Rz | Reinheitszahl |
| S.E.M. | standard error of the mean |
| | Shimodaira Hasegawa-like |
| SH-aLRT | approximate likelihood ratio test |

| | |
|---|---|
| SN | single network |
| SSM | site saturation mutagenesis |
| SSN | sequence similarity network |
| TCS | transitive consistency score |
| TF | transcription factor |
| $T_m$ | thermal unfolding transition temperature |
| UFBoot | ultrafast bootstrap |
| UPO | unspecific peroxygenase (variant PaDaI) |
| WGD | whole genome duplication |
| ΔG | change in Gibbs free energy |
| ΔH | change in enthalpy |
| ΔS | change in entropy |

# List of Tables

# List of Figures

# Chapter 1: Introduction

## 1.1 Protein engineering using evolutionary methods

Proteins are molecules that have been shaped by evolution to act in a wide variety of functions such as in catalysing a plethora of reactions, acting as sensors, forming various structures with different properties, and as transport vesicles for other molecules. As such, proteins form an attractive target for engineering so that we can utilise these versatile molecules for our own desired functions. Furthermore, in the current climate crisis, proteins are especially attractive tools as they are generally less environmentally harmful than many commonly used catalysts and materials[3-5]. This stems from their biodegradability and that they function at ambient temperatures and in mild conditions to carry out their functions. However, while it has been long known that proteins are comprised of amino acid building blocks encoded by a genetic sequence, fully *de novo* rational design of protein function from sequence still proves challenging[6,7].

One class of methods that can be utilised for protein engineering with relatively limited prior knowledge are evolution and sequence-based techniques, such as directed evolution and ancestral sequence reconstruction[8,9] (ASR). These strategies use either evolutionary information present in modern-day sequences (ASR) or attempt to mimic natural evolution processes, driven by a desired selection criterion (directed evolution). Thus, evolutionary methods can explore the protein sequence-function landscape in different ways for engineering novel protein functions or capabilities (Fig. 1.1). Perhaps the main advantage of both evolutionary techniques is that relatively little information is needed for engineering, compared to other engineering strategies[8,9]. This is because many methods, like *de novo* design, site saturation mutagenesis and computational-guided design, often require detailed structural and/or functional information or large, labelled training datasets in the case of machine learning-based strategies. Furthermore, as well as using evolutionary methods directly for engineering novel proteins, the relatively distinct ways in which ASR and directed evolution 'explore' the protein sequence-function landscape can provide unique insights into how changes in sequence can affect changes in function. This can greatly aid both in the fundamental understanding of proteins as well as in providing information useful for other engineering strategies, like rational design[10], semi-rational design[11], machine learning-guided design[12] and bioinformatics-guided design[13].

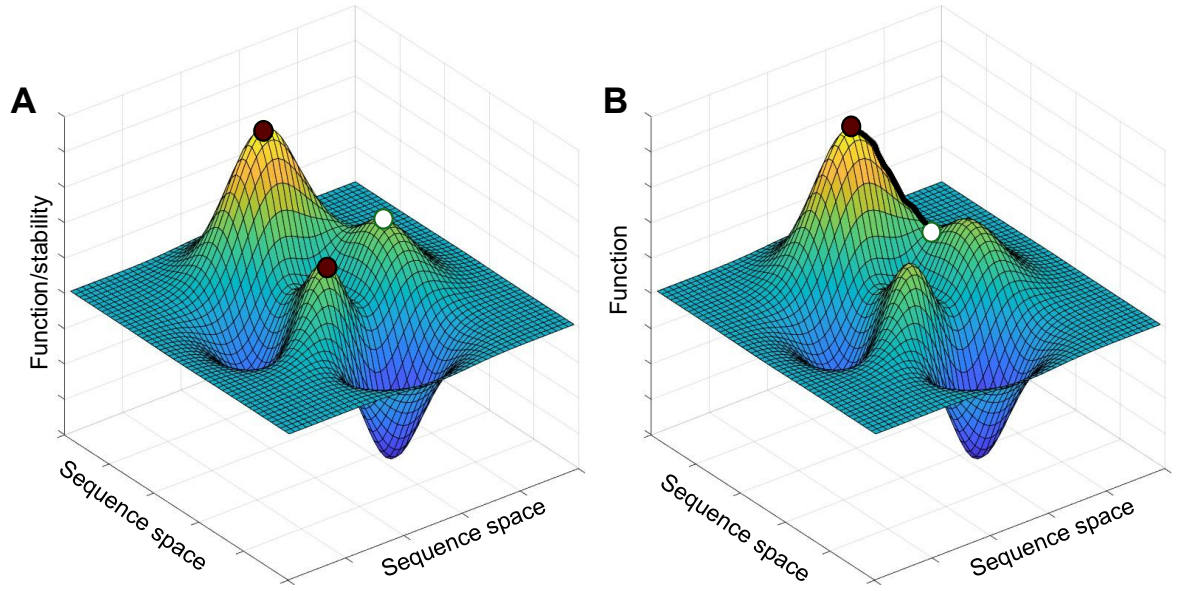



**Figure 1.1:** Protein fitness landscape exploration using A) ancestral sequence reconstruction and B) directed evolution. A study may begin with a certain protein, represented by the white circle, and aims to reach the 'peak' protein(s), represented by a red circle. Ancestral reconstruction predicts ancestral proteins, which are likely relatively stable and functional, thus, are likely 'peaks' in the fitness landscape. Directed evolution generally attempts to improve a certain function of the protein; thus, can be approximated to 'hill climbing' the fitness landscape. Figure made using MATLAB.

The evolution-guided protein engineering techniques of ancestral sequence reconstruction and directed evolution, which were used to engineer the proteins described in this thesis, will be reviewed in more detail below. Then, the emerging field of use for proteins, and thus of engineered proteins, in the creation of materials through enzyme-mediated polymerisation will be explored. This will be followed by the main aims of this thesis.

## 1.2 Ancestral sequence reconstruction

Ancestral sequence reconstruction (ASR) aims to reconstruct ancestral protein sequences to gain further insight into protein evolution, better understand modern day proteins, and to identify ancestral proteins that might have novel properties, be relatively promiscuous and/or have increased stability[14-17]. Essentially, the ancestral sequences for each node in a phylogenetic tree (Fig. 1.2A) are predicted using the information from the extant sequences, the predicted evolutionary relationship between these sequences, and an evolutionary amino acid substitution model. The concept of ASR was formalised by Pauling and Zuckerlandl in 1963[18]; however, its implementation was limited until the 2000s, with the increase in genome sequencing, gene synthesis and computational power[19,20]. Since then, ASR has been utilised in numerous ways, such as to demonstrate how catalytic activity can arise from a non-

enzymatic ancestor[21], to reveal how ligand specificity changed in steroid receptors[22] and even to engineer non-natural Kemp eliminases when combined with rational design[10].

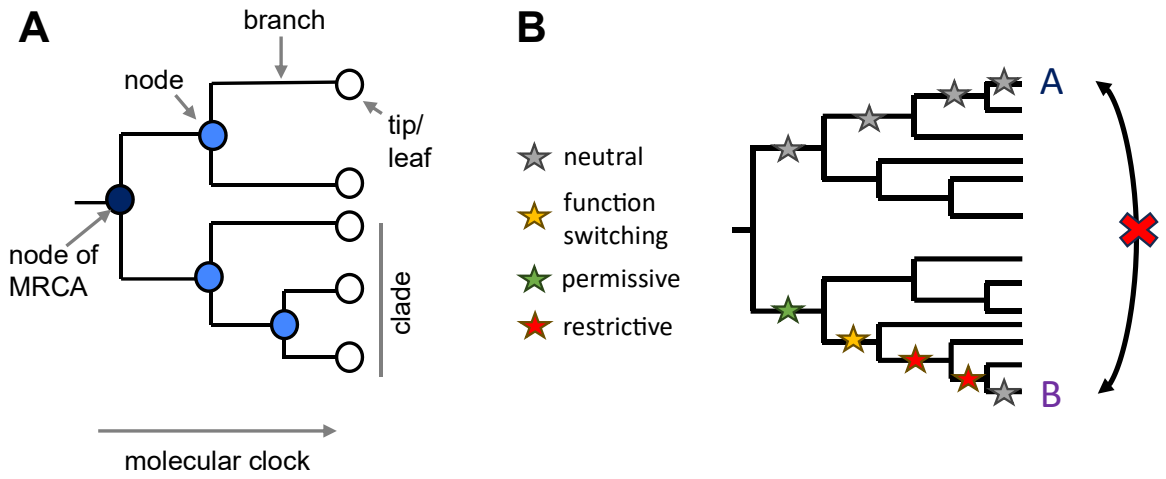



**Figure 1.2:** A) A sequence phylogeny represents the evolutionary relationship between extant protein sequences. Extant sequences are represented at the tips and are related through nodes representing common ancestors. The common ancestor of the phylogeny is the most recent common ancestor (MRCA). Branches connect the tips and nodes, and their relative length can indicate time, commonly through average number of substitutions per site. B) The 'vertical' approach of ancestral sequence reconstruction facilitates identification of functional changes by tracking the order of mutations with time. 'Horizontal' approaches may be obfuscated by the build-up of neutral mutations and epistatic interactions, like permissive and restrictive mutations. Figure was made using Microsoft PowerPoint.

Perhaps the largest advantage of ASR, compared to other techniques that only utilise extant protein sequences, is that with ASR the additional dimension of time is considered, which allows for a relatively limited number of mutational changes to be tracked in chronological order. This is beneficial because studying mutations in the order they occur can allow for much easier identification of epistatic interactions, such as permissive and restrictive mutations to a given function, as well as reducing the 'noise' of neutral mutations, which accumulate over time in modern day sequences[16,19,23,24] (Fig. 1.2B). Thus, ASR is more easily able to characterise function switching mutations, thereby 'exploring' the protein genotype-phenotype landscape in an efficient manner. This ability of ASR to predict ancestral sequences, which are likely 'peaks' on the sequence-function landscape, also allows for identification of novel proteins that are likely stable and functional. In this manner, ASR can further be utilised as an engineering tool as well as for gaining insights into how protein sequence can affect changes in protein function.

The process of ancestral sequence reconstruction typically consists of five main steps: i) sequence collection, ii) multiple sequence alignment, iii) phylogeny reconstruction, iv)

ancestral sequence reconstruction, and v) characterisation of ancestral protein(s) (Fig. 1.3). The first four steps of this process will be reviewed below to provide background for the ASR conducted in Chapter 3. Following this, relevant studies utilising ASR will be briefly discussed.

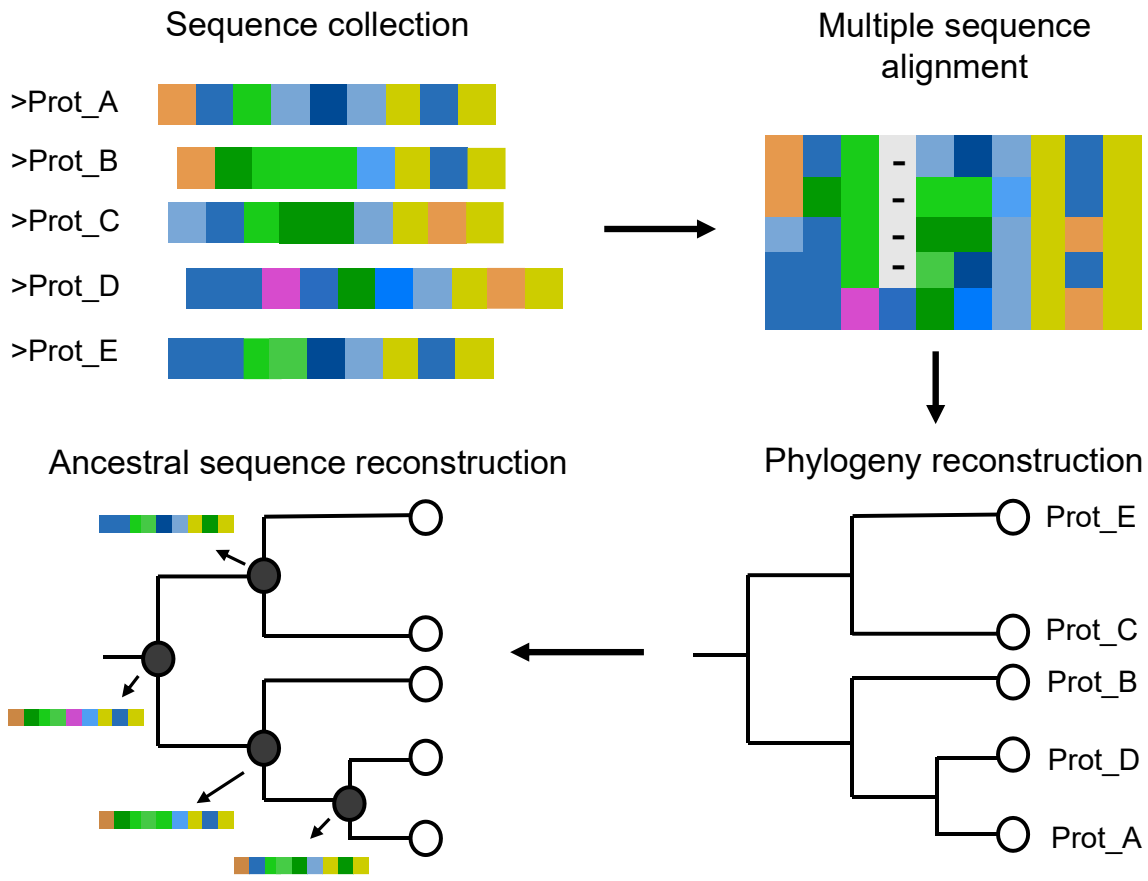



**Figure 1.3:** Overview of the computational process during an ancestral sequence reconstruction study, which consists of sequence collection, multiple sequence alignment, phylogeny reconstruction, and prediction of the ancestral sequences. Following this, typically ancestral sequences(s) will be selected for experimental characterisation. Figure was made using Microsoft PowerPoint.

### *1.2.1 Sequence collection and curation*

Once a research question has been posed that can be investigated using ASR, the first step is to find sequences of interest and then possibly to curate this sequence set before the downstream processes. In order to reconstruct a phylogeny and ancestral sequences, extant (modern-day) sequences need to be identified that have a common ancestor, and thus are part of the same 'family'. Such sequences, which are evolutionarily related, are said to be homologous. Homologous sequences can most commonly be related by orthology, paralogy

or xenology (Fig. 1.4). Orthologous sequences have descended from the same common ancestral gene but have diverged over time, paralogous sequences have diverged following duplication of an ancestral gene and xenologous sequences result from horizontal gene transfer between species[17,25,26]. As xenologous sequences are challenging for many evolutionary models used during phylogeny and ancestor reconstruction, typically a sequence set for ASR should ideally contain orthologous and perhaps paralogous sequences in order to build a more accurate phylogenetic relationship between the sequences using current models[27-29]. This can be done by searching for sequences in databases of relatively similar identity to the sequence(s) of interest, by searching curated databases of orthologous sequences for sequences of the family of interest. Following identification of a potential sequence set, software can then be used to attempt to determine the nature of the evolutionary relationships in your sequence set for further filtering of xenologous or non-homologous sequences[30-32]. It can also be informative to compare a protein sequence set with the organismal phylogeny to ensure there are no sequences from unexpected organisms. Following the identification of potentially homologous sequences, it can be beneficial to remove any sequences from the set that are fragmented, potentially frameshifted, or which do not contain highly conserved residues[19,33,34].

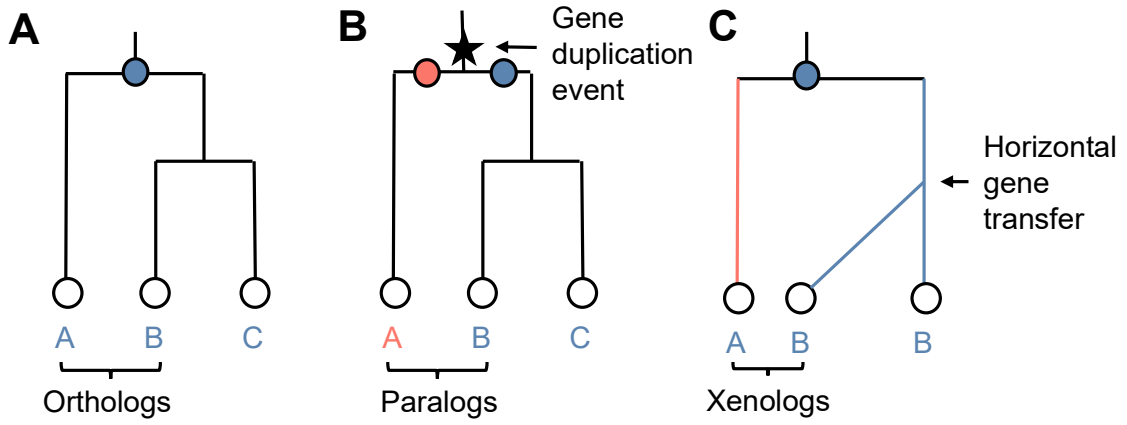



**Figure 1.4:** Homologous sequences, which are evolutionary related, can be related through different mechanisms such as A) orthology, B) paralogy and C) xenology. Figure was made using Microsoft PowerPoint.

Other than attempting to identify evolutionarily related sequences and filtering out clearly unsuitable sequences, like gene fragments, other aspects of sequence collection are typically more dependent on the research question. For example, if one is interested in more recent ancestors of closely related protein sequences, then relatively few extant sequences will be needed. However, if one is interested in studying older ancestors of early diverging sequences

a much larger, more diverse sequence set likely will be needed to reconstruct the phylogeny and older ancestral sequences more accurately[35,36]. Typically, the sequence set should be relatively diverse because sequences with very high identity may bias downstream processes and may not contain enough information to accurately reconstruct the phylogeny and ancestral sequences[37]. Also, accurate phylogenetic reconstruction may be aided by ensuring there are protein sequences from several organismal lineages[25,38,39]. On the other hand, sequences that have very low identity with other sequences in the set may also negatively affect downstream processes[37,40,41]. This is believed to be because current evolutionary models may struggle to identify the relationship between the different sequence and the rest of the set, resulting in the different sequence adding 'noise' to the reconstruction process.

A final consideration for sequence collection is that the larger the sequence set, both in terms of number of sequences and length of the individual sequences, the greater the computational power will be needed. If there is only access to a typical desktop with 4 CPU cores, conducting ASR using a sequence set of more than a few hundred sequences of medium length (~300 residues) may be impractical in terms of time requirement. Reducing the number of sequences can be conducted by removing sequences of high sequence identity, such as by using CD-HIT, or by attempting to identify sequences which are potentially the most informative, such as by using T-COFFEE[42]. However, it should also be considered that the greater the number of sequences, typically the more evolutionary 'information' will be available in the set and the greater the accuracy[43-45]. Also, it has been found that excessive sequence curation can lead to biases as well as loss of information[34,46]. Thus, it is recommended for typical ASR studies to have at least 100 sequences in the final sequence set[36].

### *1.2.2 Multiple sequence alignment*

Before phylogenetic and ancestral sequence reconstruction analysis can be carried out, the homologous (evolutionarily related) positions of the sequences need to be identified. This is because insertions and deletions (indels) in protein genes, which are particularly common in eukaryotes[47-49], result in homologous residues not being in the same position for each protein sequence. For the purposes of phylogeny construction, multiple sequence alignment (MSA) algorithms attempt to 'align' protein sequences such that each column of the alignment contains homologous residues, and any indels in a particular protein sequence are represented as 'gaps'. Common alignment algorithms attempt to identify homologous residues between different protein sequences by using a scoring matrix, which can be based on empirical

and/or statistical knowledge, combined with an affine gap penalty[49-51]. For example, empirical knowledge could include that the last codon is a 'wobble' codon, which can lead to some substitutions occurring relatively frequently. Statistical knowledge can be gained from protein sequences that are known to be homologous, which can be used to determine the likelihood of a given substitution or indel and thus weight the scoring matrix. Typical amino acid scoring matrices include percent accepted mutations (PAM)[52] and BLOcks SUbstitution Matrix (BLOSUM)[53]. While alignments can be conducted using either amino acid or DNA sequences, it is typically better to use amino acids for protein sequences because they have a slower rate of change than nucleotides, allowing deeper timescales to be studies, and the greater number of amino acids (~20) compared to nucleotide bases (4) leads to less ambiguity and thus higher accuracy[25,36,37].

Main algorithm types for protein MSA include progressive, consistency-based, structure-based and phylogeny aware algorithms. Progressive algorithms, such as ClustalΩ[54] and Muscle[55], rely on a 'guide-tree', which shows roughly how similar the sequences within the set are. Similar sequence pairs at the tips of the tree are aligned first using pairwise alignment, then progressively more distal sequences are aligned to these 'profiles' until the full MSA is built up[56,57]. Progressive algorithms are typically relatively fast, which can make them useful for large datasets of >1000 sequences, particularly if certain heuristics are used[58-60]. However, the accuracy of purely progressive alignments can be relatively low for complex datasets with relatively high substitution rates and number of indels[35,51,61,62]. This is possibly due to these algorithms getting stuck in local optima, which is particularly problematic for indel handling.

To overcome this issue, consistency-based progressive methods were developed, where in conjunction with the tree-based progressive algorithm, each pairwise comparison is also scored against a 'library' of all possible pairwise alignments to maximise compatibility of the alignment with the rest of the library[42,50]. Thus, global consistency is optimised, thereby potentially reducing the number of times the greedy pairwise algorithm is trapped in local optima. Examples of consistency-based MSA software include T-COFFEE[42], ProbCons[63], MAFFT INS-I variations[64] and MSAProbs[65]. For complex datasets, consistency-based algorithms tend to be more accurate than progressive, however, with the cost of computational time[25,57]. Hence, for many datasets, consistency-based algorithms may only be practical for less than a few hundred sequences[50].

With the increase in the number of protein structures available, alignment methods that combine the consistency-based approach with structural information have been developed. These include Expresso[41], ProMals3D[66] and MAFFT DASH[67]. While the exact algorithms between structure-based software vary, typically where possible structures of each sequence, or of a protein relatively similar in sequence, are identified from databases and used to aid the alignment. For datasets containing sequences of relatively low identity, it is possible that structure-based approaches may be superior to other approaches because protein structures are more conserved than sequences[41,68]. However, for large datasets structure-based algorithms will likely be slower than many consistency and progressive based algorithms. Structure-based approaches are also limited for protein families which lack determined tertiary structures.

To potentially combine the benefits and 'cancel-out' the disadvantages of several alignment algorithms, an integrated approach can be used where the consistency of several MSAs is maximised to generate a 'combined' MSA. An example of a software capable of doing this is M-coffee, which uses a similar consistency-based approach to T-COFFEE to combine alignments[69]. It has been found that the integrated approach can be highly accurate for certain datasets, possibly due to the 'majority rules' leading to better indel handling[68]. Thus, for protein families with few determined structures, the integrated approach could be especially useful. However, it has been found that the integrated approach can become quite inaccurate if it is biased with similar alignments that are correlated with each other, or with very inaccurate alignments that possibly lead to too much noise[68,69].

However, progressive algorithms and their more recent modifications do not explicitly model evolution when aligning sequences for the purposes of phylogeny construction. As such, they may struggle to correctly identify homologous positions, particularly when insertions or deletions have occurred[70,71]. Thus, phylogeny-aware software was developed that models substitutions, insertions and deletions in an evolution-aware manner by considering information from the guide phylogeny for software like PRANK[72], or by co-estimating the alignment and phylogeny at the same time for software including BAli-Phy[40] and Saté[73]. Based on simulated datasets using evolutionary models, the accuracy of phylogeny-aware software has been found to be very high, possibly making these algorithms a good choice for phylogeny and ancestral state reconstruction. However, it should be noted that the accuracy of PRANK is greatly reduced without providing a relatively accurate guide phylogeny[35,49],

and that software which co-estimate the alignment and phylogeny tend to be extremely computationally expensive, hence, can be impractical with sets of >100 sequences[40,62].

Overall, the selection of an MSA algorithm to use for ancestral sequence reconstruction is ultimately dependent on the dataset of protein sequences of interest. For relatively simple sets consisting of a small number of sequences with low rates of substitution and low numbers of indels, it has been found that most alignment algorithms will be equally accurate for phylogeny and ancestral reconstruction. For sets with a small number of sequences but with greater evolutionary complexity, possibly phylogeny-aware algorithms could perform best. However, most recent ASR studies contain hundreds to thousands of sequences of varying evolutionary complexity in their dataset; thus, the selection of MSA algorithm will depend on the computational power accessible, the relative complexity of the sequences and the availability of structural information. Generally, sequence sets containing a high number of indels have been found to be the most challenging because gap penalty scoring is a consistent issue for non-evolutionary-based methods[57,68].

For non-simple sequence sets it may be best to compare several MSAs. This can be done visually by inspecting MSAs for qualities like 'gappiness' and how well conserved residues are aligned; however, such an approach is subjective and irreproducible. Thus, reproducible metrics of assessing alignment quality have been developed, including the transitive consistency score (TCS)[2], GUIDANCE2[74], and MetAl[61]. Furthermore, less informative or potentially erroneous parts of the alignment can be filtered out using software such as TrimAl[75] or Gblocks[76], although it should be noted that a few studies have found that the trimming process may actually lead to decreased phylogeny and reconstruction accuracy[2,77,78].

While the relative accuracy of the multiple sequence alignment has been shown to have potentially very large impacts on downstream phylogeny and ancestral reconstruction accuracy[35,68], a few recent ASR studies have still used alignment tools that could potentially be somewhat inaccurate for their dataset[79-86]. Therefore, during the ASR conducted in Chapter 3 several alignment tools representing different algorithm types were investigated and compared, and the potential accuracy of each of the 'best' two MSAs was further characterised by separately reconstructing a phylogeny using each.

### *1.2.3 Phylogeny Reconstruction*

The next stage in ancestral sequence reconstruction is to predict how all the extant sequences of interest are related to each other through phylogenetic reconstruction. Phylogenetic reconstruction allows for the identification of how ancestral 'nodes' will be related to extant sequences through their relative position on the predicted phylogeny, as well as giving an indication of how long ago each ancestral node was predicted to exist, relative to the extant sequences or more recent ancestors, through the relative branch lengths. It has been found that the accuracy of the phylogeny can be important to the accuracy of the subsequent ancestral sequence reconstruction[87,88], although it has also been shown that the ancestral reconstruction may be robust to some degree of uncertainty in the phylogeny[89,90].

However, before a phylogeny is reconstructed, a sequence evolution model to use during the reconstruction is selected based on the sequence set. Most models are 'general time reversible', such that a substitution has the same probability of occurring in either direction, e.g. alanine -> valine is considered the same as valine -> alanine[91]. Such models consist of an amino acid substitution matrix representing the likelihood of substitutions and a vector of the equilibrium frequencies of each amino acid. In several commonly used amino acid models, these values are derived largely empirically from databases[92]. In order of their creation, amino acid models include JTT[93], Dayhoff[52], VT[94], WAG[95], DCMut[96], JTTDCMut[96] and the Q matrix family[97]. However, the use of these models alone assumes that all sites in a sequence evolve at a uniform rate, which can be violated for some sequence sets. Thus, to account for different parts of the sequence evolving at different rates, amino acid substitution models are combined with methods that account for rate heterogeneity amongst sites according to a gamma distribution (Γ)[98,99] or more complex mixture model (R)[100], with a given number of categories and with parameters typically being estimated from the dataset. Additionally, amino acid substitution models can also be modified by models which account for a proportion of sites being invariable (I)[101,102], which can be beneficial for sequences with highly conserved regions. For certain sequence sets it may also be desirable to determine custom amino acid frequencies empirically from the given sequence set (F)[103,104].

Given that many sequence evolution models exist, especially when considering the number of rate categories to use for site heterogeneity, a selection process is typically used to identify the best model for a given dataset. There are conflicting views on the importance of model selection, with some studies claiming that model selection has a limited effect on phylogeny accuracy[105-107], while other, perhaps more compelling, studies have shown that model

selection can be relatively impactful on accuracy of both phylogenetic and ancestral reconstruction[45,89,108-113]. Indeed, some studies have shown that model misspecification can lead to incongruences in phylogeny reconstruction such as topology bias and long branch attraction[114-118]. Therefore, it is generally recommended to attempt to select the 'best' sequence evolution model for a given sequence set and MSA. This is typically conducted using the corrected Akaike information criterion (AICc)[119-121] or Bayesian information criterion (BIC)[122]. The AICc calculates and compares log likelihoods of models with maximised parameters while the BIC calculates and compares approximate likelihoods for each model given the optimal parameters[113,118]. Both the AICc and BIC attempt to penalise models that might underfit or overfit the sequence dataset; however, the extent to which they do so differs between them (Equations 1.1&1.2). This can lead to a slight bias of the AICc towards more complex models and a slight bias of the BIC towards simpler models[118]. Nonetheless, it has been found that using either the BIC or AICc will generally yield similar results for phylogenetic model selection for many datasets[45]. Software which can be used for model selection include Modelfinder[103], ProtTest[104] and jModelTest[123].

$$\mathrm{AIC}c_i = -2\ \ln L_i + 2\ k_i + \frac{2k_i(k_i+1)}{n-k_i-1} \qquad (1)$$

$$\mathrm{BIC}_i = -2\ \ln L_i + \ k_i \ln n \qquad (2)$$

**Equations 1.1 and 1.2:** The corrected Akaike criterion (AICc) (1) and the Bayesian information criterion (BIC) (2), where k is the number of parameters in the model, n is the number of sites and L is either the maximum log likelihood for the AICc, or the maximum likelihood score for the BIC.

Once a model has been selected, phylogenetic reconstruction can begin. There are three main classes of phylogeny reconstruction using character-based methods: maximum parsimony, maximum likelihood (ML) or Bayesian inference. It should be noted that a fourth class, using machine learning, has recently begun to be developed[124-127]. However, testing of some current machine learning-based approaches (with non-biased datasets) indicated that they are not yet competitive with the classical methods[43,128]. Rather, the most promising machine learning-based software currently developed is designed to be used as an aid for aspects of the phylogeny reconstruction process[129-133]. Nonetheless, the focus in this work will be directed towards the most used methods of phylogenetic reconstruction at the time of writing.

Of these principal phylogenetic reconstruction methods, maximum parsimony was the earliest to be developed[134] and is essentially an 'Occam's razor' approach[20]. The aim of maximum parsimony is to identify the phylogeny that would result in the least number of possible mutations to explain the data (Fig. 1.5). However, it has been found that this model is an overly simplistic model of evolution that does not account for different branch lengths and substitution rate heterogeneity, and thus is relatively inaccurate[135-139]. Consequently, maximum parsimony has been largely superseded by the probabilistic methods of ML and Bayesian, once computational power increased in the early 2000s[19,20,34]. Nonetheless, maximum parsimony is still utilised today by some ML software to generate an initial candidate set of phylogenies in a relatively quick manner[140-142].

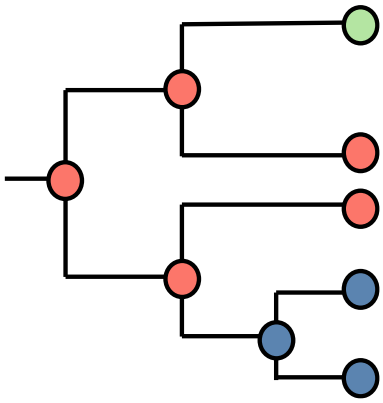


**Figure 1.5:** Maximum parsimony inference of a phylogeny attempts to minimise the number of state changes, here represented by green, red and blue circles. Figure was made using Microsoft PowerPoint.

Currently, the most commonly utilised methods for phylogeny building due to their relatively high accuracy are the statistical methods of maximum likelihood (ML) and Bayesian inference[25,28]. Both of these methods utilise more complex models of evolution, such as incorporating branch length and the sequence evolution models with rate parameters described earlier. ML attempts to identify the phylogeny with the 'maximum likelihood' of fitting the data given the optimal parameters, such as topology and branch length (Equation 1.3). ML programs include IQ-TREE[143], RAxML[141] and PhyML[142]. It should be noted that as ML programs rely on a 'hill-climbing' principle, they can at times get 'stuck' in local optima[37]. Therefore, it is generally advisable to reconstruct multiple 'maximum likelihood' phylogenies to attempt to have more confidence that the 'best' phylogeny has been found. Phylogeny construction using Bayesian inference utilises Bayesian logic to calculate the posterior probability distribution of a given phylogeny using its likelihood multiplied by factors describing a prior distribution of parameters, such as topology, branch length and

substitution model[37,144,145] (Equation 1.4). Thus, in contrast to ML, Bayesian inference is able to incorporate uncertainty in parameter estimation throughout the phylogeny reconstruction process, which results in the calculation of posterior probability confidence supports for each node. Typically, these supports will be calculated for the maximum *a posteriori* phylogeny (MAP), which is often selected for further study[145,146]. Software developed for Bayesian phylogenetic reconstruction include BEAST[147], MrBayes[148] and PhyloBayes[149].

$$L(data|tree) \quad (3)$$

$$P(tree|data) = \frac{P(data|tree) \times P(tree)}{P(data)} \quad (4)$$

**Equations 1.3 & 1.4:** Maximum likelihood phylogeny reconstruction compares the likelihoods of each phylogeny during construction (3), whereas Bayesian inference instead considers the posterior probability of a given phylogeny during construction (4).

The accuracy of ML compared to Bayesian is debated[34,87,90,150,151]. Nonetheless, Bayesian inference is typically slower than ML, especially for relatively large sequence sets, which somewhat constrains the utility of Bayesian-based phylogeny reconstruction in ASR investigation for researchers lacking access to computational power.

### *1.2.4 Phylogenetic supports for maximum likelihood phylogenies*

Once a phylogeny has been reconstructed, typically some calculations known as supports will be made. Due to the complexity of evolution, current evolutionary models for phylogenetic inference are not completely accurate and, due to the extremely large search space when reconstructing a given phylogeny, the reconstruction algorithms rely on assumptions known to be untrue and use stochastic elements to speed up the search process[37,45,46,132]. As such, phylogenetic supports are typically calculated following phylogenetic reconstruction to estimate how probable or likely a given phylogeny might be, given the information in the dataset used to create it. It is particularly important to note that low phylogenetic supports can indicate incongruence in the phylogeny[46,152]. Supports tend to be interpreted as either node (branch) supports, when examining individual nodes in the phylogeny, or topology supports, when examining the overall phylogeny, although topology support is ultimately derived from node and/or branch support calculations. Either way, the supports are generally designed to give an indication for how likely a given tree topology might be, given the dataset. While supports are commonly interpreted as ‘confidence’ in each node position or overall tree topology, in reality, due to assumptions used in calculating supports as well as their reliance

on the evolutionary model used being correct amongst other factors, the interpretation of a given support should depend on the specific support used and is never exactly equivalent to confidence. It should also be emphasised that no phylogenetic support is an explicit measure of accuracy. For most support measurements, a better interpretation of their values would be that they show how robust a given phylogenetic topology is to small changes in the dataset[153].

As with the broader field of statistics, phylogenetic supports fall into using either frequentist or Bayesian statistics. Common supports using frequentist logic include the bootstrap, the approximate likelihood ratio test (aLRT), approximately unbiased (AU) test and quartet puzzling. The bootstrap is the oldest phylogenetic support developed and is still commonly used today, albeit with more modern algorithms for its calculations[154,155]. During the bootstrapping process, pseudosamples are first generated from the data set through random selection of characters and simulation using the parameter values, then a ML phylogeny is reconstructed for each pseudosample. The bootstrap is then calculated as the percentage of samples where the original node position is supported. Historically, bootstrap calculations have been found to be conservative, with a standard nonparametric bootstrap of 80 approximately representing 95% confidence[155-157]. However, the ultrafast bootstrap (UFBoot) developed by Minh *et al*.[155] and optimised by Hoang *et al*.[158] was found to be much less conservative and the bootstrap value was found to be approximately equivalent to the confidence level for bootstraps greater than 70. Nevertheless, when interpreting bootstraps, it should be taken into consideration that the generation of pseudosamples relies on the assumption that all sites in a given sequence alignment are independent, which has been found to be categorically violated[152]. Hence especially for datasets containing large amounts of interdependence amongst sites during evolution, the bootstrap test may not be ideal. Furthermore, it should be noted that bootstrapping only becomes useful when the dataset is large enough to generate informative pseudosamples and at least 1000 bootstrap replicates should be calculated[153].

Another well used phylogenetic support is the approximate likelihood ratio test (aLRT), specifically typically the Shimodaira Hasegawa-like approximate likelihood ratio test (SH-aLRT)[142]. The SH-aLRT test was adapted from the Kishino-Hasegawa (KH) test by improving the calculation of the null hypothesis such that the SH test could be used with *a posteriori* selected candidate set of phylogenies that included the ML phylogeny[142,159,160]. The SH-aLRT compares the likelihood of a candidate set of phylogenies, including the ML phylogeny, to the null hypothesis[161]. As in bootstrapping, the SH-aLRT test utilises a

resampling of the data and then re-estimates phylogeny parameters for these samples. Following this, the SH-aLRT test then estimates log likelihood differences between the distribution calculated from the samples to the distribution calculated from the original candidate set. For a given node in a given phylogeny, the SH-aLRT value represents the proportion of replicates for which the node position has a significant log likelihood difference from the log likelihood of the define null hypothesis[157]. However, due to the assumptions made by evolutionary models and notably the assumption made during the SH-aLRT test that all trees in the candidate set are equally likely, the test tends to be quite conservative[144]. It has been found that an SH-aLRT value of 80 for a given node is approximately equivalent to a confidence of 95%[155]. Also, as with bootstrapping, the larger the number of replicates generated, the greater the statistical power of the test[161].

Due to the conservative nature of the SH-aLRT test, Shimodaira then later developed the approximately unbiased (AU) test[162]. This test attempts to correct for selection bias by changing procedure during the selection of bootstrap replicates. In the AU test, bootstraps of many different lengths are sampled, which allows for the estimation of the curvature of the selection bias and the calculation of the null distribution is then modified by this estimated parameter. While the AU test has found to be less conservative than the SH-aLRT, it has been found to lead to overconfidence in certain situations[163].

Finally, perhaps the least commonly used frequentist support algorithm is that of quartet puzzling[164]. In quartet puzzling, for each set of four taxa in the dataset (‘quartet’), the likelihood of each of the three possible unrooted topologies is evaluated and the best or multiple best topologies are stored. Then, many intermediate phylogenies are constructed by adding taxa in random order for each set of supported quartet topologies. Finally, a consensus phylogeny is generated from the intermediate phylogenies, and each node in this consensus phylogeny is given a ‘puzzle support value’ based on the percentage of times its positions occurred in the intermediate phylogenies. An advantage of quartet puzzling is that it uses relatively little computational time for relatively small datasets, particularly when compared to standard nonparametric bootstrapping or Bayesian analysis[163,165]. However, quartet puzzling has a number of issues such as being strongly influenced by how taxa are sampled and being particularly prone to incongruences such as long branch attraction[166]. As such, quartet puzzling is not commonly used for phylogenetic support.

Changing to a different school of thought, phylogenetic supports can instead be estimated using Bayesian inference. If the phylogeny itself is reconstructed using Bayesian inference, then naturally the posterior probabilities for each node position will have already been calculated for a given dataset and evolutionary model. However, as with bootstrapping and aLRT, Bayesian inference can also be used to compare the observed dataset to resampled and simulated datasets. Supports derived in this manner can be used on any phylogeny selected *a posteriori.* This method of generating supports differs from frequentist methods by generating pseudosamples using parameter values derived from the posterior distribution, rather than just the maximum likelihood parameter values[167]. Ideally, supports generated using parametric Bayesian inference should be unbiased[155]. However, it has been generally found that Bayesian supports can often be inaccurate due to their large sensitivity to model violations[157,160,168]. As evolution is complex and differs between datasets, current evolutionary models are unable to perfectly capture the process, e.g. polytomies, heterotachy, horizontal gene transfer, hybridisation, etc. Thus, Bayesian supports should be interpreted carefully given their sensitivity to inadequate evolutionary models. Furthermore, calculating posterior probabilities using full MCMC Bayesian algorithms can become very computationally expensive, especially for large datasets[157]. Hence, approximations like aBayes may be used, which takes advantage of the relative speed of the aLRT but applies the Bayes theorem to it[157]. Nevertheless, like other Bayesian implementations, aBayes has also been found to generally overestimate posterior probabilities, especially when there are large model violations[155,157].

The utility of all phylogenetic supports has nevertheless improved greatly during the last two decades; partly due to increased computational power, and partly due to the use of improved time-saving algorithms, including resampling estimated log likelihoods (RELL) and nearest neighbour interchange (NNI), amongst others[27,155,169]. Furthermore, improvements in defining the hypotheses tested by these supports, such as the null hypothesis, and improvements in evolutionary models have also increased the utility of supports[152,168]. Notwithstanding, while more recent support algorithms, such as the UFBoot, AU test and aBayes, tend to have improved performance that better represents the confidence or posterior probability one can have in the topology of a given phylogeny given the dataset, they should still be interpreted in the context of the hypotheses and assumptions that they make; *e.g.* the evolutionary model used should be ‘true’ for the dataset, which is highly unlikely to be the case. Therefore, phylogenetic supports should never be interpreted as absolute measures of confidence or

accuracy. It should also be considered that the statistical power of the support values is dependent on the amount of informative data in the dataset, as well as the number of replicates sampled, hence, supports for datasets with small numbers of taxa or supports derived from <1000 samples are generally not as meaningful.

### *1.2.5 Phylogeny curation*

Although not direct measures of accuracy, low phylogenetic supports can indicate conflicting topologies in the phylogeny, which is known as incongruence. Incongruence can result from certain assumptions made during the phylogenetic reconstruction process, including that all sequences identified are homologous, that the taxa have been sufficiently sampled, that the evolutionary model selected is accurate for the data, that substitutions are equally likely in both directions, that sites in each sequence evolve independently, and that the substitution rate is homogenous over time[167]. While xenologous sequences resulting from horizontal gene transfer are technically homologous, because their evolution can be quite different from the evolution of orthologs or paralogs, these sequences can also lead to incongruence and/or inaccuracies[170-172]. Other sources of incongruence for protein gene phylogenies include convergent evolution, where similar protein sequences are distally related but have evolved independently to be similar, recombination events, events leading to combination of domains from separate evolutionary origins, and the occurrence of a high number of indel events[25,46,152,173].

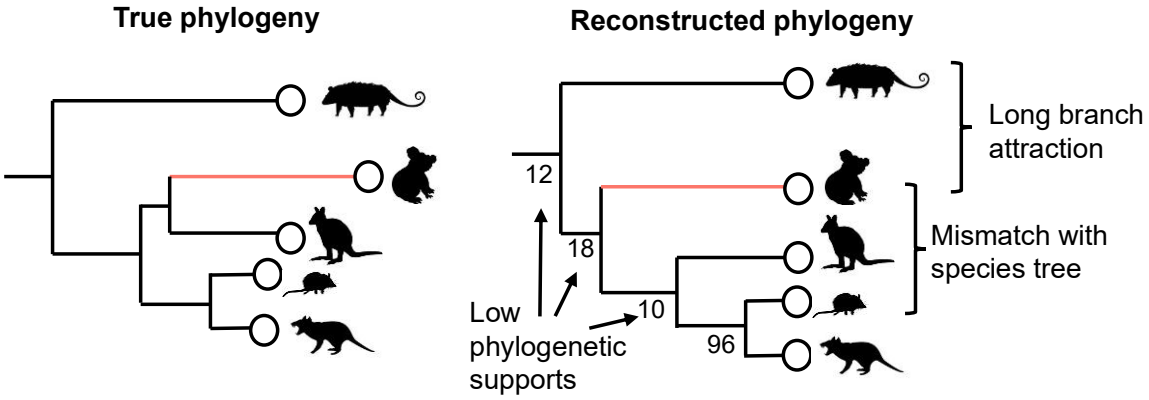



**Figure 1.6:** Potential indications of inaccuracy in a reconstructed phylogeny can include long branch attraction, low phylogenetic supports, and large mismatches between the reconstructed gene phylogeny and the corresponding species phylogeny. Figure was made using Microsoft PowerPoint.

Such incongruence and/or inaccuracy in the reconstructed phylogeny can be indicated through low phylogenetic supports, large mismatches between the protein phylogeny and the

organismal phylogeny, or grouping of long branches together, known as long branch attraction (LBA)[46,117,152,174] (Fig. 1.6). At this stage, it can be informative to remove ('trim') potential sequences that have been identified as perhaps leading to incongruence, before building the phylogeny again and comparing the old and new phylogenies. Trimming can be conducted manually; however, as it is subjective, this can lead to irreproducibility of results. Alternatively, programs have been developed for trimming like TreeShrink[175] and TreSpEx[176]. Nevertheless, any trimming process should be conducted with care because it has been found that trimming can lead to further losses in accuracy in some cases[44,78], and low branch supports or long branches do not always signify inaccuracy. Rather than removing taxa from the phylogeny, instead, sometimes incongruences can be ameliorated by adding taxa to improve sampling of clades with low support or to resolve LBA[38,39].

### *1.2.6 Ancestral sequence reconstruction*

Once the multiple sequence alignment and phylogeny have been inferred with an evolutionary model, this information can then be used to predict the sequences of the ancestral proteins. As for phylogenetic reconstruction, there are three main classes of ancestral sequence reconstruction: i) maximum parsimony (MP), ii) maximum likelihood and iii) Bayesian inference.

As for phylogenetic reconstruction, maximum parsimony sequence reconstruction is largely historical. MP attempts to reconstruct ancestral sites based on the character that would lead to the smallest number of changes[134]. However, as for phylogenetic reconstruction using MP, this approach to ASR is limited by being an 'oversimplification' because it does not incorporate an explicit model of evolution, does not account for branch lengths, and tends to arbitrarily assign sites with ambiguous character potentials[20,34,37]. Likely due to these factors, it has been found that MP reconstruction is typically much less accurate than probabilistic methods of reconstruction[136,138,139,177].

Because of the limitations of MP sequence reconstruction, the most commonly used methods of ASR are maximum likelihood and Bayesian inference. Maximum likelihood (ML) aims to identify the characters of the ancestral nodes that maximise the probability of the data (extant sequences) given a phylogeny and evolutionary model, which are assumed to be correct. Using the ML method, ancestral character reconstruction can either be conducted for a given site per each ancestral node at a time, starting from the most recent ancestors of extant sequences, known as marginal reconstruction, or for a given site at all ancestral nodes at the

same time, known as joint reconstruction (Fig 1.7). Marginal reconstruction is computationally cheaper than joint reconstruction; however, marginal methods are more prone to optimising local solutions so may not capture the global optimum like joint reconstruction[19,178].

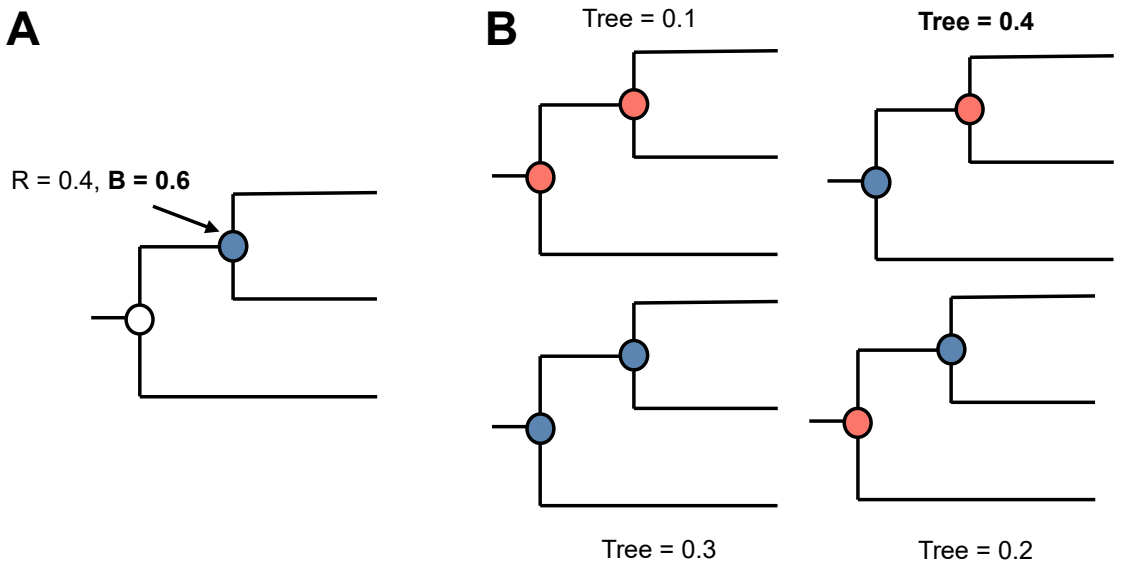



**Figure 1.7:** Sequences of ancestral nodes can be predicted using either A) marginal or B) joint reconstruction. Potential character states for a site at a given node are represented by either red (R) or blue (B). Marginal reconstruction considers only one given ancestral node at a time; however, joint reconstruction takes into account all ancestral nodes at the same time. Figure was made using Microsoft PowerPoint.

ML sequence reconstruction relies on the evolutionary model and phylogeny being correct and does not calculate the relative uncertainty associated with the predicted ancestral sequences. Thus, Bayesian inference methods of reconstruction can be utilised to incorporate such uncertainties. The most commonly used Bayesian inference methods are empirical Bayes (EB) and hierarchical Bayes (HB). Empirical Bayes, which is also referred to as the maximum marginal likelihood, determines the posterior probability (PP) distribution of characters for a given site in an internal node given the data. However, like ML, EB does not incorporate uncertainties in the phylogeny or evolutionary model, but rather fixes these parameters when defining the prior distribution[20,29]. Thus, for each site the maximum *a posteriori* (MAP) residue can be determined, as well as the posterior probability distribution, which provides an idea of confidence in the prediction. Programs that implement EB include PAML[179], PhyML[142], RAxML[141], FastML[180], GRASP[181] and ProtASR[182], which additionally incorporates structural information. However, it should be considered when using EB that both the MAP and PP are calculated under the assumption that the phylogeny and

evolutionary model are correct. Nonetheless, the reporting of the PP can be advantageous over purely ML methods because it provides some idea of the uncertainty in the prediction. While the PP distribution is useful, it has been found by Sennett *et al.*[45] that use of evolutionary models with data violations will lead to large overestimations of PP. Nonetheless, if a reasonable evolutionary model is utilised, it has generally been found that EB tends to predict ancestral sequences that are relatively close to the 'true' sequence in identity and function[45,151,183,184]. On the other hand, it has been demonstrated that EB, as well as ML, tends to overpredict ancestral stability, possibly due to the 'consensus rules' effect[185]. As for ML, EB can be conducted using either marginal or joint reconstruction methods. Marginal may be preferred for studies investigating a small number of ancestors along a subclade because the posterior probability distribution for each ancestor will be calculated, whereas joint may be preferred for applications involving characterising character changes across the whole tree as the joint posterior probability for changes across the entire tree will be calculated[29,34].

To fully account for uncertainties in the phylogeny and evolutionary model when predicting ancestral sequences, hierarchical Bayes can be utilised. This utilises a joint reconstruction method to calculate the posterior probabilities of ancestral characters at equivalent sites across the tree given the data using prior distributions of the evolutionary model and phylogeny. HB ancestral sequence reconstruction is implemented in PhyloBayes[149], MrBayes[148] and BEAST[147]. However, while HB can incorporate uncertainty in all the parameters, this results in it being computationally expensive. Therefore, use of HB is typically restricted to relatively small sequence sets. It has also been found that the conservative nature of HB when predicting ancestral sequences can lead to it being less accurate at predicting the 'true' ancestral sequence and function than EB or ML methods[151]. However, HB may yield ancestral proteins that are arguably closer in stability to the 'true' ancestor than EB or ML[87,186].

A final consideration when reconstructing ancestral sequences is how to treat and/or predict insertions and deletions (indels). Many EB reconstruction methods, such as those implemented in PAML[179], RAxML[141] and PhyML[142], treat 'gaps' representing indels as unknown characters, which results in ancestral sequences that are longer than expected because each ancestral sequence will contain a residue for each column in the multiple sequence alignment. Thus, following ASR using these methods, an additional step needs to be conducted to infer the position of gaps, which is typically conducted using maximum

parsimony[28,29]. As an alternative, ASR can instead be conducted using programs such as FastML[180,187] or GRASP[181] which explicitly model gaps and use ML-based methods to infer indels in ancestral sequences.

Regardless of the method used for reconstruction, due to the inherent uncertainty in the ML or MAP sequence of the ancestral proteins, it can be beneficial to experimentally characterise other alternative ancestral sequences if time and resources permit. This gives a measure of the robustness of the ancestral function and/or stability to uncertainties in the ancestral sequence. These alternative sequences can be sampled from the posterior probability distribution, or from using alternative method during the ancestral sequence reconstruction process, such as joint *vs.* marginal reconstruction, using a different ancestral reconstruction method, using ML *vs.* Bayesian phylogeny reconstruction, or using a different MSA method. Traditionally, for a given ancestor the sequence with the second highest posterior probability for each site was tested, known as the 'AltAll'[188-190]. However, some studies have suggested that the AltAll version is usually quite far from the 'true' ancestral sequence, and that other methods of sampling the possible sequence space are more useful[45,191]. If time and resources restrict the ability to characterise alternative ancestral sequences, several studies have suggested that the ML protein (often referring to the MAP protein determined through EB) will be close in function to the 'true' ancestral protein if a decent evolutionary model is used, particularly for more recent ancestors[45,87,89,90,151].

### *1.2.7 ASR to understand molecular evolution*

ASR studies of how proteins can change function over time have greatly expanded our knowledge of potential mechanisms of protein evolution[15,21,192-195]. As well as increasing the fundamental knowledge of proteins, such insights have also proven valuable for engineering novel proteins[11,26,196-198]. Of specific interest to Chapter 2 is how substrate specificity and dynamics can change through evolution.

Protein functional preference has been found to possibly result from either gradual changes or large changes, which has been described as either 'creeping' or 'leaping'[199]. 'Creeping' describes when the ancestral protein contains some level of a given activity that under natural selection becomes progressively stronger and/or relatively more specialised in its descendent modern-day protein. Mechanistically, gradual changes towards the modern-day function can be enabled by processes including neutral drift and gene duplication[200-203]. After gene duplication this process is often called subfunctionalization[204]. For example, a detailed study

by Harris *et al.*[205] demonstrated that the metabolism of several xenobiotics and steroids by members of the cytochrome P450 family 1 had gradually increased from activity present in the oldest ancestor studied along the mammalian lineage, likely reflecting increasing needs to metabolise these relatively novel xenobiotics and steroids. On the other hand, 'leaping' describes a process whereby the descendent protein acquires a capability that is largely not present in its ancestor. One hypothesis for retention of the novel function is through active natural selection if the novel function confers a fitness advantage, but other mechanisms may also be possible[206]. As an example of rapid functional change, Tran *et al.*[207] found that loss of antibacterial activity in a human RNase (HsR2) from an ancestor occurred relatively quickly through only four mutational changes. This study further identified that the cytotoxic activity of HsR2 also appeared to arise relatively rapidly from almost no activity present in an ancestor. It is thought that one likely method of such 'leaping' is gene duplication, because this allows for one copy of the gene to conduct the original function whilst the other can undergo larger functional changes through mutation due to not being under as strong selection pressures[195,208,209]. Following gene duplication, such rapid functional changes are often termed neofunctionalizations[204].

As well as tracking overall functional changes through ASR, it is also possible to characterise how those functional changes may have been driven through effects of mutations on protein dynamics. While the protein structure-function paradigm is important, increasingly it has been found that the dynamics and conformational landscapes of proteins also plays a major role their function[186,210-214]. How functional changes can largely result from changes in dynamics was shown by research of Kim *et al.*[215], where it was found that functional shift from a green fluorescent ancestor to a red fluorescent extant protein of GFP-like proteins from corals was almost entirely by a shift in the dynamics of the protein. Hence, Kim et al. termed the change in function as an example of 'dynamics-driven evolution'. Several other studies utilising ASR to examine evolutionary mechanisms have also demonstrated that functional shifts can be mostly driven by changes to conformational dynamics or loop dynamics, such as that of the chalcone isomerases[192], diterpene cyclases[216], solute binding proteins[194], steroid hormone receptors[217] and tyrosine kinases[218]. Furthermore, studied by Damborsky and coworkers not only demonstrated the importance of dynamics in the bifunctional nature of an ancestral luciferase/dehalogenase compared to an extant luciferase[219], but also that the ancestral luciferase/dehalogenase could be engineered to form a luciferase with very high catalytic efficiency through altering its dynamics using an indel

library[220]. Thus, these studies highlight the power of ASR to gain insights into protein functional change, which can then be utilised to engineer proteins for desired purposes.

### *1.2.8 ASR for engineering thermostable proteins*

Another application of ASR, particularly using maximum likelihood-based methods, is to predict potentially thermostable ancestral proteins that still maintain a relatively high level of a given activity[186,221-223]. It should be noted that not every predicted ancestral protein will necessarily have increased thermostability over its extant descendants. However, it has been commonly found that using ML-based methods many predicted ancestors, especially older ancestors, are relatively thermostable[189,224-226]. Proteins with increased stability can be desirable as scaffolds for other protein engineering techniques, like directed evolution, rational design and/or computational-guided design[221,222]. This is because the majority of mutations to a given protein, including function-switching mutations, tend to be destabilising[227-229]. Thermostability is also useful for applications requiring proteins with increased heat tolerance, solvent tolerance and/or product tolerance, like in industrial enzymatic chemical production[222]. Furthermore, pertinent to Chapter 3, thermostable proteins can have increased recombinant expression yield[9,26,221], which can aid with recombinant protein production.

Several studies highlighting the ability of ASR to identify thermostable proteins that still maintain similar catalytic activity to selected descendants have been conducted by Yamagishi, Akanuma and coworkers. They have been able to show that ASR can be used to increase thermostability of a range of protein families, including the β-subunit of DNA gyrase[230], 3-isopropyl dehydrogenase[231,232], glycyl t-RNA synthetase[233], and β-amylase[234]. Recently, these findings were consolidated by a relatively large-scale study by Livada *et al.*[226] comparing 56 ancestral ene reductases with 57 extant homologs, which showed that there is a likely a general trend of increased ancestral thermostability whilst maintaining a range of functional abilities. Indeed it has been found by both Risso *et al.*[224] and Gumulya *et al.*[235] for ancestral β-lactamases and ancestral cytochrome P450s or ketol-acid reductoisomerases respectively that their thermal unfolding transition temperature was up to 35 °C greater than extant sequences. Increases in thermostability of ancestral proteins has even been observed for enzyme families with relatively complex catalysis, such as for carboxylic acid reductases[236] and class I diterpene cylases[237]. While relatively few ASR studies have investigated plant protein families[34,221], those by Hendrikse *et al.*[237] and Kaltenbach *et al.*[192] looking at the class I diterpene cyclases and chalcone isomerases respectively have also demonstrated that plant

protein ancestors typically display increased thermostability compared to their modern counterparts.

Commonly, increased thermostability of ancestral proteins is identified as a benefit for industrial processes[222,223]. However, other applications have included for use of the ancestral protein in medical applications[238] and for the ability to more easily crystallise the ancestral protein for structural studies[239]. Another important benefit of thermostability is that it can lead to increased recombinant expression, which has been shown by many ASR studies[14,221], including that by Wilding *et al.*[240] for ancestral transaminases and that by Babkova *et al.*[241] for ancestral haloalkane dehalogenases.

### *1.2.9 Conclusion*

Ancestral sequence reconstruction (ASR) is a powerful tool for gaining insights into protein functional changes. ASR has revealed that such changes can occur either relatively progressively and gradually, or sometimes quite rapidly. Furthermore, ASR has often highlighted the importance of protein dynamics in the functional change of many protein families. As well as increasing our fundamental knowledge, such insights can also be valuable aid for further protein engineering efforts. In many cases, ASR can also be used to directly 'engineer' proteins by identifying more stable ancestors. Nonetheless, when analysing the results from an ASR study it should be considered that the ancestral sequence is a prediction, and not the 'true' ancestor. Also, it may be difficult to engineer a protein with a given desired function using ASR alone.

## 1.3 Directed evolution

To engineer a protein with a specific function, the technique of directed evolution can be used. This is also known as laboratory evolution, because directed evolution attempts to mimic the evolutionary processes of variation and selection in a laboratory environment. As such, directed evolution has the potential to be able to 'engineer' a protein with almost any reasonable property or function. Additionally, relatively limited prior knowledge can be necessary for a directed evolution experiment compared to other engineering techniques[9,242-244]. Furthermore, directed evolution experiments can provide important insights into how proteins are able to evolve and change function or property[227,245-248]. This is because it is possible to track all the protein 'steps' along the evolutionary trajectory, or even to characterise alternative trajectories[183,249-251]. Such characterisation not only increases our

fundamental understanding of proteins, but can also inform other protein engineering techniques like rational design[252-254] and machine-leaning guided design[255,256].

Directed evolution consists of generating genetic diversity, followed by screening or selection to identify variants with the desired function, which are then amplified such that the iterative process can be repeated until the desired target is reached (Fig. 1.8). Early 'directed evolution' experiments were conduct by Spiegelman and coworkers in the late 1960s to investigate evolution of RNA molecules[257]. However, the practice of using directed evolution to engineer proteins was formalised by Arnold and others in the early 1990s[258]. Progressive advances in techniques for directed evolution, which are still occurring, have continued to widen its range of applications[242].

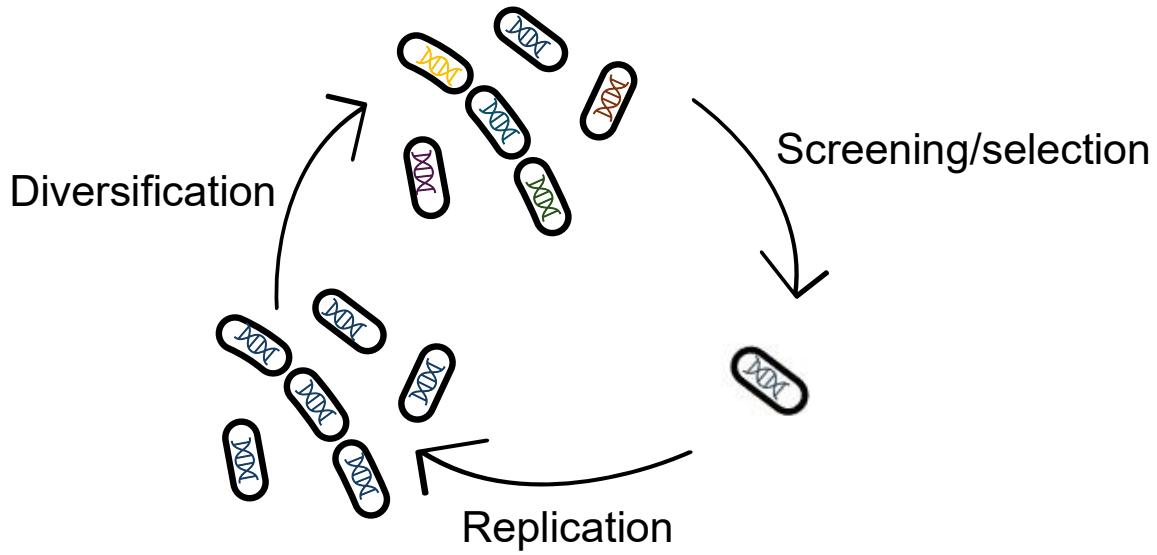



**Figure 1.8:** Main stages of a directed evolution experiment. Figure was made using Microsoft PowerPoint.

Before beginning a directed evolution experiment to engineer a protein with a desired function or property, a 'starting point' needs to be identified. While theoretically any reasonable trait could be acquired from any starting point by random mutagenesis and insertions/deletions, the enormity of the protein sequence-function space results in the probability of this occurring being vanishingly small. Thus, typically many directed evolution experiments attempt to identify a protein which has a similar function or property to the target one, or perhaps which already has some level of the target trait[9]. Another strategy can be to use structural, chemical, dynamic and/or machine-learning guided information to identify a suitable 'scaffold' protein as a starting point, where likely relatively few mutational

changes will be needed to confer the desired trait. Evolution-guided approaches can also be used to guide selection of a starting point, since promiscuous and thermostable proteins tend to be the most evolvable[8,9,259,260]. Thermostable proteins are theorized to be advantageous because most mutations have been found to result in a destabilising effect, including those conferring functional or property changes[227,261]. Promiscuous or generalist proteins are thought to be more evolvable because they already have the capability to conduct several functions; hence, generally fewer mutational changes will likely be needed to 'specialise' one of these pre-existing functions[262,263]. One method of identifying a suitable promiscuous and thermostable protein that has shown success is through ancestral sequence reconstruction[231,234,237].

Following the selection of a starting protein(s), the two main processes to consider when designing a directed evolution experiment are how to generate genetic diversity and how to screen or select for the desired function or property. Variation in a given population of genes is needed as the source upon which selection can act, yet genetic variation occurs on a much slower timescale than is practical for a laboratory setting[244]. It should also be considered that it is impractical to sample the full sequence space of a given gene, for example, there would be $4^{1000}$ permutations for a gene of 1000 nucleotides. Hence, directed evolution experiments require processes that both accelerate genetic diversification and sample the sequence space in a reasonable manner. Different ways to generate genetic diversity include through random mutagenesis, recombination and through insertions and deletions (indels) (Fig. 1.9). Such genetic variation can either be conducted using *in vitro* techniques, like error prone PCR[264], Mu transposon-based techniques[265,266] and DNA shuffling[267], or *in vivo* techniques, such as error prone plasmid/polymerase systems[268,269], multiplex automated genome engineering (MAGE)[270] and CRISPR/Cas9-based systems[271,272]. *In vitro* techniques tend to be more commonly used, perhaps because they are typically more controllable[242-244,273]. However, *in vitro* techniques are more laborious because they necessitate more steps like transformation, and consequently the library size can be limited for practical reasons. On the other hand, *in vivo* techniques can be comparatively more efficient and thus allow larger library sizes. However, when attempting to evolve a single protein gene the *in vivo* methods can result in high background mutations to the plasmid or host organism, or the mutation window can be relatively short[242-244,273].

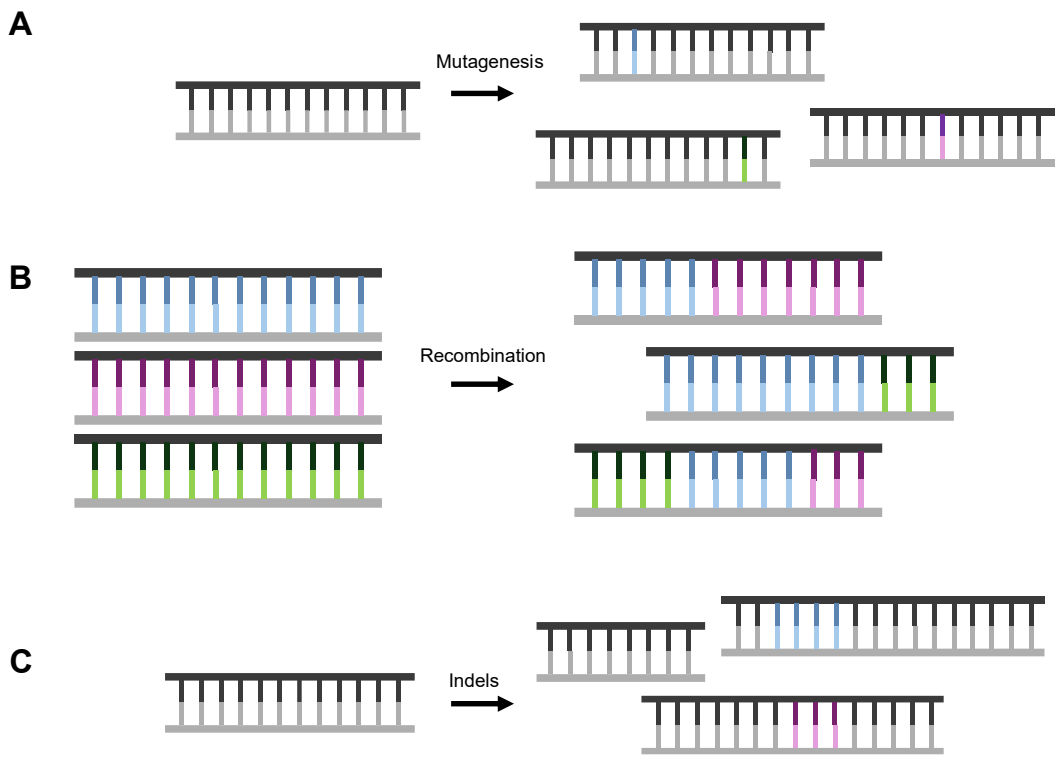



**Figure 1.9**: Overview of methods for generating variant libraries during directed evolution. A) Random mutagenesis. B) Recombination. C) Creating insertions and delections (indels). Figure was made using Microsoft PowerPoint.

Overall, using random mutagenesis, random indel incorporation and recombination enables for random diversification across a target gene, which can be beneficial if little is known about that gene or if sequence-function relationship for the desired trait is not well understood. However, due to the large sequence-function space of a given gene and limited practical library size, important areas of the sequence-function 'landscape' may be missed. Therefore, one can limit the sequence 'search' space to regions that are believed to be important for the target property or function through site saturation mutagenesis (SSM)[274]. In SSM, prior knowledge of the sequence-function relationship is utilised to identify sites that are most likely to lead to the desired trait, and these sites are then mutated randomly to generate a library of variants. Prior knowledge can include protein structure, molecular dynamics simulations of the function or property of interest, and machine-learning guided predictions[242,275]. Mutation of the selected site(s) is conducted using a library of primers, which can be randomised over all four nucleotides for each site per codon or constrained using prior knowledge to be smaller and more efficient. For example, an NNK (N= A,G,C,T and K = T,G) library for a given codon may be employed due to the redundancy of the genetic code[276]. Additionally, an even smaller, potentially more beneficial, library can be created using additional knowledge of a given codon site, such as chemical, structural or

dynamic information. In order to potentially combine beneficial effects of multiple sites, SSM of one site may be followed by SSM of other sites of interest, in a process known as iterative saturation mutagenesis[275].

Following generation of a library of variants, the next important step is then to either screen or select for variants with the desired function or property. Selection tends to be faster and is closer to natural evolutionary processes, yet this relies on the ability to link the desired trait with organismal fitness[242,244,277]. This can include by directly linking the desired protein function or property to the survival or growth of the organism. Alternatively, it can be possible to artificially link the target trait with selection, such as by physical separation of cells displaying proteins which can bind to a desired substrate[278,279] or through fluorescent activated cell sorting of cells that display fluorescence[280]. By combining selection with *in vivo* genetic diversity creation, continuous directed evolution can be conducted, which is very efficient and relatively much less laborious[281,282].

However, as it is not always possible to link a desired phenotype to fitness, a screening process can be utilised instead[242,244,273]. During screening, cells with a given variant are isolated, such as by plating or microfluidics, then an assay or other technique is used either directly on the cells, or on the cell lysate. For certain properties, the protein variants may have to be buffer exchanged before screening, although this is less desirable due to the extra time and input required. Screening techniques aim to detect a desired substrate affinity, product production, or other trait through methods including imaging, colorimetric assays, fluorometric assays, high performance liquid chromatography (HPLC) and mass spectrometry (MS). Screening can be greatly aided by using high-throughput techniques such as colony picking and liquid handling robots, microfluidics and fluorescence activated cell sorting[283]. Recently, relatively high throughput MS methods have also been developed to screen for substrates or products that would not be detected by other techniques like fluorimetry or colorimetry[242]. Nonetheless, certain properties, like enantioselectivity, still remain difficult to screen for in a high throughput way[284].

Directed evolution has been widely utilised to engineer proteins with a large variety of phenotypes[242,243,258]. Relevant to Chapter 4 is the use of directed evolution to improve recombinant protein expression. Early experiments by Arnold and coworkers demonstrated that directed evolution could be utilised to improve the recombinant expression of proteins in organisms like *E. coli* and *S. cerevisiae*, including *Fusarium* galactose oxidase and

horseradish peroxidase[285,286]. Many studies have since shown that directed evolution can be utilised to improve recombinant expression of many proteins. Indeed studies by Alcalde and coworkers have shown that directed evolution of several fungal enzymes with interesting redox catalysis can improve their expression in *S. cerevisiae*[287-292]. Many of these studies also show that this *S. cerevisiae* directed evolution strategy can be combined 'in tandem' with *K. phaffii* or other yeasts to further improve expression yields[288,289,291-293].

***Conclusion***

Directed evolution can be a powerful technique to engineer proteins with a desired function and/or property if a suitable protein as a starting point is identified as well as a suitable selection or screening method. It should also be considered that many directed evolution experiments require the iterative screening of relatively large numbers of protein variants, which can be relatively laborious and resource intensive. Furthermore, it may be necessary to 'allow' some variants with relatively neutral or even slightly deleterious functions in some cycles due to the 'rugged' fitness landscape of many proteins. Nonetheless, directed evolution has been successfully used numerous times to engineer a plethora of proteins with various properties and activities; some of which would be useful for the enzymatic-driven polymerisation of materials.

## 1.4 Enzyme-mediated radical polymerisation

An emerging application of proteins, and thus of evolutionary-guided protein engineering, is in enzyme-mediated radical polymerisation (EMRP). EMRP describes the formation of materials due to enzyme catalytic action. While technically this process could also incorporate many *in vivo* material processes, of interest to this thesis is the use of enzymes to catalyse backbone polymerisation primarily *in vitro*. The advantages of utilising enzymes to catalyse the formation of such materials over traditional chemical approaches include that enzymes can work under relatively mild conditions, are non-toxic, are biodegradable, and can be relatively selective[5,294,295]. Enzymes can mediate several different polymerisation reactions, including condensation and ring opening; here the focus will be directed towards radical polymerisations.

Radical polymerisation describes processes where polymers are formed through progressive radical additions of free radical monomeric units. Typically, radical polymerisation reactions follow three main processes: initiation, propagation and termination (Fig. 1.10). Many vinyl

monomers can undergo radical polymerisation, for example, vinyl chloride, acrylamide and styrene. Polyphenols, like catechin, can also form through oxidative coupling of radical species[5,296,297]. Typically, enzymes can 'mediate' radical polymerisation processes either by participating in the initiation stage (Fig. 1.10), and/or by oxygen scavenging, because oxygen can 'quench' the radical propagation process[294,298,299]. Hence, enzymes can allow radical polymerisations to take place in open air, which otherwise can be a limitation of these reactions due to equipment and time requirements for oxygen purging. Furthermore, using enzymes to initiate polymerisation rather than chemicals can be advantageous because on top of being more environmentally friendly and non-toxic, enzymes can catalyse the production of radicals at a relatively steady rate over time, such that the polymerization rate can be more easily controlled[298,300].

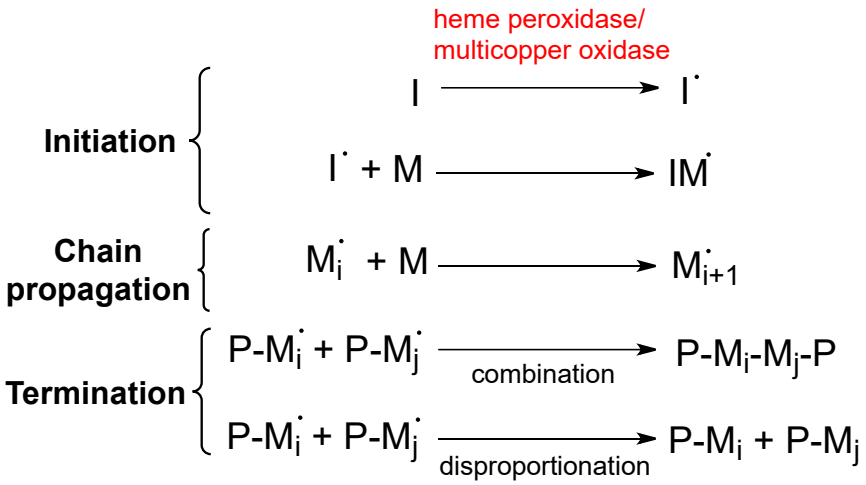



**Figure 1.10:** Basic stages of a chain growth radical polymerisation. Initiation can be catalysed by enzymes from the heme peroxidase or multicopper oxidase superfamily, typically through a small molecular mediator. Termination can occur either through combination or disproportionation. Figure was made using ChemDraw.

Consequently, since an early experiment by Parravano in 1951[301], a rapidly growing number of studies have been utilizing enzymes for various radical polymerisations *in vitro*[5,298,302]. It has been demonstrated that enzymes can be used to form a variety of materials, from relatively simple polymers like polystyrene[303] and polyacrylamide[304,305], to more complex triblock copolymer vesicles[306,307] and diblock nanoobjects[307]. It has also been shown that enzymes can be utilised in more complex radical polymerization reactions, such as reversible addition−fragmentation chain-transfer polymerisation-induced self-assembly (RAFT PISA)[308] and photo-enzymatic RAFT[309].

Despite the growing use of enzymes in radical polymerisations, a relatively limited range of enzymes tend to be used for most studies; most commonly horseradish peroxidase (HRP) and glucose oxidase (GOx). Notwithstanding, most enzymes used for enzyme-mediated radical polymerisation fall under the oxidoreductase enzyme class. While other enzyme classes have been 'utilised', typically the cofactors of these enzymes are extracted from the enzyme and used to drive the reaction, which is not strictly 'enzymatic' as claimed[310-312]. Nonetheless, oxidoreductases can enzymatically mediate radical polymerisation processes due to their ability conduct the desired functions of catalysing radical generation and/or reacting with oxygen for oxygen scavenging. For the role of catalysing polymerisation initiation either directly or using mediators, most enzymes utilised generally fall under the heme peroxidase or multicopper oxidase superfamilies[5,313] (Fig. 1.11).

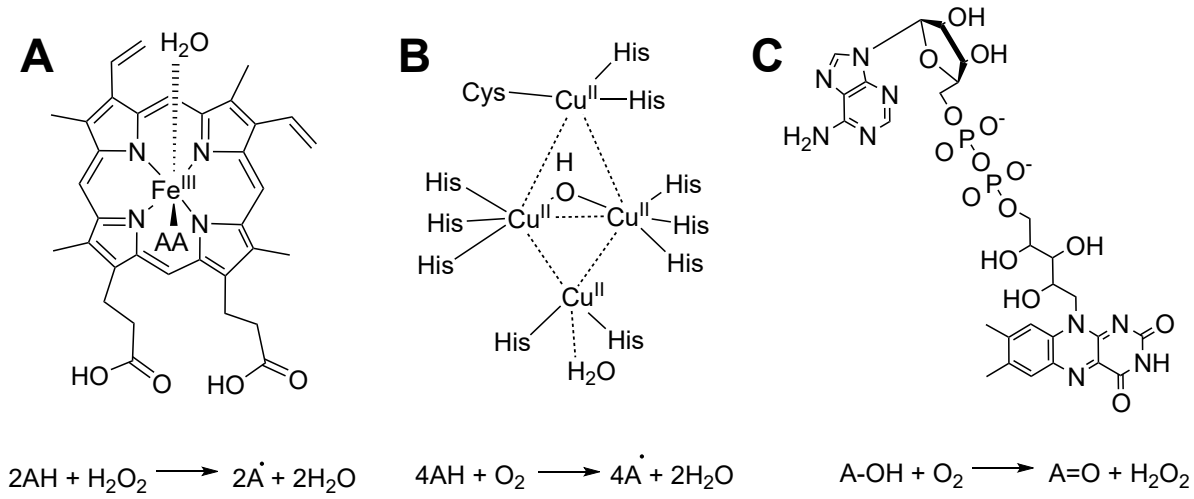



**Figure 1.11:** Overview of the chemically important moieties in enzymes commonly used for enzyme-mediated radical polymerisation and schema of relevant reactions they can catalyse. A) Active site of heme peroxidases contains an Fe-protoporphyrin IX[314]. This allows for their catalysis of two single electron redox transfer. B) Active site of multicopper oxidases contains a multicopper centre[315]. They are capable of catalysing four single electron redox transfers. C) FAD cofactor, which is present in the active site of several oxidases. Some oxidases are able to oxidise a substrate through reducing oxygen to hydrogen peroxide. Figure was made using ChemDraw.

Heme peroxidases (EC 1.11.1) include enzymes that encompass a range of enzymatic activities, but commonly many can catalyse the oxidation of a substrate through the reduction of hydrogen peroxide (Fig. 1.11A). This redox reaction can involve single electron transfer, which can accordingly generate a free radical product[313,316]. Of the heme peroxidases, horseradish peroxidase (HRP) is by far the most commonly utilised[5,294,313]. This is possibly due to its relatively low cost, reasonable thermostability and solvent stability, relatively good substrate range and high activity. HRP has been utilised in a wide array of polymerisations,

which broadly cover many vinyl monomer and phenol polymerisations[5]. To name but a few examples, HRP has been used to catalyse polymerisation of poly(ethylene glycol) methacrylate (PEGMA)[308], poly(N-(2-hydroxypropyl) methacrylamide)[306], poly(N-isopropylacrylamide) (PNIPAm)[317], polyaniline[318] and poly(hydroquinone)[319]. While HRP has been the most commonly utilised, a few other studies have demonstrated that other heme peroxidases can also initiate radical polymerisations, such as soybean peroxidase[320], *Coprinus cinereus* peroxidase[321], *Caldariomyces fumago* chloroperoxidase[322,323] and an *E. coli* catalase[324].

Some multicopper oxidase superfamily proteins are also able to generate radicals to mediate polymerisations. The most used enzyme class has been laccases, which are able to oxidise substrates through single electron transfers by reducing oxygen (Fig. 1.11B). This can be an advantage, as heme peroxidases typically require $H_2O_2$ addition. However, the requirement of oxygen for laccases can be detrimental, because excess oxygen can lead to quenching of the radical reaction[5,313]. This may have led to their relatively lesser popularity, compared to HRP. Nonetheless, laccases have still been utilised in a wide range of polymerisation reactions, including many vinyl monomers and phenols, and furthermore their different chemistry can open up different applications to those which are possible for HRP[325,326]. For example, laccases have been used to polymerise polyacrylamide[304], *N*-vinylimidazole[327] and artificial urushi[328]. While less utilised, a few studies have also shown that another multicopper oxidase, bilirubin oxidase, is also capable of mediating polymerisation[329-331].

Another class of enzymes also commonly used in EMRP are oxidases[294,298,316]. Except for the multicopper oxidases, many of the oxidases used contain an flavin adenine dinucleotide (FAD) cofactor and catalyse the reduction of oxygen to hydrogen peroxide and the oxidation of a substrate (EC 1.1, 1.2 and 1.5; Fig 1.11C). As most of these redox reactions involve two electron transfers, a free radical substrate is not generated. Hence, most of these oxidases used in EMRP indirectly initiate polymerisation through a cascade and/or act as oxygen scavengers[294,298,316]. The ability of oxidases to scavenge oxygen is especially beneficial as it can prevent oxygen quenching and thereby allow the radical polymerisation reaction to occur in open air. Furthermore, as several oxidases catalyse the production of hydrogen peroxide, a bienzymatic cascade can be formed using such an oxidase with a heme peroxidase. As well as scavenging atmospheric oxygen, using an oxidase to generate $H_2O_2$ *in situ* also reduces the probability of the heme peroxidase being deactivated by excess $H_2O_2$[332]. The most used oxidase for EMRP is glucose oxidase (GOx), which has been used for oxygen

scavenging[333,334], by itself as part of chemical cascades[335,336], and as part of bienzymatic cascades like with HRP[308,337]. The common use of GOx could be due to its relatively low cost, high thermostability, high solvent tolerance and high activity[338]. Nonetheless, other oxidases which have been utilised include pyranose oxidase[309,339], alcohol oxidase[340] and sarcosine oxidase[341]. Also, Li *et al*.[342] recently showed that formate oxidase could be used with HRP, but with the added benefit that its product, $CO_2$, was easily separated from the solution.

Overall, enzyme-mediated radical polymerizations are becoming increasingly utilised, largely resulting from the 'greener' nature of enzyme-driven chemistry that is of vital importance given the current climate crisis[5,294,313]. However, challenges to the field include that the relatively low stability of enzymes compared to chemical equivalents can result in lower reusability and lower solvent tolerance, which in turn means that reactions with hydrophobic monomers are less tenable. Such issues could perhaps be overcome using enzyme engineering, especially through evolutionary methods. Nevertheless, even using natural enzymes it has been shown that EMRP can be used to form a wide range of interesting materials. As many enzyme-mediated polymerisation reactions involve relatively hydrophilic monomers, often the resultingly hydrophilic polymers can absorb water to form hydrogels.

### *1.4.1 Hydrogels*

A class of materials that can be formed through enzyme-mediated radical polymerisation are hydrogels. Hydrogels are materials which consist of water and polymer network(s), which are at least partly hydrophilic. The aqueous content of hydrogels can range from at least 10% of their weight to over 99%[343,344]. This relatively high water content of hydrogels contributes to some of their interesting properties, including their viscoelasticity, hysteresis and biocompatibility. Additionally, many hydrogels display stimuli responsive behaviour like swelling or deswelling in response to factors like pH[345], temperature[346], light irradiation[347] and ionic strength[348]. Accordingly, hydrogels have been used in numerous applications in fields ranging from medicine to soft robotics and engineered living materials[349]. Following early characterisation studies in the 1960s[350], one of the earliest commercial uses of a hydrogel material was in contact lenses[351]. Since then, hydrogels have also been used in numerous applications including wound dressings[352,353], drug delivery[354,355], as scaffolds for tissue engineering[356,357], to form 'smart' window covers[358], as part of living bioreactors[359] and as soft robotic actuators[360].

Often, the polymer network of a hydrogel is crosslinked, which can occur through noncovalent and/or covalent bonding interactions. These are also referred to as physical and chemical crosslinks[344,349]. Noncovalent hydrogels may be crosslinked through ionic interactions, hydrogen bonding, and hydrophobic interactions (e.g. London dispersion forces). On the other hand, as the name suggests, covalent crosslinks are created through covalent bonds. These covalent bonds can be created through a wide variety of chemical reactions, including through free radical chemistry. Thus, enzyme-mediated radical polymerisation can be utilised to catalyse the formation of both the polymer backbone and/or crosslinks of many hydrogels. For example, the enzyme cascade of glucose oxidase and horseradish peroxidase (HRP) was demonstrated to catalyse the radical polymerisation of PNIPAm with *N,N′*-Methylenebisacrylamide (MBA) crosslinks[361]. As another example, it has also been shown that bienzymatic systems of horseradish peroxidase with various oxidases can also polymerise just the diphenol crosslinks of a modified alginate through a radical mechanism[362].

While crosslinking can improve the mechanical properties of a hydrogel, further benefits can often be gained by combining two (or more) polymer 'networks' together to form a double network (DN)[363-365] (Fig. 1.12A). In this way, double network hydrogels have been synthesised with up MPa tensile and compressive strength, thereby increasing the range of applications for which hydrogels can be used. Double networks may consist of two covalently crosslinked networks, such as the first double network synthesised by Gong *et al.*[366] that comprised of poly(2-acrylamido-2-methylpropanesulfonic acid) and polyacrylamide. Alternatively, double networks can consist of an ionically crosslinked network and a covalently crosslinked network, like the polyacrylamide and $Ca^{2+}$-alginate hydrogel synthesised by Sun *et al.*[367]. For both types of double network, the increased mechanical strength compared to the constitutive single networks may derive from the increased number of bonding interactions between the two networks, and the ability of one network to act as a 'sacrificial' network and/or 'hold' the other broken network together, as in the Brown-Tanaka model[363,365,368,369]. However, it has generally been found that double networks consisting of both a physically crosslinked network and covalently crosslinked network show greater elastic recovery than those with two covalently crosslinked networks, likely because non-covalent bonds are much more easily reformed than covalent bonds[363,370]. Nonetheless, the formation of both types of double networks can be a complex and somewhat laborious process, which is important to optimise to maximise the mechanical strength. Also,

many double networks have been formed through radical mechanisms using relatively toxic catalysts and with at least one stage requiring oxygen removal, however, these limitations could possibly be resolved through the use of enzymes[294,298,300,371,372].

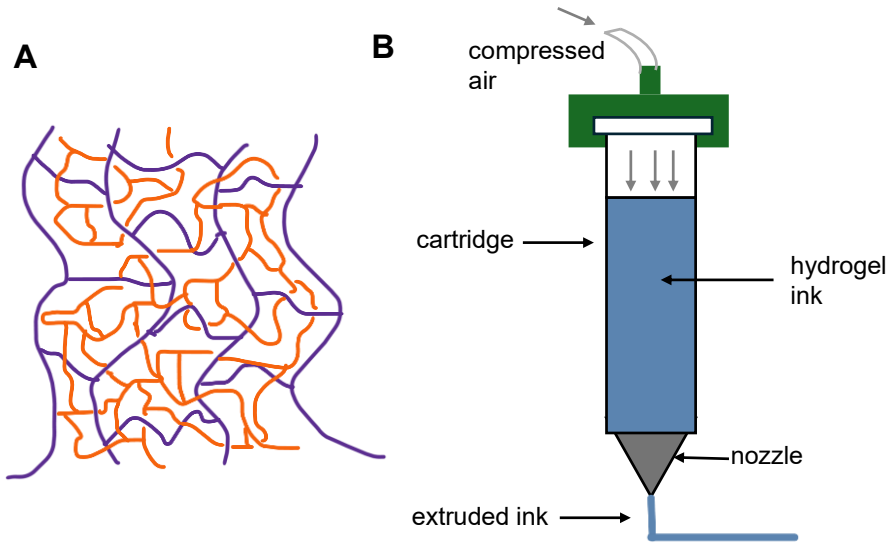



**Figure 1.12:** A) Double network hydrogels consist of two polymeric networks, represented by the orange and purple colours. B) Overview of pneumatic-driven extrusion 3D printing. Figure was made using Microsoft PowerPoint.

Another factor that can widen the scope of applications for hydrogels is their 3D printability. 3D printing is an additive manufacturing process using computer-aided design to build custom 3D constructs[373-375]. 3D printing allows for precise spatial control, which can be beneficial for building complex materials for tissue engineering[376], cell co-cultures[377], specific actuation responses[378], *etc*. 3D printing can be conducted in a variety of different ways. For hydrogels, extrusion-based printing is commonly used, likely due to its affordability and simplicity[373,374,379]. For hydrogels that are viscous and shear-thinning, extrusion printing can be conducted using a pneumatic-driven system (Fig. 1.10B). However, for bioprinting applications that involve cells it should be considered that printing cell-laden hydrogels in this manner will lead to cellular stresses through shear forces[359,373]. Nonetheless, by incorporating cells in potentially 3D-printed hydrogels, engineered living materials can be created.

### *1.4.2 Engineered living materials*

The biocompatibility of hydrogels has resulted in their frequent use in engineered living materials (ELMs)[359,380,381]. Engineered living materials are a relatively recent class of

materials that attempt to incorporate the benefits of synthetic biology and living cells with advantageous material properties. Some aims of ELMs include that they possess properties shown by living organisms like self-healing, self-assembly, responding to environmental cues, and dynamically switching between different states[359,381,382] (Fig. 1.13). These properties are then combined with the bulk properties of the material, such as mechanical strength, structure and stimuli responsiveness. It is hoped that such 'smart' materials could prove more adaptable than traditional materials and thus, on top of a wider range of applications, may have a longer lifespan and be more ecofriendly. Furthermore, this 'green' nature of ELMs can be improved by utilising biodegradable materials, which can even be synthesised by the cells using relatively low energy 'food' in some cases[383,384]. Indeed, often synthetic ELMs are inspired by natural 'ELMs' that consist of such 'smart' materials made and maintained by living cells using relatively little energy; for example, wood, coral, nacre, biofilms, mycelium and diatom glass [382].

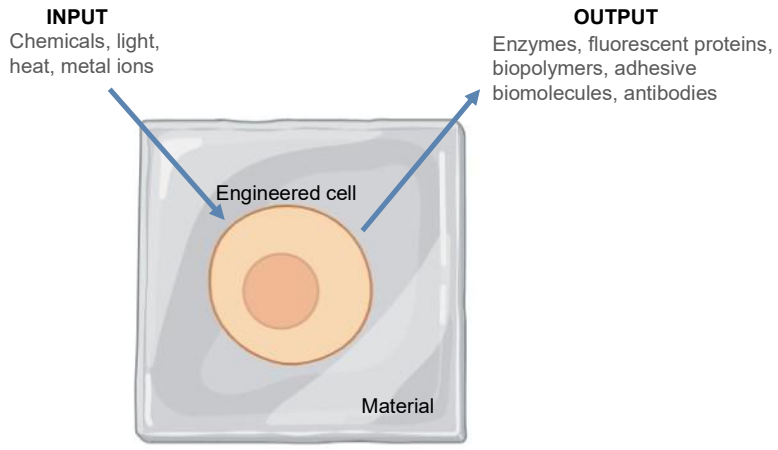



**Figure 1.13:** Engineered living materials can take advantage of biological properties by incorporating engineered living cells into a material. Figure was made using BioRender.

ELMs have a wide range of applications, including in bioremediation, infrastructure, healthcare and green energy production[359,382,385,386]. An early example of an ELM was the development of a 'self-healing' concrete using microbially induced calcium carbonate precipitation[387]. Since then, various ELMs have been developed, such as various bioreactors that can produce proteins or catalyse the production of value-added chemicals 'on demand', biosensors like one that could sense and remove from the environment the biotoxin and pollutant mercury[388], the light responsive living material biosensor composed of bacterial cellulose[384], a blood-responsive living glue[389], a living gel to treat fungal skin infections, and

algal hydrogels that can produce hydrogen using solar energy[390]. However, while there are many novel and exciting applications of ELMs, they also face some challenges. Perhaps the biggest concern is ensuring the safety of ELMs given that they contain living cells, which are often genetically modified. Hence, containment of the cells within the material is important, as well as possibly encoding a 'kill-switch' as a failsafe[359,391,392]. Also, ELMs tend to be more costly than traditional materials, and many could be potentially hard to scale for industrial use[381,382]. Nonetheless, if these challenges can be faced then ELMs have much potential for use in many fields.

To develop an ELM for such varied applications, a system needs to be developed that consists of both living cells and a material 'scaffold'. For the strict definition of an ELM, the cells should be 'engineered' in some way to confer a desired behaviour or property[380]. Many commonly used material scaffolds in ELMs tend to be hydrogels or to be hydrogel-like because of the ability of hydrogels to diffuse the water, nutrients and gasses needed by (most) living cells[359]. This often hydrogel-like scaffold of an ELM can be a biopolymer synthesised by the cells themselves, hence having the property of being a 'self-assembling' ELM, or can be a separately synthesised material to which the cells are added. These latter ELMs are often classed as 'hybrid' ELMs. Hybrid ELMs can still be composed of natural polymers, like the commonly used alginate, or they may be composed of synthetic polymers to take advantage of certain properties these polymers may have[382,385]. Formation of hybrid ELMs can be conducted using a variety of methods, such as through moulding, electrospinning, dip coating or 3D-printing[382,385]. However, 3D-printing can be particularly beneficial in allowing customisability for various applications, as well as spatial control, which can be beneficial for hierarchical organisation of the material and positioning the microbes as desired[373,386]. The ability to spatially control the relative positions of the microbes can be especially advantageous for co-cultures[393]. Lastly, addition of cells to the material scaffold for hybrid ELMs can be conducted either before formation of the material by mixing the cells into the material precursor solution, or after material formation through absorption or adsorption.

One class of cells that can be utilised in hybrid ELMs for various applications are the single celled fungi known collectively as yeast. Advantages of utilising yeast in ELMs include that they are relatively easy and cheap to maintain, are relatively easy to manipulate using synthetic biology, can produce eukaryotic posttranslational modifications, possess eukaryotic secretion pathways, can undergo anaerobic fermentation, can potentially be stored for long time periods and have relatively decent 're-usability'[393-395]. Additionally, many yeast strains

are generally regarded as safe (GRAS)[396-398]. By far the most commonly utilised yeast species in ELMs is *Saccharomyces cerevisiae*[386], perhaps due to its availability and the wide range of synthetic biology tools developed for it. Nonetheless, some examples of other yeast species include the study by Yuan *et al.*[399], where an ELM bioreactor using *Komagataella phaffii* was developed, that by Wang *et al.*[400], who developed a self-assembling biosilica-yeast ELM for removal of biotoxins using *Yarrowia liplytica* and the recently the study by Yuan *et al*.[401] where *Kluyveromyces marxianus* was used to produce fragrance molecules within calcium-alginate beads.

Perhaps one of the earliest studies where a yeast hybrid ELM was developed was by Fine *et al*.[402] in 2006, before the term 'ELM' had been adopted[380]. In this study it was shown that *S. cerevisiae* within calcium-alginate beads or PVA/PEG could produce luciferase if oestrogen was detected, thereby acting as a biosensor for potential applications in healthcare. Other yeast hybrid ELMs include those developed by Johnston *et al.*[393], where *S. cerevisiae*-laden F127- bisurethane methacrylate hydrogels were either hand extruded or extrusion 3D-printed to form a yeast ELM bioreactor that was capable of producing 2,3-butanediol. It was then further shown that *S. cerevisiae* co-cultures could be formed within an ELM to increase fermentation efficiency or, when combined with *E. coli*, to produce betaxanthins efficiently. Other key examples of yeast hybrid ELMs are described in two related studies by Sugianto *et al.*[403] and Altin-Yavuzarslan *et al*.[404], where *S. cerevisiae* were embedded in bovine serum albumin-polyethylene glycol diacrylate and the resulting cell-laden gel ink was 3D-printed using stereolithography. It was then demonstrated that the resulting ELM could act as a bioreactor to produce betaxanthins or proteinase A in a controlled, repeatable manner and that the ELM could even be printed into functional gears. Nonetheless, relatively few yeast hybrid ELMs have been developed thus far, hence, there are many potential directions for these materials in the future.

### *1.4.3 Conclusions*

Overall, enzyme-mediated radical polymerisation (EMRP) can be used to form a wide variety of polymeric materials for many applications in a relatively eco-friendly manner. However, it is currently somewhat restricted by its use of largely natural enzymes, which may not be thermostable, solvent stable and/or substrate specific for certain types of polymerisation reactions. Engineering such proteins could possibly by conducted using evolutionary-guided techniques, because both ancestral sequence reconstruction and directed evolution can identify stabler proteins with novel activities. Nonetheless, one recent but rapidly growing

field where EMRP could be utilised is in generating hydrogel ‘scaffolds’ for engineered living materials in a relatively ‘green’ manner compared to traditional UV/chemical initiator approaches. Furthermore, because ELMs often utilise proteins expressed by the contained cells in various ways, protein engineering through evolutionary approaches could further expand the range of properties and uses of ELMs

## 1.5 Research Aims

The overarching goal of this thesis is to push the boundaries of ancestral sequence reconstruction (ASR) both as a framework for understanding fundamental protein properties, such as the evolution of ligand binding thermodynamics, and as a technology to engineer “challenging” enzymes, particularly those with complex post-translational requirements like plant glycoproteins. By examining how ancient versions of modern proteins might exhibit altered stability, binding affinity, or catalytic properties, this work aims to reveal new insights into protein structure–function relationships and open up innovative applications in biotechnology. These include expressing plant-derived enzymes that are difficult to produce in standard hosts, integrating engineered proteins into advanced biomaterials, and ultimately forming engineered living materials capable of novel functions.

Building on this motivation, Chapter 2 investigates whether changes in ligand binding observed among modern members of the *lac* repressor (LacI) family could be more clearly understood by resurrecting ancestral LacI proteins. Specifically, the chapter addresses the thermodynamic underpinnings of ligand specificity and whether shifts in binding affinity over evolutionary time are governed by distinct enthalpic or entropic contributions. Individual mutations are also explored to clarify the sequence–function relationship, aiming to map how particular substitutions alter the energetic landscape of ligand binding.

A separate application of ASR is taken up in Chapter 3, where it is hypothesized that recombinant expression of horseradish peroxidase (HRP), a glycosylated plant enzyme, can be improved using ancestral variants. Since HRP is commonly employed in enzyme-mediated radical polymerization (EMRP), it is further tested whether the recombinant ancestral peroxidase can fulfil the same role as modern HRP in forming poly(*N*-isopropylacrylamide) (PNIPAm), a thermo-responsive hydrogel with diverse industrial and biomedical uses.

Finally, Chapter 4 expands upon the use of engineered enzymes in EMRP by examining whether a peroxidase evolved via directed evolution could also be harnessed to catalyse PNIPAm formation within an alginate/PNIPAm double network. This chapter then explores the possibility of establishing a 3D printed engineered living material, in which embedded microorganisms (*S. cerevisiae* or *K. phaffii*) could transform the hydrogel environment into a bioreactor or biosensor.

In summary, these three chapters illustrate both the fundamental value of ancestral reconstruction in dissecting how protein functions, such as ligand binding, emerge and diversify, and the practical potential of leveraging ancestral or engineered enzymes to advance contemporary biotechnological applications. By demonstrating ASR's utility with respect to difficult to express plant enzymes and integrating newly designed proteins into self-assembling biomaterials, this thesis seeks to further bridge the study of protein evolution with synthetic biology and materials engineering.

# Chapter 2: Evolution of effector binding in LacI family transcription factors

## 2.1 Abstract

Transcription factors of the *lac* repressor (LacI) family are important tools in bacterial synthetic biology for applications including biosensing, protein production, and genetic circuits. However, engineering novel transcription factors that respond to desired 'signal' ligands has been challenging, largely due to the allosteric networks of these transcription factors. To overcome this difficulty in traversing the complex fitness landscape of ligand binding in transcription factors, here the ligand binding of previously reconstructed LacI family ancestors was characterised in detail along the LacI trajectory using ligand binding screens, isothermal titration calorimetry and tryptophan fluorescence assays of single mutants of the oldest ancestor. This revealed a gradual specialisation towards binding of β-galactosides from weak binding activity present in the oldest ancestor. The importance of considering thermodynamics and enthalpic-entropic 'trade-offs' for future engineering efforts was demonstrated through changes in the binding thermodynamics accompanying this binding strength change. Additionally, it was found that single mutations had a complex effect on binding changes in LacI family ancestors, thereby highlighting the capability of ancestral sequence reconstruction to gain insights into difficult protein fitness landscapes.

## 2.2 Introduction

*Lac* repressor (LacI) family transcription factors (TFs) have many applications in synthetic biology, for example, in biosensing[405-409], controlling gene expression 'on demand'[410-413], controlling metabolic flux for complex chemical production[414-417], and for developing novel biological circuits[418-421]. LacI family proteins, like other transcription factors, regulate gene expression by binding to certain DNA elements, known as operators or promoters. This binding can result in the repression, activation and/or initiation of genes controlled by that element[422,423]. One of the earliest transcription factor systems characterised was that of the *lac* repressor protein and the *lac* operon by Jacob & Monod[424] (Fig. 2.1). They found that LacI regulated expression of the *lac* operon in *E. coli* by binding to an operator sequence, which

prevented transcription of the downstream genes. However, if certain metabolites were present, such as allolactose, it would bind to LacI and result in LacI losing its affinity for the DNA operator sequence, thereby allowing transcription to occur. This regulation of the *lac* operon by LacI enables the *E. coli* cell to only produce proteins necessary for β-galactose metabolism if they are present in the environment, which reduces unnecessary energy expenditure[425-427].

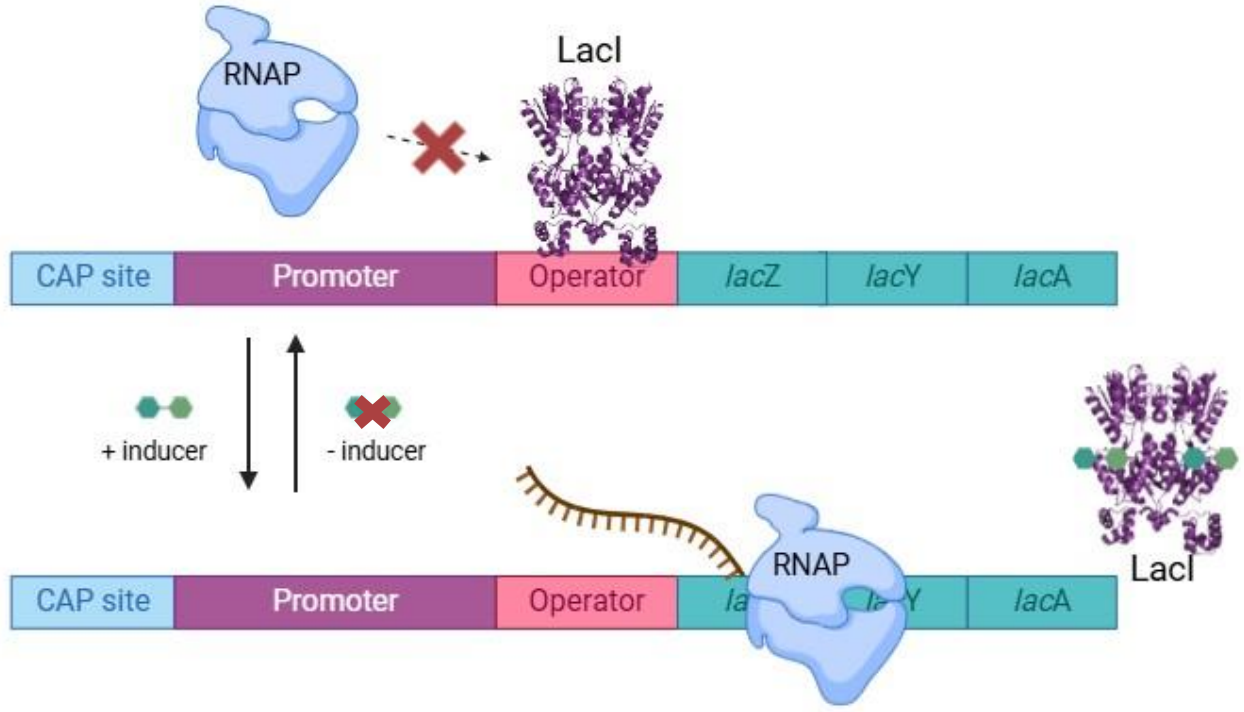



**Figure 2.1:** Overview of the regulation of expression of the lac operon in *E. coli* by the *lac* repressor (LacI). Without its inducer, LacI represses gene expression by 'blocking' RNA polymerase (RNAP) from transcribing through binding of LacI to a segment of DNA known as an operator. When an inducer is present in the *E. coli* cell, such as allolactose, LacI loses its affinity for the operator sequence and the downstream genes are transcribed. Note that dimeric LacI is shown here for clarity. Figure made using BioRender and PyMOL.

More generally, it has been found that the LacI family of transcription factors is a widespread and diverse protein family, with members possibly present in all extant bacteria[428,429]. LacI family proteins typically regulate aspects of the carbon metabolism in bacteria[430,431]. LacI TFs can act as repressors or co-repressors, like LacI or the purine repressor (PurR)[432,433], or as both repressors and activators, such as the *gal* repressor (GalR) and carbon catabolite protein A (CcpA)[434-436]. Whilst some LacI family members modulate the expression of only a handful of genes, such as the *rbs* repressor (RbsR) and *gal* isorepressor (GalS)[437,438], others are 'global' regulators which control the expression of hundreds of genes, such as the catabolite repressor/activator (Cra)[439,440].

Despite the wide diversity of LacI family proteins and their relatively low sequence similarity, the tertiary structure has found to be largely conserved[428,441] (Fig. 2.2A). This structure consists of a DNA binding domain (DBD) and ligand binding domain (LBD) per monomer unit, with the monomer unit forming part of a homodimer or homotetramer in order to bind to DNA[428,442-445]. The DBD of each monomer consists of a highly conserved, likely evolutionary ancient, helix-turn-helix motif, which is able to bind to half of an operator sequence, with the other half being bound by the DBD of another monomer unit[446-448]. Tetrameric LacI family members typically consist of a 'dimer of dimers', with each dimer being capable of binding to an operator sequence at the same time[449,450](Fig. 2.2B). This capability of binding to two operators can cause the DNA loop, thereby making it more difficult for downstream genes to be transcribed and expressed[429,451]. The ligand binding domain of LacI family members is composed of two subunits, termed the N- and C-terminal subdomains[452]. The two subdomains are quite similar in structure, typically consisting of four α-helices surrounding a β-sheet[428].

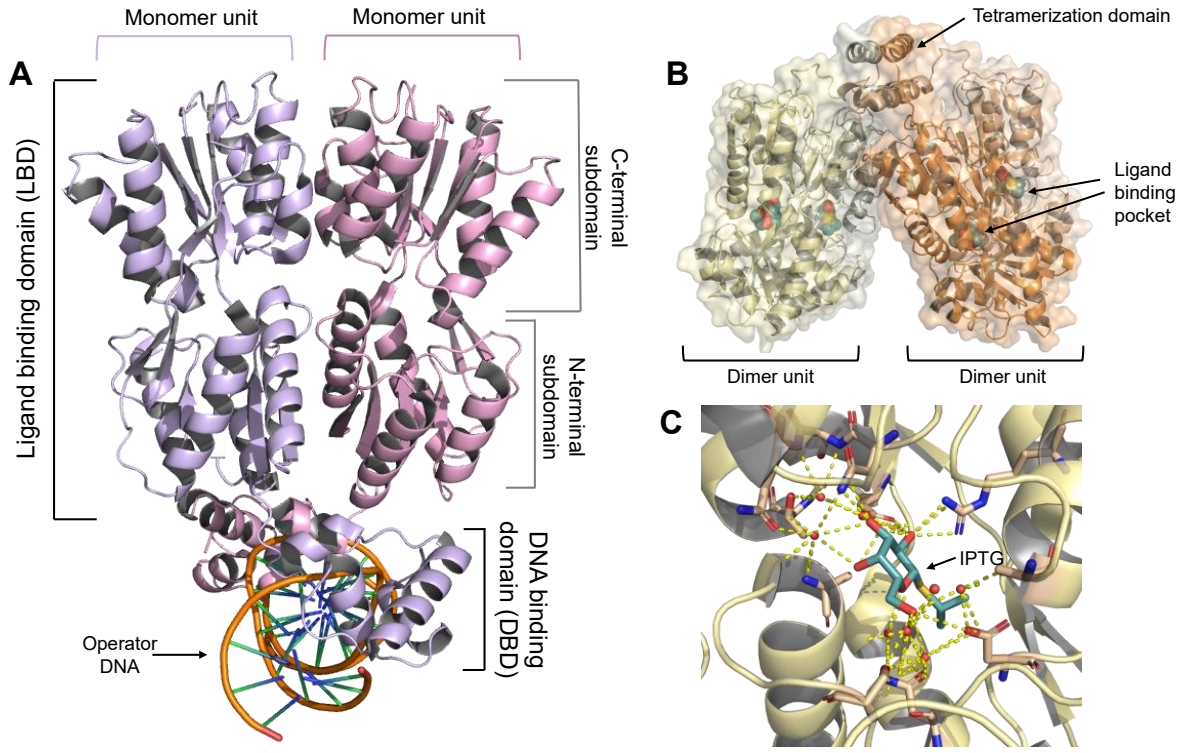



**Figure 2.2***:* Structural features of typical LacI family members as shown by LacI. A) Structure of dimeric LacI (PDB: 1EFA) bound to DNA[453]. The homodimer consists of two monomeric units, which each have a ligand binding domain (LBD) and DNA binding domain (DBD). The ligand binding domain contains and N-terminal subdomain and a C-terminal subdomain, which are somewhat symmetrical. B) Some LacI family members like LacI form homotetramers. Shown here is the ligand binding domain of tetrameric LacI[454] (PDB: 1LBH), which consists of a dimer of dimers. The ligand binding pocket for each monomeric unit is shown to be in between the N- and C-terminal subdomains. C) Structure of LacI bound to a

gratuitous inducer, isopropyl β-D-thiogalactopyranoside (IPTG)[452] (PDB: 2P9H). A relatively extensive hydrogen bonding network forms upon binding of IPTG, shown by the dashed yellow lines. Water molecules involved in this network are shown by the red spheres. Figure made using PyMOL.

Binding of 'effectors' to the LBD is how the ability of the DBD to bind DNA is modulated, which is key to transcriptional regulation. Effectors of LacI family members include sugars, purines and small proteins[428,430,450]. Many LacI family members possess a ligand binding pocket in between the two subdomains of the LBD where small molecule ligands bind (Fig. 2.2C). However, larger effectors like proteins typically bind to the outside of the LBD[443,455,456]. Most LacI family members will bind to one effector per monomer, however, it has been found that some LacI family proteins are able to bind to two distinct effectors per monomer unit. For example, it was found that one way in which the *in vivo* function of CcpA can be modulated is by whether just the phosphorylated small protein HPr binds to each monomer unit, or if both HPr and glucose-6-phosphate bind to each monomer unit[443,457]. Based on studies of LacI[458], the DNA binding ability of LacI family members may only be modulated if both monomers of a dimer unit bind to an effector.

The ability of LacI family members to change their DNA binding strength through binding of an effector to a relatively distal site results from allosteric communication between the ligand binding site(s) and the DNA binding domain[459-461]. Such allostery is key to their function as regulatory proteins. However, identifying residues involved in allosteric networks and how they are able to convey the 'signal' of ligand binding to cause functional change in a relatively distal part of the protein can be difficult[431,462,463]. For LacI, inducer binding causes the DBD to become relatively disordered through networks that cause a shift in the ensemble of conformations sampled by LacI[452,464,465]. While the static structure of the LBD does not change much with inducer binding[452-454], it was found that regions around the ligand binding site and regions of the N-terminal subdomain become more rigid, resulting in more interactions between the N-terminal subdomains of each monomeric unit[452,464]. Thus, Glasgow *et al.*[464] proposed that propagation of the 'signal' of inducer binding in LacI occurs through the ligand binding making regions near the ligand binding pocket becoming more rigid. This in turn makes the regions of the N-terminal subdomain near the monomer-monomer interface more rigid, which then makes the hinge helix of each DBD, present in the monomer-monomer interface, more flexible and disordered. However, allosteric communication networks likely differ between different LacI family members because it has been found that many 'chimeras' of various LBDs and DBDs from LacI family members no

longer show allosteric responses to their respective effectors[419,449,450,466]. Furthermore, for several LacI family members it has been shown that only certain ligands are able to induce an allosteric response, with other ligands binding but resulting in a neutral or anti-inducing effect[467-471].

Due to this relatively fine-tuned allostery of LacI family members[450,472], it has proven difficult to engineer new transcription factors using a LacI family scaffold. Several studies have created chimeras of various combinations of LacI family LBDs and DBDs; however, many of these were found to be non-functional or to have a relatively weak allosteric response compared to native family members[419,449,450]. Other approaches to engineer novel ligand binding into LacI family TFs have included rational-based design such as computer aided design and site saturation mutagenesis[406,418,420,473]. However, limitations of these studies included that only binding of ligands very similar to the native ligand could be engineered, the allosteric response of the engineered proteins was sometimes weaker, the engineered protein was sometimes found to not be specific to its new function, and all studies required high-throughput screening of large libraries of variants. Rondon *et al.*[405] utilised another protein engineering technique, directed evolution, to engineer novel caffeine binding into a TF based on PurR. However, this also required screening of large libraries to design binding of a ligand that is chemically and structurally similar to the native ligand of PurR.

As these previous engineering efforts to design novel ligand binding in LacI family TFs have faced difficulties, here it was proposed to use the different technique of ancestral sequence reconstruction (ASR) to characterise changes in ligand binding in the LacI family and provide insights for future engineering efforts. Unlike methods probing only modern-day proteins, ASR can track mutations and their effects on function over time; thus, providing an extra dimension of information. As many residues in a given protein will have intragenic epistatic interactions[213,474,475], the ability of ASR to track the order of mutational changes and their effect on function makes it easier to characterise how function switches take place compared to 'horizontal' studies of extant proteins, where the context of permissive and restrictive mutations that have led to changes in function is much more difficult to discern[19,23,28,29,34]. Additionally, extant proteins have typically accumulated many neutral mutations, which makes it harder to identify which mutations between two given extant proteins have led to functional changes than by comparing related ancestral proteins, because these ancestors will typically have relatively few mutations between them[19,28,248]. Furthermore, using ASR to predict ancestral sequences, which are likely 'peaks' in the

protein fitness landscape[186,476,477], is a relatively low-throughput, 'smart' method of exploring and characterising the most informative parts of the protein fitness landscape. This ability of ASR to identify 'peaks' in the fitness landscape is particularly powerful for proteins with a relatively complex genotype-phenotype landscape, like the LacI family TFs[472,478].

Previously, ASR of the LacI family was conducted by Meger *et al*.[479] and utilised to characterise the ability of ancestral DBDs to bind to the *lac* operator when combined with the LacI LBD. It was found that there was large variation in binding of the *lac* operator by the ancestral DBD-LacI chimeras, thereby providing support to other studies showing that the DNA-binding fitness landscape of the LacI family is relatively 'rugged'[472,480-482]. However, as it has been found to be less difficult to engineer novel DNA binding than ligand binding in LacI family TFs, this study instead focused on ligand binding characterisation of the ancestors previously reconstructed by Meger *et al*.[479].

In this Chapter, selected full-length ancestors along the LacI trajectory were characterised to better understand changes in ligand binding in the LacI family and how these changes occur for future engineering efforts. It was found that the relatively weak binding of the last universal common ancestor to methyl β-D-galactopyranoside gradually became stronger over time along the LacI trajectory, likely reflecting increasing specialisation towards binding β-substituted D-galactosides. The importance of thermodynamics in this change in ligand binding was highlighted through a shift from entropically driven binding in the oldest ancestor to enthalpically driven binding in the extant LacI. Finally, by characterising the effect of single mutations on function between two of the oldest selected ancestors, further evidence was shown for the complex genotype-phenotype landscape of LacI family proteins and that this might be driven by epistatic residue interaction networks.

## 2.3 Results

The ancestral sequence reconstruction of the LacI family was conducted previously by Meger *et al.*[479] and utilised here for further investigation of ligand binding changes. Due to the size of the constructed LacI phylogeny and corresponding number of ancestral nodes, in-depth characterisation of the ligand binding of all reconstructed ancestors would be difficult. Therefore, four ancestral proteins of decreasing age were selected along the LacI trajectory and named Anc1 to Anc4 accordingly (Fig. 2.3). This clade was selected because LacI is the most studied extant protein[450,483,484]. The four selected ancestral sequences were calculated by

Meger *et al.*[479] to have posterior probabilities >0.84 (Suppl. Fig. 2.1) and their respective nodes were supported by an UltraFast Bootstrap >90 (Fig. 2.3), indicating reasonable confidence in the sequences of the selected ancestors and their respective positions in the phylogeny.

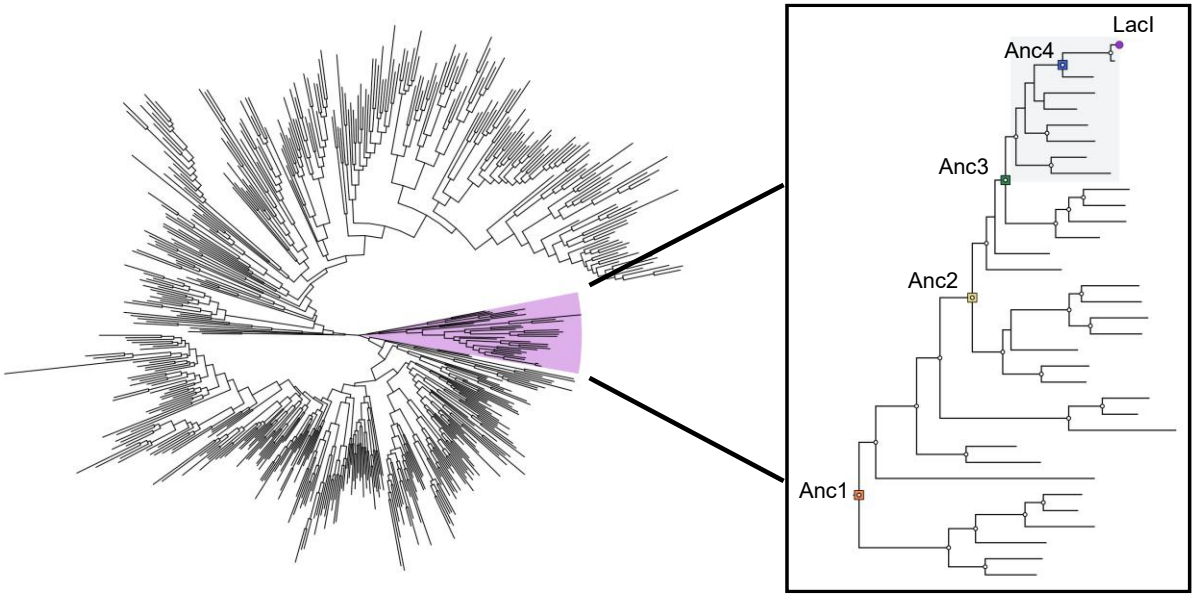



**Figure 2.3:** Phylogeny of LacI family proteins[479] with the LacI clade highlighted in purple and show in more detail on the right. Four ancestors were selected along the LacI trajectory, shown on the right. The presence of a white circle on a node indicates that it was predicted to have an Ultrafast Bootstrap[155] >90, and thus that the position is fairly likely given the dataset and evolutionary model selected. Figure made using the ggtree package[485] in R.

### *2.3.1 Basic structural and thermostability characterisation of the selected ancestral transcription factors*

Firstly, given that the selected ancestral sequences are predicted, it was important to verify that the selected ancestral proteins could be expressed in a soluble, folded form. It was found that the ancestral proteins could be recombinantly expressed in *E. coli* with relatively high yields (Suppl. Fig. 2.2; Suppl. Table 2.1), and that their circular dichroism spectra were similar to that of the extant LacI (Fig. 2.4A). As the structure of LacI family members is relatively highly conserved[428,450,486,487], the similarity in the circular dichroism spectra of LacI and the ancestors suggested that the ancestors folded into similar structures to that of LacI. However, prediction of the relative proportions of secondary structure elements for both LacI and the ancestors (Suppl. Fig. 2.3) likely over-predicted β-strand content, possibly due to the relatively high number of turns present in LacI and likely present in the ancestors[488,489].

Despite the similarity in the circular dichroism spectra between LacI and the selected ancestors, it was found that they displayed different unfolding behaviour with temperature (Fig. 2.4B). While LacI and the most recent ancestor characterised, Anc4, were found to possess similar thermal unfolding transition temperatures ($T_m$) ~66 °C and appeared to have mostly unfolded by 90 °C, the older ancestors, Ancs1-3, had relatively higher predicted $T_m$s and appeared to have not fully unfolded by 90 °C (Fig. 2.4B&C). This relatively high thermostability of Ancs1-3, especially Anc1, could indicate that they were present in a hotter environment. Indeed, the age of Anc1 has been estimated to be around three billion years old[479], when potentially life was present in a hotter environment[490-493]. However, the relative thermostability of the older ancestors could also result from biases in the evolutionary algorithms used during the ASR process[9,34,68].

Having established that the selected ancestors possessed secondary structure elements, their tertiary structure was then predicted using AlphaFold2[494,495]. As there are several LacI family structures present in the protein data bank used for AlphaFold's training[428,496,497], it was expected to be fairly accurate. The dimeric form of each ancestor was predicted because this is the minimal functional unit for LacI family transcription factors[428,498,499]. As for the greater LacI family where sequences with relatively low pairwise sequence identity typically fold into the canonical LacI family structure[428,450,486,487], it was found that all of the selected ancestral proteins were predicted to have very similar structures (Fig. 2.5A-D). Pairwise root mean square deviations (RMSDs) less than 2 Å were calculated, despite pairwise sequence identities of less than 50% between Anc1 and Anc4 (Suppl. Table 2.2). The predicted ancestral structures were also very similar to extant LacI family members like LacI (Suppl. Fig. 2.4). For Anc1 and Anc4, X-ray structures of the ligand binding domain (LBD) have been characterised by Georgelin[500] and it was found that the full protein predicted structures of these respective ancestors were also relatively similar to the experimentally characterised ligand binding domains (Suppl. Fig. 2.4). This, thus supported the likelihood of the structure predictions.

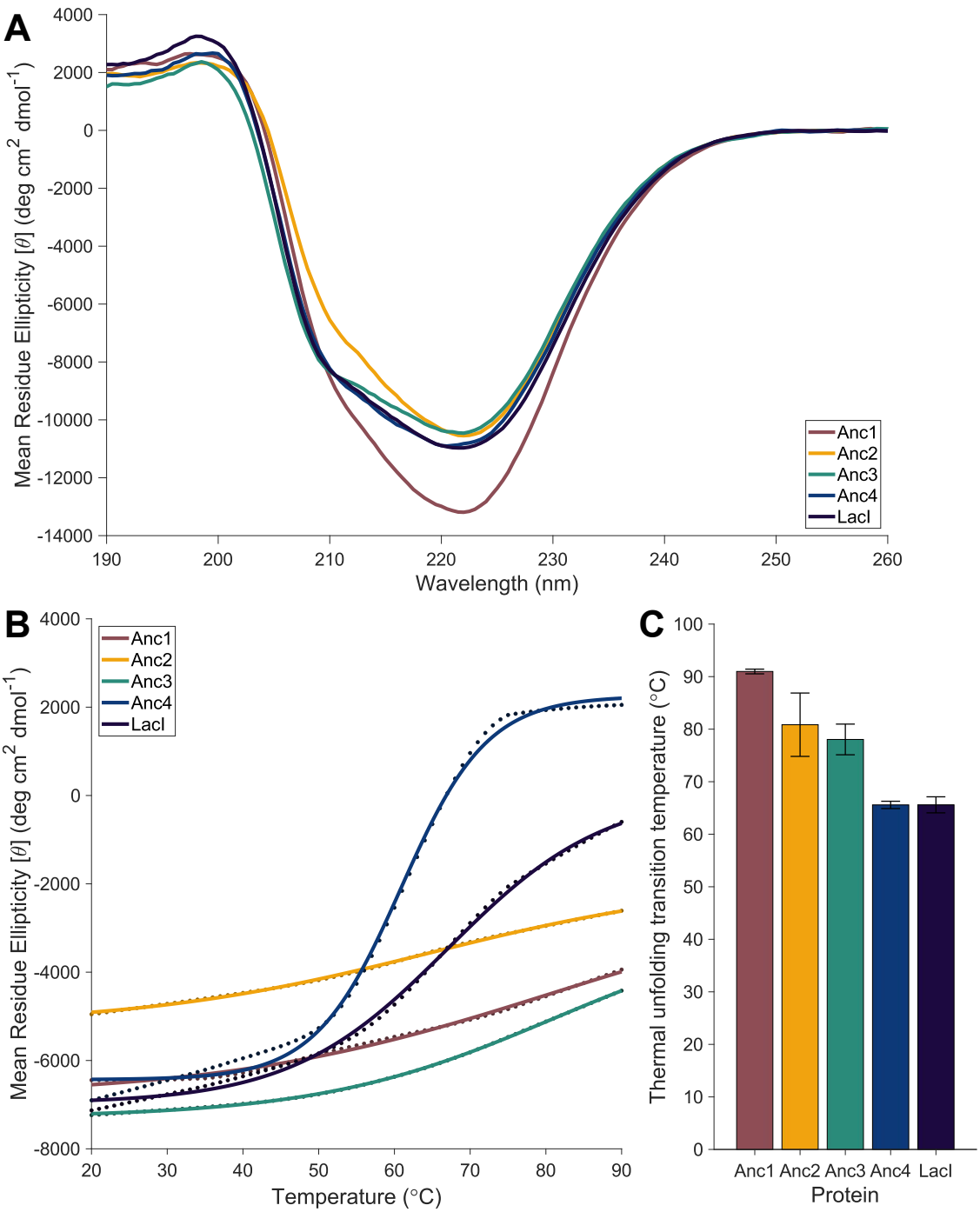



**Figure 2.4:** Characterisation of selected ancestral LacI family proteins and LacI by circular dichroism (CD). A) Spectra of the selected ancestors and LacI measured at 20 °C indicate that they possess relatively similar secondary structure. B) Thermal unfolding of the selected ancestors and LacI shown at 208 nm. C) Thermal unfolding transition temperatures of the selected ancestors and LacI. Note that the values for Anc1, Anc2 and Anc3 are extrapolations as they did not fully unfold by 90 °C. Error bars represent S.E.M. of 3 replicates. Figure made using MATLAB.

For an allosteric response to occur *in vivo*, the ancestral proteins need to form as a homodimer or greater order oligomeric structure in order to have the ability to bind to a DNA

operator sequence. Therefore, preliminary investigation into the oligomeric state of the selected ancestors was conducted *in vitro* using analytical size exclusion chromatography (Fig. 2.5E). Interestingly, it was found that the oligomeric preference of the selected ancestors fluctuated between dimeric and tetrameric over the course of the LacI trajectory. For the transcriptional repression function of LacI family members, forming a tetrameric complex typically enhances repressive ability[451,501]. Therefore, the changing oligomeric state preference of the selected ancestor could reflect changing conditions for how tightly gene expression needed to be regulated. Nonetheless, although characterising the oligomeric state of the selected ancestors *in vitro* was supported by the characterisation of LacI in its *in vivo* state as tetrameric (Fig. 2.5E)[453,483], the different conditions *in vivo* could result in the ancestral proteins favouring different oligomeric states than those measured *in vitro*. For example, the ancestral proteins would likely be present in lower concentrations *in vivo* and experience different thermodynamic effects due to molecular crowding[422,502]. Thus, to establish the oligomeric state *in vivo*, perhaps Förster resonance energy transfer (FRET) tags could be utilised in future to measure if monomeric units were close enough to form oligomers, and the relative level of fluorescence from FRET could be used to determine which oligomeric state had likely formed[503,504].

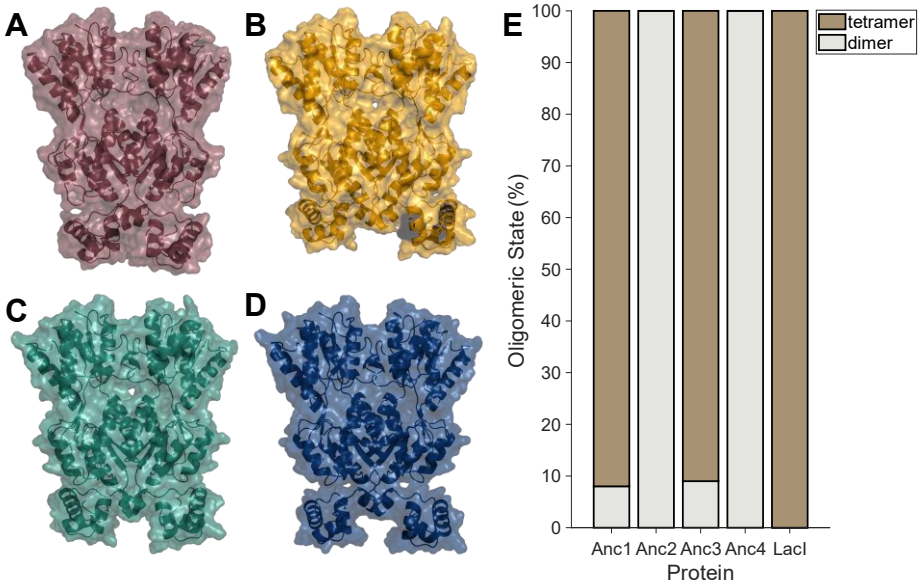



**Figure 2.5:** A-D) Predicted tertiary structures of A) Anc1, B) Anc2, C) Anc3 and D) Anc4 using AlphaFold2. The structure were predicted as dimers, because this is the minimum functional unit of LacI family transcription factors *in vivo*[428,429,445,498]. The predicted structures all are similar to the canonical LacI family structure[428,437,450,454,456,487]. E) The oligomeric state and proportion of the population in that oligomeric state of the selected ancestors and the *E. coli lac* repressor (LacI) was determined through analytical size exclusion chromatography. It was found that Anc1, Anc3 and LacI preferred the tetrameric state, whereas Anc2 and Anc4 formed into the dimeric state. Figure made using PyMOL and MATLAB.

### *2.3.2 Ligand binding characterisation of the selected ancestral transcription factors*

Having established that the selected LacI family ancestors each likely folded into a canonical LacI family structure and could form functional oligomers, the ligand binding of the ancestors was then investigated. As effectors of extant LacI family members mostly consist of small, carbon-based molecules[428,450,505], to identify potential ligands of the ancestral transcription factors a screen of 192 biological, small, carbon-based molecules was used with differential scanning fluorimetry (DSF) to characterise changes to the thermal unfolding transition temperature ($T_m$) . This is because increases in $T_m$ can indicate increased protein stability due to ligand binding in a relatively high-throughput manner[506-509]. However, it should be noted that DSF is prone to false positives and negatives because non-ligands can also lead to increases in $T_m$ through structure stabilisation and some ligand binding events may not necessarily increase the $T_m$[510-512]. Notwithstanding, use of the screen with extant LacI as a control identified three sugars that have been shown by previous studies to act as inducers of LacI[467,471,473] (Fig. 2.6E) (Suppl. Fig. 2.5). Thus, having established the efficacy of the screen, it was used to identify potential ligands for the LacI family ancestors. It was found that methyl β-D-galactopyranoside (MβDG) could be capable of binding to Anc1, Anc3 and Anc4. Additionally, it was found that D-fucose could be a ligand of Anc1 (Fig. 2.6A-D). Anc1 was also stabilised by a few other sugars, suggesting they could also be ligands, though to a lesser degree. On the other hand, no potential ligands were identified for Anc2, which highlighted the limitations of the DSF screen. In future, it would be beneficial to screen a larger library of compounds, including purines and possibly some extant small protein effectors, and possibly to attempt more accurate ligand binding screening techniques like surface plasmon resonance[513-517]. Nevertheless, it was interesting that most of the ancestral proteins were identified as possibly binding to MβDG, as this molecule is structurally similar to the native effector of LacI and can also act as an effector of LacI[467]. On the other hand, the primary potential ligand of Anc1, D-fucose, is structurally somewhat different from its possibly weaker ligands like MβDG (Suppl. Fig. 2.5). This possible capability of Anc1 to bind to multiple ligands could suggest that it is relatively promiscuous or generalist in its effector binding; however, this would need to be better established in future using more accurate binding characterisation like surface plasmon resonance.

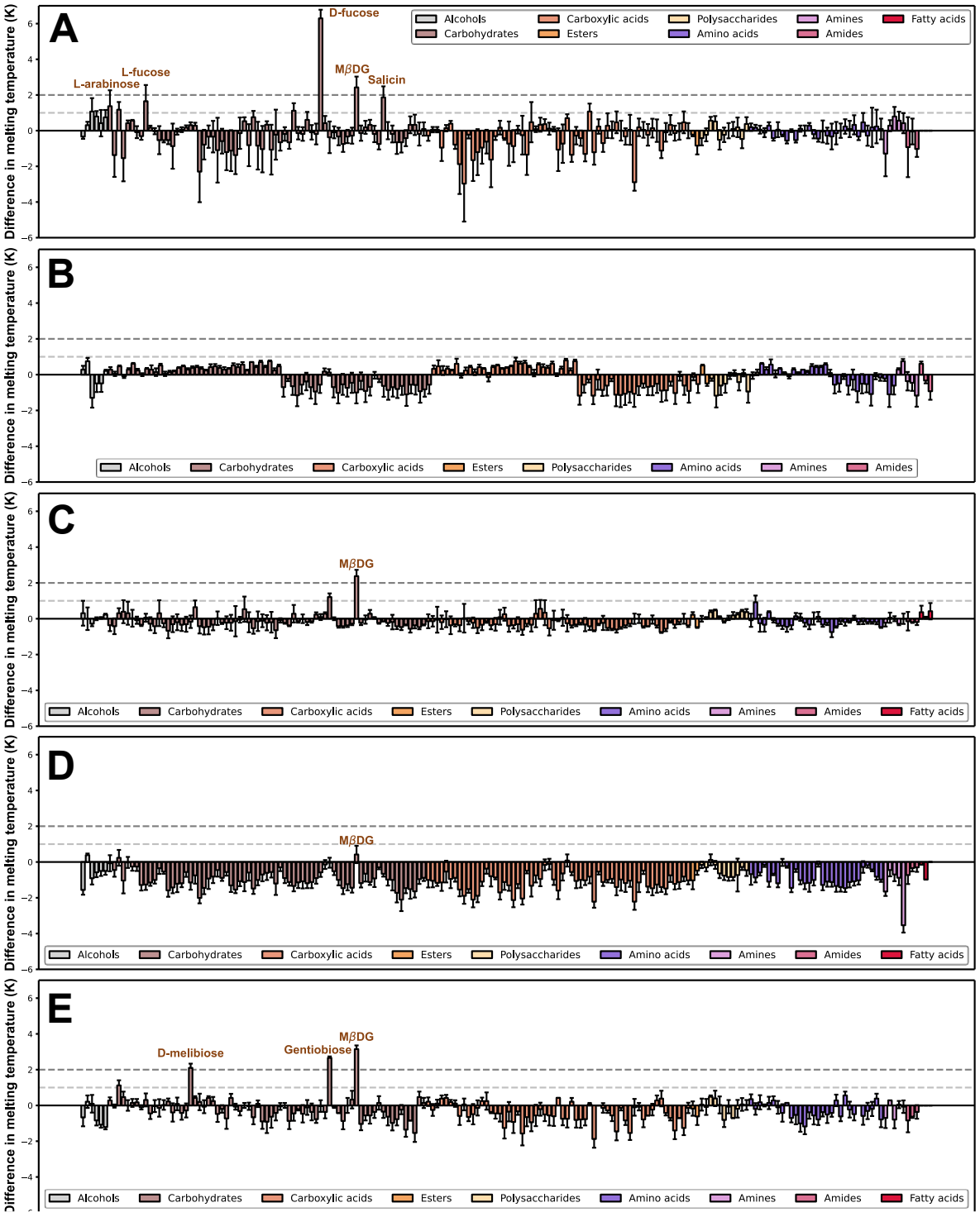



**Figure 2.6:** Binding screen of 192 small, biological, carbon-based molecules by thermal unfolding transition temperature ($T_m$) stabilisation in differential scanning fluorimetry of A) Anc1, B) Anc2, C) Anc3 and D) Anc4 to identify potential ligands. D-fucose was identified as a potential ligand of Anc1, and methyl β-D-galactopyranoside (MβDG) a potential ligand of Ancs1-4 respectively. E) The binding screen was conducted with extant LacI as a control, which identified known effectors. Figure made using Python.

As MβDG was identified as a potential ligand for most ancestors and D-fucose a potential ligand of Anc1, the binding of MβDG and D-fucose to the selected ancestors and LacI was characterised in more detail using isothermal titration calorimetry (ITC). For MβDG binding (Fig. 2.7A), it was found that the oldest ancestor, Anc1, bound MβDG relatively weakly, with an equilibrium dissociation constant ($K_D$) of 4.2 mM. However, progressively the binding of the selected ancestral proteins to MβDG increased in strength with decreasing ancestor age, such that the most recent ancestor, Anc4, was found to have a $K_D$ of 550 μM, which is only slightly higher than that of the extant LacI. Therefore, the increase in binding strength of MβDG over the course of the LacI trajectory suggests a change in functionalisation over time towards binding β-galactosides. As LacI family members typically regulate carbon metabolism in bacteria, perhaps the change in binding could reflect increasing specialisation along the LacI trajectory towards regulating the metabolism of β-galactosides.

On the other hand, whilst binding of D-fucose was measured to be relatively strong for Anc1, with a KD of 127 μM, the binding of D-fucose was too weak to be fully characterised for Anc2 using ITC, was not detected for Anc3, and was very weak for Anc4 and LacI (Fig. 2.7B; Suppl. Fig. 2.7). This shows that there was a relatively rapid decrease in D-fucose binding ability between Anc1 and Anc2, indicating a relatively quick change in binding function between Anc1 and Anc2. Thus, perhaps Anc2 and the other ancestors along the LacI trajectory were no longer involved in the regulation of the metabolism of fucose-like carbon sources. This would have resulted in a lack of positive selection to maintain D-fucose binding ability and hence it was mostly lost, perhaps through negative selection or neutral drift.

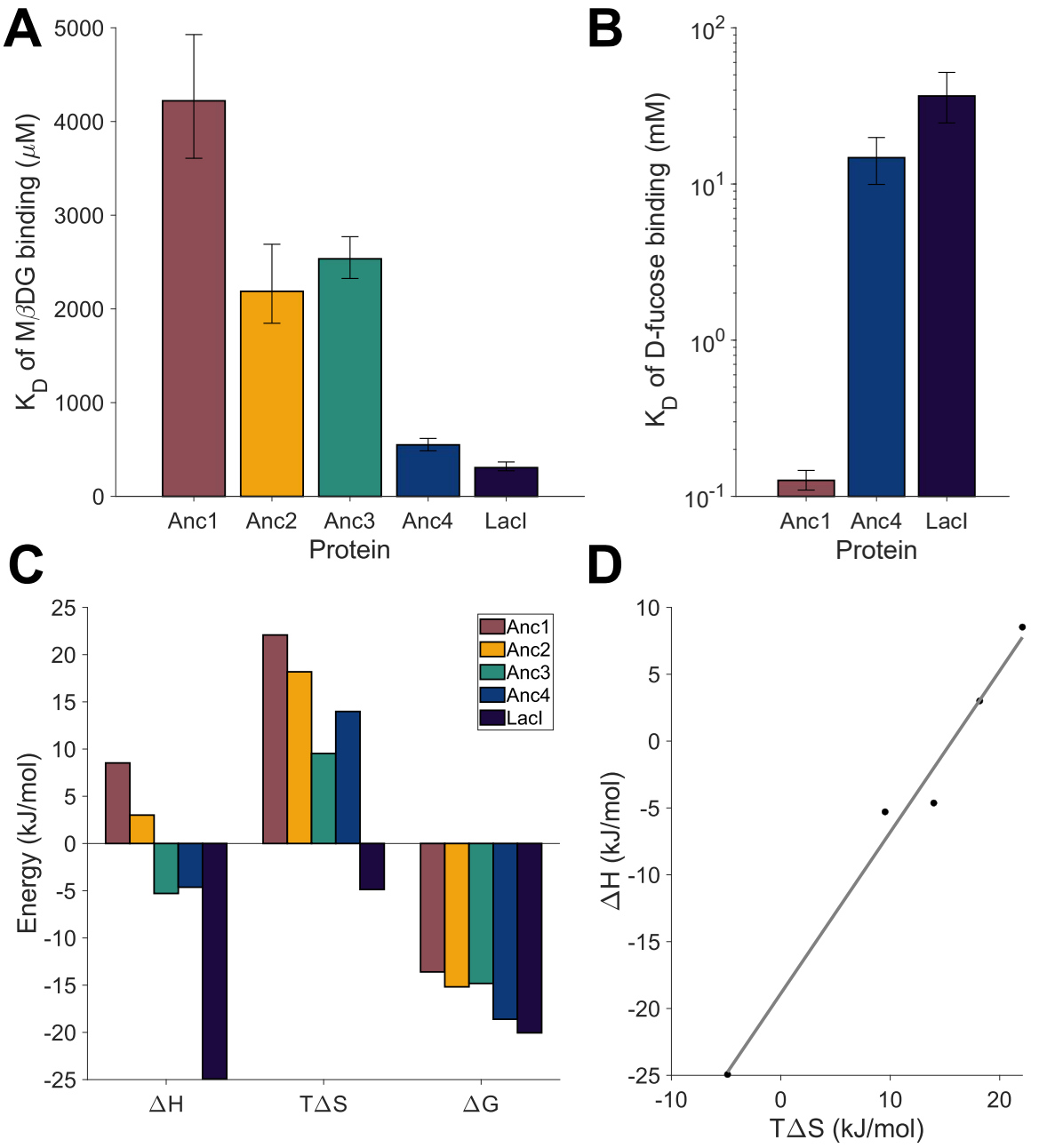



**Figure 2.7:** Characterisation of the equilibrium dissociation constant ($K_D$) of binding of A) methyl β-D-galactopyranoside (MβDG) and B) D-fucose to the selected ancestral transcription factors and the extant *E. coli* lac repressor (LacI). Binding strength for MβDG increases over the LacI trajectory whereas binding strength of D-fucose decreases. $K_D$s were measured using isothermal titration calorimetry (ITC) and error bars represent the 68% confidence interval from fitting at least three replicates of at least two biological replicates. C) The thermodynamics of MβDG binding for the selected ancestors and LacI as measured and calculated from ITC. Anc1 binds MβDG in an entropically driven reaction, but LacI binds MβDG in an enthalpically driven reaction. D) Relationship between change in enthalpy and change in entropy of MβDG binding by the selected ancestors appeared linear (Pearson correlation coefficient of 0.99). Figure made using MATLAB.

To further characterise how the changes in D-fucose and MβDG binding occurred from the oldest ancestor to LacI, the thermodynamics of binding were examined. When characterising

the entropic and enthalpic contributions to binding for both hexoses, it was revealed that the binding reactions of the extant LacI were enthalpically driven with entropic cost (Fig. 2.7C; Suppl. Fig. 2.6). This has been supported by dynamics simulations by Campitelli *et al.*[518], which found that although isopropyl β-D-thiogalactopyranoside (IPTG) binding to LacI resulted in entropic gain in the DNA-binding domain (DBD), the increase in rigidity in the ligand binding domain (LBD) resulted in net entropic loss. Perhaps for this reason, the relatively greater entropic cost of D-fucose binding compared to MβDG binding found here could result from only MβDG and not D-fucose inducing an allosteric response. This is because the increase in flexibility of the DBD, and thus entropic gain from this, would only occur if an allosteric response were induced. More generally, the enthalpic gain and entropic cost of the LacI binding reactions found here could also be explained by the research of Glasgow *et al.*[464]. By using hydrogen deuterium exchange with mass spectrometry, they were able to observe upon IPTG binding increased rigidification of residues near the ligand binding pocket, thereby experimentally determining a potential cause for the observed net entropic cost of binding. Additionally, they showed that upon IPTG binding there was an increase in the number of bonding interactions formed, both with the ligand as well as between subdomains and between the monomer units. This is likely driving the net enthalpic gain observed for LacI binding reactions[422,519].

However, interestingly it was found that the binding reactions of the selected ancestral proteins with MβDG and D-fucose progressively had increasing net entropic gains and decreasing enthalpic contributions (Fig. 2.7C; Suppl. Fig. 2.8A). Therefore, it was found that the thermodynamics of ligand binding have progressively changed from being entropically driven in the oldest ancestor, Anc1, to more and more enthalpically driven in Anc4 and especially in the extant LacI. These differences in the thermodynamics of the binding reactions over time for the selected LacI family ancestors indicate that each ancestor likely undergoes different changes in its dynamics and/or bonding networks upon ligand binding. The net entropic gain for the binding reactions of the ancestors could possibly result from relatively little rigidification of the LBD upon ligand binding while the DBD still increases its flexibility. The net enthalpic cost of the oldest two ancestors could suggest that binding of the ligand results in breaking of some of the bonding networks within the protein. It should be noted that Anc1 still forms bonds with its ligand upon binding[500] and likely Anc2 would as well, however, the net enthalpic cost observed for these ancestors could result from an overall greater number of bonds breaking than forming during the binding reaction. Nonetheless, it

should be noted that the characterisation of the thermodynamics of binding was only conducted at a single temperature of 25 °C, thus, to more broadly establish trends in the thermodynamics of the ancestor binding reactions more temperatures should be investigated.

Notwithstanding, it was found that by comparing the changes in entropy and enthalpy for both MβDG and D-fucose respectively, a linear relationship could be observed (Fig. 2.7D, Suppl. Fig. 2.8B). This highlights there was an almost linear trade-off between the change in entropy and change in enthalpy of the ligand-protein binding interactions observed. This supports wider literature where such a trade-off has also been relatively commonly observed in ligand-protein binding reactions[519-524].

### *2.3.3 Effect of single mutations from Anc2 on the thermostability and ligand binding of Anc1*

As a relatively large change was observed between Anc1 and Anc2 in the binding of MβDG and D-fucose respectively, it was subsequently investigated which mutations between Anc1 and Anc2 may have contributed to such a large shift in binding function to better understand the sequence-function relationship over the evolutionary trajectory. To this end, residues around the binding pocket that had changed between Anc1 and Anc2 were selected (Suppl. Fig. 2.9), and single mutants of Anc1 containing one of these residues from Anc2 were expressed and characterised.

Despite the Anc1 variants only containing a single mutation from Anc2, a relatively high variation in the recombinant expression yields was observed when using the same expression conditions (Fig. 2.8A; Suppl. Fig. 2.10). Interestingly, several of the mutants were found to have lower expression than both Anc1 and Anc2, and one mutant was found to have higher expression than both Anc1 and Anc2. As recombinant expression yield can correlate with relative protein stability, the relative thermostability of the Anc1 mutants was investigated using differential scanning fluorimetry (DSF) (Fig. 2.8B). This showed that the relatively thermostability of each mutant indeed positively correlated with expression yield (Suppl. Fig. 2.12). However, there were three mutants that had higher thermostability than expected given their expression yield, supporting that expression yield is not purely determined by relative thermostability[14,185,190,221]. Notwithstanding, the relative thermostabilities of the mutants compared to Anc1 suggested that most mutations were destabilising (~52%), many were neutral (~43%) and one was stabilising (~5%). This relatively high number of destabilising

and neutral mutations may have contributed to the lower thermostability and expression yield of Anc2 compared to Anc1 (Fig. 2.4C; Suppl. Table 2.1). The pattern of most mutations resulting in a destabilising or neutral effect, with very few in a stabilising effect, also supports the findings of the wider literature, where this trend has been commonly found for many proteins, including LacI[248,472,478,525,526].

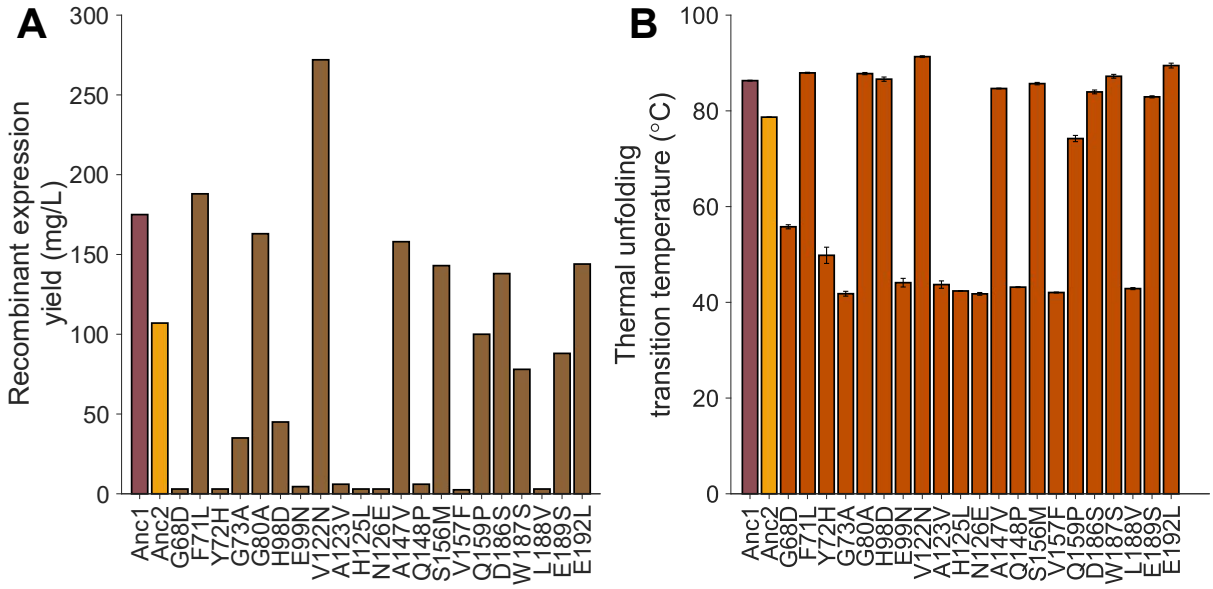



**Figure 2.8:** A) Recombinant *E. coli* expression yields of Anc1, Anc2 and the Anc1 single mutants shows significant variation. B) Thermal unfolding transition temperatures ($T_m$) determined by differential scanning fluorimetry (DSF) of Anc1, Anc2 and the Anc1 single mutants also show significant variation. Note that 300 mM guanidinium hydrochloride was used in the buffer during the DSF experiments such that the $T_m$s could be characterised within the 100 °C limit of the machine. Figure made using MATLAB.

As there was found to be relatively large change in D-fucose and MβDG binding between Anc1 and Anc2, the effect of individual mutations from Anc2 on the ligand binding of Anc1 was then investigated. In order to characterise the binding of all of the mutants, the relatively high throughput method of measuring changes in tryptophan fluorescence with different ligand concentrations was utilised[527,528]. This was possible because Anc1 contains a tryptophan residue in its binding pocket[500]. The use of tryptophan fluorescence was also validated by comparing calculated equilibrium binding dissociation constants ($K_D$s) of Anc1 to those obtained using ITC (Suppl. Fig. 2.11). As the results were not significantly different, this supported the use of tryptophan fluorescence to characterise binding. Thus, using these assays it was shown that single mutations from Anc2 had varied effects on the binding of D-fucose and MβDG to Anc1 (Fig. 2.9). For D-fucose binding (Fig. 2.9A), it was found that most mutants had significantly worse (52%) or similar (29%) binding strength compared to

Anc1, with relatively few mutants displaying slightly improved binding strength (19%). As most of the individual Anc2 mutations lead to lower D-fucose binding strength, this could explain the net effect of very weak D-fucose binding to Anc2. That most mutations resulted in a deleterious or neutral effect to the D-fucose binding of Anc1 also follows the pattern found both for LacI and in wider literature, where it has been found that the vast majority mutations will be deleterious or neutral to the native function of a given protein[248,472,478,525,526].

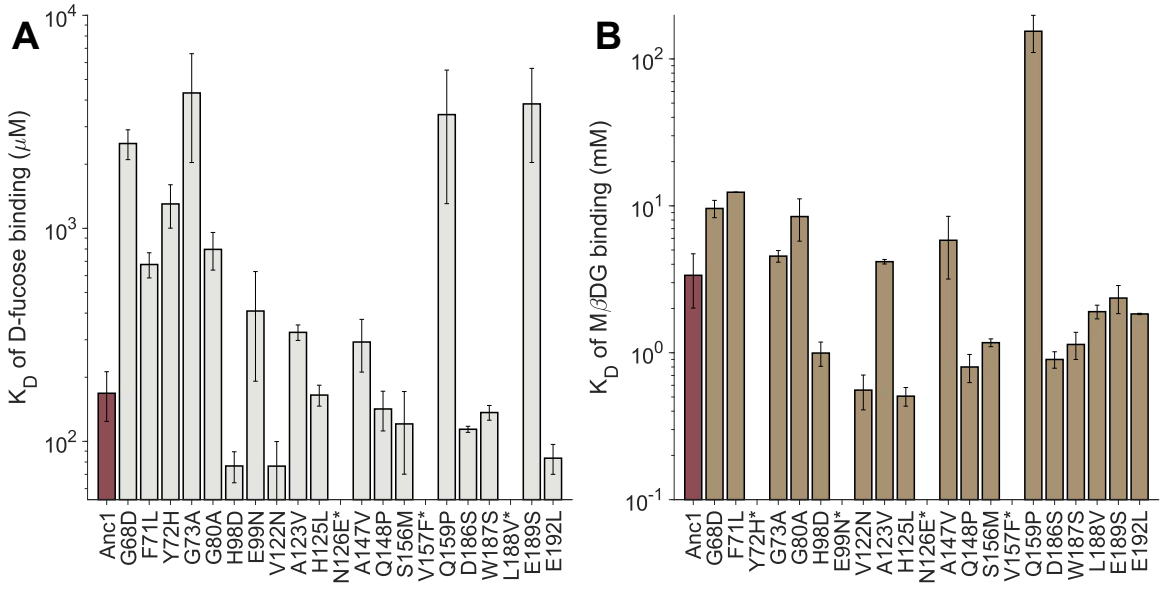



**Figure 2.9:** The equilibrium binding dissociation constant ($K_D$) of binding of A) D-fucose and B) methyl β-D-galactopyranoside (MβDG) to Anc1 and Anc1 mutants containing single mutations from Anc2. If binding was not detected within the limits of the assay, then an asterisk (*) is shown after the name of the mutation. Binding assays were conducted by measuring changes in tryptophan fluorescence at different concentrations of D-fucose and MβDG respectively. Error bars represent S.E.M. of 3 replicates. Figure made using MATLAB.

For the binding of MβDG to the Anc1 mutants (Fig. 2.9B), it was also found that compared to Anc1 many mutants had weaker binding strength (43%). On the other hand, only a few mutants displayed similar binding strength to Anc1 (14%), and many mutants showed greater binding strength (43%). The net effect of all the mutations in Anc2 has been found to result in a $K_D$ of MβDG binding of ~2.2 mM, approximately two-fold lower than that of Anc1 (Fig. 2.7A). However, from the binding strength characterisation of the single mutations of Anc2 in Anc1, how this net effect is achieved is not immediately apparent from the dataset, given that an approximately equal number of mutations led to worse MβDG binding strength as improved strength. Furthermore, there is less difference in the $K_D$s for those mutants with improved MβDG binding, compared to Anc1, than for those mutants with worse MβDG

binding, compared to Anc1. This suggested that adding the effects of all the individual mutations likely would not lead to the MβDG binding strength observed in Anc2. Therefore, either the other mutations not characterised between Anc1 and Anc2 positively contributed to the binding strength of MβDG to 'outweigh' the effect of the deleterious mutations, or the effect of mutations between Anc1 and Anc2 on MβDG binding were non-additive when combined in Anc2. If so, this could indicate intragenic epistatic interactions between the mutations[213,475,529].

Given that the net effect of binding changes between Anc1 and Anc2 resulted in decreased D-fucose binding strength and increased MβDG binding strength, in the simplest case it would be expected that any given mutation from Anc2 would likely result in a deleterious or neutral effect on D-fucose binding and a beneficial or neutral effect on MβDG binding strength. However, such an effect was only observed for 33% of the Anc1 single mutants, with the other mutants either showing weaker binding of both sugars (33%), stronger binding of both sugars (19%), similar D-fucose binding strength but worse MβDG binding strength (10%) or one mutant (5%) even showing better D-fucose binding strength and worse MβDG binding strength. Overall, this highlighted that the net change in binding between Anc1 and Anc2 resulted from relatively complex effects of the mutations between them. Thus, indicating that the sequence-function landscape between and Anc1 and Anc2 is not a gradual, smooth slope but somewhat 'rugged'.

## 2.4 Discussion

Firstly, it was shown that all of the selected ancestral LacI transcription factors (TFs) likely folded into canonical LacI family tertiary structures, formed into potentially functional oligomers of dimers or tetramers, and were able to bind to potential effector ligands. This highlights the power of ancestral sequence reconstruction to identify functional 'peaks' in the protein fitness landscape, especially given that a large-scale study by Tack *et al.*[472] determined that the genotype-phenotype landscape of LacI was unpredictable and that screening during engineering efforts have identified many dysfunctional LacI family transcription factors[405,406,449,473]. These difficulties faced by previous studies attempting to engineer LacI family proteins, and more generally protein engineering, could stem from the deleterious or netural effect of most single mutations to a given protein sequence[248,525,526]. In turn, this could be caused by the relatively high degree of epistasis between many residues in

a given protein[213,475,529-531]. In fact, comprehensive studies of the effect of individual mutations on LacI by Miller *et al*.[478] demonstrated the mostly neutral or deleterious effect of single mutations on proteins[248,525]. Since then, later studies have shown that there are likely several epistatic networks in LacI, which possibly resulted in the effects observed for many of the single mutations[472,480,532,533]. However, to fully validate the utility of ASR for exploring the LacI family genotype-phenotype landscape, in future the ability of the ancestral LacI family TFs to act as transcriptional regulators *in vivo* needs to be assessed.

Nonetheless, it was demonstrated that the selected LacI family ancestors likely folded into the canonical LacI family structure. Interestingly, it was found the thermostability of the older ancestors was significantly greater than that of the extant LacI, with the predicted common ancestor of the phylogeny, Anc1, being especially thermostable. This highlighted the ability of ASR to predict more thermostable yet still functional protein sequences, especially for ancestors with older predicted ages[14,23]. As Anc1 could be around three billion years old[479], its relatively high thermostability could perhaps support the theory that early life was present in relatively hot environments[490-493]. However, the relatively high thermostability of the ancestral TFs could also be an artefact and not reflective of the 'true' ancestral proteins because other studies have suggested that the method of ASR used, maximum likelihood, tends to overestimate protein stability[9,29,68,185]. Nonetheless, although the 'true' ancestral proteins may not have been as thermostable, the relatively high thermostability of the older LacI family ancestors could still be useful in future protein engineering applications. This is because typically it is easier to start with a thermostable protein scaffold for engineering new functions though directed evolution or even rational design approaches because most mutations are destabilising[28,227,229,248,534].

Although the 'true' LacI family ancestors may not have been as thermostable as the predicted ones, it has nevertheless been found that ASR typically predicts sequences close in function to the 'true' ancestor if a suitable evolutionary model is used[28,151,183,184,535]. Furthermore, regardless of its accuracy, it still provides a powerful way to explore functional 'peaks' in the genotype-phenotype landscape and thus better characterise and understand it[34,186,476,477,479]. These abilities of ASR were supported by the finding that all the predicted ancestral LacI family TFs were shown to form into dimers or tetramers, because extant LacI family members also form dimers or tetramers in order to bind to their respective DNA operators[443,451,499,536]. Interestingly no particular trend was found between the preference for the dimeric or tetrameric state with ancestral age. For the extant *E. coli* proteins LacI and *gal*

repressor (GalR), it has been found that formation of a tetramer likely leads to tighter gene repression than the equivalent dimer[444,536,537]. This is beneficial to the *E. coli* cell because there is relatively less 'leaky' expression of the operons and genes controlled by these respective TFs, which reduces expenditure of energy and resources when not necessary[427,538-540]. If the tetrameric form of the ancestral LacI family TFs also resulted in tighter gene regulation compared to the dimeric form, then the changing oligomeric state preference with ancestral age could indicate fluctuation between tighter and weaker regulation of the genes they control. Perhaps this could reflect fluctuating needs for metabolic regulation over time depending on the availability of carbon sources in the environment as well as changes in the rest of the metabolic network. If the oligomeric state was indicative of the relative number of promoter sequences for a given operon, then the changing oligomeric state could highlight the interplay between evolution of *cis*-acting and *trans*-acting elements for regulation[541-543]. Perhaps it is due to this relationship that the trend of oligomeric states found along the LacI trajectory is relatively different to other studies investigating oligomeric state evolution, where more gradual changes over time were observed[193,203,544,545]. However, in future the relative strength of repression or activation of each of the selected ancestral TFs needs to be characterised to validate whether the oligomeric state correlates with strength of regulation. Furthermore, the oligomeric state of the ancestral proteins was only characterised *in vitro*, which may not reflect the *in vivo* oligomeric state due to effects including lower concentrations of the given TF *in vivo* and molecular crowding. Thus the oligomeric state should be characterised *in vivo* in future, such as by using a Förster resonance energy transfer (FRET) tag system[503,504,546].

Nonetheless, having established that the selected ancestral proteins were forming likely functional oligomeric units *in vitro*, it was then shown that the selected ancestors were capable of binding certain sugar ligands, namely methyl β-D-galactopyranoside (MβDG) and additionally D-fucose for Anc1. As MβDG is structurally and chemically similar to the native effectors of LacI[467,547], this suggests that perhaps the ligand binding activity of most of the ancestors along the LacI trajectory has not changed dramatically over time, but rather gradually strengthened towards binding of β-galactosides. However, the oldest ancestor, Anc1, was found to bind more strongly to D-fucose, a deoxy galactopyranose. This could suggest that the function of Anc1 was relatively different to the later ancestors. Perhaps Anc1 was involved in the regulation of carbohydrate metabolism of monosaccharides more similar to D-fucose. Indeed, it has been found that the formation of hexoses from formaldehyde is

possible prebiotically[548-553], thereby raising the possibility that D-fucose or similar sugars were available in the environment at the predicted age of Anc1 of around three billion years[479]. However, D-fucose has not been found to be metabolizable by modern day *E. coli* and possibly other bacteria[554], hence raising the question as to whether the ancestral bacterium of Anc1 metabolized D-fucose or not and consequently whether D-fucose was the native effector of Anc1 or not. In any case, the binding of D-fucose to Anc1 was found to be relatively strong through isothermal titration calorimetry (ITC) and D-fucose binding is exhibited even by modern LacI family members. For example, it has been found that D-fucose acts as a gratuitous inducer of extant GalR and GalS[554-556], as well as an anti-inducer of extant AraC[557,558]. This ability of D-fucose to bind to extant LacI family members supports the findings that the 'true' ancestral protein may also have been capable of D-fucose binding. Furthermore, the ability of the predicted Anc1 to bind to D-fucose, and thus potentially act as a transcription regulator in response to D-fucose, could make Anc1 a useful tool for synthetic biology. This is because D-fucose is not metabolized *in vivo*, which makes it relatively biorthogonal[418,554,559]. Indeed, for this reason, development of a D-fucose binding LacI family TF was previously attempted in two separate studies, one by Taylor *et al.*[473] and the other by Groseclose *et al.*[418]. However, the process to identify a D-fucose TF was highly laborious in both cases, requiring screening of large libraries. In contrast, Anc1 was identified by characterising four ancestral proteins. Nonetheless, it would be interesting in future to more fully characterise the selected ancestral LacI TFs by testing a larger library of potential ligands and using a more accurate but still high throughput binding screening method, like surface plasmon resonance[513-517].

The differential scanning fluorimetry binding screen indicated that Anc1 was potentially capable of binding to a few different pyranoses. This could suggest that Anc1 was relatively generalist in nature, or relatively promiscuous beyond its native function. As Anc1 was predicted to be the last common ancestor of the relatively diverse LacI family, the relatively generalist or promiscuous nature of Anc1 could thus support a model of evolution whereby the main function observed in certain extant proteins has evolved through positive selection of an existing activity present in Anc1, which has been described as 'creeping'[199]. If this change occurred after a gene duplication, then it could also be an example of subfunctionalization[186,202,217,247,560]. This may provide support to the studies suggesting that one of the common mechanisms of transcription evolution is through subfunctionalization and optimisation of suboptimal ancestral function[561-567]. However, given the wide diversity of

ligand binding functionality in extant LacI family members[428,450] and the likely much more limited functionality of Anc1, it is unlikely that this evolutionary method occurred along all branches of the extensive LacI phylogeny. For some quite distinct extant ligand binding functions, such as binding of the small protein HPr to CcpA[455]. It is possible that these functions may have arisen through neutral drift and/or neofunctionalization instead following gene duplication [9,208,209,526,568,569].

Nevertheless, the progressive increase in strength of MβDG binding from Anc1 to extant LacI indicated that the binding of β-galactosides observed in modern day LacI evolved through relatively gradual optimisation of activity present in the ancestor. Therefore, along the LacI trajectory it is likely that there was evolving specialisation towards the function of binding β-galactosides like MβDG. Georgelin[500] showed that this increasing preference for binding of β-galactosides between Anc1 and Anc4 may have resulted from an opening of the binding pocket, which allowed the β-substitution group to 'fit' more easily, as well as possibly an increased bonding network between MβDG, binding pocket residues and water molecules in Anc4 compared to Anc1. In future, it would be interesting to investigate if a native ligand of LacI, allolactose, also follows the same binding strength trend as MβDG. If so, as LacI regulates metabolism of lactose and other β-galactosides[539,554,570,571], this could potentially suggest increasing specialisation along the LacI trajectory towards acting as regulators of β-galactoside metabolism. Compared to D-fucose, β-galactosides like lactose are relatively more complex. Hence, their synthesis and availability as a carbon source may have arisen later. Certainly lactose synthesis is thought to have only arisen in mammals, a relatively recent branch on the tree of life[572-575]. Therefore, perhaps the increasing specialisation towards β-galactosides along the LacI trajectory is reflective of an increasing number and complexity of sugars available in the environment over time, which in turn drove increasing need to regulate carbohydrate metabolism in prokaryotes in more complex ways with a greater diversity of transcriptional regulators.

In contrast to this relative progressive and gradual trend of increasing MβDG binding strength, it was found that D-fucose binding was mostly lost along the LacI trajectory after the last common ancestor of the phylogeny. This rather rapid loss of old function but gradual gain of new function is somewhat distinct from some previously characterised protein function changes. For example, several studies have observed relatively gradual loss of old function during gain of new function[247,474,576] and others have found that old function was maintained with the new function[577,578]. Therefore, the relatively rapid loss of old function

found here emphasises that there are many evolutionary 'paths' and that the protein fitness landscape can vary greatly between different evolutionary trajectories. Perhaps the relatively sharp loss of D-fucose binding function from Anc1 to Anc2 found here resulted from a combination of mutations which increased MβDG binding ability but completely restricted D-fucose binding at the same time. It would be interesting to investigate this further in future by characterising the structure and dynamics of Anc1 and Anc2 and establishing mechanistically how the changes in ligand binding have occurred. For instance, this could be conducted using molecular dynamics simulations and hydrogen deuterium exchange mass spectrometry, amongst other techniques.

Along with the changes in ligand preference observed from Anc1 to LacI, it was found that the thermodynamics of the ligand binding reactions progressively changed, with Anc1 binding ligands in a strongly entropically driven manner which gradually shifted over the course of the LacI trajectory to LacI binding ligands in enthalpically driven way. For LacI, its net entropic cost ligand binding has been found to result primarily from rigidification of parts of the ligand binding domain[464,518], whilst its net enthalpic gain during ligand binding may result from the formation of the bonding networks that have been found to occur upon ligand binding[452,464]. However, the different net changes observed in the ligand binding reactions of the selected LacI family ancestors suggested that different changes occurred upon ligand binding compared to LacI, especially for the older ancestors. The entropically favourable ligand binding reactions of the ancestors could result from less rigidification of the ligand binding domain compared to LacI, while still increasing the flexibility of the DNA binding domain and other regions. For the increasingly less favourable changes in enthalpy observed between LacI and Anc1, while all of the ancestors are likely forming bonds with the ligand in an enthalpically favourable way[452,500], perhaps for the older ancestors more bonding networks are disrupted during ligand binding than bonds formed. In future to validate these theories, the differences in the bonding networks and dynamics upon ligand binding of the ancestral transcription factors should be characterised experimentally; for example, by using a combination of hydrogen deuterium exchange mass spectrometry (HDX MS), molecular dynamics simulations and structural determination of the ancestors with and without ligand bound. Furthermore, if the selected ancestors can act as repressors *in vivo*, then ligand binding to the ancestor bound to DNA should also be investigated to fully characterise the ligand binding reaction as it might occur *in vivo*.

Interestingly, the net differences in the thermodynamics of ligand binding found over the course of the LacI trajectory correlated with the relative thermostability and age of the protein. When considering the thermodynamics of ligand binding, hotter temperatures would result in entropically driven reactions being much more favourable than entropically costly reactions because entropy is directly modified by temperature[516,579]. Therefore, the increasingly entropically driven ligand binding shown from Anc4 and Anc1, as well as their relatively thermostabilities, could reflect increasing environmental and intracellular temperatures from Anc4 to Anc1. However, this hypothesis is difficult to corroborate because it is difficult to determine the climate up to three billion years ago[490,580,581] and it is unknown what the microclimate of the respective host bacterium of each ancestor may have been. Also, as the changes in entropy and enthalpy of a given reaction are not fixed over different temperatures[582], the ligand binding capability of the older ancestors compared to the more recent ancestors and LacI should be determined over a range of higher temperatures in future.

As well as correlating with relative ancestral age, the net thermodynamic changes during ligand binding of the selected ancestors and LacI determined here also reveal an almost linear relationship between the change in enthalpy and change in entropy terms of binding. This adds support to the theory that there may be a universal negative 'trade-off' or compensation between the enthalpic and entropic components of protein binding interactions[519,522-524,534,583]. While the exact level of this compensation has been found to vary between different protein-ligand systems[584-587], overall, for the protein binding systems characterised it was found that any gains in enthalpy come with an entropic cost, and *vice versa*. Therefore, this study highlights the importance of considering net entropic and enthalpic gains for engineering new protein-substrate binding reactions. For example, if it were desired that a LacI family-like transcription factor bind a new effector, then it would commonly be assumed that new bonding interactions between the effector and protein need to be created in order to bind that effector. However, if these new bonding interactions result in an especially large enthalpic gain during the binding reaction, then it needs to be considered that there will be a corresponding large entropic cost, such as by rigidifying more of the protein with the increased number of bonding interactions. This entropic cost could negatively impact the function of the engineered transcription factor. Hence, this study emphasises the importance of the entropy-enthalpy 'trade-off' in protein binding interactions for future protein engineering efforts, especially given that currently many engineering studies do not consider

how changes in thermodynamics may have affected the function and activity of their engineered protein[8,29,214,431,583].

Another important consideration for protein engineering, as well as for gaining an increased understanding of proteins, is the relationship between sequence and function[8,186,588]. To this end, the change in ligand binding function between Anc1 and Anc2 was investigated through characterising the effect of single mutations from Anc2 on Anc1. Despite both Anc1 and Anc2 being relatively thermostable proteins capable of binding to their respective ligands, the single mutant Anc1 variants showed large diversity in both thermostability and binding function. No mutation was found to result in a relatively neutral impact on binding of both D-fucose and MβDG, indicating that all the selected positions impacted ligand binding in some manner. The importance of these positions to ligand binding is supported through the study by Suckow *et al*.[478], which found that the majority of the equivalent positions in LacI when mutated led to loss of ligand binding and/or loss of allostery. It would be interesting to characterise whether these positions are more broadly important in ligand binding in the greater LacI family, or if they are specific to the LacI trajectory. Certainly, five of the characterised mutations between Anc1 and Anc2 could be specific to the LacI trajectory because the equivalent residues in LacI have been found to participate in bonding with IPTG[452]. However, many of the other mutations are not present in the binding pocket, thus, their effect on ligand binding is likely through changing bonding network interactions or through changes to the dynamics of the protein. Such long-range effects on ligand binding have also been characterised in detail in LacI[464,465,481,518,533] and found in other LacI family proteins too[428,457,466,480,589]. Thus, the relatively large effects on ligand binding found between Anc1 and Anc2 by residues not in the binding pocket highlighted the complex process of ligand binding by LacI family transcription factors.

This complex, relatively unpredictable, and possibly unproportional, effect of the individual mutations from Anc2 on the binding ability of Anc1 indicated that these positions form interactions with other residues. As networks of interacting residues have also been found in LacI[464,480,481,533] and possibly other LacI family TFs[419,449,590], this suggests that such networks are important to the function of LacI family TFs. These residue networks have likely led to the possibly non-additive effects on binding of the mutations from Anc1 to Anc2[591,592]. This supports previous findings where the effects on function of mutating modern day LacI family proteins have also been epistatic[405,472,478-480,589]. This relatively high degree of epistasis between residues in LacI family proteins has likely led to the difficulty in engineering new

LacI family transcription factors, because the effect of mutations is very difficult to predict, even using coevolutionary models or deep learning[472,518,592]. Therefore, this study further highlighted that LacI family proteins have complex residue interaction networks and indicated the power of ASR to explore the 'peaks' of such a complex genotype-phenotype in a rational way. In future, to gain further insights into the fitness landscape along the LacI trajectory and its degree of 'ruggedness', or epistasis, perhaps more variants between Anc1 and Anc2 could be characterised with different numbers of combinations of the mutations between the two. This could then be combined with machine learning to predict the full fitness landscape, which could then possibly be used to engineer novel transcription factors by predicting functional 'peaks' in the landscape.

In conclusion, it was found that the binding of β-substituted D-galactosides to extant LacI likely evolved gradually over time from the most recent common ancestor of the phylogeny (MRCA) and that this binding change was accompanied by a progressive shift from entropically driven binding in the MRCA to increasingly enthalpically driven binding in LacI. It was also shown that the effects of the mutations leading to these binding changes are complex and likely epistatic in nature. This characterisation of the evolution of binding along the LacI trajectory of the LacI family highlighted the utility of ASR as a relatively low-throughput method to explore the protein genotype-phenotype landscape, which in turn informs future transcription factor engineering efforts. In future, to gain deeper insights into the LacI family and protein evolution it would be informative to also investigate how the *in vivo* function and allosteric communication of the LacI family has changed over time. While this chapter demonstrated how ASR could be applied to understanding protein functional changes, the next chapter shows how ASR can be applied to engineer proteins with improved recombinant expression.

## 2.5 Methods

### *2.5.1 Expression and purification of selected LacI family ancestral sequences and LacI*

The genes for the selected LacI family ancestors[593] were codon optimised using the GeneArt tool (Thermofisher) before being synthesised by Twist Bioscience in the pET28a(+) vector. The gene for *E. coli* LacI in the pETMCSIII vector was kindly provided by Dr Matthew Spence (Australian National University). Genes were transformed into electrocompetent *E. coli* BL21(DE3). A single transformed colony was selected for a 10 mL overnight LB culture

(10 g/L tryptone, 10 g/L sodium chloride, 5 g/L yeast extract), incubated at 37 °C, 200 rpm. This was then used to inoculate 1 L of LB, which was incubated at 37 °C, 200 rpm until the $O.D._{600}$ reached ~0.8. At this point, a working concentration of 1 mM isopropyl β-D-thiogalactopyranoside (IPTG) was used to induce recombinant expression. Based on expression trials (data not shown), for LacI and Anc3 expression, cultures were incubated at room temperature for 48 hrs, for Anc1 and Anc2 cultures were incubated at 30 °C for 24 hrs, and for Anc4 expression cultures were incubated at 18 °C for 48 hrs. Cultures were then centrifuged (4 °C, 4000 rpm, 15 min), the supernatant discarded and the pellet stored at -20 °C until use.

For purification, cell pellets were defrosted on ice before being resuspended in Buffer A (20 mM Tris, 500 mM NaCl, 20 mM imidazole, pH 8.0) and lysed by sonication. Following centrifugation (4 °C, 12000 rpm, 60 min), the filtered lysate was then loaded onto a 5 mL HisTrap HP column (Cytiva). Bound protein was eluted using Buffer B (20 mM Tris, 500 mM NaCl, 250 mM imidazole, pH 8.0). Protein fractions were pooled, concentrated and filtered before being loaded onto a HiLoad 26/600 Superdex 200 pg (Cytiva) column and eluted using Buffer C (0.2 M NaCl, 40 mM sodium phosphate, pH 7.4 + 3 mM 2-mercaptoethanol for Anc3, 4 and LacI + 2% v/v glycerol for Anc4 and LacI). Protein content of SEC fractions was determined by SDS-PAGE and fractions with relatively high purity of the expected molecular weight were pooled and concentrated. Protein concentration was determined by UV/Vis using a Nanodrop (Thermofisher) at 280/260 nm with extinction coefficients predicted by Expasy Protparam[594]. Protein sample was then stored at 4 °C for <2 days or snap frozen in liquid nitrogen and stored at -80 °C.

### *2.5.2 Circular dichroism of LacI family ancestors and LacI and 3D-structure prediction*

Purified samples of the selected ancestors and LacI were buffer exchanged into distilled water. Using a quartz cuvette (0.1 cm pathlength), the blank spectrum was first recorded followed by the protein spectrum at 20 °C from 180-280 nm with a stepsize of 0.5 nm, and using 0.5 s per point. The thermal response was then measured from 20 to 90 °C at 208 and 222 nm with a stepsize of 1 °C and speed of 2 °C per minute. A sigmoidal curve was fitted to the data for each wavelength respectively using the Curve Fitter app in Matlab and the transition point for each wavelength calculated from the curve fit. The thermal transition unfolding temperature was calculated by taking the mean of the transition points from each wavelength.

The dimeric 3D-structures of the selected ancestral sequences were predicted using AlphaFold2[494] *via* ColabFold[495].

### *2.5.4 Oligomeric state determination by analytical size exclusion chromatography*

Firstly, a calibration curve of a 10/300 GL superdex200 column (GE Healthcare) was created using a LMW and HM2 calibration kit (Cytiva) at 20 °C with a running speed of 0.25 mL/min in Buffer D (150 mM NaCl, 40 mM sodium phosphate, pH 7.4). Samples of the selected ancestors and LacI were prepared at 3 mg/mL in Buffer D, before being loaded onto the column, which was run at 0.25 mL/min at 20 °C. The approximate mass of each eluted protein was determined by comparison of the maximum of each peak's elution volume to a standard curve. Proportion of protein present in each oligomeric state (if relevant) was determined by integrating each peak and comparing the relative peak areas.

### *2.5.5 Small molecule binding screen using differential scanning calorimetry*

Carbon-based small molecule screening plates PM1 and PM2A were purchased from Biolog. Each compound was dissolved in distilled water to make a 10x stock of between 10-20 mM. Purified protein samples were prepared consisting of 0.5 mg/mL protein with 1 x ThermalShift dye (Thermofisher). Due to the relatively high thermostability of Anc1, 300 mM guanidinium hydrochloride was also added to the Anc1 sample. The protein sample was then added to the small molecule stock and the changes in fluorescence measured using a qPCR machine from 20 °C to 100 °C. These melting curves were then analysed using the ThermalShift software, and relative melting temperatures of each sample were determined. Four to six replicates of each condition were conducted.

### *2.5.6 Isothermal titration calorimetry*

D-fucose or MβDG stocks were prepared in Buffer C using a volumetric flask for concentration determination, before being diluted to the desired concentration. Protein samples and ligand stocks were then degassed for at least 10 min. Experiments were conducted using a TA instruments NanoITC machine at 25 °C using 1 x 1 µL dummy injection followed by 22 x 2 µL injections of ligand stock. Stirring speed used was between 250-350 rotations/s, and initial and final baselines were collected for 120 s. At least two biological replicates and three total replicates were collected for all ligand-protein interactions characterised. Baseline was removed using NITPIC[595,596] and analysed using SEDPHAT[597] using a heteroassociation model. As most experiments fell within the 'low c'

range, the stoichiometry was fixed at 1 to improve the accuracy of the analysis[598]. Thermograms and titration curves were plotted using Gussi[599].

### *2.5.7 Expression, purification and characterisation of single mutants of Anc1*

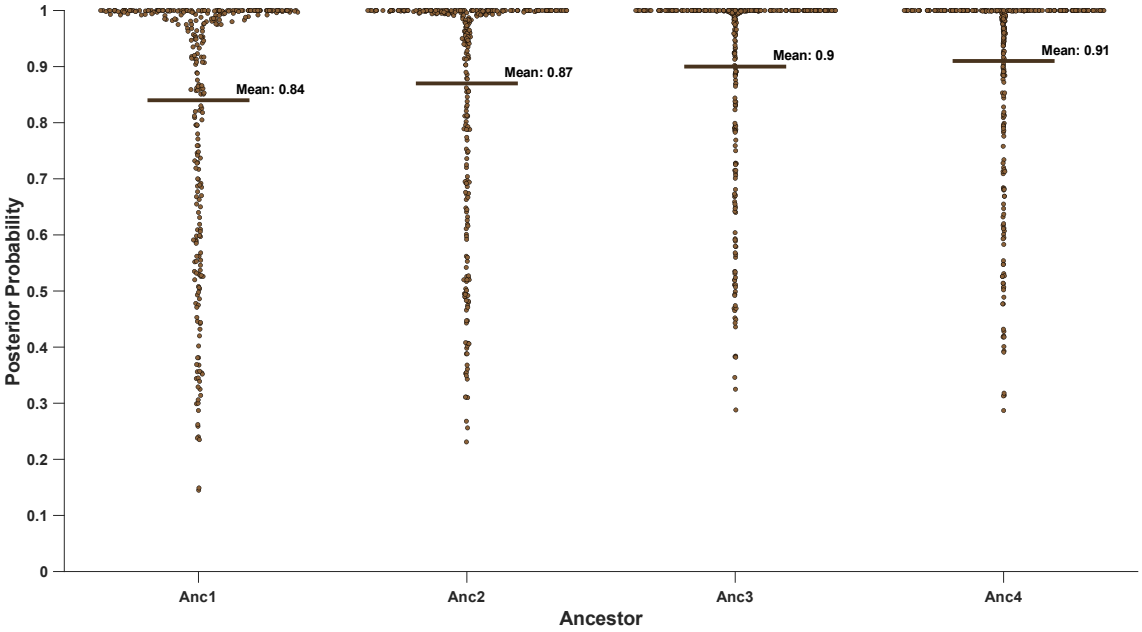
The genes for the Anc1 mutants containing selected mutations from Anc2 were synthesised by Twist Bioscience in the pET28a(+) vector and expressed in *E. coli* BL21(DE3) as described previously for Anc1. The Anc1 mutants were IMAC purified as described previously for Anc1 before being buffer exchanged into Buffer C using a HiPrep 26/10 Desalting column (Cytiva). Thermostability of the mutants, Anc1 and Anc2 was characterised using differential scanning fluorimetry with samples of 0.5 mg/mL protein with 1 x ThermalShift dye (Thermofisher) and 300 mM guanidinium hydrochloride. Changes in fluorescence with temperature were measured as described previously and data analysed as described previously. At least 6 replicates for each protein were conducted.

Binding of the mutants to D-fucose and MβDG respectively was determined used a tryptophan fluorescence assay (tryptophan is present in the binding pocket of Anc1). Stocks of D-fucose and MβDG were prepared in Buffer C. Protein was then added to the stocks at a working concentration of 10 µM before the assay plate was centrifuged briefly and incubated for 10 min at room temperature. Tryptophan fluorescence was then measured using a TECAN infinite M200 plate reader with excitation of 260 nm and emission of 288-400 nm. Optimum gain and z height were calculated from one of the sample wells. At least three replicates for each condition were conducted. Binding curves were fitted using a one site binding model, where $B_{max}$ is maximum specific binding and $K_D$ is the equilibrium dissociation constant:

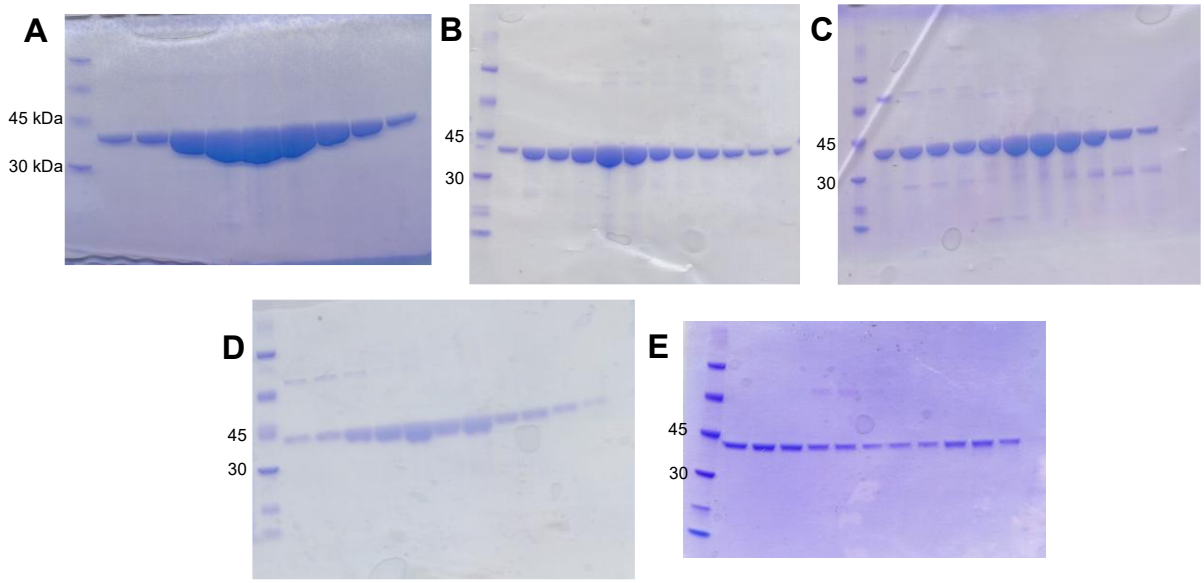
$$y = \frac{Bmax \times x}{K_D + x} + background.$$

## 2.6 Supplementary

**Suppl. figure 2.1:** Distribution of the posterior probabilities of the predicted residues for each of the selected ancestral sequences. Whilst most residues of the selected ancestral sequences had a relatively high posterior probability, some were relatively low.



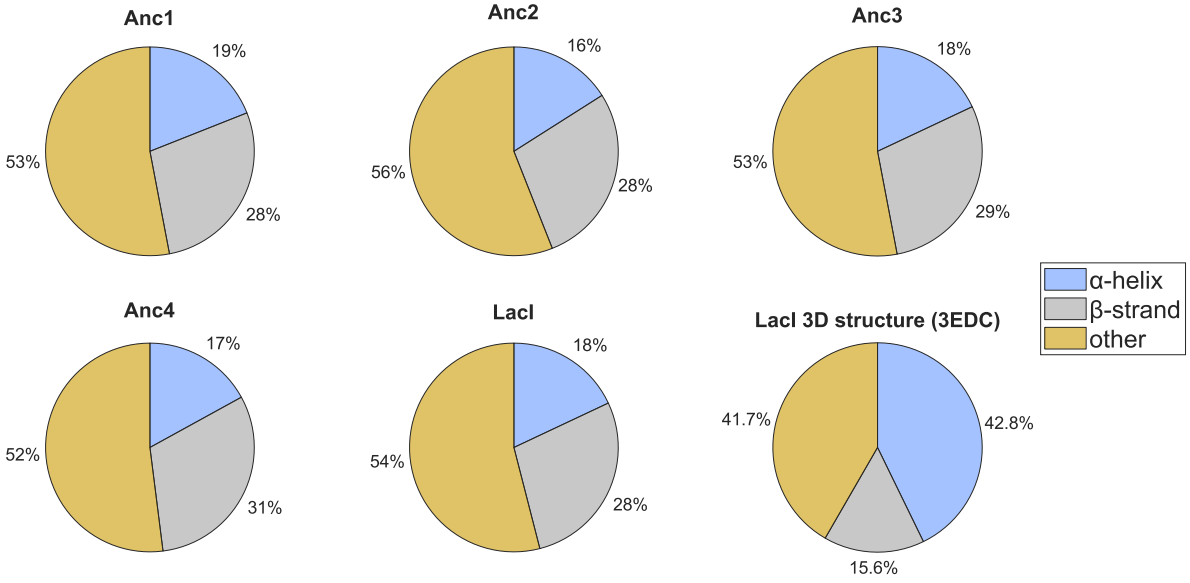

**Suppl. figure 2.2:** SDS-PAGE of size exclusion chromatography fractions of A) Anc1, B) Anc2, C) Anc3, D) Anc4 and E) LacI. First lane in each gel contains the ladder.



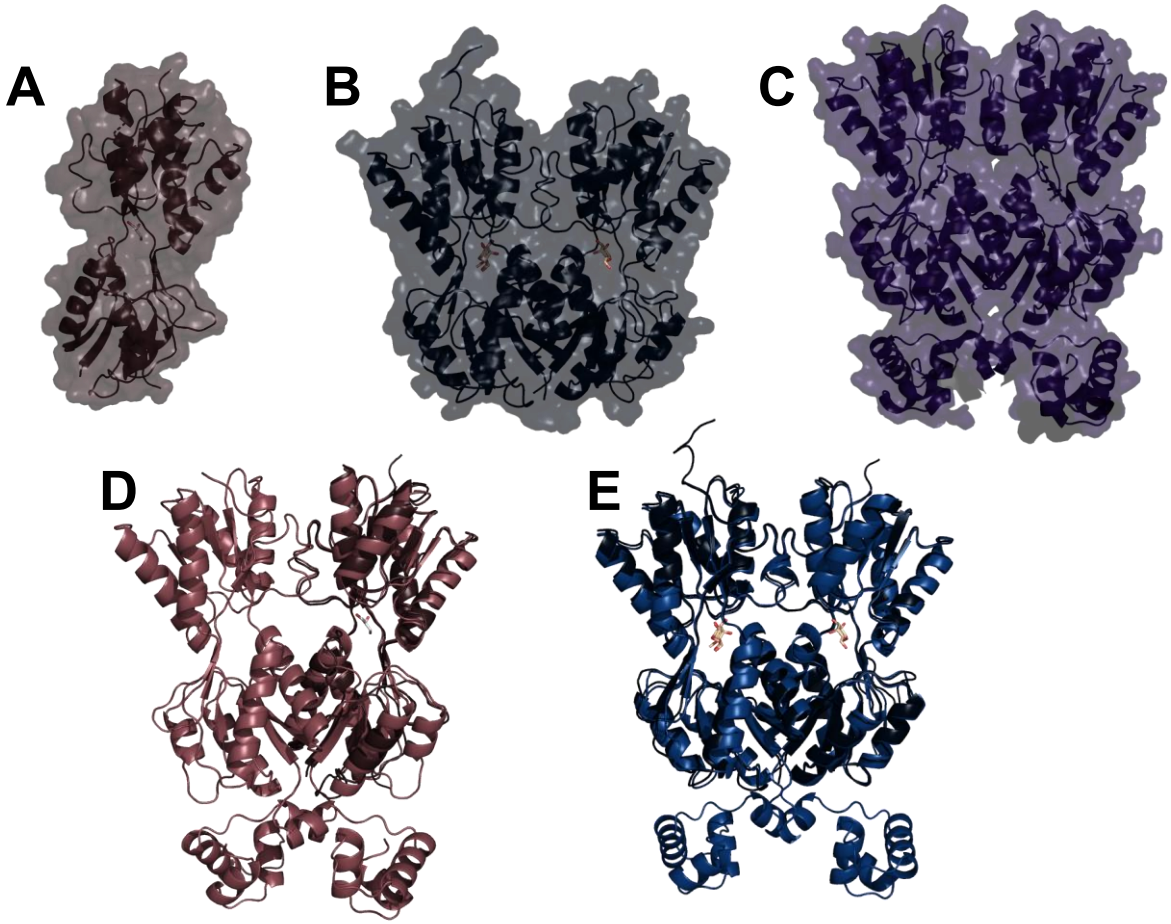
**Suppl. table 2.1:** Yields of selected LacI family ancestors and LacI when overexpressed in *E. coli.*

| Protein | Yield (mg/L culture) |
|---|---|
| Anc1 | 135±23 mg/L culture |
| Anc2 | 71±15 mg/L culture |
| Anc3 | 64±11 mg/L culture |
| Anc4 | 57±13 mg/L culture |
| LacI | 46±15 mg/L |

**Suppl. figure 2.3:** Prediction of secondary structure elements for the selected ancestors and LacI from their respective circular dichroism (CD) spectra using the CDSSTR method. The relative proportion of secondary structure elements of the X-ray structure of LacI[537] (PDB: 3EDC) is also shown for comparison. The predictions from the CD spectra seem to have underpredicted α-helical content and overpredicted β-strand content.



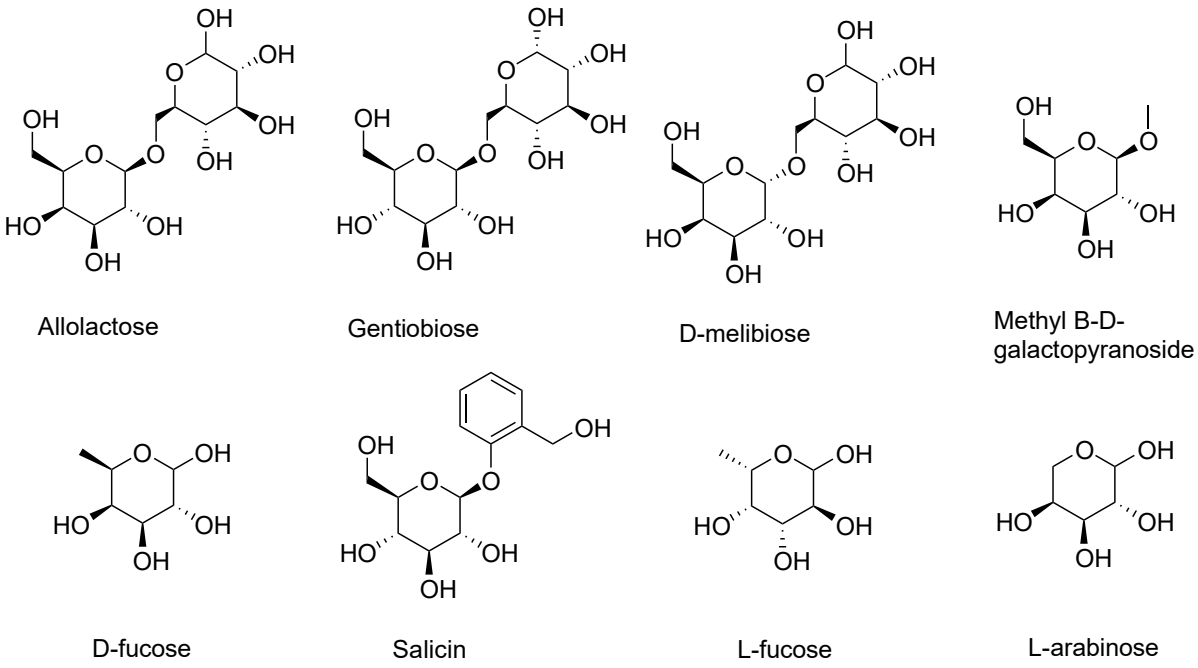
**Suppl. figure 2.4:** A) X-ray structure of Anc1 ligand binding domain determined (LBD) by Georgelin[500]. B) X-ray structure of Anc4 LBD determined by Georgelin[500]. C) X-ray structure of dimeric LacI[453] (PDB: 1EFA). D) Overlay of predicted Anc1 structure with the X-ray Anc1 LBD structure. E) Overlay of the predicted structure of Anc4 with the X-ray structure of Anc4 LBD.




**Suppl. table 2.2:** Comparison of the root mean square deviations of atomic positions (RMSDs) of the predicted ancestral structures with the X-ray structures of Anc1 ligand binding domain and Anc4 ligand binding domain[500], as well as the X-ray structure of dimeric LacI[453] (PDB: 1EFA).

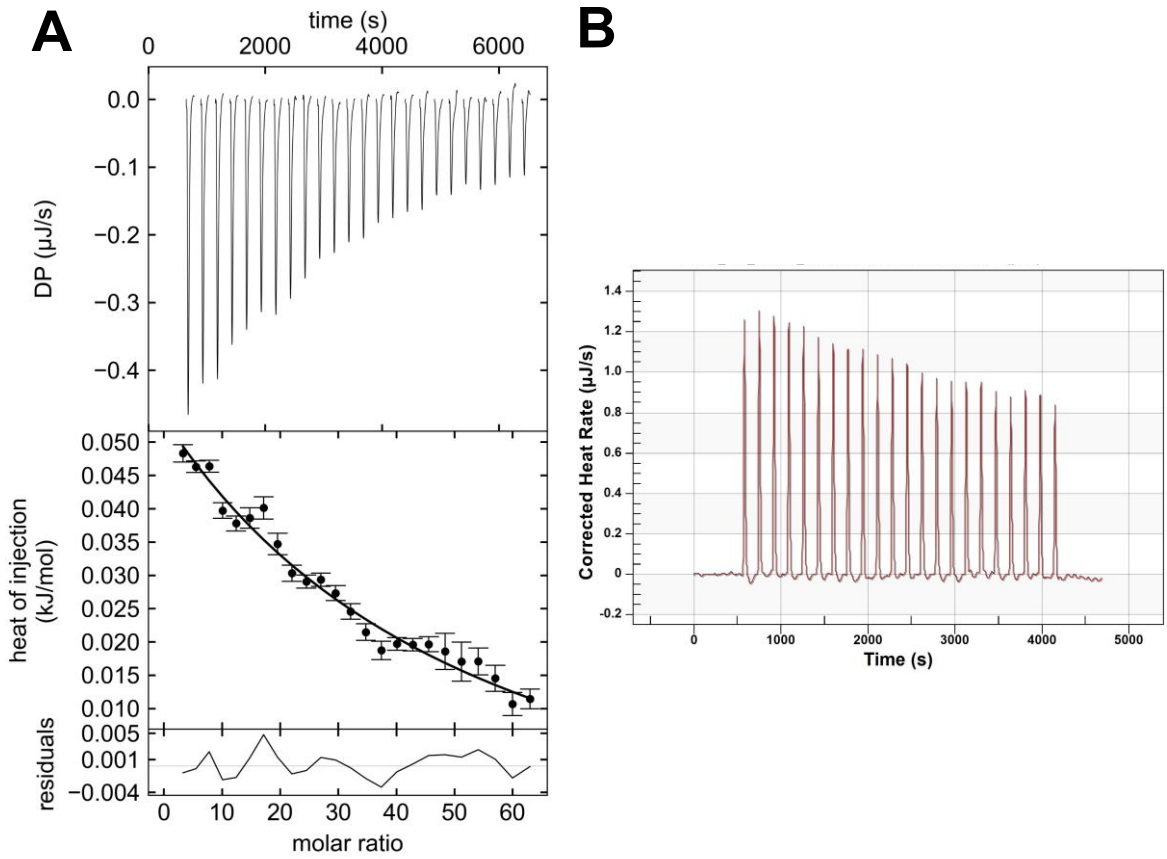
| Structure 1 | Structure 2 | RMSD (Å) | Sequence identity |
|---|---|---|---|
| Anc1 | Anc2 | 1.408 | 59% |
| Anc1 | Anc3 | 0.863 | 54% |
| Anc1 | Anc4 | 1.973 | 48% |
| Anc1 | LacI X-ray | 1.064 | 37% |
| Anc1 | Anc1 LBD X-ray | 1.236 | N/A |
| Anc2 | Anc3 | 0.819 | 87% |
| Anc2 | Anc4 | 0.903 | 71% |
| Anc2 | LacI X-ray | 2.154 | 54% |
| Anc3 | Anc4 | 1.462 | 76% |
| Anc3 | LacI X-ray | 1.218 | 55% |
| Anc4 | Anc4 LBD X-ray | 0.611 | N/A |
| Anc4 | LacI X-ray | 2.475 | 69% |

**Suppl. figure 2.5:** Structures of the sugars identified as potential ligands of the ancestors and LacI in the binding screen and the structure of a native ligand of LacI, allolactose.



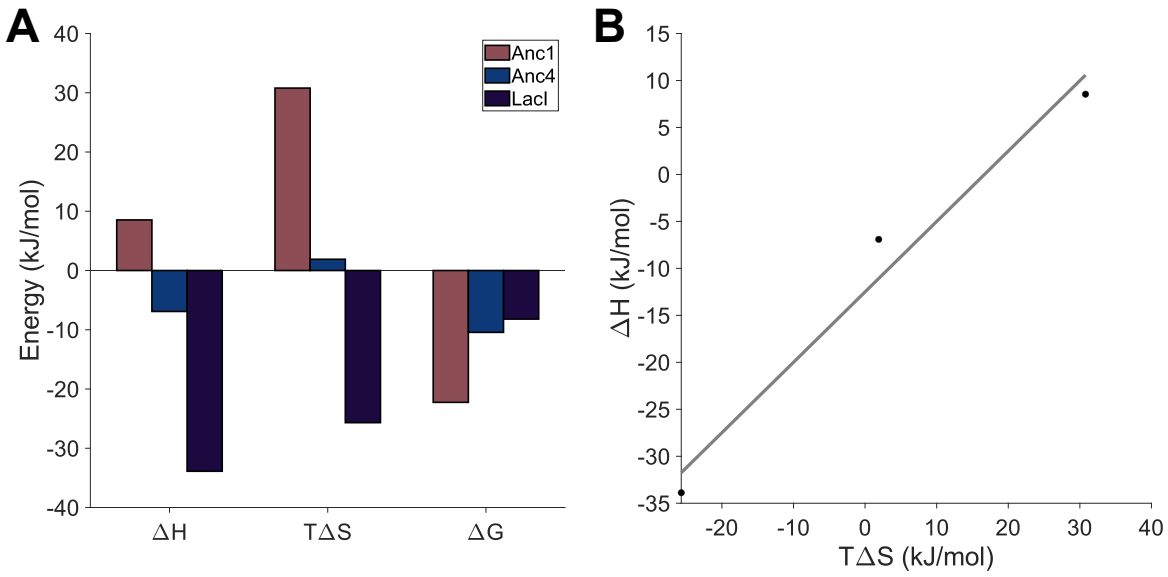

**Suppl. figure 2.7:** A) Thermogram and titration curve for D-fucose binding to Anc2. A concentration of 866 µM Anc2 was used in the cell and 153.5 mM D-fucose was used in the syringe, thus concentrations were nearing precipitation levels for Anc2 and saturation levels for D-fucose respectively. Reducing the concentration of Anc2 resulted in a signal too small to be measured (data not shown). The binding reaction did not saturate, such that it could not be characterised. B) Thermogram of D-fucose binding to Anc3. A concentration of 157 µM Anc3 was used, nearing its precipitation limit, and 150.1 mM D-fucose. The change in heat over the injections was too small to be characterised. Other experiments with lower D-

fucose concentrations showed no change in heat rate for each injection, thus were likely measuring the heat of dilution (data not shown).

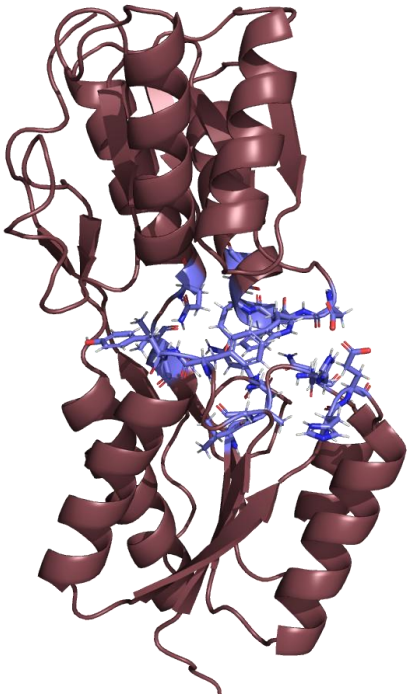



**Suppl. figure 2.8:** A) The change of enthalpy, change of entropy and change in Gibbs free energy of the binding of D-fucose to Anc1, Anc4 and LacI. LacI binds in an enthalpically-driven manner whereas the binding reaction of Anc1 is entopically-driven. B) The

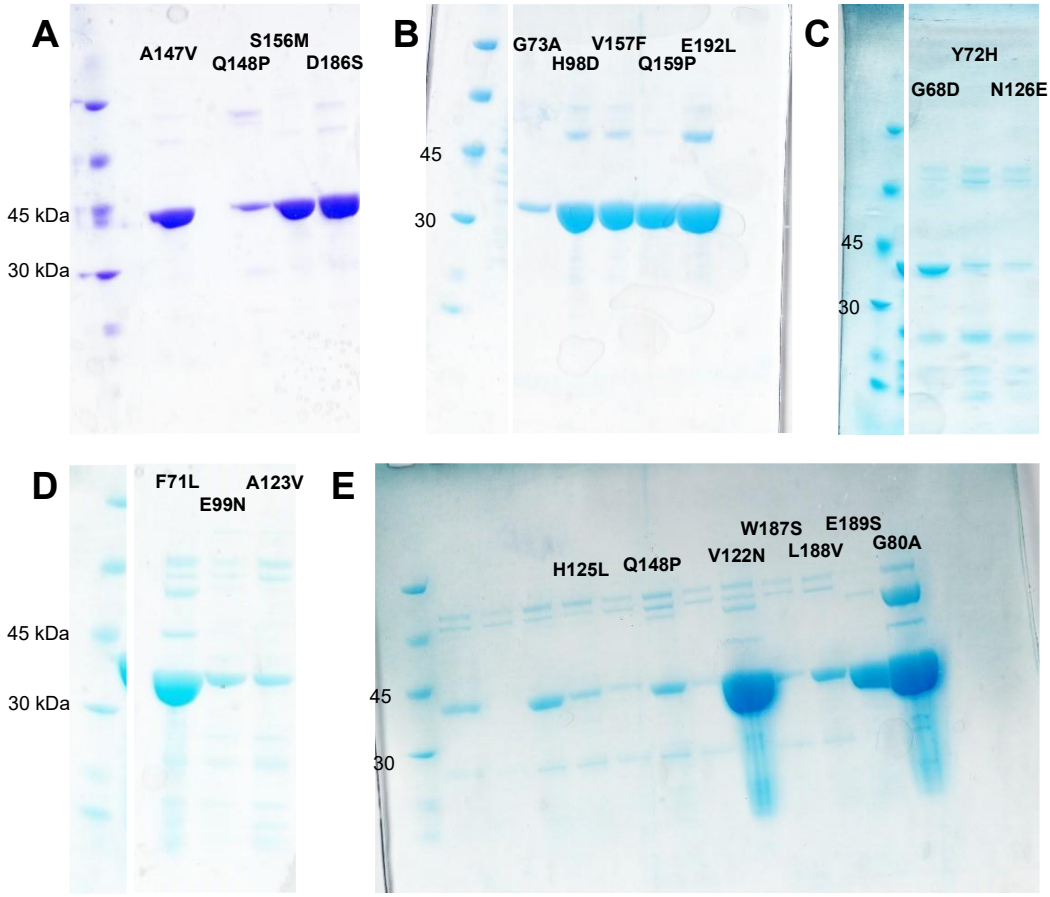
relationship between the change in enthalpy and change in entropy of D-fucose binding to Anc1, Anc4 and LacI. A Pearson correlation coefficient of 0.99 was calculated.



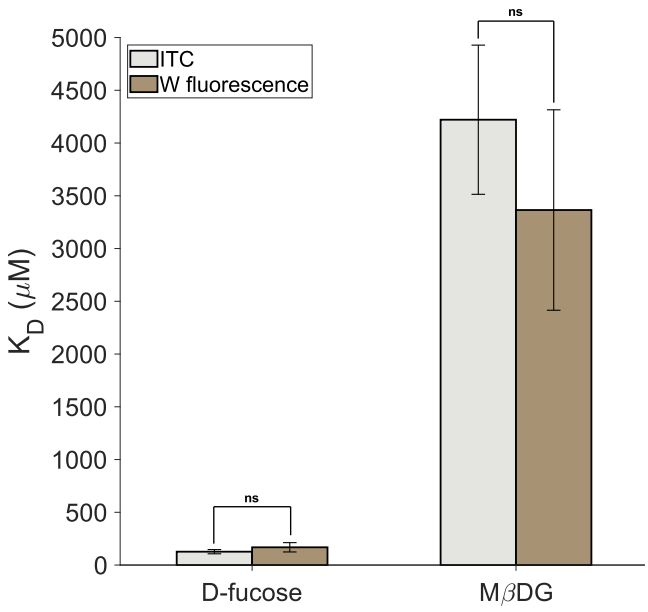

**Suppl. figure 2.9:** The positions of the mutations near the ligand binding pocket between Anc1 and Anc2 are shown in purple on the background of the predicted monomeric unit structure of Anc1, shown in maroon.



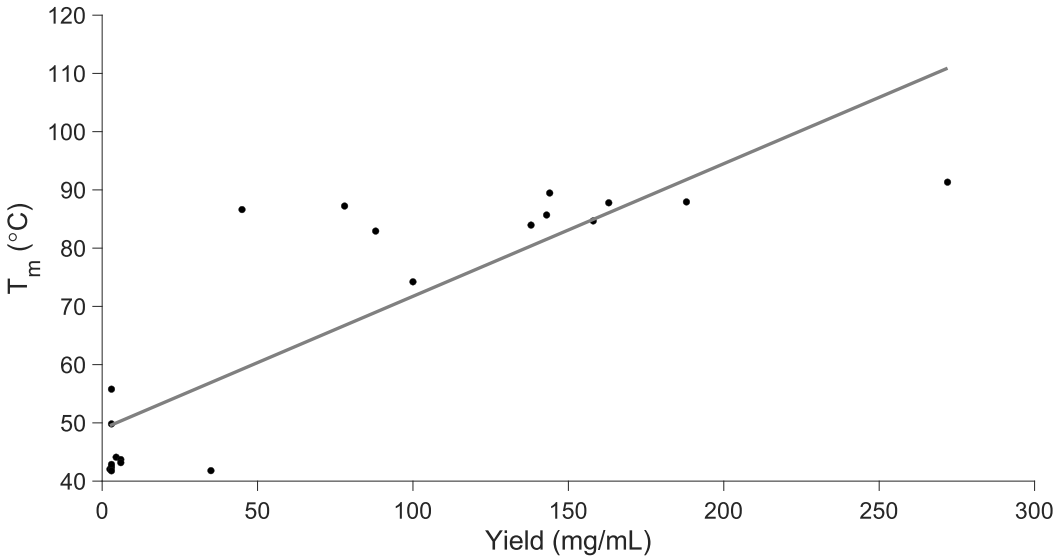
**Suppl. figure 2.10:** SDS-PAGE of Anc1 single mutants with Anc2 mutations. First lane in each gel is the ladder, and fractions after IMAC purification are labelled according to their mutation.




**Suppl. figure 2.11:** Comparison of the equilibrium binding dissociation constants (KDs) determined using isothermal titration calorimetry (ITC) and tryptophan (W) fluorescence for binding of D-fucose and methyl β-D-galactopyranoside (MβDG). Significance was determined by unpaired t test.



**Suppl. figure 2.12:** Positive correlation between the recombinant expression yield and thermostability, as measured by the thermal unfolding transition temperature ($T_m$), of the Anc1 single mutants. A Pearson correlation coefficient of 0.85 was calculated.



# Chapter 3: Engineering a bacterial expressing peroxidase using ancestral sequence reconstruction

## 3.1 Abstract

Horseradish (*Armoracia rusticana)* peroxidase (HRP) is a widely used enzyme with many applications, including in the enzyme-mediated polymerisation of materials like poly(*N*-isopropylacrylamide) (PNIPAm). However, HRP is commonly sourced from its native plant root, which is a laborious, low-yielding process and results in large batch-to-batch variation. Previously, there have been many attempts to ameliorate this by expressing HRP recombinantly. However, heterologous expression of soluble and active, homogenous horseradish peroxidase has proven to be difficult in bacterial systems. Here we utilised ancestral sequence reconstruction of the class III peroxidase family to predict ancestral peroxidases of HRP that had improved soluble, active expression in *Escherichia coli*. It was found that recombinant ancestors were able to act as peroxidases and possessed high thermostability. We further demonstrated that one of the ancestral peroxidases, Anc9, was able to catalyse the polymerisation of PNIPAm in a similar manner to commercial plant-derived HRP. Overall, this work advanced our understanding of the evolution of the class III peroxidase family, as well as demonstrating that ancestral sequence reconstruction can be used as a tool to improve heterologous expression industrially relevant enzymes.

## 3.2 Introduction

Horseradish peroxidase (HRP), which is produced natively in the *Armoracia rusticana* plant root, is an important and widely used enzyme due to its ability to catalyse a single electron redox reaction transfer on its substrate. This enzyme has a long history, with the effects of HRP having been observed in 1810, before 'enzymes' were even known of, by Louis Antoine Planche[600,601]. Planche found that various plant roots, including that of *A. rusticana*, were able to turn the *Guaiacum* root blue. Since then, it has been found that HRP is indeed a useful enzyme, with redox chemistry that can be used to drive many reactions, including coupling reactions, polymerisation, hydroxylation, N- and O-dealkylation and oxygen transfer[602]. In particular, HRP can be used to catalyse the production of a chromogenic, fluorogenic or electrochemical signal[603]. HRP has a relatively wide range of substrates and has good thermal and oxidative stability compared to other peroxidases[604]. Consequently, HRP is utilised in a

wide range of applications, including in immunoassays[605-607], biosensing systems[608-610], in bioremediation to degrade phenol-containing pollutants[611-613] and increasingly in material chemistry to drive polymerisation reactions[614-616](Fig. 3.1).

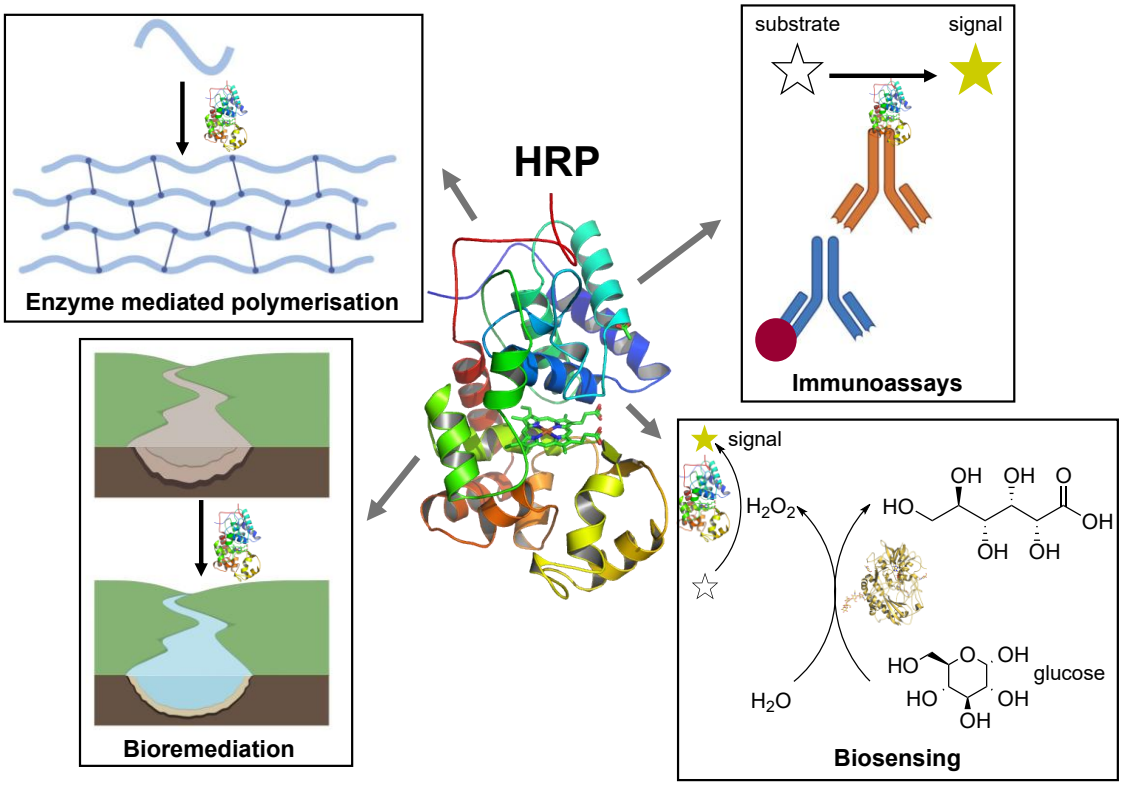



**Figure 3.1:** Horseradish peroxidase (HRP) has a wide range of applications due to its catalytic activity involving single electron transfers, which can initiate radical polymerisation, degrade phenol-containing pollutants, and catalyse the formation of products which generate a chromogenic signal, light signal, difference in charge, difference in pH, etc. The production of a signal underlies its widespread use in biosensing, e.g. glucose sensing, and in immunoassays, where HRP is attached to an antibody. Figure made using BioRender, ChemDraw, PyMOL and Microsoft PowerPoint.

One major use of HRP is to catalyse polymerisation reactions through enzyme-mediated radical polymerisation (EMRP). In EMRP, polymers are formed through the actions of oxidoreductase enzymes[294]. These enzymes are usually involved in the initiation stage by catalysing the formation of free radical compounds, which then leads to the polymerisation of the material. Currently, some of the most utilised enzymes in free-radical EMRP include horseradish peroxidase (HRP), glucose oxidases (GOx) and laccases[316]. These enzymes have been used to catalyse the formation of various materials including polyphenols, polyacrylamides and polyvinyls. Some advantages of using EMRP over traditional chemical methods are that enzymes are non-toxic, have high regio-, chemo-, and stereo-selectivity, and perform their catalysis under relatively mild conditions[298]. Furthermore, as the enzymatic reaction will not start without the substrate(s), the reaction can be performed ‘on demand’.

These benefits of EMRP have resulted in its common use to form hydrogels[316], which are polymeric materials with considerable aqueous content. Hydrogels are frequently used in biomedical or environmental applications due to their biocompatibility, thus, their *in situ* formation using non-toxic enzymes is especially beneficial. For example, Wei *et al.*[617] developed a double network hydrogel with controllable drug release properties for wound healing, where the polymerisation of both networks was initiated by glucose oxidase.

Another example of a hydrogel that can be formed through EMRP is poly(*N*-isopropylacrylamide) (PNIPAm)[340]. PNIPAm, often crosslinked by *N*,*N*′-methylenebisacrylamide (MBA) (Fig. 3.2A), is a biocompatible hydrogel[618,619], with the relatively special behaviour of 'contracting' at temperatures above its lower critical solution temperature (LCST)[620]. This thermosensitivity thereby adds an extra dimensionality to the material. The observed 'contraction' results from a coil to globule phase transition above the LCST, originating from the presence of both hydrophobic and hydrophilic moieties in the polymer. At lower temperatures, where entropy is less weighted, it is more favourable for water molecules to bond to the hydrophilic groups at the entropic cost of losing some of their freedom near the hydrophobic groups. However, at higher temperatures, where entropy is weighted more greatly, the entropic cost outweighs the enthalpic benefit of binding, such that water is largely excluded from the material (Fig. 3.2B). Typically, the LCST of PNIPAm is around 32 °C[621], making it ideal for applications in biomedicine and soft robotics[622]. For example, many PNIPAm-based gels have been developed for drug release[623] and increasingly PNIPAm-based gels have also been used as actuators[624]. However, the chemically mediated formation of PNIPAm is disadvantageous because it is usually conducted using toxic chemical initiators under nitrogen or vacuum[625]. As an alternative, PNIPAm has instead been formed by EMRP using a paired system of HRP and GOx[308,361] (Fig. 3.2A), which uses the non-toxic enzymes as initiators and can be conducted under atmospheric oxygen, due to the oxygen scavenging ability of GOx. However, the batch-to-batch variation of *A. rusticana* HRP (*Aru*HRP) results in gel formulations needing to be adapted for each *Aru*HRP batch. Consequently, a recombinant, homogeneous source of a relatively stable peroxidase would be beneficial for the formation of PNIPAm, and more generally for the field of EMRP.

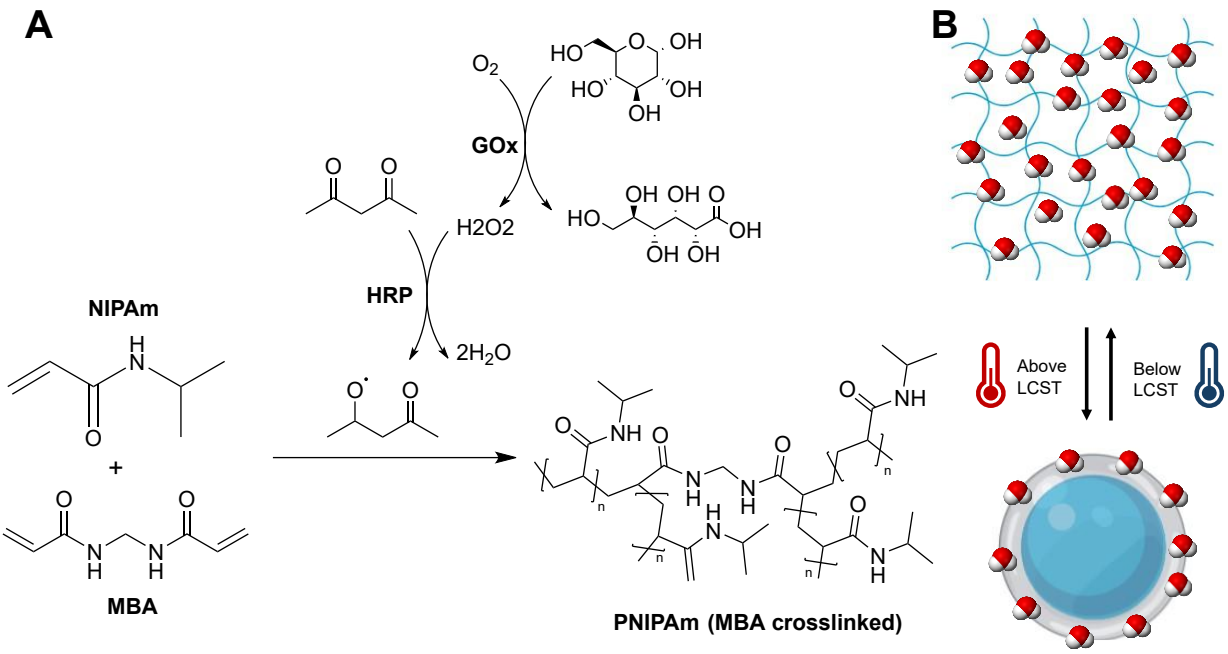



**Figure 3.2:** A) Schematic of the formation of *N,N*-methylenebisacrylamide (MBA) crosslinked poly(*N*-isopropylacrylamide) (PNIPAm) hydrogel initiated using glucose oxidase (GOx) and horseradish peroxidase (HRP). B) The thermosensitivity of PNIPAm is shown. At temperatures below the lower critical solution temperature (LCST) the chains of PNIPAm bind water, however, at higher temperatures water is excluded from the PNIPAm and the gel appears to contract. Figure made using ChemDraw and BioRender.

Peroxidases are oxidoreductase enzymes that catalyse the oxidation of substrate(s) by reducing hydrogen peroxide to water. Peroxidases are split into two classes based on whether they coordinate iron-protoporphyrin IX (heme): heme peroxidases and non-heme peroxidases. Heme-containing peroxidases, such as HRP, contain a widely conserved proximal histidine to coordinate the iron ion due to its importance in the catalysis[626,627]. There are two main superfamilies of heme-containing peroxidases: the animal and non-animal superfamilies[628]. It should be noted that these classifications are primarily structure-based and should not be taken literally, as while most proteins in the animal superfamily are present in animals, some can also be found in fungi and bacteria, and conversely, some non-animal superfamily proteins can be found in animals[629,630]. HRP is found within the non-animal superfamily, which is further divided into three classes of peroxidases: class I, II and III (Fig. 3.3). The peroxidases of class I tend to be intracellular and can be found in a wide range of organisms from plants and fungi to prokaryotes[628]. The class I family is the most catalytically diverse, consisting of catalases, cytochrome c peroxidases and ascorbate peroxidases, amongst others. It is also the most diverse in subcellular localisation and function, with members being localised to the cytosol, mitochondrion, chloroplast or secreted, and exhibiting functions including electron shuffling in the mitochondrial membrane, stress

response regulation and $H_2O_2$ degradation[628]. As such, it has been predicted that the class I peroxidase family is the oldest of the three classes, possibly originating in an ancestral bacterium[631], and that the class II and class III peroxidase families both arose from ancestral class I proteins[629]. The class II peroxidase family is found in fungi and consists of secretory peroxidases that have generally been found to have a lignin-degrading function[630]. Members of this family include lignin peroxidases, versatile peroxidases and manganese peroxidases. It has been posited by Mathé *et al.*[632] that the class II family of peroxidases may have evolved from a class I peroxidase ancestor sometime after the separation of the Ascomycetes and Basidiomycetes fungal divisions. The final class is the class III peroxidases, of which HRP is part of. Members of this class are only found in plants and some algae, and as such Mbadinga *et al.*[629] posited that this class evolved in the ancestor of Streptophyte algae, potentially from an ancestral cytochrome c peroxidase. Perhaps as a result of extensive gene duplication in Viridiplantae, there now exists a relatively large number of class III peroxidases[627,633]. Members of this class are commonly secreted into the cell wall or vacuole and have many functions including cell wall metabolism, wound healing, auxin metabolism, $H_2O_2$ scavenging, defence of the plant, cell growth and in ROS pathways[634]. The physiological functions of HRP alone are believed to be quite diverse, from lignin and suberin formation to indole-3-acetic acid metabolism and infection resistance[603]. The structure of the class III peroxidases is generally highly conserved, with several conserved residues for heme binding, calcium binding, disulfide bonds, substrate binding and for the peroxidase catalysis[627]. Nevertheless, the glycosylation patterns tend to be much more variable.

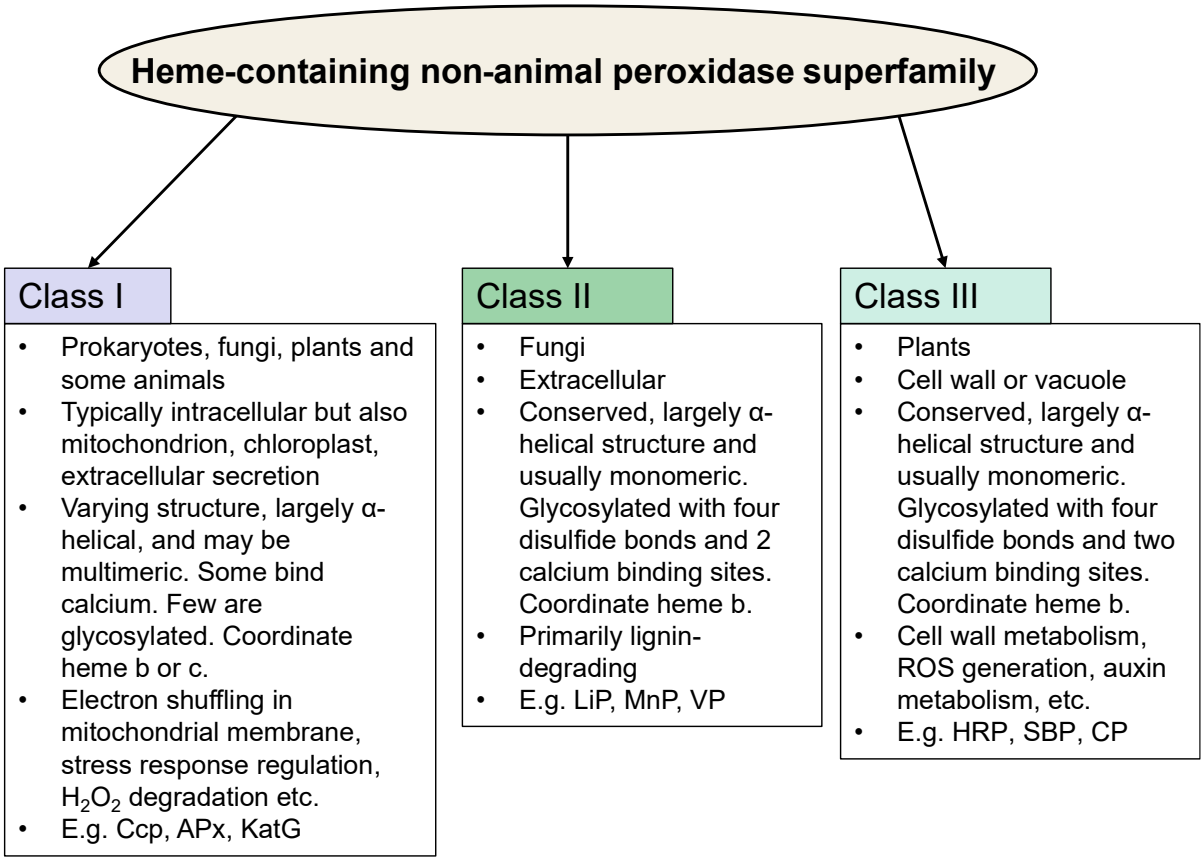



**Figure 3.3:** Of the heme-containing peroxidases, the non-animal superfamily is thought to be phylogenetically split into three main families: class I, class II and class III. A brief description of each class is provided, detailing their native organism, subcellular localisation, structural characteristics, functions and some examples of members. Ccp are cytochrome c peroxidases, APx are ascorbate peroxidases, KatG are catalases, LiP are lignin peroxidases, MnP are manganese peroxidases, VP are versatile peroxidases, HRP is horseradish peroxidase, SBP is soybean peroxidase, CP is cationic peanut peroxidase. Figure made using Microsoft PowerPoint.

Of the class III peroxidases, HRP is the most utilised and studied[627]. However, when HRP is sourced natively from the horseradish root, it is present as a mixture of isoenzymes. Of these, the isoenzyme C1A is present in the largest amounts[603], and thus has been the most widely investigated. Hence, this study will also focus on the C1A isoform due to the availability of previous knowledge, as well as its desirable catalytic properties. As a peroxidase, the HRP C1A isoenzyme catalyses the reduction of hydrogen peroxide to water by oxidising two equivalents of a reducing substrate[635]. A simplified explanation of the catalytic cycle[314] (Fig. 3.4) is that hydrogen peroxide oxidises the iron centre from Fe(III) to Fe(V) (although in reality there is considerable electron delocalisation over the heme cofactor) to form compound I. Then, this reduced iron-heme centre returns to its resting state of Fe(III) in two steps: first by oxidising one substrate molecule to form compound II, then second by oxidising another substrate molecule to return to the resting state of the enzyme. This

catalytic cycle of HRP C1A acts upon a relatively wide range of substrates, which, although typically phenolic[603,636], can also include other molecules with sufficient redox potential including acetyl acetate[637] and ferricyanide[638]. However, a limitation of the chemistry of HRP C1A is that the enzyme may become inactivated if there is too much $H_2O_2$ present, as this favours formation of compound III[332] (Fig. 3.2). Consequently, HRP C1A generally functions best when $H_2O_2$ is generated *in situ*. Such $H_2O_2$ generation can be conducted enzymatically through enzymes such as glucose oxidase (GOx), or otherwise can be achieved through use of a feed[316].

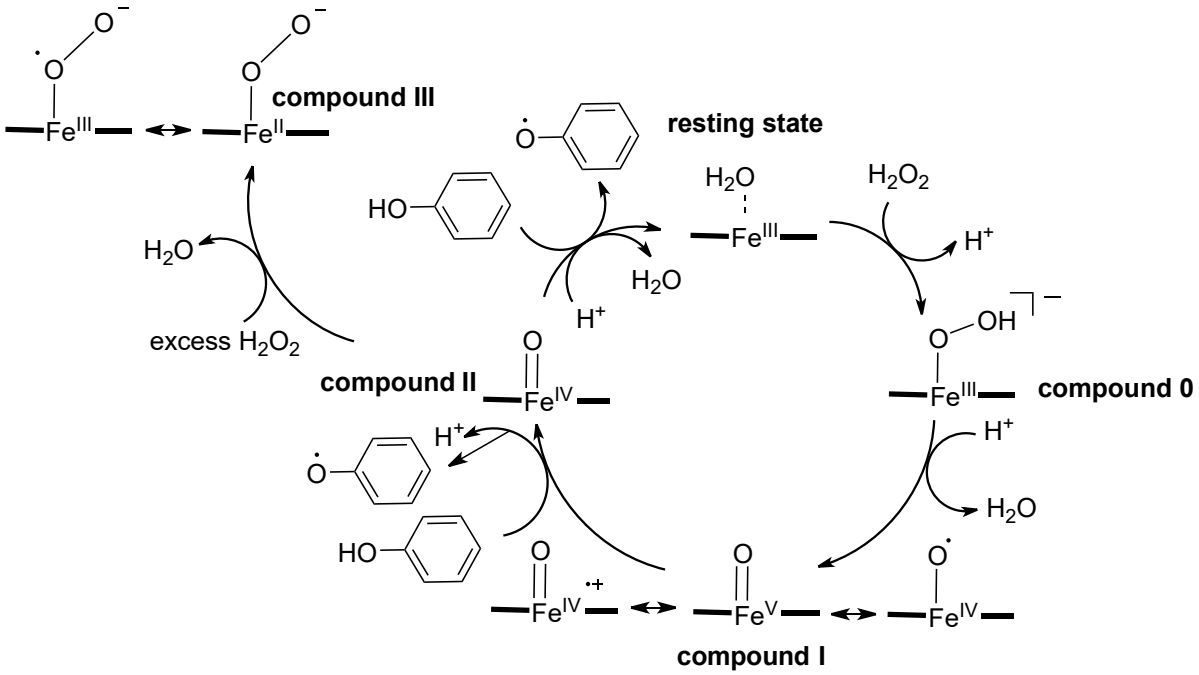



**Figure 3.4**: Simplified catalytic cycle of horseradish peroxidase C1A[603]. From the resting state, $H_2O_2$ reacts with the iron centre to form compound 0 which subsequently becomes compound I. Compound I has considerable delocalisation over the coordinated heme group and iron centre. Subsequently, a substrate, such as phenol, is then able to conduct a redox transfer with compound I to form compound II. A second substrate molecule, then undergoes a redox reaction again to return the enzyme to its resting state. Under conditions of excess $H_2O_2$ the formation of compound III is favoured, thereby limiting the catalytic ability of the enzyme, although this inhibition is reversible. Figure made using ChemDraw.

In terms of structure, HRP C1A is typical of a class III peroxidase, with a globular, largely α-helical monomeric structure that coordinates an iron-protoporphyrin IX (heme) cofactor[639] (Fig. 3.3). The iron ion in the cofactor is coordinated to a conserved 'proximal' histidine (His170) in the protein. HRP C1A also contains other features conserved in the class III peroxidase family, including two calcium ion binding sites and four disulfide bonds. HRP C1A is a glycoprotein, with eight sites that are natively N-glycosylated[640], and a total carbohydrate content of ~19%[641]. Another important feature of HRP C1A is that it is

processed by the endoplasmic reticulum and Golgi secretory pathway, before being localised to either the tonoblast, plasma membrane or cell wall *via* its N-terminal signal peptide and C-terminal propeptide[642,643].

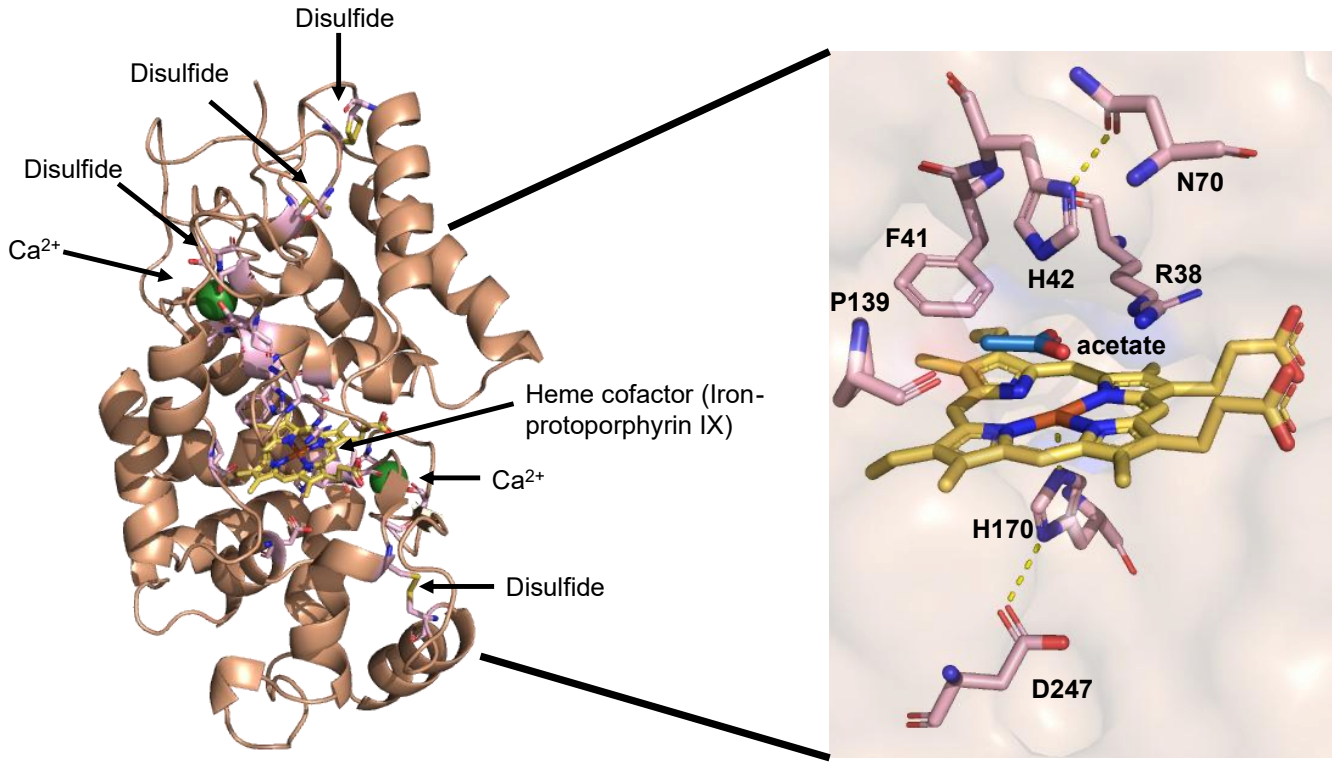



**Figure 3.5:** Structure of horseradish peroxidase (HRP) isoform C1A. The full structure on the left (PDB: 1HCH) demonstrates the positions of the calcium ions (green), disulfide bonds (yellow for sulfur and pink for carbons), heme cofactor (yellow for carbons, blue for nitrogen, orange for iron, red for oxygen) and other important residues (pink). On the right is a closer view of the catalytic centre (PDB: 1H5A) with acetate (blue) bound, the heme cofactor (yellow) and some of the residues important for catalysis (pink). Residue R38 stabilises both the substrate and compound I, F41 prevents substrate access to the ferryl oxygen of compound I, H42 accepts a proton from $H_2O_2$ and stabilises the substrate, N70 increases basicity of H42, P139 is thought to stabilise some substrates, H170 coordinates the iron centre and D247 increases the basicity of H170[603]. The surface of HRP is also shown in orange (with high transparency), highlighting that the substrate entrance/product exit pocket is facing out of the page towards the viewer. Figure made using PyMOL.

Although HRP is a powerful catalyst and consequently so often utilised, for enzyme-mediated radical polymerisation and other applications, its current sourcing from the *A. rusticana* plant root[644] is problematic. The two main issues are firstly, that the extraction process is time-consuming and results in relatively poor yields[645], and secondly, that the final product is a heterogeneous mixture of isoenzymes[644,646]. As of the time of writing, it is still not fully

understood which environmental or internal factors might be contributing to the differential expression of the HRP isoenzymes in the plant root[647,648] . Furthermore, it has been found that the over 40 different isoenzymes have different physical and catalytic properties[646,649]. As a result, the sourcing of HRP from the plant root is inadequate, especially given that for its applications it would be best if the enzyme displayed the same properties and catalytic activities every time. The relatively high cost of *A. rusticana* sourced HRP also limits its industrial use[602]. Therefore, to achieve a homogeneous, high-yielding source of HRP, recombinant expression needs to be used.

Recombinant expression of HRP C1A in heterologous hosts has proven to be difficult. This is likely due to factors involved in its native expression and processing in the horseradish plant that are not present or are different in heterologous hosts. These factors include codon usage, disulfide bond formation, glycosylation and differences in chaperones. The frequency of use of different codons tends to differ significantly between distantly related organisms, such as *A. rusticana* and *Escherichia coli* [650]. This is thought to affect protein expression due to several factors, including by leading to differences in elongation speed and translation efficiency[651]. While codon optimisation can be used for heterologous expression, it generally favours the most frequently used codons of a given organism and as such may not capture the nuances required for optimal protein folding and expression. Another factor to consider is that HRP C1A and other class III peroxidases contain disulfide bonds, which are important for their structure. However, for some heterologous hosts, such as *E. coli*, it is difficult to overexpress disulfide-bond containing proteins because they typically need be directed to the periplasm[652], which may result in lower expression yields than cytoplasmic proteins, or strains such as SHuffle T7 Express[653] can be used, which also may result in relatively low yields depending on the target protein for various reasons. Another fundamental difference between the expression of HRP C1A in its native host *versus* heterologous hosts are the posttranslational modifications like glycosylation[654], which vary considerably between distant eukaryotes, such as plants and yeast, and may not even be present in some prokaryotes. The glycosylation pattern of a protein is important to its structure and function, and when HRP is produced lacking glycosylation or with different glycosylation patterns this was found to affect its solubility[644] and catalytic activity[655]. Similarly, chaperones can also vary greatly between distantly related organisms[26,656]. Hence, if the folding of HRP C1A is chaperone-assisted in its native organism, then without the horseradish chaperones in heterologous organisms, the folding of HRP C1A might tend to get stuck in intermediate folding states[657].

Consequently, HRP C1A could be more prone to aggregation and/or proteolysis when expressed heterologously. Previously, Kondo *et al*.[658] attempted to improve the expression of HRP C1A by co-expressing it with various *E. coli* chaperones. However, only very small gains in yield were obtained, perhaps due to the differences between the chaperones in *E. coli* and *A. rusticana*.

Due to the aforementioned barriers amongst other factors, recombinant expression of HRP C1A has often resulted in low yields and/or low enzymatic activity. Previous attempts to improve expression and other properties have included adding various solubility tags[659], expressing with various chaperones[658], optimising expression conditions[644,660], attempting to express in various organisms or cell-free[661-664] and directed evolution[286,665]. Of all these strategies, the greatest soluble yield success was obtained either by using the non-GRAS organism *Cryptococcus sp.*[663], or by using a very large, uncleaved, solubilising protein tag[659]. However, neither of these strategies achieved catalytic activity close to the native HRP C1A. Humer *et al.*[644] were able to recombinantly express HRP C1A with high yields and activity, however, as the expression was in *E. coli* as inclusion bodies, the refolding and purification process was relatively laborious and long. Perhaps the most successful recombinant expression in both yield and activity was developed by Krainer *et al.*[664] using *Pichia pastoris (Komagataella sp.)* grown in a bioreactor. Recombinant HRP C1A yields of 132 mg/L were achieved, with activity similar to that of native HRP. However, such a feat is not readily repeatable without the availability of the HRP expressing strain developed by Krainer *et al.*, because *P. pastoris* transformation is highly variable[666]. Furthermore, the HRP C1A produced by *P. pastoris* will always be inhomogeneous due to varying amounts of glycosylation[666]. Therefore, to develop a homogeneous recombinant peroxidase, the use of a prokaryotic expression host would be preferred. We propose to use laboratory strains of *E. coli* because they are GRAS organisms that are relatively cheap to work with, are quick to grow and produce protein, and they have many existing synthetic biology tools designed for them[538]. To improve the soluble expression and activity of HRP expressed in *E. coli*, we propose to use ancestral sequence reconstruction (ASR) as a tool to predict ancestral proteins, which could have improved expression in *E. coli* whilst retaining similar catalytic properties.

Ancestral sequence reconstruction is relatively underutilised in the Viridiplantae, with many studies instead focusing on genome wide association studies (GWAS) to better understand plant proteins[34,629,667]. However, using techniques that rely only on extant proteins, such as GWAS, can result in information being obfuscated by factors like the highly epistatic nature

of residues in proteins and/or neutral mutations accumulating over time[19]. Furthermore, some extant proteins can be relatively specialised and can be thought to be on 'peaks' in the sequence-function landscape[263]. This can then make it difficult to engineer new properties or functionality using extant proteins as a starting point, because the vast majority of mutations will be deleterious or neutral[248]. In contrast, ASR contains the additional dimension of time, thereby allowing a considerably clearer and deeper understanding of how a given protein family has changed and evolved over time along its different lineages. As such, it can be elucidated which residues are most important for functional change and which are only ancillary or neutral, as well as the order of the mutations over time, which is critical for epistatic interactions. By using ASR, the various 'peaks' of the sequence-function landscape can be explored in a much more productive and informative manner. We hypothesise that the use of ASR for the class III peroxidase family will predict ancestral proteins with the desired HRP-like peroxidase function due to the highly conserved nature of peroxidase catalysis in extant class III family members[634]. Additionally, as predicted ancestral proteins are often more stable[14,23], some of the predicted class III ancestral peroxidases may also be relatively stable. As relatively high stability has been found to lead to improved heterologous expression in some cases[668-670], some of the predicted class III ancestral peroxidases, therefore, may also exhibit increased recombinant expression in *E. coli* compared to HRP. It should be noted that protein stability is not necessarily correlated with improved heterologous folding, but that in some cases they may be linked due to the stabilising mutations also changing the folding landscape of the protein[26]. Regardless, whether directly linked or not, many ancestral proteins exhibit increased foldability as well as stability[9,26,186].

The advantages of using phylogenetic information to better understand and engineer proteins have already been shown for some of the peroxidase families, including the fungal ligninolytic peroxidases (class II)[671]. ASR of this family was used to better understand how structure and catalysis of these peroxidases may have changed over time, thereby providing tools to engineer novel lignin-degrading peroxidases in the future as well as providing evidence for the coevolution of plants and fungi. Another peroxidase ASR study of the dye-decolorising peroxidases (DyP)[672] was able to achieve relatively high yields of *E. coli* recombinant expression of an ancestral DyP, thereby providing support for the use of ASR to improve heterologous peroxidase expression. An additional example of using evolutionary information to engineer a peroxidase with improved heterologous expression was shown by

Barber-Zucker *et al.*[673], where PROSS[674] was used with the versatile peroxidase family to improve their expression in yeast.

The use of ASR with the class III plant peroxidase family should unveil evolutionary mechanisms, as well as potentially yielding more stable ancestral peroxidases, which might have better recombinant expression in *E. coli* than the extant HRP. It should be noted that phylogenetic reconstruction of the class III peroxidase family has been previously attempted by several studies including that of Duroux *et al.*[675], upon which ASR was later performed[676,677]. However, due to the age of the study by Duroux *et al.*[675], the sequence collection and curation was limited, with sequences from only two organisms making up 42% of their sequence set. Furthermore, the phylogenetic reconstruction methods were also suboptimal compared to modern methods. For example, ClustalX[678] was used for sequence alignment and the quartet puzzling algorithm[164] was used during maximum likelihood tree inference, both of which have since been superseded[20,57,166,167]. Consequently, due to the relatively inaccurate methods of phylogenetic reconstruction by Duroux *et al.*[675], the later ancestral sequence reconstruction was likely inaccurate by modern standards[19,679]. Perhaps as a result, a low recombinant yield of 1.4 mg/mL in *E. coli* was reported for the ancestral peroxidase, and its activity was not reported[677]. Whilst later phylogenetic studies used improved methods, the sequence sets were still limited both in number of sequences and in their taxonomic diversity due to the narrower focus of the studies[634,680]. Consequently, here the phylogenetic inference and ancestral sequence reconstruction of the class III peroxidase family will be conducted again *ab initio* using modern sequence databases and modern software.

Despite evolutionary models improving substantially over time, the complexity of the evolutionary process results in it being extremely difficult to perfectly capture in a model[167]. This leads to what are known as incongruences in either the phylogeny topology or in the ancestral sequence reconstruction. General sources of incongruence are discussed in more detail in Chapter 1.2.5. Here, only those that should be especially considered with respect to plant phylogenetics will be mentioned. Plants differ from other kingdoms of life in their relatively high rate of polyploidy, hybridisation and genome duplication, which can lead to incongruences[25]. As such, during plant protein phylogenetics particular care should be taken when collecting and curating the protein sequences, as well as during the phylogenetic reconstruction. Please refer to Chapter 1.2 for a more detailed description of best practices. Nevertheless, when using best practices, ASR is still a powerful technique for building a

model of plant protein evolution, as well as for predicting ancestral proteins that may have different and desirable properties, like increased stability and recombinant expression.

In summary, while HRP is a widely used enzyme, it is commonly extracted from the native plant root, which leads to heterogeneity, low yields, and batch-to-batch variation. These problems could be resolved by recombinant expression of HRP C1A in *E. coli*, which is also relatively easy to use and low cost. However, previous attempts at *E. coli* expression have resulted in low yields, low activity, laborious refolding from inclusion bodies or use of large fusion proteins as solubilising tags. Here, we instead utilised ASR of the class III peroxidase family to identify ancestral peroxidases that could be expressed solubly in *E. coli* without the use of fused solubilising proteins. We found that the recombinantly expressed ancestral peroxidase 'Anc9' displayed higher thermostability than native HRP, as well as possessing comparable enzymatic activity to native HRP. It was demonstrated that Anc9 was able to catalyse the polymerisation of PNIPAm to form a material with similar properties to that catalysed by native HRP. In future, if the yields and heme incorporation of recombinant Anc9 are improved, it could provide a replacement for HRP in many applications. Here we demonstrated the advantages of using ASR as a tool to improve heterologous expression, thereby adding to the growing number of studies also showing the benefits of ASR for ameliorating heterologous protein expression[9,26].

## 3.3 Results

### *3.3.1 Sequence collection and curation.*

The first stage for the ancestral sequence reconstruction of the class III peroxidases was the collection and curation of homologous class III peroxidase sequences. Initial sequence collection and curation was conducted as described in Methods (section 3.5.1). To attempt to ensure appropriate diversity of sequences for accurate phylogenetic inference and ancestral sequence reconstruction[25,46], sequences were then clustered by 80% identity and representative sequences selected using CD-HIT[681]. To curate the sequence set such that it contained homologous sequences and a good representation of different protein sequences, the set representation was compared to the Viridiplantae organismal phylogeny[682-684]. Class III plant peroxidases are theorised to have evolved in the common ancestor of plants and Streptophyta[629], hence, the sequence set should ideally contain all major plant clades except Algae. The sequence set did indeed contain all major plant clades except Algae; however, when compared to the organismal phylogeny, the peroxidase sequence set lacked sequences from Polypodiophyta and Bryophyta. This may be because these clades are currently underrepresented in genotyping efforts[628,685]. Nevertheless, some additional sequences from these clades for predicted class III were manually identified from the RedoxiBase database[628] and added to the sequence set to better represent the organismal phylogeny and ensure good diversity of sequences.

The resulting final sequence set of class III peroxidases mostly contained proportional representation of the Viridiplantae clades, except for the possible over-representation of core rosid sequences and under-representation of asterid sequences (Suppl. Fig. 3.1). As the core rosid clade is predicted to have undergone many whole genome duplications (WGDs), this may have increased peroxidase diversity in this clade[682,686]. Furthermore, many plants are known to contain multiple, distinct class III peroxidases[628,685]; thus, the relatively large number of core rosid peroxidase sequences compared to core rosid plants could result from this. Consequently, it was decided not to remove any core rosid sequences from the set. Nevertheless, this bias towards core rosid sequences may also have resulted from using *A. rusticana* HRP C1A as a seed sequence, because *A. rusticana* is from the core rosid clade. On the other hand, the reason for the underrepresentation of asterid sequences in the curated peroxidase sequence set is somewhat unclear. The asterid clade is very large, representing ~33% of extant land plants[684,687], and there are a similar number of asterid peroxidase sequences as core rosid peroxidase sequences in the UniProt and RedoxiBase databases[628,685].

Furthermore, the asterid clade is predicted to have undergone a similar number of WGDs to the core rosid clade[682]. Therefore, perhaps the underrepresentation of asterid clade peroxidase sequences may instead arise from a lack of diversity of peroxidase sequences within this clade, resulting in their removal earlier during the sequence curation process. As diversity of peroxidase sequences is important for phylogeny inference and ancestral sequence reconstruction[25,46], it was ultimately decided not to add any more asterid clade peroxidase sequences into the final set.

The relative diversity of the class III peroxidase sequence set was further characterised by constructing sequence similarity networks (SSNs)[33,688], where the edges represent varying cutoffs of pairwise sequence identity (Fig. 3.6). The SSNs revealed that most sequences shared somewhat less than 60% identity, suggesting relatively high levels of sequence conservation amongst the class III peroxidases, but that few sequences shared greater than 75% identity, which is expected given that the set was curated to contain sequences of 80% identity at most. The clusters of the SSNs with varying sequence identity cutoffs tended to mostly contain sequences from organisms of relatively closely related clades, however, there were a few exceptions, which could indicate either paralogs or unexpected rates of evolution[46]. However, this is better validated through the phylogenetic construction process[689]. The SSNs also showed the diversity present within the peroxidase sequences originating from the Rosid clade, as even with 65% identity many Rosid originating sequences had already been separated into different clusters. This further supports the utilisation of a relatively high number of Rosid peroxidase sequences within the class III peroxidase set due to their relatively high diversity. The SSNs also demonstrated the relative divergence of sequences originating from organisms from the Lycophyte, Bryophyte and Fern clades, as these sequences tended to have low connectivity. This could be due to the relatively early divergence of these clades in the Viridiplantae[682], which may have led to relatively unique evolutionary changes in their class III peroxidases. The final curated set of class III peroxidase sequences contained 328 sequences of relatively high sequence diversity and of relatively diverse organismal origin (Suppl. Table 3.1).

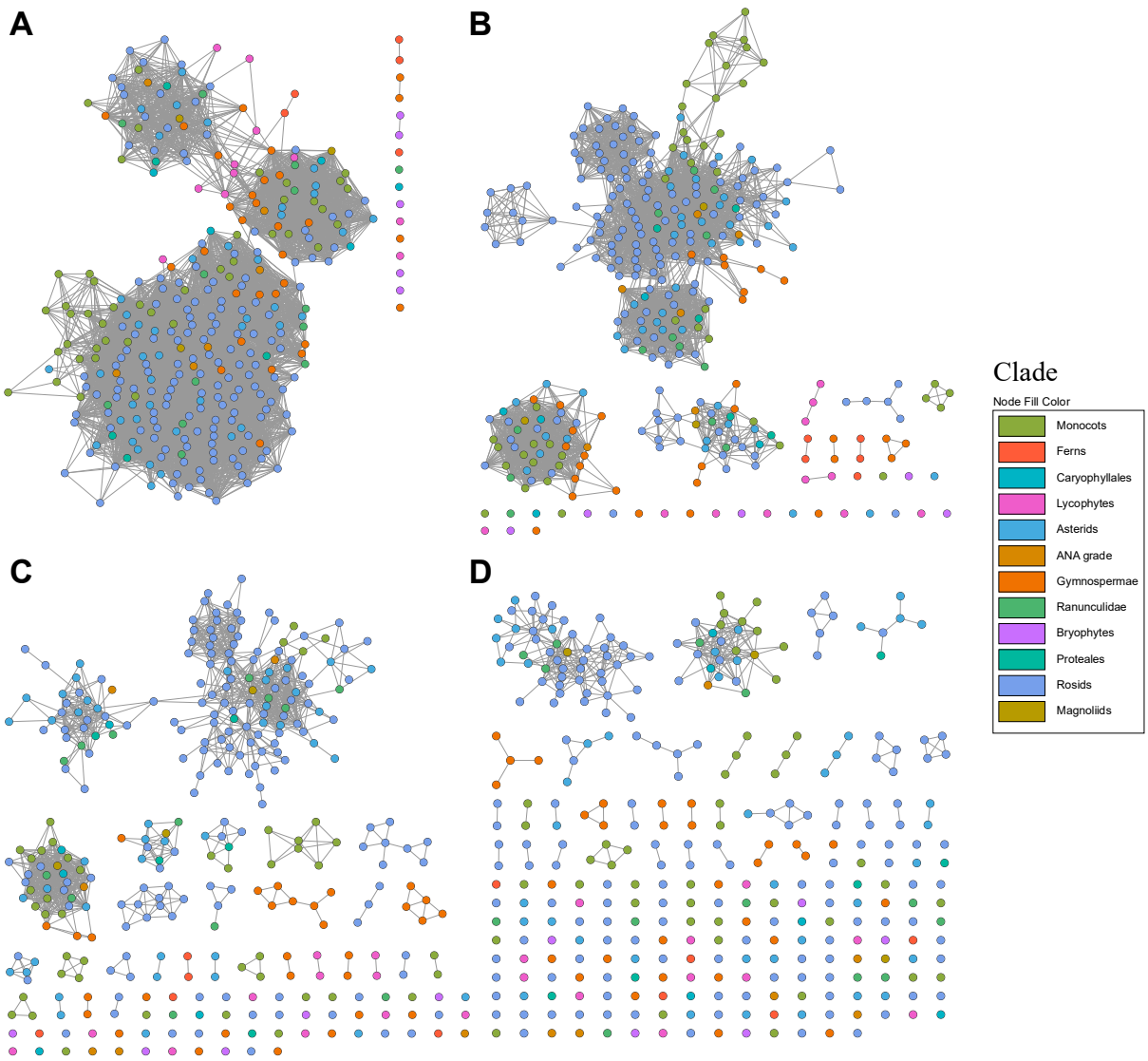



**Figure 3.6:** Sequence similarity networks (SSNs) constructed from the curated class III peroxidase sequence set with nodes representing sequences and coloured according to the clade of the organism of the sequence, and edges representing A) 60% pairwise identity, B) 65% pairwise identity C) 70% pairwise identity and D) 75% pairwise identity. Figure made using Cytoscape.

### *3.3.2 Multiple sequence alignment*

Before phylogenetic and ancestral sequence reconstruction analysis can be carried out, the homologous positions of the sequences need to be identified using multiple sequence alignment (MSA). It has previously been found that for a given a sequence set, the MSA has a large impact on the accuracy of the subsequent phylogeny inference[46,57,68]. As different alignment software uses different algorithms and scoring matrices, each software will predict different homologous positions. Therefore, it is important to use multiple alignment

algorithms and attempt to find the 'best' for a certain sequence set, which hopefully will be the most accurate. To this end, several MSAs were conducted on the curated sequence set using Clustal Omega[54], Muscle[55], DialignTX[690], MAFFT G-INS-I[64], T-COFFEE[42], ProbCons[63], MAFFT DASH[67], PROMALS3D[66], and BAli-Phy[691] respectively. These software cover each of the main algorithm types of MSA (see Chapter 1.2.2). Furthermore, as previous studies have found that combining MSA results from different software can sometimes lead to the most accurate result[68], the MSAs from all aforementioned programs, except BAli-Phy, were combined into a single MSA using M-COFFEE[69].

The MSAs were first visually inspected to validate if they had aligned highly conserved residues of the class III peroxidase family. It was observed that the BAli-Phy MSA failed to align conserved peroxidase residues for all sequences and appeared to contain many large sections of fragmented 'insertions' (Suppl. Fig. 3.2). This poor handling of insertions/deletions (indels) was unexpected given that in contrast to the other MSA software tested, the evolutionary aware BAli-Phy software jointly infers the MSA and phylogeny, and explicitly models insertions/deletions (indels). Nevertheless, other programs, namely Clustal Omega and DialignTX, also failed to align the conserved peroxidase residues for all sequences (Suppl. Fig. 3.3&3.4).

Following visual inspection, all MSAs, except BAli-Phy, were then tested to validate whether they might fulfil certain assumptions made during phylogenetic inference known as stationarity and homogeneity. Currently, most phylogeny building software assumes that during the protein evolutionary process amino acid frequencies have remained constant over time (stationarity assumption) and that substitution rates have remained constant over time (homogeneity assumption), amongst other assumptions[1,186]. In reality, neither the stationarity or homogeneity assumption hold true, as amino acid frequencies are species-dependent and tend to change over time, and substitution rates are known to differ across different parts of the genome and sometimes even within a single gene[25,46]. Despite this, for some MSAs the assumptions of homogeneity and stationarity may nevertheless be approximated, which can be tested for using an initial symmetry test[1]. It should be highlighted that this symmetry test is only an initial check, because even if the MSA is predicted to fail these assumptions, then a more complex evolutionary model can be used during phylogeny reconstruction involving rate parameters, which can implicitly account for violations in these assumptions[25,167]. It should be noted that the symmetry test cannot be used with evolution aware software, like

BAli-Phy, because in order to model indels it can no longer assume site independence, thus breaking the assumption of homogeneity[691].

For each aforementioned MSA except BAli-Phy, the symmetry test was used as an initial investigation for their suitability for phylogeny inference (Table 3.1). Most MSAs passed the assumption of homogeneity, except for the Clustal Omega and DialignTX MSAs where it could not be calculated by the test. The fact that most MSAs passed the homogeneity test may suggest that different sections of the peroxidase genes changed at a relatively similar rate over time. In contrast to the homogeneity test, the stationarity assumption test was only passed by the Muscle and ProbCons MSAs respectively. This may suggest that for the class III peroxidase sequence set, amino acid frequencies may have changed over time, or that many of the models used for the construction of the MSAs are perhaps not well suited to plant amino acid substitution frequencies. Additionally, the failure of the stationarity test by many MSAs could also have been caused by the relatively large indels present in the peroxidase sequences, as these can lead to relatively low symmetry and consequently cause difficulties both in MSA and phylogenetic inference. This is likely to be because most current MSA and phylogeny software does not explicitly model indels. However, as previously mentioned, the failure of the stationarity assumption can instead be implicitly modelled by using rate parameters during phylogenetic inference.

Finally, to achieve a more unbiased and repeatable selection process of which MSA to choose, the MSAs were scored using the transitive consistency score (TCS)[2]. Interestingly, the MSAs with the lowest TCSs, by a margin of ~300, were the same MSAs that appeared inaccurate by visual inspection, thereby supporting the use of the TCS (Table 3.1). Generally, the TCS showed that progressive MSA algorithms performed poorly for this sequence set, supporting other studies which have also found that progressive algorithms can be relatively inaccurate[35,68]. The progressive algorithm is relatively computationally inexpensive, although this study supports the growing body of literature that this speed comes at the cost of alignment accuracy. In contrast, several studies have supported the use of evolution aware alignment algorithms as being the most accurate[35,68], whereas in this study BAli-Phy was unable to align highly conserved residues for all sequences and had the lowest TCS. This may have been due to the peroxidase sequence set being too complex for BAli-Phy to converge on an accurate solution[40]. Therefore, perhaps other evolution aware software such as SATé-II[73] or PRANK[72] could be investigated in future.

| Program | Basic algorithm type | Marginal symmetry P value (stationarity) | Internal symmetry P value (homogeneity) | Maximum test of symmetry P value | Transitive consistency score |
|---|---|---|---|---|---|
| **ClustalΩ** | Progressive | Could not calculate | Could not calculate | 0.30 | 486 |
| **Muscle** | Progressive | 0.21 | 0.32 | 0.24 | 896 |
| **DialignTX** | Consistency-based | Could not calculate | Could not calculate | 0.40 | 617 |
| **MAFFT G-INS-I** | Consistency-based | 0.00004 | 0.70 | 0.04 | 916 |
| **T-COFFEE** | Consistency-based | 0.0003 | 0.90 | 0.17 | 900 |
| **ProbCons** | Consistency-based | 0.70 | 0.50 | 0.56 | 913 |
| **MAFFT DASH** | Structure-based | 0.00004 | 0.10 | 0.05 | 915 |
| **PROMALS3D** | Structure-based | 0.0007 | 0.90 | 0.22 | 905 |
| **BAli-Phy** | Evolution aware | NA | NA | NA | 465 |
| **All alignments except BAli-Phy combined through T-COFFEE** | Mixture, consistency-based | 0.0002 | 0.90 | 0.17 | 321 |

**Table 3.1:** Comparison of multiple sequence alignment algorithms used by their symmetry[1] and transitive consistency score[2]. For the symmetry test, if P <0.05 then the assumption is rejected. For the transitive consistency score, the higher the score, the more consistent the MSA is calculated to be.

As with evolution-aware algorithms, previous studies have also generally supported the use of combined algorithm MSAs[69]. For example, a study by Aadland *et al.*[68] found they were 1.2-fold more accurate than structure-based algorithms for the datasets tested. Theoretically, combined MSAs should reduce the rate of error when predicting indels and be more accurate due to using a 'majority rules' algorithm. However, for the class III peroxidase sequence set, the opposite was found, with the combined MSA performing relatively poorly as assessed by TCS. Visual inspection showed that the combined MSA generally seemed to align standard sections well, however, did not appear to handle indels well, with unusually some large sections in the middle and end of the alignment not even containing any residues in any of the columns (Suppl. Fig. 3.5). The use of the combined MSA for the class III peroxidase sequence set may have performed poorly due to the observed large differences in indel treatment by the individual MSAs used to form the combined MSA.

The three MSAs with the best TCSs by relatively close margins were the MAFFT G-INS-I, MAFFT DASH and ProbCons. These represent software using consistency and structure-based algorithms respectively. Generally, consistency-based algorithms have usually only

been found to be more accurate than purely progressive algorithms[68]; but for this particular sequence set, consistency-based algorithms seemed to perform well. However, it should be noted that the 'good' performance of the consistency-based algorithms could be inaccurate, as the TCS essentially scores based on maximal consistency, which is the same principle used to construct the consistency-based MSAs in the first place[2,57]. On the other hand, structure-based MSA algorithms have consistently been found to be relatively accurate for families with determined structures[41,66,68], hence, making it less surprising that they perform well for the class III peroxidase sequence set.

As the ProbCons MSA possibly performed well and had also been predicted to pass both the stationarity and homogeneity assumptions, it was selected for phylogeny inference. Additionally, because it has been previously found that structure-based algorithms tend to be more accurate than consistency-based algorithms, the MAFFT DASH MSA was also selected for phylogeny construction in parallel to the ProbCons MSA (Suppl. Fig. 3.6&3.7).

### *3.3.3 Phylogeny reconstruction*

Iterative phylogeny reconstruction of the class III peroxidase sequence set was used to predict the evolutionary relationship between the sequences and the positions of the ancestral nodes. Phylogenetic reconstruction was conducted as described in Methods (section 3.5.3) for both the ProbCons and MAFFT DASH MSAs respectively using the maximum likelihood (ML) method through IQ-TREE 2[143]. Briefly, certain sequences were removed from the sequence set if they were assessed as leading to long branch attraction, a common issue with phylogenetic inference, which can result from paralogous or xenologous sequences[25,168], or from the evolutionary model not capturing some sequences well[25,168]. Some sequences were also removed if they had low statistical support and their corresponding species of origin was not closely related to the origin species of nearby protein sequences, as this again indicates some sort of incongruence (Suppl. Table 3.3). For both the ProbCons MSA-based and MAFFT DASH MSA-based phylogeny reconstructions, the majority (59% and 57% respectively) of sequences removed originated from core rosid organisms, which were potentially overrepresented in the sequence set in any case (Suppl. Fig. 3.1). Notably, several sequences were removed in both the ProbCons MSA-based and MAFFT DASH MSA-based phylogenetic reconstructions, suggesting that these sequences may not have had a typical evolutionary history. This could be caused by incongruence in their evolutionary history such as hybridisation events, or due to incongruences in the evolutionary model used which did not capture e.g. certain aspects of amino acid substitution rates[692]. Of particular interest, *Vitis*

*vinifera* peroxidases were removed in both the ProbCons MSA and MAFFT DASH MSA-based phylogenetic processes, when previously the One Thousand Plant Transcriptomes Initiative phylogenomic project classed *V. vinifera* as a 'rogue' taxon during their phylogenomic analysis[682]. Thus, perhaps the genome of *V. vinifera* may have been especially affected by human selective breeding to the extent that it is difficult to model its selected evolution over time.

Following the sequence trimming of both the ProbCons MSA-based and MAFFT DASH-based phylogenetic sets, a total of 10 phylogenies were constructed for each MSA due to the heuristic nature of the ML algorithms[692]. Interestingly, the best predicted model for each of the 10 runs of both the MAFFT DASH-based and ProbCons-based phylogeny reconstructions always contained between eight to ten FreeRate categories[100,103,693] (Suppl. Table 3.3). This high number of rate categories could be due to certain assumptions in the base evolutionary model used not being accurate for the sequence set, as predicted earlier by the symmetry tests of the respective MSAs. Instead, a high number of rate categories can be used to model any heterogeneous aspects of the data violating the assumptions.

Finally, from the sets of ProbCons MSA-based and MAFFT DASH MSA-based phylogenies respectively, a single phylogeny needed to be selected as the 'best' to go forward with. It was decided to select the 'best' phylogeny by validating which phylogeny was the most 'central', or closest to the most other phylogenies by the most distance metrics (Fig. 3.7, Suppl. Table 3.4& 3.5). This method of selection was used because the most 'central' phylogeny represents the solution which the algorithm has converged upon most frequently, hence it was theorised to be the most accurate. It could also be thought of as the 'consensus' phylogeny. This method was used over the more conventional method of using statistical confidence tests, like bootstrapping and the approximate likelihood ratio test, as the confidence test results for each set of 10 phylogenies were very similar and did not clearly indicate which phylogeny best fit each respective data set (Suppl. Fig. 3.8&3.9). Similarly, comparing the log likelihoods of each final set of 10 phylogenies did not clearly indicate which phylogeny was the 'best', especially given the rather large standard errors (Suppl. Fig. 3.10&3.11).

Instead, the distance metrics between each respective set of 10 phylogenies were calculated[694]. It was found that the distances between the MAFFT DASH-based final phylogenies were relatively close by several distance metrics (Suppl. Fig. 3.12A; Suppl. Table 3.4), suggesting that reconstructed phylogenies were fairly similar. In contrast, the

distances between the final ProbCons-based phylogenies tended to be somewhat greater (Suppl. Fig. 3.12B; Suppl. Table 3.5), suggesting that the phylogenetic reconstruction was more influenced by the stochastic nature of the Monte Carlo algorithm, and thus perhaps that the phylogeny 'landscape' was less smooth. Given that several studies have found that model selection has relatively little impact on topology[105,106], the differences observed in the relative distances between the final set of phylogenies of the MAFFT DASH MSA-based process *versus* the ProbCons MSA-based process highlight the significant impact of the MSA on phylogeny reconstruction.

Ultimately, the 7$^{th}$ phylogeny of the MAFFT DASH MSA-based phylogeny process and the 6$^{th}$ phylogeny of the ProbCons MSA-based phylogeny process were selected as the 'consensus' phylogenies, and thus potentially the most accurate, to move forward with for further analysis and ancestral sequence reconstruction (Fig. 3.7A&B). However, before ASR, the statistical measurements of ultrafast bootstrap (UFBoot)[155] and SH-approximate likelihood ratio test (SH-aLRT)[142] were examined to determine how robust the respective MAFFT DASH-based and ProbCons-based phylogenies were in terms of their node positions (Fig. 3.7C&D, Suppl. Table 3.6). Both the MAFFT DASH-based and ProbCons-based final selected phylogenies had a similar average UFboot of ~91 and a similar average SH-aLRT value of ~82 respectively. These represent reasonable confidence in most nodes; for high confidence in node positions a UFboot value greater than or equal to 95 and an SH-aLRT value equal to or greater than 80 being recommended[155]. This is because the SH-aLRT test has generally been found to be a more conservative measure of accuracy than the UFBoot test. For both the MAFFT DASH-based and ProbCons-based final selected phylogenies respectively, the majority (74%) of nodes had very high confidence based on their SH-aLRT values. This suggests that most nodes are robust given the dataset. However, for both respective phylogenies the UFBoot confidence was somewhat lower, with only around two-thirds of nodes being of very high confidence (>95). Nevertheless, for both cases the majority (>80%) of nodes were of reasonable confidence (>90%) by UFBoot, mostly aligning with the SH-aLRT test that most of the topology is likely to be robust. Comparing the confidence of the MAFFT DASH-based phylogeny to the ProbCons-based phylogeny by UFBoot and SH-aLRT suggests that perhaps a few more of the node positions in the MAFFT DASH-based phylogeny are in robust positions. However, as the ProbCons-based phylogeny has more taxa, and thus nodes, than the MAFFT DASH-based phylogeny, this could have contributed to a slightly larger number of nodes being of lower confidence. Interestingly, the MAFFT DASH-

based phylogeny and ProbCons-based phylogeny are relatively different despite mostly containing common taxa, with a weighted Robinsons-Foulds distance between them of 33 (Suppl. Table 3.7). Therefore, this study further highlights the importance and impact of the MSA process on the phylogeny reconstruction. As both phylogenies are relatively distinct but have relatively similar likelihoods, it was decided to conduct ancestral sequence reconstruction using both phylogenies in parallel respectively.

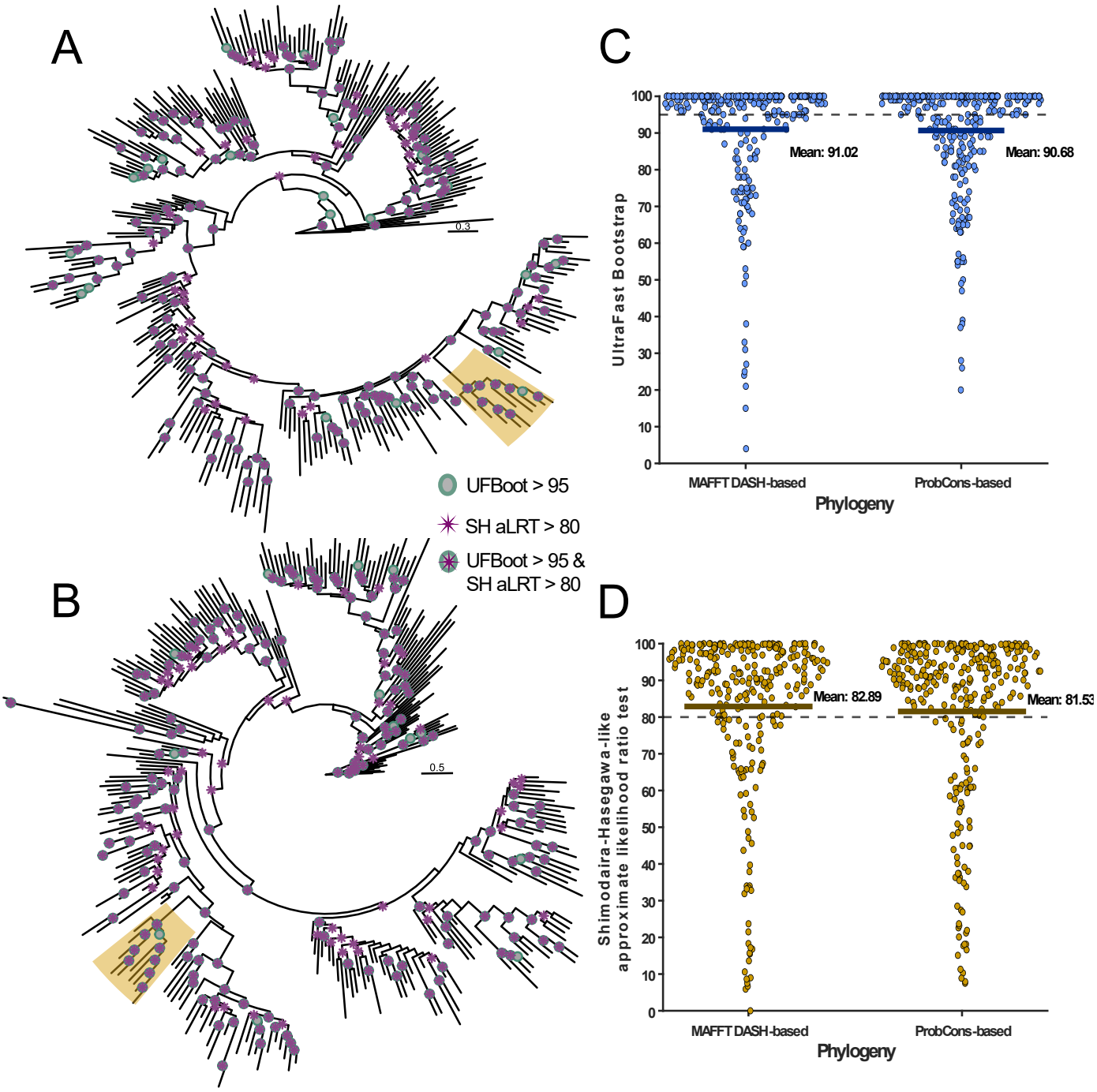



**Figure 3.7:** Phylogenies use *Marchantia polymorpha* peroxidase (MpPrx179) sequence as an outgroup and highlighted in gold is the local clade of A. rusticana HRP C1A (AruPrx01 to ThasPrx69). Scalebar shows number of substitutions per site. A) Final selected phylogeny based on the MAFFT DASH MSA. B) Final selected phylogeny based on the ProbCons MSA. C) Distribution of UltraFast Bootstrap values for each node of each phylogeny respectively. Dashed line shows desired UFBoot of >95. D) Distribution of SH-like approximate likelihood values for each node of each phylogeny respectively. Dashed line shows desired SH-aLRT of >80. Figure made using ggtree and MATLAB.

### *3.3.4 Ancestral sequence reconstruction*

Ancestral sequence reconstruction was conducted on both the final ProbCons and final MAFFT DASH MSA-based phylogenies respectively using the CodeML package of PAML[179], as described in Methods (section 3.5.4). It should be noted that positions resulting from indels of extant sequences ('gaps') were removed from predicted ancestral sequences using a parsimony method. In future, perhaps indel modelling of this ASR of the class III peroxidases could be improved by using software such as GRASP[181], which explicitly models indels in the ancestral sequences in an evolutionary aware manner.

By characterising the average posterior probability of each maximum *a posteriori* reconstructed ancestral sequence, it can be evaluated statistically how 'likely' the reconstruction was for each respective sequence. Both the MAFFT DASH-based and ProbCons-based reconstructions had very high mean average posterior probabilities of ~0.95 (Fig. 3.8), suggesting that the reconstructions have high confidence and thus are robust. The distribution for both reconstructions also showed that most (95%) sequences had an average posterior probability >0.9, indicating that most sequences have high confidence. This high mean average posterior probability and narrow distribution of average posterior probabilities is better than that of the LacI family (Chapter 2.6.1) and towards the higher end of posterior probabilities reported in literature[17,87,695,696]. This high confidence in reconstruction by posterior probability for this reconstruction of class III peroxidase ancestors may have resulted from the relatively highly conserved nature of sequences in this family[17]. This conservation of key catalytic and heme binding residues was reflected in the reconstruction of the ancestral sequences, which should make it likely that the ancestral proteins will have the desired peroxidase function (Suppl. Fig. 3.13). However, it should be emphasised that the accuracy of a given reconstruction cannot be determined *a priori*[150], and that the posterior probabilities take the evolutionary model to be completely accurate, which it never is.

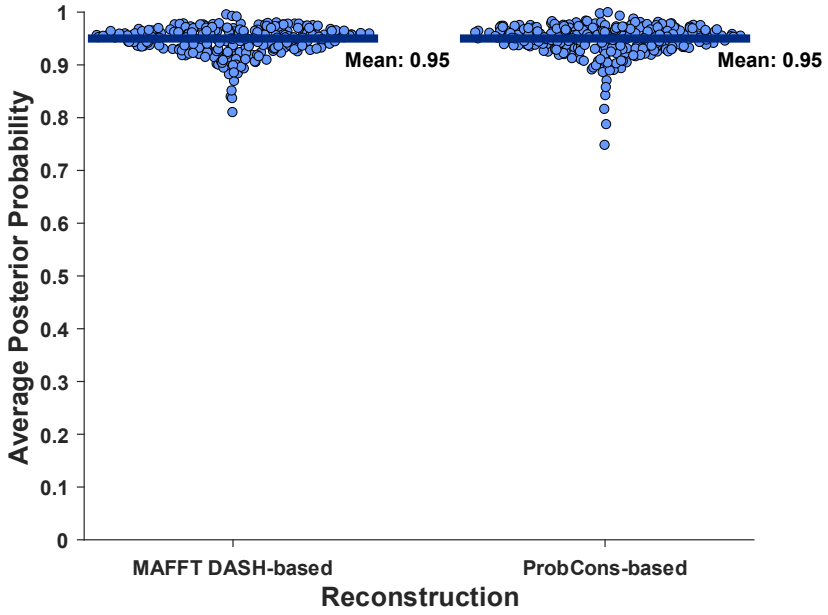



**Figure 3.8:** Distribution of the average posterior probabilities for each node of each ancestral sequence reconstruction. Figure made using MATLAB.

Overall, it was decided that the MAFFT DASH-based phylogeny and reconstruction were slightly more statistically robust, so these were selected as the final phylogeny and reconstruction. It is interesting that the structure-based alignment algorithm proved to be the most statistically robust for both this ASR and the previous ASR of the LacI family[479]. This further supports previous studies that have found structural-based MSA algorithms to generally perform well if structures of the given protein family have been determined[68].

### *3.3.5 Selection of ancestral sequences for experimental characterisation*

Following the ancestral sequence reconstruction, certain ancestors were selected for experimental characterisation to determine whether the process had been able to 'engineer' peroxidases which express in *E. coli* as desired. Due to time constraints, it was not feasible to express and characterise all 282 predicted ancestral sequences. As HRP is the protein of interest, it was decided to select ancestors along its lineage. The HRP lineage ancestral sequences were then assessed based on their average reconstruction posterior probability, as well as whether their corresponding node in the phylogeny possessed high UF-boot and SH-aLRT supports (Suppl. Table 3.8). The predicted most recent common ancestor (MRCA) for this phylogeny was also selected for further investigation, though its node does not contain any statistical support values as the phylogeny is unrooted. Ultimately, 14 ancestors along the HRP lineage were selected for expression trials in *E. coli,* from the MRCA of all the class III

peroxidases to the MRCA of HRP (Fig. 3.9). These were named Anc1 to Anc14 according to their relative age.

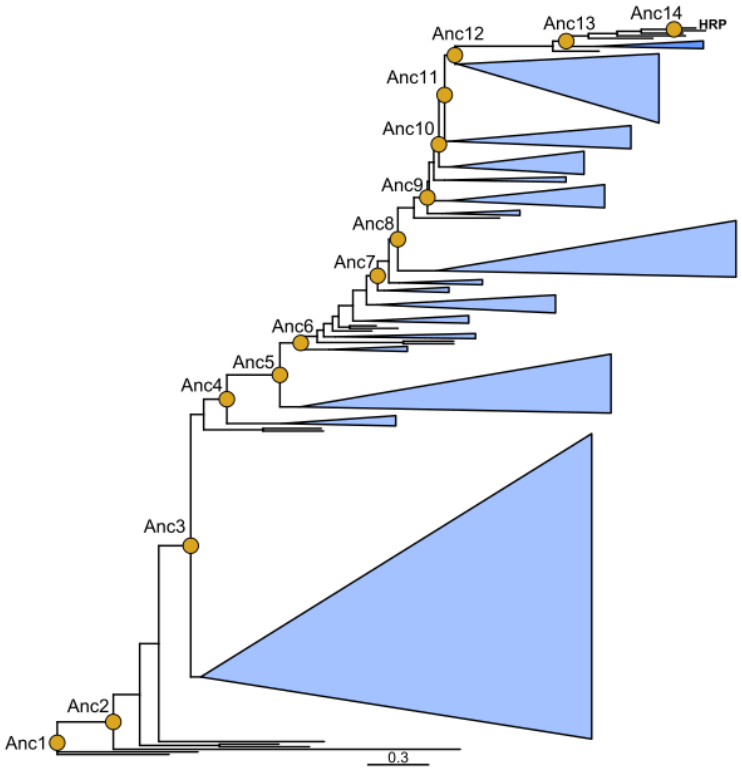



**Figure 3.9:** Collapsed phylogeny (MAFFT DASH-based) of the class III peroxidase family showing the ancestors along the *Armoracia rusticana* HRP lineage selected for further characterisation. Scalebar represents the number of substitutions per site. Figure made using ggtree.

### *3.3.6 Ancestral peroxidase E. coli expression test*

Following the ancestral sequence reconstruction of the class III peroxidases, it was then validated whether the selected ancestral peroxidases could be expressed recombinantly in *E. coli* in a soluble, functional form as desired. Gene design, transformation and expression conditions were conducted as described in Methods (section 3.5.5). Two *E. coli* strains were tested: SHuffle T7 Express[653] and BL21(DE3)pLysS[410,411]. The SHuffle T7 Express strain was tested as it facilitates the formation of disulfide bonds, which the ancestral proteins are likely to possess given that their respective sequences all contain cysteines in homologous positions to disulfide-forming cysteines in many extant proteins (Suppl. Fig. 3.16). The BL21(DE3)pLysS strain was tested as it can aid with expression of potentially toxic proteins, which the peroxidase ancestors may be due to the ability of peroxidases to catalyse free radical formation. Peroxidase ancestor expression was attempted at a temperature of 30 °C

for 24 hrs or at a lower temperature of 18 °C for 48 hrs because these conditions are common in literature[652]. First, it was investigated whether there was soluble overexpression of any of the ancestors under the various expression conditions *via* SDS-PAGE (Suppl. Fig. 3.17). Unfortunately, no obvious soluble overexpression was observed and, for some of the ancestral peroxidases, only insoluble expression was observed (Suppl. Fig. 3.18). Nevertheless, despite the lack of visible soluble overexpression of any of the selected ancestral peroxidases, use of a colorimetric peroxidase assay revealed that each of the selected ancestors had expressed solubly to some amount and were functional to some extent (Fig. 3.10). In previous studies, *A. rusticana* HRP C1A (*Aru*HRP) expressed in *E. coli* typically needed to be refolded or activated before activity was detected[644,659]; hence, the presence of active peroxidase in the cell lysate suggests the ancestral proteins have improved soluble expression in *E. coli* over HRP, as desired. Generally, the peroxidase activity was greater for each ancestral protein when expressed in SHuffle T7 Express as opposed to BL21(DE3)pLysS, highlighting the importance of disulfide bond formation for the class III peroxidase family.

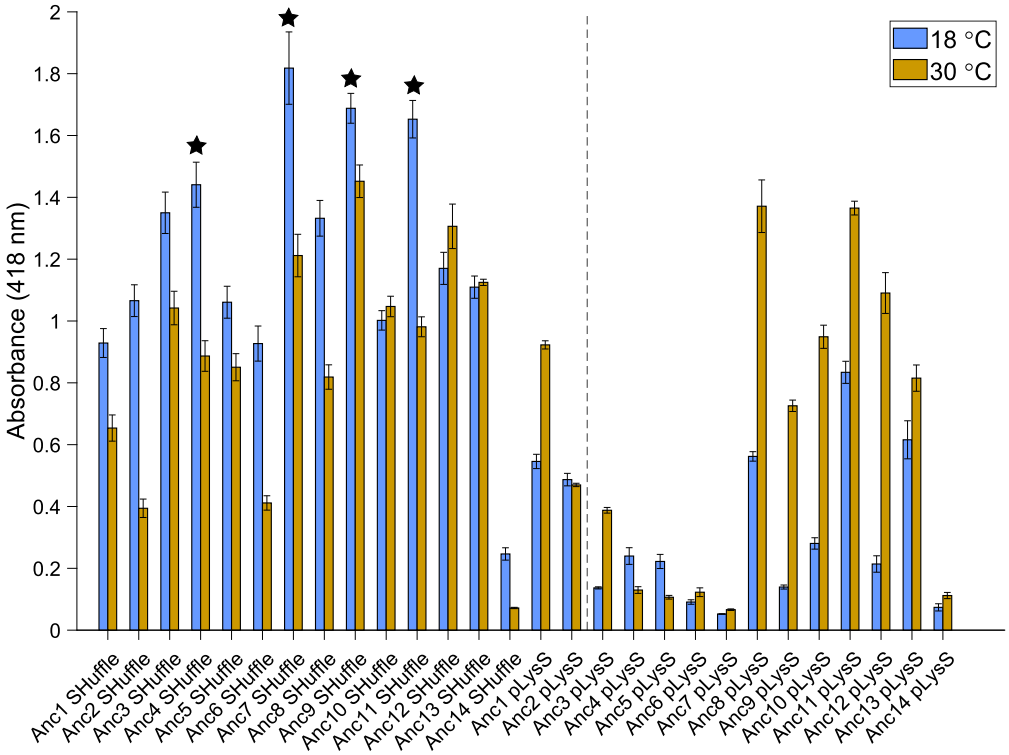



**Figure 3.10:** Relative peroxidase activity by ABTS assay representing active expression of ancestral proteins in the cytoplasm of *E. coli*. Expression conditions tested were *E. coli* strain (SHuffle T7 Express or BL21(DE3)pLysS) and incubation temperature post induction (18 °C or 30 °C). Cultures incubated at 18 °C were incubated for 48 hrs and cultures incubated at

30 °C were incubated for 24 hrs. The ancestors with the best four activity levels are shown by a star marker. Error bars represent S.E.M. of 3 replicates. Figure made using MATLAB.

Based on the ASR of the LacI family[593] (Chapter 2.3.1) and some previous ASR studies[221], it was hypothesised for this ASR that the older peroxidase ancestors might be the most thermostable, and thus most likely to express well in *E. coli*[26,668,669]. Certainly, the MRCA of HRP, Anc14, showed very low peroxidase activity and thus likely had low expression (Fig. 3.10). However, unexpectedly the ancestors with the best peroxidase activity, and thus likely expression of active protein, tended to be ancestors of middling age rather than older age. This supports previous literature which has found that protein stability is less correlated with ancestor age than other factors, such as specificity of function[227,591]. For the selected class III peroxidase ancestors, it is difficult to determine why some ancestral proteins had greater active expression than others from only their sequences. For instance, the ancestors with the highest activity contained a relatively high number of predicted glycosylation sites even though glycosylation is important to the activity of HRP[697]. In future it would be informative to experimentally determine the structure of the HRP ancestral proteins to better understand the sequence-structure relationship and its role in the variability of recombinant expression of active protein. Due to time constraints, the four ancestors with the highest level of peroxidase activity were selected for further characterisation, which were Anc4, Anc7, Anc9 and Anc11.

### *3.3.7 Characterisation of recombinant Anc4, Anc7, Anc9 and Anc11 expressed in E. coli*

Following confirmation that the ancestral reconstructed proteins could be expressed solubly in *E. coli* with some level of peroxidase activity, the four ancestors with the best peroxidase activity were selected for further characterisation. This consisted of investigation of their structural and catalytic properties in more depth in comparison to the extant *Armoracia rusticana* (horseradish) peroxidase (*Aru*HRP).

As protein structure can reveal key insights regarding function, the 3D structure of the four selected ancestors were predicted using AlphaFold2[494] *via* Colabfold[495]. Given that AlphaFold2 was trained using the protein data bank and that several structures of class III peroxidases have been deposited to the data bank[496,497,629,634], it was expected that AlphaFold2 would yield fairly accurate predictions.[698,699] The predicted structures (Fig. 3.11) showed that the ancestral proteins folded into the conserved class III peroxidase structure, with RMSDs of

less than 0.7 Å between each of the ancestral structures and the experimentally determined structure of *A. rusticana* HRP C1A[314] respectively (Fig. 3.11F, Suppl. Table 3.9). The predicted ancestral structures also provided a rationale for their peroxidase activity, because key residues highly conserved in the class III peroxidase family for catalytic function, heme binding, calcium binding and disulfide bonds were predicted to be in equivalent positions in HRP C1A as in the selected ancestral peroxidases (Fig. 3.11). Furthermore, the predicted protein surfaces of the selected ancestors revealed conserved active site pockets for the iron-heme catalytic centre (Suppl. Fig. 3.19). However, the predicted ancestral structures do each show the presence of an additional N-terminal α-helix compared to HRP C1A. This could represent a signal peptide, which class III peroxidases typically have to mark them for secretion, but which are sometimes cleaved during post-processing[680]. The presence of a signal peptide in the selected ancestral sequences was tested by using the software SignalP[700], which predicted that each of the ancestors likely possessed a cleavable N-terminal signal peptide (Suppl. Table 3.10). Interestingly, the predicted signal peptide regions differed slightly between each of the selected ancestors, with an overall sequence identity of only 28% between all four regions (Suppl. Fig. 3.20). This reflects the diversity observed in extant signal peptides, which may result from relaxed evolutionary constraints on the signal region compared to the protein region[701]. Nevertheless, all the selected ancestral signal peptide regions were predicted to contain a core hydrophobic region, as well as an N-terminal region and C-terminal region, which is standard for signal peptides of type I[700]. Hence, it is probable that each of the selected ancestors would have been processed through the secretory Golgi pathway, just as the extant members of this family are.

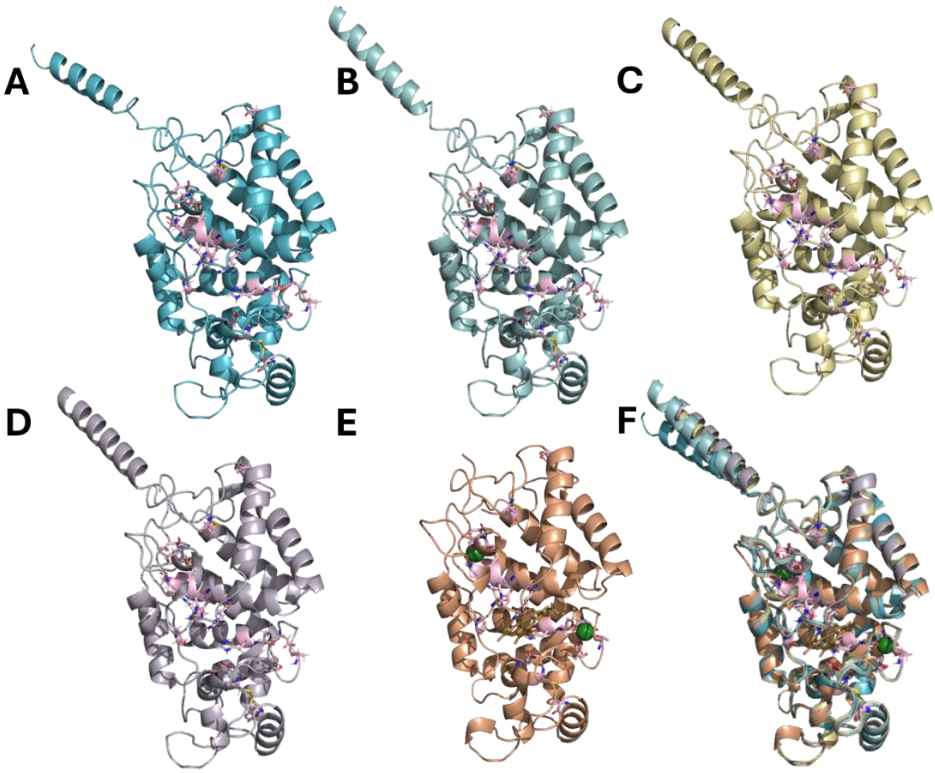



**Figure 3.11:** The predicted 3D structure of A) Anc4, B) Anc7, C) Anc9 and D) Anc11 by AlphaFold 2, with conserved class III peroxidase residues shown in pink. E) shows the structure of *A. rusticana* horseradish peroxidase (HRP) C1A (PDB: 1HCH)[314] with conserved peroxidase residues shown in pink and calcium ions in green. F) contains an overlay of the four predicted ancestral proteins' structures with the structure of HRP C1A. The overlay highlights that the predicted structures of the ancestral proteins are very similar to the structure of HRP C1A, hence, that the selected ancestors may act as peroxidases in a similar manner to the extant HRP C1A. Figure made using PyMOL.

Following confirmation of their predicted typical peroxidase structure, the four selected ancestors were recombinantly purified in *E. coli* (Suppl. Fig. 3.21). It should be noted that it was unlikely that the signal peptide region of each of the ancestral proteins was cleaved in the *E. coli* expression system used here, whereas it likely would have been cleaved in a eukaryotic expression system. Unfortunately, due to time constraints, the recombinant expression of the ancestral proteins in *E. coli* without the signal peptide was not investigated. However, as the predicted signal peptide regions are relatively hydrophobic, it is possible that soluble expression of each respective ancestral peroxidase in *E. coli* could be improved by not including their respective signal regions. Alternatively, swapping the native signal peptide for a prokaryotic peptide to direct the protein for periplasmic expression could also be investigated.

Perhaps partly due to the signal peptide region, the respective expression yields of purified Anc4, Anc7, Anc9 and Anc11 per litre of *E. coli* culture (Table 3.2) were relatively low, with Anc9 having the highest yield of 4 mg $L^{-1}$. However, these represent yields of soluble protein without a fused solubilising partner, which is still a significant improvement over previous attempts to express HRP C1A recombinantly in *E. coli*. Greater yields have only been achieved by either insoluble inclusion body expression[644], which requires laborious refolding, or by the use of large, fused solubilising protein tags, which were not cleaved and may have affected enzymatic activity[659,660,702]. For example, the PGK-HRP fusion[659] only showed 54% of the specific activity of commercial *Aru*HRP, and the DsBA-HRP fusion[702] was found to have a turnover number ~180 times worse than commercial *Aru*HRP. Therefore, although the yields of the selected ancestral peroxidases achieved here are relatively low, they still represent an improvement on the previously reported yields for HRP expressed solubly in *E. coli* without the use of a fused solubilising protein. Furthermore, the recombinant yields of the selected ancestral peroxidases could be improved in future by identifying the optimal culture and expression conditions, using a bioreactor rather than shake flasks and perhaps even by adding a solubilising fusion protein tag that could later be cleaved.

Although some *E. coli* soluble expression of each respective selected ancestral protein was achieved, measurement of their absorption spectra demonstrated that the samples had relatively low Reinheitszahl[703] (Rz) values , suggesting low heme incorporation (Table 3.2, Suppl. Fig. 3.22). This is because the Rz value is a measurement of the absorbance of the Soret peak, which is primarily due to the heme cofactor, divided by the absorbance at ~275 nm, which includes contributions from tryptophan residues. Therefore, the higher the Rz value, the greater the proportion of the sample contains a heme cofactor. As the ancestors have Soret peaks at 403 nm like HRP C1A, it would therefore be expected that for very high heme incorporation an Rz of >3.0 would be measured[655]. Previous studies where HRP has been expressed recombinantly have reported Rz values of 1.8 or greater[660,662,704]. Though the Rz values of the peroxidase ancestors may or may not be directly comparable to those of HRP C1A due to potential differences in the heme environment, the Rz values for the purified selected ancestral proteins were nevertheless comparably very low, being 0.26 or lower. These low Rz values indicated that a relatively large proportion of the respective purified ancestral protein was lacking the heme cofactor. However, it could also be due to the relatively lower purity of these samples; with low yield it is difficult to obtain high purity and proteins will lower the Rz value. Regardless, this suggests that some portion of each of the

ancestral proteins may have been inactive due to the lack of heme cofactor, as it is essential for catalysis. Therefore, in future, heme incorporation needs to be improved, perhaps by the addition of δ-aminolaevulinic acid to the culture medium[603,697] or by the addition of hemin (iron-protoporphyrin IX) to the ancestral protein sample following the initial stage of purification[659,660].

| | **Yield (mg/L culture)** | **Reinheitszahl value** |
|---|---|---|
| *Anc4* | 2.3 | 0.12 |
| *Anc7* | 1.9 | 0.02 |
| *Anc9* | 4.0 | 0.26 |
| *Anc11* | 0.8 | 0.22 |

**Table 3.2:** Yields of *E. coli* recombinantly expressed purified ancestral peroxidase and their Reinheitszahl values. Yields are low but do represent soluble peroxidase. However, the Reinheitszahl values are much lower than desired, indicating low heme content.

Although the selected *E. coli* peroxidase ancestors had relatively low heme cofactor incorporation, it was confirmed that they nevertheless possessed secondary structure, and thus were likely folded, *via* circular dichroism (Fig. 3.12A). However, prediction of the respective secondary structure elements from the spectra showed a surprisingly high β-strand content and low α-helix content for the selected ancestors, given that their predicted tertiary structures are greater than 50% α-helical and less than 2% β-sheet (Fig. 3.12A). Previously, it has been found that secondary structure prediction of circular dichroism data can erroneously overpredict β-strand content when the protein structure contains many turns[488,489], which the selected ancestors likely do based on their predicted tertiary structures. Given that the class III peroxidase tertiary structure is largely conserved[634] and that the selected ancestors were shown to act as peroxidases, it is probable that their respective secondary structure predictions of containing relatively large proportions of β-sheet and unordered sections are not especially accurate. This is further supported by the reported CD spectrum of HRP C1A being similar to those of the selected ancestral peroxidases[705,706]. Furthermore, the secondary structure prediction of the CD spectrum of HRP C1A also underpredicted α-helical content and overpredicted β-sheet content in similar proportions to the predictions of the ancestral peroxidases[707]. Thus, in future, it would be informative to experimentally determine the

tertiary structures of the selected ancestral peroxidases experimentally and to allow for comparison to both the AlphaFold predicted models and the CD spectra.

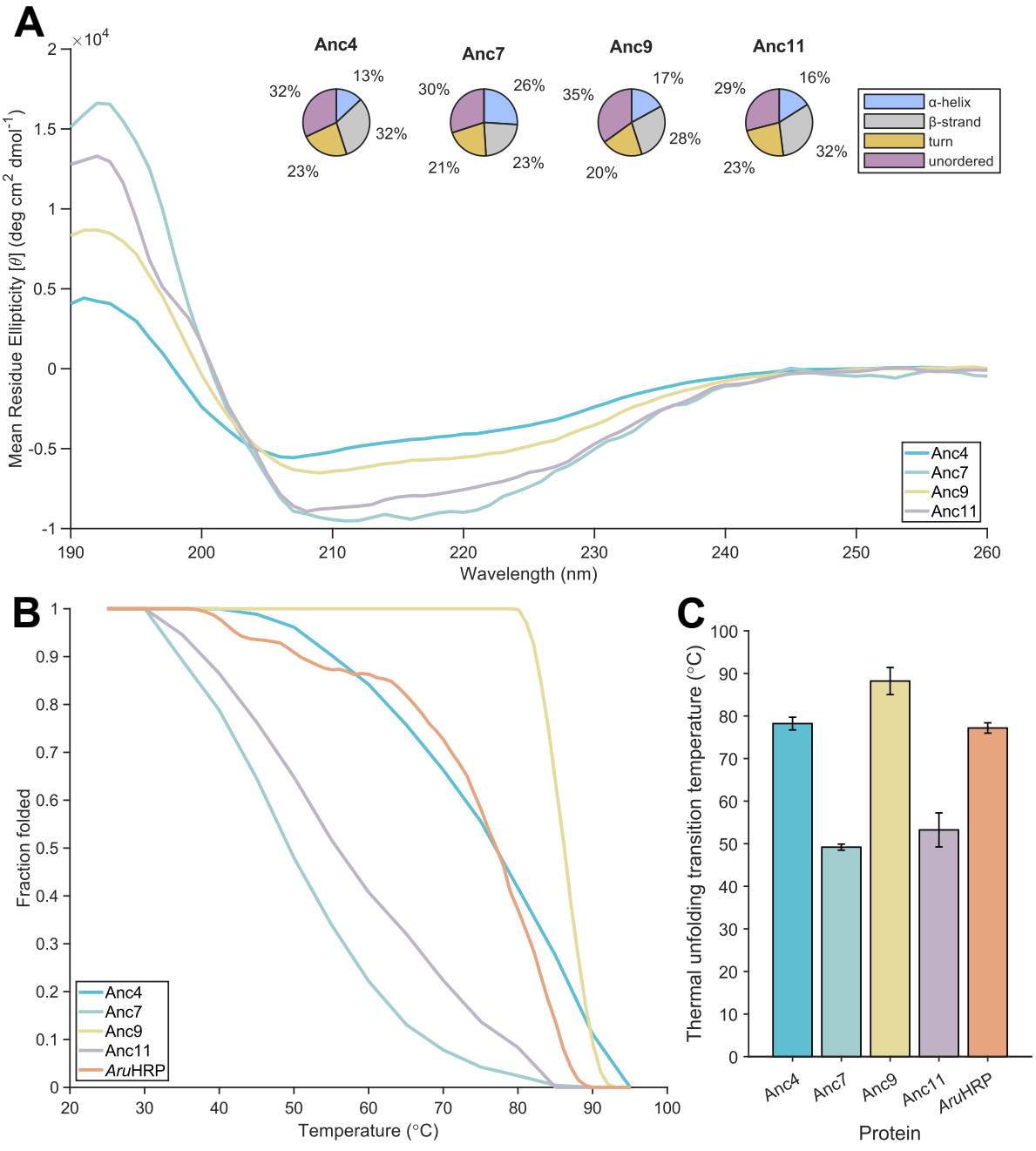



**Figure 3.12:** Characterisation of Anc4, Anc7, Anc9 and Anc11 by circular dichroism. A) Spectra of the selected ancestors at 25 °C (see Materials for concentrations) shows that they contain secondary structure. Prediction of the secondary structure by CDSSTR may have overpredicted β-strand content and underpredicted α-helical content. B) Changes in the fraction of folded protein with temperature of the selected ancestors and a commercial sample of *A. rusticana* horseradish peroxidase (*Aru*HRP). C) Thermal unfolding transition temperatures ($T_m$) of the selected ancestors and a commercial sample of *Aru*HRP measured over several wavelengths then averaged. Error bars represent S.E.M. Anc9, which was expressed recombinantly in *E. coli* without glycosylation, possesses a higher $T_m$ than the glycosylated, plant-expressed *Aru*HRP sample. Figure made using MATLAB.

After characterising the folded ancestral proteins, the thermal unfolding of the selected ancestral peroxidases and a commercial sample of *A. rusticana* HRP (*Aru*HRP) was characterised by measuring the secondary structure thermal unfolding transitions (Fig. 3.12B). All the selected ancestors were found to have a thermal unfolding transition temperature ($T_m$) above 49 °C (Fig. 3.12C), which indicates their potential for use up to this relatively high temperature. Of particular interest was that the determined $T_m$ of Anc9 was 88 °C, indicating that it is very thermostable and hence that it may have extensive utility. In fact, the $T_m$s of both Anc9 and Anc4, which were expressed recombinantly without glycosylation, were greater than that of the commercial sample of *Aru*HRP, which was expressed natively and thus was glycosylated. As native glycosylation commonly leads to increased thermostability[708,709] and the ancestral peroxidases contain glycosylation sites (Suppl. Fig. 3.14), the high $T_m$s of the recombinant Anc9 and Anc4 lacking glycosylation are especially remarkable and highlight the ability of ASR as a tool to identify thermostable proteins. Interestingly, the unfolding transition temperatures of the selected ancestral proteins showed a relatively strong correlation with recombinant expression yield (0.84 Pearson correlation coefficient) (Suppl. Fig. 3.15) , further providing support to previous studies showing that protein thermostability is correlated with expression yield for some proteins[668-670]. However, unlike with the LacI family ASR (Chapter 2.3.1) and some other previous ASR studies[189,710,711], here it was found that the relative thermostability of the selected peroxidase ancestors only very weakly correlated with relative ancestral age (0.27 Pearson correlation coefficient). Although the results are not statistically significant, this perhaps shows that empirical Bayesian or maximum likelihood ASR does not always result in increasingly thermostable ancestral proteins with age due to evolutionary model bias, as has been previously claimed by some[34,68], but instead is able to identify potential ancestral proteins with varying thermostabilities[185,186]. The reason for the pattern of thermostabilities of the selected class III peroxidase ancestors is unclear, as they do not appear to correlate with global climatic trends either (Suppl. Table 3.11). Nevertheless, it should be noted that the thermostabilities of the predicted ancestral proteins were measured without their predicted native glycosylation, and furthermore that as only four selected ancestors were characterised, any correlations are not statistically significant. Therefore, in future it would be interesting to

better characterise the thermostability of more ancestors, both glycosylated and unglycosylated.

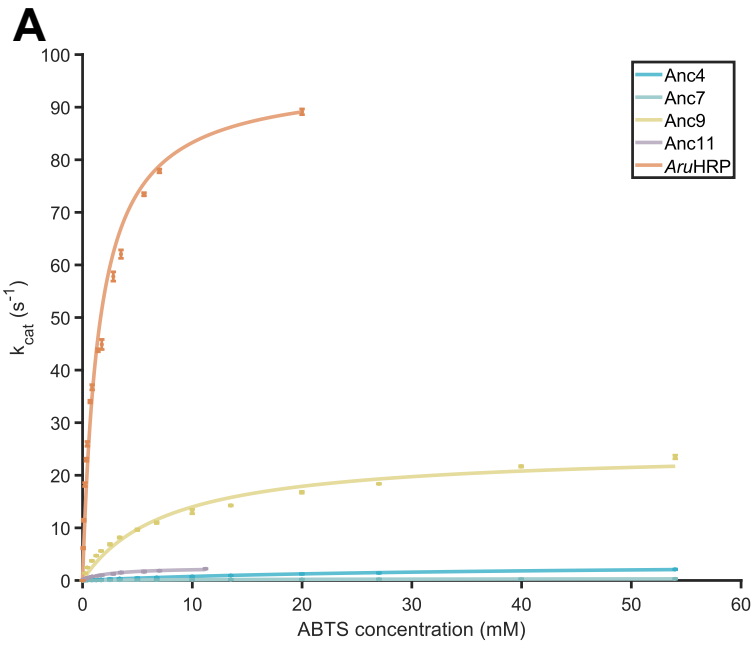
Following structural characterisation, the peroxidase activity of the selected ancestors was characterised to determine if they could act as viable alternative catalysts to *Aru*HRP. To this extent, the steady state kinetics and activity of the selected ancestral proteins and a commercial sample of *Aru*HRP were determined using 2,2'-Azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS); a substrate commonly used in peroxidase assays. First, the approximate kinetic constants under steady state kinetics were determined for each of the selected ancestral enzymes and *Aru*HRP with ABTS. It should be noted that the parameters are approximate as the catalytic cycle of class III peroxidases involves several reaction steps and $H_2O_2$ as an additional substrate. However, to calculate steady state Michaelis-Menten kinetics, the reaction was approximated to two steps and excess $H_2O_2$ was used. Using such approximations, the steady state reaction velocities were determined for the Anc4, Anc7, Anc9, Anc11 and *Aru*HRP with varying concentrations of ABTS (Fig. 3.13A), from which in turn the kinetic parameters were calculated (Fig. 3.13B). It should be noted that as *Aru*HRP is a mixture of isoenzymes, the kinetic constants should be treated as a rough average of the kinetics of the particular sample of horseradish peroxidase isoenzymes. Nonetheless, comparison of the turnover numbers ($k_{cat}$) shows that of the selected ancestors, Anc9 has the greatest turnover number by a relatively large margin. However, the turnover number of the *Aru*HRP sample tested was still about four times larger than that of Anc9. However, the respective concentration of functional ancestral enzyme was likely much lower than the total enzyme concentration used due to low heme cofactor incorporation. Thus, the functional concentrations of each respective ancestral peroxidase were estimated based on their respective Rz values, and an approximate turnover number ($k_{cat}$*) was calculated for each selected ancestral protein. The approximate turnover numbers of Anc4, Anc7 and Anc11 were still lower than that of *Aru*HRP, but certainly much more reasonable for future utility. However, the approximate turnover number of Anc9 is predicted to be even slightly greater than that of *Aru*HRP, suggesting that it could be a good replacement for HRP for future applications. Therefore, turnover numbers of the ancestral enzymes could likely be greatly improved by increasing their heme incorporation.

Unlike the turnover number, the Michaelis constant ($K_M$) is independent of enzyme concentration. The *Aru*HRP sample tested still possessed the lowest $K_M$, and thus perhaps the

best binding to ABTS assuming a slow catalytic step constant, although the $K_M$ of Anc11 is relatively close in value. Interestingly, the Michaelis constant decreased steadily with ancestral age, perhaps indicating a gradual shift in substrate binding preferences over time. This gradual change in substrate preferences was also previously observed for the LacI family of proteins (Chapter 2.3.2). This progressive change in $K_M$ for the selected class III peroxidase ancestors could perhaps result from the changes to the entrance of the active site pocket observed in the predicted ancestral structures (Suppl. Fig. 3.26). The entrance for the active site of HRP C1A is dominated by three phenylalanine residues, which may be able to provide aromatic stabilisation and thus contribute for its preference for relatively small phenolic substrates like ABTS[602]. However, while the entrances to the active sites of Anc7, Anc9 and Anc11 are still relatively hydrophobic, they only contain one phenylalanine residue that is positioned such that its aromaticity is less prominent on the respective entrance surfaces. Therefore, this change may contribute to the respective higher $K_M$ values for ABTS catalysis using these ancestral peroxidases compared to *Aru*HRP. Similarly, the highest $K_M$ of Anc4 for ABTS catalysis perhaps could be due to its active site entrance not possessing any aromatic residues, and thus not being able to form aromatic interactions with substrates like ABTS. Nevertheless, to better establish binding specificity changes over time along the HRP lineage, more experimental data needs to be collected involving at least several potential substrates. Also, docking experiments might further elucidate rationales for different substrate preferences.

Overall, the catalytic efficiencies ($k_{cat}/K_M$) of the ancestral enzymes against ABTS were lower than *Aru*HRP. Anc9 had the best catalytic efficiency of the ancestors, yet it is still about 20 times less than that of the *Aru*HRP sample. Therefore, the relatively low catalytic activities of the selected ancestors further emphasise that in future heme incorporation needs to be

improved, because this will likely increase the turnover number of the ancestral enzymes, thereby making them a viable alternative to *Aru*HRP.




| Enzyme | $k_{cat}$ ($s^{-1}$) | $k_{cat}$* ($s^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ ($s^{-1}$ $mM^{-1}$) |
|---|---|---|---|---|
| *Anc4* | 3.36 ± 0.311 | 32.6 ± 3.06 | 33.8 ± 5.53 | 0.0993 ± 0.0254 |
| *Anc7* | 0.387 ± 0.0148 | 21.1 ± 0.807 | 16.7 ± 1.47 | 0.0232 ± 0.00293 |
| *Anc9* | 24.8 ± 2.34 | 120.8 ± 11.4 | 7.75 ± 2.12 | 32.05 ± 11.8 |
| *Anc11* | 2.51 ± 0.231 | 14.1 ± 1.30 | 2.35 ± 0.555 | 1.07 ± 0.350 |
| *AruHRP* | 95.8 ± 10.7 | N/A | 1.51 ± 0.49 | 63.6 ± 27.9 |

**Figure 3.13:** Characterisation of the enzymatic catalysis of ABTS by the selected ancestral peroxidases and *A. rusticana* HRP (*Aru*HRP). A) Michaelis-Menten curves of the reaction velocities with increasing ABTS concentrations conducted in 50 mM potassium phosphate buffer, pH 6.0 at 25 °C. Error bars represent the S.E.M. of 3 replicates. Note that different concentrations of each enzyme were used as well as different concentration series of ABTS to make steady state kinetics assumptions. B) Michaelis-Menten kinetic parameters for the selected ancestral enzyme and *Aru*HRP. kcat* was calculated using the approximate concentration of functional protein given the percentage of heme incorporation. Error shown for kinetic constants is the 95% confidence interval. Figure made using MATLAB.

Following kinetic characterisation of the ancestral enzymes, their specific activity of ABTS oxidation was investigated to determine how much of each respective ancestor should be utilised to achieve similar activity to commercial *Aru*HRP. Perhaps due to poor heme incorporation, it was found that the specific activities of ancestral peroxidases were much lower than that of commercial *Aru*HRP (Suppl. Fig. 3.23). The ancestor with the best specific activity was Anc9, however, its specific activity suggests that about seven times the amount of Anc9 should be utilised to attain a similar enzymatic activity to *Aru*HRP. Nonetheless, as Anc9 can be produced relatively quickly and easily in *E. coli*, whereas HRP is still mostly produced in the more time-consuming manner from *A. rusticana*, Anc9 could still provide a

viable alternative to HRP in many applications, provided that a greater concentration of Anc9 enzyme is used.

### *3.3.8 Enzyme-mediated radical polymerisation of PNIPAm hydrogel by Anc9*

Due to Anc9 displaying the best specific activity of the selected ancestors for ABTS oxidation, its potential utility as a catalyst for the polymerisation of PNIPAm was investigated. *Aru*HRP is commonly used in enzyme-mediated radical polymerisation (EMRP) reactions, such as that of PNIPAm. However, gel formulations need to be adapted for each *Aru*HRP batch due to inter-batch variation. Thus, it would be beneficial to replace *Aru*HRP in EMRP PNIPAm formulations with a recombinantly expressed peroxidase. To this extent, it was tested whether Anc9 could be employed to catalyse the polymerisation of a PNIPAm hydrogel in a similar manner to *Aru*HRP. It was observed that Anc9 could indeed catalyse the polymerisation of PNIPAm (Fig. 3.14A&B) and that those hydrogels formed by Anc9 catalysis possessed similar thermosensitive properties to those formed by *Aru*HRP (Fig. 3.14C; Suppl. Fig. 3.24). Thermosensitivity was demonstrated by the Anc9-catalysed PNIPAm deswelling by a similar amount to the *Aru*HRP-catalysed PNIPAm when incubated at 60 °C, a temperature well above the lower critical solution temperature of PNIPAm[621]. While the deswelling of PNIPAm gels depends on their exact composition and solvent conditions, that of the EMRP PNIPAm gels presented here even falls within the range reported for similar PNIPAm materials formed using chemical initiated radical polymerisation[712,713].

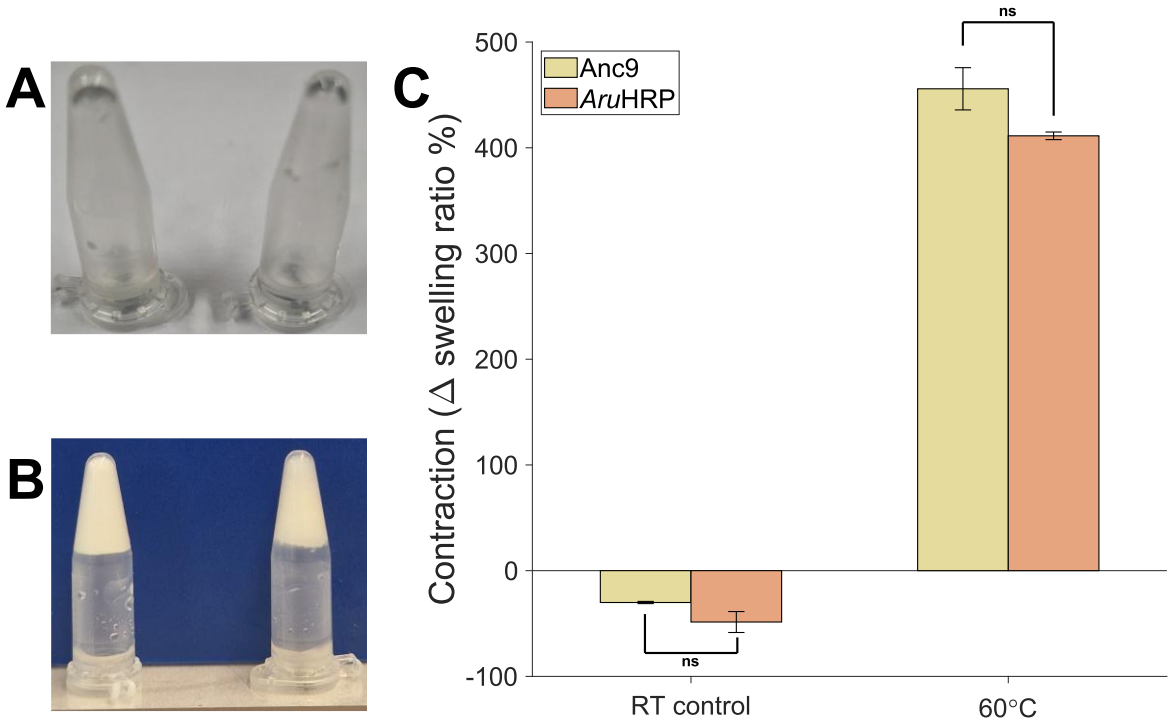



**Figure 3.14:** Enzyme-mediated formation of PNIPAm hydrogel by Anc9 compared to *A. rusticana* peroxidase (*Aru*HRP). A) PNIPAm gel mixture prior to curing containing *Aru*HRP on the left or Anc9 on the right. B) PNIPAm gel post curing containing AruHRP on the left or Anc9 on the right. C) 'Contraction' (deswelling) of Anc9 formed PNIPAm or *Aru*HRP formed PNIPAm after being incubated for 30 min at room temperature (RT) as a control or 60 °C, which is greater than the lower critical solution temperature of PNIPAm. Contraction was measured as the difference between swelling ratio % before and after incubation. Error bars represent the S.E.M. of four replicates and Anc9 gels were compared to *Aru*HRP gels using an unpaired t test. Both the Anc9 formed PNIPAm and *Aru*HRP formed PNIPAm displayed similar lack of contraction when incubated at RT, but relatively high contraction when incubated at 60 °C. Figure made using MATLAB.

Further validation that Anc9 could be used to polymerise PNIPAm formation in a similar manner to *Aru*HRP was conducted by characterising a mechanical property of PNIPAm: its behaviour under a compressive load. Both the Anc9 formed PNIPAm and *Aru*HRP formed PNIPAm displayed similar behaviour under compression, with the resulting engineering stress-strain curves containing two distinct linear regions connected by an inflection region (Fig. 3.15A). Such curves are typical of hydrogels under compression due to their poroelastic and viscoelastic properties[714]. However, these properties lead to the engineering stress-strain model used here, and commonly elsewhere for hydrogels[715], fitting relatively poorly[716,717]. Therefore, in future it would be beneficial to characterise the properties of the PNIPAm

hydrogels using different techniques, such as rheology, allowing the viscoelastic properties of the hydrogels to be better estimated using more complex models such as the Mooney-Rivlin model[718,719]. Nevertheless, the compressive test demonstrated that both PNIPAm gels formed by Anc9 or *Aru*HRP displayed similar compressive behaviour, and furthermore, possessed compressive strengths in the range of those reported in literature for chemically-initiated PNIPAm[713,720]. The elastic nature of both enzyme-formed PNIPAm gels within the compressive range tested was also observed by their recovery after the test (Suppl. Fig. 3.25). The thermosensitive and mechanical behaviours of the enzyme-formed PNIPAm highlights the ability of using enzymes to catalyse its formation over traditional chemical methods. Furthermore, it demonstrates that the *E. coli* recombinantly expressed Anc9 peroxidase could replace *A. rusticana* sourced HRP when catalysing the polymerisation of PNIPAm. Use of recombinant Anc9 should alleviate issues with the batch-to-batch variation of *Aru*HRP, which result in its concentration in the formulation needing to be optimised for each batch.

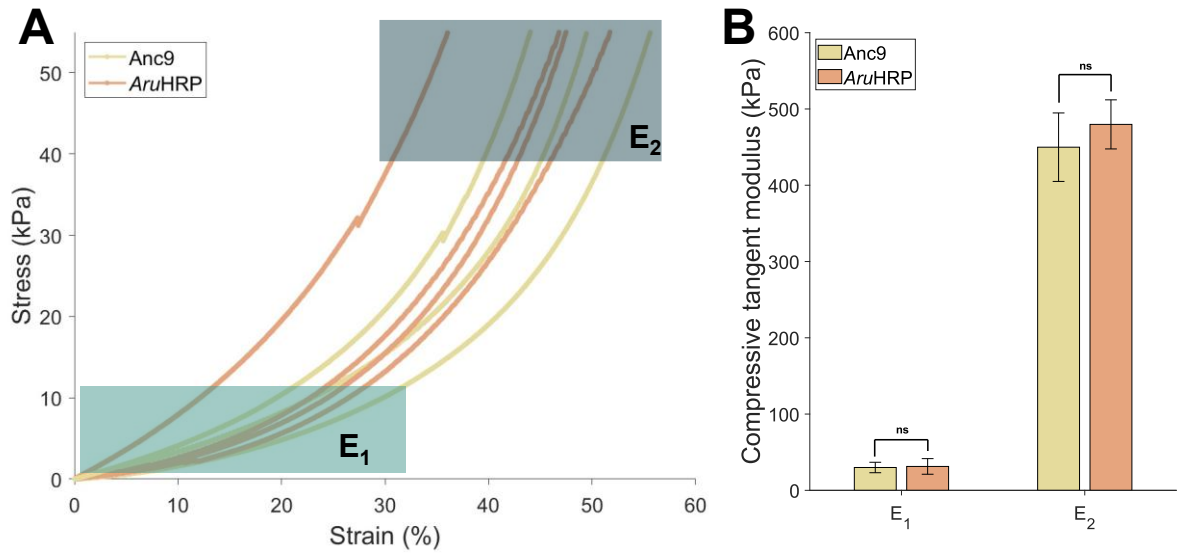



**Figure 3.15:** Compressive behaviour of PNIPAm gels formed by Anc9 or *Aru*HRP. A) Engineering stress-engineering strain curves of Anc9 and *Aru*HRP formed PNIPAm gels during compression at 1 mm/min using a 10 N load. Curves are typical of hydrogels, which do not display a single linear elastic region, but rather two linear regions, which are highlighted. This results in two tangential moduli: $E_1$ and $E_2$. B) Compressive strength using 1 mm/min compressive load as represented by the tangent moduli $E_1$, calculated from the first linear region, and $E_2$, calculated from the second linear region. Error bars represent S.E.M. from 3 replicates for Anc9 formed gels or 4 replicates from *Aru*HRP formed gels. Tangent moduli of Anc9 gels were compared to those of *Aru*HRP gels using an unpaired t test. No significant difference was found in compressive strength between Anc9 formed PNIPAm versus *Aru*HRP formed PNIPAm. Figure made using MATLAB.

## 3.4 Discussion

This study showed that ancestral sequence reconstruction (ASR) can be utilised as a tool to engineer proteins for increased heterologous recombinant expression. While many other ASR studies have demonstrated that predicted ancestral proteins may be more thermostable or have different functionality, here we provided support to the smaller number of studies that specifically use ASR to improve heterologous expression[26]. Despite ASR not having been widely used for proteins of the plant kingdom[34], this study showed that it can reveal a greater understanding of the protein evolutionary trajectory, as well as being useful as an engineering tool to generate more stable, functional ancestral proteins. Given that the vast majority of mutations to a given protein sequence have a deleterious or neutral effect on function[248], this study further highlights the ability of ASR to identify functional 'peaks' of the sequence-function landscape in a rational manner. This can be demonstrated by the fact that the selected ancestral peroxidase proteins were highly divergent from horseradish peroxidase (HRP) (sequence identities from 66% to 57% with HRP), yet each was capable of peroxidase catalysis.

Although plants undergo processes such as whole genome duplication, hybridisation and polyploidy, which can affect phylogenetics and ASR by making it more difficult to determine which genes are homologous[25,34], here it was shown that careful sequence curation throughout the process can improve the quality of the sequence set. Important curation steps include ensuring that there is a medium diversity of sequences, and trimming any sequences that lead to long branch attraction or which are especially out of place when compared to the organismal phylogeny. It should be noted that species and proteins evolve in different ways, thus the organismal phylogeny acts only as a guide for the protein phylogeny. Nevertheless, in future the sequence collection and curation process may improve even further, as more genes are added to databases, and their potential function and family perhaps could be predicted automatically by machine learning. Additionally, this study, like many others[14,186], focuses on single isoforms of genes, but in future the ASR of the class III peroxidases could be repeated using full genes containing multiple enzyme isoforms. As this is how plant proteins are encoded in nature, such a reconstruction process could be more accurate, and resultingly provide further insight into the evolution of the class III peroxidases and more generally eukaryotic protein evolution.

Another issue faced by this particular ASR study of the class III peroxidase family was that some of the sequences appeared to contain relatively large insertions. This may have been the reason why some of the multiple sequence alignment software tested failed to align highly conserved homologous positions, or generally did not perform the alignment well for the indel sections. While previous studies have shown that evolution aware algorithms which explicitly model indels can be the most accurate for some sequence sets[35,68], this study shows that the evolution aware BAli-Phy software is not well suited for the sequence set of this study. This may be due to its relatively large size[40] and also due to the indels being relatively large[71] for some sequences used in this study. Similarly, the relatively large indels in the class III peroxidase sequence set may have led to the relatively inaccurate MSA by the combined MSA algorithm, which has been found to be relatively accurate for other sequence sets[68]. Instead, for this particular sequence set, the MSA which led to the most statistically robust phylogeny and reconstruction used a structure-based algorithm. As structure-based algorithms do not necessarily penalise long indels as much as other algorithms, this may have resulted in it performing the best[36]. However, it should be noted that there are many peroxidase protein structures available which would have greatly aided using a structure guided MSA algorithm in this case, but for protein families with fewer available structures using a structure based MSA may not be as accurate. Ultimately, this study further emphasises the importance of testing and comparing multiple MSAs for a given sequence set because their performance and accuracy can vary depending on the characteristics of the sequence set. Currently in the ASR field, comparing multiple MSAs has not been widely adopted, with some of the recent examples of ASR as of the time of writing only testing a single MSA[80,81,696]. This study highlights how important it is to test different MSAs, especially given their impact on the later phylogenetic and reconstruction stages shown here and by others [35,57,68]. One hopes that in future the importance of the MSA process on phylogenetic and ancestral sequence reconstruction will be more commonly recognised, especially when undue significance is often given to phylogenetic model selection, which has repeatedly been found to be relatively less impactful[105,106].

This study also highlights the difficulties in the phylogeny selection process. While statistical supports such as bootstrap and likelihood tests are the most commonly used, they are not measures of accuracy but rather represent measures of support, with the caveats that their calculations contain several assumptions and that they are no longer as useful if an inaccurate phylogenetic model is used[155,167,168]. Here for the phylogenetic reconstruction of the class III

peroxidase family, the UFBoot and SH aLRT tests were not able to clearly indicate which phylogeny was the most robust. Consequently, a novel, simpler approach to phylogeny selection was investigated by selecting the most 'central' phylogeny in each phylogeny set for the MAFFT DASH-based and ProbCons-based approaches respectively. It was theorized that the most central phylogeny should represent the solution which the MCMC algorithm of the phylogenetic software has converged upon the most often, and thus hopefully be the most accurate. Further ancestral sequence reconstruction and ancestral protein characterisation using the phylogeny selected in this manner suggested that this approach was viable, but further validation is needed through more complex studies to evaluate how accurate this approach truly is compared to the more standard approach using topology supports.

Nevertheless, the ancestral sequence reconstruction using the 'consensus' selected phylogeny resulted in predicted ancestral sequences with very high statistical support by posterior probability for both the MAFFT DASH-based and ProbCons-based processes respectively. The average posterior probability mean for all ML ancestral sequences was greater than that of the LacI family ASR (ch2 section x) and in the higher end of the range reported by other ASR studies[17,87,695,696]. This relatively high statistical support by posterior probability for the reconstruction of the class III peroxidase family may result from the relatively high number of conserved residues in the sequences of this family[634]. This family of peroxidases catalyses relatively complex redox chemistry, hence perhaps due to this complex catalysis, certain residues of this family may be relatively highly conserved because of selection pressure to maintain their specific peroxidase catalysis function. Nevertheless, the ASR process used only parsimony to model indels, so the 'true' ancestral proteins may have possessed more variation through indels. In future, it would be informative to repeat the ASR of the class III peroxidases using a program such as GRASP which can execute explicit indel modelling throughout and characterise any resulting differences in the reconstructed ancestors[181].

Notwithstanding the lack of indel modelling in this study, ultimately this ancestral sequence reconstruction was able to identify functional ancestral peroxidase proteins. Moreover, the selected ancestral proteins could all be recombinantly expressed in a soluble, active form in *E. coli*. This is an advancement on previous attempts to engineer *A. rusticana* HRP in *E. coli,* such as the directed evolution study by Lin *et al.*[665], which resulted in very small yields of soluble HRP, or the more recent study by Chauhan *et al.*[659], which although had larger yields, ultimately expressed HRP in an inactive form with a large solubility tag and required an extra activation step. Here, it is theorised that the use of ASR predicted ancestral proteins which

were more stable, such that they could be heterologously expressed in their unglycosylated state. However, the expression of the selected ancestral proteins in *E. coli* emphasised the importance of disulfide bond formation. This was shown in the expression trial of the selected ancestors, where varying expression temperature and time had less effect on yield of active protein than using an *E. coli* strain which facilitated cytoplasmic disulfide bond formation.

The expression trial of the selected ancestral proteins also revealed considerable variation in peroxidase activity and/or expression levels among the selected ancestral proteins. This emphasises the strength in testing multiple ancestral proteins of varying ages, as it is currently difficult to predict which ancestor will possess the properties best suited to the desired application. For instance, here ancestors in the middle of the phylogeny were found to have the best activity and/or expression level when expressed recombinantly in *E. coli,* but for the LacI family the last ancestor of the family had the greatest expression yields and most interesting ligand binding properties for future applications (Chapter 2.3). Although characterising multiple ancestors for the desired function is not ideal, currently other protein engineering methods such as *de novo* design[7] and machine learning-based design[721,722] require many, sometimes hundreds, of protein variants to be tested to find the protein with the desired function. Therefore, ASR remains a useful protein engineering tool that allows functional proteins to be identified relatively easily for various desired applications.

The ability of ASR to generate highly functional sequences in a 'one-shot' approach was further demonstrated by the more detailed characterisation of recombinant Anc4, Anc7, Anc9 and Anc11, which revealed that they were predicted to have tertiary structures similar to HRP C1A, possessed secondary structure similar to HRP C1A, were thermostable, and all were able to act as peroxidase catalysts. Firstly, the predicted tertiary structures of the selected ancestral peroxidases showed that each of the selected ancestors likely folded into very similar tertiary structures to each other as well as to the extant HRP C1A. Given that the extant members of the class III peroxidase family tend to have a conserved structure[634], that the predicted structures of the selected ancestral proteins are also similar suggests that the class III peroxidase family structure has not changed much over time, and thus must be relatively important to peroxidase function. It is interesting that both the class III peroxidase family and the LacI family (Chapter 2.2) maintained relatively similar structures and functions over evolutionary time, while the individual protein sequences were relatively varied. This provides support to the theory that while nucleotides are the units mediating

evolutionary changes, the selection pressure is stronger on the protein structural and functional level than on the actual DNA and sequence level, where there tends to be greater neutral drift[212,723,724].

Examination of the predicted tertiary structures of Anc4, Anc7, Anc9 and Anc11 also revealed that they likely contained a signal peptide on their N-terminus, which was supported by the SignalP prediction that the same region was likely a cleavable signal peptide. Extant members of the class III peroxidase family are known to also contain a similar N-terminal signal peptide region, mainly directing their localisation to the plant cell wall[627,680]. Therefore, as the ancestors are also predicted to contain a similar signal peptide region, this then suggests that the localisation of the class III peroxidase family has likely remained highly conserved over time. In future, it would be interesting to investigate when and how the three main classes of heme-containing non-animal peroxidases split in their respective localisations. More specifically for the *E. coli* recombinant expression of the selected class III peroxidase ancestors, in future it would be beneficial to test if their expression in *E. coli* could be improved by not including the signal peptide region, as this would normally be cleaved in a plant expression system and is relatively hydrophobic[680].

Perhaps partly due to their respective signal peptides, the respective soluble yields of the *E. coli* expressed Anc4, Anc7, Anc9 and Anc11 are relatively low. However, the yields are nevertheless an improvement on previous attempts for soluble expression without the use of a large solubilizing tag, which previously has not been cleaved[659,702]. Notwithstanding, in future expression tests, the levels of the ancestral peroxidases need to be improved; possibly by not expressing the signal peptide region as mentioned before, by thoroughly optimising expression conditions, by supplementing with heme precursors and/or by adding a cleavable solubilisation tag. Resources permitting, the use of a bioreactor may also greatly improve ancestral peroxidase yield in *E. coli* as well as scalability for commercial use, as was demonstrated for HRP[660]. Given that the selected ancestral peroxidases express in measurable yields in *E. coli* using relatively unoptimized expression protocols unlike HRP C1A, it is probable that their optimised expression yields will be much greater than that of HRP C1A. However, despite soluble expression of the selected ancestral peroxidases, currently a relatively large proportion of each respective recombinant ancestor does not contain heme, thereby lowering the functional yield of each ancestor. Therefore, in future heme incorporation needs to be improved, possibly by incubating purified ancestral protein with hemin as has previously been conducted for recombinant HRP[659,660,702,706], by adding heme

precursors to the expression medium[725,726], or by using a bacterial strain engineered for heme incorporation[727]. It should also be considered, however, that the different selected ancestral proteins may have different binding preferences for the heme cofactor, so this should be characterised in future too.

It was interesting to find that the current expression yields of Anc4, Anc7, Anc9 and Anc11 correlate with their respective thermal unfolding transition temperatures, providing support to previous studies which have also suggested a correlation between protein stability and recombinant expression[668-670]. Thus, the hypothesis of this study that use of ASR could predict a more thermostable ancestor, which would express better in *E. coli* than the extant HRP, proved relatively sound. More generally, thermostable proteins are very useful as they tend to be more stable over a wide range of temperatures and provide a good starting point for directed evolution, as stability often decreases with new functional gains[227,229]. Hence, the ability of ASR to identify potential thermostable ancestors, as demonstrated here for the class III peroxidase family and in the previous LacI family ASR (Chapter 2.3.1), is a strength and one which could be more laborious to achieve in other engineering methods such as *de novo* design, directed evolution or machine-learning assisted design[221]. Interestingly, it was found here that the relatively thermostability of the selected ancestral peroxidases did not correlate with their relative ages, unlike for the LacI family ancestors and other ancestors of different families[189,710,711] (Chapter 2.3.1). They also do not appear to correlate with world climatic trends[728] based on their roughly estimated ages[729-734], though it should be noted that the predicted global temperature average does not capture the nuances of different climatic zones worldwide or intracellular temperature. However, as only four ancestors were characterised, any correlations or lack thereof are not statistically significant. Also, the absence of clear patterns in the respective thermostabilities of the selected ancestral peroxidases could result from the lack of their native glycosylation; hence perhaps the pattern of their relatively thermostabilities could differ with glycosylation. As lack of native glycosylation tends to have a destabilising effect on many proteins[655,708,709], it is particularly impressive that the thermal transition unfolding temperatures of the selected recombinant ancestors are as high as they are and especially that that of Anc9 is greater even than that of the glycosylated *Aru*HRP. This thereby indicates that Anc9 could have promising future utility.

A further benefit of using ASR as a protein engineering tool was demonstrated through the selected ancestral peroxidases all demonstrating peroxidase catalytic function. More detailed characterisation of their catalysis of ABTS, however, revealed that while their respective

Michaelis constants were relatively similar to the commercial *Aru*HRP, their respective turnover numbers were significantly lower than that of *Aru*HRP. It is hypothesised that the turnover numbers of the selected ancestral enzymes are low partly due to poor heme cofactor incorporation, which likely resulted in the functional enzyme concentration being much lower than the total enzyme concentration. By estimating the functional enzyme concentration of the selected ancestors, the turnover numbers did improve, however, only Anc9 was predicted to be in the range of the commercial *Aru*HRP for ABTS catalysis. Therefore, both the turnover number and Michaelis constant of Anc9 suggests that it might be the most suitable replacement for *Aru*HRP.

Interestingly, the respective Michaelis constants for ABTS catalysis consecutively increased with ancestral age. Perhaps this could result from the small changes in the binding pocket size and polarity over time, as predicted by the AlphaFold structures. This consecutive decrease in the ABTS catalysis Michaelis constant over time could suggest that binding preferences have gradually increased towards ABTS-like molecules along the HRP lineage, in a similar manner to binding preferences gradually increasing towards β-D-galactosides along the LacI lineage (Chapter 2.3.2). The decrease in Michaelis constants of ABTS catalysis of the class III peroxidase ancestors provides further support to the growing literature that shows that substrate binding preference appears to change relatively gradually over time[202,560,735]. This trend of gradual shift in binding preference over time is important for future protein design efforts, as it suggests that in some cases if a novel substrate is desired then the best starting framework could be a protein that already has a somewhat similar substrate. For example, Rondon & Wilson[405] were able to develop a novel caffeine sensor using the *E. coli* purine repressor as a starting point, as its native ligand of hypoxanthine is chemically similar to caffeine. In future, it would be interesting to further investigate how the binding profiles of the selected ancestors have changed over time by characterising the equilibrium binding dissociation constants and kinetics of the selected ancestral proteins with a wide range of potential substrates. It may be that each ancestor could have different applications for future use if they each have distinct substrate profiles. It would also be informative to investigate the catalysis of each ancestor over a range of pHs and temperatures, as they may have different optimal catalytic conditions, which could inform their future use. Additionally, it would be beneficial to characterise the activity of the selected ancestral peroxidases under different solvents and conditions, such as hydrophobic solvents and with surfactants, as some of the desired chemical catalysis applications of peroxidases, e.g. polystyrene catalysis, occur in

non-biological conditions[736]. Finally, on a more fundamental level it would be beneficial to characterise the binding and catalysis of hydrogen peroxide to the respective ancestral proteins, as this molecule is key to their function.

While the specific activity of recombinant Anc9 is not as high as that of commercial *Aru*HRP, it was shown that it can nevertheless catalyse the polymerisation of PNIPAm hydrogels in a similar manner to *Aru*HRP. Previously, the enzyme-mediated formation of PNIPAm has been catalysed typically by *Aru*HRP and glucose oxidase (GOx)[308,361], but this required formulation optimisation for each different batch of *Aru*HRP used. Instead, it was shown that the *E. coli* expressed ancestral peroxidase Anc9, which should not have batch-to-batch variation, could replace *Aru*HRP to form PNIPAm gels which had similar thermosensitivity and mechanical properties to both those formed by *Aru*HRP as well as those formed using chemical mediated synthesis. The use of the Anc9 enzyme, in combination with GOx, to mediate PNIPAm synthesis is advantageous over chemical mediated synthesis as it does not need to be conducted under nitrogen using toxic chemicals[625]. Another advantage of enzyme use is that they can be expressed in cells, thus, in future it would be interesting to develop an engineered living material by programming *E. coli* for the surface display of Anc9[737], then using the *E. coli* in the PNIPAm gel formulation to mediate its formation. Furthermore, post formation the gels could be used as bioreactors or potentially in soft robotics[381,738]. As GOx is paired with HRP to mediate PNIPAm formation, in future it would be beneficial to investigate if an ancestral GOx might also be able to be expressed recombinantly in *E. coli*. It would also be interesting to investigate in future if Anc9 could be used to replace HRP in other applications such as in reactive oxygen species sensing systems, glucose sensing systems, immunoassays, bioremediation, *et cetera*.

Overall, it was demonstrated that ancestral sequence reconstruction could be used to ‘engineer’ ancestral peroxidases that had similar structures and functions to the extant horseradish peroxidase but could be expressed solubly in *E. coli.* Use of the recombinantly expressed ancestral peroxidases over *Aru*HRP could save time and money, as well as greatly reduce the batch-to-batch variation. However, expression conditions and heme incorporation of the ancestral peroxidases when expressed in *E. coli* still need to be optimised. Nevertheless, as an example of application, it was shown that the ancestral peroxidase Anc9 could be used in place of *Aru*HRP for enzyme-mediated polymerisation of PNIPAm. Therefore, ancestral sequence reconstruction proved a useful tool for not only better understanding the evolution of the plant class III peroxidase protein family, but also for

predicting ancestral proteins with desired functions and characteristics, such as improved recombinant expression in *E. coli*. This study provides further support to the growing number of studies where it has been demonstrated that ASR can be used to ameliorate recombinant protein expression and thus, this study highlights another application of ASR for future protein engineering efforts. Subsequently, the utility of evolutionary methods of protein engineering was further shown in the next chapter by using an enzyme that was engineered through directed evolution to catalyse the formation of an alg/PNIPAm engineered living material.

## 3.5 Methods

### *3.5.1 Class III peroxidase sequence collection and curation*

The coding sequence for horseradish peroxidase isoenzyme C1A was used as an initial seed sequence. This sequence was obtained from entry P00433 in UniProt[685]. This seed sequence was then used to run a protein BLAST (blastp) (v1.11.1.1) search on both the UniProt and RedoxiBase[628] databases using the default settings, except to increase the number of hits to 1000 and 500 respectively. From this sequence set, any redundant sequences (occurred in both the UniProt and RedoxiBase search) and sequences which were partial or fragmented were removed. Then, a simple multiple sequence alignment was conducted using MAFFT DASH, and sequences which did not contain key heme binding or catalytic residues for peroxidase activity were removed. In HRP C1A without its secretion signal these residues are R38, F41, H42, D43, G48, D50, S52, E64, N70, H170, T171, D222, and T225[603].

CD-HIT[681] was utilised to cluster sequences by 80% identity and the representative sequences were taken from each cluster, resulting in 253 sequences. The sequence set was compared to the Viridiplantae organismal phylogeny. RedoxiBase was manually searched for class III peroxidases from ferns and bryophytes and some additional sequences were added to the set, resulting in 328 sequences.

Sequence similarity networks were constructed using the EFI-EST webserver[33,688] and images of the networks were made using Cytoscape (v 3.10.3)[739].

### *3.5.2 Multiple sequence alignment*

Multiple sequence alignment was conducted on the curated sequence set using Clustal Omega[54] (webserver accessed 17/6/22), Muscle[55] (webserver accessed 17/6/22),

DialignTX[690] (v 2), MAFFT G-INS-I[64] (v 7.4.9), T-COFFEE[42] (v 13.46), ProbCons[63] (v 1.1.2), MAFFT DASH[67] (v 7.4.9), PROMALS3D[66] (webserver accessed 20/6/22) and BAli-Phy[40] (3.6.1). T-COFFEE was also used to combine the results of all alignments except BAli-Phy into a single alignment through a customized M-COFFEE[69] function. To compare the resulting MSAs, a method developed by Chang *et al.*[2] was used which calculated a 'transitive consistency score' for each alignment. This was conducted for each MSA individually within T-COFFEE. A symmetry test algorithm within IQ-TREE (2.2.0) was used to predict if the assumptions of homogeneity and stationarity could be made for each MSA except BAli-Phy.

### *3.5.3 Phylogeny construction*

Iterative phylogeny construction was conducted for the MAFFT DASH and ProbCons MSAs respectively using IQ-TREE[143] (v2.2.0). Each phylogeny was rooted using *Marchantia polymorpha* peroxidase sequence (Mp179) as the outgroup, because this organism is predicted to have split off the earliest in the Viridiplantae out of my sequence set[682]. Also, each time a phylogeny was constructed, 10000 ultrafast bootstrap replicate tests (Ufboot)[155] and 10000 SH-like approximate likelihood ratio tests (SH alrt)[142] were run and the model for phylogeny construction was predicted by ModelFinder[103].

After each round of phylogeny construction based on the MAFFT DASH and ProbCons MSAs respectively, a custom Python script was used to extract the IDs of sequences with UF-boot or SH-alrt values less than 50. Then, for each of these low support sequences respectively, it was manually validated in the phylogeny if the clade it was in contained sequences originating from closely related plants or not. If the clade contained peroxidase sequences all originating from plants distantly related to the origin plant of the low support sequence, then the low support sequence was removed from the sequence set. After this process, the phylogeny was also manually inspected for the presence of long branch attraction. Any particularly obvious cases were removed from the sequence set. Following this, the abridged sequence set was aligned through their respective MSA software again, then the phylogeny construction repeated. This whole cycle was repeated until there appeared to be minimal long branch attraction in the phylogeny. For the ProbCons-based phylogeny construction process, this took 5 rounds, and for the MAFFT DASH-based phylogeny construction process, this took 11 rounds.

Following this, 9 further phylogenies were generated, yielding 10 phylogenies based on the ProbCons and MAFFT DASH MSAs respectively. The Visual TreeCmp webserver[694]

(accessed 31/1/23) was used to calculate the weighted Robinsin Foulds distance, quartet distance, path difference distance, matching triplet metric, matching split distance, unrooted maximum agreement subtree distance and geodesic unrooted distance between the 10 trees for each set respectively. For each set, the phylogeny which was the closest to the most other phylogenies by the most distance metrics was selected as the final phylogeny. This was phylogeny #7 for the MAFFT DASH MSA-based process, and phylogeny #6 for the ProbCons-based process.

Network diagram was made using the Networkx package in Python, and phylogeny diagrams were created using the ggtree[485] package in R.

### *3.5.4 Ancestral sequence reconstruction*

Ancestral sequence reconstruction was conducted using the empirical Bayes (aka marginal maximum likelihood) method through PAML CODEML[179]. Before reconstruction, custom Python scripts were created to format a given MSA and tree file respectively into the format required by PAML. For the ProbCons-based reconstruction, the VT model[94] with 10 categories and empirical model was used because previously ModelFinder had predicted this model to be the best for this MSA. Similarly, for the MAFFT DASH-based reconstruction, the WAG model[95] with 8 categories and empirical + F model was selected based on the results of ModelFinder.

Following the reconstruction process, custom MatLab and Python scripts were created to extract the maximum a posteriori sequences for each ancestral node, and to extract and process the corresponding posterior probability data. Finally, as the reconstruction process used by PAML cannot handle indels, ancestral sequences are predicted to contain residues for every column of the MSA of extant sequences. As such, a custom Python script was created to predict positions corresponding to ‘gaps’ (insertions) according to parsimony by checking whether the position contained a residue in over 50% of the MSA of extant sequences. If not, the position was marked as a ‘gap’ position, and the residue in this position removed from all the ancestral sequences. The posterior probability data was correspondingly updated using a custom Python script to remove the ‘gap’ position data.

As discussed in Results 3.3.5, 14 ancestral sequences were chosen for further characterisation from the *Armoracia rusticana* HRP C1A lineage based on their average posterior probability, and the likelihood of the position of the ancestral node (UF-Boot and SH-aLRT).

### *3.5.5 Expression trial of selected ancestral sequences*

The selected ancestral sequences were codon optimized for expression in *E. coli* using ThermoFisher's GeneArt. The genes were then synthesised and cloned into the pET-29b(+) vector by Twist Bioscience. The resulting plasmids containing the selected ancestral genes were transformed into SHuffle T7 express chemically competent *E. coli* and OneShot BL21(DE3)pLysS chemically competent *E. coli* respectively according to the supplier's protocol. A single colony from each transformant was then used to inoculate a 5 mL overnight culture. Each overnight culture was used to inoculate two 20 mL TB cultures respectively using a 200 µL aliquot. When the $O.D._{600}$ of the TB cultures reached 0.4, they were then induced with IPTG at a working concentration of 1 µM. It should be noted that the SHuffle cultures took significantly longer to reach the desired optical density than the BL21(DE3)pLyS cultures, as expected from the supplier's protocols. After induction, two expression conditions were tested for each plasmid and bacterial strain: 24 hrs at 30 °C or 48 hrs at 18 °C. Once these conditions had been met, the culture was pelleted through centrifugation for 15 min at 4000 rpm, 4 °C. Once the supernatant had been discarded, the wet weight of the pellet was measured, and then the pellet was stored at -20 °C.

To lyse the cells, each pellet was defrosted on ice and then the cells resuspended and lysed using BugBuster according to the supplier's protocol. It should be noted that 25 U/mL of benzonuclease was additionally used during this process. The soluble and insoluble protein fractions for each sample were separated, and the insoluble fraction further washed according to the BugBuster protocol. Following this, the resulting insoluble fraction pellets were solubilized using 100 µL of solubilisation solution (8 M urea, 5 mM TCEP).

Peroxidase expression in both the soluble and insoluble protein fractions for each sample were characterised through SDS-PAGE. For SDS-PAGE, 16% Tris-glycine gels (ThermoFisher Novex) were used with PageRuler unstained broad range protein ladder (ThermoFisher). Electrophoresis was run at 140 V for 60 min. Post this, gels were stained using InstantBlue (Abcam) overnight. Soluble fractions were also characterised for peroxidase activity using a colorimetric assay with 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS). The assay conditions for the colorimetric ABTS assay were 0.3 mM ABTS (Sigma-Aldrich), 0.03% w/w H2O2 (Sigma-Aldrich), 1:10 dilution of sample and assay buffer (50/100 mM citrate/phosphate, pH 4.5). Three replicates for each sample were conducted, as well as three replicates of buffer as a negative control. After setting up the assay in a clear 96 well plate (Costar Corning), the plate briefly centrifuged to mix and then

incubated at room temperature for approximately 30 min. After incubation, the absorbance at 418 nm of each well in the plate was measured using a plate reader.

### *3.5.6 Modelling of 3D protein structures of Anc4, Anc7, Anc9 and Anc11 and signal peptide prediction*

Models were created from the sequences of Anc4, Anc7, Anc9 and Anc11 using the webserver ColabFold[495]. Default settings were used, except the top five structures were relaxed with Amber, and the pdb100 template mode was selected. Protein models were viewed, and figures made in open source PyMol.

The sequences for Anc4, Anc7, Anc9 and Anc11 were tested for containing a signal peptide using SignalP 6.0[700] with 'eukaryotic' prediction selected.

### *3.5.7 Expression and purification of Anc4, Anc7, Anc9 and Anc11*

Plasmids containing the respective ancestral gene were transformed into SHuffle T7 Express according to the supplier's protocol. Transformants were used to inoculate an overnight culture of 100 mL LB medium, which was incubated at 30 °C, 200 rpm shaking. A 10 mL aliquot of overnight culture was then used to inoculate a 1 L culture of TB medium. This was incubated at 30 °C, 200 rpm shaking until $O.D._{600}$ reached 0.4, when protein expression was induced through the addition of IPTG. Cultures were then incubated at 18 °C, 200 rpm for 48 hrs. After incubation, the culture was pelleted through centrifugation (4 °C, 5000 rpm, 15 min), the supernatant discarded, and the pellet collected and frozen at -20 °C.

For purification, cell pellets were defrosted on ice, then resuspended in Buffer A (50 mM potassium phosphate, 300 mM sodium chloride, 20 mM imidazole, 2 mM calcium chloride, 5% v/v glycerol, pH 7.4) before being lysed by sonication. The lysate containing the soluble protein fraction was then separated from the cell debris and insoluble protein through centrifugation (4 °C, 18000 rpm, 60 min). The lysate was filtered through a 0.45 um filter before being loaded onto a 5 mL HisTrap HP column (Cytiva). Protein bound to the column was eluted with buffer B (50 mM potassium phosphate, 300 mM sodium chloride, 250 mM imidazole, 2 mM calcium chloride, 5% v/v glycerol, pH 7.4) before being changed into Buffer C (50 mM potassium phosphate, 150 mM sodium chloride, 2 mM calcium chloride, 5% v/v glycerol, pH 7.4) using 5*5 mL HiTrap Desalting columns (Cytiva). The protein was then further purified by size exclusion chromatography using a HiLoad 16/600 Superdex pg (Cytiva) column and eluting with Buffer C. Fractions were tested for peroxidase activity using the ABTS assay (0.3 mM ABTS [Sigma-Aldrich A1888], 0.03% w/w $H_2O_2$ [Sigma-

Aldrich H1009], 50 mM potassium phosphate (pH 6.0) buffer) and fraction protein content was investigated through SDS-PAGE. Fractions with the highest peroxidase activity were combined, concentrated using 10 kDa cutoff centrifuge filters, and concentration measured using the BCA assay with bovine hemoglobin standards.

### *3.5.8 Absorbance spectrum and Reinheitszahl value of Anc4, Anc7, Anc9, Anc11 and commercial horseradish peroxidase*

The absorbance spectrum from 200-800 nm of the selected ancestral proteins and commercial horseradish peroxidase (Alfa Aesar/Thermo Scientific J60026.MC) was measured using a CARY spectrophotometer at 60 nm/min with a 0.5 nm interval. Buffer signal was also measured and subtracted from the respective protein signals. The Reinheitszahl value of the selected ancestral proteins and HRP was calculated by dividing the absorbance at 403 nm by the absorbance at 275 nm.

### *3.5.9 Circular dichroism of Anc4, Anc7, Anc9, Anc11 and commercial horseradish peroxidase*

The selected ancestors were exchanged into di$H_2O$ using 10 kDa cutoff centrifugal filters. Concentration of the selected ancestors was then measured through the BCA assay as previously and molecular weights of the selected ancestors were calculated using ProtParam[594]. The molecular weight of the commercial horseradish peroxidase was assumed to be 44 kDa[602]. Throughout the CD experiments, a 10 mm pathlength quartz cuvette was used. First, the spectrum of di$H_2O$ at 25 °C was measured from 190 – 280 nm at 5 nm/min with 1 nm interval and 1 nm bandwidth to determine the baseline. Using the same conditions, the spectrum of each selected ancestor and commercial HRP was then measured. Mean residue ellipticity was calculated and secondary structure predicted by the CDSSTR, SELCON and K2D methods through the DichroWeb server[740] using sets 4 or 7 when appropriate. The prediction with the lowest NRMSD was selected as the ‘best’.

Thermal denaturation of each selected ancestor was measured from 205 nm to 225 nm from 25 °C to 90 °C at 1 or 5 °C intervals. A sigmoidal curve was fitted to the data for each wavelength respectively using the Curve Fitter app in Matlab and the transition point for each wavelength calculated from the curve fit. The thermal transition unfolding temperature was calculated by taking the mean of the transition points from each wavelength. Following this, from the fitted sigmoidal curves the predicted constants for the maximum and minimum CD signals at each wavelength were obtained and used to calculate the fraction of folded protein

for each wavelength for each temperature. Subsequently, the mean fraction of folded protein was calculated across the wavelengths for each temperature, omitting outliers (more than 3 median absolute deviations from the median).

### *3.5.10 Kinetics and activity assays of Anc4, Anc7, Anc9, Anc11 and commercial horseradish peroxidase using ABTS*

Kinetics assays used excess 2,2'-Azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) and $H_2O_2$ such that the Michaelis-Menten model could be used as an estimate for kinetic calculations. All assays were conducted in 50 mM potassium phosphate, pH 6.0 buffer and had a working concentration of 0.03% w/w $H_2O_2$. $H_2O_2$ was added immediately prior to measurement because long incubation of peroxidases with $H_2O_2$ can lead to their partial inactivation. ABTS concentrations used were higher for Anc4, Anc7 and Anc9 and lower for Anc11 and HRP (Alfa Aesar/Thermo Scientific J60026.MC) based on their higher and lower $K_M$ values respectively (data not shown). Enzyme concentration used was determined based on experiments which showed if the linear response could be measured for the highest and lowest ABTS concentration tested (data not shown). For Anc4 this was a working concentration of 500 nM, for Anc7 2.5 µM, for Anc9 50 nM, for Anc11 500 nM and for the commercial HRP this was 5 nM. During the kinetics assays, changes in absorbance at 415 nm with 4 s intervals were monitored using a plate reader. This was converted into the concentration of $ABTS^{\cdot}$ product formed per second through the Beer-Lambert law assuming a pathlength of 0.56 cm and extinction coefficient of 36000 $M^{-1}$ $cm^{-1}$ [741]. A custom Matlab script was made to process the data and the Michaelis-Menten equation was used in the Curve Fitter app of Matlab to predict the values of $K_M$ and $v_{max}$. The approximate percentage of heme incorporation for each of the respective ancestors was calculated from a standard curve of Rz *vs.* heme incorporation[604,742]. This was used to estimate the respective concentrations of functional ancestor, which in turn was used to calculate $k_{cat}$*.

Activity assay of each of the selected ancestral proteins and commercial HRP was conducted using a working concentration of 60 mM ABTS, 0.03% w/w $H_2O_2$ and 5 nM of the respective protein all in 50 mM potassium phosphate buffer, pH 6.0. Change in absorbance at 415 nm with 4 s intervals was measured using a plate reader. Activity was calculated from the linear response using the same constants as in the kinetics assays.

### *3.5.11 Enzyme-mediated polymerisation of PNIPAm by Anc9 and commercial HRP*

Poly(*N*-isopropylacrylamide) (PNIPAm) was prepared according to the formula in Table 3.3 below, which is adapted from the alginate/PNIPAm double network formulation of Klemperer *et al.*[361]. Gels were stored in di$H_2O$ after formation.

| Component | Concentration |
|---|---|
| *N*-isopropylacrylamide | 15% wt/wt |
| *N,N'*-Methylenebisacrylamide | 7.5 mg/mL |
| D-glucose | 1 mg/mL |
| glucose oxidase (type VII from *Aspergillus niger*) | 0.5 mg/mL |
| HRP or Anc9 | 50 µg/mL or 110 µg/mL |
| acetylacetone | 1% v/v |

**Table 3.3:** Formulation for enzyme-mediated PNIPAm hydrogel

The thermosensitivity of the PNIPAm hydrogels was tested by measuring their contraction at 60 °C. Gels were dried on filter paper, before being weighed and transferred to 1 mL of di$H_2O$. Gels were then either incubated at 60 °C in a heat block, or at room temperature as a control, for 30 min. Gels were then dried on filter paper and their weight measured again. Finally, gels were dried in a drying oven overnight, and their dried weight measured the next day. The weight ratio of the gels before and after incubation was calculated according to the following formula:

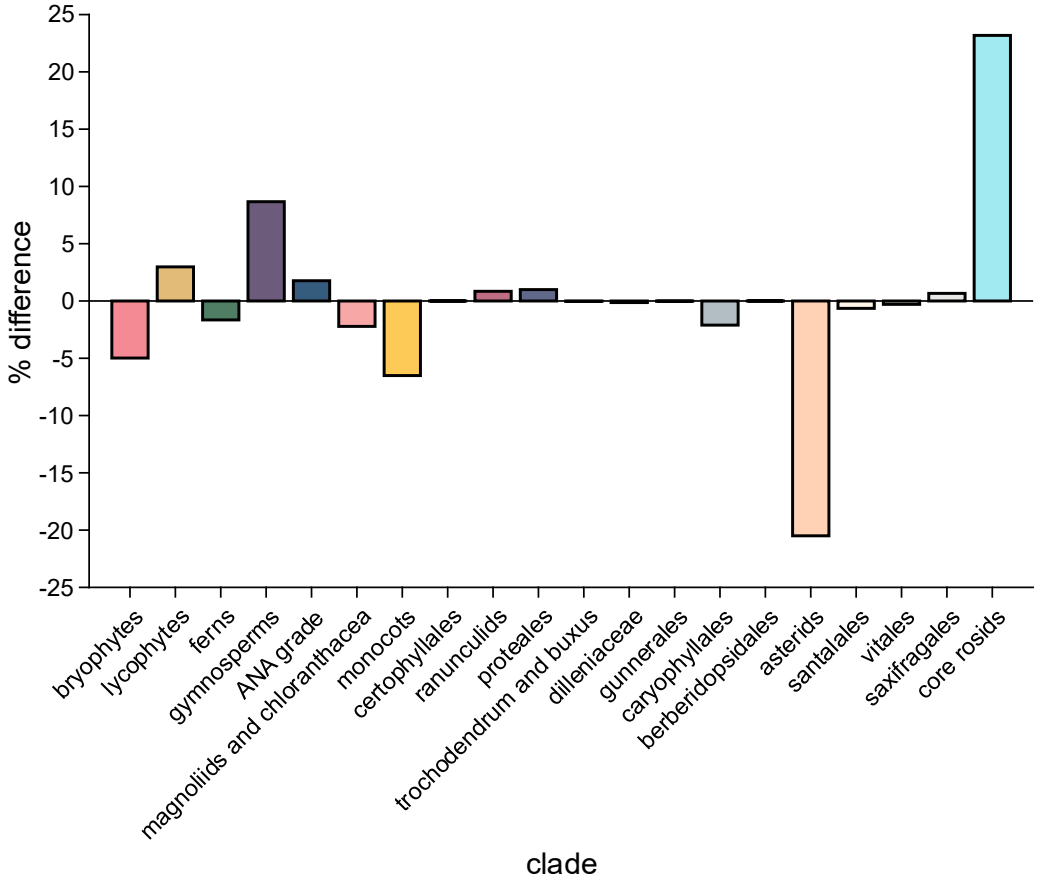
$$(de)swelling\ ratio\ (\%) = \frac{W_s - W_d}{W_d} \times 100$$

*Equation 3.1: where $W_s$ is the weight of the hydrogel sample and $W_d$ is the weight of the dried hydrogel*

The compressive strength of the PNIPAm hydrogels was characterised using an Starrett FMS500 universal testing machine fitted with a 10 N compressive load cell. PNIPAm gels were formed using custom 3D-printed cylindrical moulds with an inner height of 5 mm and inner diameter of 10 mm. The height and diameter of the gels was measured using electronic callipers. Compression was conducted with a speed of 1 mm/min. Data was analysed assuming engineering strength and strain using a custom Python script. Tangential compressive strength was calculated from the first and second linear parts of the curve.

## 3.6 Supplementary information

**Suppl. figure 3.1:** Relative representation of land plant clades in the final class III peroxidase sequence set as compared to the organismal phylogeny. A negative percentage difference indicates there is a lower proportion of clade sequences in the peroxidase set compared to the organismal phylogeny and vice versa for a positive percentage difference.




**Suppl. table 3.1***:* List of sequence IDs post trimming of sequence set using CD-HIT

| ID | Main clade | Sub clade | Species |
|---|---|---|---|
| **A0A067FBN7** | Eudicot | Core rosids | *Citrus sinensis* |
| **A0A067KDU9** | Eudicot | Core rosids | *Jatropha curcas* |
| **A0A067L780** | Eudicot | Core rosids | *Jatropha curcas* |
| **A0A067LI86** | Eudicot | Core rosids | *Jatropha curcas* |
| **A0A068UP04** | Eudicot | Asterids | *Coffea canephora* |
| **A0A087GQG4** | Eudicot | Core rosids | *Arabis alpina* |
| **A0A0A0KWW3** | Eudicot | Core rosids | *Cucumis sativus* |
| **A0A0C9RWS7** | Gymnosperm | Araucariales | *Wollemia nobilis* |
| **A0A0D2MVT2** | Eudicot | Core rosids | *Gossypium raimondii* |
| **A0A0D3E2V6** | Eudicot | Core rosids | *Brassica oleracea var. oleracea* |
| **A0A0D6QT46** | Gymnosperm | Araucariales | *Araucaria cunninghamii* |

| | | | |
|---|---|---|---|
| **A0A0D6R4B9** | Gymnosperm | Araucariales | *Araucaria cunninghamii* |
| **A0A0J8BPV6** | Eudicot | Caryophyllales | *Beta vulgaris* |
| **A0A0K9RGQ6** | Eudicot | Caryophyllales | *Spinacia oleracea* |
| **A0A0L9VDP0** | Eudicot | Core rosids | *Phaseolus angularis* |
| **A0A0S3S4A2** | Eudicot | Core rosids | *Phaseolus angularis* |
| **A0A0S3S4B6** | Eudicot | Core rosids | *Phaseolus angularis* |
| **A0A151RQY9** | Eudicot | Core rosids | *Cajanus cajan* |
| **A0A164ZWV6** | Eudicot | Asterids | *Daucus carota* |
| **A0A199UT12** | Monocot | Commelinids | *Ananas comosus* |
| **A0A199UUS6** | Monocot | Commelinids | *Ananas comosus* |
| **A0A199VKN8** | Monocot | Commelinids | *Ananas comosus* |
| **A0A199W3C6** | Monocot | Commelinids | *Ananas comosus* |
| **A0A1E5W9N7** | Monocot | Commelinids | *Dichanthelium oligosanthes* |
| **A0A1R3GSG1** | Eudicot | Core rosids | *Corchorus olitorius* |
| **A0A1R3HRT4** | Eudicot | Core rosids | *Corchorus capsularis* |
| **A0A1R3HRW9** | Eudicot | Core rosids | *Corchorus capsularis* |
| **A0A1S2Z1S4** | Eudicot | Core rosids | *Cicer arietinum* |
| **A0A1S3BNG1** | Eudicot | Core rosids | *Cucumis melo* |
| **A0A1S3E245** | Eudicot | Core rosids | *Cicer arietinum* |
| **A0A1S3E376** | Eudicot | Core rosids | *Cicer arietinum* |
| **A0A1S3TC61** | Eudicot | Core rosids | *Vigna radiata* |
| **A0A1S3V148** | Eudicot | Core rosids | *Vigna radiata* |
| **A0A1S3VGP7** | Eudicot | Core rosids | *Vigna radiata* |
| **A0A1S3VK04** | Eudicot | Core rosids | *Vigna radiata* |
| **A0A1S3XRT6** | Eudicot | Asterids | *Nicotiana tabacum* |
| **A0A1S3YJK7** | Eudicot | Asterids | *Nicotiana tabacum* |
| **A0A1U7X466** | Eudicot | Asterids | *Nicotiana sylvestris* |
| **A0A1U7Z6F6** | Eudicot | Proteales | *Nelumbo nucifera* |
| **A0A1U8FYA1** | Eudicot | Asterids | *Capsicum annuum* |
| **A0A200QRH9** | Eudicot | Ranunculids | *Macleaya cordata* |
| **A0A251UQX4** | Eudicot | Asterids | *Helianthus annuus* |
| **A0A287KS33** | Monocot | Commelinids | *Hordeum vulgare* |
| **A0A2G3BHR2** | Eudicot | Asterids | *Capsicum chinense* |
| **A0A2G9HZ15** | Eudicot | Asterids | *Handroanthus impetiginosus* |
| **A0A2H5PC80** | Eudicot | Core rosids | *Citrus unshiu* |
| **A0A2H5PC88** | Eudicot | Core rosids | *Citrus unshiu* |
| **A0A2H5PC89** | Eudicot | Core rosids | *Citrus unshiu* |
| **A0A2H5PCB9** | Eudicot | Core rosids | *Citrus unshiu* |
| **A0A2I0KYR8** | Eudicot | Core rosids | *Punica granatum* |

| | | | |
|---|---|---|---|
| **A0A2I0X575** | Monocot | Asparagales | *Dendrobium catenatum* |
| **A0A2I4FS82** | Eudicot | Core rosids | *Juglans regia* |
| **A0A2I4FS99** | Eudicot | Core rosids | *Juglans regia* |
| **A0A2I4H5W6** | Eudicot | Core rosids | *Juglans regia* |
| **A0A2I4HS78** | Eudicot | Core rosids | *Juglans regia* |
| **A0A2J6K2F9** | Eudicot | Asterids | *Lactuca sativa* |
| **A0A2K1KX56** | Bryophyte | Mosses | *Physcomitrium patens* |
| **A0A2K3L1L5** | Eudicot | Core rosids | *Trifolium pratense* |
| **A0A2P5BAI3** | Eudicot | Core rosids | *Trema orientale* |
| **A0A2P5ECE6** | Eudicot | Core rosids | *Trema orientale* |
| **A0A2P5WSA1** | Eudicot | Core rosids | *Gossypium barbadense* |
| **A0A2P5YG01** | Eudicot | Core rosids | *Gossypium barbadense* |
| **A0A2P6QG18** | Eudicot | Core rosids | *Rosa chinensis* |
| **A0A2P6QMU1** | Eudicot | Core rosids | *Rosa chinensis* |
| **A0A2P6SLM3** | Eudicot | Core rosids | *Rosa chinensis* |
| **A0A2R6P897** | Eudicot | Asterids | *Actinidia chinensis var. chinensis* |
| **A0A2T7C983** | Monocot | Commelinids | *Panicum hallii var. hallii* |
| **A0A2T7EZT3** | Monocot | Commelinids | *Panicum hallii var. hallii* |
| **A0A2Z7CUJ2** | Eudicot | Asterids | *Dorcoceras hygrometricum* |
| **A0A3S3P255** | Magnoliids | Laurales | *Cinnamomum micranthum* |
| **A0A3S3PQZ1** | Magnoliids | Laurales | *Cinnamomum micranthum* |
| **A0A443N798** | Magnoliids | Laurales | *Cinnamomum micranthum* |
| **A0A444FNK8** | Monocot | Commelinids | *Ensete ventricosum* |
| **A0A444FTW3** | Monocot | Commelinids | *Ensete ventricosum* |
| **A0A444X1G0** | Eudicot | Core rosids | *Arachis hypogaea* |
| **A0A444ZDM4** | Eudicot | Core rosids | *Arachis hypogaea* |
| **A0A444ZDQ5** | Eudicot | Core rosids | *Arachis hypogaea* |
| **A0A445BMR3** | Eudicot | Core rosids | *Arachis hypogaea* |
| **A0A445DIS9** | Eudicot | Core rosids | *Arachis hypogaea* |
| **A0A445DIV1** | Eudicot | Core rosids | *Arachis hypogaea* |
| **A0A446QPC3** | Monocot | Commelinids | *Triticum turgidum subsp. Durum* |
| **A0A452ZW37** | Monocot | Commelinids | *Aegilops tauschii subsp strangulata* |
| **A0A453BZ35** | Monocot | Commelinids | *Aegilops tauschii subsp strangulata* |
| **A0A453IJK6** | Monocot | Commelinids | *Aegilops tauschii subsp strangulata* |
| **A0A484K3I1** | Eudicot | Asterids | *Cuscuta campestris* |
| **A0A498IGV2** | Eudicot | Core rosids | *Malus domestica* |

| | | | |
|---|---|---|---|
| **A0A498K6B9** | Eudicot | Core rosids | *Malus domestica* |
| **A0A498K8W0** | Eudicot | Core rosids | *Malus domestica* |
| **A0A498K954** | Eudicot | Core rosids | *Malus domestica* |
| **A0A4D8Y928** | Eudicot | Asterids | *Salvia splendens* |
| **A0A4D8Z9M1** | Eudicot | Asterids | *Salvia splendens* |
| **A0A4P8JNH7** | Bryophyte | Mosses | *Pohlia nutans* |
| **A0A4S4ETG8** | Eudicot | Asterids | *Camellia sinensis* |
| **A0A4S8IN62** | Monocot | Commelinids | *Musa balbisiana* |
| **A0A4S8IW54** | Monocot | Commelinids | *Musa balbisiana* |
| **A0A4S8JBP1** | Monocot | Commelinids | *Musa balbisiana* |
| **A0A4S8KA43** | Monocot | Commelinids | *Musa balbisiana* |
| **A0A4U6T1Y7** | Monocot | Commelinids | *Setaria viridis* |
| **A0A4U6TA82** | Monocot | Commelinids | *Setaria viridis* |
| **A0A4Y7KWG0** | Eudicot | Ranunculids | *Papaver somniferum* |
| **A0A4Y7LG51** | Eudicot | Ranunculids | *Papaver somniferum* |
| **A0A540LHP9** | Eudicot | Core rosids | *Malus baccata* |
| **A0A565CIG9** | Eudicot | Core rosids | *Arabis nemorensis* |
| **A0A5A7SKP8** | Eudicot | Core rosids | *Cucumis melo* |
| **A0A5B6UNN7** | Eudicot | Core rosids | *Gossypium australe* |
| **A0A5B6VHL9** | Eudicot | Core rosids | *Gossypium australe* |
| **A0A5C7HJB2** | Eudicot | Core rosids | *Acer yangbiense* |
| **A0A5C7HRP9** | Eudicot | Core rosids | *Acer yangbiense* |
| **A0A5C7IG44** | Eudicot | Core rosids | *Acer yangbiense* |
| **A0A5C7IV37** | Eudicot | Core rosids | *Acer yangbiense* |
| **A0A5C7IX76** | Eudicot | Core rosids | *Acer yangbiense* |
| **A0A5J4ZJ22** | Eudicot | Asterids | *Nyssa sinensis* |
| **A0A5J5BWC5** | Eudicot | Asterids | *Nyssa sinensis* |
| **A0A5J5C6V6** | Eudicot | Asterids | *Nyssa sinensis* |
| **A0A5J9V7M5** | Monocot | Commelinids | *Eragrostis curvula* |
| **A0A5J9WHR0** | Monocot | Commelinids | *Eragrostis curvula* |
| **A0A5N5P1B1** | Eudicot | Core rosids | *Salix brachista* |
| **A0A5N6L6C2** | Eudicot | Core rosids | *Carpinus fangiana* |
| **A0A5P1EUX1** | Monocot | Asparagales | *Asparagus officinalis* |
| **A0A660KKX8** | Eudicot | Core rosids | *Carpinus fangiana* |
| **A0A6A3CD50** | Eudicot | Core rosids | *Hibiscus syriacus* |
| **A0A6A5P0D9** | Eudicot | Core rosids | *Lupinus albus* |
| **A0A6A6MRF6** | Eudicot | Core rosids | *Hevea brasiliensis* |
| **A0A6I9RJ83** | Monocot | Commelinids | *Elaeis guineensis* |
| **A0A6I9RN05** | Monocot | Commelinids | *Elaeis guineensis* |

| | | | |
|---|---|---|---|
| **A0A6I9S386** | Monocot | Commelinids | *Elaeis guineensis* |
| **A0A6I9S9K4** | Monocot | Commelinids | *Elaeis guineensis* |
| **A0A6I9SV98** | Eudicot | Asterids | *Sesamum indicum* |
| **A0A6J0ZTD6** | Eudicot | Core rosids | *Herrania umbratica* |
| **A0A6J0ZVL5** | Eudicot | Core rosids | *Herrania umbratica* |
| **A0A6J1DIP9** | Eudicot | Core rosids | *Momordica charantia* |
| **A0A6J1DJS5** | Eudicot | Core rosids | *Momordica charantia* |
| **A0A6J1DLR6** | Eudicot | Core rosids | *Momordica charantia* |
| **A0A6J1I1V1** | Eudicot | Core rosids | *Cucurbita maxima* |
| **A0A6J1J318** | Eudicot | Core rosids | *Cucurbita maxima* |
| **A0A6J5WT12** | Eudicot | Core rosids | *Prunus armeniaca* |
| **A0A6J5XFK0** | Eudicot | Core rosids | *Prunus armeniaca* |
| **A0A6P3ZQ77** | Eudicot | Core rosids | *Ziziphus jujuba* |
| **A0A6P3ZTV3** | Eudicot | Core rosids | *Ziziphus jujuba* |
| **A0A6P4ATK4** | Eudicot | Core rosids | *Ziziphus jujuba* |
| **A0A6P4D4Q3** | Eudicot | Core rosids | *Arachis duranensis* |
| **A0A6P4E0H2** | Eudicot | Core rosids | *Arachis duranensis* |
| **A0A6P5SX62** | Eudicot | Core rosids | *Prunus avium* |
| **A0A6P6G281** | Eudicot | Core rosids | *Ziziphus jujuba* |
| **A0A6P6S8Q1** | Eudicot | Asterids | *Coffea arabica* |
| **A0A7J6X8Q5** | Eudicot | Ranunculids | *Thalictrum thalictroides* |
| **A0A7J7CNR9** | Eudicot | Core rosids | *Tripterygium wilfordii* |
| **A0A7J7CTD7** | Eudicot | Core rosids | *Tripterygium wilfordii* |
| **A0A7J7DR46** | Eudicot | Core rosids | *Tripterygium wilfordii* |
| **A0A7J7L9C9** | Eudicot | Ranunculids | *Kingdonia uniflora* |
| **A0A7J7P3N0** | Eudicot | Ranunculids | *Kingdonia uniflora* |
| **A0A7J8XWZ9** | Eudicot | Core rosids | *Gossypium aridum* |
| **A0A7J9HLZ5** | Eudicot | Core rosids | *Gossypium harknessii* |
| **A0A7N0THG3** | Eudicot | Saxifragales | *Kalanchoe fedtschenkoi* |
| **A0A7N2KLK8** | Eudicot | Core rosids | *Quercus lobata* |
| **A0A7N2KLM6** | Eudicot | Core rosids | *Quercus lobata* |
| **A0A803L6Y2** | Eudicot | Caryophyllales | *Chenopodium quinoa* |
| **A0A804HN12** | Monocot | Commelinids | *Musa acuminata* |
| **A0A804HUG2** | Monocot | Commelinids | *Musa acuminata* |
| **A0A804HV05** | Monocot | Commelinids | *Musa acuminata* |
| **A0A804I645** | Monocot | Commelinids | *Musa acuminata* |
| **A0A804JFN0** | Monocot | Commelinids | *Musa acuminata* |
| **A0A804JUK9** | Monocot | Commelinids | *Musa acuminata* |
| **A9PD65** | Eudicot | Core rosids | *Populus trichocarpa* |

| | | | |
|---|---|---|---|
| **AcoPrx85** | Eudicot | Ranunculids | *Aquilegia coerulea* |
| **AfiliPrx01** | Fern | Polypodiidae | *Azolla filiculoides* |
| **AfiliPrx02** | Fern | Polypodiidae | *Azolla filiculoides* |
| **AfiliPrx05** | Fern | Polypodiidae | *Azolla filiculoides* |
| **AfiliPrx06** | Fern | Polypodiidae | *Azolla filiculoides* |
| **AfpPrx06** | Eudicot | Ranunculids | *Aquilegia 141ormosa x Aquilegia pubescens* |
| **AncPrx01** | Monocot | Commelinids | *Ananas comosus* |
| **AruPrx01-1** | Eudicot | Core rosids | *Armoricana rustica* |
| **AruPrx22** | Eudicot | Core rosids | *Armoricana rustica* |
| **AtPrx36** | Eudicot | Core rosids | *Arabidopsis thaliana* |
| **AtrPrx08** | ANA grade | Amborellales | *Amborella trichopoda* |
| **AtrPrx27** | ANA grade | Amborellales | *Amborella trichopoda* |
| **AtrPrx48** | ANA grade | Amborellales | *Amborella trichopoda* |
| **B9GYK2** | Eudicot | Core rosids | *Populus trichocarpa* |
| **BdiPrx19** | Monocot | Commelinids | *Brachypodium distachyon* |
| **BrPrx23-1Aa_other** | Eudicot | Core rosids | *Brassica rapa* |
| **BvPrx21** | Eudicot | Caryophyllales | *Beta vulgaris* |
| **CcampPrx73** | Eudicot | Asterids | *Cuscuta campestris* |
| **CclPrx111** | Eudicot | Core rosids | *Citrus clementina* |
| **CclPrx30** | Eudicot | Core rosids | *Citrus clementina* |
| **CmichPrx01** | Gymnosperm | Cycads and gingko | *Cycas micholitzii* |
| **CpapPrx05** | Eudicot | Core rosids | *Carica papaya* |
| **CpapPrx06** | Eudicot | Core rosids | *Carica papaya* |
| **CroPrx19** | Eudicot | Asterids | *Catharanthus roseus* |
| **CroPrx22** | Eudicot | Asterids | *Catharanthus roseus* |
| **CsaPrx57** | Eudicot | Core rosids | *Cucumis sativus* |
| **CsPrx58** | Eudicot | Core rosids | *Citrus sinensis* |
| **D7LB84** | Eudicot | Core rosids | *Arabidopsis lyrata* |
| **D8RT10** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **D8S6M6** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **EcamPrx53** | Eudicot | Core rosids | *Eucalyptus camaldulensis* |
| **EferPrx127** | ANA grade | Nymphaeles | *Euryale ferox* |
| **EferPrx131** | ANA grade | Nymphaeles | *Euryale ferox* |
| **EferPrx31** | ANA grade | Nymphaeles | *Euryale ferox* |
| **EglPrx137** | Eudicot | Core rosids | *Eucalyptus globulus* |
| **EgrPrx191** | Eudicot | Core rosids | *Eucalyptus grandis* |

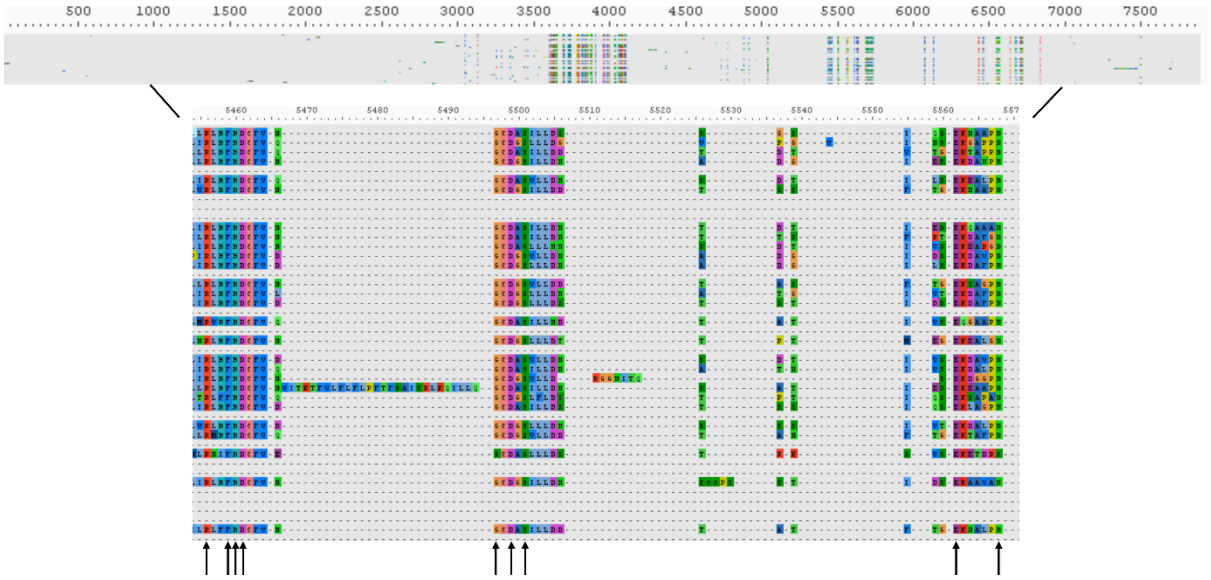
| | | | |
|---|---|---|---|
| **EguPrx55** | Eudicot | Core rosids | *Eucalyptus gunnii* |
| **FcPrx01** | Eudicot | Core rosids | *Ficus carica* |
| **GbPrx07** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx10** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx14** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx32** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx42** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx46** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx52** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx54** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx56** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx59** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GbPrx63** | Gymnosperm | Cycads and gingko | *Ginkgo biloba* |
| **GmPrx19** | Eudicot | Core rosids | *Glycine max* |
| **GmPrx262** | Eudicot | Core rosids | *Glycine max* |
| **GmPrx273** | Eudicot | Core rosids | *Glycine max* |
| **GmPrx57** | Eudicot | Core rosids | *Glycine max* |
| **HaPrx101** | Eudicot | Asterids | *Helianthus annuus* |
| **HaPrx31** | Eudicot | Asterids | *Helianthus annuus* |
| **HbPrx01** | Eudicot | Core rosids | *Hevea brasiliensis* |
| **HvPrx17** | Monocot | Commelinids | *Hordeum vulgare* |
| **IaquPrx06** | Eudicot | Asterids | *Ipomoea aquatica* |
| **IaquPrx55** | Eudicot | Asterids | *Ipomoea aquatica* |
| **IaquPrx57** | Eudicot | Asterids | *Ipomoea aquatica* |
| **InPrx36** | Eudicot | Asterids | *Ipomoea nil* |
| **InPrx47** | Eudicot | Asterids | *Ipomoea nil* |
| **InPrx55** | Eudicot | Asterids | *Ipomoea nil* |
| **J3LLQ6** | Monocot | Commelinids | *Oryza brachyantha* |

| | | | |
|---|---|---|---|
| **KfedPrx06** | Eudicot | Saxifragales | *Kalanchoe fedtschenkoi* |
| **KfedPrx38** | Eudicot | Saxifragales | *Kalanchoe fedtschenkoi* |
| **KfedPrx58** | Eudicot | Saxifragales | *Kalanchoe fedtschenkoi* |
| **KfedPrx60** | Eudicot | Saxifragales | *Kalanchoe fedtschenkoi* |
| **LaPrx01** | Eudicot | Core rosids | *Lupinus albus* |
| **LaPrx03** | Eudicot | Core rosids | *Lupinus albus* |
| **LcPrx02** | Eudicot | Core rosids | *Litchi chinensis* |
| **LePrx05** | Eudicot | Asterids | *Lycopersicon esculentum* |
| **LePrx28** | Eudicot | Asterids | *Lycopersicon esculentum* |
| **LePrx35** | Eudicot | Asterids | *Lycopersicon esculentum* |
| **LjPrx26** | Eudicot | Core rosids | *Lotus japonicus (corniculatus var japonicus)* |
| **LjPrx47** | Eudicot | Core rosids | *Lotus japonicus (corniculatus var japonicus)* |
| **LperPrx22** | Eudicot | Asterids | *Lactuca perennis* |
| **LuPrx148** | Eudicot | Core rosids | *Linum usitatissimum* |
| **LuPrx21** | Eudicot | Core rosids | *Linum usitatissimum* |
| **LuPrx42** | Eudicot | Core rosids | *Linum usitatissimum* |
| **LuPrx43** | Eudicot | Core rosids | *Linum usitatissimum* |
| **M1BAK2** | Eudicot | Asterids | *Solanum tuberosum* |
| **M5B2X1** | Gymnosperm | Cupressales | *Chamaecyparis obtusa* |
| **M7ZJD9** | Monocot | Commelinids | *Triticum urartu* |
| **M7ZJS0** | Monocot | Commelinids | *Triticum urartu* |
| **MePrx07** | Eudicot | Core rosids | *Manihot esculenta* |
| **MePrx135** | Eudicot | Core rosids | *Manihot esculenta* |
| **MePrx16** | Eudicot | Core rosids | *Manihot esculenta* |
| **MePrx57** | Eudicot | Core rosids | *Manihot esculenta* |
| **MguPrx101** | Eudicot | Asterids | *Mimulus guttatus* |
| **MguPrx17** | Eudicot | Asterids | *Mimulus guttatus* |
| **MguPrx86** | Eudicot | Asterids | *Mimulus guttatus* |
| **MpalPrx27** | Bryophyte | Liverworts | *Marchantia paleacea* |
| **MpalPrx94** | Bryophyte | Liverworts | *Marchantia paleacea* |
| **MpPrx179** | Bryophyte | Liverworts | *Marchantia polymorpha* |
| **MsPrx03** | Eudicot | Core rosids | *Medicago sativa* |
| **MtPrx17** | Eudicot | Core rosids | *Medicago truncatula* |
| **MtPrx23** | Eudicot | Core rosids | *Medicago truncatula* |
| **MtPrx30** | Eudicot | Core rosids | *Medicago truncatula* |
| **MtPrx35** | Eudicot | Core rosids | *Medicago truncatula* |
| **MtPrx37** | Eudicot | Core rosids | *Medicago truncatula* |

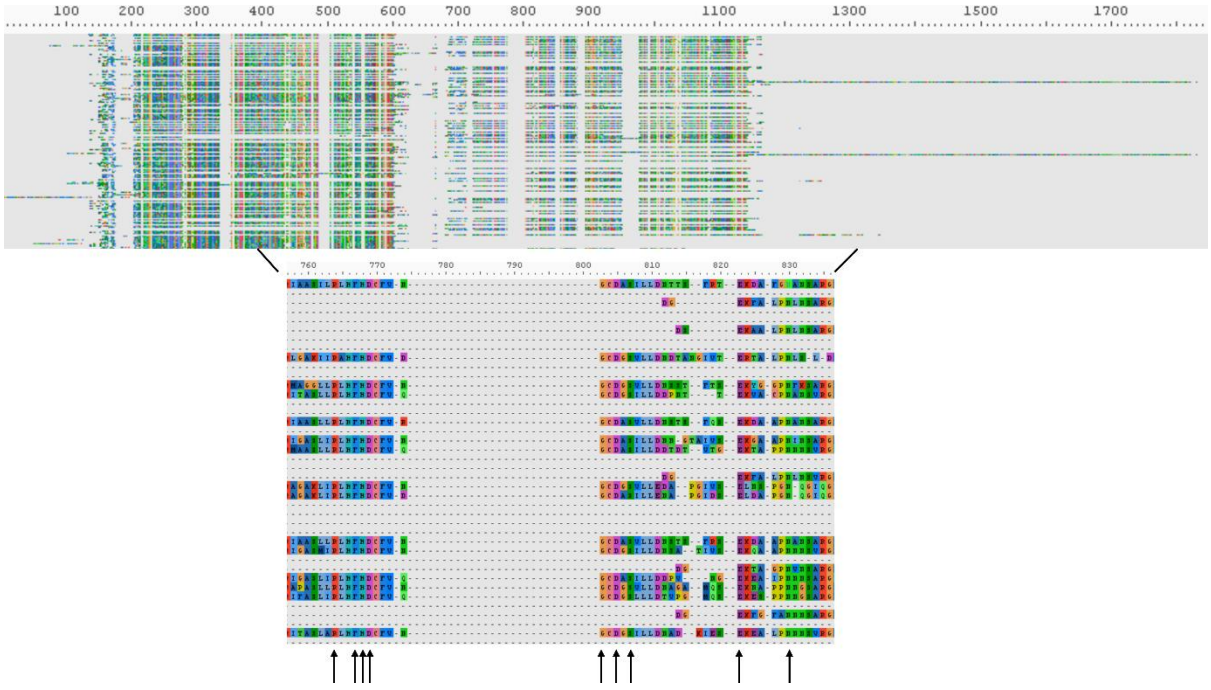
| | | | |
|---|---|---|---|
| **MtPrx71** | Eudicot | Core rosids | *Medicago truncatula* |
| **NnPrx41** | Eudicot | Proteales | *Nelumbo nucifera* |
| **NnPrx54** | Eudicot | Proteales | *Nelumbo nucifera* |
| **NnPrx55** | Eudicot | Proteales | *Nelumbo nucifera* |
| **NnPrx84** | Eudicot | Proteales | *Nelumbo nucifera* |
| **NtPrx30-1A** | Eudicot | Asterids | *Nicotiana tabacum* |
| **OsPrx39** | Monocot | Commelinids | *Oryza sativa ssp 144ormosa144 cv Nipponbare* |
| **PabPrx05** | Gymnosperm | Pinaceae | *Picea abies* |
| **PabPrx163** | Gymnosperm | Pinaceae | *Picea abies* |
| **PabPrx164** | Gymnosperm | Pinaceae | *Picea abies* |
| **PabPrx188** | Gymnosperm | Pinaceae | *Picea abies* |
| **PabPrx40** | Gymnosperm | Pinaceae | *Picea abies* |
| **PpePrx22** | Eudicot | Core rosids | *Prunus persica* |
| **PpePrx79** | Eudicot | Core rosids | *Prunus persica* |
| **PsPrx13** | Eudicot | Core rosids | *Pisum sativum* |
| **PsPrx14** | Eudicot | Core rosids | *Pisum sativum* |
| **PtaPrx125** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtaPrx137** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtaPrx144** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtaPrx20** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtaPrx33** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtaPrx66** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtaPrx68** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtaPrx83** | Gymnosperm | Pinaceae | *Pinus taeda* |
| **PtPrx100** | Eudicot | Core rosids | *Populus trichocarpa* |
| **PtPrx101** | Eudicot | Core rosids | *Populus trichocarpa* |
| **PtPrx102** | Eudicot | Core rosids | *Populus trichocarpa* |
| **RcPrx16** | Eudicot | Core rosids | *Ricinus communis* |
| **RcPrx27** | Eudicot | Core rosids | *Ricinus communis* |
| **RsPrx01** | Eudicot | Core rosids | *Raphanus sativus* |
| **SaPrx07** | Eudicot | Asterids | *Striga asiatica* |
| **SaPrx58** | Eudicot | Asterids | *Striga asiatica* |
| **ScucPrx01** | Fern | Polypodiidae | *Salvinia cucullata* |
| **ShPrx06-1B** | Eudicot | Core rosids | *Stylosanthes humilis* |
| **SmPrx07-2_11** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SmPrx31-2_131** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SmPrx33-1_84** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SmPrx34-1_84** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |

| | | | |
|---|---|---|---|
| **SmPrx36-2_44** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SmPrx37-2_44** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SmPrx50-1_126** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SmPrx56-2_33** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SmPrx75-2_233** | Lycophyte | Selaginellales | *Selaginella moellendorffii* |
| **SoPrx15** | Eudicot | Caryophyllales | *Spinacia oleracea* |
| **SpolPrx22** | Monocot | Alismatales | *Spirodela polyrhiza* |
| **ThasPrx11** | Eudicot | Core rosids | *Tarenaya hassleriana* |
| **ThasPrx17** | Eudicot | Core rosids | *Tarenaya hassleriana* |
| **ThasPrx31** | Eudicot | Core rosids | *Tarenaya hassleriana* |
| **ThasPrx36** | Eudicot | Core rosids | *Tarenaya hassleriana* |
| **ThasPrx49** | Eudicot | Core rosids | *Tarenaya hassleriana* |
| **ThasPrx60** | Eudicot | Core rosids | *Tarenaya hassleriana* |
| **ThasPrx69** | Eudicot | Core rosids | *Tarenaya hassleriana* |
| **TsPrx59** | Eudicot | Core rosids | *Thellungiella salsuginea* |
| **VvPrx17** | Eudicot | Core rosids | *Vitis vinifera* |
| **VvPrx34** | Eudicot | Core rosids | *Vitis vinifera* |
| **VvPrx36** | Eudicot | Core rosids | *Vitis vinifera* |
| **W9QDB2** | Eudicot | Core rosids | *Morus notabilis* |
| **W9RTT4** | Eudicot | Core rosids | *Morus notabilis* |
| **W9S432** | Eudicot | Core rosids | *Morus notabilis* |
| **W9SE23** | Eudicot | Core rosids | *Morus notabilis* |
| **WmPrx08** | Gymnosperm | Gnetales | *Welwitschia mirabilis* |
| **ZmaPrx47** | Monocot | Alismatales | *Zostera marina* |
| **ZmPrx31** | Monocot | Commelinids | *Zea mays* |

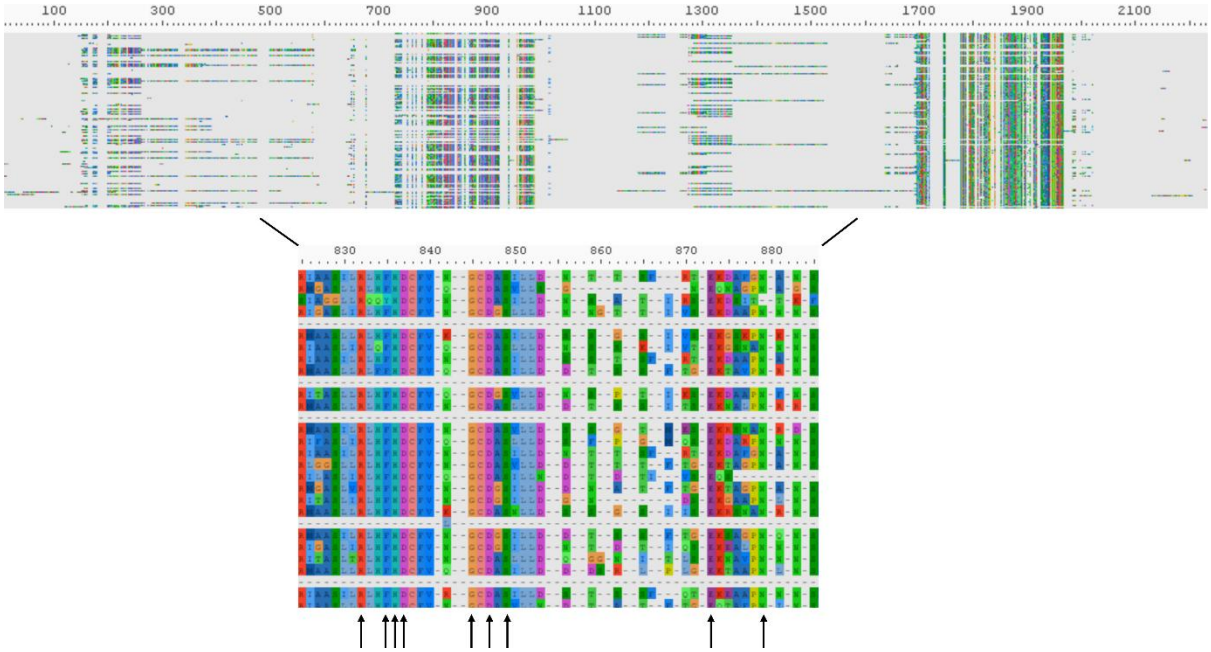
**Suppl. figure 3.2:** BAli-Phy MSA. Different colours represent different amino acids. Arrows indicate some sections where MSA failed to align conserved residues for some sequences.



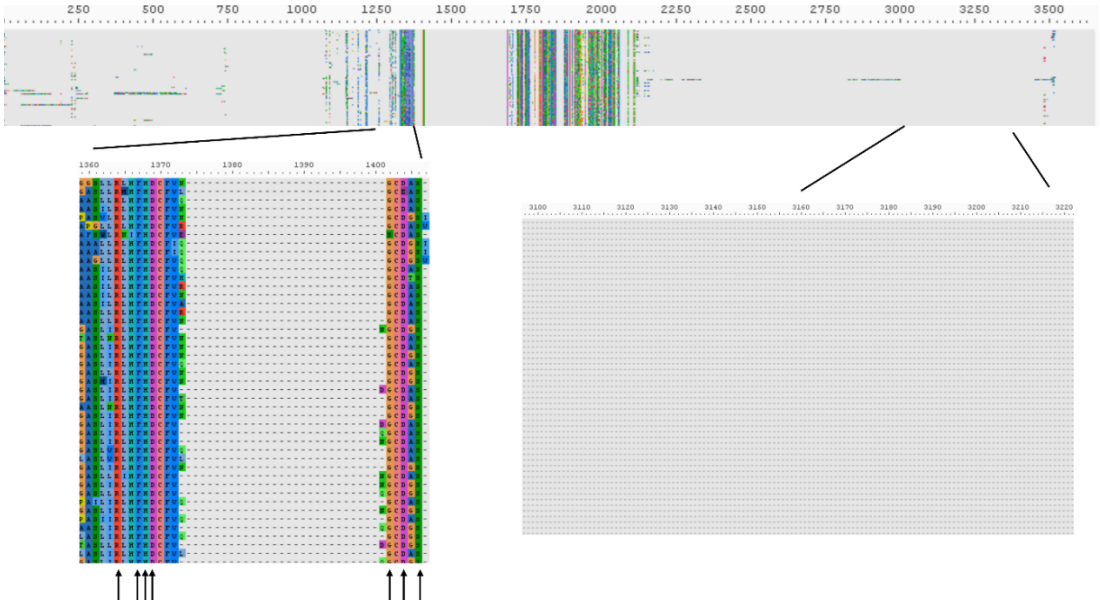

**Suppl. figure 3.3:** Clustal Omega MSA. Different colours represent different amino acids. Arrows indicate some sections where MSA failed to align conserved residues for some sequences.



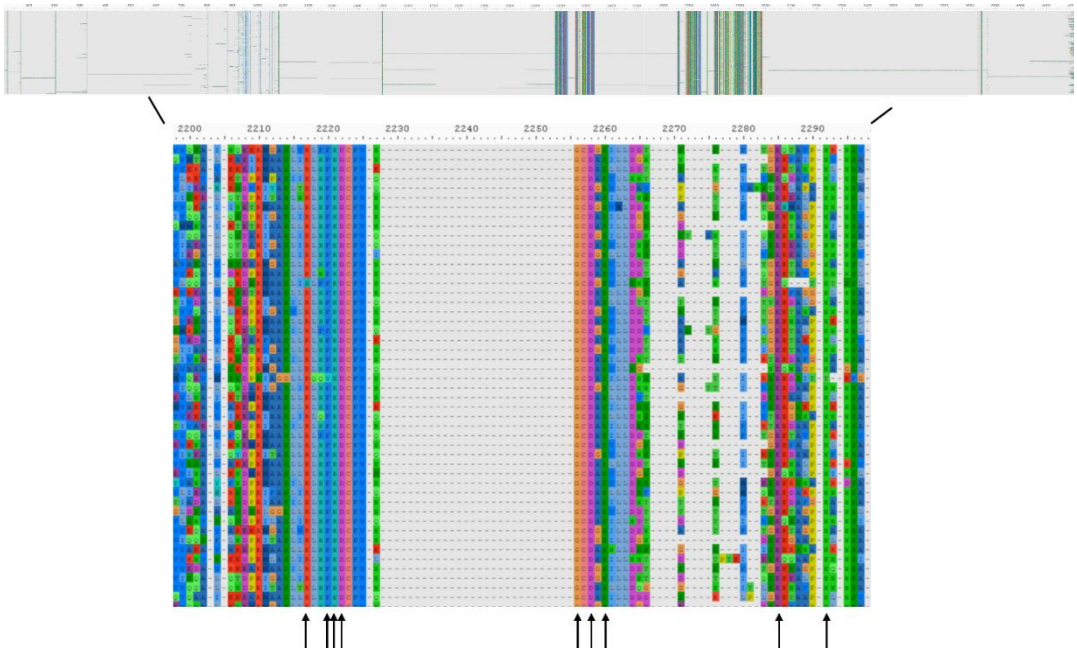
**Suppl. figure 3.4***:* DialignTX MSA. Different colours represent different amino acids. Arrows indicate some sections where MSA failed to align conserved residues for some sequences.



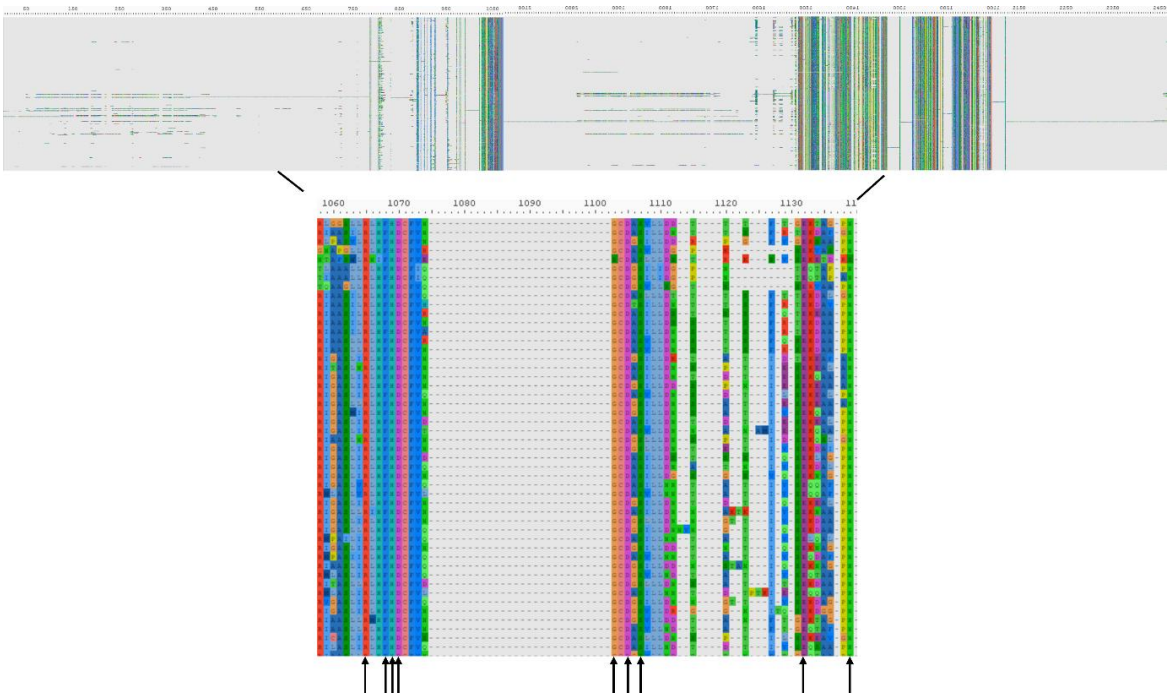

**Suppl. figure 3.5***:* Combined MSA. Different colours represent different amino acids. Arrows indicate some sections where MSA aligned conserved residues. On the bottom right is an indication of a section where the MSA contained no residues in any of the columns, likely due to indel handling issues.



**Suppl. figure 3.6:** ProbCons MSA. Different colours represent different amino acids. Arrows indicate some sections where MSA aligned conserved residues.




**Suppl. figure 3.7***:* MAFFT DASH MSA. Different colours represent different amino acids. Arrows indicate some sections where MSA aligned conserved residues.



**Suppl. table 3.2***:* List of IDs of the sequences removed during phylogeny construction.

| **MAFFT DASH-based** | | | **ProbCons-based** | | |
|---|---|---|---|---|---|
| | | | | | |
| **ID** | **Clade** | **Species of origin** | **ID** | **Clade** | **Species of origin** |
| **a0a0c9rws7** | gymnosperm | *Wollemia nobilis* | A0A0D6QT46 | gymnosperm (araucaliales) | *Araucaria cunninghamii* |
| **A0A0D6QT46** | gymnosperm | *Araucaria cunninghamii* | A0A1S3TC61 | core rosid | *Vigna radiata* |
| **A0A0D6R4B9** | gymnosperm | *Araucaria cunninghamii* | A0A1U7Z6F6 | proteales | *Nelumbo nucifera* |
| **A0A164ZWV6** | asterid | *Daucus carota* | A0A2I4FS82 | core rosid | *juglans regia* |
| **A0A2I4FS82** | core rosid | *Juglans regia* | A0A2I4HS78 | core rosid | *juglans regia* |
| **A0A2I4FS99** | core rosid | *Juglans regia* | A0A5B6UNN7 | core rosid | *Gossypium australe* |
| **A0A4Y7KWG0** | ranunculid | *Papaver somniferum* | A0A5C7HJB2 | core rosid | *Acer yangbiense* |
| **A0A5B6UNN7** | core rosid | *Gossypium australe* | A0A5N6L6C2 | core rosid | *Carpinus fangiana* |
| **A0A5C7HRP9** | core rosid | *Acer yangbiense* | A0A660KKX8 | core rosid | *Carpinus fangiana* |
| **A0A5N6L6C2** | core rosid | *Carpinus fangiana* | A0A6I9RJ83 | monocot commelinid | *Elaeis guineensis* |
| **A0A6A3CD50** | core rosid | *Hibiscus syriacus* | A0A6I9S386 | monocot commelinid | *Elaeis guineensis* |
| **A0A6I9RJ83** | monocot commelinid | *Elaeis guineensis* | A0A6I9S9K4 | monocot commelinid | *Elaeis guineensis* |
| **A0A6P4ATK4** | core rosid | *Ziziphus jujuba* | AtPrx36 | core rosid | *Arabidopsis thaliana* |
| **A0A6P4D4Q3** | core rosid | *Arachis duranensis* | BvPrx21 | caryophyllales | *Beta vulgaris* |

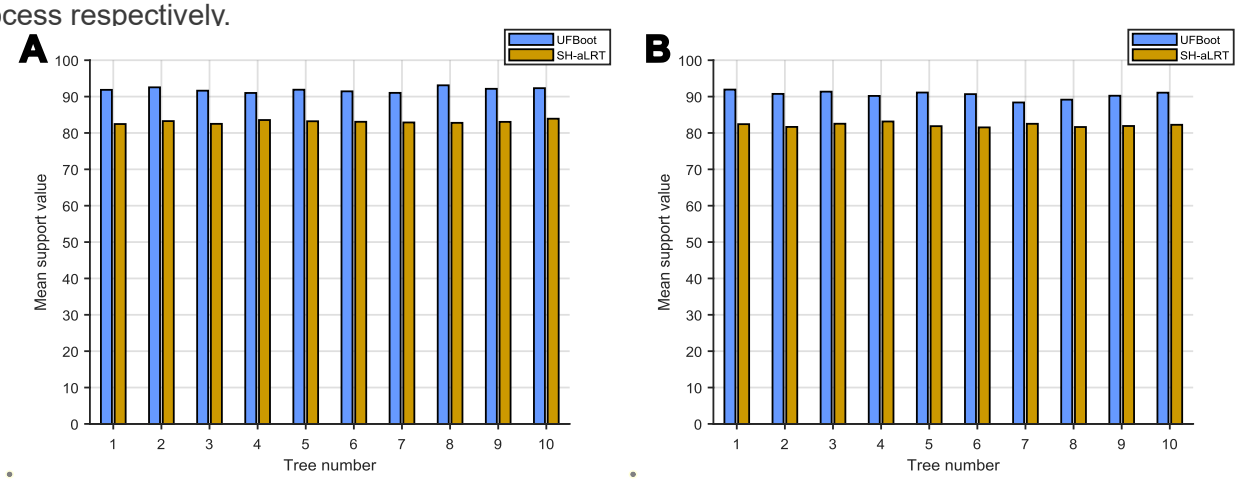
| | | | | | |
|---|---|---|---|---|---|
| **A0A7J7CNR9** | core rosid | *Tripterygium wilfordii* | CclPrx111 | core rosid | *Citrus clementina* |
| **A0A7J7CTD7** | core rosid | *Tripterygium wilfordii* | GbPrx63 | gymnosperm | *Gingko biloba* |
| **A0A7J8XWZ9** | core rosid | *Gossypium aridum* | KfedPrx38 | saxifragales | *Kalanchoe fedtschenkoi* |
| **A9PD65** | core rosid | *Populus trichocarpa* | PabPrx164 | gymnosperm | *Picea abies* |
| **AfpPrx06** | ranunculid | *Aquilegia 150ormosa x Aquilegia pubescens* | ThasPrx60 | core rosid | *Tarenaya hasslerian a* |
| **AtPrx36** | core rosid | *Arabidopsis thaliana* | VvPrx17 | core rosid | *Vitis vinifera* |
| **AtrPrx08** | ANA grade | *Amborella trichopoda* | VvPrx34 | core rosid | *Vitis vinifera* |
| **AtrPrx48** | ANA grade | *Amborella trichopoda* | W9RTT4 | core rosid | *Morus notabilis* |
| **B9GYK2** | core rosid | *Populus trichocarpa* | | | |
| **BvPrx21** | caryophyllales | *Beta vulgaris* | | | |
| **CclPrx30** | core rosid | *Citrus clementina* | | | |
| **CpapPrx05** | core rosid | *Carica papaya* | | | |
| **CpapPrx06** | core rosid | *Carica papaya* | | | |
| **EferPrx131** | ANA grade | *Euryale ferox* | | | |
| **EferPrx31** | ANA grade | *Euryale ferox* | | | |
| **EglPrx137** | core rosid | *Eucalyptus globulus* | | | |
| **EgrPrx191** | core rosid | *Eucalyptus grandis* | | | |
| **GbPrx63** | gymnosperm | *Ginkgo biloba* | | | |
| **HaPrx101** | asterid | *Helianthus annuus* | | | |
| **KfedPrx38** | saxifragales | *Kalanchoe fedtschenkoi* | | | |
| **KfedPrx58** | saxifragales | *Kalanchoe fedtschenkoi* | | | |
| **M5B2X1** | gymnosperm | *Chamaecyparis obtusa* | | | |

| | | | | | |
|---|---|---|---|---|---|
| **PpePrx22** | core rosid | *Prunus persica* | | | |
| **RcPrx16** | core rosid | *Ricinus communis* | | | |
| **SmPrx07-2_11** | lycophyte | *Selaginella moellendorffii* | | | |
| **SmPrx75-2_233** | lycophyte | *Selaginella moellendorffii* | | | |
| **ThasPrx60** | core rosid | *Tarenaya hassleriana* | | | |
| **VvPrx17** | core rosid | *Vinis vitifera* | | | |
| **VvPrx34** | core rosid | *Vitis vinifera* | | | |
| **W9QDB2** | core rosid | *Morus notabilis* | | | |

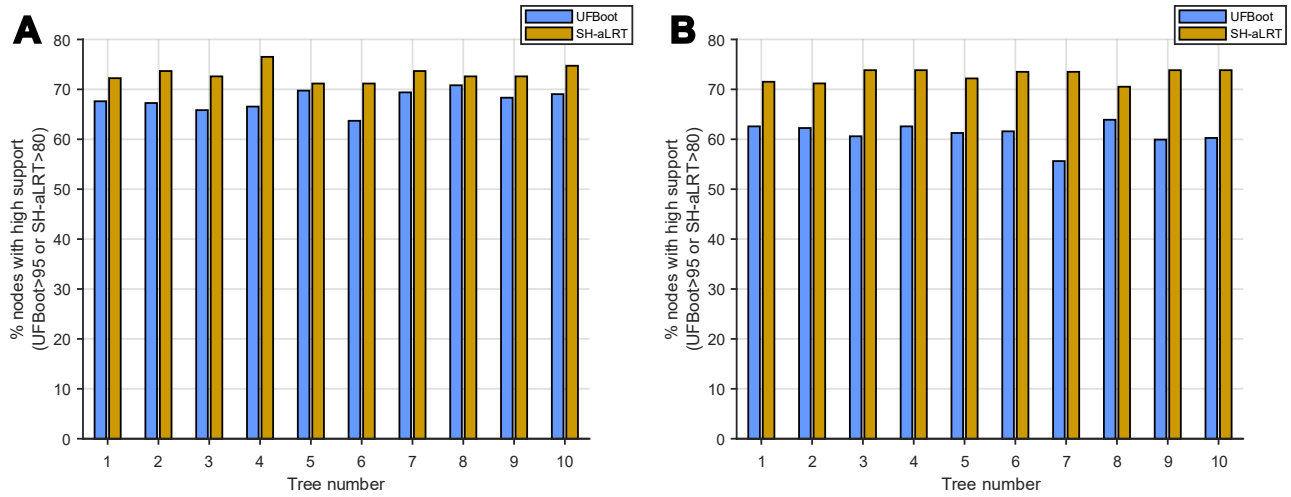
**Suppl. table 3.3:** Models used during the final set of phylogeny construction as predicted by the corrected Akaike information criterion in ModelFinder[103].

| **MAFFT DASH-based** | | **ProbCons-based** | |
|---|---|---|---|
| **Phylogeny number** | **Model** | **Phylogeny number** | **Model** |
| **1** | WAG+F+I+I+R8 | 1 | VT+R10 |
| **2** | WAG+F+I+I+R8 | 2 | VT+R10 |
| **3** | WAG+F+I+I+R10 | 3 | VT+I+I+R9 |
| **4** | WAG+F+R9 | 4 | VT+I+I+R8 |
| **5** | WAG+F+I+I+R10 | 5 | VT+R10 |
| **6** | WAG+F+I+I+R8 | 6 | VT+I+I+R10 |
| **7** | WAG+F+I+I+R8 | 7 | VT+I+I+R10 |
| **8** | WAG+F+I+I+R10 | 8 | VT+I+I+R10 |
| **9** | WAG+F+I+I+R8 | 9 | VT+R10 |
| **10** | WAG+F+I+I+R8 | 10 | VT+I+I+R10 |

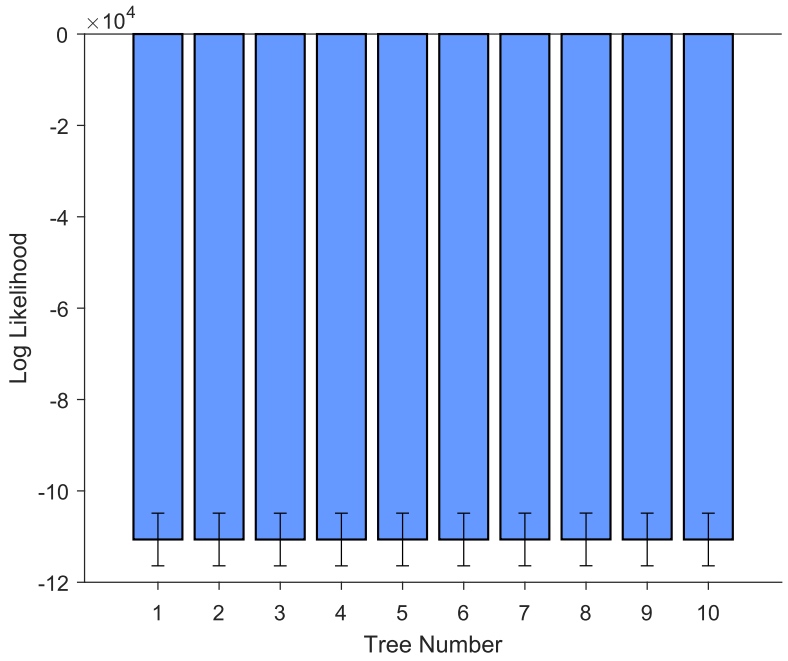
**Suppl. figure 3.8:** Mean UFBoot and SH-aLRT values for each of the 10 phylogenies reconstructed during A) MAFFT DASH MSA-based process and B) ProbCons MSA-based process respectively.



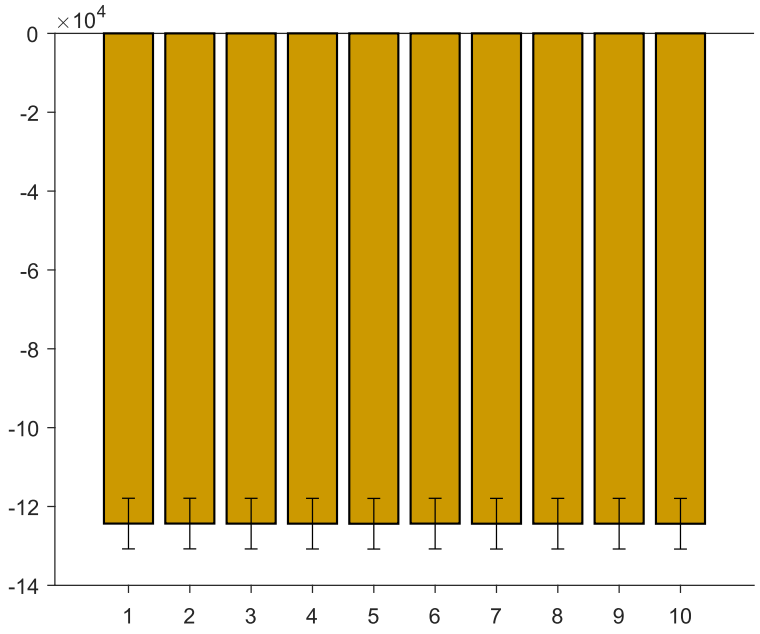

**Suppl. figure 3.9:** Percentage of nodes with high support values per UFBoot (>95) or SH-aLRT (>80) for each of the 10 phylogenies reconstructed during A) MAFFT DASH MSA-based process and B) ProbCons MSA-based process respectively.



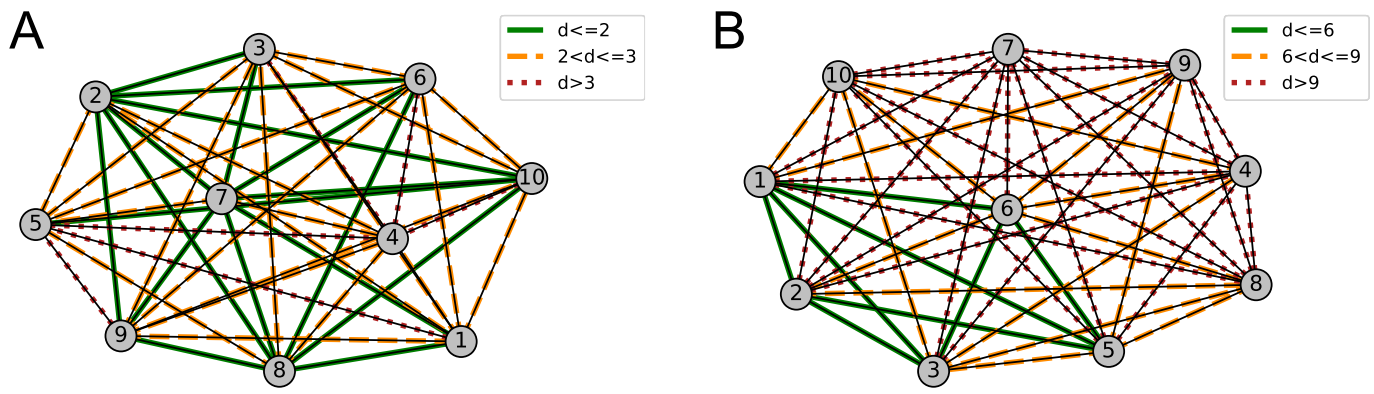
**Suppl. figure 3.10:** Comparison of the log likelihoods of the ML trees for the MAFFT DASH MSA-based phylogeny reconstruction. Error bars represent standard error.




**Suppl. figure 3.11:** Comparison of the log likelihoods of the ML trees for the ProbCons MSA-based phylogeny reconstruction. Error bars represent standard error.



**Suppl. figure 3.12:** Network diagrams representing the weighted Robinson-Foulds distance between the final 10 phylogenies, where distance (d) is shown as different coloured edges. A) 10 final phylogenies based on the MAFFT DASH MSA. B) 10 final phylogenies based on the ProbCons MSA.




**Suppl. table 3.4:** Pairwise tree distance metrics for MAFFT DASH-based phylogenies in final set of 10.

| Tree1 | Tree2 | RFWeighted (0.5) | Quartet | Path Difference | RF(0.5) | Matching Triplet | Matching Split | UMAST | GeoUnrooted |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2.2 | 1509187 | 350 | 30 | 167352 | 310 | 25 | 0.55 |
| 1 | 3 | 2.8 | 1460706 | 279 | 35 | 141780 | 201 | 22 | 0.66 |
| 1 | 4 | 3.0 | 20456041 | 739 | 31 | 650149 | 426 | 39 | 0.75 |
| 1 | 5 | 3.1 | 1267438 | 229 | 31 | 136288 | 107 | 21 | 0.59 |
| 1 | 6 | 2.6 | 1412597 | 213 | 32 | 140638 | 157 | 24 | 0.63 |
| 1 | 7 | 1.8 | 1245664 | 219 | 23 | 117022 | 172 | 14 | 0.50 |
| 1 | 8 | 1.4 | 211780 | 113 | 15 | 35930 | 35 | 8 | 0.40 |
| 1 | 9 | 2.2 | 1302672 | 326 | 25 | 135898 | 287 | 15 | 0.58 |
| 1 | 10 | 2.5 | 1288642 | 273 | 33 | 134194 | 121 | 27 | 0.60 |
| 2 | 3 | 2.0 | 339983 | 232 | 21 | 58772 | 159 | 15 | 0.52 |
| 2 | 4 | 2.6 | 19608342 | 794 | 28 | 571129 | 398 | 34 | 0.72 |
| 2 | 5 | 2.9 | 1028666 | 334 | 25 | 135797 | 291 | 22 | 0.53 |
| 2 | 6 | 2.0 | 2246818 | 431 | 22 | 216908 | 369 | 26 | 0.53 |
| 2 | 7 | 0.8 | 264085 | 211 | 9 | 51444 | 144 | 13 | 0.26 |
| 2 | 8 | 1.8 | 1304605 | 336 | 20 | 134193 | 285 | 20 | 0.42 |
| 2 | 9 | 0.8 | 208197 | 182 | 9 | 32572 | 51 | 12 | 0.37 |
| 2 | 10 | 2.0 | 1014561 | 327 | 24 | 137286 | 289 | 17 | 0.50 |
| 3 | 4 | 3.5 | 19623244 | 757 | 36 | 575916 | 357 | 38 | 0.84 |
| 3 | 5 | 2.8 | 931301 | 234 | 27 | 96803 | 174 | 15 | 0.60 |
| 3 | 6 | 2.4 | 2279270 | 332 | 28 | 204513 | 266 | 22 | 0.62 |
| 3 | 7 | 1.7 | 220799 | 185 | 18 | 25623 | 55 | 10 | 0.50 |
| 3 | 8 | 2.1 | 1256915 | 259 | 28 | 108368 | 172 | 16 | 0.59 |
| 3 | 9 | 2.2 | 277807 | 240 | 20 | 45055 | 168 | 11 | 0.59 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 10 | 2.8 | 1201810 | 299 | 35 | 146022 | 218 | 23 | 0.66 |
| 4 | 5 | 4.0 | 20141279 | 735 | 36 | 637101 | 447 | 46 | 0.80 |
| 4 | 6 | 3.3 | 21260507 | 789 | 34 | 728488 | 523 | 46 | 0.81 |
| 4 | 7 | 2.4 | 19567335 | 753 | 24 | 570259 | 326 | 35 | 0.70 |
| 4 | 8 | 2.9 | 20347463 | 736 | 29 | 636170 | 419 | 41 | 0.74 |
| 4 | 9 | 2.7 | 19534631 | 805 | 25 | 555375 | 383 | 34 | 0.76 |
| 4 | 10 | 3.6 | 20375673 | 743 | 39 | 676995 | 471 | 46 | 0.81 |
| 5 | 6 | 3.0 | 1560393 | 249 | 27 | 160677 | 168 | 24 | 0.56 |
| 5 | 7 | 2.7 | 910598 | 245 | 23 | 103122 | 191 | 18 | 0.50 |
| 5 | 8 | 2.4 | 1065096 | 212 | 25 | 104661 | 90 | 19 | 0.51 |
| 5 | 9 | 3.1 | 967606 | 351 | 25 | 122108 | 304 | 19 | 0.59 |
| 5 | 10 | 1.9 | 273865 | 180 | 13 | 51182 | 62 | 12 | 0.38 |
| 6 | 7 | 1.7 | 2066755 | 288 | 19 | 183896 | 245 | 20 | 0.50 |
| 6 | 8 | 1.9 | 1203611 | 184 | 21 | 106937 | 128 | 18 | 0.52 |
| 6 | 9 | 2.1 | 2123763 | 402 | 21 | 202497 | 352 | 21 | 0.58 |
| 6 | 10 | 2.2 | 1665963 | 296 | 27 | 177207 | 174 | 24 | 0.54 |
| 7 | 8 | 1.3 | 1042198 | 197 | 14 | 84439 | 149 | 10 | 0.35 |
| 7 | 9 | 0.7 | 57570 | 147 | 4 | 19996 | 119 | 3 | 0.33 |
| 7 | 10 | 2.0 | 1109856 | 295 | 25 | 138078 | 201 | 24 | 0.51 |
| 8 | 9 | 1.7 | 1099206 | 313 | 16 | 103341 | 264 | 11 | 0.47 |
| 8 | 10 | 1.9 | 1262118 | 259 | 24 | 137557 | 104 | 23 | 0.49 |
| 9 | 10 | 2.4 | 1166864 | 395 | 27 | 157059 | 312 | 25 | 0.59 |

**Suppl. table 3.5:** Pairwise tree distance metrics for ProbCons-based phylogenies in final set of 10.

| RefTree | Tree | RFWeighted (0.5) | Quartet | Path Difference | RF(0.5) | Matching Triplet | Matching Split | UMAST | GeoUnrooted |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 5 | 3.8797 | 738250 | 358.6419 | 34 | 115315 | 189 | 33 | 0.9035 |
| 1 | 6 | 4.8071 | 1955768 | 318.8354 | 26 | 101638 | 197 | 21 | 0.9622 |
| 1 | 3 | 5.1632 | 763825 | 357.4717 | 33 | 103577 | 155 | 23 | 1.1139 |
| 1 | 2 | 5.973 | 6547841 | 395.0924 | 42 | 343351 | 232 | 40 | 1.2869 |
| 1 | 10 | 7.9277 | 3206092 | 448.7783 | 43 | 138789 | 244 | 26 | 1.4889 |
| 1 | 9 | 8.4703 | 42695942 | 711.1863 | 61 | 1393895 | 742 | 100 | 1.549 |
| 1 | 8 | 9.2068 | 8152269 | 468.647 | 61 | 364650 | 370 | 53 | 1.6727 |
| 1 | 4 | 9.5004 | 38320674 | 604.685 | 48 | 1266274 | 672 | 89 | 1.5753 |
| 1 | 7 | 10.861 | 14333776 | 671.8988 | 70 | 597332 | 554 | 66 | 1.9245 |
| 2 | 3 | 5.5124 | 5954856 | 218.055 | 25 | 266571 | 127 | 21 | 1.2088 |
| 2 | 5 | 5.6947 | 6538385 | 347.5917 | 43 | 330943 | 203 | 34 | 1.3165 |
| 2 | 6 | 7.7292 | 7954664 | 318.4839 | 50 | 370500 | 281 | 40 | 1.3778 |
| 2 | 8 | 8.1484 | 6157514 | 365.1712 | 56 | 329391 | 274 | 51 | 1.5174 |
| 2 | 9 | 9.332 | 37654424 | 664.3147 | 66 | 1206933 | 650 | 85 | 1.7651 |
| 2 | 7 | 10.6224 | 8678294 | 548.9135 | 70 | 405147 | 472 | 61 | 1.814 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 10 | 10.6923 | 9321564 | 488.9581 | 65 | 429604 | 352 | 48 | 1.7837 |
| 2 | 4 | 10.7293 | 36965602 | 586.0512 | 64 | 1315875 | 645 | 94 | 1.6191 |
| 3 | 6 | 4.429 | 2162716 | 276.2535 | 37 | 130381 | 204 | 24 | 1.1258 |
| 3 | 5 | 6.3246 | 912369 | 348.9842 | 43 | 126944 | 176 | 27 | 1.2695 |
| 3 | 8 | 6.9578 | 7708647 | 369.3102 | 49 | 325803 | 317 | 45 | 1.3926 |
| 3 | 10 | 7.7033 | 3611610 | 489.6162 | 53 | 184674 | 271 | 30 | 1.5955 |
| 3 | 4 | 8.9254 | 38553473 | 595.1353 | 59 | 1296997 | 682 | 93 | 1.6059 |
| 3 | 7 | 9.5233 | 14066804 | 552.6427 | 64 | 582056 | 489 | 64 | 1.7654 |
| 3 | 9 | 10.2297 | 42864336 | 718.3913 | 69 | 1416265 | 739 | 95 | 1.778 |
| 4 | 6 | 7.0479 | 37751836 | 478.3472 | 37 | 1216346 | 510 | 74 | 1.4319 |
| 4 | 10 | 7.8033 | 37771720 | 571.8496 | 44 | 1258549 | 579 | 81 | 1.7069 |
| 4 | 8 | 9.3408 | 32554572 | 575.8941 | 59 | 1181033 | 573 | 94 | 1.5763 |
| 4 | 7 | 10.4173 | 38260704 | 790.8856 | 67 | 1302674 | 769 | 110 | 1.677 |
| 4 | 5 | 10.5883 | 38556148 | 618.2912 | 57 | 1312058 | 684 | 91 | 1.6877 |
| 4 | 9 | 11.1196 | 23196250 | 579.8327 | 59 | 758111 | 497 | 75 | 1.7525 |
| 5 | 6 | 5.7356 | 2157042 | 329.8333 | 33 | 136849 | 208 | 29 | 1.1228 |
| 5 | 9 | 7.4708 | 42179311 | 647.6836 | 48 | 1328166 | 645 | 82 | 1.5119 |
| 5 | 8 | 8.3588 | 7769259 | 375.6115 | 52 | 315011 | 307 | 47 | 1.6065 |
| 5 | 10 | 9.0635 | 3442446 | 463.0421 | 52 | 176029 | 273 | 35 | 1.6152 |
| 5 | 7 | 12.2716 | 14684279 | 662.4983 | 82 | 655452 | 591 | 81 | 2.0221 |
| 6 | 10 | 6.4315 | 4498252 | 376.9138 | 37 | 150075 | 117 | 18 | 1.5392 |
| 6 | 8 | 8.5299 | 9426029 | 404.8975 | 55 | 367726 | 364 | 53 | 1.6396 |
| 6 | 9 | 8.6231 | 41854238 | 629.9254 | 51 | 1368613 | 617 | 92 | 1.5646 |
| 6 | 7 | 10.5838 | 15715214 | 635.8978 | 71 | 615128 | 549 | 65 | 1.8757 |
| 7 | 8 | 9.3264 | 7695357 | 568.695 | 69 | 427143 | 487 | 62 | 1.6595 |
| 7 | 9 | 10.9743 | 37195460 | 847.2981 | 78 | 1299031 | 789 | 107 | 1.8826 |
| 7 | 10 | 12.2576 | 16099050 | 754.9874 | 77 | 655760 | 608 | 69 | 2.0596 |
| 8 | 9 | 9.4922 | 37372126 | 638.2241 | 64 | 1233437 | 638 | 89 | 1.7197 |
| 8 | 10 | 10.6303 | 9895841 | 552.9032 | 67 | 410969 | 437 | 62 | 1.8541 |
| 9 | 10 | 10.9802 | 43121216 | 602.0166 | 61 | 1401535 | 638 | 95 | 1.8398 |

**Suppl. table 3.6:** Comparison of UltraFast Bootstrap and SH-approximate likelihood ratio test values for the final selected ProbCons-based phylogeny and MAFFT DASH-based phylogeny respectively

| | ProbCons-based final phylogeny | MAFFT DASH-based final phylogeny |
|---|---|---|
| Ufbootstrap mean | 90.68 | 91.02 |

| | | |
|---|---|---|
| SH-aLRT mean | 81.53 | 82.89 |
| aBayes mean | 0.9449 | 0.9503 |
| % branches with very good Ufbootstrap (>95) | 62% | 69% |
| % branches with good Ufbootstrap (>80) | 84% | 82% |
| % branches with very good SH-aLRT (>80) | 74% | 74% |
| % branches with very good aBayes (>0.95) | 79% | 80% |

**Suppl. table 3.7:** Comparison of taxa and Robinsin Foulds weighted distance between the final selected MAFFT DASH-based and ProbCons-based phylogenies.

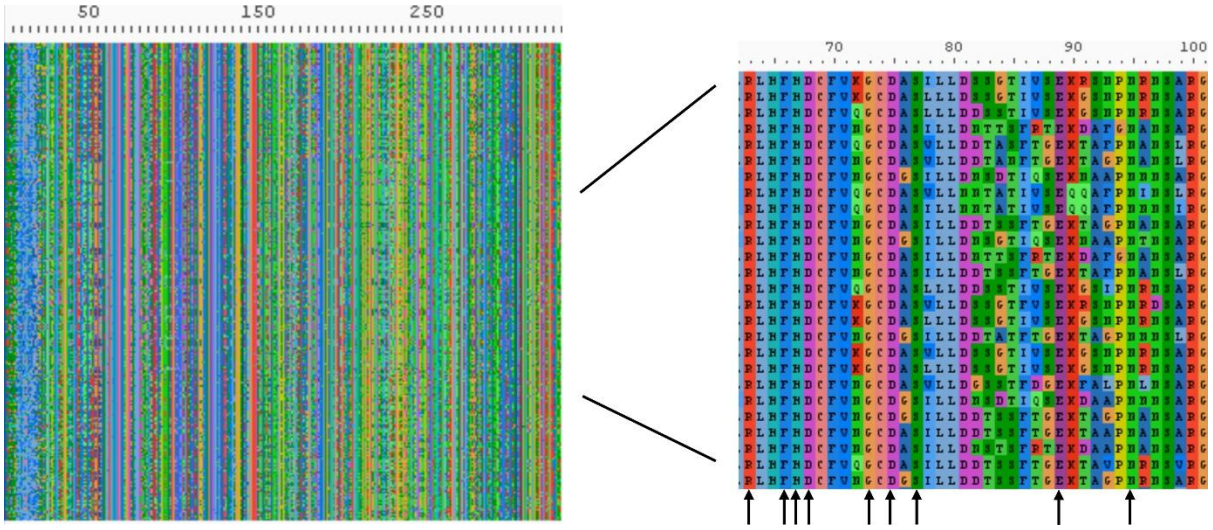
| Probcons Tree Taxa | Mafft dash Tree taxa | Common taxa | Robinsin Foulds weighted distance |
|---|---|---|---|
| 305 | 284 | 263 | 33.3 (cf. 1-10 for very similar trees) |

**Suppl. table 3.8:** Statistical supports for ancestors along the horseradish peroxidase lineage

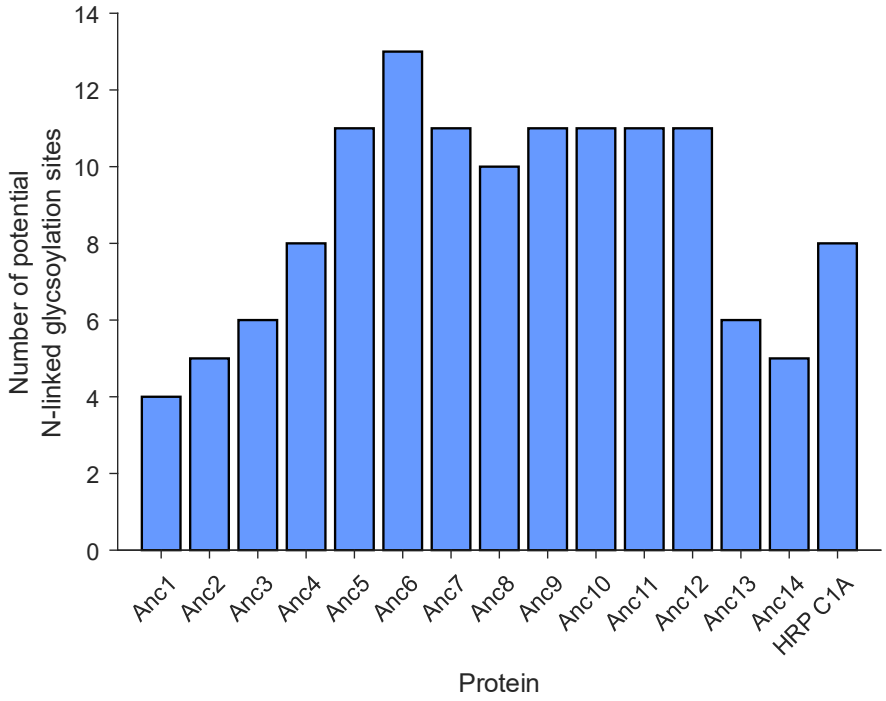
| Node number | Average ML reconstruction posterior probability | node SH-aLRT | node Ufboot |
|---|---|---|---|
| 285 (Anc1) | 0.840 | N/A (unrooted) | N/A (unrooted) |
| 286 (Anc2) | 0.837 | 96.4 | 100 |
| 287 | 0.870 | 59.2 | 99 |
| 288 | 0.884 | 72.9 | 99 |
| 289 (Anc3) | 0.919 | 86.4 | 93 |

| | | | |
|---|---|---|---|
| 290 | 0.919 | 52.9 | 82 |
| 291 (Anc4) | 0.915 | 92.6 | 98 |
| 292 (Anc5) | 0.948 | 99.9 | 100 |
| 293 (Anc6) | 0.960 | 92 | 94 |
| 294 | 0.965 | 72 | 90 |
| 295 | 0.972 | 81.8 | 68 |
| 296 | 0.978 | 88.4 | 75 |
| 297 | 0.981 | 76.7 | 64 |
| 298 | 0.974 | 92.2 | 78 |
| 299 | 0.974 | 90.4 | 70 |
| 300 (Anc7) | 0.970 | 88.8 | 98 |
| 301 | 0.965 | 91.6 | 83 |
| 302 (Anc8) | 0.957 | 87.7 | 95 |
| 303 | 0.957 | 98.4 | 99 |
| 304 (Anc9) | 0.975 | 97.1 | 99 |
| 305 | 0.977 | 71 | 89 |
| 306 | 0.977 | 92.2 | 98 |
| 307 (Anc10) | 0.978 | 88.7 | 98 |
| 308 (Anc11) | 0.970 | 85.2 | 95 |
| 309 (Anc12) | 0.954 | 93.3 | 99 |
| 310 | 0.947 | 100 | 100 |
| 311 (Anc13) | 0.957 | 84.1 | 100 |
| 312 | 0.951 | 96.4 | 100 |
| 313 | 0.944 | 97.7 | 100 |
| 314 | 0.953 | 78.3 | 100 |
| 315 (Anc15) | 0.969 | 99.7 | 100 |

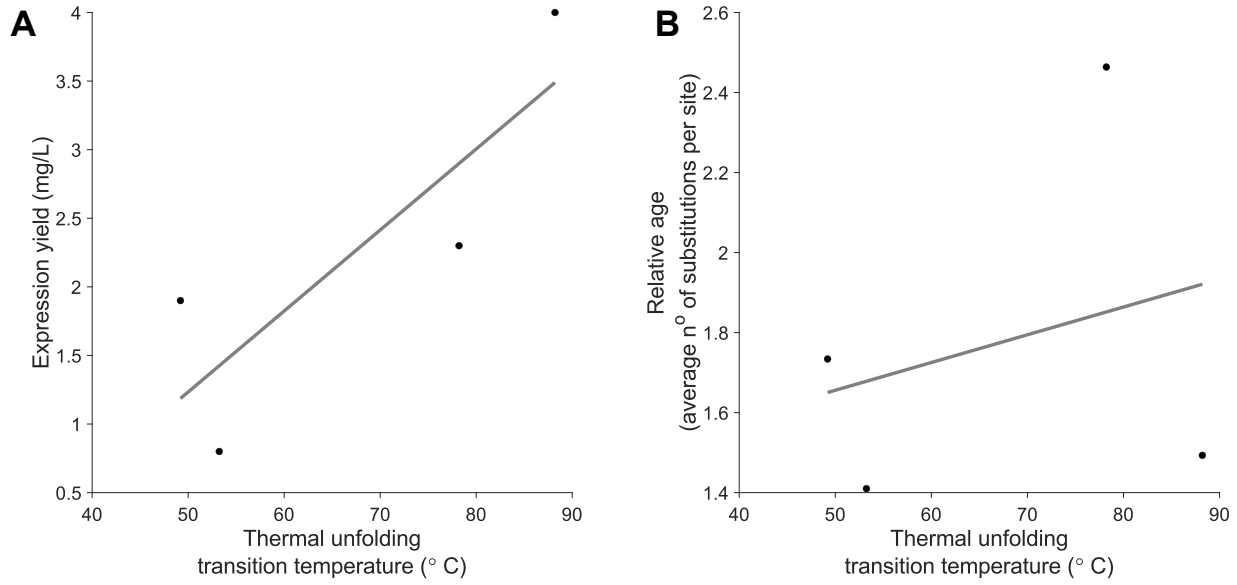
**Suppl. figure 3.13:** Maximum a posteriori ancestral sequences (MAFFT DASH-based reconstruction). Different colours represent different amino acids. On the right is shown a section with arrows indicating key functional peroxidase residues which are highly conserved in extant sequences.



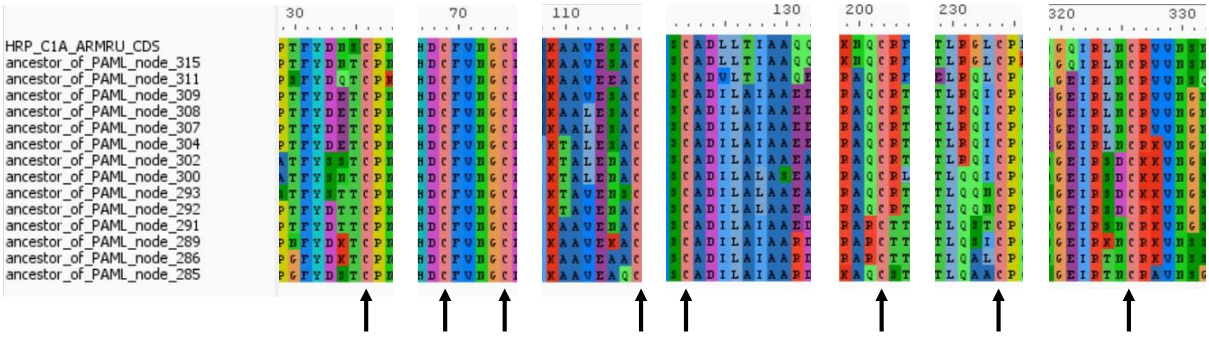

**Suppl. figure 3.14:** The number of potential sites of N-linked glycosylation (Asn-X-Ser/Thr motif[743,744]) for the selected ancestors. Note that HRP C1A has 9 potential N-linked glycosylation sites, but only 8 are actually glycosylated *in vivo*[603], so it is shown as having 8 sites.



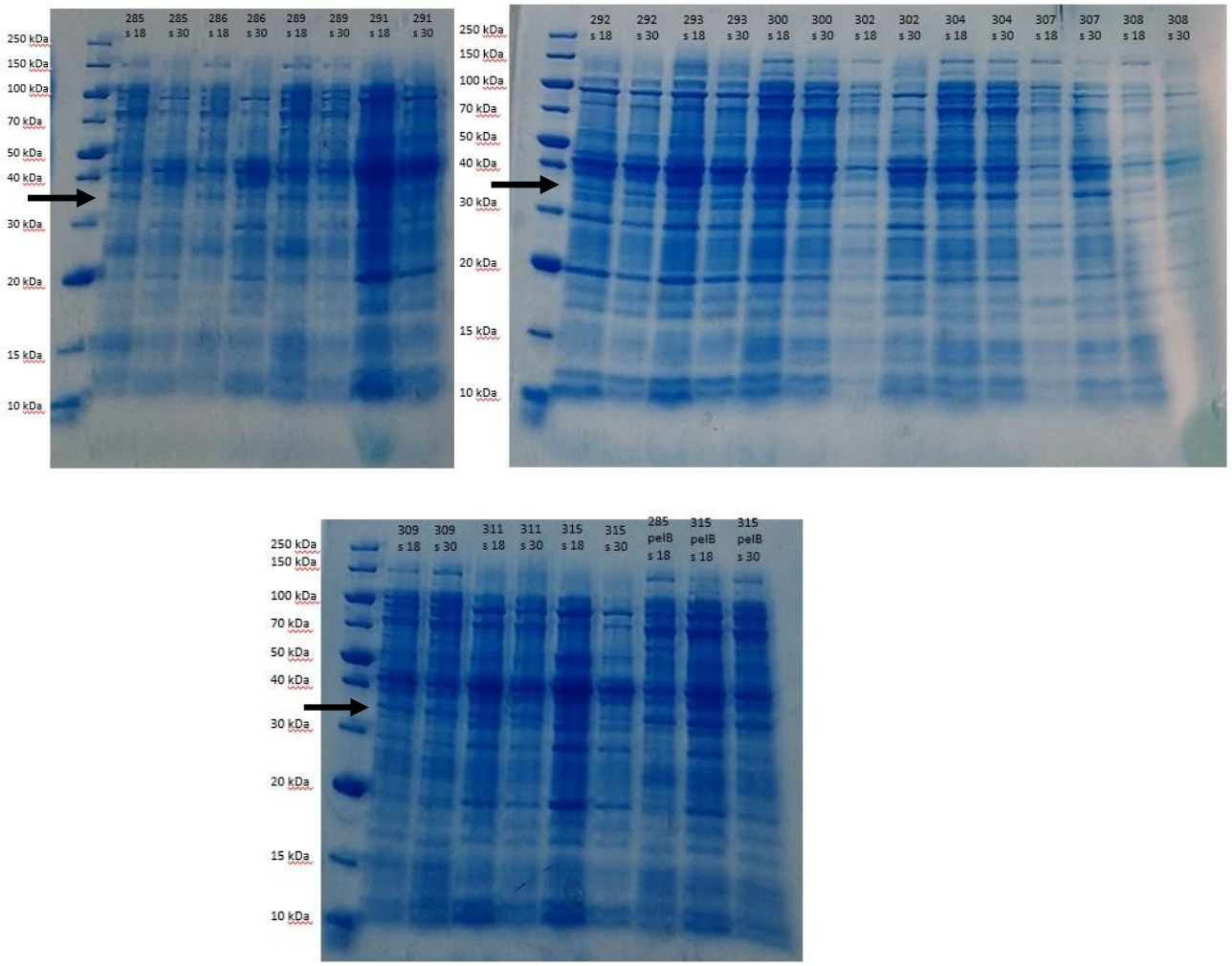
**Suppl. figure 3.15:** The thermal unfolding transition temperature of the selected ancestral peroxidases (Anc4, 7, 9, 11) correlates relatively strongly with expression yield (ρ = 0.84), but only weakly with relative ancestral age (ρ = 0.27).



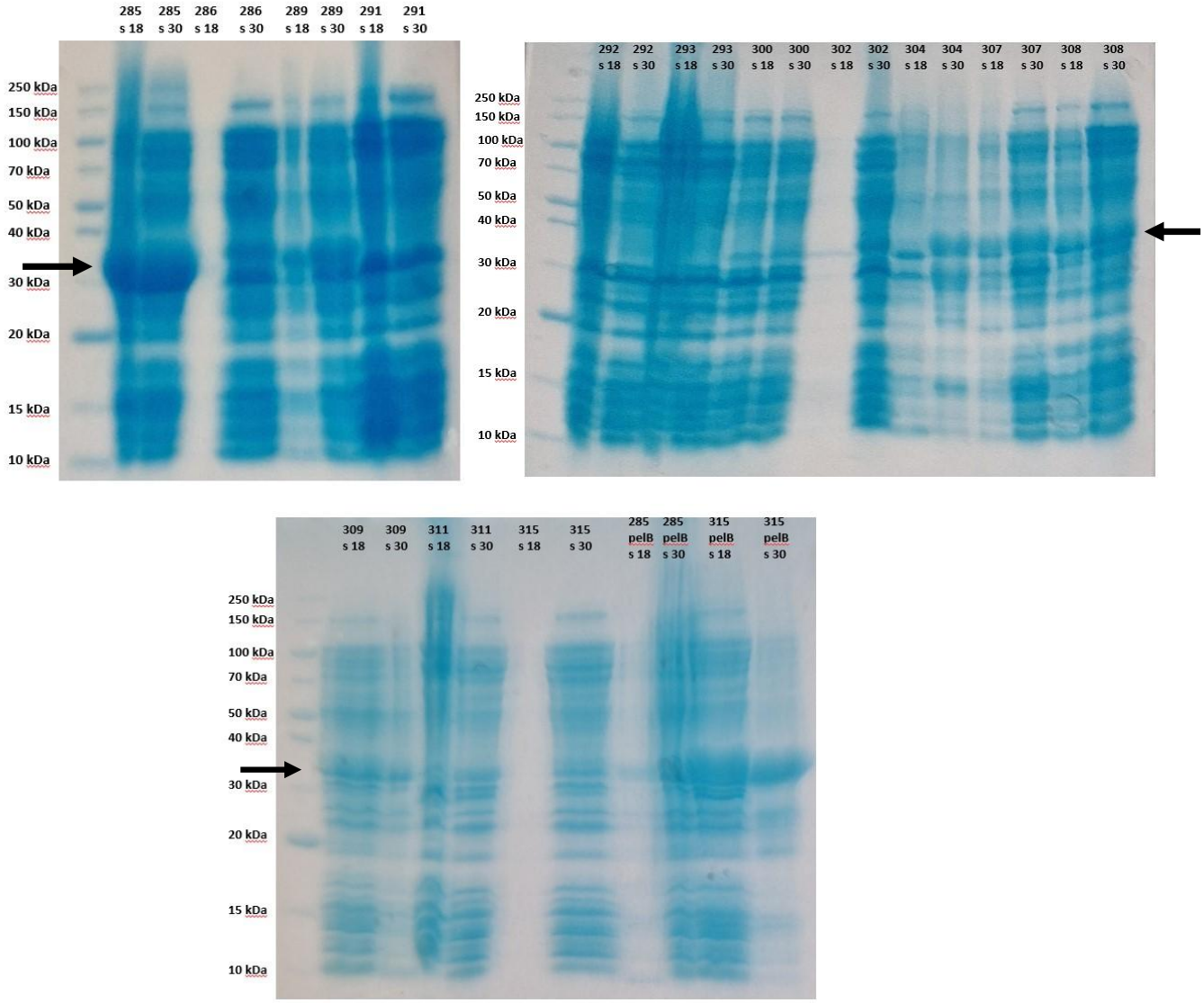

**Suppl. figure 3.16:** The ancestral peroxidase sequences contain cysteines at homologous positions to those known to form disulfide bonds in extant class III peroxidase members. Homologous cysteine positions are shown by the arrows.



**Suppl. figure 3.17:** SDS-PAGE of the soluble lysate fraction during expression trial for the selected ancestral peroxidases in SHuffle T7 Express. Arrow indicates bands of ~35 kDa, which is the expected molecular weight of the ancestral peroxidases. While many of the ancestors have a band of this molecular weight, there is no clear overexpression. Ancestors are number by their position, with 291 being Anc4, 300 being Anc7, 304 being Anc9 and 308 being Anc11.



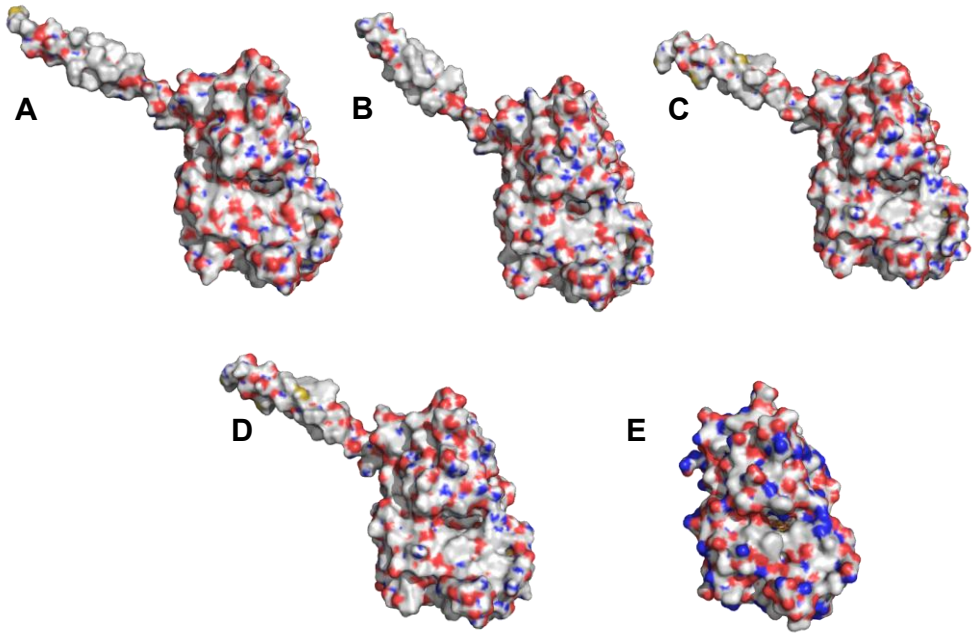
**Suppl. figure 3.18:** SDS-PAGE of the insoluble lysate fraction during expression trial for the selected ancestral peroxidases in SHuffle T7 Express. Arrow indicates bands of ~35 kDa, which is the expected molecular weight of the ancestral peroxidases. Some ancestors appear to have relatively large bands at this molecular weight, indicating relatively high amounts of insoluble expression. Ancestors are number by their position, with 291 being Anc4, 300 being Anc7, 304 being Anc9 and 308 being Anc11.



**Suppl. table 3.9:** RMSD values between predicted structures for the selected ancestors and the experimentally-determined structure for horseradish peroxidase (HRP) C1A (PDB: 1HCH). RMSD values were calculated using PyMol

| | | RMSD (Å) |
|---|---|---|
| **Anc4** | **HRP (C1A)** | 0.617 |
| **Anc7** | **HRP (C1A)** | 0.55 |
| **Anc9** | **HRP (C1A)** | 0.684 |
| **Anc11** | **HRP (C1A)** | 0.641 |
| **Anc4** | **Anc11** | 0.276 |
| **Anc7** | **Anc11** | 0.212 |
| **Anc9** | **Anc11** | 0.098 |
| **Anc4** | **Anc9** | 0.29 |
| **Anc7** | **Anc9** | 0.2 |
| **Anc4** | **Anc7** | 0.243 |

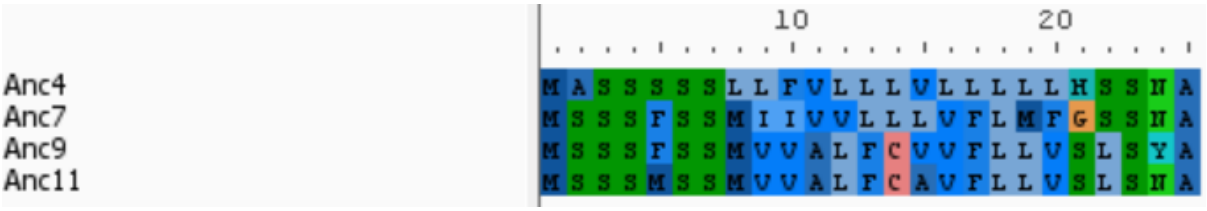
**Suppl. figure 3.19:** Surfaces of the predicted structures of A) Anc4, B) Anc7, C) Anc9, D) Anc11 and the experimentally determined structure of E) HRP C1A (PDB: 1HCH). The middle 'pocket' of each structure likely is the substrate access and product egress to the catalytic iron-heme centre.



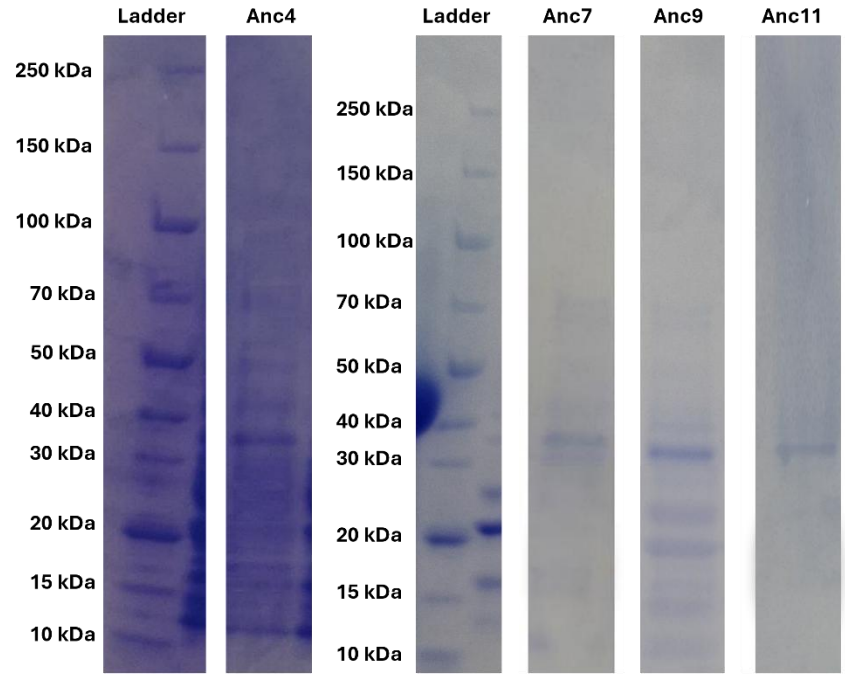
**Suppl. table 3.10:** Prediction of the probability that the sequences for Anc4, Anc7, Anc9 and Anc11 contain a signal peptide using SignalP 6.0

| | signal peptide probability (Sec/SPI) | cleavage site position | probability of cleavage site position |
|---|---|---|---|
| *Anc4* | 0.999802 | 25-26 | 0.9796 |
| *Anc7* | 0.999773 | 25-26 | 9.9768 |
| *Anc9* | 0.999794 | 25-26 | 0.9834 |
| *Anc11* | 0.999785 | 25-26 | 0.9825 |

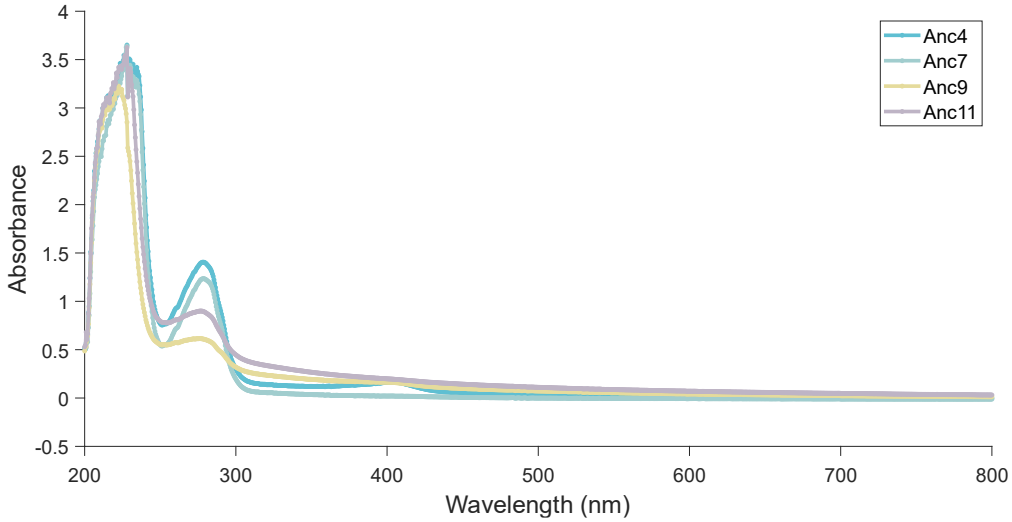
**Suppl. figure 3.20:** Sequences of the predicted signal peptide regions of Anc4, Anc7, Anc9 and Anc11.



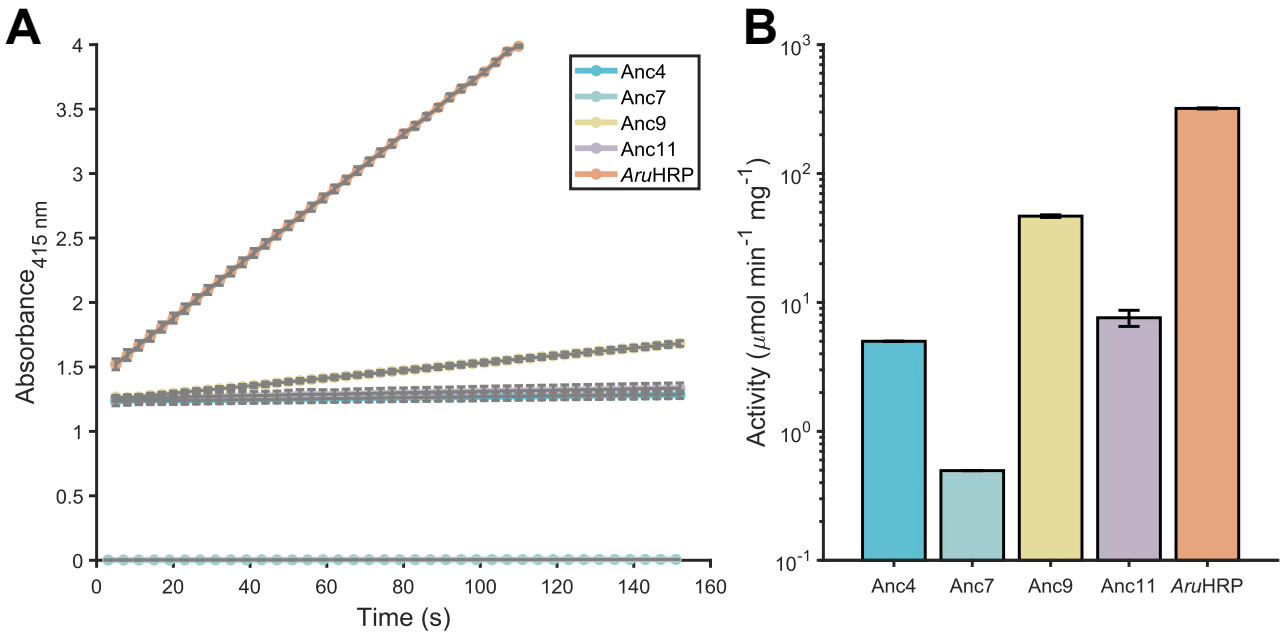

**Suppl. figure 3.21:** SDS-PAGE of purified Anc4, Anc7, Anc9 and Anc11. There is a band present at the expected molecular weight of ~36 kDa for each ancestor.



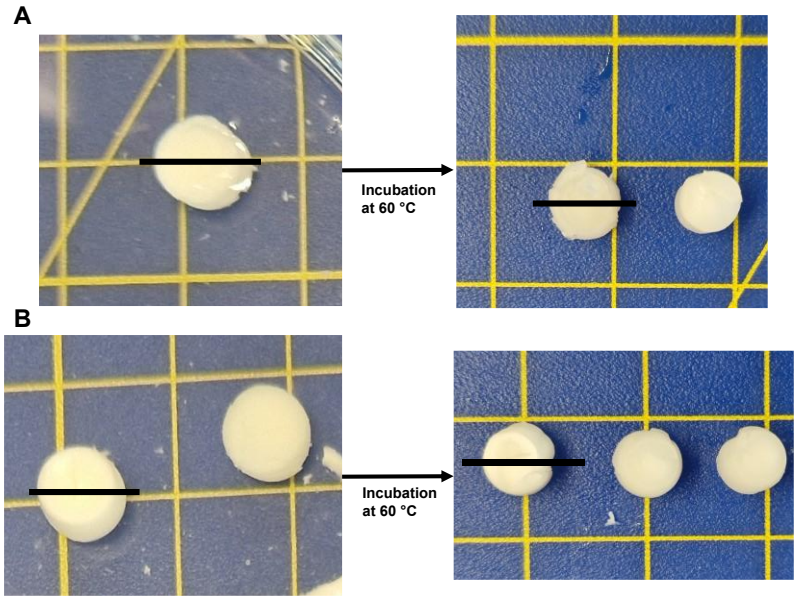
**Suppl. figure 3.22:** Absorbance spectra of Anc4, Anc7, Anc9 and Anc11 from 200-800 nm. Measured with a concentration of 2.3 mg/mL for Anc4, 1.9 mg/mL for Anc7, 0.18 mg/mL for Anc9 and 0.75 mg/mL for Anc11.



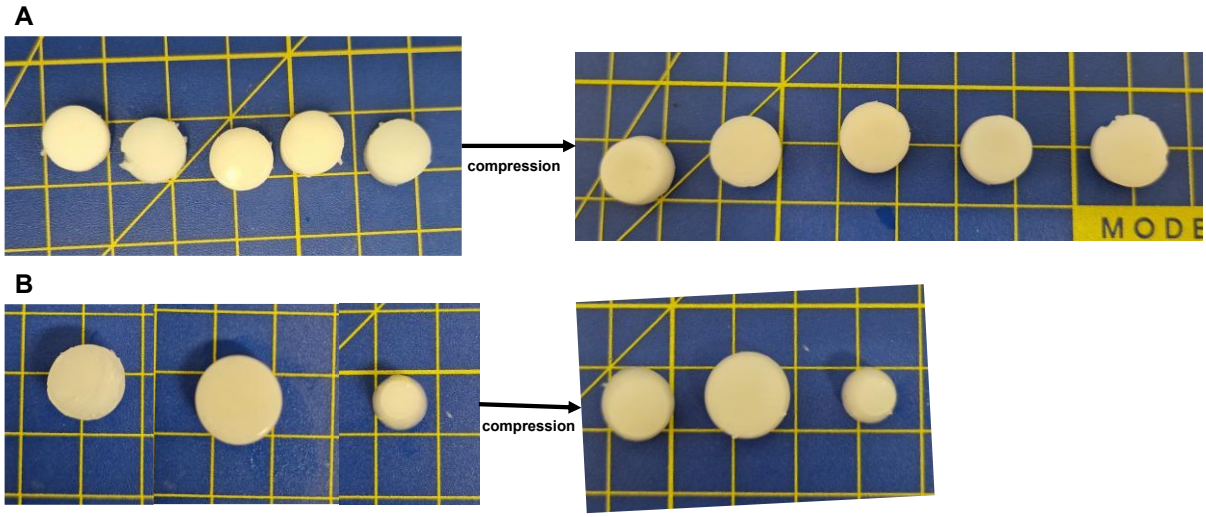

**Suppl. figure 3.23:** A) A high concentration of ABTS and low enzyme concentrations were used to determine the initial linear response of ABTS radical product increase over time in 50 mM potassium phosphate buffer, pH 6.0 at 25 °C. Error bars represent S.E.M. of 3 replicates. B) The calculated specific activities of the selected ancestral proteins and *Aru*HRP with ABTS. Error bars represent S.E.M. of 3 replicates. The specific activities of the ancestors were much lower than *Aru*HRP, such that a larger quantity of ancestral enzyme would be needed to replace *Aru*HRP.



**Suppl. figure 3.24***:* PNIPAm gels formed by A) HRP or B) Anc9 before and after incubation at 60 °C. Contraction was observed for both gels, indicating that PNIPAm had formed due to its relatively unique thermosensitivity property.




**Suppl. figure 3.25:** PNIPAm gels formed by A) HRP or B) Anc9 before and after the compressive test. Both the HRP and Anc9 formed gels showed high elastic recovery.



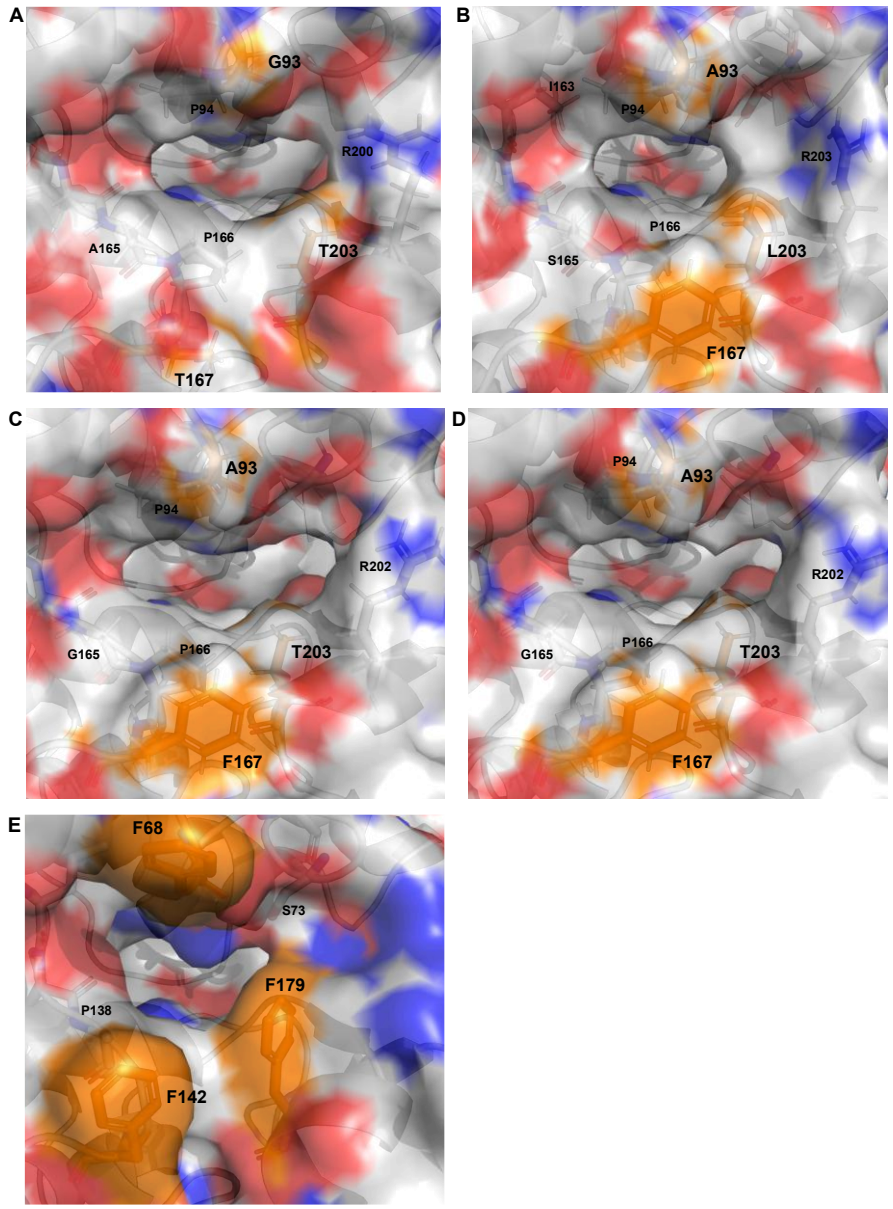
**Suppl. table 3.11:** Roughly estimated ages of Anc4, Anc7, Anc9 and Anc11 based on their respective most recent common ancestor of their respective extant plant taxa. This age is then used to find the approximate average global temperature. Overall, the rough predictions suggest that all of the selected ancestral peroxidases may have evolved during a relatively warm climatic period.

| Ancestral peroxidase | Most recent organism taxa ancestor | Approximate age (organism ancestor) | Approximate average global temperature predication (derived from Scotese *et al.*[728]) |
|---|---|---|---|
| Anc4 | Gymnosperm ancestor | ~380 mya[729] | 17-21 °C |
| Anc7 | Eudicot ancestor | ~125-128 mya[730,731] | 20-22 °C |
| Anc9 | Fagales ancestor | ~124 mya [745] | 20-22 °C |
| Anc11 | Ancestor of sapindales and malvales | ~104 mya[732] | 21-22 °C |
| HRP | N/A | >3000 ya and <0.2mya[733,734] | 14 °C |

**Suppl. table 3.12:** Pairwise sequence identities of selected ancestors and horseradish peroxidase.

| | Anc4 | Anc7 | Anc9 | Anc11 | HRP |
|---|---|---|---|---|---|
| Anc4 | - | | | | |
| Anc7 | 78.1 | - | | | |
| Anc9 | 76.5 | 84.3 | - | | |
| Anc11 | 75.2 | 79.7 | 93.1 | - | |
| HRP | 56.6 | 57 | 62.8 | 65.7 | - |

**Suppl. figure 3.26:** Differences in the entrances to the active sites of A) Anc4 B) Anc7 C) Anc9 D) Anc11 and E) HRP C1A (PDB: 1HCH). The entrance site of HRP is dominated by three phenylalanine residues, which are shown in orange. Equivalent residues in the ancestral peroxidases are also shown in orange. Other residues forming part of the entrances are labelled.



# Chapter 4: Development of a 3D-printable yeast 'on demand' bioproducer formed through enzyme-mediated radical polymerisation

## 4.1 Abstract

The ability to produce eukaryotic proteins in response to a desired signal or environmental cue could be useful for applications such as biopharmaceutics, bioremediation and 'on demand' chemical production. As engineered living materials (ELMs) contain living cells, they are able to act as 'on demand' bioproducers. However, despite 3D printing offering manufacturing flexibility, few 3D-printable ELMs have been developed, thereby limiting their future utility. Here we showed that polymerisation of the 3D-printable double network hydrogel alginate/PNIPAm could be mediated in a reproducible and non-toxic manner for ELM use by the novel bienzymatic system of unspecific peroxygenase PaDa-I and glucose oxidase. We then demonstrated that chemically induced 'on demand' production of both a reporter protein for biosensing applications and an enzyme for bioreactor applications could be achieved by living yeast cells within 3D-printed yeast alg/PNIPAm. These results highlight the ability of yeast ELMs to produce eukaryotic proteins when desired in a material that can be 3D-printed into custom shapes for diverse applications.

## 4.2 Introduction

Engineered living materials (ELMs) are a relatively novel class of materials which aim to incorporate cells into materials such that the materials possess the desirable properties of living organisms with the mechanical behaviour of traditional materials[295,381,385]. One property of living organisms which is desirable in a material is the ability for organisms to respond to environmental stimuli. For example, some animals will change their fur colour to white in winter in response to the low light to improve their camouflage in snow[746]. Similarly, it would be useful to have a material which would produce desired proteins in response to certain cues, for example, a drug delivery system which only produced and secreted a desired protein pharmaceutical at the site of injury[747-749] or a biosensor system which only produced a fluorescent signal protein if a biotoxin was present in the environment[750-752]. This property, where a material produces desired proteins 'on demand', can be driven by living cells within

an ELM due to the ability of cells to change gene expression in response to environmental cues. An example of such an ELM is that developed by Tay *et al.*[388], where mercury-absorbing curli fibres were only produced by *E. coli* if a certain concentration of mercury was present in the environment. Other examples of 'on demand' bioreactors include the ELMs developed by Sugianto *et al.*[403] where it was shown that engineered *E. coli* in F127 bis-urethane methacrylate could produce a biosynthetic cascade of enzymes capable of producing pyruvoyl tetrahydropterin under a regulatory circuit, and additionally that *S. cerevisiae* in BSA-PEGDA could be induced to produce either proteinase A enzyme or an enzyme system for production of betaxanthins. Relatively few ELM bioreactors have been developed which utilise 3D-printing[393,399,753]. Although the *S. cerevisiae* BSA-PEGDA system[403,754] was 3D-printed, it utilised stereolithography which is relatively expensive and limited in the materials it can print compared to extrusion 3D-printing[373,755,756]. Notwithstanding, it has been demonstrated that extrusion printing can be used to form ELMs in a few studies[757-760]. Additionally, it was recently demonstrated that an extra dimension of behaviour could be added to such an ELM bioreactor by using the 3D printed thermoresponsive double network alginate/poly(*N*-isopropylacrylamide) (alg/PNIPAm)[361]. This study by Klemperer *et al.* was also one of very few to utilise enzyme-mediated radical polymerisation during the formation of the ELM, which is advantageous over the more commonly used UV/chemical initiator process because the enzyme-mediated reaction is non-toxic, can occur under mild conditions, and the enzymes are biodegradable[5,294]. Therefore, it would be beneficial to develop a thermoresponsive, 3D-printable alg/PNIPAm ELM that forms through enzyme-mediated radical polymerisation and only produces proteins when desired in response to an input signal (Fig. 4.1).

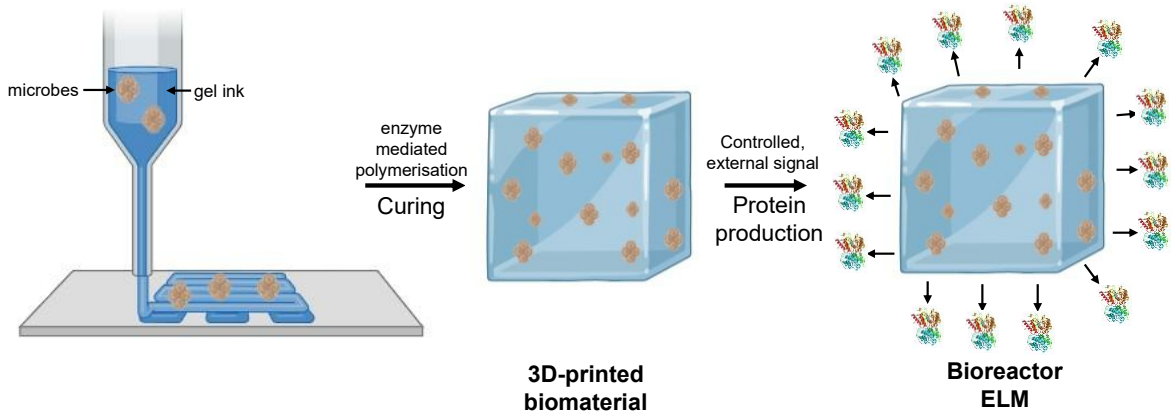


**Figure 4.1:** The aim of this project is to develop a 3D-printable engineered living material biogel which could be formed 'on demand' by utilising the ability of cells to secrete proteins in response to environmental signals. Figure made using BioRender and PyMOL.

Alg/PNIPAm is a double network hydrogel composed of both alginate and PNIPAm polymer networks, ideally in an entangled state to form an interpenetrating network[761] (Fig. 4.2B). The advantages of using alg/PNIPAm to develop a hydrogel that can form 'on demand' include that it has been found to be relatively biocompatible[762], that PNIPAm can form through enzyme-mediated radical polymerisation (EMRP)[361], that alg/PNIPAm is thermosensitive, which increases its usability in several applications like in soft robotics[370,763], that alg/PNIPAm is 3D-printable[361], and that alg/PNIPAm has relatively high mechanical strength[370,764]. While single network PNIPAm can be formed through EMRP and is thermosensitive and biocompatible, it is not 3D-printable, has a very slow volume transition, and has relatively weak mechanical strength[765,766]. Therefore, an alg/PNIPAm biomaterial is more attractive for future applications.

The alginate network of alg/PNIPAm is a natural polysaccharide composed of 1,4-linked β-D-mannuronate (M) and α- L -guluronate units (G) in MM, GG or alternating M and G blocks[767]. Alginate can be crosslinked using positively charged ions because the carboxyl groups of the M and G units are negatively charged at neutral or basic pHs[768]. The PNIPAm network is an acrylamide-based polymer which undergoes a volume phase transition at temperatures above its lower critical solution temperature, which is typically ~32 °C. PNIPAm is composed of N-isopropylacrylamide (NIPAm) chains crosslinked with *N,N′*-Methylenebisacrylamide (MBA)[621]. Due to the free radical polymerisation mechanism of PNIPAm, the initiation of its polymerisation is able to be catalysed using horseradish peroxidase and glucose oxidase[361] (Fig. 4.2C). The alg/PNIPAm double network can be formed by first polymerising the PNIPAm network in the presence of alginate, then later incubating the hydrogel in an ion solution to crosslink the alginate[764,769,770]. However, as this is time consuming, it has also been shown that alg/PNIPAm can be formed by crosslinking both the PNIPAm and alginate networks at the same time[367,771,772]. Klemperer *et al*.[361] showed that enzyme-mediated polymerisation of PNIPAm within alg/PNIPAm could occur using the method where both networks are formed at the same time, which is the most attractive for an alg/PNIPAm ELM.

Regardless of the formation method, the relative mechanical strength and elasticity of alg/PNIPAm double networks has been found to derive from the use of the ionically

crosslinked alginate with the more rigid, covalently crosslinked PNIPAm[370,773]. The mechanical strength has been found to partly derive from the alginate network being able to 'sacrifice' its ionic bonds to dissipate energy when under stress, whilst the PNIPAm network retains its structure and can 'bridge' the cracks. However, unlike other double networks consisting of two covalent networks, because the ionic bonds in the alginate network are physical, they are able to be re-formed. This thereby makes alg/PNIPAm relatively less brittle and able to recover elastically.

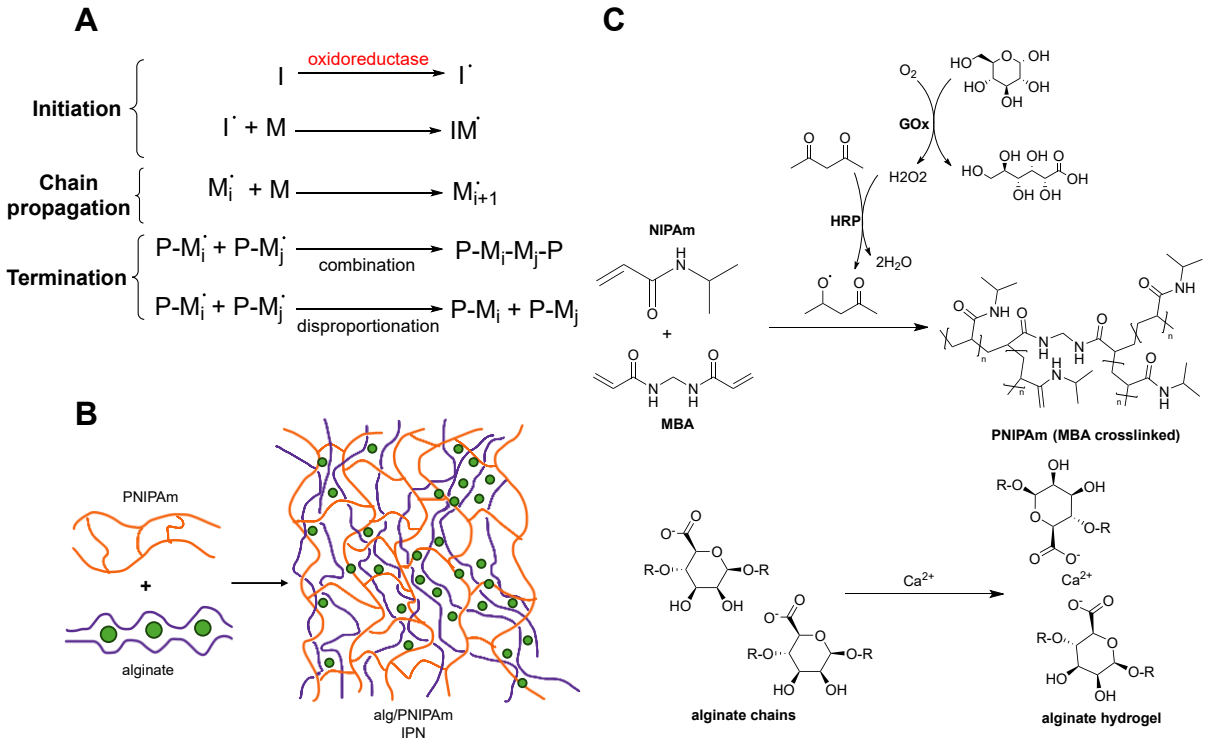



**Figure 4.2:** Polymerisation of materials using enzymes. A) Oxidoreductase enzymes can catalyse the initiation stage of chain growth radical polymerisation reactions by catalysing the production of a free radical initiator (I). This radical is then able to form a free radical of some monomers (M), e.g. vinyl monomers. Polymerisation proceeds through chain propagation, usually radical addition, before terminating either through combination or disproportionation processes (polymer-monomer is represented as P-M). B) Alginate and PNIPAm polymers can be combined under certain conditions to form an interpenetrating double network. Calcium(II) ions are represented by green spheres. C) Polymerisation of PNIPAm can be catalysed using a cascade of horseradish peroxidase (HRP) and glucose oxidase (GOx). Polymerisation of alginate chains into a network can be conducted using calcium(II) ions to form ionic crosslinks. Figure made using ChemDraw and Microsoft PowerPoint.

One method that can be used to cure alg/PNIPAm in a relatively non-toxic and 'green' manner compared to traditional initiators is enzyme-mediated polymerisation using

oxidoreductases[294,316,361]. Many oxidoreductase enzymes can catalyse single electron transfers from an electron donor substrate to an electron acceptor substrate, thereby resulting in the production of free radicals, which can initiate the polymerisation of vinyl monomers *via* chain growth addition (Fig. 4.2A). Advantages of utilising enzymes to catalyse polymerisation of an ELM include that they require relatively mild conditions, are non-toxic, and are biodegradable[294,298,316]. Previously, commonly used enzymes in enzyme-mediated polymerisation systems have mainly included peroxidases and oxidases, such as horseradish peroxidase (HRP), chloroperoxidase, glucose oxidase and laccase[294,316]. However, dual enzyme systems have become more attractive recently in order to both deoxygenate the reaction mixture, thereby allowing for improved radical polymerisation under atmospheric conditions, and initiate polymerisation[296]. These dual enzyme systems typically utilise an oxidase, which scavenges oxygen and catalyses *in situ* $H_2O_2$ production, and a peroxidase, which utilises the $H_2O_2$ to catalyse the formation of free radicals. For example, the bienzymatic cascade of glucose oxidase and horseradish peroxidase has been used under atmospheric conditions to polymerise materials including polyphenol[296], poly (3,4-ethylenedioxythiophene) (PEDOT)[774] and the gelatin-based hydrogel (gelatin-poly(ethylene glycol)-tyramine[775].

Previously, the peroxidase enzyme used to catalyse the polymerisation of alg/PNIPAm was commercial plant-derived HRP[361], which has considerable batch-to-batch variation in its composition[603]. This variation can cause irreproducibility in the formation of alg/PNIPAm between each batch of HRP. To reduce this, it would be best to use a single isoform of HRP purified recombinantly; however, it has proven difficult to recombinantly express HRP in a soluble, active form[644,660]. Whilst in Chapter 3 it was shown that predicted ancestral peroxidases could be expressed in a soluble, active form in *E. coli* (Chapter 3.3.6), their yields were still relatively low. In contrast, Molina-Espeja *et al.*[290] were able to utilise directed evolution to engineer an enzyme with peroxidase activity to have high levels of recombinant expression in GRAS yeast strains[293]. This enzyme was termed unspecific peroxygenase PaDa-I (UPO) and was engineered from an *Agrocybe aegerita* unspecific peroxygenase. Although unspecific peroxygenases possess the ability to catalyse both peroxidase and oxygenation reactions[776], most research focuses on their oxygenation abilities[777-779]. However, they still possess relatively proficient peroxidase activity[293], and unlike many peroxidases including HRP, UPO is able to be recombinantly expressed in an active, soluble form with relatively high yields[293,780].

Whilst both HRP (EC 1.11.1.7) and UPO (EC 1.11.2.1) have peroxidase activity, HRP falls under the peroxidase-catalase superfamily whereas UPO is part of the peroxidase-peroxygenase superfamily[781]. The latter family, which appears to primarily be present in fungi, is relatively unique in that its members possess a range of catalytic activity including 'classical' peroxidase activity, peroxygenase activity, and haloperoxidase activity (Fig. 4.3A). UPO itself can catalyse both peroxidative and peroxygenative reactions, and has a fairly broad substrate range[779]. It has also been shown to catalyse a wide range of peroxygenative reactions, including aromatic hydroxylation, aliphatic hydroxylation and alkene epoxidation[782]. While the native role of unspecific peroxygenases has not been fully established, it has been linked to lignin and humus transformation[783]. The structure of the engineered UPO is fairly typical of a heme-thiolate peroxidase, consisting primarily of 10 α-helices and five short β-strands[784-786] (Fig. 4.3B). Like many heme peroxidases, the heme binding pocket of UPO is fairly central and binds to a protoporphyrin IX heme cofactor[781,784]. Molina-Espeja *et al.*[787] found that both the peroxidative and peroxygenative catalysis of UPO occurs near the heme cofactor, which is accessed through a relatively hydrophobic tunnel[784]. However, unlike 'classical' peroxidases, the heme cofactor of UPO and other heme-thiolate peroxidases is coordinated at the proximal site by a cysteine residue rather than the histidine present in 'classical' peroxidases. Additionally, UPO and other heme-thiolate peroxidases typically only contain one cation binding site, as opposed to the two sites commonly present in 'classical' peroxidases[629]. Nonetheless, the overall peroxidase catalytic cycle of UPO is identical to that of HRP (Chapter 3 Fig. 3.4), though UPO is additionally able to catalyse several peroxygenation reactions as well through various other catalytic cycles (the reader is referred to a review on UPO catalysis by Hofrichter *et al.*[777]). Due to its impressive catalytic capabilities as well as its engineered high yielding expression in yeast, the engineered UPO has been increasingly utilised. Recent examples include the study by Struwe *et al.*[778], where UPO was utilised to catalyse the oxidation of terpene scaffolds, and that by Stenner *et al.*[788] where it was shown that UPO could catalyse selective oxidation of cyclohexane to KA oil. The increasing use of engineered UPO highlights the power of directed evolution as a protein engineering technique. However, few studies have utilised the peroxidative ability of UPO, and, to the best of our knowledge, no unspecific peroxygenase has been used for enzyme-mediated polymerisation. On the other hand, previously it was shown that other heme-thiolate peroxidases were able to be used in EMP systems, including a chloroperoxidase[322], thereby indicating the potential of UPO for EMP. Furthermore, a biomaterial formed using UPO

could later also serve as a bioreactor resulting from the relatively wide catalytic range of UPO and its ability to be easily secreted by certain cells[293,777].

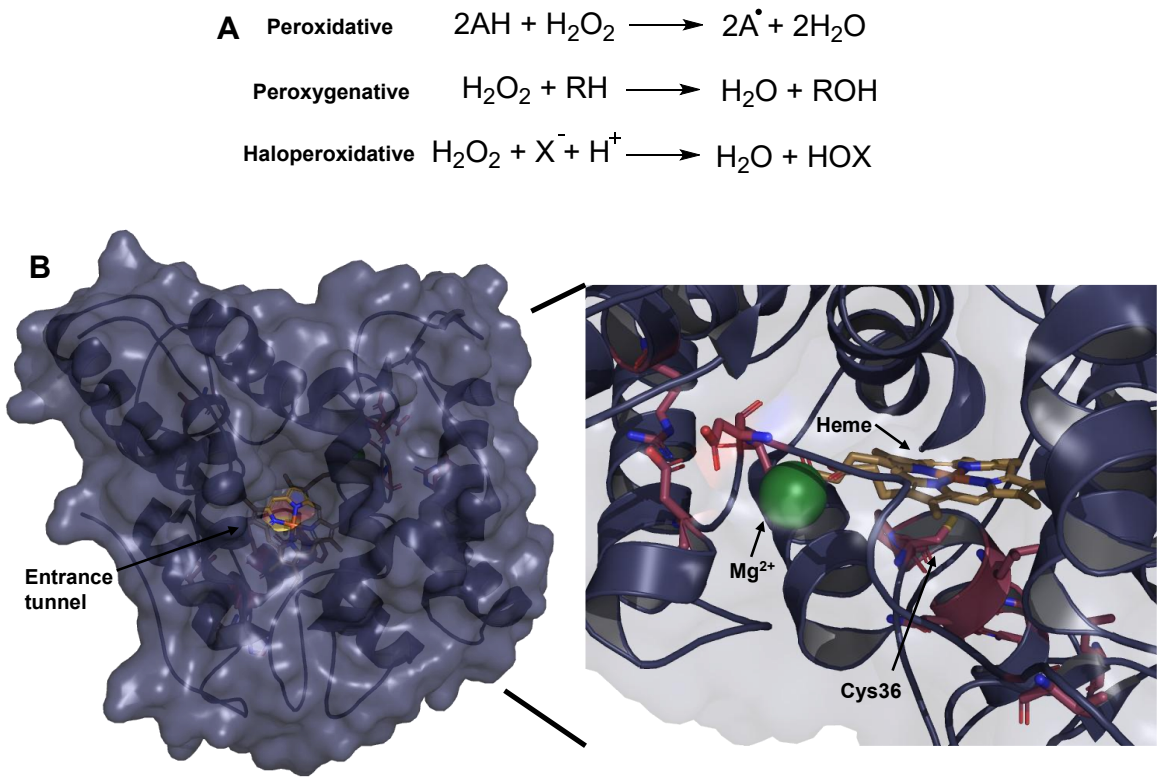



**Figure 4.3:** A) Enzyme-catalysed reactions conducted by members of the peroxidase-peroxygenase superfamily of which *A. aegerita* unspecific peroxygenase (UPO) is part of. UPO itself can catalyse both peroxidative and peroxygenative reactions B) Structure of UPO mutant PaDa-I (PDB: 5OXU) highlighting the substrate access tunnel, the heme cofactor, cation binding site and proximal cysteine residue. The $Mg^{2+}$ ion is shown in green, carbons of the heme cofactor are shown in gold and carbons of conserved residues are shown in pink. Iron is shown in orange, nitrogen in red, oxygen in blue and sulphur in yellow. Figure made using ChemDraw and PyMOL.

While enzymes could be key in catalysing the polymerisation of a 3D-printable ELM in a relatively non-toxic manner, in order for the ELM to act as a bioreactor or biosensor it must possess the ability to maintain living cells. These cells are critical because they possess the ability to produce desired proteins in response to an environmental cue of choice. It is also important for certain applications that the desired proteins be secreted or displayed such that they can interact with the environment around them, for example, catalysing the degradation of a biotoxin in the environment of the material[789,790]. It could also be beneficial for some use cases if proteins with eukaryotic modifications could be produced, such as in the production of biopharmaceuticals[749].

Previously, it has been shown by Klemperer *et al.*[361] that *E. coli* cells are viable within alg/PNIPAm. However, secretory expression in *E. coli* can be difficult[791], as well as

eukaryotic post translational modifications[792,793]. On the other hand, secretory expression of many proteins has been found to be relatively high yielding in the budding yeast species *S. cerevisiae* and *K. phaffii*[398]. As yeast is eukaryotic, these yeast species are also capable of eukaryotic modifications and certain strains of *K. phaffii* have been engineered to produce biopharmaceutics[794,795]. Additionally, both *S. cerevisiae* and *K. phaffii* are generally regarded as safe (GRAS), with *S. cerevisiae* having commonly been used in baking and brewing for thousands of years[796] and *K. phaffii* isolated from trees and noted for its ability to use methanol as a carbon source[797,798]. It should be noted that *K. phaffii* was previously known as *Pichia pastoris* before being renamed due to the phylogenetic assessment that *P. pastoris* represented several yeast strains[799,800]. Notwithstanding, both *S. cerevisiae* and *K. phaffii* are now used for recombinant protein production in synthetic biology, and as eukaryotic organisms are able to produce glycosylated proteins through a secretory pathway. Furthermore, recombinant protein production in both *S. cerevisiae* and *K. phaffii* is comparatively high yielding, cheap, quick and less laborious than other eukaryotic systems like insect cells or Chinese hamster ovary cells[398,801,802]. *S. cerevisiae* is considered a model organism, hence it is relatively well studied and has had many synthetic biology tools developed for it[397]. Indeed, it has been utilised in several ELMs previously[384,393,395,403,803] (see Chapter 1.4.2). However, although *K. phaffii* has been less studied, it has been shown to yield greater amounts of UPO[293] and other recombinant proteins[804,805] compared to *S. cerevisiae.* This could result from the lack of fermentation in *K. phaffii* and the relative strength of its *AOX1* promoter[804-806]. Additionally, *K. phaffii* has been found to hyperglycosylate recombinant proteins to a lesser degree than *S. cerevisiae*, and to secrete less proteins natively, leading to relatively higher purity of the secreted recombinant protein. While *K. phaffii* has been less utilised than *S. cerevisiae* in ELMs, recently Yuan *et al*.[399] demonstrated that an ELM bioreactor could be made utilising *K. phaffii*. For both yeast species, their immobilisation within a hydrogel is advantageous over liquid culture because product separation is easier, there is increased re-usability, co-cultures are more easily utilised due to the separation provided by the material, and the material may provide a form of microbial containment to prevent escape of the modified yeast cells into the environment[395,399]. Furthermore, if the yeast cells are able to be 3D-printed within alg/PNIPAm, then the spatial distribution of the cells can be controlled and complex biomaterial structures can be formed[385].

Therefore, as a result of the scarcity of 3D-printable living yeast materials that can be formed through relatively 'green' methods and produce protein in response to a desired signal, here we aim to develop a yeast alg/PNIPAm ELM which can form through enzyme-mediated polymerisation using UPO, and act as a 'bioproducer' of proteins 'on demand'. To this end, it was demonstrated that UPO could be recombinantly expressed and secreted in both *K. phaffii* and *S. cerevisiae*, though with better yields in *K. phaffii*. It was shown for the first time, to the best of our knowledge, that UPO could be utilised to form alg/PNIPAm with relatively similar thermosensitive and mechanical properties to that formed by HRP. Furthermore, it was demonstrated for the first time, to the best of our knowledge, that yeast laden alg/PNIPAm could be made and that recombinant expression of both EGFP and UPO could be induced by both *S. cerevisiae* and *K. phaffii* respectively within the alg/PNIPAm hydrogel. However, in future the levels of protein production by yeast and their viability within the ELM need to be further improved. Additionally, the ELM's storability, reusability, mechanical properties and biocontainment should be established in future too.

## 4.3 Results

The development of an enzyme-formed alg/PNIPAm ELM which produced protein 'on demand' required i) the ability to recombinantly express *Agrocybe aegerita* unspecific peroxygenase PaDa-I (UPO) and the capability of UPO to catalyse the formation of alg/PNIPAm in place of HRP and ii) that alg/PNIPAm is biocompatible with yeast and protein production can be induced by yeast within alg/PNIPAm. Thus, towards the first aim, UPO was recombinantly expressed, its basic biochemical properties assessed and then its ability to catalyse the formation of alg/PNIPAm investigated in detail.

### *4.3.1 Comparison of UPO PaDa-I expressed in S. cerevisiae and K. phaffii*

Initially, UPO was recombinantly expressed in both *S. cerevisiae* and *K. phaffii* to characterise which yeast species possessed superior recombinant expression of UPO for future use (Suppl. Fig. 4.1). While previously it was found by Molina-Espeja *et al.*[293] that *K. phaffii* yielded greater amounts of UPO, here both species were investigated again because different strains of *S. cerevisiae* and *K. phaffii* were utilised here due to their availability (Methods 4.5.1&4.5.2).

Despite these differences, it was nevertheless found that the yield of UPO expressed in *K. phaffii* was over double that expressed in *S. cerevisiae* (Table 4.1). Notwithstanding, the

yields of UPO expressed in both yeast species were both a marked improvement over the yields of ancestral class III peroxidases expressed in *E. coli* (Chapter 3.3.7) and HRP expressed in *E. coli*[660,704] or in *K. phaffii* in shake flasks[648]. Moreover, both the *S. cerevisiae* and *K. phaffii* UPO yields generated here are higher than those reported previously by Molina-Espeja *et al.*[293] for shake flask expression. In particular, the yield obtained here for *K. phaffii* grown in shake flasks is almost four times greater than that reported by Molina-Espeja *et al.*[293] and even 1.25 times greater than that reported by Pullman *et al.*[780], who further optimised the signal peptide. The relatively high yield obtained here for *K. phaffii* shake flask expression could be due to the use of the $Mut^S$ (slow methanol utilisation) strain KM71H over the $Mut^+$ (methanol utilisation plus) strains used by the other studies. Here both the X33 and KM71H strains of *K. phaffii* were tested for UPO expression (Suppl. Fig. 4.2) because previously it has been found that certain proteins can express better in strains with $Mut^S$ phenotypes like KM71H[807,808]. It was theorized that this may be due to the capability of $Mut^S$ strains to dedicate more resources to recombinant protein expression than $Mut^+$ strains because $Mut^S$ lack the *AOX1* gene possessed by $Mut^+$ strains. Although some recombinant proteins still express best in $Mut^+$ strains, here the use of a $Mut^S$ strain was found to result in higher yields for UPO. Although bioreactor expression of UPO in *K. phaffii* X33 was found to result in significantly improved yields of up to 217 mg/L[293], much greater than reported HRP expression yields[663,697,702], the results here suggest that UPO yields could be further improved in future by using the KM71H strain within a bioreactor.

However, although relatively high yields were obtained for shake flask expression in both yeast species here, the Reinheitszahl (Rz) values and heme incorporations were relatively low (Table 4.1). This was especially true for UPO expressed in *S. cerevisiae*, with a measured heme incorporation of only 8%. As the heme cofactor is essential for catalysis, less than 100% heme incorporation results in the functional concentration of enzyme being correspondingly lower than the total enzyme concentration. Previously, UPO PaDaI has been purified with Rz values of 2.4[293], indicating much higher heme incorporation. Therefore, heme incorporation of the UPO expressed here needs to be improved in future. It was found that addition of hemoglobin to the expression medium of *K. phaffii* resulted in no change to the Rz value of the final UPO sample (Suppl. Fig. 4.5). Therefore, perhaps the lower heme incorporation found here compared to that obtained by Molina-Espeja *et al.*[293] could result from the slightly different purification method used here due to limited availability of purification matrices (Methods 4.5.3). In future, heme incorporation should thus be improved

by attempting different purification techniques, or by incubating purified UPO with hemin, as has been conducted for other heme-containing enzymes[644,725]. Nonetheless, the heme incorporation of the UPO expressed in *K. phaffii* here is much greater than of any of the characterised ancestral class III peroxidases expressed in *E. coli* (Chapter 3.3.7) and of HRP purified without hemin incubation[644,659,661].

| | Yield (mg/L culture) | Soret peak (nm) | Reinheitszahl ($A_{418}/A_{280}$) | Heme incorporation (%) | ABTS specific activity (μmol $min^{-1}$ $mg^{-1}$) |
|---|---|---|---|---|---|
| UPO expressed in *S. cerevisiae* | 11.5 | 418 | 0.2 | 8 | 32.6±1.33 |
| UPO expressed in *K. phaffii* | 30 | 418 | 1.2 | 65 | 130±6.86 |
| Commercial *A. rusticana* HRP | N/A | 403 | N/A | N/A | 484±9.45 |

**Table 4.1:** Purification of UPO PaDa-I in *S. cerevisiae* vs. *K. phaffii* resulted in different yields of purified protein, Reinheitszahl (Rz) values, heme incorporation and ABTS specific activities. Yields represent purified protein and Rz values use Soret peak for UPO. Heme incorporation was measured by pyridine hemochromagen assay. ABTS specific activity was conducted in 50/100 mM citric acid/phosphate buffer, pH 4.5 at 25 °C and error in ABTS specific activity represents S.E.M. of 4 replicates.

Perhaps partially resulting from the relatively low heme incorporation, the 2,2'-Azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) specific activities of UPO purified in both yeast species were significantly lower than that of the commercial *A. rusticana* HRP (Table 4.1). However, even accounting for the predicted functional concentration of the respective UPO enzymes, the ABTS specific activities would still be significantly lower than that of HRP. This could be due to the multifunctional nature of UPO as both a peroxidase and oxygenase, perhaps resulting in less efficient peroxidase catalysis, and/or the fact that HRP is relatively optimised for aromatic substrates like ABTS[602,603], whereas UPO is more generalist[809]. Alternatively, the lower ABTS oxidation specific activity of UPO could be due to the conditions used, such that its activity may be improved using a different temperature and pH. Previously, the catalysis of ABTS by both HRP and UPO respectively has been found to be optimal at pH 4.0 and at a temperature of ~30 °C[290,810,811], therefore in future it would be interesting to compare the specific activities of HRP and UPO under these conditions as well.

Nonetheless, UPO purified in *K. phaffii* exhibited better yields, heme incorporation and ABTS specific activity than that purified in *S. cerevisiae*. Therefore, it was decided to characterise *K. phaffii* UPO in more detail for its potential use in enzyme-mediated polymerisation. However, in future it would be interesting to also characterise *S. cerevisiae* expressed UPO further and investigate its potential in polymerisation of alg/PNIPAm due to the greater availability of synthetic biology tools available for *S. cerevisiae* than *K. phaffii*[397]. These tools could potentially allow for a greater range of future applications for an *S. cerevisiae* alg/PNIPAm engineered living material. It may be possible to improve the expression of UPO in *S. cerevisiae* in future by using a genomically integrated expression system[812], rather than the current episome plasmid system, and by using a bioreactor[293]. Nonetheless, due to time constraints, only UPO expressed in *K. phaffii* was characterised in more depth.

### *4.3.2 Characterisation of mass, secondary structure, thermostability and kinetics of UPO expressed in K. phaffii*

Before investigating the use of *K. phaffii* expressed unspecific peroxygenase PaDa-I (UPO) in enzyme-mediated polymerisation, further biochemical characterisation was conducted to explore its properties in more detail. These properties were compared with horseradish peroxidase (HRP) as a standard for which it would be desirable for UPO to meet.

Firstly, the molecular weight of the recombinantly purified UPO was determined to be approximately 44.2 kDa by MALDI-TOF mass spectrometry (Suppl. Fig. 4.2), with a relatively broad range of masses likely due to the variable nature of eukaryotic glycosylation[813]. This molecular weight was measured to be less than those previously reported for UPO PaDaI purified in *K. phaffii* X33 of 51.1 kDa reported by Molina-Espeja *et al.*[293] and 46.6. kDa reported by Pullman *et al.*[780] The lower molecular weight determined here could perhaps result from the use of the KM71H strain, which perhaps may have resulted in a lower degree of glycosylation and thereby a lower molecular weight. Nevertheless, the molecular weight of the recombinant UPO purified here is relatively similar to the wild type *A. aegerita* UPO, which has a molecular weight of roughly 46 kDa as predicted by SDS-PAGE[290]. Given that any differences in molecular weight of a given protein are commonly caused by post-translational modifications[814,815], the similarity in molecular weight between the UPO purified here and the wild type UPO could suggest that they have a similar degree of glycosylation, which would likely be beneficial to the structure and/or function of recombinant UPO[816]. Nonetheless, in future the relative glycosylation degree of

UPO expressed in *K. phaffii* KM71H should be experimentally characterised to fully validate that a lower degree of glycosylation has led to the lower molecular weight. This would allow a more accurate comparison of the glycosylation degree to the wild type UPO and other recombinant UPO PaDaI proteins.

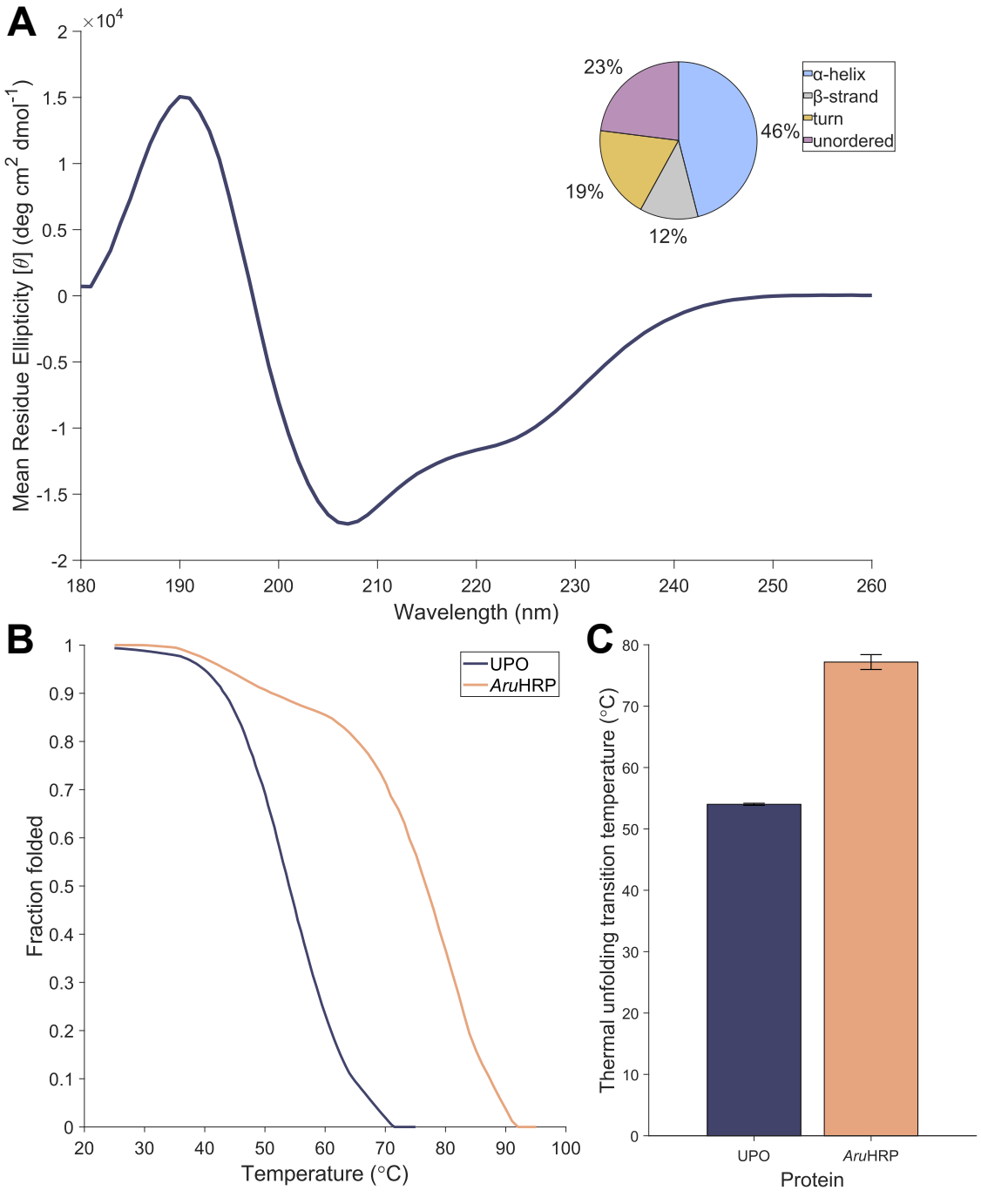



**Figure 4.4:** A) Circular dichroism (CD) spectrum of *K. phaffii* expressed unspecific peroxygenase (UPO) at 25 °C showing that it contains significant secondary structure. This spectrum was used with the CDSSTR method to predict the relative elements of secondary structure shown as the pie chart. B) Change in the fraction of folded protein with temperature of UPO and commercial *A. rusticana* HRP (*Aru*HRP). C) Thermal unfolding transition

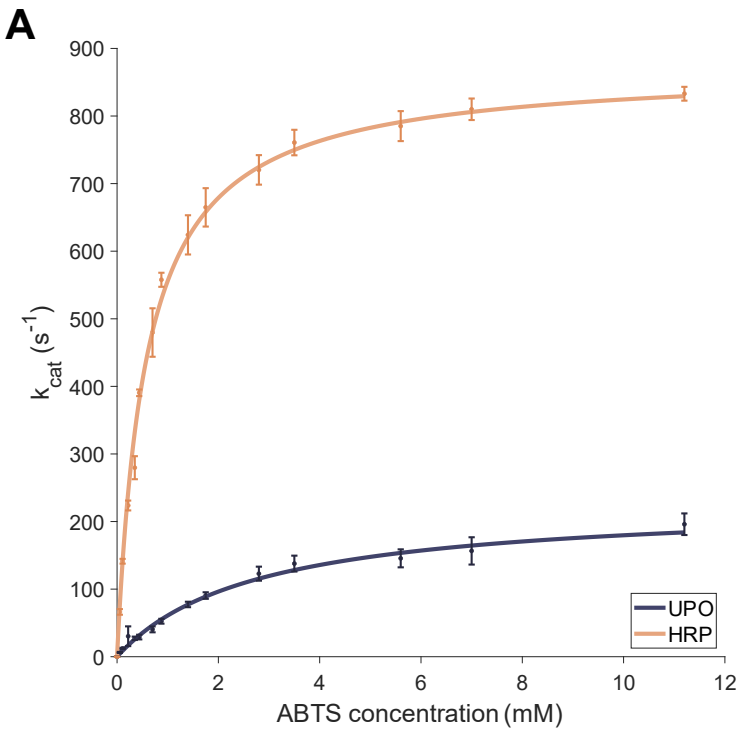
temperatures of UPO and *Aru*HRP. Error bars represent S.E.M. Figure made using MATLAB.

Subsequently, it was shown that UPO possessed secondary structure and thus was likely folded *via* circular dichroism (CD) (Fig. 4.4A). The prediction of the relative proportions of secondary structure elements from the CD spectrum proved relatively more accurate when compared to the experimentally determined tertiary structure[784] (Suppl. Table 4.2) than the predictions for the LacI family ancestors and the class III ancestral peroxidases (Chapter 2.3.1; Chapter 3.3.7). This is somewhat unexpected given that UPO also contains significant amounts of turns/disordered regions, which were theorized to reduce the accuracy of the previous secondary structure predictions[488,489]. Perhaps the disparity in accuracy could result from the standard reference datasets containing proteins more similar to UPO than to the ancestral peroxidases and transcription factors[740]. It would be informative in future to further explore the effect of turn length and reference dataset on the accuracy of secondary structure prediction from CD spectra to validate whether this trend is more broadly applicable or not.

By monitoring the change in the circular dichroism signal (Fig. 4.4B) of UPO and HRP, their respective thermal unfolding transition temperatures ($T_m$) were determined (Fig. 4.4C). While the $T_m$ of UPO was found to be ~20 °C lower than that of HRP, it is nevertheless high enough such that UPO should be relatively stable for most applications. This includes for its desired use in PNIPAm hydrogels, as the LCST of PNIPAm is ~32 °C. Interestingly, the $T_m$ measured here of 54 °C is similar to the activity half-inactivation temperature (T50) of ~58 °C measured by Molina-Espeja *et al*.[290] This could suggest a relatively strong degree of correlation between the loss of activity and unfolding for UPO.

Having established that the UPO PaDa-I recombinantly expressed in *K. phaffii* was folded and relatively thermostable for most purposes, its kinetics were investigated in more detail using ABTS as a substrate (Fig. 4.5A&B). It should be noted that the calculated kinetic constants are approximate both due to the multi-step catalysis cycle of UPO and HRP[603,777] and the range of molecular weights of both UPO and HRP due to variable glycosylation of both and different isoforms present in the HRP sample (Suppl. Fig. 4.2&4.6). Nonetheless, with such considerations in mind, it was found that the Michaelis constant ($K_M$) of UPO was approximately five times larger than that of commercial HRP, suggesting that UPO may be somewhat worse at binding ABTS than HRP and/or slightly worse at catalysing the oxidation of ABTS than HRP under the conditions tested. However, previously UPO PaDa-I expressed in *K. phaffii* has been reported to have much lower $K_M$s of ABTS catalysis of 0.12 mM to

even 0.05 mM under similar buffer conditions but potentially different temperatures (not reported)[293,784]. This thereby suggests that the $K_M$ of UPO catalysis may be relatively sensitive to temperature. Similarly, while the catalytic rate constant of the catalysis of ABTS by UPO found here was approximately four times less than that of HRP, it has previously been found by Ramirez-Escudero *et al.*[784] under similar conditions, but possibly different temperatures (not reported), to be up to 965 $s^{-1}$. Therefore, in future the kinetics of *K. phaffii* expressed UPO should be investigated over a range of temperatures and pHs and with various different substrates to paint a better picture of its kinetic abilities. In particular, it would be beneficial to characterise the kinetics of UPO using acetylacetone because this is the substrate of interest for the desired application of catalysing the polymerisation of PNIPAm.




| Kinetic constant | UPO | HRP |
|---|---|---|
| $K_M$ (mM) | 2.76 ± 1.11 | 0.567 ± 0.194 |
| $k_{cat}$ ($s^{-1}$) | 229 ± 37.7 | 871 ± 77.9 |
| $k_{cat}^*$ ($s^{-1}$) | 351 ± 57.9 | N/A |
| $k_{cat}/K_M$ ($s^{-1}$ $mM^{-1}$) | 83.1 ± 47.0 | 1537 ± 664 |

**Figure 4.5:** ABTS steady state kinetics with *K. phaffii* expressed unspecific peroxygenase PaDa-I (UPO) and commercial *A. rusticana* horseradish peroxidase (HRP). Kinetics were conducted in 50/100 mM citric acid/phosphate buffer, pH 4.5 at 25 °C. A) Michaelis-Menten curves of the steady-state reaction turnover numbers at different concentrations of ABTS. Error bars represent S.E.M. B) Kinetic constants of catalysis of ABTS by UPO and HRP respectively. Error represents 95% CI. $k_{cat}^*$ is calculated from the predicted functional enzyme concentration of UPO rather than the total enzyme concentration. Figure made using MATLAB.

Interestingly, the kinetic constants of HRP catalysis of ABTS in the lower pH buffer used here were significantly improved to those found previously using a higher pH buffer (Chapter 3.3.7), even though the same batch of commercial *A. rusticana* HRP was used. This trend has also been reported by other studies, with optimum catalysis of HRP tending to occur from pH

4.5 to 6.0 for several substrates[811,817,818]. The improvement of peroxidase catalysis in slightly acidic pHs could result from these pHs leading to an optimal hydrogen bonding network within the heme catalytic pocket and/or optimum formation of compounds II and III during the catalytic cycle[819-821]. Therefore, for the enzyme-mediated polymerisation of PNIPAm, in future it would be interesting to test how changing the pH of the precursor solution affects the efficiency of polymerisation to establish an optimal polymerisation reaction. Nonetheless, as both the ABTS specific activity assay and kinetics assay indicate that the catalysis of UPO may differ from that of HRP, the kinetics of UPO at different pHs should be further investigated in future before adapting the alg/PNIPAm formulation.

### *4.3.3 Formation of alg/PNIPAm with UPO*

Characterisation of UPO PaDa-I expressed in *K. phaffii* indicated that it may exhibit different catalytic properties to HRP, and therefore that the alg/PNIPAm gel formulation would need to be adapted for UPO. For alg/PNIPAm formation, the relative speed of formation of each network is important[363,364,822]. As formation of the alginate network is primarily diffusion limited[768,823], it is important that PNIPAm polymerisation occurs relatively quickly as well. The impact of insufficiently fast PNIPAm formation during alg/PNIPAm gelation was demonstrated by the synthesis of hydrogels with properties more similar to alginate when insufficient enzyme concentrations were used (Suppl. Fig. 4.7).

Consequently, the kinetics of PNIPAm formation with UPO were optimised to have a similar rate to the polymerisation of PNIPAm by HRP in the alg/PNIPAm formulation developed by Klemperer *et al.*[361]. To this end, it was established that the kinetics of PNIPAm formation could be tracked by measuring the increase in turbidity at 700 nm over time (Suppl. Fig. 4.8). Subsequently, based on the specific activity of ABTS oxidation by UPO and HRP it was predicted that approximately 3.7 times the amount of UPO should be used compared to HRP for a similar PNIPAm gelation rate (Table 4.1). However, at this concentration of UPO, the gelation rate of PNIPAm did not improve despite increasing the glucose oxidase concentration, which indicated that the concentration of UPO was rate-limiting (Fig. 4.6A). Therefore, a higher concentration of UPO was tested, of approximately 4.2 times the amount of HRP, with increasing concentration of glucose oxidase until a formulation was identified with a similar PNIPAm gelation rate to the original HRP/GOx formulation (Fig. 4.6B&C). It was interesting that a higher UPO concentration was needed than predicted from the ABTS specific activity assay. This may result from the difference in substrate and conditions when curing PNIPAm compared to the specific activity assay. However, it was not predicted that

the glucose oxidase concentration would need to be increased at an amount greater than the peroxidase concentration, with the optimal glucose oxidase concentration for UPO increasing by a factor of 10 from the original HRP formulation. Perhaps this could be due to the dual role of glucose oxidase during PNIPAm formation as both catalysing the formation of hydrogen peroxidase and as an oxygen scavenger, which may have led to the need to greatly increase its concentration such that both roles could be carried out sufficiently. The large increase needed in glucose oxidase concentration relative to UPO concentration could also result from differences in the kinetics of $H_2O_2$ reduction by UPO compared to the previously used HRP. While under standard conditions of phosphate buffer of pH 6-7 the kinetics of $H_2O_2$ reduction by both enzymes has been found to be fairly similar[293,811,824], perhaps UPO and HRP display relatively distinct kinetics in the environment of the gel ink, thereby necessitating greater production of $H_2O_2$, and thus GOx concentration, for UPO. Overall, the effect of both the concentration of UPO and GOx was shown to have a significant impact on the gelation rate of PNIPAm, however, a combination of UPO and GOx was identified which resulted in a similar gelation rate to the original formulation of HRP and GOx.

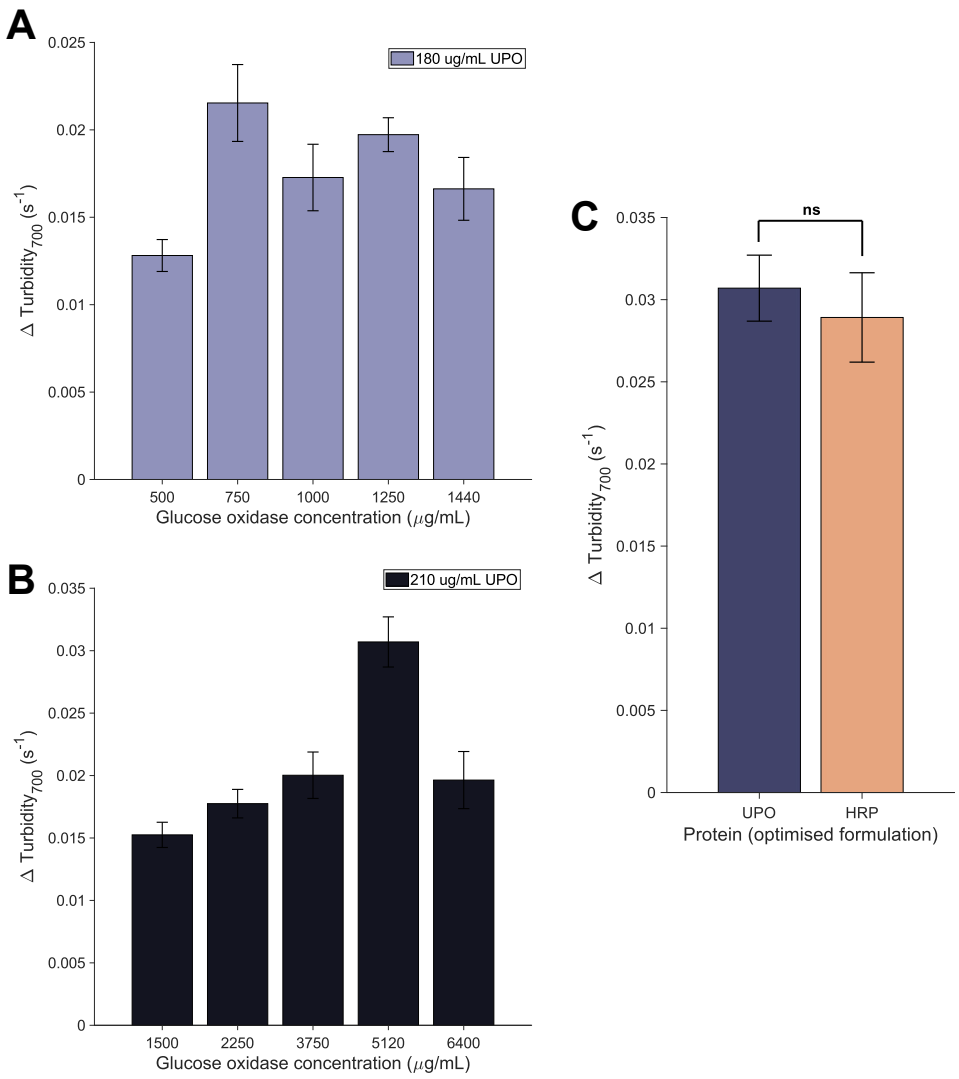



**Figure 4.6:** Optimisation of the gelation rate of PNIPAm using unspecific peroxygenase PaDa-I (UPO) by measuring the change in turbidity at 700 nm. All error bars represent S.E.M. of 3 replicates. A) Testing 180 µg/mL UPO (based on specific activity assay) with various concentrations of glucose oxidase (GOx) suggested that UPO was the limiting factor at this concentration. B) Testing a higher concentration of UPO at 210 µg/mL with various combinations of GOx identified a formulation with a PNIPAm gelation rate similar to that of the HRP/GOx formulation. C) Comparison of the PNIPAm gelation rate of the HRP/GOx formulation (50 µg/mL / 500µg/mL) with the best UPO/GOx formulation (210 µg/mL / 5120 µg/mL) showed a non-significant difference as analysed using an unpaired t test. Figure made using MATLAB.

Using the optimised UPO and GOx formulation, it was then validated whether UPO could be employed to form the 3D-printed alg/PNIPAm hydrogel with similar properties to that formed by HRP. Firstly, the printability of the UPO alg/PNIPAm appeared similar to that of the HRP alg/PNIPAm, with identical printing pressures of ~18 kPa used for both when printing with a 20 gauge nozzle (0.61 mm diameter). Furthermore, the UPO alg/PNIPAm ink was observed to be able to print various 3D structures in a similar manner to the HRP

alg/PNIPAm (Fig. 4.7). Nonetheless, it future it would be beneficial to quantitatively compare the printability of the UPO alg/PNIPAm ink with the corresponding HRP ink through rheology characterisation and assays of 3D printability, for example the grid method and filament collapse test[375,379,825].

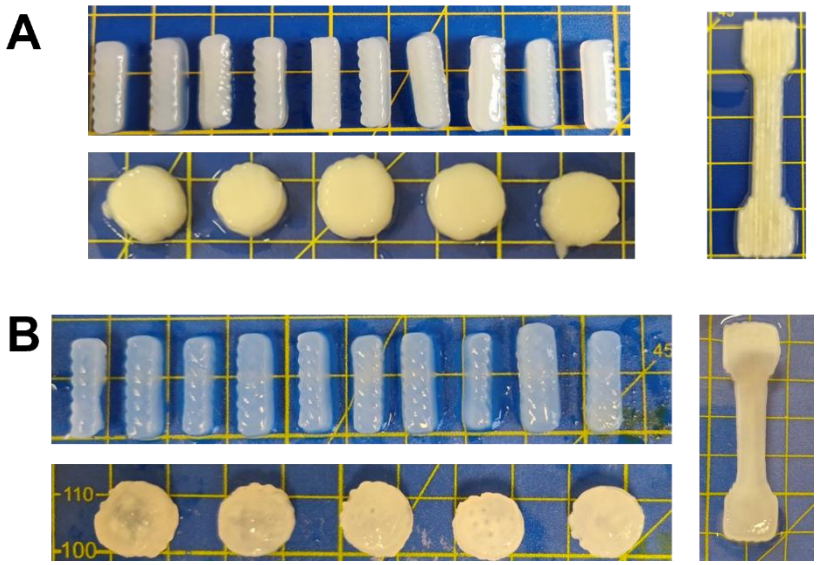



**Figure 4.7:** Comparison of 3D-objects printed using A) UPO alg/PNIPAm and B) HRP alg/PNIPAm. Both inks were able to be 3D-printed into a variety of objects. Yellow squares are 1 cm by 1 cm for scale.

### *4.3.4 Comparison of thermosensitive response between UPO-formed and HRP-formed alg/PNIPAm*

Following confirmation that the UPO-formed alg/PNIPAm could be 3D-printed and cured in a similar manner to the HRP-formed alg/PNIPAm formulation, it was then investigated whether the UPO formulation retained similar thermosensitive properties to the HRP formulation. In contrast to the earlier work by Klemperer *et al*.[361], here the thermosensitivity of the hydrogel was measured using gravimetric methods commonly used by other hydrogel studies[769,772,826,827] rather than dimension measurements. This is because the relatively 'soft' nature of hydrogels makes them difficult to measure using callipers, hence accuracy is reduced[344,365].

Using such experimental methods, it was first established that the UPO alg/PNIPAm formulation possessed similar 'contractability', or deswelling, at the high temperature of 60 °C to the HRP formulation (Fig. 4.8A). Therefore, this indicates that the PNIPAm network had formed to a similar extent in the UPO alg/PNIPAm formulation as in the HRP

alg/PNIPAm. The thermosensitivity of the UPO and HRP alg/PNIPAm formulations was further explored by examining the relative speed of 'contraction' of each respective formulation at 60 °C (Fig. 4.8B). Interestingly, it was found that the HRP formulation deswelled slightly faster than the UPO formulation, although both formulations had mostly 'contracted' in the relatively rapid time of 15 minutes. It should be noted that it was found that the alg/PNIPAm hydrogel also deswells slightly at room temperature (Suppl. Fig. 4.9), likely due to diffusion of calcium ions into water, but that the deswelling is significantly less and occurs at a much slower rate. Therefore, it is likely that the deswelling kinetics of both the UPO-formed and HRP-formed alg/PNIPAm hydrogels were largely driven by the thermosensitive behaviour of the PNIPAm network. The slightly faster deswelling of the HRP formulation suggests that it may have a slightly less compact structure than the corresponding UPO formulation[769]. This may have resulted from in differences in the extent of the PNIPAm network, length of the PNIPAm chains and overall pore size, amongst other properties, which can all influence the 'contractile' property of the alg/PNIPAm.

Perhaps also due to these differences between the formulations, it was found that the UPO alg/PNIPAm hydrogel showed slightly different temperature sensitivity to the corresponding HRP formulation (Fig. 4.8C). It was found that the deswelling of the UPO formulation had a somewhat more gradual transition range with temperature than the HRP formulation, and that the lower critical solution temperature (LCST) of the UPO formulation is ~30 °C whereas that of the HRP formulation is ~31 °C. Additionally, full 'contraction' of the HRP-formed hydrogel was found to occur by ~37 °C, whereas, due to the more gradual temperature sensitivity of the UPO-formed hydrogel, full 'contraction' occurred at the higher temperature of ~41 °C. These differences in temperature response could impact future applications such as biomedical applications, where precise temperature response may be required. Consequently, the UPO formulation may need to be further optimised in future to ensure a relatively 'sharp' response for some applications, such as temperature-controlled drug

release[828].

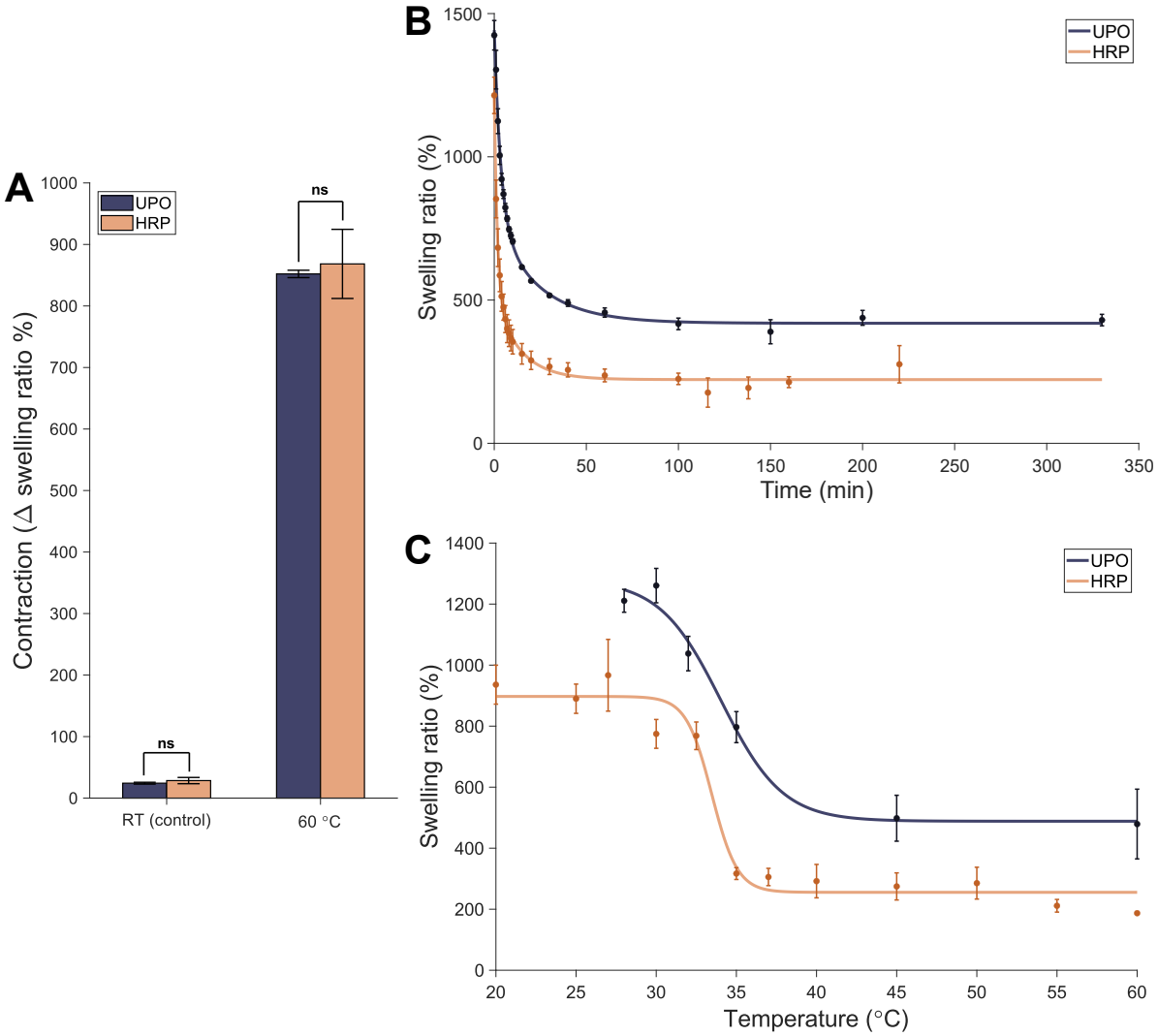



**Figure 4.8:** Exploring the thermosensitive properties of the alg/PNIPAm formed by unspecific peroxygenase PaDa-I (UPO) compared to that formed by horseradish peroxidase (HRP). A) The alg/PNIPAm hydrogels formed by both UPO and HRP displayed similar high deswelling when incubated at 60 °C in di$H_2O$. In contrast, at room temperature (RT), both gel formulations showed similar minimal deswelling. Significance was calculated by unpaired t test, and error bars represent S.E.M. of 3 replicates. B) The deswelling kinetics of alg/PNIPAm formed by both HRP and UPO at 60 °C follow a similar trend, with equilibrium deswelling reached by 50 min for both formulations. Error bars represent S.E.M. of 3 replicates. C) Deswelling of the alg/PNIPAm hydrogels formed by HRP or UPO is temperature dependent. The midpoint of the transition is ~33.5 °C for the HRP-formed hydrogel and ~34 °C for the UPO-formed hydrogel. Error bars represent S.E.M. of 3 replicates. Figure made using MATLAB.

Due to the slight differences in thermosensitive behaviour observed between the HRP-formed alg/PNIPAm and UPO-formed alg/PNIPAm, it would be beneficial in future to investigate which properties may have led to these differences, such as by studying the chemical composition of the respective hydrogels and examining their structures using cryo-scanning electron microscopy. A greater understanding of how the relative amount of UPO impacts the

properties of the alg/PNIPAm hydrogel could allow for fine-tuning of the formulation for each desired application. Nonetheless, the overarching thermosensitive behaviour of the UPO alg/PNIPAm hydrogel is relatively similar to the corresponding HRP hydrogel, indicating that the *K. phaffii* recombinantly expressed UPO could be used in place of *A. rusticana* HRP. As previously discussed, use of recombinantly expressed UPO would result in much higher reproducibility of the material, because plant-derived HRP suffers from batch-to-batch variation[603].

As reusability is important for many applications, it would be desirable if the 'contractile' property of the alg/PNIPAm hydrogel were repeatable. The repeatability of the deswelling of the UPO-formed alg/PNIPAm was compared to that of the HRP-formed alg/PNIPAm over several days (Fig. 4.8A & B), where once per day the respective hydrogel was incubated at 60 °C before being incubated overnight at room temperature to recover. This experiment demonstrated that the 'contractability' of both the HRP and UPO alg/PNIPAm was fairly repeatable up to the 11 times tested, indicating high reusability potential. However, the 'contraction' of the UPO alg/PNIPAm does decrease significantly from the first to second day before stabilising, whereas that of the HRP alg/PNIPAm is slightly more consistent (Fig. 4.9C). The significant decrease in the amount of deswelling of the UPO alg/PNIPAm between the first and second repetition appears to be largely due to the failure for the hydrogel to fully recover to its initial swelling weight after the first 'contraction'. This is shown firstly by the significant difference between the initial swelling weight ratio of day one compared to the swelling weight ratio of day two post recovery (Suppl. Fig. 4.10A). Secondly, this is shown by the fact that the difference between the control weight ratio and the recovery weight ratio is insignificant on day one (Suppl. Fig. 4.10C) but is very significant from day two to eleven (Fig. 4.10D). The inability to fully recover to the initial swelling weight could suggest that a small amount of degradation of one or both of the polymeric networks within the hydrogel occurred after the first incubation. Interestingly, it was also found that the deswelled weight ratio of UPO alg/PNIPAm, post incubation at 60 °C, gradually decreased with each repetition, with a significant difference between the first and final repetitions (Suppl. Fig. 4.10B). This could suggest that there is slight degradation of the alginate network with each incubation, as the greater 'contraction' could indicate a higher ratio of PNIPAm to alginate within the alg/PNIPAm hydrogel.

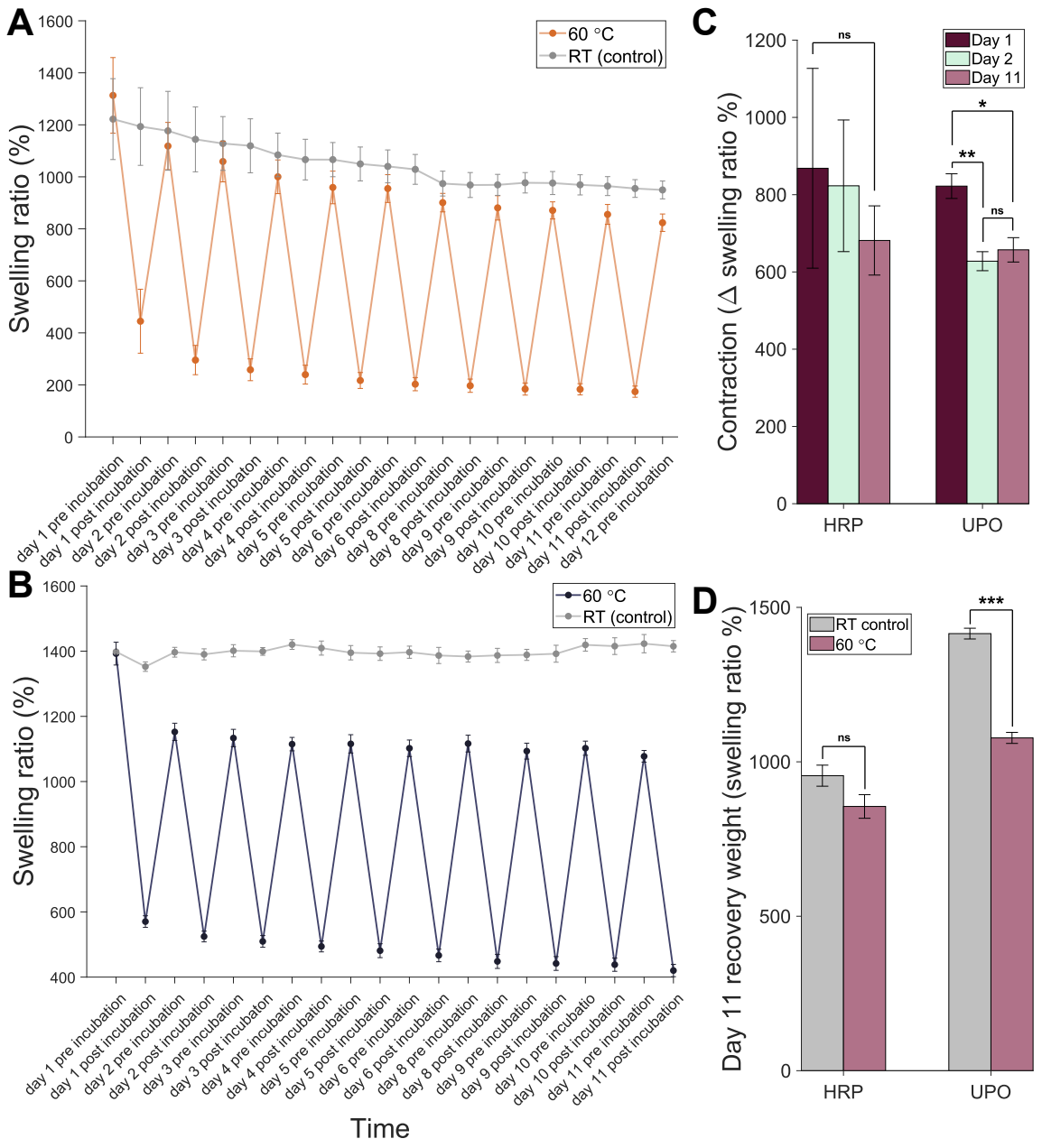



**Figure 4.9:** Repeatability of the 'contractability', or thermosensitive deswelling behaviour, of the UPO-formed and HRP-formed alg/PNIPAm hydrogels. A) The repeatability of 'contraction' for the HRP alg/PNIPAm when incubated at 60 °C (post incubation) then recovered at room temperature (RT) (pre incubation) for several days. As a control, HRP alg/PNIPAm was also continuously incubated at RT for several days. Error bars represent S.E.M. of 3 replicates. B) A) The repeatability of 'contraction' for the UPO alg/PNIPAm when incubated at 60 °C then recovered at room temperature (RT) for several days. As a control, UPO alg/PNIPAm was also continuously incubated at RT for several days. Error bars represent S.E.M. of 3 replicates. C) The HRP alg/PNIPAm deswelled by a similar amount each day, whereas the UPO alg/PNIPAm deswelled more on the first day compared to the later days. Error bars represent S.E.M. of 3 replicates and significance was calculated by unpaired t test. D) On the final day of testing (day 11), there was found to be no significant difference between the HRP alg/PNIPAm which had been repeatedly incubated at 60 °C and recovered at RT than the control which had only been incubated at RT. However, for the UPO alg/PNIPAm the repeated incubation at 60 °C and recovery appears to have reduced

its recovery weight ratio compared to the control only incubated at RT. Error bars represent S.E.M. of 3 replicates and significance was calculated by unpaired t test. Figure made using MATLAB.

For HRP alg/PNIPAm, although the 'contraction' did appear to gradually decrease with each repetition, the decrease was non-significant. This may be partly due to the relatively higher variability of the HRP alg/PNIPAm hydrogel samples tested. Thus, in future it would be informative to repeat the experiment using a greater number of samples. Nonetheless, the recovery swelling weight ratio for the HRP alg/PNIPAm was found to significantly decrease between the second and eleventh repetition in a manner not dissimilar to the UPO alg/PNIPAm. Additionally, the 'contracted' weight ratio of the HRP alg/PNIPAm also decreased slightly with each repetition similar to UPO alg/PNIPAm, although for the HRP-formed hydrogel this decrease is not significant likely due to the larger variability of the samples tested. This decrease in recovery swelling weight and 'contracted' weight could also suggest small degradation of the alginate network with each incubation.

Overall, despite the possible small amount of degradation of the UPO-formed alg/PNIPAm hydrogel, it was demonstrated that it has relatively high contractile repeatability of up to the eleven repetitions tested, such that it likely would have good reusability for desired applications.

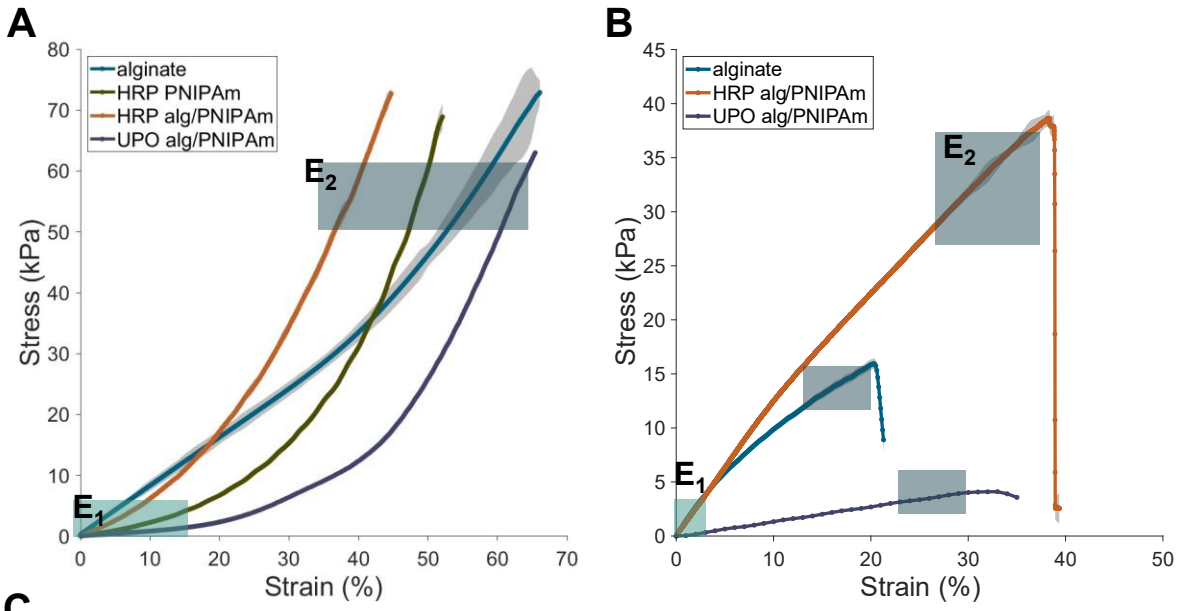
### *4.3.5 Mechanical behaviour of UPO alg/PNIPAm compared to HRP alg/PNIPAm*

Having established that the UPO alg/PNIPAm showed relatively similar thermocontractile behaviour as the HRP alg/PNIPAm, the mechanical behaviour of UPO alg/PNIPAm compared to HRP alg/PNIPAm under compressive and tensile forces was then characterised. Firstly, under compressive loads it was found that both HRP and UPO alg/PNIPAm, as well as HRP PNIPAm single network (SN) and alginate SN all displayed behaviour classical of hydrogels, with two distinct linear regions connected by a transition region[716,717] (Fig. 4.10A). This behaviour results from viscoelastic properties of hydrogels and their biphasic nature[717,829]. However, despite a general similarity, the compressive stress-strain curve of the alginate SN was slightly different in shape from the other formulations tested. Furthermore, the alginate hydrogel samples were not observed to make an elastic recovery, unlike the other formulations (Suppl. Fig. 4.11). It is hypothesised that this lack of recovery was likely due its reliance on physical ionic bonds compared to the covalent bonds present in the other networks[363,367].

Analysis of the linear regions of the different hydrogel formulations tested under compression revealed that HRP alg/PNIPAm showed compressive elastic moduli in between those of the alginate SN and PNIPAm single network for both the first ($E_1$) and second ($E_2$) linear regions (Fig. 4.10C). Typically, the compressive strength of fully interpenetrating double networks is greater than either of the single networks it is composed of[363,364] and other studies have indeed found this for alg/PNIPAm hydrogels[370,764,773]. However, it should be taken into consideration that the HRP alg/PNIPAm was printed whilst PNIPAm was moulded such that perhaps a direct comparison cannot be made. Notwithstanding, the relatively low compressive strength of HRP alg/PNIPAm could suggest that it has not formed a fully interpenetrating double network. Indeed, the lack of formation of such a fully interpenetrating double network was supported cryo-scanning electron microscopy (SEM) conducted by Klemperer[830], which found that interlacing of the alginate and PNIPAm networks only appeared to occur near the surface of the hydrogels, whereas, in the core of the hydrogels the two networks appeared to be separated. Furthermore, the porosity of alg/PNIPAm interpenetrating networks has been reported to be at least 50 μm[769,773,831,832], whereas the pore size of the HRP alg/PNIPAm was found to be only 2 μm by Klemperer *et al*.[361], which is more indicative of phase separated PNIPAm and alginate networks. Formation of a more homogenous alginate network within the PNIPAm network could be improved in future through the use of $CaSO_4$ to provide $Ca^{2+}$ ions at a slower rate than the currently used $CaCl_2$[773], or by initially using $Na^+$ ions then later replacing with $Ca^{2+}$ ions through displacement[770].

In contrast to HRP alg/PNIPAm, UPO alg/PNIPAm displayed a significantly lower $E_1$ elastic modulus, which was also less than those of either single network as well. However, for the second linear region, although the elastic modulus of UPO alg/PNIPAm was also lower than HRP alg/PNIPAm, it was in between the values found for alginate and PNIPAm single networks, making it more similar to HRP alg/PNIPAM. Unfortunately, due to limitations of the equipment used, the breaking behaviour of each formulation was not established. Therefore, in future a larger load or higher speed should be tested to investigate the behaviour of the respective hydrogels upon breaking. Nevertheless, before breaking, the compressive behaviour of UPO alg/PNIPAm was found to be somewhat different from that of HRP alg/PNIPAm, especially in the first linear region. This suggests that the UPO alg/PNIPAm had a somewhat different structure to HRP alg/PNIPAm. Perhaps UPO alg/PNIPAm may have had a slightly different ratio of alginate to PNIPAm, the respective alginate and PNIPAm

polymers may have been more ‘fragmented’, or the amount of interweaving of the alginate and PNIPAm networks was different.




C

| Hydrogel | Load | $E_1$ (kPa) | $E_2$ (kPa) | $\varepsilon_{frac}$ (%) | $\sigma_{frac}$ (kPa) | Strain energy density [area under curve] (kPa) |
|---|---|---|---|---|---|---|
| HRP alg/PNIPAm | Compressive | 59.7±10.1 | 335±5.1 | N/A | N/A | N/A |
| UPO alg/PNIPAm | Compressive | 13.3±1.66 | 284±12.5 | N/A | N/A | N/A |
| alginate | Compressive | 91.6±11.4 | 227±31.1 | N/A | N/A | N/A |
| HRP PNIPAm | Compressive | 31.3±10.2 | 480±32.2 | N/A | N/A | N/A |
| HRP alg/PNIPAm | Tensile | 117±8.21 | 77.0±4.77 | 40.8±2.98 | 38.4±2.16 | 8.42±1.10 |
| UPO alg/PNIPAm | Tensile | 14.6±2.26 | 14.0±0.823 | 36.0±0.895 | 4.87±0.309 | 0.951±0.0111 |
| alginate | Tensile | 163±14.6 | 68.5±7.28 | 21.8±0.466 | 13.8±0.0204 | 1.25±0.0198 |

**Figure 4.10:** Comparison of the behaviour of various hydrogel formulations under a compressive or tensile load. Samples of alginate, HRP alg/PNIPAm and UPO alg/PNIPAm were 3D printed whilst samples of HRP PNIPAm were cast in moulds, as PNIPAm cannot be 3D printed. A) Average stress-strain curves under a compressive load at 1 mm/min for alginate single network, HRP PNIPAm single network, horseradish peroxidase (HRP) alg/PNIPAm and unspecific peroxygenase (UPO) alg/PNIPAm. The 95% confidence interval (CI) is shown with light grey shading. The two linear regions of the curves are shown as $E_1$ and $E_2$. B) Average stress-strain curves under a tensile load at 1 mm/min for alginate single network, HRP alg/PNIPAm, and UPO alg//PNIPAm. The 95% CI is shown with light grey shading. The two linear regions of the curves are shown as $E_1$ and $E_2$. C) Table showing the elastic moduli, $E_1$ and $E_2$, the stress and strain at fracture and the strain energy density. Errors are S.E.M of at least 3 replicates. Figure made using MATLAB.

Differences between the UPO alg/PNIPAm and HRP alg/PNIPAm were also found in their respective behaviour under a tensile load (Fig. 4.10B). While both, as well as the alginate SN, displayed non-linear responses typical of viscoelastic materials like hydrogels[833,834], the

relative strength of the HRP alg/PNIPAm, as measured by its elastic moduli, was found to be significantly greater than UPO alg/PNIPAm. Furthermore, the stress and strain at fracture as well as the strain energy density were all significantly higher for HRP alg/PNIPAm than UPO alg/PNIPAm. In fact, UPO alg/PNIPAm was shown to have lower elastic moduli, lower stress and strain at fraction and lower strain energy density than even the alginate single network. This could further indicate that the polymeric chains in the UPO alg/PNIPAm are shorter or more fragmented than in both HRP alg/PNIPAm and alginate, leading to its relatively weak behaviour under tensile loads. It is posited that more fragmented chains would overall result in a smaller number of bonds, especially the hydrogen bonds that occur between the PNIPAm and alginate networks[773], thereby reducing the amount of energy that could be absorbed. While hydrogels in general display tensile-compressive asymmetry due to their viscoplasticity under compression[715], the relatively weak behaviour of UPO alg/PNIPAm under tensile load compared to compressive load was found to be more significant than HRP alg/PNIPAm or alginate. This could be caused by a more 'fragmented', or less 'entangled', network of both polymers in UPO alg/PNIPAm being more susceptible to tensile over compressive forces, especially because during compression the liquid phase of the hydrogel is more easily able to dissipate some of the force, which results in more plastic behaviour compared to the more purely elastic behaviour that occurs during tensile loading[715]. However, in future the structure of UPO alg/PNIPAm should be characterised, e.g. by using cryo-SEM, to better understand the differences observed in tensile and compressive strength.

Interestingly, under tensile load the HRP alg/PNIPAm possessed a larger stress and strain at fracture than alginate, as well as a greater strain energy density, indicating that when both the alginate and PNIPAm networks form well, the tensile strength of the material increases significantly. Possibly the relatively large tensile strength of HRP alg/PNIPAm may have resulted from the presence of the more rigid covalent PNIPAm network providing structure while the softer alginate network could 'sacrifice' intranetwork ionic bonds and internetwork hydrogen bonds to dissipate energy[363,367]. In future it would be informative to also characterise the behaviour of the PNIPAm SN under a tensile load to investigate whether it would be stronger or weaker than the HRP alg/PNIPAm. Work by Sasaki & Koga[835] indicates that the tensile strength of PNIPAm would likely be in the kPa range, however, here a different formulation of PNIPAm was utilised so it would be beneficial to characterise this specific formulation. Notwithstanding, all the hydrogels tested still fractured under tensile load but not compressive load within the weight limit tested. This highlights that the hydrogel

formulations tested showed relatively strong tensile-compression asymmetry, which is characteristic of hydrogels in general due to their biphasic nature[715]. Thus, for the alg/PNIPAm hydrogel, as for most hydrogels, this would suggest that for future applications environments with compressive forces would be more favourable than those with tensile forces.

Overall, UPO alg/PNIPAm was found to possess weaker 'strength' under compressive and tensile forces than HRP alg/PNIPAm. This highlights how sensitive the alg/PNIPAm formulation is to very slight changes and indicates that the concentration of UPO may need to be further optimised in future to achieve an alg/PNIPAm with similar mechanical properties to the HRP alg/PNIPAm. However, the sensitivity of the mechanical strength of alg/PNIPAm could be advantageous for applications requiring differing hydrogel strengths, as the strength of the alg/PNIPAm could potentially be changed depending on the application by the relative amount of UPO. Nevertheless, further characterisation of the hydrogels would be beneficial to inform future applications, such as the behaviour of the materials under shear forces, the behaviour of the materials under cyclic compressive, tensile and shear forces in order to investigate their relative durability, and rheological characterisation to better understand their viscoelastic behaviour. Notwithstanding, it was demonstrated that alg/PNIPAm formed by UPO still displayed relatively similar thermosensitive behaviour to that formed by HRP and possessed compressive and tensile strength in the same kPa range as HRP alg/PNIPAm. Therefore, it was next investigated whether yeast cells would be viable and could be induced to produce protein in UPO alg/PNIPAm such that it could form an engineered living material.

### *4.3.6 Towards the development of a yeast alg/PNIPAm engineered living material*

Following the establishment that UPO could be utilised to form 3D-printed alg/PNIPAm, to then develop a 3D-printable UPO-formed yeast alg/PNIPAm ELM which could produce proteins 'on demand' it was next determined whether yeast cells would be viable in alg/PNIPAm. Subsequently, it was then investigated if the yeast cells could be induced to produce protein within the alg/PNIPAm hydrogel, because this is critical for its future use as a bioreactor or biosensor, etc.

Initially the viability of yeast cells within the alg/PNIPAm hydrogel was established because previously only *E. coli* cells had been utilised in this formulation and, more generally, there have been relatively few yeast hydrogels developed, especially for *K. phaffii*[399,836]. Confocal

microscopy was utilised to assess viability with the SYTOX nuclear stain showing if the plasma membrane was compromised, the Calcofluor White M2R stain showing the cell wall and cytosolic EGFP expression showing prevalence of protein within the cytosol. To begin with, liquid culture controls of EGFP-expressing *S. cerevisiae, K. phaffii KM71H, K. phaffii X33* and nutrient deprived *K. phaffii X33* were examined to establish that the selected dyes could be used and to establish controls to compare against for cells in the alg/PNIPAm hydrogel (Fig. 4.11A).

The liquid culture controls showed that the selected stains worked as expected with minimal autofluorescence detected after background removal in post-processing. Notably, Calcofluor White M2R staining revealed the presence of many bud scars on most of the yeast cells of both species (Suppl. Fig. 4.13). This indicated relatively high rates of reproduction, which was expected in liquid cultures. The relative percentage of cells expressing EGFP was almost 100% for both *K. phaffii* strains, highlighting that a robust, inducible EGFP expression system was achieved. However, only 71% of *S. cerevisiae* cells were observed to express EGFP. This lower number of *S. cerevisiae* cells expressing EGFP may have resulted from the use of an episomal plasmid expression system rather than a genomically integrated expression system, which was used for *K. phaffii*. An episomal plasmid expression system can result in lower recombinant protein production as the plasmid has a relatively high loss rate with each new generation[812]. While chromosome integration of recombinant genes can be more laborious and costly if one does not have the necessary plasmid construct, in future it would be beneficial to attempt a chromosomal recombinant expression system of both EGFP and UPO PaDaI in *S. cerevisiae* to investigate whether this could improve the expression of both proteins. Nonetheless, most of the *S. cerevisiae* cells were shown to express EGFP. More generally, the cytosolic EGFP expression in both *S. cerevisiae* and *K. phaffii* revealed the presence of round structures lacking EGFP, which were likely nuclei and vacuoles (Suppl. Fig. 4.13). The ability to observe these structures highlights the relatively high resolution of fluorescent confocal microscopy compared to brightfield microscopy[837]. Both *S. cerevisiae* and *K. phaffii* were also shown to possibly display high levels of flocculation, as has previously been found to be characteristic of these yeast species[666,838]. The average cell diameters for both *S. cerevisiae* and *K. phaffii* respectively (Fig. 4.11C) were also found to be within the range of those previously reported[796,839,840], thereby establishing that the images collected by confocal microscopy could be utilised to characterise changes in cell size.

The nutrient deprived EGFP-expressing *K. phaffii* X33 revealed that many of the cells with compromised plasma membranes still contained significant levels of cytosolic EGFP and still possessed some degree of cell wall (Fig. 4.11). As EGFP is relatively stable *in vivo*[841-843], the presence of EGFP may not necessarily indicate that the cells are metabolically active and expressing EGFP, but rather that EGFP previously expressed over 24 hrs ago has not yet degraded, been digested or leached from the cell due to large ruptures of both the plasma membrane and cell wall. Nonetheless, the presence of SYTOX fluorescence inside many of the cells strongly suggests that these cells are likely dead, though ideally in future this should be corroborated with other experiments including a plating assay[844,845]. Although further experiments would be needed to validate true cell death, if death had occurred then it is proposed that apoptotic regulated cell death may have occurred in response to the starvation conditions due to i) the diameter of the nutrient deprived *K. phaffii* was significantly smaller than that of healthy cells indicating cell shrinkage, which occurs during apoptosis (Fig. 4.11C), ii) the cell wall and plasma membrane were not so greatly ruptured as to cause leakage of EGFP as may occur under necrosis, and iii) a round organelle lacking SYTOX fluorescence, possibly the nucleus, was visible in many of the cells (Suppl. Fig. 4.13), which is typical of apoptosis but not necrosis or autophagy[846,847]. This mechanism of regulated cell death is supported by previous work by Ruckenstuhl *et al.*, who were able to demonstrate apoptosis following starvation under a similar timeframe, albeit for *S. cerevisiae*[848]. Interestingly, some (<15%) of the cells in the nutrient-deprived condition were not found to show significant cell wall fluorescence while still displaying SYTOX and/or EGFP fluorescence. Assuming that staining and the relevant laser power were sufficient, the presence of EGFP and/or SYTOX fluorescence without Calcofluor flourescence could suggest that the intracellular contents have not ‘leaked’ from the cells despite the lack of a cell wall containing chitin and cellulose, to which Calcofluor binds to. Therefore, perhaps some yeast cells may have formed into protoplasts, perhaps due to autolysis[849].

Taken as a whole, the liquid culture controls demonstrated that aspects of yeast cell viability could be assessed with fluorescent microscopy and thus, that this technique could perhaps be used to observe the state of cells within a hydrogel.

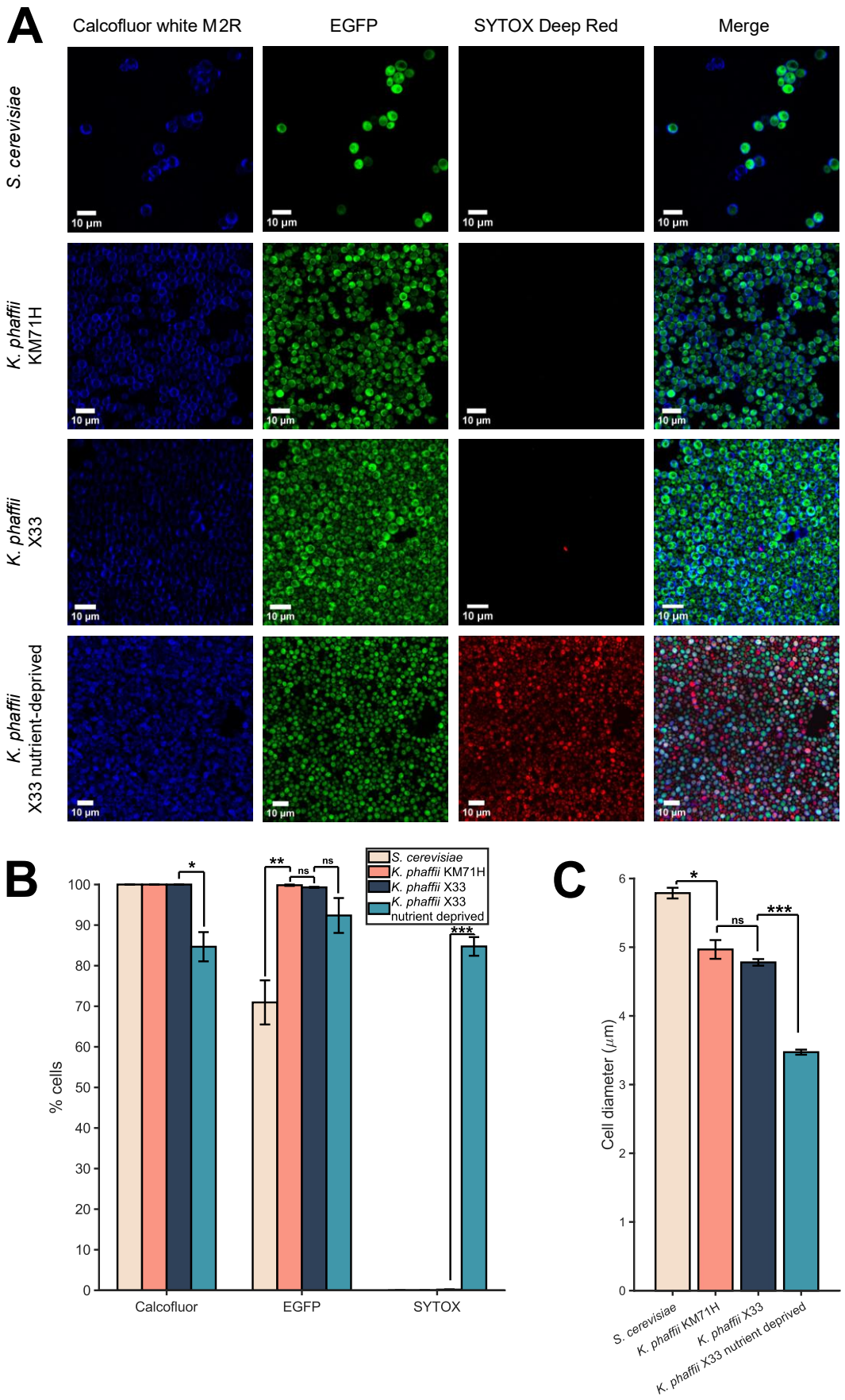

A
Calcofluor white M2R
EGFP
SYTOX Deep Red
Merge
S. cerevisiae
K. phaffii KM71H
K. phaffii X33
K. phaffii X33 nutrient-deprived
10 µm
B
% cells
S. cerevisiae
K. phaffii KM71H
K. phaffii X33
K. phaffii X33 nutrient deprived
Calcofluor
EGFP
SYTOX
*
**
ns
ns
***
C
Cell diameter (µm)
*
ns
***
S. cerevisiae
K. phaffii KM71H
K. phaffii X33
K. phaffii X33 nutrient deprived

**Figure 4.11:** Fluorescent images of *S. cerevisiae*, *K. phaffii* KM71H and *K. phaffii* X33 yeast cells expressing EGFP and stained with Calcofluor White M2R (Calcofluor) and SYTOX Deep Red. A) Representative images showing that Calcofluor could be used to stain the cell wall, EGFP was expressed and that SYTOX Deep Red could be used to show compromised plasma membranes. Scale bars represent 10 μm. B) The percentage of cells which were expressing EGFP, or which showed fluorescence from Calcofluor or SYTOX. Only the nutrient deprived X33 cells showed compromised cell walls and plasma membranes. EGFP expression was somewhat lower in *S. cerevisiae*, possibly due to the plasmid expression system used. Error bars represent S.E.M. of three samples and statistical significance was calculated by ANOVA or unpaired t test as appropriate. C) The average cell diameter of each of the yeast cell types as measured using the 405 nm (Calcofluor) channel. Error bars represent S.E.M. and statistical significance was calculated by ANOVA. Figure made using Fiji and MATLAB.

Initially, it was assessed whether yeast cells were viable in casted hydrogels of alginate (alg) single network (SN), PNIPAm SN, alg/PNIPAm and alg/PNIPAm with reduced *N*,*N*′-Methylenedi(prop-2-enamide) (MBA) following four hours of recovery post curing. The latter formulation was characterised because MBA is relatively toxic[850,851] and because reduction of MBA may increase pore size in the gel, thereby possibly allowing more 'space' for yeast colonies to form. Due to time constraints, only EGFP expressing *K. phaffii* X33 was investigated, because of the two *K. phaffii* strains it had higher EGFP expression. However, in future it would be informative to also investigate *K. phaffii* KM71H and *S. cerevisiae.* Nonetheless, using *K. phaffii* X33 within the various hydrogel formulations casted using a rectangular mould, it was found that the cells could be both stained and viewed by relatively high-resolution confocal microscopy while embedded inside the hydrogel without the need for sectioning (Fig. 4.12A).

However, it was found that the percentage of cells showing Calcofluor stained cell wall fluorescence was generally lower than expected given the relatively high percentages of cells with EGFP fluorescence and generally low percentages of cells with SYTOX fluorescence (Fig. 4.12B). Therefore, the concentration of Calcofluor used may have been insufficient for this experiment, and hence its concentration was subsequently increased in following experiments. Despite the relatively low Calcofluor fluorescence detected, for both the alginate and PNIPAm single networks almost all cells were observed to likely have an intact plasma membrane and possibly be viable, as indicated by low levels of intracellular SYTOX fluorescence and very high levels of cytosolic EGFP fluorescence (Fig. 4.12B). Thus, here it was shown that *K. phaffii* retain high viability in alginate, supporting the limited number of previous studies which have investigated *K. phaffii* in alginate[852-854]. Similarly, only one other study was found which demonstrated that *K. phaffii* are typically viable in acrylamide-based hydrogels[399] and here it was shown for the first time, to the best of our knowledge, that *K.*

*phaffii* retained high viability in PNIPAm, thereby supporting the biocompatibility of enzyme-polymerised PNIPAm. This is despite the formation of PNIPAm involving processes which could be damaging to yeast cells, such as the generation of $H_2O_2$ and radicals, and local oxygen deprivation[855].

Despite the likely high viability of *K. phaffii* cells in the alginate and PNIPAm single networks, the viability of the cells in both the alg/PNIPAm and alg/PNIPAm with reduced MBA respectively was relatively lower (Fig. 4.12B). This was demonstrated by the slightly lower number of cells displaying EGFP fluorescence, possibly indicating serious cell wall rupture or autophagy in some cells, and the relatively higher number of cells showing SYTOX fluorescence, indicating that the plasma membrane had been compromised in ~40 % of cells. The alg/PNIPAm formulation with reduced MBA concentration showed similar numbers of 'unhealthy' cells to the original alg/PNIPAm formulation, suggesting that toxicity of MBA and its crosslinking role within the hydrogel was not leading to the decreased viability of cells observed within the alg/PNIPAm. This was also supported by the relatively high viability of cells within the PNIPAm SN, which also contained the original concentration of MBA. Rather, the decrease in cell viability in the alg/PNIPAm could be due to the lengthier process required to make the alg/PNIPAm compared to its constitutive SNs, which would have resulted in the *K. phaffii* cells being subjected to the relatively 'harsh' environment of the gel ink, with likely relatively high osmotic stress, starvation conditions and possibly pH stress, for a longer time compared to the SNs. Furthermore, formation of the alg/PNIPAm requires more mixing steps using a dual asymmetric centrifuge than formation of either of the SNs, which may have resulted in the cells in alg/PNIPAm experiencing shear forces for a greater length of time, and thus, greater pressure and strain stresses. Nevertheless, ~60% of cells were observed to still have maintained an intact plasma membrane in the alg/PNIPAm and a reasonable cell density was also observed, therefore, a reasonable number of cells had likely survived the process to form the alg/PNIPAm.

Interestingly, however, for all the formulations the average cell diameter was approximately the same (Fig. 4.12C) and significantly smaller than the average diameter measured in liquid culture by ~1 μm. A decrease in yeast cell size can result from many factors, but commonly occurs under conditions of stress, such as nutrient deprivation[839,856-858]. Therefore, the decrease in cell diameter of *K phaffii* within the hydrogel formulations tested may suggest that the cells were experiencing some sort of stress in the hydrogel environment. To the best of our knowledge, the stresses that may be experienced by yeast cells within hydrogels has

not been well studied, thus, in future it would be interesting to characterise in more detail which aspects of the hydrogel environment could possibly be acting as external stress triggers for the yeast cells.

Despite using the same amount of *K. phaffii* culture in each of the gel formulations, it was found that the PNIPAm hydrogel possessed the greatest cell density when observed (Fig. 4.12D). Its higher density than the alginate SN could arise from a difference in pore size, as Klemperer *et al.*[361] found that the pore size of the alginate hydrogels was larger than the PNIPAm hydrogel, which would facilitate relatively rapid diffusion of the yeast cells into the surrounding medium in alginate. However, this would need to be validated in future through measurements of the respective pore sizes of the yeast-laden hydrogels and measurement of the respective numbers of colony forming units in the media of the hydrogels. The relatively low cell density of the alg/PNIPAm formulations could also have been partially caused by its slightly larger pore size than PNIPAm, although again this would need to be experimentally validated. However, given that the alg/PNIPAm networks were shown to result in significantly higher numbers of SYTOX permeable cells than either of the single networks, the low cell density observed in the alg/PNIPAm hydrogels may have instead largely been caused by complete lysis and/or denaturation of many cells such that they could not be viewed by the selected stains using fluorescent confocal microscopy. Nonetheless, in future this would also need to be confirmed; perhaps by attempting to detect the presence of EGFP in the medium, which could suggest relatively high levels of cell lysis.

While the viability of *K. phaffii* in the alg/PNIPAm hydrogel was not as high as in the alginate or PNIPAm single networks, nonetheless, ~60% of cells were observed to maintain their plasma membrane integrity and due to the interesting and desirable properties of the alg/PNIPAm, further experiments continued to investigate the feasibility of utilising yeast cells within this material.

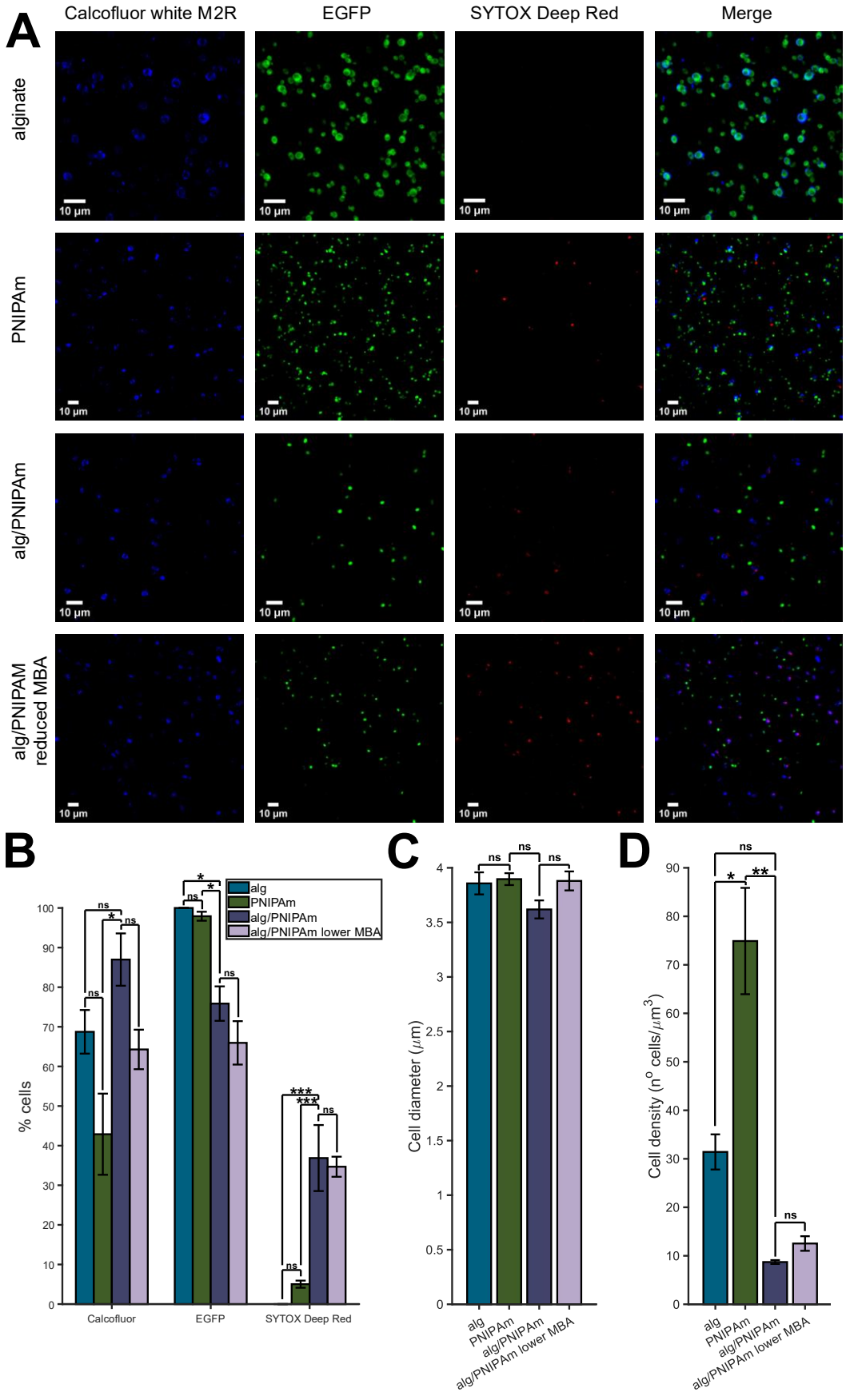

A
Calcofluor white M2R
EGFP
SYTOX Deep Red
Merge
alginate
PNIPAm
alg/PNIPAm
alg/PNIPAm reduced MBA
10 µm
B
alg
PNIPAm
alg/PNIPAm
alg/PNIPAm lower MBA
% cells
Calcofluor
EGFP
SYTOX Deep Red
C
Cell diameter (µm)
D
Cell density (n° cells/µm3)
ns

**Figure 4.12:** Fluorescent images of *K. phaffii* X33 yeast cells expressing EGFP in alginate, PNIPAm, alg/PNIPAm and alg/PNIPAm reduced MBA hydrogels formed using moulds. Cells are stained with Calcofluor White M2R (Calcofluor) and SYTOX Deep Red. A) Representative images showing the yeast cells within the various gel formulations. Images were taken of the cells within the hydrogels without sectioning and following 4 hrs of recovery since curing of the gels. Scale bars represent 10 μm. B) The percentage of cells which were expressing EGFP, or which showed fluorescence from Calcofluor or SYTOX. The percentages of cells which had retained EGFP and which were impermeable to SYTOX were relatively high for the alginate and PNIPAm single networks, indicating high viability. However, the percentages of cells which were permeable to SYTOX for both alg/PNIPAm formulations indicated less than 62% of the cells in each hydrogel formulation remained viable. Error bars represent S.E.M. of three samples and statistical significance was calculated by ANOVA. C) The average cell diameter of each of the yeast cell types as measured using the 405 nm (Calcofluor) channel. Error bars represent S.E.M. and statistical significance was calculated by ANOVA. D) The density of cells in each respective formulation. The density of cells in PNIPAm was the highest, possibly due to smaller pore size than alginate and higher viability than alg/PNIPAm. Error bars represent S.E.M. of three samples and statistical significance was calculated by ANOVA. Figure made using Fiji and MATLAB.

Subsequently, it was investigated whether alg/PNIPAm gel ink containing *K. phaffii* X33 could be 3D-printed, and the effect of printing, number of cells in the gel ink, and incubation medium on yeast cell viability. To this end, two concentrations of *K. phaffii*, $10^6$ cells/mL gel ink and $10^7$ cells/mL gel ink, were added to the alg/PNIPAm ink during the first stage of the gel making process. Cells which had not been induced to express EGFP were utilised in case the overproduction of EGFP could be contributing to a high cell load, which potentially could reduce cell health. Following printing and curing, the hydrogels were then incubated for 24 hrs in either a minimal (BMG) or complex (BMGY) medium. Typically, complex yeast media contain yeast extract and peptone, which are prone to batch-to-batch variation[859]. Thus, the use of minimal medium, which does not contain these components, could be more reproducible if it does not detrimentally affect cell viability.

Firstly, the *K. phaffii* laden alg/PNIPAm was observed to be 3D printable (Suppl. Fig. 4.14), although a higher printing pressure was needed (~20 kPa) compared to the acellular gel ink (~18 kPa). The higher pressure could indicate that the addition of cells has increased the viscosity of the ink, however, this should be validated using rheology in future. Following curing and incubation in media for 24 hrs, the printed yeast-laden hydrogel structures were observed to have mostly maintained their printed structure of either rectangles or boxes (Fig. 4.12A; Suppl. Fig. 4.14) and the boxes could be easily handled with tweezers (Suppl. Fig. 4.14), suggesting some degree of mechanical strength under compressive forces. Thus, the addition of *K. phaffii* to alg/PNIPAm appears to have had relatively minimal impact on printability and medium incubation does not appear to majorly impact the shape of the

printed alg/PNIPAm structure. Nonetheless, in future it will be important to characterise the impact that the cells may have had on the alg/PNIPAm material, such as the effect of the cells on the thermosensitive and mechanical properties.

However, it was found that the impact of the printing process on the cells was likely quite detrimental, with almost all cells showing plasma membrane permeabilization in all the conditions tested (Fig. 4.13B). There was a significant increase in the percentage of *K. phaffii* cells which displayed compromised plasma membranes in the printed alg/PNIPAm compared to the previously characterised moulded alg/PNIPAm, suggesting that the printing process was primarily the cause. During printing, the cells would have been subjected to both relatively large pressure stresses and shear forces[359]. Additionally, the printing process added time to the gel formation process, which would have lengthened the time the cells are exposed to stresses caused by the environment of the gel ink, which might have included osmotic stress, pH stress and starvation.

Despite the fact that almost all cells were shown to possess a compromised plasma membrane, and the cells were subject to shear forces and pressure during printing, many cells were observed to have still maintained some degree of cell wall. Interestingly, the *K. phaffii* alg/PNIPAm incubated in complex BMGY medium had a significantly higher percentage of cells with a cell wall than those incubated in minimal BMG medium, which perhaps could have resulted from either an increased level of reproduction or perhaps even cell wall repair[860]. Furthermore, the cell density was found to be higher in the alg/PNIPAm gels incubated in BMGY medium compared to BMG medium, even for samples with the same initial cell concentration (Fig. 4.13D). This could suggest that more cells were able to reproduce using BMGY medium, or that more cells died and denatured in BMG medium, such that they could not be viewed with the selected stains.

Nevertheless, for both incubation media tested the average cell diameter was found to be very similar (Fig. 4.13C), and also similar to that measured in all previous experiments for the *K. phaffii* cells embedded within a hydrogel. This is despite the fact that most of the cells in the printed alg/PNIPAm were shown to have a compromised plasma membrane, whereas previously most of the cells still had an intact plasma membrane. It is hypothesised that the lack of difference could possibly result from the fact that yeast cell size is controlled by many factors, such that perhaps cell shrinkage due to stress would be indistinguishable from cell shrinkage caused by regulated cell death, as previously discussed for ‘viable’ *K. phaffii*

within the hydrogels[844,845]. Perhaps such stress could result from mechanical forces exerted by the hydrogels onto the cells[359].

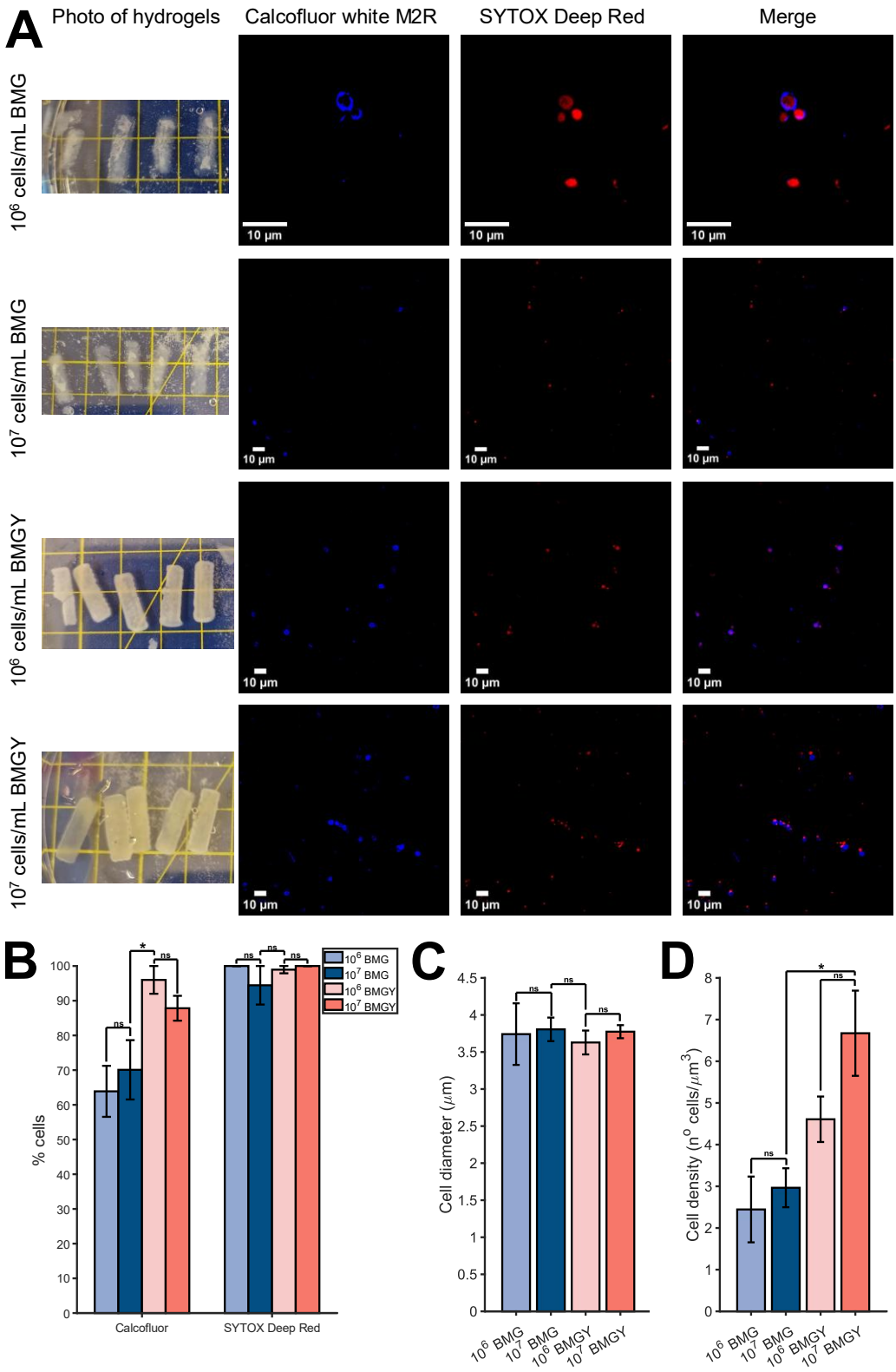



**Figure 4.13:** Investigating the printability of *K. phaffii* X33 laden alg/PNIPAm and viability of the cells within the hydrogel. Two different concentrations of cells were used in the gel

formulation ($10^6$ cell/mL gel ink and $10^7$ cells/mL gel ink) and two different incubation media (BMG and BMGY) were examined. A) Leftmost column shows photos of the yeast-laden hydrogels following 24 hrs of incubation in BMG or BMGY medium. Other columns show representative images from fluorescent microscopy showing the yeast cells within alg/PNIPAm. Images were taken of the cells within the hydrogels without sectioning and after staining with Calcofluor White M2R (Calcofluor) and SYTOX Deep Red. Scale bars represent 10 μm. B) The percentage of cells which showed fluorescence from Calcofluor or SYTOX. Error bars represent S.E.M. of three samples and statistical significance was calculated by ANOVA. C) The average cell diameter of each of the yeast cell types as measured using the 405 nm (Calcofluor) channel. Cell diameter was not significantly different between the conditions tested. Error bars represent S.E.M. and statistical significance was calculated by ANOVA. D) The density of cells in each respective formulation and medium incubation. Incubation in BMGY and forming the hydrogel with $10^7$ cells/mL gel ink led to the highest cell density. Error bars represent S.E.M. of three samples and statistical significance was calculated by ANOVA.

As both the cell density and the percentage of cells with some degree of cell wall was higher with BMGY medium, it was decided to only use this medium for subsequent experiments. However, it would be beneficial in future to also test rich YPD medium to investigate whether this could improve the recovery of yeast cells within alg/PNIPAm following the hydrogel formation process. Subsequent experiments used a concentration of $10^7$ cells/mL in the alg/PNIPAm gel ink to maximise the number of cells in the gel given the potentially very high death rate during the gel formation process. To attempt to improve yeast cell viability when forming alg/PNIPAm, the effects of changing aspects of the gel formation process and curing process were next investigated.

Due to the low viability of *K. phaffii* within the printed alg/PNIPAm, it was attempted to improve the viability by changing aspects of the hydrogel making process. To this end, rather than adding the cells in the initial stages of the process, the cells were instead only added immediately before the printing step and mixed into the gel using a lower speed. It was hypothesised that this would reduce the length of time that the cells are exposed to various stresses, and thereby perhaps improve their survival rate. Also, because the curing solution contains a high concentration of $CaCl_2$ (300 mM), which likely would lead to osmotic stress on the cells, this concentration was halved. Finally, it was investigated whether eliminating $CaCl_2$ from the gel ink could improve viability, as this perhaps might reduce the osmotic stress on the cells. *K. phaffii* X33 cells which had been induced to express EGFP were utilised, because while EGFP expression may increase cell load as previously mentioned, the fluorescent signal of EGFP provides additional information as to the state of the cell. Nonetheless, in future it would possibly be more beneficial to instead utilise a dye which indicates whether the yeast cell is metabolically active, such as FUN-1[861]. Two timepoints

were used to observe the state of the cells after a shorter and longer recovery incubation time: 4 hrs and 24 hrs.

Firstly, while both the gel inks containing $CaCl_2$ and not containing $CaCl_2$ could be printed, the printing pressure needed for the ink lacking $CaCl_2$ was relatively low (~10 kPa) and its printability was observed to be somewhat worse, due to its lower viscosity (Suppl. Fig. 4.15). Therefore, eliminating $CaCl_2$ from the alg/PNIPAm ink would not be practical for future applications of the biogel which need to utilise the adaptability and customisability of 3D-printing.

Nonetheless, the *K. phaffii* cells within the alg/PNIPAm hydrogels formed from both respective formulations were characterised by confocal microscopy at 4 hrs and 24 hrs of incubation post hydrogel formation (Fig. 4.14A). As found previously, the average cell diameter did not significantly differ between the formulations and timepoints tested (Fig. 4.14B) and was similar to those measured in the preceding experiments.

In contrast to previous experiments, in this experiment two timepoints were examined, thereby allowing characterisation of any changes in cell density over time. Interestingly, for both formulations the cell density was not found to significantly differ over the period tested, which suggests both that the cells were not reproducing, and thus were either in some form of cell arrest or even dead, but also that the cells were not being lysed or degraded to the extent that their structure could not be viewed by the selected dyes and EGFP. It should be noted that this model assumes a closed system where the cells are contained within the hydrogel and not able to diffuse into the surrounding medium, which needs to be validated in future using, for example, a plating assay of the medium. However, if the cells had diffused into the surrounding medium, the observation of constant cell density would imply reproduction. However, reproduction could be less likely due to most of the cells being observed to be isolated, because *K. phaffii* reproduces through budding[666], and the relatively high number of cells with compromised plasma membranes, indicating that many of the cells were 'unhealthy' and thus perhaps unable to have reproduced (Fig. 4.14C&D). Furthermore, daughter cells would not possess EGFP fluorescence due to the inducible protein production system used, however, the relative number of cells with EGFP fluorescence was observed to remain constant for both formulations (Fig. 4.14C&D).

Nevertheless, the percentage of cells with a compromised plasma membrane, shown by SYTOX fluorescence, was found to decrease within alg/PNIPAm from 4 hrs to 24 hrs. As the

cell density was found to have remained constant over this time, then this could either indicate that some cells were able to repair their plasma membrane[844,862], or that daughter cells with intact plasma membranes were produced and some of the 'unhealthy' cells diffused into the medium. While the first scenario is posited here as more likely; however, without having characterised the presence of *K. phaffii* cells in the surrounding incubation medium, either scenario or even both could have occurred. Interestingly, for the formulation which lacked $CaCl_2$ in the alg/PNIPAm ink the percentage of cells with compromised plasma membranes remained close to 100% after 24 hrs. As the cell density had also remained constant and as any daughter cells would be unlikely to possess compromised plasma membranes, this would suggest that the cell population within the hydrogel remained either 'unhealthy' or even dead over the time period.

On the other hand, the percentage of cells which had maintained EGFP was observed to stay constant for both respective formulations. This suggests that the cell walls of the observed cells did not significantly deteriorate over time, which would have resulted in leakage of intracellular content. However, the alg/PNIPAm formulation lacking $CaCl_2$ in the ink was found to have a lower percentage of cells with EGFP fluorescence compared to the formula containing $CaCl_2$ in the ink. As the relative percentage of cells displaying some degree of cell wall remained similar between both formulations, the lower EGFP fluorescence in the formulation lacking $CaCl_2$ in the ink could perhaps indicate increased autophagy or other protein degradation processes, or could indicate that while the cell walls could be stained with Calcofluor, they were more seriously compromised or ruptured leading to EGFP leakage. In either case, this would indicate that the cells in the formulation lacking $CaCl_2$ in the ink were perhaps less viable than those in the ink containing $CaCl_2$. This perhaps could provide a rationale for the possible recovery of plasma membrane observed for the *K. phaffii* in the alg/PNIPAm containing $CaCl_2$ and lack of recovery for the cells in the formulation lacking $CaCl_2$. The lower viability of the cells in the formulation lacking CaCl2 in the ink could result from the greater osmotic shock when incubated in the curing solution compare to cells that had already been exposed to some level of $CaCl_2$ and thus could adapt[863]. However, this theory would need to be validated in future work.

Despite the differences in the relative percentages of cells which contained EGFP and SYTOX between the formulations, both were found to have a similar percentage of cells with cell walls, and this percentage increased over time by a similar amount. Assuming measurement of the number of cell walls was correct and cells were not reproducing, the

increase in the percentage of cells with cell walls over time suggests that some of the cells which were initially in a spheroblast state were possibly able to repair their cell wall. While this would need to be more definitively characterised in future, several studies have shown that cell wall recovery is possible in yeast and identified proteins important for cell wall repair[860,862,864]. Therefore, here preliminary evidence was shown that some *K. phaffii* cells may be able to 'recover' from any stresses experienced during hydrogel formation and printing after incubation of the biogel in medium. Thus, it is possible that a 3D printed *K. phaffii* ELM can be formed if sufficient recovery time is allowed.

Overall, it was found that removing $CaCl_2$ from the alg/PNIPAm ink worsened printability and did not improve *K. phaffii* viability, hence, in future the alg/PNIPAm ink containing $CaCl_2$ should be utilised. Additionally, for the alg/PNIPAm ink containing $CaCl_2$, a lower percentage of cells with a compromised plasma membrane after 24 hrs was observed compared to the previous experiment of printing *K. phaffii* in alg/PNIPAm. This indicated that the addition of cells immediately before printing and/or halving the concentration of $CaCl_2$ in the curing solution may have improved cell viability. Nonetheless, cell viability following printing was still found to likely be quite low. Therefore, in future perhaps a lower pressure nozzle gauge could be used to reduce the pressure stress on the cells, thereby perhaps increasing survival rates. Nonetheless, as there was some indication that the cells could recover after printing, it was instead next investigated if protein production could be induced in the yeast cells within the alg/PNIPAm hydrogel.

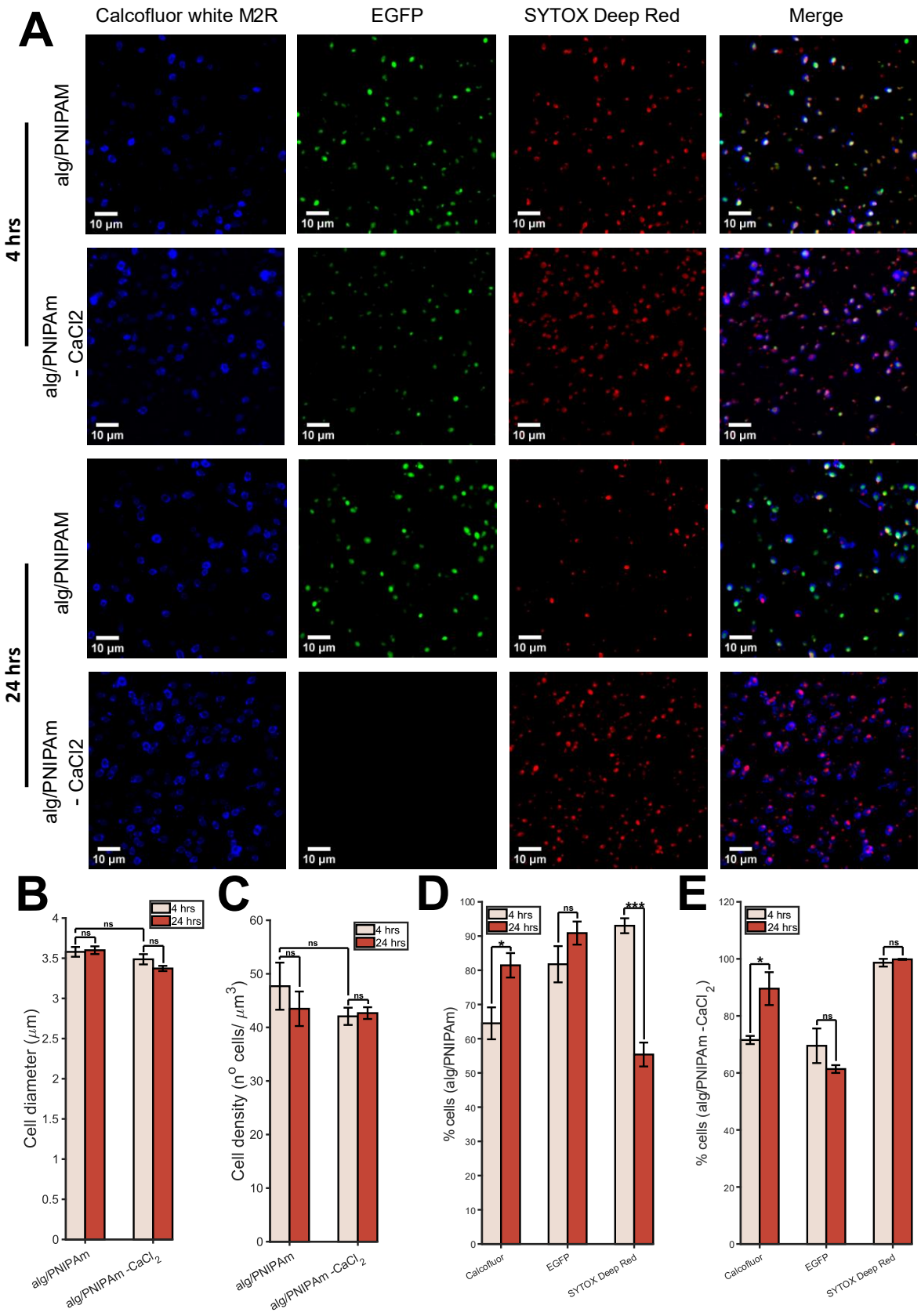



**Figure 4.14:** Attempting to improve the viability of *K. phaffii* X33 within alg/PNIPAm. The hydrogel making process was adapted such that *K. phaffii* was only added directly before printing and the concentration of $CaCl_2$ in the curing solution was halved to 150 mM. Two

different formulations were examined: one with $CaCl_2$ in the gel ink and one without $CaCl_2$ in the gel ink. Observations of the cells were made at 4 hrs of incubation (post curing) and 24 hrs of incubation. A) Representative images from fluorescent microscopy showing the EGFP expressing K. phaffii cells within alg/PNIPAm with and without $CaCl_2$ (-$CaCl_2$) in the gel ink (note that both formulations were cured with $CaCl_2$). Images were taken of the cells within the hydrogels without sectioning and after staining with Calcofluor White M2R (Calcofluor) and SYTOX Deep Red. Scale bars represent 10 μm. B) The average cell diameter of each of the conditions tested as measured using the 405 nm (Calcofluor) channel. Cell diameter was not significantly different between the conditions tested. Error bars represent S.E.M. and statistical significance was calculated by ANOVA. C) The density of cells in each respective condition. No significant difference in cell density was found for any of the conditions. Error bars represent S.E.M. of three samples and statistical significance was calculated by ANOVA. D) The percentage of cells which showed fluorescence from Calcofluor, EGFP or SYTOX for the normal alg/PNIPAm formulation. Cells possibly show some recovery of viability after 24 hrs. Error bars represent S.E.M. of three samples and statistical significance was calculated by unpaired t test. E) The percentage of cells which showed fluorescence from Calcofluor, EGFP or SYTOX for the alg/PNIPAm formulation where $CaCl_2$ was not used in the gel ink. Cells possibly show similar viability after 24 hrs. Error bars represent S.E.M. of three samples and statistical significance was calculated by unpaired t test. Figure made using FIJI and MATLAB.

After adaptations to the alg/PNIPAm making and curing process were found to somewhat improve *K. phaffii* viability within the hydrogel, next it was investigated whether protein production could be induced in yeast cells within alg/PNIPAm. This is critical for the alg/PNIPAm ELM to be able to act as an 'on demand' bioproducer. Induction of the production of enhanced green fluorescent protein (EGFP) in the cytosol and UPO for secretion were each characterised. EGFP was chosen because it could act as a reporter signal for biosensor applications[608,750,751] and UPO because its relatively broad chemical functionality would enable it to be used in bioreactor systems to e.g. catalyse the formation of value-added chemical compounds[777,778,865] or catalyse the removal of biotoxins from the environment[611].

While it was found that *K. phaffii* yielded a higher titre of UPO in liquid culture expression and the yeast viability experiments focused also on *K. phaffii*, it was decided to also characterise protein induction by *S. cerevisiae* because it could have advantages over *K. phaffii* in the production of certain proteins and because it has been more commonly been utilised in hydrogels than *K. phaffii*[359,399,866]. For protein induction in *K. phaffii*, it was decided to use the strain which had resulted in the highest yield of each respective protein during expression trials, which was X33 for EGFP and KM71H for UPO. The relative viability of the yeast cells within printed alg/PNIPAm rectangles and presence of any EGFP fluorescence was characterised before protein induction (24 hrs post curing) and after protein induction (86 hours post curing). To test for the production of UPO, a peroxidase activity

assay with ABTS as a substrate was utilised on the incubation medium of the UPO-expressing *K. phaffii* and *S. cerevisiae* hydrogels after protein induction.

Firstly, measurement of the cell diameter of both *K. phaffii* and *S. cerevisiae* cells within alg/PNIPAm both before and after induction (Suppl. Fig. 4.16B) revealed that surprisingly the diameter of both yeast species was found to be much more similar in alg/PNIPAm than was previously found in liquid culture, where ~1 µm difference in diameter between the yeast species was observed. The reduction of both *K. phaffii* and *S. cerevisiae* cells within alg/PNIPAm to a similar, smaller size could suggest that both yeast species possibly experienced relatively high levels of cell size reduction, possibly from the hydrogel environment leading to stresses on the cell or even regulated cell death[844,856], as previously discussed for *K. phaffii*.

Similarly, no significant difference was also observed in the percentage of cells with cells walls at both time points between *K. phaffii* and *S. cerevisiae* (Suppl. Fig. 4.16C). This indicated that an equivalent amount of *K. phaffii* and *S. cerevisiae* maintained their cell wall following the alg/PNIPAm printing and curing process. The lack of significant difference between the timepoints measured for both yeast species respectively could suggest that cells with a compromised cell wall were unable to repair it, but also that cells with compromised cell walls were not further degrading or denaturing, perhaps due to being 'trapped' within the hydrogel.

Unlike the similarity in cell diameter and the percentage of cells with largely intact cell walls between *K. phaffii* and *S. cerevisiae* within alg/PNIPAm, the cell density of each respective yeast species showed different trends (Fig. 4.15B). While the difference in initial density between *K. phaffii* and *S. cerevisiae* could be due to the imprecise nature of using $O.D._{600}$ measurements to estimate cell number, the cell density of *K. phaffii* remained similar after a further 72 hrs of incubation, whereas that of *S. cerevisiae* was observed to significantly increase. As previously discussed, the constant cell density of *K. phaffii* over time could suggest that the cells were not reproducing, and thus, may be relatively 'unhealthy'. This is supported by the relatively high percentage of *K. phaffii* with SYTOX fluorescence at both time points (Fig. 4.15C), which suggests that most of the cells have compromised plasma membranes and thus are not in a 'healthy' state. In contrast, the increase in cell density observed for *S. cerevisiae* could suggest that the cells were able to reproduce, and thus were 'healthier'. However, this somewhat conflicts with the relatively high levels of SYTOX

fluorescence observed for most *S. cerevisiae* cells at both time points, which indicates that many of the cells have compromised plasma membranes even at 86 hrs post curing. Thus, to confirm whether the *S. cerevisiae* cells were able to reproduce, or if the apparent difference in cell density results from the small number of samples, further experiments need to be conducted in future.

Despite the relatively high percentages of cells with compromised plasma membranes observed for both yeast species after protein induction, somewhat unexpectedly it was found that ~30% of EGFP-expressing *K. phaffii* cells and ~40% of EGFP-expressing *S. cerevisiae* cells showed EGFP fluorescence (Fig. 4.15A&D). While yeast cells do display some autofluorescence in the range of EGFP[867], after image processing only the aforementioned numbers of cells within the EGFP-expressing population were found to possess fluorescence in the range of wavelengths for EGFP, indicating that EGFP had likely been expressed by certain cells. Furthermore, the media of the UPO-expressing variants of both yeast species also demonstrated some level of UPO peroxidase activity, albeit very low for *K. phaffii* (Fig. 4.15E). As a negative control of media from the EGFP-expressing hydrogel was used in the ABTS assay, it is most likely that the peroxidase activity detected in the assays results from UPO production by cells within the respective alg/PNIPAm biogels. Therefore, for both species of yeast it was demonstrated that recombinant protein production could be induced within the alg/PNIPAm hydrogel. This result adds support to previous experiments showing that yeast protein induction is possible within hydrogels, such as the study by Yuan *et al.*[399], where it was demonstrated that protein production by *K. phaffii* could be induced by methanol in the acrylamide based hydrogel F127-BUM and the study by Sugianto *et al.*[403], where it was shown that *S. cerevisiae* protein production could be induced by galactose within a BSA-PEGDA hydrogel. Therefore, here the capability of alg/PNIPAm yeast-laden hydrogels to act as ELMs, for example bioreactors, was demonstrated. Nonetheless, as for both yeast species tested less than half of the cells were found to have been recombinantly expressing EGFP, and also that levels of recombinant UPO expression for both yeast species, but especially for *K. phaffii*, were lower than desired, the viability and metabolic activity of the yeast within alg/PNIPAm still needs to be improved in future.

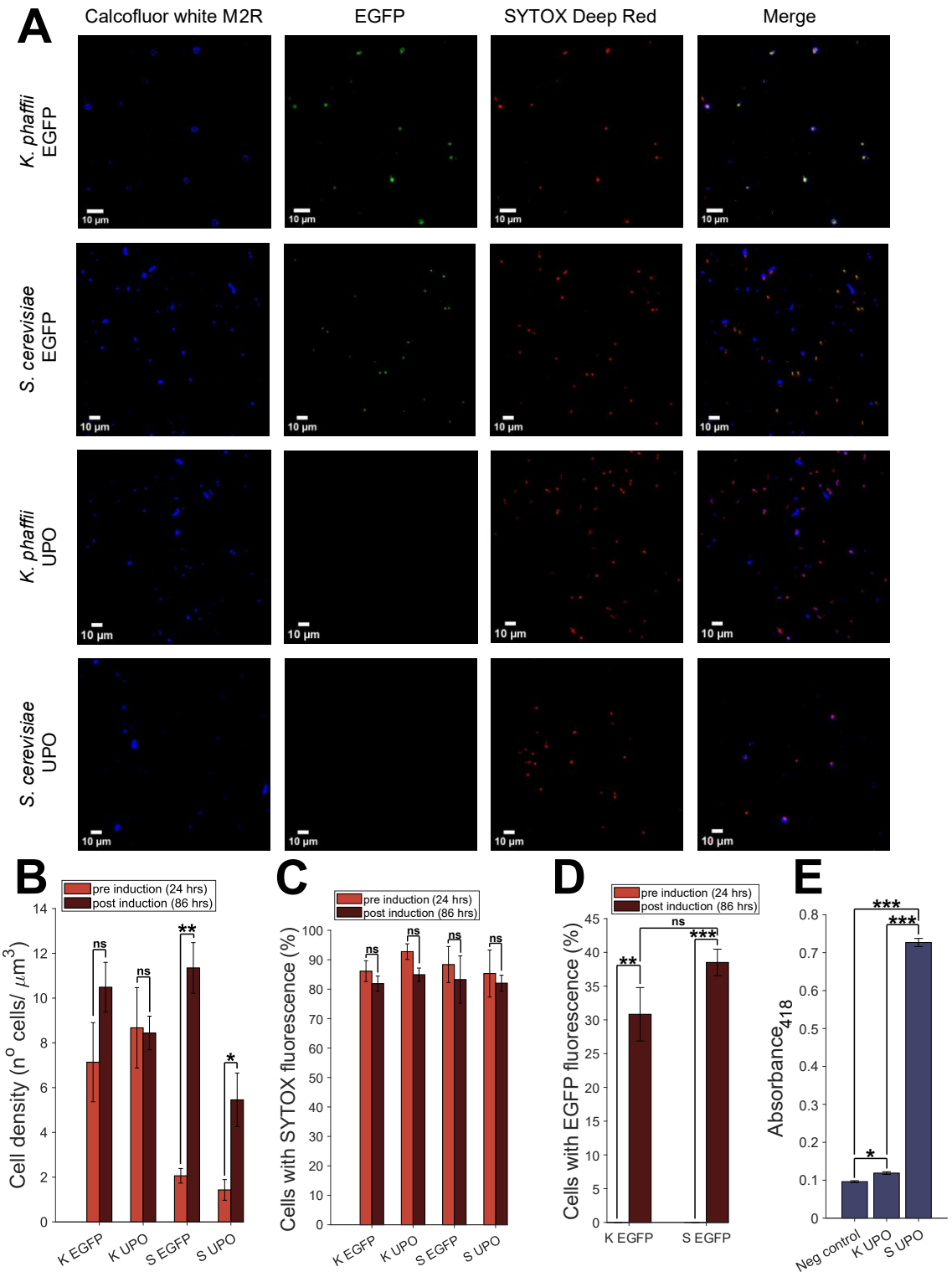



**Figure 4.15:** Characterising if recombinant protein production could be induced in *K. phaffii* and *S. cerevisiae* within an alg/PNIPAm hydrogel. The inducible expression of EGFP was investigated using *K. phaffii* X33 (K EGFP) and *S. cerevisiae* (S EGFP). The inducible expression of UPO was investigated using *K. phaffii* KM71H (K UPO) and *S. cerevisiae* (S UPO). A) Representative images from fluorescent microscopy taken after protein production had been induced (86 hrs after hydrogel curing). Images were taken of the cells within the hydrogels without sectioning and after staining with Calcofluor White M2R (Calcofluor) and SYTOX Deep Red. Scale bars represent 10 µm. B) The density of cells for each expression

system before induction (24 hrs after hydrogel curing) and after induction. No significant difference in cell density was found for *K. phaffii*, however, cell density appeared to increase for *S. cerevisiae*. Error bars represent S.E.M. of three samples and statistical significance was calculated by unpaired t test. C) The percentage of cells which were found to possess SYTOX fluorescence for each expression system before and after protein induction. No significant difference was found between any of the expression systems or time points. Error bars represent S.E.M. of three samples and statistical significance was calculated by unpaired t test to compare time points or ANOVA to compare expression systems. D) The percentage of cells observed to contain EGFP fluorescence for the EGFP expression systems before and after induction. EGFP expression was induced in some of the cells of both *K. phaffii* and *S. cerevisiae*. Error bars represent S.E.M. of three samples and statistical significance was calculated by unpaired t test. E) For the UPO expression systems, the presence of any UPO was characterised by measuring peroxidase activity in the surrounding medium of the alg/PNIPAm biogels using ABTS. A negative control (neg control) of the medium of *S. cerevisiae* EGFP biogel was used. *K. phaffii* expressed a small amount of UPO, and *S. cerevisiae* expressed a larger amount of UPO. Figure made using FIJI and MATLAB.

## 4.4 Discussion

The aim of this study was to develop a 3D-printable alg/PNIPAm ELM 'on demand' bioreactor, which could be formed through enzyme-mediated polymerisation in a reproducible manner. To this end, it was first characterised if recombinant unspecific peroxygenase PaDa-I (UPO), a protein engineered through directed evolution, could replace plant-derived HRP in the formation of alg/PNIPAm. It was then determined whether UPO-formed alg/PNIPAm could be utilised to form a yeast ELM, which could produce proteins in an inducible manner.

Towards the first aim, it was characterised whether HRP could potentially be replaced in EMP applications due to its lack of reproducibility. Previously, plant-derived horseradish peroxidase (HRP) has been commonly used in enzyme-mediated polymerisations[294,316,614], including in the polymerisation of alg/PNIPAm by Klemperer *et al.*[361]. However, HRP is difficult to express recombinantly in a soluble and active form[655,660], which is necessary for developing a reproducible alg/PNIPAm material. Therefore, here it was investigated whether HRP could be replaced by *A. aegerita* unspecific peroxygenase PaDa-I (UPO) because this enzyme was engineered through directed evolution for secreted, active and soluble recombinant expression in the GRAS yeast *S. cerevisiae*[290]. Here, as in previous studies[293,780], it was demonstrated that *A. aegerita* unspecific peroxygenase PaDa-I (UPO) could be recombinantly expressed and secreted in a soluble, active form by the GRAS yeast species *S. cerevisiae* and *K. phaffii*, thereby adding support to the utility of directed evolution as an engineering technique. Indeed, the number of studies published using the engineered UPO

has been increasing relatively rapidly[777,778,865], showing how powerful directed evolution can be to engineer desired properties into proteins, like improved recombinant expression. Similar to previous research conducted by Molina-Espeja *et al*.[293], it was also found here that yields of recombinantly expressed UPO were greater in *K. phaffii* than *S. cerevisiae*, as well as the *K. phaffii* expressed UPO having better heme incorporation and peroxidase activity. The higher yield of recombinant UPO in *K. phaffii* could result from its higher cell density than *S. cereivisae* as well as the strength of the *AOX* promoter utilised in *K. phaffii* compared to the *GAL1* promoter system used in *S. cerevisiae*[801,804,805].

However, to the best of our knowledge all previous studies have used $Mut^+$ strains of *K. phaffii* to express UPO, but here it was demonstrated that use of the $Mut^S$ strain KM71H resulted in higher UPO yields. This highlights the importance of testing both a $Mut^+$ and $Mut^S$ strain for the expression of a given recombinant protein in *K. phaffii*[807,808]. Notwithstanding, the *K. phaffii* KM71H expressed UPO was found to have somewhat lower heme cofactor incorporation than that previously reported by Molina-Espeja *et al.*[293], hence the higher yield may have come at the cost of reduced heme cofactor incorporation. As the heme cofactor is vital for the catalysis of UPO, to improve the yield of functional protein in future perhaps either the expression medium could be supplemented with hemin or δ-aminolevulinc acid, or purified UPO could be incubated with hemin[725]. Additionally, the yields of UPO could be improved in future using a bioreactor, as has previously been shown by Molina-Espeja *et al*.[293]. Nonetheless, the current recombinant expression of UPO in a soluble, mostly active form achieved in both yeast species is still an improvement over recombinant HRP expression[655,659], as well as the recombinant expression of the ancestral peroxidases in *E. coli* (Chapter 3.3.7).

Subsequently, basic biochemical characterisation of the *K. phaffii* recombinantly expressed UPO was conducted to compare its properties to plant-derived HRP. Initially, the molecular weight of UPO was determined for concentration measurements and to indicate its relative degree of glycosylation. The average molecular weight was found to be similar to that of wild-type (WT) UPO[290]. This may suggest that the UPO expressed here possessed similar glycosylation levels to native UPO, which could be beneficial for the structure and activity of the recombinant UPO expressed here[816]. Notwithstanding, in future the relative degree of glycosylation of UPO expression in *K. phaffii* KM71H should be verified using an endoglycosidase assay[666].

Characterisation of the *K. phaffii* KM71H expressed UPO by circular dichroism revealed that it had likely folded into the correct structure, with secondary structure elements in similar proportions to that found in the experimentally determined 3D structure[784]. Furthermore, the thermal unfolding of UPO was characterised for the first time, to the best of our knowledge, and interestingly it was shown that it possessed a thermal unfolding transition temperature ($T_m$) similar to the 50% activity inactivation temperature ($T_{50}$) measured by Molina-Espeja *et al.*[293]. This suggests that for UPO there is a relatively strong correlation between the $T_m$ and $T_{50}$, thereby perhaps supporting that there is a broader positive trend between $T_m$ and $T_{50}$ for many proteins[868,869]. As it is desired that UPO replace plant-derived HRP, the finding that the Tm of UPO is somewhat lower than that of HRP could suggest that HRP would still be more suitable than UPO for applications requiring relatively high temperatures (>50 °C). Nevertheless, as the enzyme-mediated polymerisation of PNIPAm within the alg/PNIPAm formulation is conducted at room temperature, the measured $T_m$ of UPO of 54 °C indicates that the protein should be thermostable during polymerisation. Furthermore, if in future a yeast alg/PNIPAm ELM were to be used as a UPO-expressing bioreactor, it would be desirable for UPO to be stable above the lower critical solution temperature (LCST) of the material. As the LCST of alg/PNIPAm was shown to be ~31 °C, UPO would still be relatively stable at temperatures somewhat greater than the LCST. Therefore, UPO was demonstrated to be sufficiently thermostable for the applications of interest of enzyme-mediated polymerisation and catalysis within an alg/PNIPAm bioreactor or its surroundings.

Although UPO was shown to be thermostable, it was found that its kinetics of peroxidase activity, as measured by ABTS oxidation, were somewhat worse than the commercial plant-derived HRP. While the kinetic constants of UPO were of the same order of magnitude as HRP, the Michaelis constant ($K_M$) was approximately 5 times larger and the turnover number ($k_{cat}$) was approximately 2.5 times smaller. The relatively poor kinetic constants for UPO found in this study contrast with the better values reported in several studies by researchers of the Alcalde group[293,784], where $K_M$ and $k_{cat}$ values of ABTS catalysis were reported to be lower and higher, respectively, than those found here. As the kinetics assay here was conducted in the same buffer and pH as those reported in the publications by the Alcalde group, either the UPO expressed and purified here has worse kinetics of peroxidase catalysis than that purified by researchers of the Alcalde group, or perhaps the previous assays were conducted at a different temperature to the temperature used here. It should also be noted that the kinetic constants of ABTS oxidation by UPO purified in *K. phaffii* under the same assay

conditions were reported to be different in different studies by researchers in the Alcalde group[293,784], perhaps suggesting a lack of precision or repeatability in their characterisation. Notwithstanding, possibly the differences in kinetics observed between the UPO purified here and that purified previously could suggest that the UPO purified here is somewhat less optimal catalytically. If this is the case, then perhaps in future UPO should be expressed using the X-33 strain of *K. phaffii* and purified in an identical manner to previous studies to explore whether this would result in UPO with catalytic activity more competitive with HRP. As the application of interest here for UPO is to replace HRP in the formation of alg/PNIPAm, it would also be beneficial in future to characterise the kinetics of UPO and HRP using acetylacetone as the substrate in assay conditions with a similar pH and range of temperatures at which alg/PNIPAm is typically formed.

Given that the peroxidase catalysis kinetics of UPO were different to those of HRP, in order to 'replace' HRP with UPO, the concentration of UPO should be adjusted for the desired application. Here, the application of interest is in the enzyme-mediated polymerisation of PNIPAm within alg/PNIPAm. As the relative speed formation of each network is important to form alg/PNIPAm with high mechanical strength and thermosensitivity, the concentration of UPO, and accordingly glucose oxidase (GOx), were optimised such that PNIPAm formation occurred at a similar speed to that catalysed using HRP. In order to characterise the speed of PNIPAm formation the increase in turbidity at 700 nm was tracked, which was relatively quick and required more readily available laboratory equipment than some previously used methods, such as using a piezoelectric-excited millimetre cantilever (PEMC) sensor[831] or rheology[870]. However, tracking formation by the increase in turbidity is less informative and precise than, for example, using a PEMC sensor, thus it would be beneficial in future to use more precise methods. Nevertheless, by tracking the relative rate that turbidity increased, it was found that approximately four times the amount of UPO to HRP needed to be utilised to achieve a comparable rate of PNIPAm formation. This was similar to the results from the specific activity assay of ABTS catalysis of UPO and HRP, which showed that the specific activity of UPO was also approximately four times less than that of HRP. Therefore, about four times the amount of UPO would likely be needed to replace HRP in various future applications. While using a relatively larger amount of UPO is not ideal, this amount could possibly be lessened by improving the heme incorporation of UPO as this would raise the functional concentration of protein. Furthermore, UPO is still advantageous in that it can be recombinantly expressed solubly in relatively high yields, unlike HRP.

As in the enzyme-mediated polymerisation of PNIPAm, GOx cannot be used in excess because this would inactivate peroxidase catalysis[779,787], the concentration of GOx was optimised concurrently with the concentration of UPO. While the increase in the amount of UPO was expected from its kinetics and activity assays, it was not hypothesised that the amount of GOx would have to be increased by a relatively larger amount than UPO, with approximately ten times more GOx needed for an optimal PNIPAm polymerisation rate. This relatively larger increase in GOx concentration required could result from the relatively slower kinetics of UPO catalysis, because perhaps the rate of oxygen scavenging needed to be much higher for UPO such that less radicals are 'quenched' by oxygen, which slows the polymerisation process[316,333]. Another potential reason for the relatively large GOx concentration could be that UPO is worse at binding $H_2O_2$ than HRP or has slower kinetics of $H_2O_2$ reduction, as either of these would then lead to a relatively higher amount of $H_2O_2$ generation being needed for UPO to achieve a comparable catalytic speed of acetylacetone oxidation to HRP. Thus, in future it would be informative to characterise the binding strength of hydrogen peroxide to UPO and HRP as well as the kinetics of hydrogen peroxide reduction by UPO and HRP. It would also be beneficial to characterise the speed of PNIPAm formation with a much larger range of UPO and GOx concentrations to better characterise and understand the relationship between the concentration of these two proteins and the speed of PNIPAm formation.

Notwithstanding, the properties of alg/PNIPAm formed using UPO were characterised and compared to those formed using HRP to verify that UPO could 'replace' HRP in the enzyme-mediated polymerisation of PNIPAm in alg/PNIPAm. To this end, it was first demonstrated that the UPO alg/PNIPAm hydrogel displayed similar printability to the HRP alg/PNIPAm. However, in future more detailed printability assays should be carried out, such as the filament collapse test and grid test, as well as rheology[375,379,825]. Nonetheless, characterisation of several aspects of the thermosensitive 'contractile' behaviour of UPO alg/PNIPAm revealed that it possessed relatively similar thermosensitive behaviour to HRP alg/PNIPAm, with near identical 'contraction' at high temperatures shown by both formulations and repeatability of contraction and recovery up to the 11 times tested over 11 days. However, characterisation of the thermosensitivity of UPO alg/PNIPAm in more detail did reveal slight differences in behaviour to HRP alg/PNIPAm. It was found that the UPO alg/PNIPAm deswelled at a slightly lower rate than HRP alg/PNIPAm and that UPO alg/PNIPAm had slightly lower LCST of 30 °C *vs.* the 31 °C measured for HRP alg/PNIPAm. These

differences highlight that even relatively small changes to the alg/PNIPAm formulation can result in different material properties. Perhaps the new UPO formulation resulted in a slightly different PNIPAm polymerisation rate, which in turn affected how PNIPAm polymerised within the alginate network. Therefore, it is possible that the UPO-catalysed alg/PNIPAm could have a different ratio of alginate to PNIPAm, a different amount of entanglement vs separation of the alginate and PNIPAm networks and/or different pore sizes. In future, it would be important to attempt to better characterise the differences between the UPO-formed and HRP-formed alg/PNIPAm hydrogels using techniques including cryo-scanning electron microscopy (cryo-SEM) and Fourier transform infrared microscopy (FTIR).

Indeed, characterisation of the bulk mechanical properties of the respective UPO and HRP alg/PNIPAm hydrogels also revealed differences in behaviour. While the relationship between the stress and strain response of both respective hydrogels was similar under both compressive and tensile loads, overall, HRP alg/PNIPAm showed better compressive and tensile 'strength' compared to UPO alg/PNIPAm. Like for the differences in thermosensitive behaviour, the differences in mechanical properties suggest that the structure of the UPO-formed hydrogel differs from that of the HRP-formed hydrogel in some way, thereby, emphasising the importance in future of characterising the structure of the respective hydrogels. Nevertheless, the mechanical 'strength' of UPO alg/PNIPAm was still found to be in the same kPa range as HRP alg/PNIPAm. Furthermore, the mechanical behaviour of UPO alg/PNIPAm should be sufficient for many ELM applications, which most likely only experience relatively weak loads from handling and shaking during incubation. Nevertheless, to extend the utility of the alg/PNIPAm, both as an acellular hydrogel and ELM, it would be interesting in future to investigate if its strength could be further improved. For example, perhaps $CaSO_4$ could be utilised rather than $CaCl_2$, as $CaSO_4$ tends to lead to more homogenous formation of the alginate network[773], and perhaps the PNIPAm network could be polymerised before excess $Ca^{2+}$ ions are added to the hydrogel constructs to better form an interpenetrating network[764,770]. Nonetheless, in future to better understand the properties of alg/PNIPAm for future applications, as well as structural characterisation, it would also be informative to characterise the shear response of UPO alg/PNIPAm, conduct cyclic testing under compressive and shear loads to characterise relative durability, and use the data from the compression, tensile and shear tests to fit the Mooney-Rivlin model[718,719] for a more accurate and complete understanding of the UPO alg/PNIPAm hydrogel.

Overall, this study provided support to the relatively high yielding recombinant expression of UPO, and demonstrated for the first time, to the best of our knowledge, the utility of UPO in enzyme-mediated polymerisation by catalysing the formation of alg/PNIPAm with UPO. While HRP had been previously used to catalyse this polymerisation[361], and many others[294,316,614], it suffers from being a plant-derived mixture of isoenzymes, which has high batch-to-batch variation and is relatively laborious and lengthy to produce. In contrast, UPO can be relatively easily and quickly expressed recombinantly in yeast with high yields, and here it was demonstrated that it can provide a viable alternative to HRP in enzyme-mediated polymerisation systems. However, the thermosensitive and mechanical behaviour of the UPO-formed alg/PNIPAm was somewhat different from the HRP-formed alg/PNIPAm, highlighting that concentrations of UPO in the formulation may need to be further optimised in future. Nonetheless, the differences in properties between the hydrogels highlighted that in future the properties of alg/PNIPAm could potentially be changed 'on demand' by adapting the UPO concentration in the formulation, though a better understanding of the relationship between enzyme concentration and material properties would need to be first established. It would also be informative for future applications to study the storability of UPO alg/PNIPAm, as it can be easier to store materials than make them *in situ* each time[393,394]. Notwithstanding, UPO alg/PNIPAm is a reproducible, biocompatible and 3D-printable material with beneficial mechanical properties and high thermosensitivity, such that it could have potential uses as a soft actuator or in temperature inducible systems like a temperature-induced drug delivery system[363,385,763,871].

To further extend the utility of UPO alg/PNIPAm for future applications, it was then investigated whether it could be used as the basis for a yeast ELM that produced proteins 'on demand'. To characterise the relative viability of cells within alg/PNIPAm, fluorescent laser scanning confocal microscopy was utilised with cells expressing EGFP and stains to detect cell membrane permeability and the cell wall. This technique was found to be fairly powerful, being able to detect live yeast cells at fairly high resolution within relatively thick and opaque hydrogel samples. While higher-throughput methods like flow cytometry and plating assays provide greater statistical power when measuring cell viability[844,846], these methods are not possible for hydrogels. Previously, several ELM studies characterised only the bulk protein production property of the cells within the material, such as bulk GFP fluorescence or the presence of a given protein in the surrounding medium[384,395,403]. However, such measurements are not as informative as directly observing several desired aspects of the cells

within the material using fluorescent confocal microscopy, as found here and by other ELM studies[391,394,759,872]. However, here it was found that the cells within the hydrogel could still be observed without the use of sectioning and by adding the fluorescent stains to the medium surrounding the hydrogel, thereby minimally interfering with the cell material system before imaging. Notwithstanding, as cell death and viability are complex[844,845,847], here it was found that using only EGFP and SYTOX fluorescence was insufficient to fully determine whether the yeast cell was dead or alive. Thus, in future to determine the viability of the cell it would be best to use a marker for metabolic activity in place of EGFP, such as FUN-1[861]. Also, to better determine cell death it would be beneficial to use additional markers to SYTOX, such as Annexin V which detects the presence of externalised phosphatidylserine during apoptosis[844,873,874], and also to view cell samples using higher resolution scanning and transmission electron microscopy to determine the state of their cellular structures, like the cell wall and organelles.

Using fluorescent microscopy, it was demonstrated for the first time, to the best our knowledge, that *K. phaffii* are relatively viable within moulded PNIPAm and alg/PNIPAm. While the viability of yeast within alginate has been well established[359,875], fewer studies have used acrylamide-based hydrogels[359,385,876], and most studies also utilise *S. cerevisiae* rather than *K. phaffii*[359,399,404]. Therefore, here the biocompatibility of both PNIPAm and alg/PNIPAm and their respective corresponding suitability for yeast ELM development were supported by demonstrating that these hydrogels were not toxic to yeast like *K. phaffii.* Nonetheless, in future it would be beneficial to also characterise the viability of *S. cerevisiae* within PNIPAm and alg/PNIPAm due to its more common use and potential in forming the 'on demand' alg/PNIPAm. Furthermore, for future applications of a yeast ELM it would ideally be able to be used over relatively long time periods and be able to be stored. Therefore, it would be important to characterise the viability of both species of yeast over a longer time period, for example up to two weeks, and after implementing various storage methods[394].

Despite the relatively high viability of *K. phaffii* observed within the moulded alginate, PNIPAm and alg/PNIPAm hydrogels, it was found that the average cell diameter within all three of these hydrogels was significantly smaller than that measured for *K. phaffii* cells in liquid culture. Additionally, it was also found that both *S. cerevisiae* and *K. phaffii* within printed alg/PNIPAm possessed an average cell diameter smaller than in liquid culture. While the control of cell size is a complex process and changes throughout the normal cell

cycle[839,856,877], the smaller average size of cells within the hydrogels could indicate that many cells had activated some sort of stress response[845,878]. This reduction in yeast cell size has also been observed in other studies using several different hydrogel matrices[359,879]. Sugianto *et al*.[403] suggested that the hydrogel environment could lead to yeast exhibiting lower growth and lower flux through their metabolic cycles. They posited that this was likely due to the mechanical forces the hydrogel exerts on the microbes, as well as variations in oxygen transport in the hydrogel. Therefore, possibly the stress response by *K. phaffii* and *S. cerevisiae* cells within hydrogels observed here may also have resulted from the mechanical force stress of the materials on the cells and the metabolic stress of not having enough access to oxygen. Interestingly, the average cell diameter of cells both with and without permeable cell membranes was not found to be significantly different, though this could result from the fact that the regulated cell death processes and cell stress processes in budding yeast can both lead to cell shrinkage[845,878]. Hence, as previously discussed, it would be informative in future to use additional fluorophores and higher resolution microscopy to better establish cell death and viability.

Nonetheless, using the current markers for presence of cell wall, plasma membrane integrity and cytosolic EGFP, the relative viability of cells within 3D-printed alg/PNIPAm was characterised to investigate the usability of 3D-printed alg/PNIPAm as an ELM. Unfortunately, it was found that the viability of *K. phaffii* in 3D-printed alg/PNIPAm was relatively low, though it was improved somewhat by adapting the alg/PNIPAm fabrication process. Preliminary evidence was also found that some cells were able to recover their cell wall and plasma membrane after the yeast laden hydrogel was incubated in medium for 24 hrs. This is supported by a growing number of studies which have shown that viable yeast cells may appear ‘dead’ using a membrane permeability dye when placed under conditions of stress, possibly resulting from a transient, recoverable membrane potential loss[880-885]. However, the ability of yeast cells within a hydrogel to repair their cell membranes after external stress would need to verified in future by tracking individual yeast cells over time with markers for cell membrane repair[862].

As a whole, it was found that potential adaptations to the alg/PNIPAm formulation to reduce cell toxicity, like reducing the concentration of toxic MBA or potential osmotic stress inducer $CaCl_2$, had little effect on *K phaffii* viability within the final hydrogel. Instead, it was found that the printing process itself likely had resulted in the relatively low cell viability. However, previously, Alper et *al.*[393,399] showed that both *S. cerevisiae* and *K. phaffii* respectively in

F127 bis-urethane methacrylate (F127-BUM) could be extrusion printed using a smaller nozzle diameter and higher pressure to yield ELMs with relatively high levels of protein production, and thus likely cell viability. Additionally, Saha *et al.*[759] demonstrated that *S. cerevisiae* in F127-dimethacrylate could be pneumatically 3D printed using higher pressures than were used in this study but with relatively low cell death following printing. However, as the curing methods for these respective hydrogels were very quick (<120 s), possibly the relatively low levels of protein production and cell viability observed in the yeast alg/PNIPAm ELMs results from the combination of both printing and the relatively long curing time of one hour. Therefore, in future it would be beneficial to investigate if shorter curing times could increase yeast cell viability whilst still resulting in hydrogel formation. Furthermore, it could be investigated in future if the viability of the yeast could be improved following capsulation because a study by Li *et al*.[760] found very high viability of encapsulated *S. cerevisiae* cells post printing using a similar pneumatic-driven extrusion 3D printing method. Additionally, as it was found here that the cells potentially showed better recovery after the stresses of the printing process in more complex medium than minimal medium, perhaps the even richer medium of yeast extract peptone dextrose (YPD) could be utilised in future for initial recovery after printing. It would also be informative in future to characterise the viability of *S. cerevisiae* in printed alg/PNIPAm because of its wider range of synthetic biology tools, which could widen the scope of future applications.

Despite the relatively low viability observed of *K. phaffii* cells within printed alg/PNIPAm, it was nevertheless investigated whether protein production could be induced by cells within printed alg/PNIPAm to further characterise its ability to act as an ELM bioreactor or biosensor. To this end, both *K. phaffii* and *S. cerevisiae*-laden alg/PNIPAm hydrogels were examined, as while *K. phaffii* tends to produce greater yields of recombinant protein, *S. cerevisiae* is more commonly used, hence might be more accessible for some, and has had more synthetic biology tools developed for it. It was found that production of both cytosolic EGFP and secreted UPO could be induced by both *K. phaffii* and *S. cerevisiae* within alg/PNIPAm. This demonstrated that printed yeast-laden alg/PNIPAm has potential for use as an ELM both as a biosensor, as demonstrated by the expression of the reporter protein EGFP, and as an ‘on demand’ bioreactor, as demonstrated by the expression and secretion of UPO with relatively broad catalytic abilities. However, it was found that *S. cerevisiae* expressed more EGFP and UPO respectively within the printed alg/PNIPAm than *K. phaffii*. As recombinant protein production of UPO was previously shown to be greater in *K. phaffii*, it is

possible that the increased protein production by *S. cerevisiae* within the hydrogel results from it having a greater number of metabolically active cells. This perhaps suggests that *S. cerevisiae* was better able to withstand and/or recover from the stress of printing and curing than *K. phaffii.* Thus, *S. cerevisiae* cells may be more suitable for use within 3D-printed alg/PNIPAm in future. However, the viability of *K. phaffii,* as well as *S. cerevisiae,* could likely be improved by reducing the printing pressure, which may improve protein production by *K. phaffii* within the hydrogel.

Interestingly, it was observed that many cells of both yeast species within alg/PNIPAm in both the EGFP-expressing and UPO-expressing populations had compromised plasma membranes, and indeed even that cells which had been induced to express EGFP still showed a compromised plasma membrane. Commonly, if a yeast cell possesses a compromised plasma membrane, as shown using hydrophobic marker fluorophores, it is considered 'dead'[844,845]. However, the presence of EGFP fluorescence observed here would suggest that the cell was metabolically active such that it could be induced to express EGFP. While it is possible that the cells may have produced EGFP and then later 'died' before observation, the relatively high number of cells with SYTOX fluorescence both before and after protein induction could suggest that some cells with compromised plasma membranes were nevertheless metabolically active enough to recombinantly produce EGFP, and thus were not 'dead'. A similar finding was observed by Rocchi *et al.*[886], where yeast cells which had stained positive with propidium iodide nevertheless had produced their chemical of interest, thiamine, and had retained this chemical intracellularly. Therefore, the preliminary findings here provide support to the growing number of studies which show that for budding yeast cells, unless the cells are characterised using multiple markers and experiments to confirm death, they may only be confirmed to be truly dead when completely lysed or denatured[844,847,873,884]. Thus, the use of plasma membrane permeability stains alone to determine yeast cell death should be strongly cautioned against in future.

In summary, it was demonstrated that 3D-printed alg/PNIPAm could be formed containing *K. phaffii* and *S. cerevisiae* respectively, and that protein production could be induced using chemical mediators in both yeast species respectively within alg/PNIPAm. Therefore, the potential of yeast-laden alg/PNIPAm to act as an ELM bioreactor or biosensor was demonstrated. However, the amount of protein produced by the 3D-printed alg/PNIPAm needs to be further improved in future. Furthermore, the effect of the yeast cells on the material properties of alg/PNIPAm needs to be characterised to inform future applications, as

it is likely that the cells may have affected the structure and thus the properties of the hydrogel[385,738,887]. Additionally, it would be informative to characterise the relative viability of the yeast cells at a temperature slightly above the LCST of the hydrogel to verify whether the thermocontractive property of the material could be utilised without the cells experiencing too much heat stress. Also, an important factor to consider when developing an ELM is that the cells within are contained and cannot 'escape' into the environment[381,875,888]. Therefore, in future it would also be important to test whether the surrounding incubation medium of the yeast alg/PNIPAm hydrogels contains living yeast cells, for example, by using a plating assay[844,847,855].

Altogether, it was demonstrated that alg/PNIPAm could be polymerised using an unspecific peroxygenase expressed recombinantly as a single isoform by *K. phaffii* to form a material with very similar thermosensitive properties and somewhat similar mechanical properties to alg/PNIPAm polymerised by commercial plant-derived HRP. It was then shown that alg/PNIPAm could be printed and polymerised with *S. cerevisiae* and *K. phaffii* and that these respective yeasts were able to produce both EGFP and UPO respectively 'on demand'. Hence, demonstrating that yeast laden alg/PNIPAm could form an ELM bioreactor or biosensor. However, the viability and protein production yield of the cells within 3D-printed alg/PNIPAm was lower than desired, thus, in future perhaps this could be improved by decreasing the printing resolution. Furthermore, future development of the alg/PNIPAm ELM system would require further characterisation of its storability, reusability, microbial containment and material properties.

## 4.5 Methods

### *4.5.1 Transformation and expression of UPO PaDa-I in Saccharomyces cerevisiae*

*Agrocybe aegerita* unspecific peroxygenase (UPO) variant PaDa-I was designed by Molina-Espeja *et al*.[290]. The sequence for UPO PaDa-I (hereafter UPO) with its signal peptide was codon optimised for *S. cerevisiae* expression using GeneArt (ThermoFisher), avoiding HindIII and XbaI cleavage sequences. The HindIII site was added prior to the N-terminus and the XbaI site was added posterior to the C-terminus. This construct was synthesised by Eurofins Genomics in the pEX-A258 vector. Restriction enzyme cloning was utilised to change the UPO construct into the pYES2/CT vector, which was kindly provided by Prof. Paul Curnow (University of Bristol). The sequence for UPO was inserted behind the *GAL1*

promoter for inducible expression. The HindIII, XbaI and ligase enzymes were sourced from NEB and restriction digest and ligation was conducted according to the supplier's protocols. Ligation product was transformed into OneShot TOP10 chemically competent *E. coli* (ThermoFisher). Plasmid was extracted from transformants, and cloning was confirmed by sequencing (Eurofins Genomics).

UPO-pYES2/CT was transformed into *S. cerevisiae* YPH499 using a protocol adapted from Gietz & Schiestl[889]. *S. cerevisiae* YPH499 and transformation protocol was kindly provided by Dr Ian Prosser (University of Bristol). Eight transformants were checked for presence of the UPO gene using colony PCR. Three of the transformants, which were confirmed by colony PCR to contain the UPO gene, were selected for expression trials. A colony of each transformant was each used to inoculate 10 mL of SC -ura (6.7 g/L yeast nitrogen base, 1.92 g/L Kaiser synthetic complete dropout -uracil, 2% w/v D-glucose). Incubated for 48 hrs, 28 °C, 210 rpm. Added aliquot of each respective culture to 50 mL SC -ura 0.1% glucose such that had $O.D._{600}$ of 0.25. Incubated for ~6 hrs, 28 °C, 210 rpm. Then each of the 50 mL cultures was added to 450 mL of expression medium[290] (1% w/v yeast extract, 2% w/v peptone, 2% w/v D-galactose, 71 mM potassium phosphate (pH 6.0), 6.5 mM $MgSO_4$, 0.1 mg/mL bovine hemoglobin, 3.33% v/v ethanol). The cultures were incubated at 28 °C, 210 rpm for 72 hrs, with samples taken every 24 hrs and frozen at -20 °C. Samples were defrosted on ice, before being centrifuged and the supernatant extracted. Relative peroxidase activity and thus expression was determined by ABTS assay of the respective supernatants (0.3 mM ABTS, 0.3% w/w $H_2O_2$, 1:10 dilution of sample, 50/100 mM citrate/phosphate buffer, pH 4.5). Absorbance was measured at 418 nm using a plate reader. The transformant with the highest absorbance was selected, and used to make a freezer stock by combining saturated culture with glycerol (30% v/v final), aliquoting, snap freezing in liquid nitrogen, and storing at -70 °C.

Large-scale expression was conducted as described for the expression trials, with 72 hrs used for the expression time post induction. Cultures were centrifuged at 5000 rpm, 15 min to pellet, then the supernatant collected.

### *4.5.2 Transformation and expression of UPO PaDa-I in K. phaffii*

The gene construct for UPO was transferred from the pYESCT2 vector into the pPICZ A vector (ThermoFisher) behind the inducible *AOX1* promoter using Gibson assembly[890].

Primer design was aided by the Amplifx software[891] and primers were synthesised by Eurofins Genomics. Primers are listed in Suppl. Fig. 4.1. Gibson assembly mix was sourced from NEB. Cloning success was confirmed by sequencing (Eurofins Genomics).

Linearised UPO-pPICZ A plasmid was generated at the PmeI restriction site using restriction digest with PmeI (NEB) according to the supplier's protocol. Complete digestion was confirmed by gel electrophoresis, then the linearized plasmid was extracted from the reaction mixture using a Wizard PCR cleanup kit (ProMega). Competent cells of *K. phaffii* X33 and KM71H (Invitrogen/ThermoFisher) were made and linearized UPO-pPICZ A was transformed into them according to the supplier's protocol. Following transformation, 10 colonies of each strain were ascertained to contain the UPO gene by colony PCR. For those shown to possess the gene, two were selected for UPO expression trials. A single colony of each of the two transformants was used to inoculate 5 mL BMGY (1% w/v yeast extract, 2% w/v peptone, 100 mM potassium phosphate, pH 6.0, 1.34% w/v yeast nitrogen base, $4x10^{-5}$% v/v biotin, 1% v/v glycerol) medium. Cultures were incubated overnight at 30 °C, 210 rpm. Then, the cultures were pelleted by centrifugation before being resuspended in 5 mL BMMY (same as BMGY except 0.8% v/v methanol rather than glycerol) and incubated for 72 hrs, 30 °C, 210 rpm. Methanol (0.8% v/v) was added every 24 hrs during this period. Cultures were then pelleted by centrifugation, and the supernatant collected. Peroxidase activity, and thus relative UPO expression, was measured by ABTS assay as previously described. The X33 and KM71H transformants with the highest respective UPO expression were selected for future experiments, and glycerol stocks made as previously described for *S. cerevisiae*.

A large-scale expression trial was conducted to find the best expression time and strain for UPO in X33 and KM71H respectively using the respective strain expression protocols recommended for each strain from the supplier's protocol. For both the strains, it was found that 72 hrs led to the greatest UPO expression per mL of lysate, as measured by ABTS assay. Comparing the strains, KM71H was found to have much greater UPO expression, as measured by peroxidase activity, than X33.

Large-scale UPO expression was conducted by inoculating 50 mL BMGY with a single UPO KM71H colony and incubating overnight, 30 °C, 210 rpm. Then 5 mL aliquots of the overnight culture were each used to inoculate eight 500 mL BMGY cultures. These were incubated overnight, 30 °C, 210 rpm. The overnight cultures were then pelleted (4000 rpm, 15 min), before four pellets were all resuspended in 400 mL BMMY medium with 10 mg/mL

bovine hemoglobin (Hb). This resulted in two 400 mL BMMY + Hb cultures, which were incubated for 72 hrs, 210 rpm, 30 °C. Methanol was added every 24 hrs to 0.8% v/v. Cultures were then centrifuges at 4000 rpm, 15 min, 4 °C to pellet, and the supernatant collected.

### *4.5.3 Purification of UPO PaDa-I*

The supernatant was concentrated using ammonium sulfate precipitation (90% saturation was determined best by experiment) before being resuspended in cation exchange chromatography (CEX) buffer A (20 mM citric acid, pH 3.3) and further buffer-exchanged into CEX buffer A through dialysis. It should be noted that the following CEX purification protocol was adapted from Martin-Diaz *et al.*[779]. After buffer exchange, the supernatant was loaded onto a pre-equilibrated 5 mL HiTrap CaptoS (Cytiva) and protein eluted using 50% buffer B (20 mM citric acid, 1 M NaCl, pH 3.3). Fractions with absorbance at 418 nm were collected and concentrated, then loaded onto a HiLoad 26/600 Superdex pg column pre-equilibrated with buffer C (10 mM potassium phosphate, 0.15 M NaCl, pH 7.0). SDS-PAGE was run of the SEC fractions according to standard protocol; however, protein samples were not heated due to the glycosylation of UPO. Fractions with relatively high purity and ABTS activity were combined, then buffer exchanged into buffer D through dialysis (10 mM potassium phosphate, pH 7.0). Yield of purified protein was determined using the BCA assay according to the supplier's (ThermoFisher) protocol with bovine hemoglobin as standards.

Molecular weight of *S. cerevisiae* UPO was estimated from both the SDS-PAGE gel (Suppl. Fig. 4.2) and the weight of wildtype *A. aegerita* UPO[290] to be ~46 kDa. Molecular weight of *K. phaffii* UPO was measured to be ~44.2 kDa by MALDI-TOF MS (Suppl. Fig. 4.3), which was conducted on a UltrafleXtreme (Bruker Daltonics) calibrated using bovine serine albumin and using Super-DHB as a matrix. Dr Chris Arthur (University of Bristol) kindly provided assistance for the MALDI-TOF experiment.

### *4.5.4 Absorbance and heme content of UPO PaDa-I*

The absorbance from 200-800 nm of the respective UPO samples was measured using a CARY UV/Vis spectrometer at 60 nm/min in a 1 cm pathlength quartz cuvette. The *S. cerevisiae* UPO was measured at 2.3 mg/mL and the *K. phaffii* UPO was measured at 0.14 mg/mL. Blanks (buffer D) were recorded for both spectra respectively and subtracted from each of the protein spectra. The Reinheitszahl value for both UPOs was calculated by dividing the Soret peak at 418 nm by the peak at 280 nm.

Heme concentration in the protein was determined by pyridine assay. Measurements were made using a CARY UV/Vis spectrometer at 100 nm/min from 500 to 600 nm. First, the blank spectrum was recorded (buffer D, 100 mM NaOH, and 10% v/v pyridine). Then, the oxidised spectrum was measured of the UPO sample was measured, before a small amount of dithionite was added and the reduced spectrum measured. Dithionite was added twice more, and the spectrum measured until no further gain in absorbance was observed. The concentration of type B heme was determined by the following equation:

$$[heme](mM) = \frac{absorbance\ of\ reduced\ heme\ at\ 556\ nm - absorbance\ of\ oxidised\ heme\ at\ 540\ nm}{23.98}$$

Approximate percentage of heme incorporation in the protein sample was then determined by calculating the concentration of protein in the assay, then dividing the heme concentration by the protein concentration and multiplying by 100.

### *4.5.5 Specific ABTS activity of UPO PaDa-I and A. rusticana HRP*

The specific activity against ABTS of UPO expressed in *S. cerevisiae* and *K. phaffii* respectively and a commercial sample of *A. rusticana* HRP was measured using 5 nM of the respective protein in 50/100 mM citric acid/phosphate buffer, pH 4.5 with 15 mM ABTS and 0.03% w/w $H_2O_2$. Absorbance was measured at 25 °C using a plate reader at 418 nm with 3 s time intervals for 15 min. Pathlength of 0.56 cm and extinction coefficient of 36000 $M^{-1}cm^{-1}$ for the ABTS radical[290,741] were used to calculate the specific activity.

### *4.5.6 Circular dichroism of UPO PaDa-I expressed in K. phaffii and A. rusticana HRP*

Circular dichroism was conducted on UPO and a commercial sample of HRP (source x) in di$H_2O$ using a 10 mm pathlength quartz cuvette. First, the spectrum of di$H_2O$ at 25 °C was measured from 190 – 280 nm at 5 nm/min with 1 nm interval and 1 nm bandwidth to determine the baseline. Using the same conditions, the spectrum of UPO and commercial HRP was then measured. Mean residue ellipticity was calculated and secondary structure predicted by the CDSSTR, SELCON and K2D methods through the DichroWeb server[740] using all relevant sets when appropriate. The prediction with the lowest NRMSD was selected as the 'best'.

Change in the CD signal with temperature was measured from 25 °C to 95 °C with 0.5 °C intervals for UPO and 1 °C intervals for HRP. The spectrum of UPO was monitored from 190-260 nm, and that of HRP monitored at 208 and 222 nm. Thermal unfolding transition

temperature was determined by fitting a sigmoidal curve at each wavelength with a change in signal, then calculating the average.

### *4.5.7 Enzyme kinetics of UPO PaDa-I expressed in K. phaffii and A. rusticana HRP with ABTS*

Kinetics assays were conducted as described in Chapter 3.5.10, however, the buffer was changed to 50/100 mM citric acid/phosphate, pH 4.5 for comparison to other kinetic studies of UPO[290,293]. Working concentrations of 1 nM HRP (assumed M.W. of 44 kDa[602]) and 4.4 nM of UPO were used. Results were analysed as described in Chapter 3.5.10.

### *4.5.8 Kinetics of PNIPAm formation*

First it was established that PNIPAm formation could be tracked by changes in absorbance/scattering. To this extent, the spectrum of PNIPAm (see Chapter 3.5.11 Table 2.3 for formulation) was measured prior to gelation without acetylacetone added from 200 to 800 nm at 60 nm/s, 0.5 nm intervals in a 10 mm plastic cuvette. Acetylacetone was then added, the liquid mixed and 30 min were allowed for gelation. Post gelation, the spectrum was then measured again under the same conditions. This showed that ~600-800 nm would show the greatest change in absorbance/scattering during gelation. Resultingly, the change in absorbance/scattering was measured at 700 nm over time at 25 °C using a plate reader by adding acetylacetone immediately prior to measurement.

### *4.5.9 Characterisation of thermosensitive properties of UPO alg/PNIPAm compared to HRP alg/PNIPAm*

The HRP alg/PNIPAm gel ink was prepared as described in Klemperer *et al.*[361] and the optimised UPO alg/PNIPAm formulation is shown in Table 4.2. Boxes of 4x15x3 mm were printed using a Cellink INKREDIBLE+ bioprinter using custom gcode, which was prepared from a CAD made in Autodesk Fusion 360 and sliced in Cellink Heartware using custom settings before being post-processed using a custom Python script. A 20 gauge nozzle was used during printing with pressure of ~18 kPa for both UPO and HRP alg/PNIPAm formulations. Hydrogels were cured in 300 mM calcium chloride with 10 mg/mL D-glucose for at least 3 hours, before being stored in 10 mM calcium chloride at room temperature until use.

| Component | Concentration |
|---|---|
| *N*-isopropylacrylamide | 15% wt/wt |
| *N,N'*-Methylenebisacrylamide | 7.5 mg/mL |
| sodium alginate | 3.5% wt/wt |
| xanthan gum | 0.75% wt/wt |
| D-glucose | 1 mg/mL |
| glucose oxidase (type VII from *Aspergillus niger*) | 5.12 mg/mL |
| UPO PaDa-I (expressed in *K. phaffii*) | 210 µg/mL |
| acetylacetone | 1% v/v |
| calcium chloride | 25 mM |

**Table 4.2:** Optimised hydrogel formulation for UPO alg/PNIPAm

Note that all the thermosensitivity experiments were conducted by placing the ~4x15x3 mm hydrogel in 1 mL of di$H_2O$ within a small Eppendorf tube. This tube was then incubated at the desired temperature in a heat block or at room temperature. Weights of hydrogels were measured after briefly drying the gel on filter paper using sensitive weighing scales. After the experiment, hydrogel samples were dried in a drying oven overnight, before their dried weight was measured and recorded.

The deswelling ('contraction') experiment of the hydrogels at 60 °C was conducted as described in Chapter 3.5.11. The swelling ratio formula is also shown in Chapter 3.5.11. Controls were also measured by incubating samples at room temperature (RT). The kinetics of deswelling experiment was conducted in a similar manner to the deswelling at 60 °C, except that the weight of the samples was measured at given time intervals during incubation at 60 °C. For HRP alg/PNIPAm, further control experiments were conducted where firstly the weight was measured for room temperature incubation, and secondly where dried gel samples were incubated in di$H_2O$ at room temperature and the weight measured at given time intervals. The change in the amount of deswelling of the HRP and UPO alg/PNIPAm hydrogels with temperature was measured by incubating samples at the given temperature in a heat block for 2 hrs (to reach equilibrium at all the measured temperatures) before measuring their weight. The repeatability of the 'contraction' of the HRP and UPO

alg/PNIPAm hydrogels was measured by incubating samples at 60 °C for 30 min, or room temperature for controls, then measuring their weight before and after the incubation. Post incubation, samples were incubated at room temperature overnight to 'recover'. This was repeated once a day for 11 days.

### *4.5.10 Compression and tensile testing of UPO and HRP alg/PNIPAm hydrogels.*

HRP alg/PNIPAm, UPO alg/PNIPAm and alginate single network gels were prepared and printed as described in the previous section, except that 10 mm diameter by 5 mm height cylinders were made for compression testing and 'dog bone' shapes were prepared for tensile testing. Gcode for the cylinders was made as described in the previous section and gcode for the 'dog bone' shape was kindly provided by Dr Charles de Kergariou (University of Bristol). HRP PNIPAm single network compression testing samples were cured in PLA plastic moulds with 10 mm inner diameter and 5 mm inner height.

Compression testing was conducted and results were analysed as described in Chapter 3.5.11. Tensile testing was conducted using a Starrett FMS500 with a 10 N load. Tensile testing for HRP alg/PNIPAm was conducted by Dr Charles de Kergariou and that for UPO alg/PNIPAm and alginate single network were conducted by the author in the same manner. Photographs of samples were taken for digital measurement, which was conducted by Dr Charles de Kergariou. Samples were then very carefully placed in the 'grips' so as not to stretch or compress them in any way. A tensile load was then applied at 1 mm/min. Data were analysed for HRP alg/PNIPAm by Dr Charles de Kergariou and data for UPO alg/PNIPAm and alginate single network were analysed by the author using custom Python and MatLab scripts.

### *4.5.11 Transformation and expression of EGFP in K. phaffii and S. cerevisiae*

The gene for enhanced green fluorescent protein (EGFP)[892] was codon optimised for *S. cerevisiae* expression using the GeneArt webserver (ThermoFisher) and synthesised in the pEX-A128 vector by Eurofins Genomics. Restriction digest cloning was used to transfer the EGFP gene into the pYESCT2 vector for *S. cerevisiae* inducible expression (*GAL1* promoter) as described in section 4.5.1. Gibson assembly was used to transfer the EGFP construct from the pYESCT2 vector into pPICZ A for *K. phaffii* inducible expression (*AOX1* promoter). Primer design was aided by the Amplifx software[891] and primers were synthesised by Eurofins Genomics. Gibson assembly mix and the HindIII, XbaI and ligase enzymes were sourced from NEB. Cloning success was confirmed by sequencing (Eurofins Genomics).

Transformation for the respective EGFP plasmids into *S. cerevisiae* and *Komagtaella sp.* X33 and KM71H was conducted as described previously (see sections 4.5.1 & 4.5.2).

Following transformation, 10 colonies were randomly selected for each yeast species and strain, and tested to have been successfully transformed using colony PCR. Colonies which were confirmed to likely contain EGFP were then compared for relative levels of EGFP using an expression trial protocol similar to those in sections 4.5.1 & 4.5.2. Fluorescence of 1 mL aliquots of each culture at various time points (24, 48 and 72 hrs) was measured in a black 96-well plate in a plate reader with excitation of 468 nm and emission of 513 nm (these values used to avoid overlap in signal between excitation and emission due to generally poor bandwidth of many plate readers). The colony which contained the highest level of EGFP expression from the expression trial was used for future experiments. Cultures of the best EGFP-expressing colony of *S. cerevisiae*, *K. phaffii* X33 and *K. phaffii* KM71H were made into glycerol stocks for storage as described in sections 4.5.1 & 4.5.2.

### *4.5.12 Confocal microscopy of liquid culture K. phaffii and S. cerevisiae*

Expression of EGFP in both *K. phaffii* strains and *S. cerevisiae* was conducted as described previously in sections 4.5.1 & 4.5.2. It should be noted that cultures were also incubated for 72 hrs post induction for EGFP expression as for UPO expression. After this time, a 1 mL aliquot of each culture was taken and concentrated by centrifugation. SYTOX Deep Red (Invitrogen) was then added to each culture at a working concentration of 1 μM, and the cultures were incubated for 30 min as recommended by the manufacturer's protocol. Meanwhile, under sterile conditions a thin layer of 50% agar was added to 9 microscope glass slides and dried. Following the 30 min incubation, under sterile conditions 20 μL aliquots of each culture was pipetted onto 3 glass slides respectively. Then, 0.5 μL of Calcofluor White M2R (Sigma-Aldrich) was pipetted onto each slide following the manufacturer's protocol. The slides were then sealed with a type I glass coverslip using clear nail varnish. Slides were viewed using a Leica SPE single channel confocal laser scanning microscope. Initial training and access to this microscope was kindly provided by the University of Bristol Wolfson Bioimaging Facility. The 63x lens was used on the microscope, and the 635 nm, 488 nm and 405 nm lasers were used to excite the Sytox Deep Red, EGFP and Calcofluor White M2R dye respectively. Emission was collected from 646-800 nm, 498-584 nm and 410-485 nm respectively. A laser power of 10% was used for the 405 and 488 nm lasers, and 15% for the 635 nm laser. Gain used was typically ~700. Zoom between 1 to 2x was used, with the

sampling for each zoom selected based on the Nyquist resolution. Typically, z stacks between 10-40 µm were imaged.

### *4.5.13 Confocal microscopy image post-processing and analysis*

The Fiji distribution v2.16.0[893] of ImageJ2[894] was used to view and process the images collected using confocal microscopy. A typical image post processing of an image proceeded as such: a maximum and minimum z projection was calculated, and the minimum was subtracted from the maximum. The subtract background filter was used and the despeckle noise filter was used. The channels were split, and additional filters (Gaussian blur or unsharp mask) were used if they were observed to improve image quality.

For quantifying the number of cells and measuring cell diameter, the binary threshold of the image was adjusted until most cells were visible and then a watershed operation was used to attempt to split any joined cells. The measurement function in Fiji was then utilised, and any cells which failed the threshold or watershed were manually selected using the ellipse tool and measured.

For generation of images, the look up table of each channel was changed to blue for 405 nm excitation channel, green for 488 nm excitation channel and red for 635 nm excitation channel. Contrast was adjusted if judged necessary. For the merged image, the merge function was used to merge the three channels. For all images, scale bars were added using the scale bar tool.

### *4.5.14 Formation of moulded alginate, PNIPAm and alg/PNIPAm hydrogels with K. phaffii*

The formulation used for alginate is shown below in Table 4.3, the formulation used for PNIPAm is shown below in Table 4.4, and the formulation used for alg/PNIPAm was listed in Table 4.2 in section 4.5.9. However, unlike previously the components of each respective formulation were either UV-sterilised (powders, dual asymmetric centrifuge (DAC) pots and moulds) or filter-sterilised through a 0.2 µm filter (liquids). Furthermore, *K. phaffii* X33, which had been induced to express EGFP as previously described, was added to each formulation at a concentration of ~$10^7$ cells/mL. The relative number of cells was estimated using $O.D._{600}$ of 1.0 = $5x10^7$ cells for *K. phaffii*[895] and $O.D._{600}$ of 1.0 = $3x10^7$ cells for *S. cerevisiae*[896].

The components and cells of each formulation were mixed by dual asymmetric centrifuge (DAC) before each ink was transferred into rectangular PVA plastic moulds, with inner diameters of 4x15 mm, under sterile conditions. The moulds were on top of sterile Petri dishes. Then the appropriate curing solution, sterilized by filtration or autoclaving as appropriate, was added to the alginate and alg/PNIPAm formulations: 300 mM $CaCl_2$ for alginate and 300 mM $CaCl_2$ with 5 mg/mL D-glucose for alg/PNIPAm. Hydrogels were cured for 1 hr before being carefully removed from their moulds and transferred to BMGY medium under sterile conditions. A single hydrogel per well in a 12 well plate was used. Biogels were then incubated at 28 °C, 220 rpm for 3.5 hrs, before a working concentration of 2 µm SYTOX Deep Red and a volume of 0.5 µL of Calcofluor White M2R was added to each well. Hydrogels were incubated for a further 30 min at 28 °C, 220 rpm before being transferred to glass microscopy slides under sterile conditions. The slides were sealed with a type I glass coverslip using clear nail varnish. Three hydrogel samples per formulation were used. Confocal microscopy was conducted as described in section 4.5.12, except that laser powers of 12% were used for the 405 nm and 488 nm lasers, and 20% for the 635 nm laser. Microscopy data were viewed and analysed as described in section 4.5.13.

| **Component** | **Concentration** |
|---|---|
| sodium alginate | 3.5% wt/wt |
| xanthan gum | 0.75% wt/wt |
| calcium chloride | 25 mM |
| *K. phaffii* | ~$10^7$ cells/mL |

**Table 4.3:** Formulation used for creating an alginate hydrogel

| **Component** | **Concentration** |
|---|---|
| *N*-isopropylacrylamide | 15% wt/wt |
| *N,N'*-Methylenebisacrylamide | 7.5 mg/mL |
| glucose oxidase (type VII from *Aspergillus niger*) | 5.12 mg/mL |
| UPO PaDa-I (expressed in *K. phaffii*) | 210 µg/mL |
| D-glucose | 1 mg/mL |
| acetylacetone | 1% v/v |

| | |
|---|---|
| *K. phaffii* | ~$10^7$ cells/mL |

**Table 4.4:** Formulation used for creating a PNIPAm hydrogel

### *4.5.15 Formation of 3D-printed alg/PNIPAm with K. phaffii*

Firstly, an overnight culture of *K. phaffii* X33 was used to form alg/PNIPAm as described in section 4.5.14, however, two concentrations of cells were tested: $10^6$ cells/mL and $10^7$ cells/mL. The ink was passed through syringes to break up any large particles, before being printed into 4x15x3 boxes and 4x15 rectangles using the Cellink INKREDIBLE+ bioprinter with custom gcode. The bioprinter was sterilized before use with 70% ethanol, however, in future it would be much better to use a bioprinter in a biosafety cabinet. Printing syringes, stoppers and 20 gauge printing nozzles were sterilized before use by UV irradiation. Following printing, hydrogels were cured for ~1 hr in 300 mM $CaCl_2$ with 5 mg/mL glucose, before being transferred into either BMG (100 mM potassium phosphate, pH 6.0, 1.34% w/v yeast nitrogen base, $4x10^{-5}$% v/v biotin, 1% v/v glycerol) or BMGY medium within 12 well plates as described in section 4.5.14. Biogels were incubated for 23.5 hrs at 28 °C, 220 rpm. For the rectangular biogels, SYTOX Deep Red was added to a working concentration of 2 µm and 5 µL of Calcofluor White M2R was added. Rectangles were incubated for a further 30 min, before three rectangles per cell seeding concentration and medium were transferred to microscopy slides and viewed using confocal microscopy as described in section 4.5.14. Microscopy data were viewed and analysed as described in section 4.5.13. The box-shaped biogels were photographed using a phone camera and handled using tweezers.

For the second attempt at printing *K. phaffii* laden alg/PNIPAm, EGFP-expressing *K. phaffii* X33 was added to the alg/PNIPAm ink at a concentration of ~$10^7$ cells/mL after all other components had been mixed and passed through syringes to break up large particles. Two formulations were attempted, that used previously containing 25 mM $CaCl_2$ and one without $CaCl_2$. The inks were then mixed by dual asymmetric centrifuge at 1700 rpm (3500 rpm was used in previous attempts), before being loaded into a sterile printing syringe and printed into rectangles as described previously. The printed constructs were cured using 150 mM $CaCl_2$ with 5 mg/mL D-glucose for 1 hr, before being transferred to BMGY as before. Biogels were either incubated for 3.5 or 23.5 hrs at 28 °C, 220 rpm before SYTOX Deep Red dye and Calcofluor White M2R were added as described previously. Following a further 30 min incubation, three biogels for each respective formulation and timepoint were viewed using confocal microscopy as described in section 4.5.14. Microscopy data were viewed and analysed as described in section 4.5.13.

### *4.5.16 Protein induction of S. cerevisiae and K. phaffii within alg/PNIPAm*

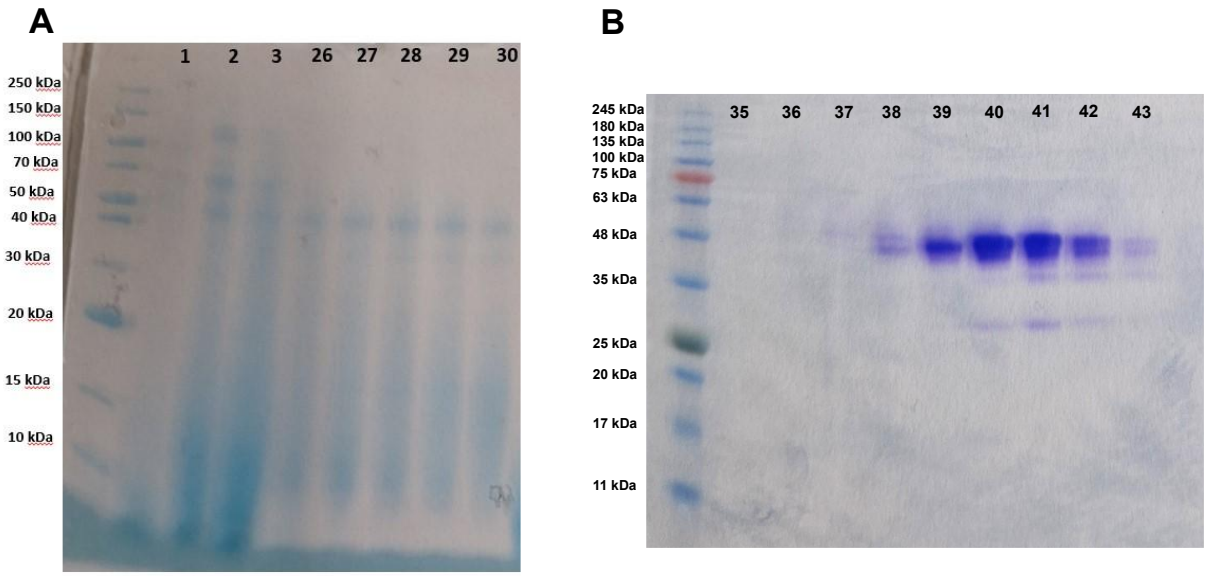
Overnight cultures were grown of *S. cerevisiae* containing the expression system for EGFP and secretory expression system for UPO respectively in SC -Ura. Overnight cultures were grown of *K. phaffii* X33 containing the expression system for EGFP and of *K. phaffii* KM71H containing the secretory expression system for UPO respectively in BMGY. Alg/PNIPAm of each yeast and expression system were prepared, printed and cured as described previously (section 4.5.15 paragraph 2). Hydrogels were transferred to BMGY medium and incubated for ~24 hrs at 28 °C, 220 rpm. Three hydrogels of each yeast type and expression system were selected and stained with SYTOX Deep Red and Calcofluor White M2R as previously described (section 4.5.15) before being observed using confocal microscopy (section 4.5.14). Meanwhile, the remaining hydrogels were washed with di$H_2O$ before their media were changed to BMMY for *K. phaffii* or expression medium for *S. cerevisiae* (section 4.5.1). The cell-laden hydrogels were incubated for a further 72 hrs, with 0.8 % v/v methanol added to the media of the hydrogels containing *K. phaffii* every 24 hrs. Three samples of each yeast species and expression system were then selected and stained as described previously, before being observed using confocal microscopy. For the UPO expression systems, three further samples were selected for each yeast species and 4*20 µL aliquots from each sample were characterised by ABTS assay for peroxidase activity (0.3 mM ABTS, 0.03 w/w% H2O2, 1/10 dilution of sample, 50/100 mM citrate/phosphate buffer). Aliquots were also taken from the medium surrounding one the *S. cerevisiae* EGFP-expressing hydrogels and also characterised at the same time by ABTS assay as a negative control. The assay plate was incubated at room temperature for 1 hr, before the absorbance at 403 nm was measured using a plate reader.

## 4.6 Supplementary

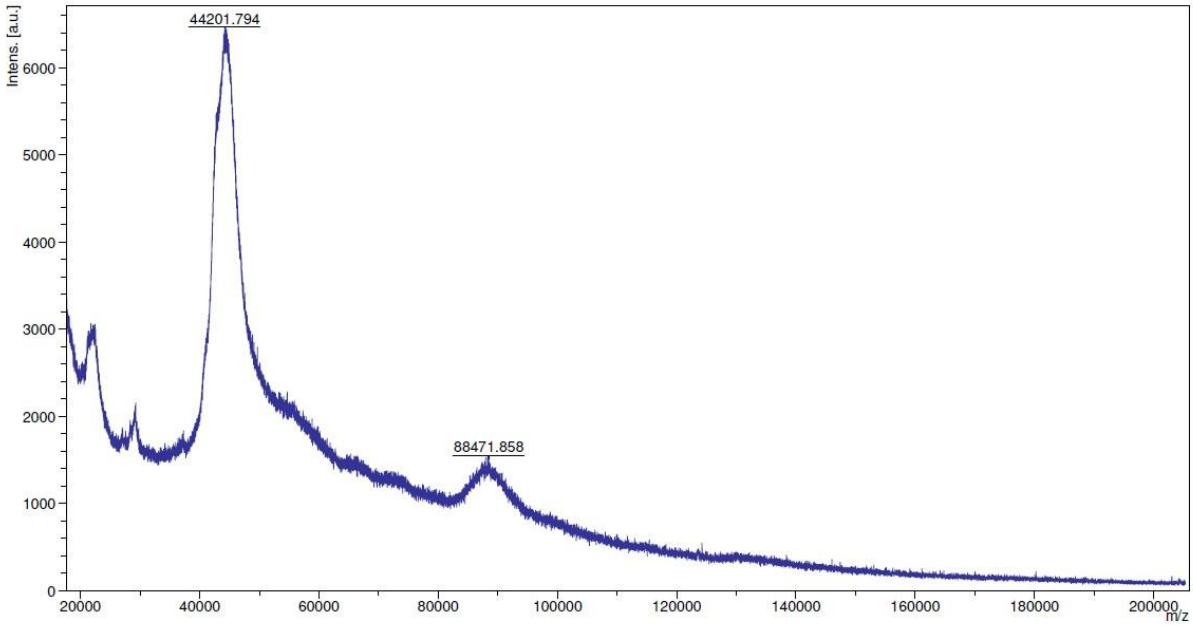
**Suppl. table 4.1***:* List of primers used for cloning by Gibson Assembly of the UPO and EGFP genes respectively into the pPICZ A vector.

| Name | Sequence (5'-3') |
|---|---|
| ppicza_A_F | GTTTGTAGCCTTAGACATGACTGTTCCTCAGTT |
| ppicza_A_R | CGTTTCGAATAATTAGTTGTTTTTTGATCTTCTCAAGTTGTCG |
| UPO_Gib_ppicz_F | GAAGATCAAAAAACAACTAATTATTCGAAACGATGAAGTACTTTCCTTTGTTCCCAACC |
| UPO_Gib_ppicz_R | CTGAGGAACAGTCATGTCTAAGGCTACAAACTTAGTCTCTACCGTATGGAAAAACTTGAGTAC |
| EGFP_Gib_ppicz_F | GAAGATCAAAAAACAACTAATTATTCGAAACGATGGTGTCTAAAGGTGAAGAGTTGTTCACTGG |
| EGFP_Gib_ppicz_R | CTGAGGAACAGTCATGTCTAAGGCTACAAACTTACTTGTATAATTCGTCCATACCTAAGGTAATACCAGCAGC |

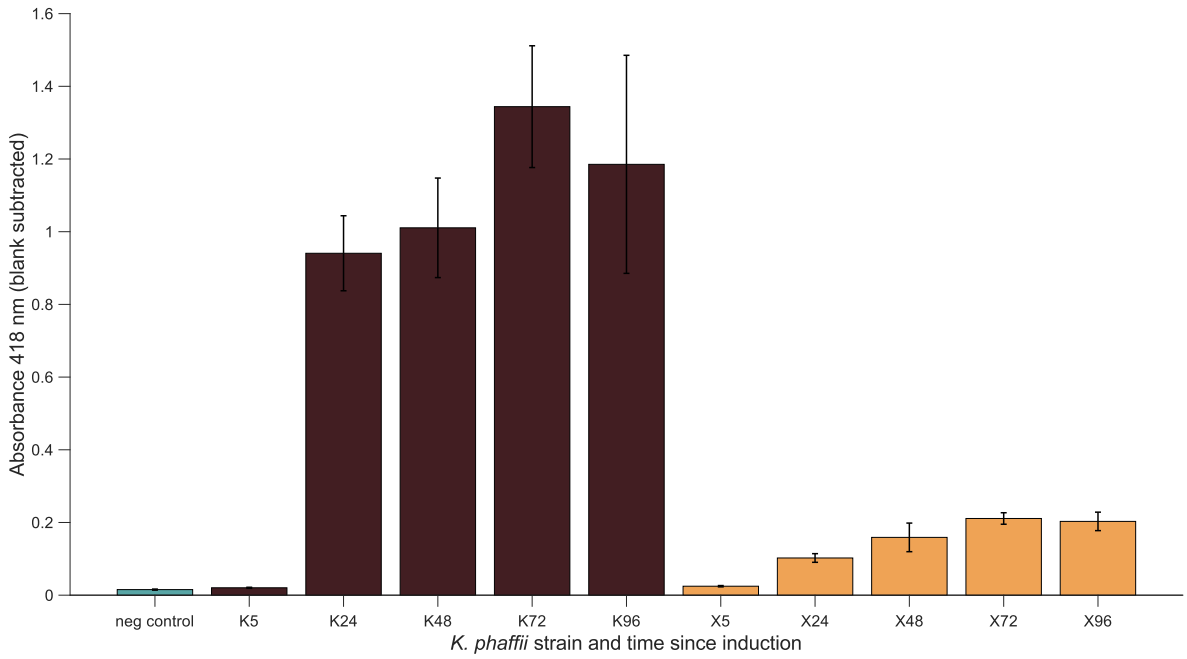
**Suppl. figure 4.1:** SDS-PAGE gels of size exclusion chromatography fractions from A) *S. cerevisiae* UPO purification and B) *K. phaffii* UPO purification. Protein ladder is on the leftmost lane of both gels.



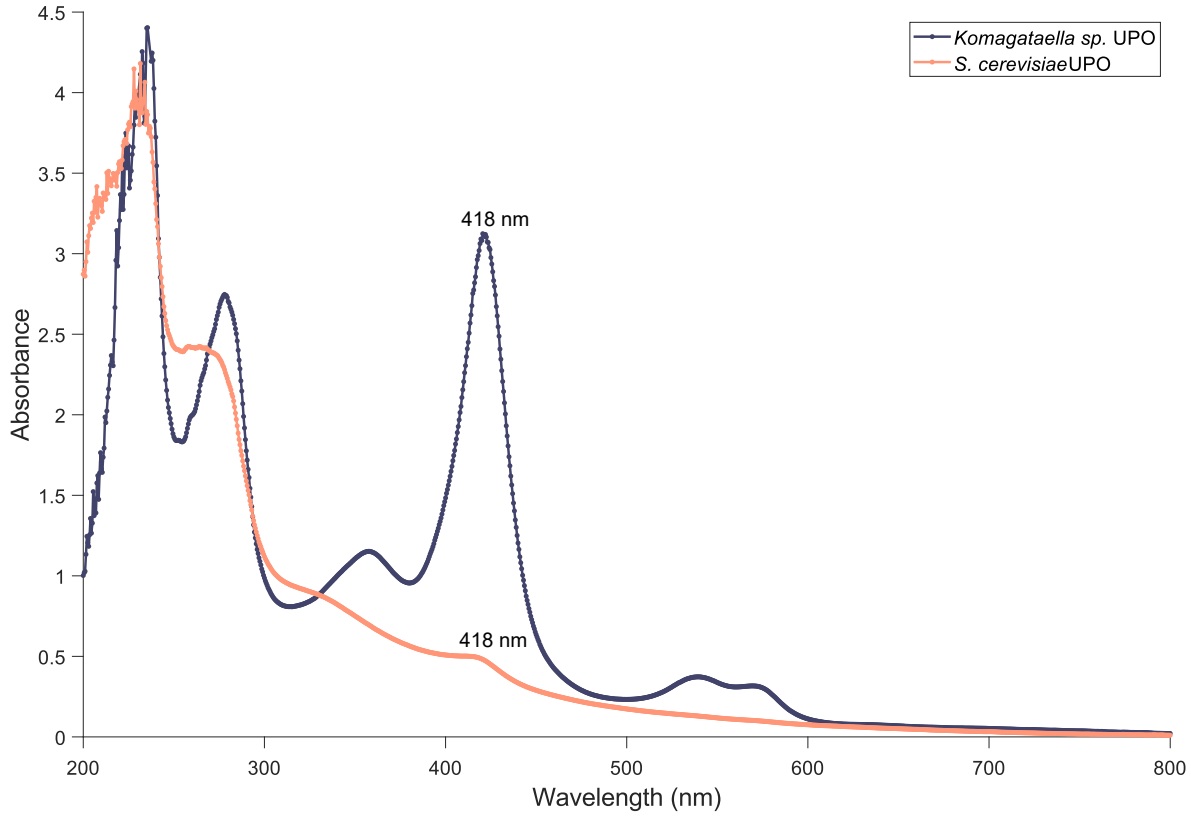
**Suppl. figure 4.2:** MALDI-TOF mass spectrometry of *K. phaffii* UPO. The base peak indicates that the molecular weight of UPO is approximately 44.2 kDa. The relative broadness of the peak is likely due to the variable glycosylation of UPO.



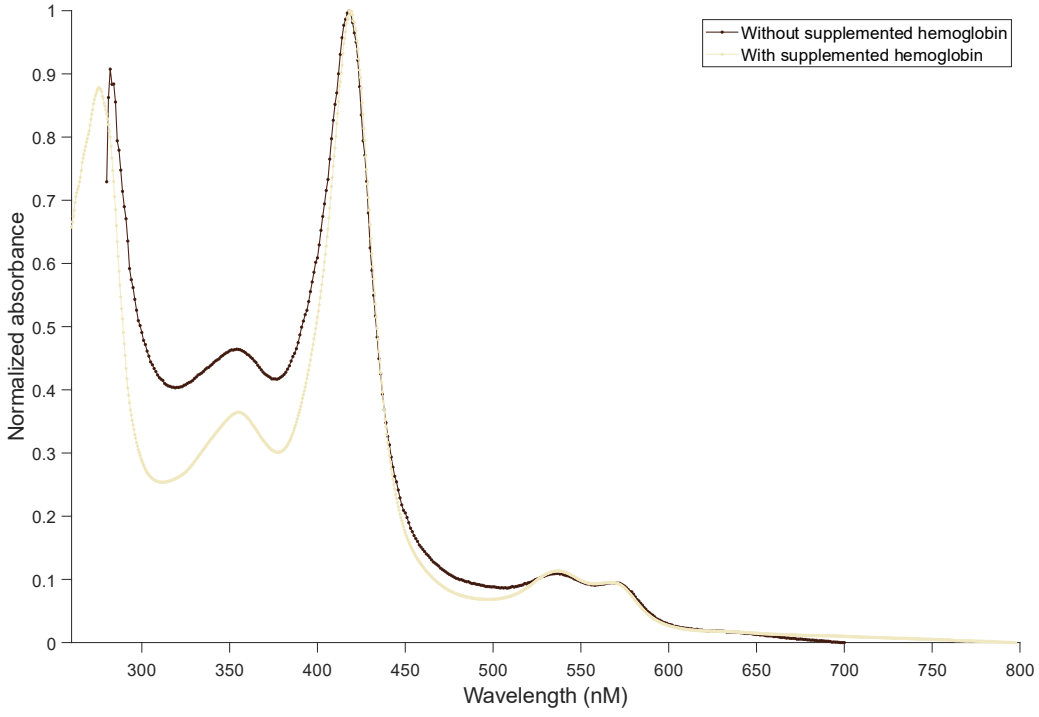

**Suppl. figure 4.3:** Comparison of UPO production by best KM71H (K) transformant versus best X33 (X) transformant as measured by ABTS assay. Samples of lysate were tested at different timepoints since induction (5 hrs, 24 hrs, 48 hrs, 72 hrs, 96 hrs). Error bars represent S.E.M. of 3 replicates. The negative control was a sample of the lysate of EGFP-expressing *K. phaffii* KM71H.



**Suppl. figure 4.4:** Absorbance spectra of UPO PaDa-I purified in *K. phaffii* and *S. cerevisiae*. Both spectra are blank subtracted. Soret peak was measured at 418 nm for both.



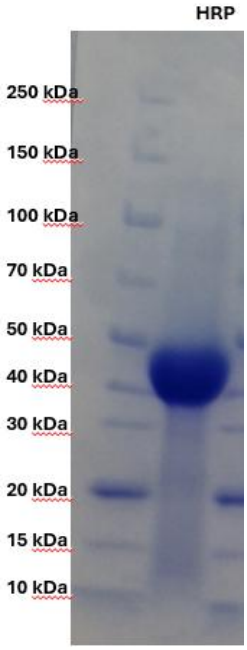

**Suppl. figure 4.5:** Normalized absorbance spectra of UPO PaDa-I purified from *K. phaffii* when the expression medium was supplemented with 0.2 mg/mL bovine hemoglobin or not. Addition of hemoglobin did not appear to improve heme incorporation, with an Rz value of 1.2 measured for both.



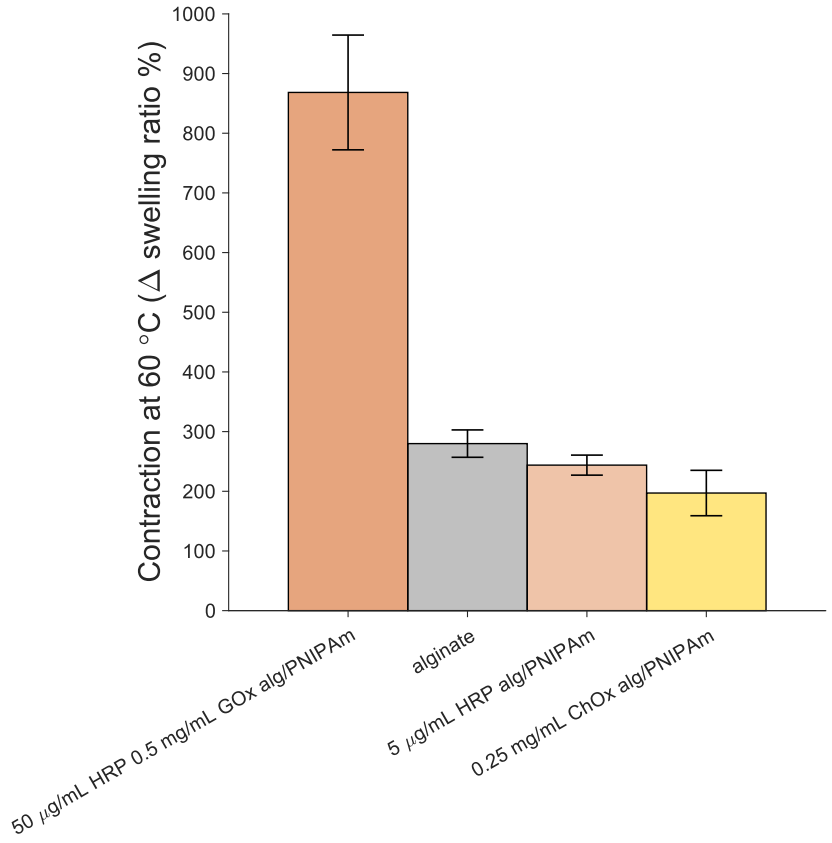
**Suppl. table 4.2:** Relative proportion of secondary structure elements of UPO PaDa-I as determined by the experimental tertiary structure (PDB:5OXU)[784]

| Secondary structure element | Proportion (%) |
|---|---|
| α-helix | 49.7 |
| β-strand | 3 |
| other | 47.3 |

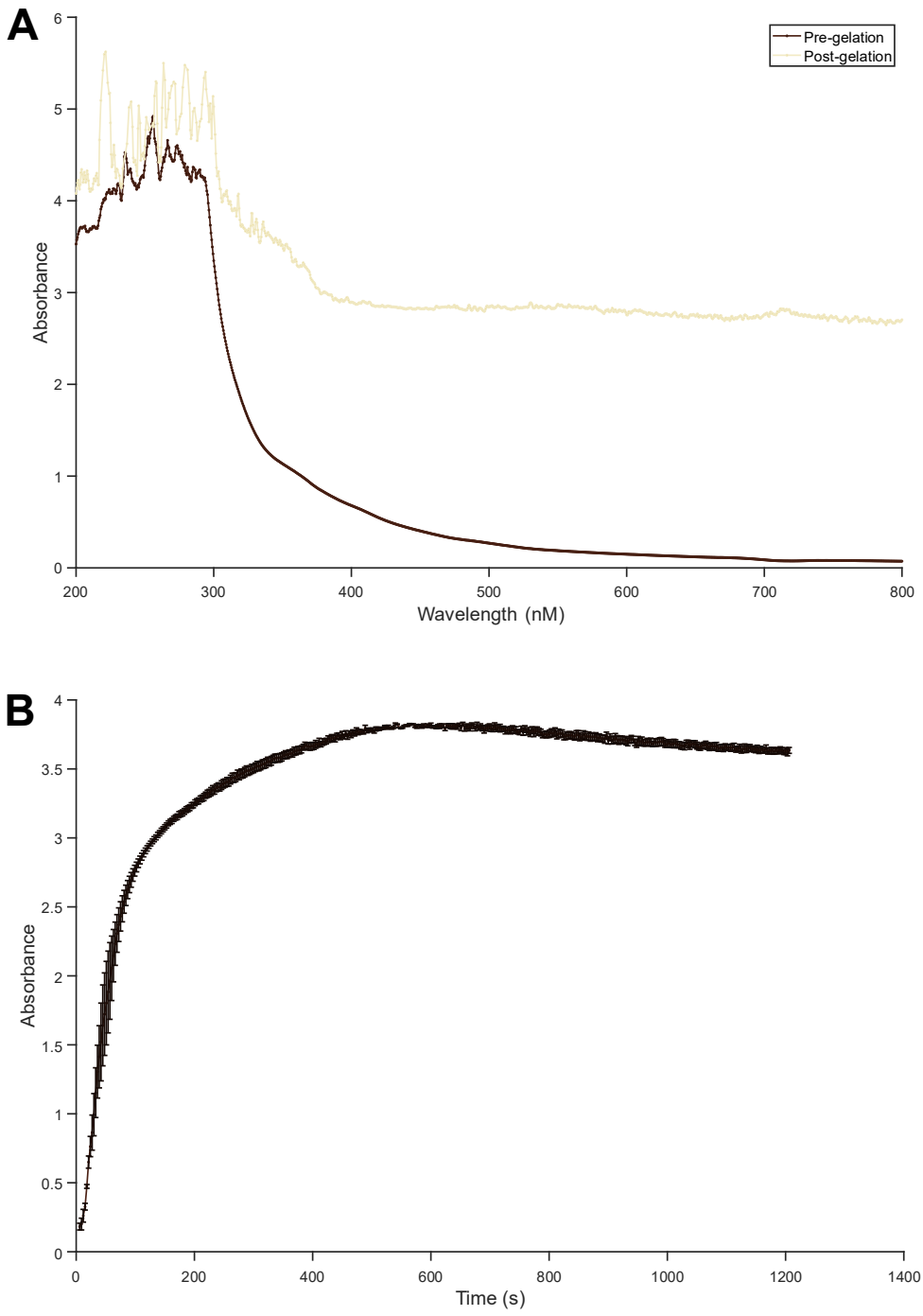
**Suppl. figure 4.6:** SDS-PAGE gel of a commercial sample of *A. rusticana* horseradish peroxidase (HRP) showing that it contains a mixture of peroxidase proteins of various molecular weights, likely due to different isoforms and variable glycosylation.



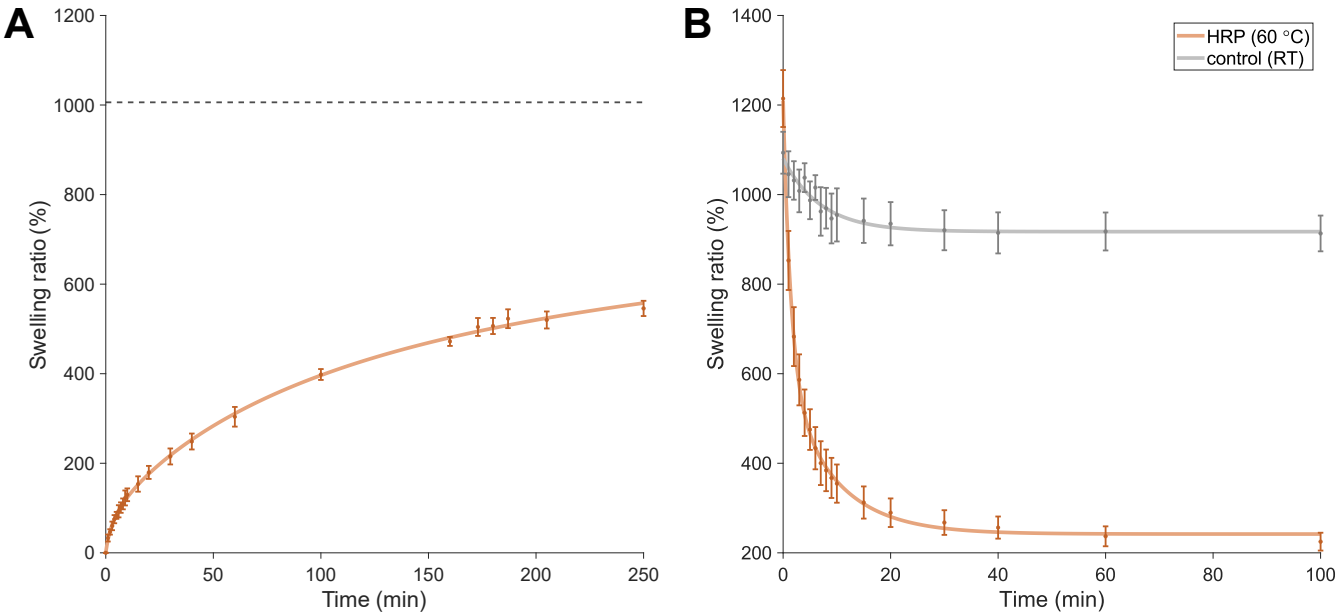
**Suppl. figure 4.7:** When insufficient concentrations of peroxidase or oxidase enzymes are used in the alg/PNIPAm formulation the resulting hydrogel has thermosensitive behaviour similar to alginate, indicating that PNIPAm likely has not formed to a significant degree. The apparent 'contraction' of alginate and the hydrogels with insufficient enzyme concentrations is likely caused by some leaching of $Ca^{2+}$ ions into the distilled water used to incubate the hydrogels during the experiment.



**Suppl. figure 4.8:** Gelation kinetics of PNIPAm. A) Absorbance/scattering spectra of PNIPAm pre-gelation and post-gelation indicated that wavelengths from ~600-800 nm may be most suitable for tracking differences in absorbance upon gelation. B) Change in absorbance/scattering at 700 nm over time during the gelation of PNIPAm showed that this method could be used to track the kinetics of PNIPAm formation. Error bars represent SEM of 3 replicates.

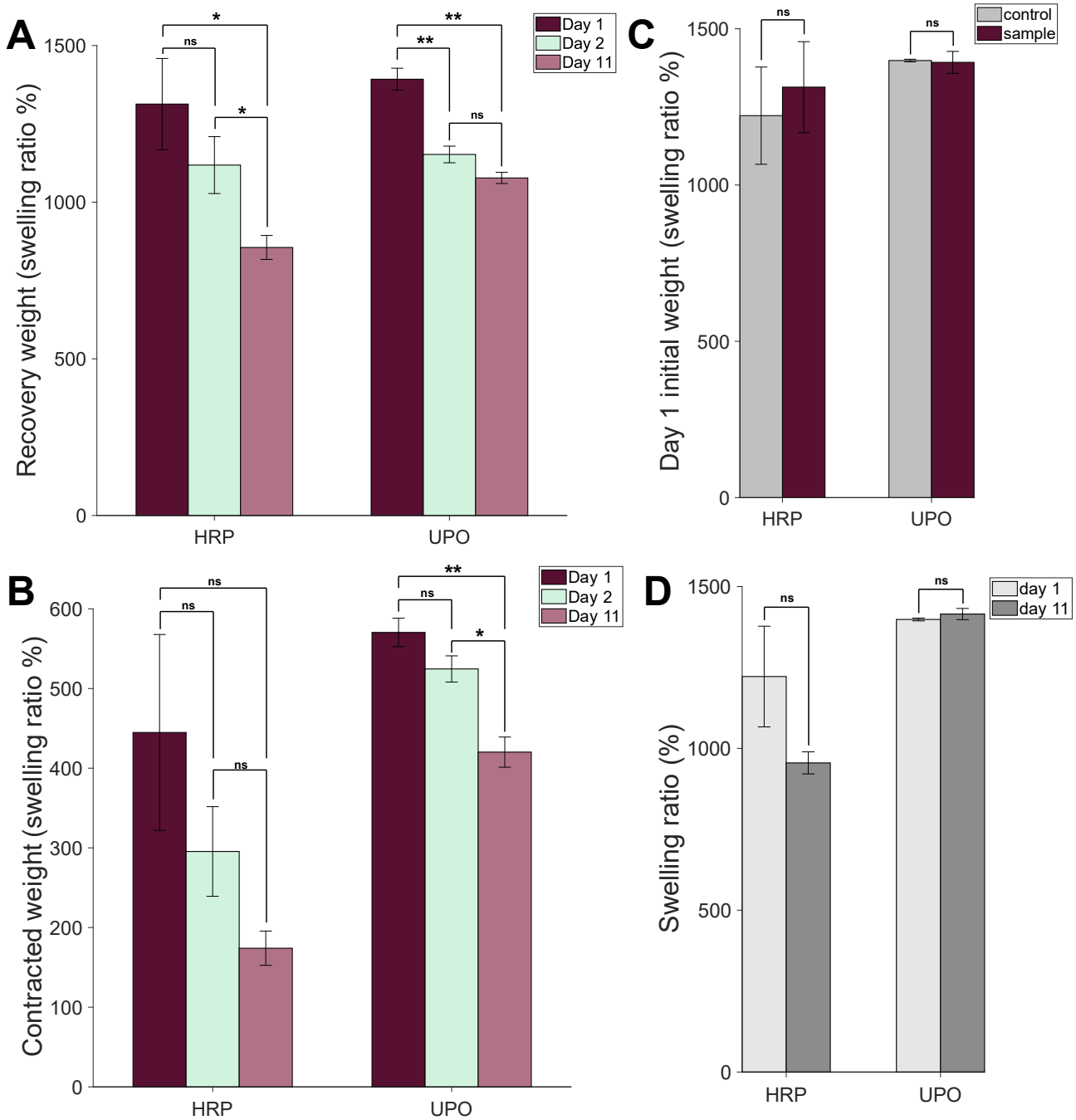


**Suppl. figure 4.9***:* A) Reswelling (hydration) kinetics of dried HRP-formed alg/PNIPAm in di$H_2O$ at room temperature. Reswelling is relatively slow, and the dried hydrogel does not reach its initial wet weight ratio (dashed grey line), perhaps due to lack of calcium chloride to crosslink the alginate component and/or leaching of alginate chains into the solution. Error bars represent S.E.M. of 3 replicates. B) Deswelling kinetics of HRP-formed alg/PNIPAm in di$H_2O$ at room temperature (RT) and 60 °C. While the hydrogel does deswell slightly at RT, perhaps due to diffusion of calcium chloride linking the alginate into the water, it deswells at a much slower rate than the deswelling at 60 °C, which is likely mostly driven by the thermosensitivity of the PNIPAm component.

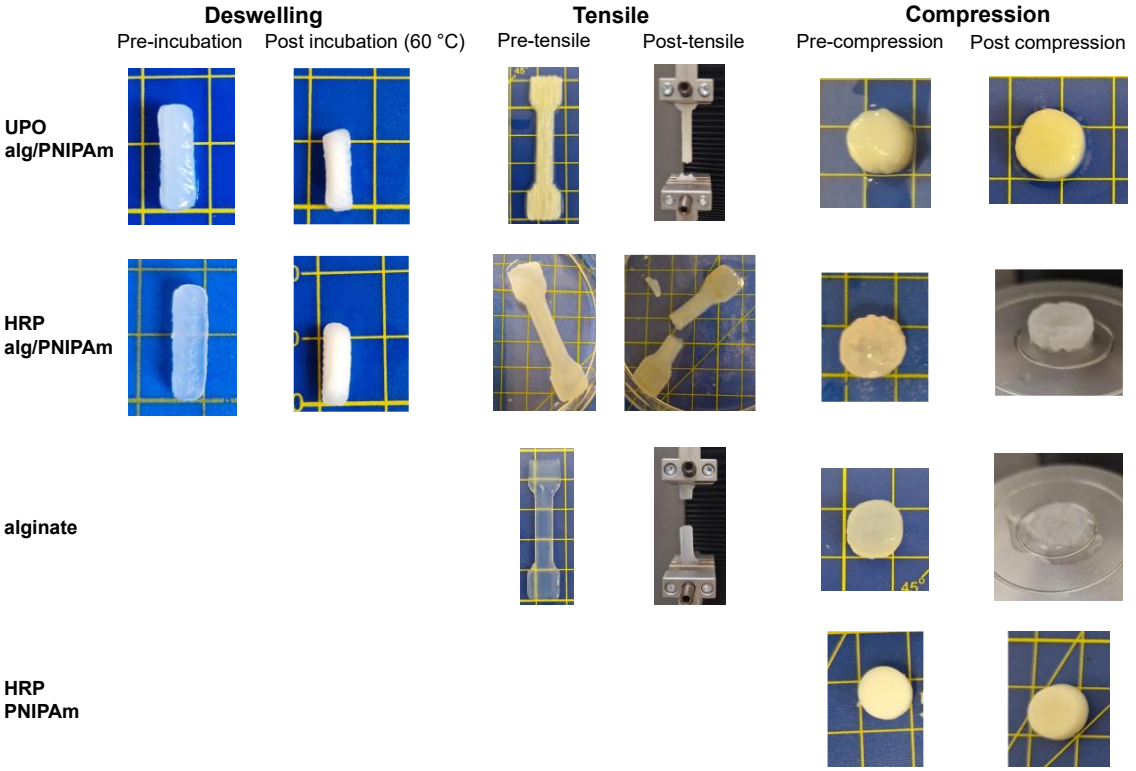



**Suppl. figure 4.10:** Repeated 'contraction' and recovery cycles were conducted on both the UPO and HRP alg/PNIPAm hydrogels respectively once a day for eleven days. Samples were incubated at 60 °C during 'contraction' and recovered at room temperature (RT) and controls were consistently incubated at RT. All error bars represent S.E.M. of 3 replicates and significance was determined by unpaired t test. A) Difference in recovery swelling weight ratios between early and final repetitions. That of HRP alg/PNIPAm gradually decreased with each repetition, whereas for UPO alg/PNIPAm the biggest change was between the first and second cycle. B) Difference in 'contracted' swelling weight ratios between early and final repetitions. That of HRP alg/PNIPAm appeared to decrease with each repetition, though this

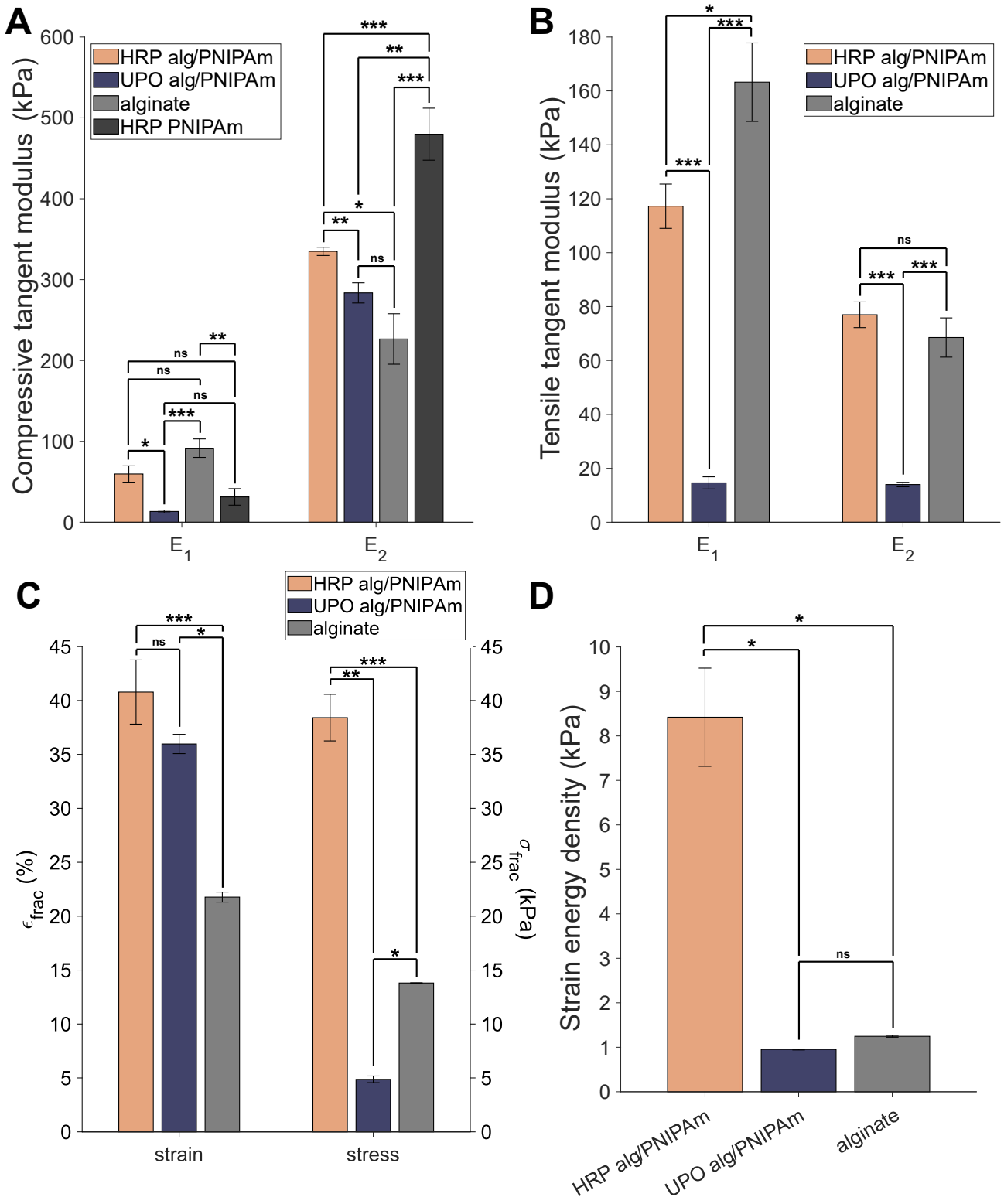
decrease was not significant. For UPO alg/PNIPAm, the 'contracted' weight ratio had significantly decreased by the 11th repetition. C) The initial swelling weight ratios for both HRP and UPO alg/PNIPAm samples and controls respectively were confirmed to be part of the same population, thereby establishing that differences over the time course are likely due to the differing incubation temperature between the samples and the controls. D) The swelling weight ratio of the controls of HRP alg/PNIPAm and UPO alg/PNIPAm respectively on the initial day and final day of measurement. The swelling ratio of HRP alg/PNIPAm controls appeared to have decreased slightly over time, however, this difference was not significant. The swelling ratio of UPO alg/PNIPAm controls did not significantly change from the first to the final day.



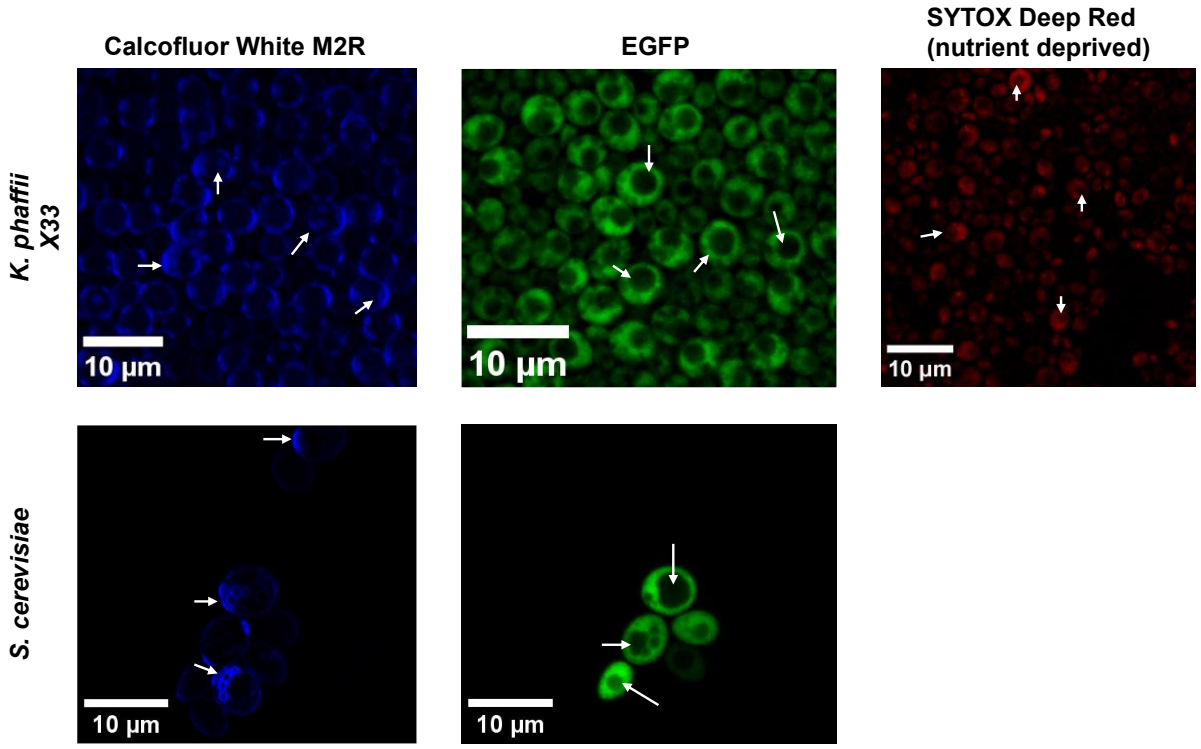
**Suppl. figure 4.13:** Representative photos of UPO alg/PNIPAm, HRP alg/PNIPAm, alginate and HRP PNIPAm hydrogels before and after high temperature incubation, tensile testing and compression testing. Following high temperature incubation, both the UPO and HRP alg/PNIPAm hydrogels became opaquer and were observed to shrink in size. During tensile testing all of the formulations tested broke. Under the compression testing conditions used, it was observed that the UPO alg/PNIPAm, HRP alg/PNIPAm and HRP PNIPAm formulations showed fairly elastic recovery, however, that the alginate hydrogel became relatively 'flat'.



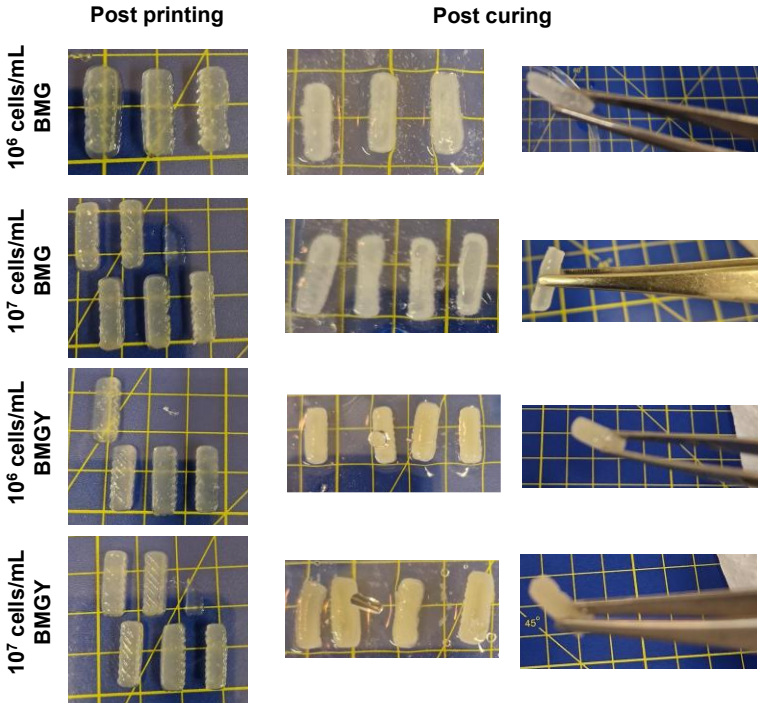
**Suppl. figure 4.12:** Comparison of key values obtained from compression (A) and tensile (B-D) testing of HRP alg/PNIPAm and UPO alg/PNIPAm as well as the single network equivalents. Error bars represent S.E.M. of at least 3 replicates and statistical analysis was conducted using one-way ANOVA with Tukey-Kramer.



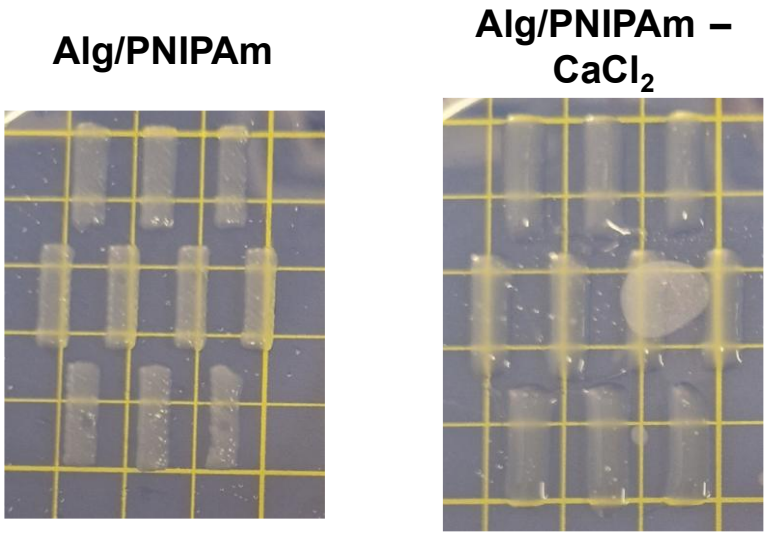
**Suppl. figure 4.13**: Bud scars (Calcofluor channel) and organelles (EGFP and SYTOX channels) were observed in the confocal fluorescent microscopy of *K. phaffii* and *S. cerevisiae* cells of liquid culture. Examples of these features are indicated with white arrows, and the scale bar represents 10 µm.



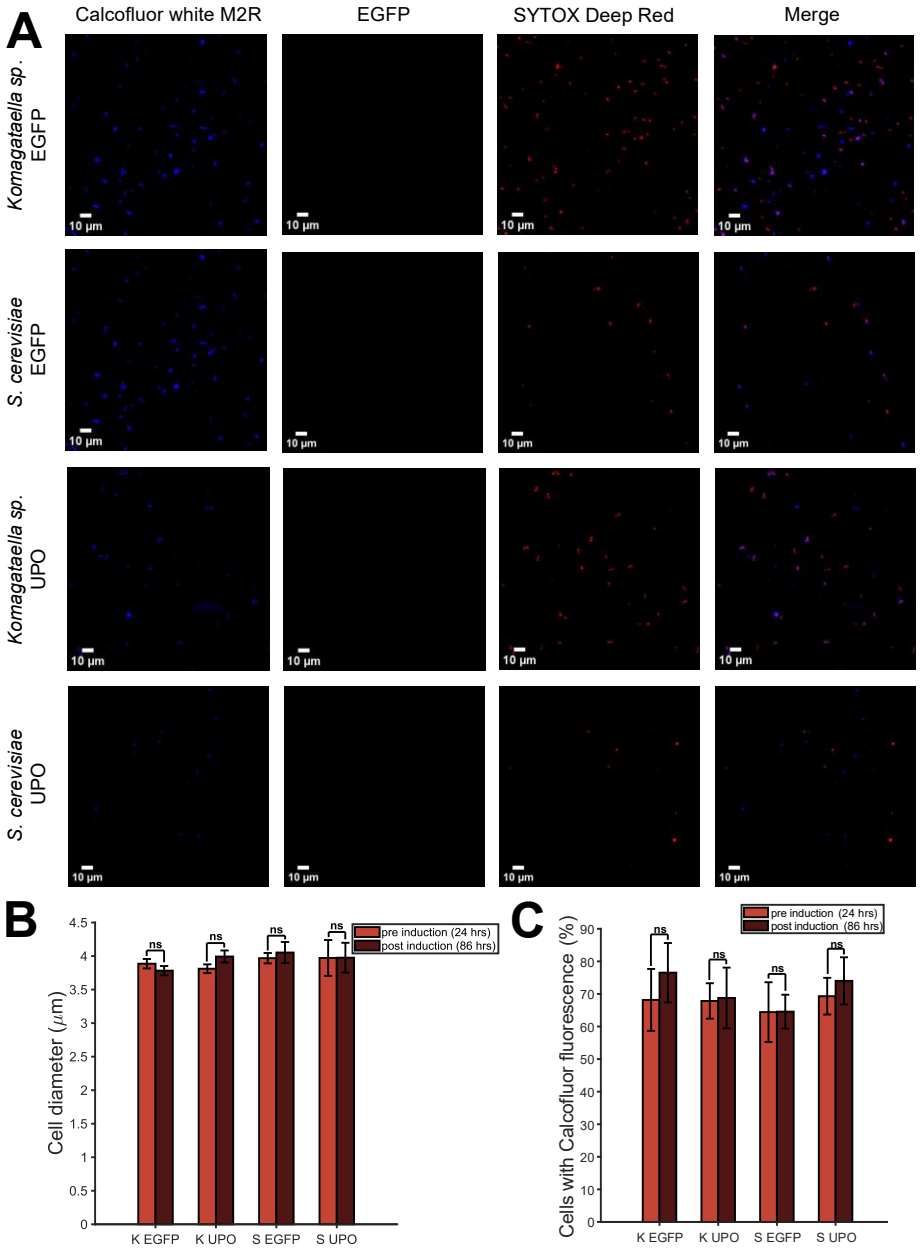

**Suppl. figure 4.14***:* Box shaped 3D-printed alg/PNIPAm with different concentrations of *K. phaffii* X33 cells photographed after 24 hrs of incubation in either BMG or BMGY medium. All prints were observed to maintain their box-like shape following incubation and could be handled with tweezers without breaking.



**Suppl. figure 4.15***:* Photos of *K. phaffii* alg/PNIPAm hydrogels after printing both with $CaCl_2$ in the ink and without $CaCl_2$ in the ink (- $CaCl_2$). It was observed that the hydrogels printed from the ink not containing $CaCl_2$ were less viscous and retained their printed shape less well. Note that both formulations were cured in a solution containing $CaCl_2$.



**Suppl. figure 4.16:** Characterising if recombinant protein production could be induced in *K. phaffii* and *S. cerevisiae* within an alg/PNIPAm hydrogel. The inducible expression of EGFP was investigated using *K. phaffii* X33 (K EGFP) and *S. cerevisiae* (S EGFP). The inducible expression of UPO was investigated using *K. phaffii* KM71H (K UPO) and *S. cerevisiae* (S UPO). A) Representative images from fluorescent microscopy taken before protein production had been induced (24 hrs after hydrogel curing). Images were taken of the cells within the hydrogels without sectioning and after staining with Calcofluor White M2R (Calcofluor) and SYTOX Deep Red. Scale bars represent 10 µm. B) Average cell diameter of each expression system before and after induction (86 hrs after curing). No significant difference was found in diameter for any of the expression systems or timepoints. Error bars represent S.E.M. of three samples and statistical significance was calculated by unpaired t test to compare time points or ANOVA to compare expression systems. C) The percentage of cells which displayed Calcofluor fluorescence for the different expression systems before and after induction... No significant difference was found in diameter for any of the expression systems or timepoints. Error bars represent S.E.M. of three samples and statistical significance was calculated by unpaired t test to compare time points or ANOVA to compare expression systems.



# Chapter 5: General Discussion

The overarching goal of this thesis was to harness evolution-guided methods as a methodological and conceptual framework for understanding protein evolution and engineering proteins with new or improved functions. While Chapter 1 outlined how evolutionary approaches can be applied in protein design for applications like material polymerisation, the subsequent chapters each tackled a different facet of this idea. Specifically, Chapter 2 focused on ancestrally reconstructed transcription factors of the *lac* repressor (LacI) family to explore how ligand specificity, binding affinity, and thermodynamics can evolve. Chapter 3 attempted to expand beyond the more commonly studied bacterial systems to reconstruct ancestors of the plant enzyme, horseradish peroxidase (HRP)[34]. HRP is widely used in applications like enzyme-mediated radical polymerization (EMRP), however, it is difficult to express recombinantly as a single isoform[603,655]. Thus, ancestral sequence reconstruction (ASR) of the class III peroxidase family was conducted with the hope of identifying an ancestral peroxidase with improved recombinant expression. Finally, Chapter 4 explores the use of a synthetically evolved enzyme, unspecific peroxygenase PaDa-I (UPO), for use in the EMRP of a hydrogel, alginate/PNIPAm, which is temperature-responsive and can be 3D printed. Additionally, Chapter 4 investigates the ability of this material to support yeast cells to form an engineered living material that can produce proteins 'on demand'.

In terms of the specific protein of interest, the shift from studying a bacterial transcription factor (LacI) in Chapter 2 to engineering a plant peroxidase (HRP) in Chapter 3 and then to utilising a synthetically evolved fungal peroxygenase (UPO) in Chapter 4 is considerable. However, all studies revolve around ancestral sequence reconstruction or evolution of long-diverged protein families, reflecting a unifying theme: using evolution engineering approaches to introduce beneficial properties or insight. In all cases, protein evolution provided routes to variants that exhibit altered or improved properties over native, extant proteins.

Below, I describe how these findings relate in a holistic sense, highlight the conceptual contributions to protein engineering, evolution and biomaterials, and propose future directions.

## 5.1 ASR as a dual tool: fundamental understanding and applied utility

### *5.1.1 Evolution of ligand binding in proteins can be gradual or rapid*

Tracking evolutionary changes increases general knowledge of how proteins can change, which can in turn inform future protein engineering efforts. Many protein families can bind to a relatively wide diversity of small molecule ligands or substrates and yet it can be difficult to discern how such diversity arose by comparing only the extant family members[15,23]. One such protein family is the *lac* repressor (LacI) family of bacterial transcription factors. Despite many studies of the modern-day proteins, including domain swapping[418,419,449], sequence comparison[590], and deep mutational scanning studies[472,478], it still remained unclear how there was such a diversity of small molecule effectors and how binding of new effectors could be designed. However, as for many other protein families[14,15,222], in Chapter 2 it was shown that ancestral sequence reconstruction could provide a new perspective on how functional changes could occur in the LacI family.

Functional changes in proteins can be classed as occurring gradually over many mutations or relatively suddenly *via* a small number of mutations[199,206,220]. This has been referred to as 'creeping' or 'leaping' in a study by Tyzack *et al.*[199]. For example, Harris *et al.*[205] demonstrated that the metabolism of several xenobiotics and steroids by members of the cytochrome P450 family 1 had gradually increased from activity present in the oldest ancestor studied along the mammalian lineage, likely reflecting increasing needs to metabolise these relatively novel xenobiotics and steroids. As an example of rapid functional change, Tran *et al.*[207] found that the cytotoxic activity of a human RNase also appeared to arise relatively rapidly from almost no activity present in an ancestor. Interestingly, in Chapter 2 it was shown that both types of functional change could take place, with specialisation towards a 'new' binding function occurring relatively gradually with many mutations, whereas the weakening of an 'old' binding function occurred relatively rapidly and was driven by comparatively fewer mutations. This highlights that the overall 'function' of a protein describes multiple different processes, such as binding to multiple ligands, and that evolutionary change can affect these different processes in different ways. This is highlighted in a study of ancestral carboxylic acid transporters by Savory *et al.*[897], where it was shown that each ancestral transporter had a preferred range of substrates that changed differently along different trajectories of the phylogeny. Hence, while it can be of interest to examine only how a single function or property of a protein can change, several studies have now demonstrated that greater insights can be gleaned from examining multiple functions or

properties can change at the same time, and potential relationships between these changes[205,207,226,897]. Nonetheless, the overall 'fitness' of a protein can be difficult to completely characterise given the vast possibilities of behaviours and properties that any given protein could display. To fully characterise any protein and its functional and property changes in the future, perhaps high-throughput experimental techniques could be combined with computational-guided techniques.

### *5.1.2 The protein sequence-function landscape*

Whilst it is informative to characterise phenotypical level changes in protein function, perhaps one of the biggest goals in protein science is to determine how the sequence of any given protein can lead to its function, or *vice versa*. The DNA sequence encoding a protein is the medium by which the variation in protein function is generated, and the sequence should encode all of the functional information of a protein[422]. With the rise of computational power that has led to the explosion of machine learning and other similar statistical techniques, large advances have been made towards predicting the structure of a protein. From this structure, molecular dynamics, machine learning-guided or other statistical-guided techniques can then be utilised to attempt to predict function. However, all these techniques remain predictive rather than completely exhaustive and accurate; even machine-learning guided techniques still need to screen thousands of candidates using molecular dynamics and then tens experimentally[898-900]. Hence, the relationship between protein sequence and function has not yet been 'solved', and it is thus still informative to tease apart this relationship through studying protein evolution. Unlike other techniques, ancestral sequence reconstruction is relatively low throughput, as typically ancestral proteins fold, are stable, and are functional. Furthermore, by including the dimension of time, ancestral sequence reconstruction allows for functional changes to be tracked over relatively few mutations. Notwithstanding, all protein engineering techniques can be used synergistically to both improve our fundamental knowledge and engineer new proteins. For example, the structure and function of ancestral proteins can be predicted using machine-learning guided techniques to reduce experimental labour and cost[12,901].

Changes in protein function derive from changes to the sequence at the genetic level, such as by random mutagenesis, insertions or deletions (indels) and/or recombination[242-244]. However, due to the evolutionary models commonly used during the ancestral sequence reconstruction process, ancestral protein sequences and their corresponding functional changes are primarily generated through mutations. By examining how a relatively limited number of mutations

lead to various changes in function over times, it has been found that often functional changes can be driven by mutations relatively distal to the ligand or substrate binding site[902]. For instance, Kaltenbach *et al*.[192] found that chalcone isomerase activity arose from a non-catalytic ancestor partly due to distal mutations that changed the dynamics of the protein. The relatively large effect of distal mutations on ligand binding was also observed by examining the effects of individual mutations between the oldest two selected ancestors of the LacI family in Chapter 2, where it was found that mutations relatively distal from the active site had large effects on ligand binding function. In future, it would be interesting to further explore the mechanistic causes behind these distal effects. Perhaps as well as changing the dynamics of the LacI ancestor, the distal sites may also have been part of networks of interacting residues, as has been found in modern-day LacI[464].

Possibly also resulting from interacting residue networks and/or changes to the overall conformational ensembles of a protein, it has increasingly been found by many studies, including that in Chapter 2, that mutations can have non-additive effects on function[213,475,531]. This has shifted the field away from the more classical model of evolutionary function changes, which assumed that most mutations that were fixed during an evolutionary trajectory conferred a neutral or positive benefit to the new function in a largely additive fashion, analogous to climbing up a hill[206]. Instead, studies such as that by Yang *et al*.[576] and Anderson *et al.*[903] have shown that epistatic interactions are highly prevalent in proteins. Furthermore, Starr *et al*.[183] and others[249,250,903] have found that historical contingencies like 'permissive' and 'restrictive' mutations determine which evolutionary trajectories are available, such that a protein will not necessarily always be able to reach every part of the sequence-function landscape. Hence, the complex nature of the functional changes in the ligand binding along the LacI trajectory further highlights the complex, multidimensional nature of the sequence-function landscape.

### *5.1.3 Importance of thermodynamics in protein function*

Any given chemical reaction or binding interaction, conducted by a protein or otherwise, is governed by thermodynamics[582,904]. To conduct useful reactions and interactions for life, proteins must utilise thermodynamics to their advantage. For example, binding of a given substrate to a protein can be driven by forming chemical bonds with that substrate, by increasing flexibility of a distal part of the protein to compensate for the increase in entropy of the binding site and substrate upon binding, and by increasing the entropy of 'trapped' water molecules in the binding site upon binding[422]. As the given function of a protein is

dependent upon thermodynamics, understanding of how proteins can have different functions and thus how to engineer proteins can be aided by gaining insights into how different functions can be driven by different thermodynamic mechanisms. One way to study functional changes over time, and thus their corresponding reaction thermodynamics, is by using ancestral sequence reconstruction[186]. For example, Wheeler *et al*.[735] characterised the thermodynamics of the binding reactions of peptides to two ancestral S100 proteins and found that the binding reactions for one ancestor were exothermic but for the other were endothermic. In Chapter 2 it was also found that functional ligand binding changes were accompanied by changes to the thermodynamics of the ligand binding reaction. It was shown that there was a general enthalpic-entropic trade-off for the characterised binding reactions, thereby providing further support that this could be a universal trend[521,522,534,586,587]. It has been posited that this trend could result from proteins conducting their reactions in largely aqueous environments[519,523]. Hence, this highlights the importance of considering thermodynamics in protein reactions.

Interestingly, it was found that the changes in the thermodynamics of the binding reactions along the LacI trajectory could not be fully explained through structural mechanisms[500], hence, may have largely resulted from changes to the protein dynamics. Many other studies have also shown that functional changes can be largely driven by dynamics for some proteins[192,194,216-219]. For example, Kim *et al*.[215] found that fluorescence changes of GFP-like proteins largely resulted from changes to the protein dynamics along different evolutionary trajectories. Thus, it would be informative to further study the dynamics of the LacI trajectory ancestors to further characterise the relationship between these changes and the thermodynamic changes found, such as by using molecular dynamics and hydrogen deuterium exchange mass spectrometry.

### *5.1.4 ASR as a tool to improve recombinant expression*

One factor that can limit the utility of many proteins is their relatively low expression levels in hosts that are relatively easy to use and cheap, like *E. coli* and *S. cerevisiae*. Many eukaryotic proteins from multicellular eukaryotes suffer this fate, often because they possess posttranslational modifications not available in microbial expression systems, are processed in different ways to microbial systems and/or require cofactors and cellular environments not present in microbial systems[398,792,793]. However, because protein expression is a complex process that is not entirely understood, it can be difficult to increase the recombinant expression levels of a given protein in a given host in a rational manner. On the other hand,

many studies have found that ancestral proteins predicted through ancestral sequence reconstruction have had improved recombinant expression compared to their extant family members[26,192,236,237,240]. It is hypothesised that this improved recombinant expression may result from the increased thermostability of many ancestral proteins, because stability often correlates with recombinant expression levels for microbial hosts[668-670]. Such a correlation was also found in Chapter 2 with the single mutants of the oldest ancestor. Furthermore, the utility of using ancestral sequence reconstruction to improve recombinant expression was shown in Chapter 3, where ancestral plant peroxidases were able to be expressed solubly in *E. coli*. However, this also highlighted some of the challenges of overexpressing eukaryotic proteins, such as cofactor incorporation and lack of signal peptide processing. Nonetheless, in future these issues could be resolved, such as through cofactor incubation and removing the signal peptide, to potentially develop ancestral proteins with peroxidase catalysis similar to horseradish peroxidase but with much greater recombinant expression yields.

Overall, it is hoped that ability of ASR to improve recombinant expression, often whilst maintaining protein activity, will be more commonly used in future for a variety of proteins and their corresponding applications. While other protein engineering techniques can also improve recombinant expression levels, including directed evolution[905-907], consensus design[674] and other computational-guided techniques[908,909], these typically require screening of many more variants than ASR, where sometimes only a single ancestral protein is screened[230,232].

## 5.2 Evolution-guided protein engineering in materials science: an underexplored aspect of biotechnology

### *5.2.1 Potential of protein engineering in enzyme-mediated radical polymerisation*

Many useful materials for diverse applications can be synthesised through free radical polymerisation, including hydrogels and value-added compounds[5]. However, many of the radical initiator catalysts are relatively toxic and non-biodegradable, such as azobisisobutyronitrile[910] and benzoyl peroxide[911]. Furthermore, typically free radical polymerisation needs to occur in an oxygen-free environment, which can be laborious to set up[300,372,912]. Both of these issues can be resolved by instead using enzymes as both catalysts and for oxygen scavenging. Indeed, over the past decade there has been increasing interest in using enzyme-mediated radical polymerisation (EMRP) for a plethora of polymers of various

complexities[300,313,316,371]. Many reviews of the fields state that the primary reason is likely the 'green' nature of enzyme chemistry compared to more traditional methods[5,294,300,316,372]. However, despite the large number of reactions that have been polymerised using EMRP, the number of enzymes investigated has been relatively limited[5,313]. Typically, either a laccase is used, most commonly *Trametes versicolor* laccase[913-915], or often horseradish peroxidase is used with a glucose oxidase or pyranose oxidase[339,916,917]. While these enzymes have been able to mediate a wide range of polymerisations[5,298], this range could be further increased through the use of different enzymes. For example, unspecific peroxygenases are capable of catalysing the production of free radicals, but have potentially different substrate ranges to laccases and class III peroxidases[655,782,918]. As another example, Li *et al.*[342] demonstrated that formate oxidase in EMRP could be advantageous over glucose or pyranose oxidase because the product of formate oxidase easily separates from the reaction mixture. Nonetheless, many commonly used enzymes, including the laccases and horseradish peroxidase, are commonly sourced natively from their host organism as multiple isoforms, which causes batch-to-batch variation[604,655,919]. Such variation can lead to irreproducibility, especially for more complex materials like double networks[830]. Furthermore, the conditions of some radical polymerisations are not well suited to native enzymes, such as reactions that use hydrophobic monomers, organic solvents and/or occur at relatively hot or cold temperatures[298,300]. Finally, enzymes are commonly not as re-usable as traditional chemical catalysts[300,316].

Many, if not all, of these challenges facing enzyme use in radical polymerisations could be met by using protein engineering techniques, such as evolutionary-guided techniques. For instance, directed evolution can 'engineer' proteins with greater solvent stability[920-922], different substrates[923-925], greater temperature stability[926-928] and improved recombinant expression[287,291,905]. Furthermore, ancestral sequence reconstruction has been shown to predict ancestral enzymes that maintain similar activity whilst being more thermostable[234-236], solvent stable[236,929,930] and/or having increased recombinant expression yields[237,240]. The utility of ASR for EMRP was demonstrated in Chapter 3 through the use of an ancestral class III peroxidase, which had been recombinantly expressed in *E. coli*, to catalyse the polymerisation of Poly(*N*-isopropylacrylamide) (PNIPAm) under open air when combined with glucose oxidase. To the best of our knowledge, this is one of few studies that a recombinantly expressed peroxidase or laccase has been utilised for EMRP[931], despite the batch-to-batch variation of these enzymes when sourced natively. Interestingly, the recombinant ascorbate peroxidase used by Frey *et al.*[931] had been engineered through directed

evolution by Ting and coworkers[932,933] for both increased recombinant expression and increased activity, thereby further highlighting the power of evolutionary techniques.

This ability of directed evolution to engineer proteins with improved recombinant expression was also found to be useful for recombinantly producing *Agrocybe aegerita* unspecific peroxygenase (UPO) in yeast[290]. In Chapter 4 it was shown that the EMRP of PNIPAm within an alginate/PNIPAm double network could be driven by recombinant UPO, thereby providing another demonstration that it is possible to 'replace' commonly used in enzymes like HRP with single isoforms optimised for recombinant expression through evolutionary-guided engineering techniques. Furthermore, the use of UPO in EMRP for perhaps the first time demonstrated many enzymes beyond HRP and laccase could be utilised in EMRP in the future. Indeed, an engineered laccase for EMRP has already been developed by Gomez-Fernando *et al*.[934], who reconstructed ancestral laccases and further used directed evolution to improve the activity of an ancestral laccase for generating β-diketone radicals. Despite the potential of this engineered ancestral laccase, surprisingly no further study has yet used it for EMRP to the best of our knowledge. Indeed, generally it seems that the field of EMRP has largely failed to capitalise on the advantages of using engineered enzymes.

Given the ability of evolutionary techniques to potentially engineer greater thermostability[935], organic solvent tolerance[920,929] and new substrates, which could be radical mediators or monomers[934], it is hoped that more research will be invested in exploring these possibilities in future to further expand the potential of enzyme-mediated radical polymerisation.

### *5.2.2 Some possible future directions for engineered living materials*

Interesting materials with a plethora of applications can be formed by combining the advantages of living cells, which have been engineered to purpose, and materials[359,380-382,385,386]. Such materials have been termed engineered living materials (ELMs), and are part of a relatively new, but rapidly growing field[380]. Many ELMs have been formed using hydrogel scaffolds because of their biocompatibility[359]. However, although the backbones and/or crosslinks of these hydrogels are often formed through radical mechanisms, commonly the hydrogels are cured using chemical initiators and UV light that are not biodegradable[394,399,759]. Hence, it could be beneficial to adopt the use of alternative curing mechanisms in future. For instance, it was shown by Klemperer *et al*.[361] for an *E. coli* ELM and in Chapter 4 for casted yeast ELMs that enzyme-mediated polymerisation can be utilised to yield alg/PNIPAm ELMs with relatively high cell viability. Another advantage of using

enzymes to mediate radical polymerisation was recently highlighted by Belluati and coworkers[737], who demonstrated 'cell'-mediated polymerisation of poly(ethylene glycol) methyl ether methacrylate (PEGMA) and NIPAm driven by *S. cerevisiae* cells expressing and displaying horseradish peroxidase. Such a system could have great potential in forming a 'self-assembling' ELM through cell controlled EMRP. While several systems have been developed that utilise natural redox processes of cells to initiate polymerisation[936], relying on natural redox processes is not controllable or inducible. By instead engineering recombinant enzyme expression, the relative level of enzyme can be controlled, *e.g.* through promoter strength, and synthetic biology can be used to induce expression in response to various inputs, such as chemicals, light and heat[937,938]. However, as Belluati *et al.*[737] did note relatively low levels of recombinant active HRP surface display, perhaps in future such a system could be improved by using proteins engineered through evolutionary techniques, like UPO variant PaDa-I[290].

Another consideration for hydrogel engineered living materials is the diversity of materials and their properties such that varied applications could be met. For instance, 3D-printable hydrogels have greater customisability, 'smart' hydrogels can respond to various stimuli and use of different polymers and double networks can form materials of varying mechanical strengths and elasticities[351,365,374,382]. Utilising the material developed by Klemperer *et al.*[361], in Chapter 4 it was shown that UPO-mediated alg/PNIPAm could form a yeast ELM that was 3D printable and potentially thermosensitive. However, future work needs to fully investigate the effects of the yeast on the mechanical properties of the alg/PNIPAm hydrogel. Nonetheless, as the diversity of materials used in ELM grows, the complexity of the materials can increase. For instance, recently Manjula-Basavanna *et al.*[383] developed a compostable engineered living material with tuneable mechanical strength using biofilm-based 'AquaPlastic'[939].

As well as using different materials, it is also beneficial to develop ELMs with different cells for different applications[940-944]. Yeast ELMs can be advantageous for being relatively easy to engineer compared to other eukaryotic cells, being able to produce eukaryotic proteins, capable of anaerobic fermentation, and with many species being generally regarded as safe[396-398]. While several yeast ELMs have been developed, the diversity of yeast species has been limited, with mostly *S. cerevisiae* being utilised[393,402,403]. However, in Chapter 4 it was shown that it was possible to form both *S. cerevisiae* and *K. phaffii* ELMs in 3D-printed alg/PNIPAm. However, it was found that the viability and correspondingly recombinant

protein production by both yeast species was relatively low, possibly driven by a combination of 3D printing and curing stresses. Nonetheless, if these processes could be optimised in future, then perhaps it could be further explored if more complex systems could be conducted by the yeast within the hydrogel, such as production of value-added chemicals in a controlled manner or detection and clean-up of a biotoxin. For example, Sugianto *et al*.[403] showed inducible betaxanthin production was possible in a yeast ELM by using yeast engineered with a catalytic cascade of enzymes. Also, Tay *et al*.[388] developed an *E. coli* ELM that could sense and absorb mercury for bioremediation.

In future, more research should also be conducted into the effects of hydrogel encapsulation on the yeast cells. Interestingly, in Chapter 4 it was observed that the size of the cells is on average smaller than in liquid culture even for cells that appeared viable. Other studies[359,757,879] have also observed an average smaller yeast cell size, even for yeast with sufficient diffusion of nutrients. This is despite other studies having established that *S. cerevisiae* in hydrogels can be viable, have metabolic activity similar to liquid culture and be capable of reproducing. Hence, to the best of our knowledge, it is still unknown why yeast cells typically grow to a smaller size in hydrogels. Sugianto *et al*.[403] and Liu *et al*.[359] have posited that the material could be exerting a mechanical force on the yeast cells, which perhaps could lead to a stress response and correspondingly smaller cell size. Perhaps such a stress response could also have led to the slower glucose consumption and different oxygen metabolism observed by Butelmann *et al.*[757] for *S. cerevisiae* in F127-BUM compared to liquid culture. Nonetheless, for future yeast ELM development it would be important to better characterise the effect of hydrogel encapsulation on yeast cells, such as by establishing if a stress response is occurring and whether this is being driven by mechanical pressure or other factors. By better understanding the relationship between materials and cells, undoubtedly the complexity of ELMs and their potential will be further increased.

## 5.3 Conclusion

In this thesis it was shown that the use of evolution-guided protein engineering techniques like ancestral sequence reconstruction and directed evolution can not only increase our understanding of how protein functional changes can occur, but can also be utilised to engineer proteins with industrially relevant properties like improved recombinant expression. It was demonstrated that evolution-guided techniques can be combined with materials science

to enzymatically-mediate the formation of hydrogels, including a complex engineered living material.

"*One never notices what has been done; one can only see what remains to be done*."

– Marie Curie

# Chapter 6: References


1 Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A. & Lanfear, R. The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biol Evol* **11**, 3341-3352 (2019). https://doi.org/10.1093/gbe/evz193

2 Chang, J. M., Di Tommaso, P. & Notredame, C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol Biol Evol* **31**, 1625-1637 (2014). https://doi.org/10.1093/molbev/msu117

3 Cipolatti, E. P. *et al.* in *Advances in Enzyme Technology* (eds Ram Sarup Singh, Reeta Rani Singhania, Ashok Pandey, & Christian Larroche) 137-151 (Elsevier, 2019).

4 Selles Vidal, L., Murray, J. W. & Heap, J. T. Versatile selective evolutionary pressure using synthetic defect in universal metabolism. *Nat Commun* **12**, 6859 (2021). https://doi.org/10.1038/s41467-021-27266-9

5 Shoda, S., Uyama, H., Kadokawa, J., Kimura, S. & Kobayashi, S. Enzymes as Green Catalysts for Precision Macromolecular Synthesis. *Chem Rev* **116**, 2307-2413 (2016). https://doi.org/10.1021/acs.chemrev.5b00472

6 Kaushik, M. *et al.* Protein engineering and de novo designing of a biocatalyst. *J Mol Recognit* **29**, 499-503 (2016). https://doi.org/10.1002/jmr.2546

7 Pan, X. & Kortemme, T. Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry* **296** (2021). https://doi.org/10.1016/j.jbc.2021.100558

8 Clifton, B. E., Kozome, D. & Laurino, P. Efficient Exploration of Sequence Space by Sequence-Guided Protein Engineering and Design. *Biochemistry* **62**, 210-220 (2023). https://doi.org/10.1021/acs.biochem.1c00757

9 Trudeau, D. L. & Tawfik, D. S. Protein engineers turned evolutionists-the quest for the optimal starting point. *Curr Opin Biotechnol* **60**, 46-52 (2019). https://doi.org/10.1016/j.copbio.2018.12.002

10 Risso, V. A. *et al.* De novo active sites for resurrected Precambrian enzymes. *Nat Commun* **8**, 16113 (2017). https://doi.org/10.1038/ncomms16113

11 Yang, D. *et al.* Engineering the Reaction Pathway of a Non-heme Iron Oxygenase Using Ancestral Sequence Reconstruction. *J Am Chem Soc* **146**, 34352-34363 (2024). https://doi.org/10.1021/jacs.4c08420

12 Brennan, C. K., Livada, J., Martinez, C. A. & Lewis, R. D. Ancestral Sequence Reconstruction Meets Machine Learning: Ene Reductase Thermostabilization Yields Enzymes with Improved Reactivity Profiles. *ACS Catalysis* **14**, 17893-17900 (2024). https://doi.org/10.1021/acscatal.4c03738

13 Chiang, C. H., Wang, Y., Hussain, A., Brooks, C. L., 3rd & Narayan, A. R. H. Ancestral Sequence Reconstruction to Enable Biocatalytic Synthesis of Azaphilones. *J Am Chem Soc* **146**, 30194-30203 (2024). https://doi.org/10.1021/jacs.4c08761

14 Gumulya, Y. & Gillam, E. M. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. *Biochem J* **474**, 1-19 (2017). https://doi.org/10.1042/BCJ20160507

15 Hochberg, G. K. A. & Thornton, J. W. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys* **46**, 247-269 (2017). https://doi.org/10.1146/annurev-biophys-070816-033631
16 Selberg, A. G. A., Gaucher, E. A. & Liberles, D. A. Ancestral Sequence Reconstruction: From Chemical Paleogenetics to Maximum Likelihood Algorithms and Beyond. *J Mol Evol* **89**, 157-164 (2021). https://doi.org/10.1007/s00239-021-09993-1
17 Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet* **5**, 366-375 (2004). https://doi.org/10.1038/nrg1324
18 Pauling, L., Zuckerkandl, E., Henriksen, T. & Lövstad, R. Chemical paleogenetics. *Acta Chem Scand* **17**, 9-16 (1963).
19 Merkl, R. & Sterner, R. Ancestral protein reconstruction: techniques and applications. *Biol Chem* **397**, 1-21 (2016). https://doi.org/10.1515/hsz-2015-0158
20 Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T. & Poon, A. F. Ancestral Reconstruction. *PLoS Comput Biol* **12**, e1004763 (2016). https://doi.org/10.1371/journal.pcbi.1004763
21 Clifton, B. E. *et al.* Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nat Chem Biol* **14**, 542-547 (2018). https://doi.org/10.1038/s41589-018-0043-2
22 Thornton, J. W., Need, E. & Crews, D. Resurrecting the Ancestral Steroid Receptor: Ancient Origin of Estrogen Signaling. *Science* **301**, 1714-1717 (2003).
23 Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* **20**, 360-366 (2010). https://doi.org/10.1016/j.sbi.2010.03.005
24 Garcia, A. K. & Kacar, B. How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radic Biol Med* **140**, 260-269 (2019). https://doi.org/10.1016/j.freeradbiomed.2019.03.033
25 Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat Rev Genet* **21**, 428-444 (2020). https://doi.org/10.1038/s41576-020-0233-0
26 Gamiz-Arco, G. *et al.* Combining Ancestral Reconstruction with Folding-Landscape Simulations to Engineer Heterologous Protein Expression. *J Mol Biol* **433**, 167321 (2021). https://doi.org/10.1016/j.jmb.2021.167321
27 *Bioinformatics for DNA Sequence Analysis*. (Springer, 2009).
28 *Computational Methods in Protein Evolution*. (Humana Press, 2019).
29 *Enzyme Engineering*. (Humana Press, 2022).
30 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019). https://doi.org/10.1186/s13059-019-1832-y
31 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003). https://doi.org/10.1101/gr.1224503
32 Altenhoff, A. M. *et al.* OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. *Nucleic Acids Res* **52**, D513-D521 (2024). https://doi.org/10.1093/nar/gkad1020
33 Oberg, N., Zallot, R. & Gerlt, J. A. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J Mol Biol* **435**, 168018 (2023). https://doi.org/10.1016/j.jmb.2023.168018

34 Scossa, F. & Fernie, A. R. Ancestral sequence reconstruction - An underused approach to understand the evolution of gene function in plants? *Comput Struct Biotechnol J* **19**, 1579-1594 (2021). https://doi.org/10.1016/j.csbj.2021.03.008
35 Vialle, R. A., Tamuri, A. U. & Goldman, N. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Mol Biol Evol* **35**, 1783-1797 (2018). https://doi.org/10.1093/molbev/msy055
36 Lichman, B. R. in *Catharanthus roseus: Methods and Protocols* (eds Vincent Courdavault & Sebastien Besseau) 165-179 (Springer US, 2022).
37 Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-314 (2012). https://doi.org/10.1038/nrg3186
38 Zwickl, D. J. & Hillis, D. M. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol* **51**, 588-598 (2002). https://doi.org/10.1080/10635150290102339
39 Graybeal, A. Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem? *Syst Biol* **47**, 9-17 (1998).
40 Redelings, B. D. BAli-Phy version 3: model-based co-estimation of alignment and phylogeny. *Bioinformatics* **37**, 3032-3034 (2021). https://doi.org/10.1093/bioinformatics/btab129
41 Armougom, F. *et al.* Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res* **34**, W604-608 (2006). https://doi.org/10.1093/nar/gkl092
42 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217 (2000). https://doi.org/10.1006/jmbi.2000.4042
43 Zaharias, P., Grosshauser, M. & Warnow, T. Re-evaluating Deep Neural Networks for Phylogeny Estimation: The Issue of Taxon Sampling. *Journal of Computational Biology* **29**, 74-89 (2022). https://doi.org/10.1089/cmb.2021.0383
44 Rangel, L. T. & Fournier, G. P. Fast-Evolving Alignment Sites Are Highly Informative for Reconstructions of Deep Tree of Life Phylogenies. *Microorganisms* **11** (2023). https://doi.org/10.3390/microorganisms11102499
45 Sennett, M. A. & Theobald, D. L. Extant Sequence Reconstruction: The Accuracy of Ancestral Sequence Reconstructions Evaluated by Extant Sequence Cross-Validation. *J Mol Evol* **92**, 181-206 (2024). https://doi.org/10.1007/s00239-024-10162-3
46 Steenwyk, J. L., Li, Y., Zhou, X., Shen, X. X. & Rokas, A. Incongruence in the phylogenomics era. *Nat Rev Genet* **24**, 834-850 (2023). https://doi.org/10.1038/s41576-023-00620-x
47 Jowkar, G., Pecerska, J., Gil, M. & Anisimova, M. Single-character insertion-deletion model preserves long indels in ancestral sequence reconstruction. *BMC Bioinformatics* **25**, 370 (2024). https://doi.org/10.1186/s12859-024-05986-1
48 Anisimova, M., Cannarozzi, G. & Liberles, D. A. Finding the balance between the mathematical and biological optima in multiple sequence alignment. *Trends in Evolutionary Biology* **2** (2010). https://doi.org/10.4081/eb.2010.e7
49 *Multiple Sequence Alignment Methods*. (Springer, 2014).
50 Chatzou, M. *et al.* Multiple sequence alignment modeling: methods and applications. *Brief Bioinform* **17**, 1009-1023 (2016). https://doi.org/10.1093/bib/bbv099
51 Chowdhury, B. & Garai, G. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics* **109**, 419-431 (2017). https://doi.org/10.1016/j.ygeno.2017.06.007
52 Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **5**, 345-352 (1978).

53 Henikoff, S. & Henikoff, J. Amino acid substitution matrices from protein blocks. *PNAS* **89**, 10915-10919 (1992).
54 Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **27**, 135-145 (2018). https://doi.org/10.1002/pro.3290
55 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004). https://doi.org/10.1093/nar/gkh340
56 Hogeweg, P. & Hesper, B. The Alignment of Sets of Sequences and the Construction of Phyletic Trees: An Integrated Method. *J Mol Evol* **20**, 175-186 (1984).
57 Chao, J., Tang, F. & Xu, L. Developments in Algorithms for Sequence Alignment: A Review. *Biomolecules* **12** (2022). https://doi.org/10.3390/biom12040546
58 Mirarab, S. *et al.* PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *J Comput Biol* **22**, 377-386 (2015). https://doi.org/10.1089/cmb.2014.0156
59 Garriga, E. *et al.* Large multiple sequence alignments with a root-to-leaf regressive method. *Nat Biotechnol* **37**, 1466-1470 (2019). https://doi.org/10.1038/s41587-019-0333-6
60 Katoh, K. & Toh, H. PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics* **23**, 372-374 (2007). https://doi.org/10.1093/bioinformatics/btl592
61 Blackburne, B. P. & Whelan, S. Measuring the distance between multiple sequence alignments. *Bioinformatics* **28**, 495-502 (2012). https://doi.org/10.1093/bioinformatics/btr701
62 Nute, M., Saleh, E. & Warnow, T. Evaluating Statistical Multiple Sequence Alignment in Comparison to Other Alignment Methods on Protein Data Sets. *Syst Biol* **68**, 396-411 (2019). https://doi.org/10.1093/sysbio/syy068
63 Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* **15**, 330-340 (2005). https://doi.org/10.1101/gr.2821705
64 Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**, 286-298 (2008). https://doi.org/10.1093/bib/bbn013
65 Liu, Y., Schmidt, B. & Maskell, D. L. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* **26**, 1958-1964 (2010). https://doi.org/10.1093/bioinformatics/btq338
66 Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, 2295-2300 (2008). https://doi.org/10.1093/nar/gkn072
67 Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res* **47**, W5-W10 (2019). https://doi.org/10.1093/nar/gkz342
68 Aadland, K. & Kolaczkowski, B. Alignment-Integrated Reconstruction of Ancestral Sequences Improves Accuracy. *Genome Biol Evol* **12**, 1549-1565 (2020). https://doi.org/10.1093/gbe/evaa164
69 Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34**, 1692-1699 (2006). https://doi.org/10.1093/nar/gkl091
70 Redelings, B. D. & Suchard, M. A. Joint Bayesian estimation of alignment and phylogeny. *Syst Biol* **54**, 401-418 (2005). https://doi.org/10.1080/10635150590947041

71 Iglhaut, C., Pecerska, J., Gil, M. & Anisimova, M. Please Mind the Gap: Indel-Aware Parsimony for Fast and Accurate Ancestral Sequence Reconstruction and Multiple Sequence Alignment Including Long Indels. *Mol Biol Evol* **41** (2024). https://doi.org/10.1093/molbev/msae109
72 Löytynoja, A. in *Multiple Sequence Alignment Methods* (ed David J. Russell) 155-170 (Humana Press, 2014).
73 Liu, K. *et al.* SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst Biol* **61**, 90-106 (2012). https://doi.org/10.1093/sysbio/syr095
74 Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res* **43**, W7-14 (2015). https://doi.org/10.1093/nar/gkv318
75 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009). https://doi.org/10.1093/bioinformatics/btp348
76 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577 (2007). https://doi.org/10.1080/10635150701472164
77 Dessimoz, C. & Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biology* **11** (2010).
78 Tan, G. *et al.* Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Syst Biol* **64**, 778-791 (2015). https://doi.org/10.1093/sysbio/syv033
79 Islam, F., Basilone, N., Yoo, V., Ball, E. & Shilton, B. Evolutionary analysis of Quinone Reductases 1 and 2 suggests that NQO2 evolved to function as a pseudoenzyme. *Protein Sci* **33**, e5234 (2024). https://doi.org/10.1002/pro.5234
80 de la Fuente, I., Manzano-Morales, S., Sanz, D., Prieto, A. & Barriuso, J. Quorum sensing in bacteria: in silico protein analysis, ecophysiology, and reconstruction of their evolutionary history. *BMC Genomics* **25**, 441 (2024). https://doi.org/10.1186/s12864-024-10355-6
81 Wiseglass, G. & Rubinstein, R. Following the Evolutionary Paths of Dscam1 Proteins toward Highly Specific Homophilic Interactions. *Mol Biol Evol* **41** (2024). https://doi.org/10.1093/molbev/msae141
82 Zeng, B. *et al.* Ancestral sequence reconstruction of the prokaryotic three-domain laccases for efficiently degrading polyethylene. *J Hazard Mater* **476**, 135012 (2024). https://doi.org/10.1016/j.jhazmat.2024.135012
83 Luu, D. D. *et al.* Evidence that the cold-- and menthol--sensing functions of the human TRPM8 channel evolved separately. *Science Advances* **10** (2024).
84 Ghosh, M., Basak, S. & Dutta, S. Evolutionary divergence of TLR9 through ancestral sequence reconstruction. *Immunogenetics* **76**, 203-211 (2024). https://doi.org/10.1007/s00251-024-01338-8
85 Mickael, M. *et al.* FEZF2 and AIRE1: An Evolutionary Trade-off in the Elimination of Auto-reactive T Cells in the Thymus. *J Mol Evol* **92**, 72-86 (2024). https://doi.org/10.1007/s00239-024-10157-0
86 Franceus, J., Rivas-Fernandez, J. P., Lormans, J., Rovira, C. & Desmet, T. Evolution of Phosphorylase Activity in an Ancestral Glycosyltransferase. *ACS Catal* **14**, 3103-3114 (2024). https://doi.org/10.1021/acscatal.3c05819
87 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). https://doi.org/10.1371/journal.pcbi.0020069

88 Park, Y., Patton, J. E. J., Hochberg, G. K. A. & Thornton, J. W. Comment on "Ancient origins of allosteric activation in a Ser-Thr kinase". *Science* (2020). https://doi.org/10.1126/science.abc8301
89 Hanson-Smith, V., Kolaczkowski, B. & Thornton, J. W. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol* **27**, 1988-1999 (2010). https://doi.org/10.1093/molbev/msq081
90 Randall, R. N., Radford, C. E., Roof, K. A., Natarajan, D. K. & Gaucher, E. A. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun* **7**, 12847 (2016). https://doi.org/10.1038/ncomms12847
91 Tavare, S. Some Probabilistic and Statistical Problems in the Analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57-67 (1986).
92 Goldman, N. Statistical Tests of Models of DNA Substitution. *J Mol Evol* **36**, 182-198 (1993).
93 Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**, 275-282 (1992).
94 MÜLLER, T. & VINGRON, M. Modeling Amino Acid Replacement. *Journal of Computational Biology* **7**, 761-776 (2000).
95 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**, 691-699 (2001).
96 Kosiol, C. & Goldman, N. Different versions of the Dayhoff rate matrix. *Mol Biol Evol* **22**, 193-199 (2005). https://doi.org/10.1093/molbev/msi005
97 Minh, B. Q., Dang, C. C., Vinh, L. S. & Lanfear, R. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Syst Biol* **70**, 1046-1060 (2021). https://doi.org/10.1093/sysbio/syab010
98 Yang, Z. Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *J Mol Evol* **39**, 306-314 (1994).
99 Uzzell, T. & Corbin, K. W. Fitting discrete probability distributions to evolutionary events. *Science* **172**, 1089-1096 (1971). https://doi.org/D - NASA: 71186188 EDAT- 1971/06/11 00:00 MHDA- 1971/06/11 00:01 CRDT- 1971/06/11 00:00 PHST- 1971/06/11 00:00 [pubmed] PHST- 1971/06/11 00:01 [medline] PHST- 1971/06/11 00:00 [entrez] AID - 10.1126/science.172.3988.1089 [doi] PST - ppublish
100 Soubrier, J. *et al.* The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol Biol Evol* **29**, 3345-3358 (2012). https://doi.org/10.1093/molbev/mss140
101 Hasegawa, M., Kishino, H. & Yano, T.-a. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *J Mol Evol* **22**, 160-174 (1985).
102 Gu, X., Fu, Y.-X. & Li, W.-H. Maximum likelihood estimation of the heterogeneity of rate substitution among nucleotide sites. *Mol Biol Evol* **12**, 546-557 (1995).
103 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589 (2017). https://doi.org/10.1038/nmeth.4285
104 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165 (2011). https://doi.org/10.1093/bioinformatics/btr088
105 Abadi, S., Azouri, D., Pupko, T. & Mayrose, I. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun* **10**, 934 (2019). https://doi.org/10.1038/s41467-019-08822-w

106 Spielman, S. J. Relative Model Fit Does Not Predict Topological Accuracy in Single-Gene Protein Phylogenetics. *Mol Biol Evol* **37**, 2110-2123 (2020). https://doi.org/10.1093/molbev/msaa075
107 Tao, Q., Barba-Montoya, J., Huuki, L. A., Durnan, M. K. & Kumar, S. Relative Efficiencies of Simple and Complex Substitution Models in Estimating Divergence Times in Phylogenomics. *Mol Biol Evol* **37**, 1819-1831 (2020). https://doi.org/10.1093/molbev/msaa049
108 Del Amparo, R. & Arenas, M. Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction. *Mol Biol Evol* **39** (2022). https://doi.org/10.1093/molbev/msac144
109 Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. Evolution of increased complexity in a molecular machine. *Nature* **481**, 360-364 (2012). https://doi.org/10.1038/nature10724
110 Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *TREE* **11**, 367-371 (1996).
111 Lemmon, A. R. & Moriarty, E. C. The importance of proper model assumption in bayesian phylogenetics.
112 Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol* **31**, 1956-1978 (2014). https://doi.org/10.1093/molbev/msu173
113 Posada, D. & Buckley, T. R. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol* **53**, 793-808 (2004). https://doi.org/10.1080/10635150490522304
114 Minin, V., Abdo, Z., Joyce, P. & Sullivan, J. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. *Systematic Biology* **52**, 674-683 (2003). https://doi.org/10.1080/10635150390235494
115 Sullivan, J. & Swofford, D. L. Are Guinea Pigs Rodents? The Importance of Adequate Models in Molecular Phylogenetics. *J Mamm Evol* **4**, 77-86 (1997).
116 Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* **7 Suppl 1**, S4 (2007). https://doi.org/10.1186/1471-2148-7-S1-S4
117 Susko, E. & Roger, A. J. Long Branch Attraction Biases in Phylogenetics. *Syst Biol* **70**, 838-843 (2021). https://doi.org/10.1093/sysbio/syab001
118 Sullivan, J. & Joyce, P. Model Selection in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **36**, 445-466 (2005). https://doi.org/10.1146/annurev.ecolsys.36.102003.152633
119 Akaike, H. in *Selected Papers of Hirotugu Akaike* (eds Emanuel Parzen, Kunio Tanabe, & Genshiro Kitagawa) 199-213 (Springer New York, 1998).
120 Sugiura, N. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods* **7**, 13-26 (1978). https://doi.org/10.1080/03610927808827599
121 Hurvich, C. M. & Tsai, C.-l. Regression and time series model selection in small samples. *Biometrika* **76**, 297-307 (1998).
122 Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461-464 (1978).
123 Posada, D. jModelTest: phylogenetic model averaging. *Mol Biol Evol* **25**, 1253-1256 (2008). https://doi.org/10.1093/molbev/msn083
124 Suvorov, A., Hochuli, J. & Schrider, D. R. Accurate Inference of Tree Topologies from Multiple Sequence Alignments Using Deep Learning. *Syst Biol* **69**, 221-233 (2020). https://doi.org/10.1093/sysbio/syz060

125 Zou, Z., Zhang, H., Guan, Y. & Zhang, J. Deep Residual Neural Networks Resolve Quartet Molecular Phylogenies. *Mol Biol Evol* **37**, 1495-1507 (2020). https://doi.org/10.1093/molbev/msz307
126 Wang, Z. *et al.* Fusang: a framework for phylogenetic tree inference via deep learning. *Nucleic Acids Res* **51**, 10909-10923 (2023). https://doi.org/10.1093/nar/gkad805
127 Smith, M. L. & Hahn, M. W. Phylogenetic inference using generative adversarial networks. *Bioinformatics* **39** (2023). https://doi.org/10.1093/bioinformatics/btad543
128 Mo, Y. K., Hahn, M. W. & Smith, M. L. Applications of machine learning in phylogenetics. *Mol Phylogenet Evol* **196**, 108066 (2024). https://doi.org/10.1016/j.ympev.2024.108066
129 Azouri, D., Abadi, S., Mansour, Y., Mayrose, I. & Pupko, T. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nat Commun* **12**, 1983 (2021). https://doi.org/10.1038/s41467-021-22073-8
130 Leuchtenberger, A. F. & von Haeseler, A. Learning From an Artificial Neural Network in Phylogenetics. *IEEE/ACM Trans Comput Biol Bioinform* **21**, 278-288 (2024). https://doi.org/10.1109/TCBB.2024.3352268
131 Tao, Q., Tamura, K., F, U. B. & Kumar, S. A Machine Learning Method for Detecting Autocorrelation of Evolutionary Rates in Large Phylogenies. *Mol Biol Evol* **36**, 811-824 (2019). https://doi.org/10.1093/molbev/msz014
132 Haag, J., Hohler, D., Bettisworth, B. & Stamatakis, A. From Easy to Hopeless-Predicting the Difficulty of Phylogenetic Analyses. *Mol Biol Evol* **39** (2022). https://doi.org/10.1093/molbev/msac254
133 Zhang, Y. *et al.* Inferring Historical Introgression with Deep Learning. *Syst Biol* **72**, 1013-1038 (2023). https://doi.org/10.1093/sysbio/syad033
134 Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology* **20**, 406-416 (1971). https://doi.org/10.1093/sysbio/20.4.406
135 SCHLUTER, D., PRICE, T., MOOERS, A. O. & LUDWIG, D. LIKELIHOOD OF ANCESTOR STATES IN ADAPTIVE RADIATION. *Evolution* **51**, 1699–1711 (1997).
136 Cunningham, C. W., Omland, K. E. & Oakley, T. H. Reconstructing ancestral character states: a critical reappraisal. *TREE* **9**, 361-366 (1998).
137 CUNNINGHAM, C. W. Some Limitations of Ancestral Character-State Reconstruction When Testing Evolutionary Hypotheses. *Syst Biol* **48**, 665-674 (1999).
138 Eyre-Walker, A. Problems with Parsimony in Sequences of Biased Base Composition. *J Mol Evol* **47**, 686-690 (1998).
139 Kolaczkowski, B. & Thornton, J. W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980-984 (2004). https://doi.org/10.1038/nature02928
140 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**, 268-274 (2015). https://doi.org/10.1093/molbev/msu300
141 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014). https://doi.org/10.1093/bioinformatics/btu033
142 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010). https://doi.org/10.1093/sysbio/syq010

143 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020). https://doi.org/10.1093/molbev/msaa015
144 Holder, M. & Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet* **4**, 275-284 (2003). https://doi.org/10.1038/nrg1044
145 Nascimento, F. F., Reis, M. D. & Yang, Z. A biologist's guide to Bayesian phylogenetic analysis. *Nat Ecol Evol* **1**, 1446-1454 (2017). https://doi.org/10.1038/s41559-017-0280-x
146 Bickel, D. R. *Phylogenetic Trees and Molecular Evolution: A Hands-on Introduction with Uncertainty Quantification Corrected*. (Springer Nature, 2022).
147 Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **10**, e1003537 (2014). https://doi.org/10.1371/journal.pcbi.1003537
148 Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference ofphylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).
149 Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286-2288 (2009). https://doi.org/10.1093/bioinformatics/btp368
150 Royer-Carenzi, M., Pontarotti, P. & Didier, G. Choosing the best ancestral character state reconstruction method. *Math Biosci* **242**, 95-109 (2013). https://doi.org/10.1016/j.mbs.2012.12.003
151 Eick, G. N., Bridgham, J. T., Anderson, D. P., Harms, M. J. & Thornton, J. W. Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. *Mol Biol Evol* **34**, 247-261 (2017). https://doi.org/10.1093/molbev/msw223
152 Som, A. Causes, consequences and solutions of phylogenetic incongruence. *Brief Bioinform* **16**, 536-548 (2015). https://doi.org/10.1093/bib/bbu015
153 Holmes, S. Bootstrapping Phylogenetic Trees: Theory and Methods. *Statistical Science* **18**, 241-255 (2003).
154 Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **39**, 783-791 (1985).
155 Minh, B. Q., Nguyen, M. A. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188-1195 (2013). https://doi.org/10.1093/molbev/mst024
156 HILLIS, D. M. & BULL, J. J. AN EMPIRICAL TEST OF BOOTSTRAPPING AS A METHOD FOR ASSESSING CONFIDENCE IN PHYLOGENETIC ANALYSIS. *Syst Biol* **42**, 182-192 (1993).
157 Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol* **60**, 685-699 (2011). https://doi.org/10.1093/sysbio/syr041
158 Hoang, D. T., Chernomor, O., Haeseler, A. v., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522 (2017). https://doi.org/10.5281/zenodo.854445
159 Anisimova, M. & Gascuel, O. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Syst Biol* **55**, 539-552 (2006).
160 Planet, P. J. Tree disagreement: measuring and testing incongruence in phylogenies. *J Biomed Inform* **39**, 86-102 (2006). https://doi.org/10.1016/j.jbi.2005.08.008
161 Goldman, N., Anderson, J. P. & Rodrigo, A. Likelihood-Based Tests of Topologies in Phylogenetics. *Syst Biol* **49**, 652-670 (2000).

162 Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492-508 (2002). https://doi.org/10.1080/10635150290069913
163 Schmidt, H. A. in *The Phylogenetic Handbook* Ch. 12, 381-404 (2009).
164 Strimmer, K. & von Haeseler, A. Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution* **13**, 964-964 (1996). https://doi.org/10.1093/oxfordjournals.molbev.a025664
165 Schmidt, H. A., Petzold, E., Vingron, M. & von Haeseler, A. Molecular phylogenetics: parallelized parameter estimation and quartet puzzling. *Journal of Parallel and Distributed Computing* **63**, 719-727 (2003). https://doi.org/10.1016/s0743-7315(03)00129-1
166 Ranwez, V. & Gascuel, O. Quartet-Based Phylogenetic Inference: Improvements and Limits. *Mol Biol Evol* **18**, 1103-1116 (2001).
167 Brown, J. M. & Thomson, R. C. Evaluating Model Performance in Evolutionary Biology. *Annual Review of Ecology, Evolution, and Systematics* **49**, 95-114 (2018). https://doi.org/10.1146/annurev-ecolsys-110617-062249
168 Simon, C. An Evolving View of Phylogenetic Support. *Syst Biol* **71**, 921-928 (2022). https://doi.org/10.1093/sysbio/syaa068
169 *Bioinformatics*. Vol. 1 (Springer, 2017).
170 Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).
171 Philippe, H. & Douady, C. J. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol* **6**, 498-505 (2003). https://doi.org/10.1016/j.mib.2003.09.008
172 Daubin, V. & Szollosi, G. J. Horizontal Gene Transfer and the History of Life. *Cold Spring Harb Perspect Biol* **8**, a018036 (2016). https://doi.org/10.1101/cshperspect.a018036
173 Kelley, L. A. & Sternberg, M. J. Partial protein domains: evolutionary insights and bioinformatics challenges. *Genome Biol* **16**, 100 (2015). https://doi.org/10.1186/s13059-015-0663-8
174 Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163-193 (2005). https://doi.org/10.1111/j.1096-0031.2005.00059.x
175 Mai, U. & Mirarab, S. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* **19**, 272 (2018). https://doi.org/10.1186/s12864-018-4620-2
176 Struck, T. H. TreSpEx-Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. *Evol Bioinform Online* **10**, 51-67 (2014). https://doi.org/10.4137/EBO.S14239
177 Krishnan, N. M., Seligmann, H., Stewart, C. B., De Koning, A. P. & Pollock, D. D. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. *Mol Biol Evol* **21**, 1871-1883 (2004). https://doi.org/10.1093/molbev/msh198
178 Pupko, T., Pe, I., Shamir, R. & Graur, D. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution* **17**, 890-896 (2000). https://doi.org/10.1093/oxfordjournals.molbev.a026369
179 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007). https://doi.org/10.1093/molbev/msm088
180 Ashkenazy, H. *et al.* FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res* **40**, W580-584 (2012). https://doi.org/10.1093/nar/gks498
181 Foley, G. *et al.* Engineering indel and substitution variants of diverse and ancient enzymes using Graphical Representation of Ancestral Sequence Predictions

(GRASP). *PLoS Comput Biol* **18**, e1010633 (2022). https://doi.org/10.1371/journal.pcbi.1010633
182 Arenas, M., Weber, C. C., Liberles, D. A. & Bastolla, U. ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. *Syst Biol* **66**, 1054-1064 (2017). https://doi.org/10.1093/sysbio/syw121
183 Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409-413 (2017). https://doi.org/10.1038/nature23902
184 Akanuma, S., Yokobori, S., Nakajima, Y., Bessho, M. & Yamagishi, A. Robustness of predictions of extremely thermally stable proteins in ancient organisms. *Evolution* **69**, 2954-2962 (2015). https://doi.org/10.1111/evo.12779
185 Trudeau, D. L., Kaltenbach, M. & Tawfik, D. S. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol Biol Evol* **33**, 2633-2641 (2016). https://doi.org/10.1093/molbev/msw138
186 Chisholm, L. O., Orlandi, K. N., Phillips, S. R., Shavlik, M. J. & Harms, M. J. Ancestral Reconstruction and the Evolution of Protein Energy Landscapes. *Annu Rev Biophys* **53**, 127-146 (2024). https://doi.org/10.1146/annurev-biophys-030722-125440
187 Moshe, A. & Pupko, T. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. *Bioinformatics* **35**, 2562-2568 (2019). https://doi.org/10.1093/bioinformatics/bty1031
188 Akanuma, S. *et al.* Experimental evidence for the thermophilicity of ancestral life. *Proc Natl Acad Sci U S A* **110**, 11067-11072 (2013). https://doi.org/10.1073/pnas.1308215110
189 Gaucher, E. A., Govindarajan, S. & Ganesh, O. K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704-707 (2008). https://doi.org/10.1038/nature06510
190 Wheeler, L. C., Lim, S. A., Marqusee, S. & Harms, M. J. The thermostability and specificity of ancient proteins. *Curr Opin Struct Biol* **38**, 37-43 (2016). https://doi.org/10.1016/j.sbi.2016.05.015
191 Dias, R., Manny, A., Kolaczkowski, O. & Kolaczkowski, B. Convergence of Domain Architecture, Structure, and Ligand Affinity in Animal and Plant RNA-Binding Proteins. *Mol Biol Evol* **34**, 1429-1444 (2017). https://doi.org/10.1093/molbev/msx090
192 Kaltenbach, M. *et al.* Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat Chem Biol* **14**, 548-555 (2018). https://doi.org/10.1038/s41589-018-0042-3
193 Pillai, A. S. *et al.* Origin of complexity in haemoglobin evolution. *Nature* **581**, 480-485 (2020). https://doi.org/10.1038/s41586-020-2292-y
194 Kaczmarski, J. A. *et al.* Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme. *Nat Commun* **11**, 5945 (2020). https://doi.org/10.1038/s41467-020-19695-9
195 Voordeckers, K. *et al.* Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol* **10**, e1001446 (2012). https://doi.org/10.1371/journal.pbio.1001446
196 Zakas, P. M. *et al.* Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat Biotechnol* **35**, 35-37 (2017). https://doi.org/10.1038/nbt.3677
197 Zinn, E. *et al.* In Silico Reconstruction of the Viral Evolutionary Lineage Yields a Potent Gene Therapy Vector. *Cell Rep* **12**, 1056-1068 (2015). https://doi.org/10.1016/j.celrep.2015.07.019

198 Joho, Y. *et al.* Enhancing PET Degrading Enzymes: A Combinatory Approach. *Chembiochem* **25**, e202400084 (2024). https://doi.org/10.1002/cbic.202400084
199 Tyzack, J. D., Furnham, N., Sillitoe, I., Orengo, C. M. & Thornton, J. M. Understanding enzyme function evolution from a computational perspective. *Curr Opin Struct Biol* **47**, 131-139 (2017). https://doi.org/10.1016/j.sbi.2017.08.003
200 Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat Genet* **37**, 73-76 (2005). https://doi.org/10.1038/ng1482
201 Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nat Rev Genet* **11**, 572-582 (2010). https://doi.org/10.1038/nrg2808
202 Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr Opin Struct Biol* **47**, 113-122 (2017). https://doi.org/10.1016/j.sbi.2017.07.003
203 Mallik, S., Tawfik, D. S. & Levy, E. D. How gene duplication diversifies the landscape of protein oligomeric state and function. *Curr Opin Genet Dev* **76**, 101966 (2022). https://doi.org/10.1016/j.gde.2022.101966
204 Birchler, J. A. & Yang, H. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* **34**, 2466-2474 (2022). https://doi.org/10.1093/plcell/koac076
205 Harris, K. L. *et al.* Ancestral Sequence Reconstruction of a Cytochrome P450 Family Involved in Chemical Defense Reveals the Functional Evolution of a Promiscuous, Xenobiotic-Metabolizing Enzyme in Vertebrates. *Mol Biol Evol* **39** (2022). https://doi.org/10.1093/molbev/msac116
206 Pillai, A. S., Hochberg, G. K. A. & Thornton, J. W. Simple mechanisms for the evolution of protein complexity. *Protein Sci* **31**, e4449 (2022). https://doi.org/10.1002/pro.4449
207 Tran, T. T. Q. *et al.* Ancestral sequence reconstruction dissects structural and functional differences among eosinophil ribonucleases. *J Biol Chem* **300**, 107280 (2024). https://doi.org/10.1016/j.jbc.2024.107280
208 Van Otterloo, E., Cornell, R. A., Medeiros, D. M. & Garnett, A. T. Gene regulatory evolution and the origin of macroevolutionary novelties: insights from the neural crest. *Genesis* **51**, 457-470 (2013). https://doi.org/10.1002/dvg.22403
209 Voordeckers, K., Pougach, K. & Verstrepen, K. J. How do regulatory networks evolve and expand throughout evolution? *Curr Opin Biotechnol* **34**, 180-188 (2015). https://doi.org/10.1016/j.copbio.2015.02.001
210 Liberles, D. A. *et al.* The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci* **21**, 769-785 (2012). https://doi.org/10.1002/pro.2071
211 Marsh, J. A. & Teichmann, S. A. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *Bioessays* **36**, 209-218 (2014). https://doi.org/10.1002/bies.201300134
212 Echave, J. & Wilke, C. O. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annu Rev Biophys* **46**, 85-103 (2017). https://doi.org/10.1146/annurev-biophys-070816-033819
213 Miton, C. M., Buda, K. & Tokuriki, N. Epistasis and intramolecular networks in protein evolution. *Curr Opin Struct Biol* **69**, 160-168 (2021). https://doi.org/10.1016/j.sbi.2021.04.007
214 Lemay-St-Denis, C., Doucet, N. & Pelletier, J. N. Integrating dynamics into enzyme engineering. *Protein Eng Des Sel* **35** (2022). https://doi.org/10.1093/protein/gzac015

215 Kim, H. *et al.* A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in GFP-like proteins. *Structure* **23**, 34-43 (2015). https://doi.org/10.1016/j.str.2014.11.011
216 Hueting, D. A., Vanga, S. R. & Syren, P. O. Thermoadaptation in an Ancestral Diterpene Cyclase by Altered Loop Stability. *J Phys Chem B* **126**, 3809-3821 (2022). https://doi.org/10.1021/acs.jpcb.1c10605
217 Harms, M. J. *et al.* Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. *Proc Natl Acad Sci U S A* **110**, 11475-11480 (2013). https://doi.org/10.1073/pnas.1303930110
218 Wilson, C. *et al.* Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **347**, 882-886 (2015). https://doi.org/10.1126/science.aaa1823
219 Chaloupkova, R. *et al.* Light-Emitting Dehalogenases: Reconstruction of Multifunctional Biocatalysts. *ACS Catalysis* **9**, 4810-4823 (2019). https://doi.org/10.1021/acscatal.9b01031
220 Schenkmayerova, A. *et al.* Engineering the protein dynamics of an ancestral luciferase. *Nat Commun* **12**, 3616 (2021). https://doi.org/10.1038/s41467-021-23450-z
221 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). https://doi.org/10.1016/j.jbc.2022.102435
222 Nicoll, C. R., Massari, M., Fraaije, M. W., Mascotti, M. L. & Mattevi, A. Impact of ancestral sequence reconstruction on mechanistic and structural enzymology. *Curr Opin Struct Biol* **82**, 102669 (2023). https://doi.org/10.1016/j.sbi.2023.102669
223 Prakinee, K., Phaisan, S., Kongjaroon, S. & Chaiyen, P. Ancestral Sequence Reconstruction for Designing Biocatalysts and Investigating their Functional Mechanisms. *JACS Au* **4**, 4571-4591 (2024). https://doi.org/10.1021/jacsau.4c00653
224 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 beta-lactamases. *J Am Chem Soc* **135**, 2899-2902 (2013). https://doi.org/10.1021/ja311630a
225 Okafor, C. D. *et al.* Structural and Dynamics Comparison of Thermostability in Ancient, Modern, and Consensus Elongation Factor Tus. *Structure* **26**, 118-129 e113 (2018). https://doi.org/10.1016/j.str.2017.11.018
226 Livada, J., Vargas, A. M., Martinez, C. A. & Lewis, R. D. Ancestral Sequence Reconstruction Enhances Gene Mining Efforts for Industrial Ene Reductases by Expanding Enzyme Panels with Thermostable Catalysts. *ACS Catalysis* **13**, 2576-2585 (2023). https://doi.org/10.1021/acscatal.2c03859
227 Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How protein stability and new functions trade off. *PLoS Comput Biol* **4**, e1000002 (2008). https://doi.org/10.1371/journal.pcbi.1000002
228 Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* **19**, 596-604 (2009). https://doi.org/10.1016/j.sbi.2009.08.003
229 Socha, R. D. & Tokuriki, N. Modulating protein stability - directed evolution strategies for improved protein function. *FEBS J* **280**, 5582-5595 (2013). https://doi.org/10.1111/febs.12354
230 Akanuma, S. *et al.* Phylogeny-based design of a B-subunit of DNA gyrase and its ATPase domain using a small set of homologous amino acid sequences. *J Mol Biol* **412**, 212-225 (2011). https://doi.org/10.1016/j.jmb.2011.07.042

231 Furukawa, R., Toma, W., Yamazaki, K. & Akanuma, S. Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties. *Sci Rep* **10**, 15493 (2020). https://doi.org/10.1038/s41598-020-72418-4
232 Miyazaki, J. *et al.* Ancestral Residues Stabilizing 3-Isopropylmalate Dehydrogenase of an Extreme Thermophile: Experimental Evidence Supporting the Thermophilic Common Ancestor Hypothesis. *J Biochem* **129**, 777-782 (2001).
233 Shimizu, H. *et al.* Extremely thermophilic translation system in the common ancestor commonote: ancestral mutants of Glycyl-tRNA synthetase from the extreme thermophile Thermus thermophilus. *J Mol Biol* **369**, 1060-1069 (2007). https://doi.org/10.1016/j.jmb.2007.04.001
234 Yamashiro, K., Yokobori, S., Koikeda, S. & Yamagishi, A. Improvement of Bacillus circulans beta-amylase activity attained using the ancestral mutation method. *Protein Eng Des Sel* **23**, 519-528 (2010). https://doi.org/10.1093/protein/gzq021
235 Gumulya, Y. *et al.* Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nature Catalysis* **1**, 878-888 (2018). https://doi.org/10.1038/s41929-018-0159-5
236 Thomas, A., Cutlan, R., Finnigan, W., van der Giezen, M. & Harmer, N. Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun Biol* **2**, 429 (2019). https://doi.org/10.1038/s42003-019-0677-y
237 Hendrikse, N. M., Charpentier, G., Nordling, E. & Syren, P. O. Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *FEBS J* **285**, 4660-4673 (2018). https://doi.org/10.1111/febs.14686
238 Whitfield, J. H. *et al.* Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci* **24**, 1412-1422 (2015). https://doi.org/10.1002/pro.2721
239 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). https://doi.org/10.1038/s41594-019-0347-2
240 Wilding, M. *et al.* Reverse engineering: transaminase biocatalyst development using ancestral sequence reconstruction. *Green Chemistry* **19**, 5375-5380 (2017). https://doi.org/10.1039/c7gc02343j
241 Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R. & Damborsky, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *Chembiochem* **18**, 1448-1456 (2017). https://doi.org/10.1002/cbic.201700197
242 Wang, Y. *et al.* Directed Evolution: Methodologies and Applications. *Chem Rev* **121**, 12384-12444 (2021). https://doi.org/10.1021/acs.chemrev.1c00260
243 Gargiulo, S. & Soumillion, P. Directed evolution for enzyme development in biocatalysis. *Curr Opin Chem Biol* **61**, 107-113 (2021). https://doi.org/10.1016/j.cbpa.2020.11.006
244 Selles Vidal, L., Isalan, M., Heap, J. T. & Ledesma-Amaro, R. A primer to directed evolution: current methodologies and future directions. *RSC Chem Biol* **4**, 271-291 (2023). https://doi.org/10.1039/d2cb00231k
245 Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929-932 (2006). https://doi.org/10.1038/nature05385
246 Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **324**, 203-207 (2009).

247 Tokuriki, N. *et al.* Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. *Nat Commun* **3**, 1257 (2012). https://doi.org/10.1038/ncomms2246
248 Bloom, J. D. & Arnold, F. H. In the light of directed evolution: Pathways of adaptive protein evolution. *PNAS* **106**, 9995–10000 (2009).
249 Karkare, K., Lai, H. Y., Azevedo, R. B. R. & Cooper, T. F. Historical Contingency Causes Divergence in Adaptive Expression of the lac Operon. *Mol Biol Evol* **38**, 2869-2879 (2021). https://doi.org/10.1093/molbev/msab077
250 Salverda, M. L. *et al.* Initial mutations direct alternative pathways of protein evolution. *PLoS Genet* **7**, e1001321 (2011). https://doi.org/10.1371/journal.pgen.1001321
251 Zheng, J., Bratulic, S., Lischer, H. E. L. & Wagner, A. Mistranslation can promote the exploration of alternative evolutionary trajectories in enzyme evolution. *J Evol Biol* **34**, 1302-1315 (2021). https://doi.org/10.1111/jeb.13892
252 Chica, R. A., Doucet, N. & Pelletier, J. N. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr Opin Biotechnol* **16**, 378-384 (2005). https://doi.org/10.1016/j.copbio.2005.06.004
253 Zhang, X. *et al.* Bacterial cytochrome P450-catalyzed regio- and stereoselective steroid hydroxylation enabled by directed evolution and rational design. *Bioresources and Bioprocessing* **7** (2020). https://doi.org/10.1186/s40643-019-0290-4
254 Martins, M. *et al.* Thermostability Enhancement of GH 62 alpha-l-Arabinofuranosidase by Directed Evolution and Rational Design. *J Agric Food Chem* **72**, 4225-4236 (2024). https://doi.org/10.1021/acs.jafc.3c08019
255 Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc Natl Acad Sci U S A* **116**, 8852-8858 (2019). https://doi.org/10.1073/pnas.1901979116
256 Chen, K. & Arnold, F. H. Engineering new catalytic activities in enzymes. *Nature Catalysis* **3**, 203-213 (2020). https://doi.org/10.1038/s41929-019-0385-5
257 Mills, D. R., Peterson, R. L. & Spiegelman, S. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *PNAS* **58**, 217-224 (1967).
258 Kuchner, O. & Arnold, F. H. Directed evolution of enzyme catalysts. *TIBTECH* **15**, 523-530 (1997).
259 Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol Direct* **2**, 17 (2007). https://doi.org/10.1186/1745-6150-2-17
260 Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* **79**, 471-505 (2010). https://doi.org/10.1146/annurev-biochem-030409-143718
261 Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *PNAS* **103**, 5869-5874 (2006).
262 Baier, F. & Tokuriki, N. Connectivity between catalytic landscapes of the metallo-beta-lactamase superfamily. *J Mol Biol* **426**, 2442-2456 (2014). https://doi.org/10.1016/j.jmb.2014.04.013
263 Payne, J. L. & Wagner, A. The causes of evolvability and their evolution. *Nat Rev Genet* **20**, 24-38 (2019). https://doi.org/10.1038/s41576-018-0069-z
264 Wilson, D. S. & Keefe, A. D. Random mutagenesis by PCR. *Curr Protoc Mol Biol* **Chapter 8**, Unit8 3 (2001). https://doi.org/10.1002/0471142727.mb0803s51
265 *Directed Evolution Library Creation*. 2 edn, (Springer, 2014).

266 Emond, S. *et al.* Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nat Commun* **11**, 3469 (2020). https://doi.org/10.1038/s41467-020-17061-3
267 Stemmer, W. P. C. DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. *PNAS* **91**, 10747-10751 (1994).
268 Selifonova, O., Valle, F. & Schellenberger, V. Rapid evolution of novel traits in microorganisms. *Appl Environ Microbiol* **67**, 3645-3649 (2001). https://doi.org/10.1128/AEM.67.8.3645-3649.2001
269 Moore, C. L., Papa, L. J., 3rd & Shoulders, M. D. A Processive Protein Chimera Introduces Mutations across Defined DNA Regions In Vivo. *J Am Chem Soc* **140**, 11560-11564 (2018). https://doi.org/10.1021/jacs.8b04001
270 Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894-898 (2009). https://doi.org/10.1038/nature08187
271 Halperin, S. O. *et al.* CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature* **560**, 248-252 (2018). https://doi.org/10.1038/s41586-018-0384-8
272 Jakociunas, T., Pedersen, L. E., Lis, A. V., Jensen, M. K. & Keasling, J. D. CasPER, a method for directed evolution in genomic contexts using mutagenesis and CRISPR/Cas9. *Metab Eng* **48**, 288-296 (2018). https://doi.org/10.1016/j.ymben.2018.07.001
273 *Directed evolution: Methods and protocols*. (Springer Nature, 2022).
274 Siloto, R. M. P. & Weselake, R. J. Site saturation mutagenesis: Methods and applications in protein engineering. *Biocatalysis and Agricultural Biotechnology* **1**, 181-189 (2012). https://doi.org/10.1016/j.bcab.2012.03.010
275 Qu, G., Li, A., Acevedo-Rocha, C. G., Sun, Z. & Reetz, M. T. The Crucial Role of Methodology Development in Directed Evolution of Selective Enzymes. *Angew Chem Int Ed Engl* **59**, 13204-13231 (2020). https://doi.org/10.1002/anie.201901491
276 Reetz, M. T. & Wu, S. Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem Commun (Camb)*, 5499-5501 (2008). https://doi.org/10.1039/b813388c
277 Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat Rev Genet* **16**, 379-394 (2015). https://doi.org/10.1038/nrg3927
278 Zahnd, C. *et al.* Directed in vitro evolution and crystallographic analysis of a peptide-binding single chain antibody fragment (scFv) with low picomolar affinity. *J Biol Chem* **279**, 18870-18877 (2004). https://doi.org/10.1074/jbc.M309169200
279 Traxlmayr, M. W. *et al.* Directed evolution of Her2/neu-binding IgG1-Fc for improved stability and resistance to aggregation by using yeast surface display. *Protein Eng Des Sel* **26**, 255-265 (2013). https://doi.org/10.1093/protein/gzs102
280 Acevedo-Rocha, C. G., Agudo, R. & Reetz, M. T. Directed evolution of stereoselective enzymes based on genetic selection as opposed to screening systems. *J Biotechnol* **191**, 3-10 (2014). https://doi.org/10.1016/j.jbiotec.2014.04.009
281 d'Oelsnitz, S. & Ellington, A. Continuous directed evolution for strain and protein engineering. *Curr Opin Biotechnol* **53**, 158-163 (2018). https://doi.org/10.1016/j.copbio.2017.12.020
282 Chen, A., Zhang, X. D., Delmas, A. D., Weitz, D. A. & Milcic, K. Systems and Methods for Continuous Evolution of Enzymes. *Chemistry* **30**, e202400880 (2024). https://doi.org/10.1002/chem.202400880
283 Ye, L., Yang, C. & Yu, H. From molecular engineering to process engineering: development of high-throughput screening methods in enzyme directed evolution.

*Appl Microbiol Biotechnol* **102**, 559-567 (2018). https://doi.org/10.1007/s00253-017-8568-y

284 Li, G. & Reetz, M. T. Learning lessons from directed evolution of stereoselective enzymes. *Organic Chemistry Frontiers* **3**, 1350-1358 (2016). https://doi.org/10.1039/c6qo00210b

285 Lin, Z., Thorsen, T. & Arnold, F. H. Functional expression of horseradish peroxidase in E. coli by directed evolution. *Biotechnol Prog* **15**, 467-471 (1999). https://doi.org/10.1021/bp990037r

286 Morawski, B., Quan, S. & Arnold, F. H. Functional Expression and Stabilization of Horseradish Peroxidase by Directed Evolution in *Saccharomyces cerevisiae*. *Biotechnology and Bioengineering* **76**, 99-107 (2001).

287 Bulter, T. *et al.* Functional expression of a fungal laccase in Saccharomyces cerevisiae by directed evolution. *Appl Environ Microbiol* **69**, 987-995 (2003). https://doi.org/10.1128/AEM.69.2.987-995.2003

288 Camarero, S. *et al.* Engineering platforms for directed evolution of Laccase from Pycnoporus cinnabarinus. *Appl Environ Microbiol* **78**, 1370-1384 (2012). https://doi.org/10.1128/AEM.07530-11

289 Mate, D. M., Gonzalez-Perez, D., Kittl, R., Ludwig, R. & Alcalde, M. Functional expression of a blood tolerant laccase in *Pichia pastoris*. *BMC Biotechnol* **13** (2013).

290 Molina-Espeja, P. *et al.* Directed evolution of unspecific peroxygenase from Agrocybe aegerita. *Appl Environ Microbiol* **80**, 3496-3507 (2014). https://doi.org/10.1128/AEM.00490-14

291 Vina-Gonzalez, J., Elbl, K., Ponte, X., Valero, F. & Alcalde, M. Functional expression of aryl-alcohol oxidase in Saccharomyces cerevisiae and Pichia pastoris by directed evolution. *Biotechnol Bioeng* **115**, 1666-1674 (2018). https://doi.org/10.1002/bit.26585

292 Gomez de Santos, P. *et al.* Functional Expression of Two Unusual Acidic Peroxygenases from Candolleomyces aberdarensis in Yeasts by Adopting Evolved Secretion Mutations. *Appl Environ Microbiol* **87**, e0087821 (2021). https://doi.org/10.1128/AEM.00878-21

293 Molina-Espeja, P., Ma, S., Mate, D. M., Ludwig, R. & Alcalde, M. Tandem-yeast expression system for engineering and producing unspecific peroxygenase. *Enzyme Microb Technol* **73-74**, 29-33 (2015). https://doi.org/10.1016/j.enzmictec.2015.03.004

294 Kumar, S., Gaddala, R., Thomas, S., Schumacher, J. & Schönherr, H. Green synthesis of polymer materials via enzyme- initiated RAFT polymerization. *Polymer Chemistry* **15**, 2011-2027 (2024). https://doi.org/10.1039/d4py00294f

295 Shannon, M. R., Zhou, B. & Perriman, A. W. Leveraging the Power of Enzymes in Engineered Dead and Living Materials. *Advanced Functional Materials* **34** (2024). https://doi.org/10.1002/adfm.202404522

296 Uyama, H., Kurioka, H. & Kobayashi, S. Novel Bienzymatic Catalysis System for Oxidative Polymerization of Phenols. *Polymer* **29**, 190-192 (1997).

297 Uyama, H. Artificial polymeric flavonoids: synthesis and applications. *Macromol Biosci* **7**, 410-422 (2007). https://doi.org/10.1002/mabi.200700005

298 Xie, W., Zhao, L., Wei, Y. & Yuan, J. Advances in enzyme-catalysis-mediated RAFT polymerization. *Cell Reports Physical Science* **2** (2021). https://doi.org/10.1016/j.xcrp.2021.100487

299 Liu, S., Ma, N., Kong, J. & Zhang, X. Enzyme-mediated controlled polymerization and its application in biomolecular analysis. *Microchemical Journal* **208** (2025). https://doi.org/10.1016/j.microc.2024.112581

300 Li, R., Kong, W. & An, Z. Enzyme Catalysis for Reversible Deactivation Radical Polymerization. *Angew Chem Int Ed Engl* **61**, e202202033 (2022). https://doi.org/10.1002/anie.202202033
301 Parravano, G. Chain Reactions Induced by Enzymic Systems. *JACS* **73** (1951).
302 Kobayashi, S. & Makino, A. Enzymatic Polymer Synthesis: An Opportunity for Green Polymer Chemistry. *Chem. Rev.* **109**, 5288–5353 (2009).
303 Singh, A., Ma, D. & Kaplan, D. L. Enzyme-Mediated Free Radical Polymerization of Styrene. *Biomolecules* **1**, 592-596 (2000).
304 Ikeda, R., Tanaka, H., Uyama, H. & Kobayashi, S. Laccase-catalyzed polymerization of acrylamide. *Macromolecular Rapid Communications* **19**, 423-425 (1998). https://doi.org/10.1002/(sici)1521-3927(19980801)19:8<423::Aid-marc423>3.0.Co;2-k
305 Emery, O., Lalot, T., Brigodiot, M. & Marechal, E. Free-Radical polymerization of acrylamide by horseradish peroxidase-mediated initiation. *Journal of Polymer Science Part A: Polymer Chemistry* **35**, 3331-3333 (1997). https://doi.org/10.1002/(sici)1099-0518(19971115)35:15<3331::Aid-pola27>3.0.Co;2-b
306 Xu, Q. *et al.* Enzyme catalysis-induced RAFT polymerization in water for the preparation of epoxy-functionalized triblock copolymer vesicles. *Polymer Chemistry* **9**, 4908-4916 (2018). https://doi.org/10.1039/c8py01053f
307 Tan, J. *et al.* Enzyme-PISA: An Efficient Method for Preparing Well-Defined Polymer Nano-Objects under Mild Conditions. *Macromol Rapid Commun* **39**, e1700871 (2018). https://doi.org/10.1002/marc.201700871
308 Zhang, B. *et al.* Enzyme-Initiated Reversible Addition–Fragmentation Chain Transfer Polymerization. *Macromolecules* **48**, 7792-7802 (2015). https://doi.org/10.1021/acs.macromol.5b01893
309 Zhou, F., Li, R., Wang, X., Du, S. & An, Z. Non-natural Photoenzymatic Controlled Radical Polymerization Inspired by DNA Photolyase. *Angew Chem Int Ed Engl* **58**, 9479-9484 (2019). https://doi.org/10.1002/anie.201904413
310 Silva, T. B. *et al.* Hemoglobin and red blood cells catalyze atom transfer radical polymerization. *Biomacromolecules* **14**, 2703-2712 (2013). https://doi.org/10.1021/bm400556x
311 Sun, Y., Zhang, J., Li, J., Zhao, M. & Liu, Y. Preparation of protein imprinted polymers via protein-catalyzed eATRP on 3D gold nanodendrites and their application in biosensors. *RSC Advances* **7**, 28461-28468 (2017). https://doi.org/10.1039/c7ra03772d
312 Hu, Q. *et al.* Coenzyme-Mediated Electro-RAFT Polymerization for Amplified Electrochemical Interrogation of Trypsin Activity. *Anal Chem* **93**, 9602-9608 (2021). https://doi.org/10.1021/acs.analchem.1c01766
313 Hollmann, F. & Arends, I. W. C. E. Enzyme Initiated Radical Polymerizations. *Polymers* **4**, 759-793 (2012). https://doi.org/10.3390/polym4010759
314 Berglund, G. I. *et al.* The catalytic pathway of horseradish peroxidase at high resolution. *Nature* **417**, 463-468 (2002).
315 Valles, M., Kamaruddin, A. F., Wong, L. S. & Blanford, C. F. Inhibition in multicopper oxidases: a critical review. *Catalysis Science & Technology* **10**, 5386-5410 (2020). https://doi.org/10.1039/d0cy00724b
316 Zavada, S. R., Battsengel, T. & Scott, T. F. Radical-Mediated Enzymatic Polymerizations. *Int J Mol Sci* **17** (2016). https://doi.org/10.3390/ijms17020195
317 Sigg, S. J. *et al.* Horseradish peroxidase as a catalyst for atom transfer radical polymerization. *Macromol Rapid Commun* **32**, 1710-1715 (2011). https://doi.org/10.1002/marc.201100349

318 Jin, Z., Su, Y. & Duan, Y. A novel method for polyaniline synthesis with the immobilized horseradish peroxidase enzyme. *Synthetic Metals* **122**, 237-242 (2001).
319 Wang, P., Martin, B. D., Parida, S., Rethwisch, D. G. & Dordick, J. S. Multienzymic Synthesis of Poly(hydroquinone) for Use as a Redox Polymer. *JACS* **117**, 12885-12886 (1995).
320 Kim, Y., An, E., Song, B., Kim, D. & Chelikani, R. Polymerization of cardanol using soybean peroxidase and its potential application as anti-biofilm coating material. *Biotechnology Letters* **25**, 1521-1524 (2003).
321 Park, S. Y., Kim, Y. H., Won, K. & Song, B. K. Enzymatic synthesis and curing of polycardol from renewable resources. *Journal of Molecular Catalysis B: Enzymatic* **57**, 312-316 (2009). https://doi.org/10.1016/j.molcatb.2008.07.001
322 Longoria, A. M., Hu, H. & Vazquez-Duhalt, R. Enzymatic synthesis of semiconductor polymers by chloroperoxidase of Caldariomyces fumago. *Appl Biochem Biotechnol* **162**, 927-934 (2010). https://doi.org/10.1007/s12010-009-8805-7
323 Román, P., Cruz-Silva, R. & Vazquez-Duhalt, R. Peroxidase-mediated synthesis of water-soluble fully sulfonated polyaniline. *Synthetic Metals* **162**, 794-799 (2012). https://doi.org/10.1016/j.synthmet.2012.02.019
324 Di Gennaro, P., Bargna, A., Bruno, F. & Sello, G. Purification of recombinant catalase-peroxidase HPI from E. coli and its application in enzymatic polymerization reactions. *Appl Microbiol Biotechnol* **98**, 1119-1126 (2014). https://doi.org/10.1007/s00253-013-4948-0
325 Lin, H. *et al.* Application of Laccase Catalysis in Bond Formation and Breakage: A Review. *Catalysts* **13** (2023). https://doi.org/10.3390/catal13040750
326 Mahuri, M., Paul, M. & Thatoi, H. A review of microbial laccase production and activity toward different biotechnological applications. *Systems Microbiology and Biomanufacturing* **3**, 533-551 (2023). https://doi.org/10.1007/s43393-023-00163-6
327 Fodor, C. *et al.* Laccase-catalyzed controlled radical polymerization of N-vinylimidazole. *Polymer Chemistry* **7**, 6617-6625 (2016). https://doi.org/10.1039/c6py01261b
328 Ikeda, R. *et al.* Man-made urushi: Preparation of crosslinked polymeric films from renewable resources via air-oxidation processes. *Proceedings of the Japan Academy, Series B* **76**, 155-160 (2000). https://doi.org/10.2183/pjab.76.155
329 Aizawa, M., Wang, L., Shinohara, H. & Ikariyama, Y. Enzymatic synthesis of polyaniline film using a copper-containing oxidoreductase: bilirubin oxidase. *J Biotechnol* **14**, 301-310 (1990).
330 Khlupova, M. E. *et al.* Enzymatic polymerization of dihydroquercetin using bilirubin oxidase. *Biochemistry (Mosc)* **80**, 233-241 (2015). https://doi.org/10.1134/S0006297915020108
331 Wang, L., Kobatake, E., Ikariyama, Y. & Aizawa, M. Regioselective oxidative polymerization of 1,5-dihydroxynaphthalene catalyzed by bilirubin oxidase in a water–organic solvent mixed solution. *Journal of Polymer Science Part A: Polymer Chemistry* **31**, 2855-2861 (1993). https://doi.org/https://doi.org/10.1002/pola.1993.080311123
332 George, P. The Third Intermediate Compound of Horseradish Peroxidase and Hydrogen Peroxide. *Journal of Biological Chemistry* **201**, 427-434 (1953). https://doi.org/10.1016/s0021-9258(18)71385-9
333 Chapman, R., Gormley, A. J., Herpoldt, K.-L. & Stevens, M. M. Highly Controlled Open Vessel RAFT Polymerizations by Enzyme Degassing. *Macromolecules* **47**, 8541-8547 (2014). https://doi.org/10.1021/ma5021209

334 Chapman, R., Gormley, A. J., Stenzel, M. H. & Stevens, M. M. Combinatorial Low-Volume Synthesis of Well-Defined Polymers by Enzyme Degassing. *Angew Chem Int Ed Engl* **55**, 4500-4503 (2016). https://doi.org/10.1002/anie.201600112
335 Iwata, H., Hata, Y., Matsuda, T. & Ikada, Y. Initiation of radical polymerization by glucose oxidase utilizing dissolved oxygen. *Journal of Polymer Science Part A: Polymer Chemistry* **29**, 1217-1218 (1991). https://doi.org/https://doi.org/10.1002/pola.1991.080290818
336 Lv, Y., Liu, Z., Zhu, A. & An, Z. Glucose oxidase deoxygenation−redox initiation for RAFT polymerization in air. *Journal of Polymer Science Part A: Polymer Chemistry* **55**, 164-174 (2016). https://doi.org/10.1002/pola.28380
337 Gormley, A. J., Chapman, R. & Stevens, M. M. Polymerization amplified detection for nanoparticle-based biosensing. *Nano Lett* **14**, 6368-6373 (2014). https://doi.org/10.1021/nl502840h
338 Schneiderman, D. K. *et al.* Open-to-Air RAFT Polymerization in Complex Solvents: From Whisky to Fermentation Broth. *ACS Macro Lett* **7**, 406-411 (2018). https://doi.org/10.1021/acsmacrolett.8b00069
339 Liu, Z., Lv, Y. & An, Z. Enzymatic Cascade Catalysis for the Synthesis of Multiblock and Ultrahigh-Molecular-Weight Polymers with Oxygen Tolerance. *Angew Chem Int Ed Engl* **56**, 13852-13856 (2017). https://doi.org/10.1002/anie.201707993
340 Derango, R. A., Chiang, L.-c., Dowbenko, R. & Lasch, J. G. ENZYME-MEDIATED POLYMERIZATION OF ACRYLIC MONOMERS. *Biotehcnology Techniques* **6**, 523-526 (1992).
341 Tsujimoto, T., Uyama, H. & Kobayashi, S. Polymerization of Vinyl Monomers Using Oxidase Catalysts. *Macromolecular Bioscience* **1**, 228-232 (2001). https://doi.org/10.1002/1616-5195(20010801)1:6<228::Aid-mabi228>3.0.Co;2-s
342 Li, R., Zhang, S., Li, Q., Qiao, G. G. & An, Z. An Atom-Economic Enzymatic Cascade Catalysis for High-Throughput RAFT Synthesis of Ultrahigh Molecular Weight Polymers. *Angew Chem Int Ed Engl* **61**, e202213396 (2022). https://doi.org/10.1002/anie.202213396
343 Hoffman, A. S. Hydrogels for biomedical applications. *Advanced Drug Delivery Reviews* **64**, 18-23 (2012). https://doi.org/10.1016/j.addr.2012.09.010
344 Raghuwanshi, V. S. & Garnier, G. Characterisation of hydrogels: Linking the nano to the microscale. *Adv Colloid Interface Sci* **274**, 102044 (2019). https://doi.org/10.1016/j.cis.2019.102044
345 Ju, H. K., Kim, S. Y., Kim, S. J. & Lee, Y. M. pH/temperature-responsive semi-IPN hydrogels composed of alginate and poly(N-isopropylacrylamide). *Journal of Applied Polymer Science* **83**, 1128-1139 (2001). https://doi.org/10.1002/app.10137
346 de Moura, M. R. *et al.* Thermo-sensitive IPN hydrogels composed of PNIPAAm gels supported on alginate-Ca2+ with LCST tailored close to human body temperature. *Polymer Testing* **25**, 961-969 (2006). https://doi.org/10.1016/j.polymertesting.2006.06.002
347 Xing, Y., Zeng, B. & Yang, W. Light responsive hydrogels for controlled drug delivery. *Front Bioeng Biotechnol* **10**, 1075670 (2022). https://doi.org/10.3389/fbioe.2022.1075670
348 Markland, P., Zhang, Y., Amidon, G. L. & Yang, V. C. A pH- and ionic strength-responsive polypeptide hydrogel: Synthesis, characterization, and preliminary protein release studies. *Journal of Biomedical Materials Research* **47**, 595-602 (1999). https://doi.org/10.1002/(sici)1097-4636(19991215)47:4<595::Aid-jbm17>3.0.Co;2-i
349 Ahmed, E. M. Hydrogel: Preparation, characterization, and applications: A review. *J Adv Res* **6**, 105-121 (2015). https://doi.org/10.1016/j.jare.2013.07.006

350 WICHTERLE, O. & LÍM, D. Hydrophilic Gels for Biological Use. *Nature* **185**, 117-118 (1960).
351 Buwalda, S. J. *et al.* Hydrogels in a historical perspective: from simple networks to smart materials. *J Control Release* **190**, 254-273 (2014). https://doi.org/10.1016/j.jconrel.2014.03.052
352 Koehler, J., Brandl, F. P. & Goepferich, A. M. Hydrogel wound dressings for bioactive treatment of acute and chronic wounds. *European Polymer Journal* **100**, 1-11 (2018). https://doi.org/10.1016/j.eurpolymj.2017.12.046
353 Liang, Y., He, J. & Guo, B. Functional Hydrogels as Wound Dressing to Enhance Wound Healing. *ACS Nano* **15**, 12687-12722 (2021). https://doi.org/10.1021/acsnano.1c04206
354 Thang, N. H., Chien, T. B. & Cuong, D. X. Polymer-Based Hydrogels Applied in Drug Delivery: An Overview. *Gels* **9** (2023). https://doi.org/10.3390/gels9070523
355 Narayanaswamy, R. & Torchilin, V. P. Hydrogels and Their Applications in Targeted Drug Delivery. *Molecules* **24** (2019). https://doi.org/10.3390/molecules24030603
356 He, W., Reaume, M., Hennenfent, M., Lee, B. P. & Rajachar, R. Biomimetic hydrogels with spatial- and temporal-controlled chemical cues for tissue engineering. *Biomater Sci* **8**, 3248-3269 (2020). https://doi.org/10.1039/d0bm00263a
357 Zhao, H., Liu, M., Zhang, Y., Yin, J. & Pei, R. Nanocomposite hydrogels for tissue engineering applications. *Nanoscale* **12**, 14976-14995 (2020). https://doi.org/10.1039/d0nr03785k
358 Zhou, Y. *et al.* Hydrogel smart windows. *Journal of Materials Chemistry A* **8**, 10007-10025 (2020). https://doi.org/10.1039/d0ta00849d
359 Liu, X., Inda, M. E., Lai, Y., Lu, T. K. & Zhao, X. Engineered Living Hydrogels. *Adv Mater* **34**, e2201326 (2022). https://doi.org/10.1002/adma.202201326
360 Lee, Y., Song, W. J. & Sun, J. Y. Hydrogel soft robotics. *Materials Today Physics* **15** (2020). https://doi.org/10.1016/j.mtphys.2020.100258
361 Klemperer, R. G., Shannon, M. R., Ross Anderson, J. L. & Perriman, A. W. Bienzymatic Generation of Interpenetrating Polymer Networked Engineered Living Materials with Shape Changing Properties. *Advanced Materials Technologies* **8** (2023). https://doi.org/10.1002/admt.202300626
362 Nakahata, M., Gantumur, E., Furuno, K., Sakai, S. & Taya, M. Versatility of hydrogelation by dual-enzymatic reactions with oxidases and peroxidase. *Biochemical Engineering Journal* **131**, 1-8 (2018). https://doi.org/10.1016/j.bej.2017.12.003
363 Chen, Q., Chen, H., Zhu, L. & Zheng, J. Fundamentals of double network hydrogels. *J Mater Chem B* **3**, 3654-3676 (2015). https://doi.org/10.1039/c5tb00123d
364 Zhang, W. *et al.* Double-network hydrogels for biomaterials: Structure-property relationships and drug delivery. *European Polymer Journal* **185** (2023). https://doi.org/10.1016/j.eurpolymj.2022.111807
365 Yan, X. *et al.* Advances in enhancing the mechanical properties of biopolymer hydrogels via multi-strategic approaches. *Int J Biol Macromol* **272**, 132583 (2024). https://doi.org/10.1016/j.ijbiomac.2024.132583
366 Gong, J. P., Katsuyama, Y., Kurokawa, T. & Osada, Y. Double-Network Hydrogels with Extremely High Mechanical Strength. *Advanced Materials* **15**, 1155-1158 (2003). https://doi.org/10.1002/adma.200304907
367 Sun, J. Y. *et al.* Highly stretchable and tough hydrogels. *Nature* **489**, 133-136 (2012). https://doi.org/10.1038/nature11409
368 Brown, H. R. A Model of the Fracture of Double Network Gels. *Macromolecules* **40**, 3815-3818 (2007).

369 Tanaka, Y. A local damage model for anomalous high toughness of double-network gels. *Europhysics Letters (EPL)* **78** (2007). https://doi.org/10.1209/0295-5075/78/56005
370 Darnell, M. C. *et al.* Performance and biocompatibility of extremely tough alginate/polyacrylamide hydrogels. *Biomaterials* **34**, 8042-8048 (2013). https://doi.org/10.1016/j.biomaterials.2013.06.061
371 Xiong, Q., Zhang, X., Wei, W., Wei, G. & Su, Z. Enzyme-mediated reversible deactivation radical polymerization for functional materials: principles, synthesis, and applications. *Polymer Chemistry* **11**, 1673-1690 (2020). https://doi.org/10.1039/d0py00136h
372 Yuan, M., Cui, X., Zhu, W. & Tang, H. Development of Environmentally Friendly Atom Transfer Radical Polymerization. *Polymers (Basel)* **12** (2020). https://doi.org/10.3390/polym12091987
373 Heinrich, M. A. *et al.* 3D Bioprinting: from Benches to Translational Applications. *Small* **15**, e1805510 (2019). https://doi.org/10.1002/smll.201805510
374 Cui, X. *et al.* Advances in Extrusion 3D Bioprinting: A Focus on Multicomponent Hydrogel-Based Bioinks. *Adv Healthc Mater* **9**, e1901648 (2020). https://doi.org/10.1002/adhm.201901648
375 Bom, S., Ribeiro, R., Ribeiro, H. M., Santos, C. & Marto, J. On the progress of hydrogel-based 3D printing: Correlating rheological properties with printing behaviour. *Int J Pharm* **615**, 121506 (2022). https://doi.org/10.1016/j.ijpharm.2022.121506
376 Zaszczynska, A., Moczulska-Heljak, M., Gradys, A. & Sajkiewicz, P. Advances in 3D Printing for Tissue Engineering. *Materials (Basel)* **14** (2021). https://doi.org/10.3390/ma14123149
377 Jammalamadaka, U. & Tappa, K. Recent Advances in Biomaterials for 3D Printing and Tissue Engineering. *J Funct Biomater* **9** (2018). https://doi.org/10.3390/jfb9010022
378 Gul, J. Z. *et al.* 3D printing for soft robotics - a review. *Sci Technol Adv Mater* **19**, 243-262 (2018). https://doi.org/10.1080/14686996.2018.1431862
379 Gillispie, G. *et al.* Assessment methodologies for extrusion-based bioink printability. *Biofabrication* **12**, 022003 (2020). https://doi.org/10.1088/1758-5090/ab6f0d
380 Gilbert, C. & Ellis, T. Biological Engineered Living Materials: Growing Functional Materials with Genetically Programmable Properties. *ACS Synth Biol* **8**, 1-15 (2019). https://doi.org/10.1021/acssynbio.8b00423
381 Srubar, W. V., 3rd. Engineered Living Materials: Taxonomies and Emerging Trends. *Trends Biotechnol* **39**, 574-583 (2021). https://doi.org/10.1016/j.tibtech.2020.10.009
382 An, B. *et al.* Engineered Living Materials For Sustainability. *Chem Rev* **123**, 2349-2419 (2023). https://doi.org/10.1021/acs.chemrev.2c00512
383 Manjula-Basavanna, A., Duraj-Thatte, A. M. & Joshi, N. S. Mechanically Tunable, Compostable, Healable and Scalable Engineered Living Materials. *Nat Commun* **15**, 9179 (2024). https://doi.org/10.1038/s41467-024-53052-4
384 Gilbert, C. *et al.* Living materials with programmable functionalities grown from engineered microbial co-cultures. *Nat Mater* **20**, 691-700 (2021). https://doi.org/10.1038/s41563-020-00857-5
385 Nguyen, P. Q., Courchesne, N. D., Duraj-Thatte, A., Praveschotinunt, P. & Joshi, N. S. Engineered Living Materials: Prospects and Challenges for Using Biological Systems to Direct the Assembly of Smart Materials. *Adv Mater* **30**, e1704847 (2018). https://doi.org/10.1002/adma.201704847

386 Lu, C. *et al.* Toward Practical Applications of Engineered Living Materials with Advanced Fabrication Techniques. *ACS Synth Biol* **13**, 2295-2312 (2024). https://doi.org/10.1021/acssynbio.4c00259
387 Adolphe, J., Loubière, J., Paradas, J. & Soleilhavoup, F. Procédé de traitement biologique d'une surface artificielle. *European patent 90400G97. 0* **1989** (1990).
388 Tay, P. K. R., Nguyen, P. Q. & Joshi, N. S. A Synthetic Circuit for Mercury Bioremediation Using Self-Assembling Functional Amyloids. *ACS Synth Biol* **6**, 1841-1850 (2017). https://doi.org/10.1021/acssynbio.7b00137
389 An, B. *et al.* Programming Living Glue Systems to Perform Autonomous Mechanical Repairs. *Matter* **3**, 2080-2092 (2020). https://doi.org/10.1016/j.matt.2020.09.006
390 Xu, Z. *et al.* Photosynthetic hydrogen production by droplet-based microbial micro-reactors under aerobic conditions. *Nat Commun* **11**, 5985 (2020). https://doi.org/10.1038/s41467-020-19823-5
391 Datta, D. *et al.* Phenotypically complex living materials containing engineered cyanobacteria. *Nat Commun* **14**, 4742 (2023). https://doi.org/10.1038/s41467-023-40265-2
392 Brooks, S. M. & Alper, H. S. Applications, challenges, and needs for employing synthetic biology beyond the lab. *Nat Commun* **12**, 1390 (2021). https://doi.org/10.1038/s41467-021-21740-0
393 Johnston, T. G. *et al.* Compartmentalized microbes and co-cultures in hydrogels for on-demand bioproduction and preservation. *Nat Commun* **11**, 563 (2020). https://doi.org/10.1038/s41467-020-14371-4
394 Brooks, S. M. *et al.* Enhancing long-term storage and stability of engineered living materials through desiccant storage and trehalose treatment. *Biotechnol Bioeng* **120**, 572-582 (2023). https://doi.org/10.1002/bit.28271
395 Herkommerova, K., Zemancikova, J., Sychrova, H. & Antosova, Z. Immobilization in polyvinyl alcohol hydrogel enhances yeast storage stability and reusability of recombinant laccase-producing S. cerevisiae. *Biotechnol Lett* **40**, 405-411 (2018). https://doi.org/10.1007/s10529-017-2485-0
396 Vieira Gomes, A. M., Souza Carmo, T., Silva Carvalho, L., Mendonca Bahia, F. & Parachin, N. S. Comparison of Yeasts as Hosts for Recombinant Protein Production. *Microorganisms* **6** (2018). https://doi.org/10.3390/microorganisms6020038
397 Nielsen, J. Yeast Systems Biology: Model Organism and Cell Factory. *Biotechnol J* **14**, e1800421 (2019). https://doi.org/10.1002/biot.201800421
398 Thak, E. J., Yoo, S. J., Moon, H. Y. & Kang, H. A. Yeast synthetic biology for designed cell factories producing secretory recombinant proteins. *FEMS Yeast Res* **20** (2020). https://doi.org/10.1093/femsyr/foaa009
399 Yuan, S. F. *et al.* Bioproduced Proteins On Demand (Bio-POD) in hydrogels using Pichia pastoris. *Bioact Mater* **6**, 2390-2399 (2021). https://doi.org/10.1016/j.bioactmat.2021.01.019
400 Wang, H. *et al.* Genetical Surface Display of Silicatein on Yarrowia lipolytica Confers Living and Renewable Biosilica-Yeast Hybrid Materials. *ACS Omega* **5**, 7555-7566 (2020). https://doi.org/10.1021/acsomega.0c00393
401 Yuan, Y. *et al.* Engineering Living Material for Controlled Fragrance Release Utilizing Kluyveromyces marxianus CBS6556 and Adaptive Hydrogel. *ACS Synth Biol* **13**, 3188-3196 (2024). https://doi.org/10.1021/acssynbio.4c00229
402 Fine, T. *et al.* Luminescent yeast cells entrapped in hydrogels for estrogenic endocrine disrupting chemical biodetection. *Biosens Bioelectron* **21**, 2263-2269 (2006). https://doi.org/10.1016/j.bios.2005.11.004

403 Sugianto, W. *et al.* Gene expression dynamics in input-responsive engineered living materials programmed for bioproduction. *Mater Today Bio* **20**, 100677 (2023). https://doi.org/10.1016/j.mtbio.2023.100677
404 Altin-Yavuzarslan, G. *et al.* Additive Manufacturing of Engineered Living Materials with Bio-augmented Mechanical Properties and Resistance to Degradation. *Adv Funct Mater* **33** (2023). https://doi.org/10.1002/adfm.202300332
405 Rondon, R. & Wilson, C. J. Engineering Alternate Ligand Recognition in the PurR Topology: A System of Novel Caffeine Biosensing Transcriptional Antirepressors. *ACS Synth Biol* **10**, 552-565 (2021). https://doi.org/10.1021/acssynbio.0c00582
406 Wu, J. *et al.* Design and application of a lactulose biosensor. *Sci Rep* **7**, 45994 (2017). https://doi.org/10.1038/srep45994
407 Zheng, C. R., Singh, A., Libby, A., Silver, P. A. & Libby, E. A. Modular and Single-Cell Sensors of Bacterial Ser/Thr Kinase Activity. *ACS Synth Biol* **10**, 2340-2350 (2021). https://doi.org/10.1021/acssynbio.1c00250
408 Sakaguchi-Mikami, A., Taniguchi, A., Sode, K. & Yamazaki, T. Construction of a novel glucose-sensing molecule based on a substrate-binding protein for intracellular sensing. *Biotechnol Bioeng* **108**, 725-733 (2011). https://doi.org/10.1002/bit.23006
409 Lee, H., Rha, E. & Kim, H. Development and optimization of a modular two-fragment LacI switch for enhanced biosensor applications. *Biotechnology and Bioprocess Engineering* **29**, 109-117 (2024). https://doi.org/10.1007/s12257-024-00020-w
410 Studier, F. W. & Moffatt, B. A. Use of Bacteriophage T7 RNA Polymerase to Direct Selective High-level Expression of Cloned Genes. *J. Mol. Biol.* **189**, 113-130 (1986).
411 Studier, F. W. Use of Bacteriophage T7 Lysozyme to Improve an Inducible T7 Expression System. *J. Mol. Biol.* **219**, 37-44 (1991).
412 Tabor, S. Expression using the T7 RNA polymerase/promoter system. *Curr Protoc Mol Biol* **Chapter 16**, Unit16 12 (2001). https://doi.org/10.1002/0471142727.mb1602s11
413 Lalwani, M. A. *et al.* Optogenetic control of the lac operon for bacterial chemical and protein production. *Nat Chem Biol* **17**, 71-79 (2021). https://doi.org/10.1038/s41589-020-0639-1
414 Zhang, Q. *et al.* De novo 2'-fucosyllactose biosynthesis using glucose as the sole carbon source by multiple engineered Bacillus subtilis. *Metab Eng* **88**, 85-93 (2024). https://doi.org/10.1016/j.ymben.2024.12.004
415 Bohnenkamp, A. C. *et al.* Multilevel optimisation of anaerobic ethyl acetate production in engineered Escherichia coli. *Biotechnol Biofuels* **13**, 65 (2020). https://doi.org/10.1186/s13068-020-01703-1
416 Xu, W. *et al.* High-Level Production of Tyrosol with Noninduced Recombinant Escherichia coli by Metabolic Engineering. *J Agric Food Chem* **68**, 4616-4623 (2020). https://doi.org/10.1021/acs.jafc.9b07610
417 Deng, C. *et al.* Refactoring transcription factors for metabolic engineering. *Biotechnol Adv* **57**, 107935 (2022). https://doi.org/10.1016/j.biotechadv.2022.107935
418 Groseclose, T. M., Hersey, A. N., Huang, B. D., Realff, M. J. & Wilson, C. J. Biological signal processing filters via engineering allosteric transcription factors. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org/10.1073/pnas.2111450118
419 Dimas, R. P., Jiang, X. L., Alberto de la Paz, J., Morcos, F. & Chan, C. T. Y. Engineering repressors with coevolutionary cues facilitates toggle switches with a master reset. *Nucleic Acids Res* **47**, 5449-5463 (2019). https://doi.org/10.1093/nar/gkz280

420 Groseclose, T. M., Rondon, R. E., Herde, Z. D., Aldrete, C. A. & Wilson, C. J. Engineered systems of inducible anti-repressors for the next generation of biological programming. *Nat Commun* **11**, 4440 (2020). https://doi.org/10.1038/s41467-020-18302-1
421 Cheng, Y. Y. *et al.* Programming bacteria for multiplexed DNA detection. *Nat Commun* **14**, 2001 (2023). https://doi.org/10.1038/s41467-023-37582-x
422 Kuriyan, J., Konforti, B. & Wemmer, D. *The Molecules of Life: Physical and Chemical Principles*. (2012).
423 *A Handbook of Transcription Factors*. (Springer Dordrecht, 2011).
424 JACOB, F. & MONOD, J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J Mol Biol* **3**, 318-356 (1961).
425 Varma, A. & Palsson, B. Metabolic Capabilities of *Escherichia coli:* I Synthesis of Biosynthetic Precursors and Cofactors. *J Theor Biol* **165**, 477-502 (1993).
426 Varma, A. & Palsson, B. Metabolic Capabilities of *Escherichia coli*: II Optimal Growth Patterns. *J Theor Biol* **165**, 503-522 (1993).
427 Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism. *Mol Syst Biol* **7**, 535 (2011). https://doi.org/10.1038/msb.2011.65
428 Swint-Kruse, L. & Matthews, K. S. Allostery in the LacI/GalR family: variations on a theme. *Curr Opin Microbiol* **12**, 129-137 (2009). https://doi.org/10.1016/j.mib.2009.01.009
429 Bondos, S. E., Swint-Kruse, L. & Matthews, K. S. Flexibility and Disorder in Gene Regulation: LacI/GalR and Hox Proteins. *J Biol Chem* **290**, 24669-24677 (2015). https://doi.org/10.1074/jbc.R115.685032
430 Ravcheev, D. A. *et al.* Comparative genomics and evolution of regulons of the LacI-family transcription factors. *Front Microbiol* **5**, 294 (2014). https://doi.org/10.3389/fmicb.2014.00294
431 Chen, J., Vishweshwaraiah, Y. L. & Dokholyan, N. V. Design and engineering of allosteric communications in proteins. *Curr Opin Struct Biol* **73**, 102334 (2022). https://doi.org/10.1016/j.sbi.2022.102334
432 Muller-Hill, B. Lac repressor. *Angew Chem Int Ed Engl* **10**, 160-172 (1971). https://doi.org/10.1002/anie.197101601
433 Sato, H. & Shiio, I. Effects of Two End Products on Enzyme Repression in Purine Nucleotide Biosynthesis. *J Biochem* **68**, 763-773 (1970).
434 Nakanishi, S., Adhya, S., Gottesman, M. & Pastan, I. Studies on the Mechanism of Action of the Gal Repressor. *Journal of Biological Chemistry* **248**, 5937-5942 (1973). https://doi.org/10.1016/s0021-9258(19)43491-1
435 Henkin, T. M., Grundy, F. J., Nicholson, W. L. & Chambliss, G. H. Catabolite repression of alpha-amylase gene expression in Bacillus subtilis involves a trans-acting gene product homologous to the Escherichia coli lacI and galR repressors. *Mol Microbiol* **5**, 575-584 (1991). https://doi.org/10.1111/j.1365-2958.1991.tb00728.x
436 TOBISCH, S., ZUHLKE, D., BERNHARDT, J., STULKE, J. & HECKER, M. Role of CcpA in Regulation of the Central Pathways of Carbon Catabolism in *Bacillus subtilis*. *Journal of Bacteriology* **181**, 6996-7004 (1999).
437 Mauzy, C. A. & Hermodson, M. A. Structural and functional analyses of the repressor, RbsR, of the ribose operon of Escherichia coli. *Protein Sci* **1**, 831-842 (1992). https://doi.org/10.1002/pro.5560010701
438 Weickert, M. J. & Adhyat, S. Isorepressor of the *gal* Regulon in *Escherichia coli*. *J Mol Biol* **226**, 69-83 (1992).
439 Ramseier, T. M. Cra and the control of carbon flux via metabolic pathways. *Forum in Microbiology* **14**, 489-493 (1996).

440 SAIER, M. H. & RAMSEIER, T. M. The Catabolite Repressor/Activator (Cra) Protein of Enteric Bacteria. *Journal of Bacteriology* **178**, 3411-3417 (1996).
441 Spisak, S. & Ostermeier, M. Engineered protein switches for exogenous control of gene expression. *Biochem Soc Trans* **48**, 2205-2212 (2020). https://doi.org/10.1042/BST20200441
442 Cortay, J. C. *et al.* In vitro asymmetric binding of the pleiotropic regulatory protein, FruR, to the ace operator controlling glyoxylate shunt enzyme synthesis. *Journal of Biological Chemistry* **269**, 14885-14891 (1994). https://doi.org/10.1016/s0021-9258(17)36548-1
443 Gosseringer, R., Kuster, E., Galinier, A., Deutscher, J. & Hillen, W. Cooperative and Non-cooperative DNA Binding Modes of Catabolite Control Protein CcpA from Bacillus megaterium Result from Sensing Two Different Signals. *J Mol Biol* **266**, 665-676 (1997).
444 Geanacopoulos, M. & Adhya, S. Genetic analysis of GalR tetramerization in DNA looping during repressosome assembly. *J Biol Chem* **277**, 33148-33152 (2002). https://doi.org/10.1074/jbc.M202445200
445 Mitra, S. *et al.* Flexible Target Recognition of the Intrinsically Disordered DNA-Binding Domain of CytR Monitored by Single-Molecule Fluorescence Spectroscopy. *J Phys Chem B* **126**, 6136-6147 (2022). https://doi.org/10.1021/acs.jpcb.2c02791
446 Nguyen, C. C. & Jr., M. H. S. Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS* **377** (1995).
447 Santos, C. L., Tavares, F., Thioulouse, J. & Normand, P. A phylogenomic analysis of bacterial helix-turn-helix transcription factors. *FEMS Microbiol Rev* **33**, 411-429 (2009). https://doi.org/10.1111/j.1574-6976.2008.00154.x
448 Rosinski, J. A. & Atchley, W. R. Molecular Evolution of Helix–Turn–Helix Proteins. *J Mol Evol* **49**, 301-309 (1999).
449 Meinhardt, S. *et al.* Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression. *Nucleic Acids Res* **40**, 11139-11154 (2012). https://doi.org/10.1093/nar/gks806
450 Groseclose, T. M. *et al.* Biomolecular Systems Engineering: Unlocking the Potential of Engineered Allostery via the Lactose Repressor Topology. *Annu Rev Biophys* **50**, 303-321 (2021). https://doi.org/10.1146/annurev-biophys-090820-101708
451 Yu, T. C. *et al.* Multiplexed characterization of rationally designed promoter architectures deconstructs combinatorial logic for IPTG-inducible systems. *Nat Commun* **12**, 325 (2021). https://doi.org/10.1038/s41467-020-20094-3
452 Daber, R., Stayrook, S., Rosenberg, A. & Lewis, M. Structural analysis of lac repressor bound to allosteric effectors. *J Mol Biol* **370**, 609-619 (2007). https://doi.org/10.1016/j.jmb.2007.04.028
453 Bell, C. E. & Lewis, M. A closer view of the conformation of the Lac repressor bound to operator. *Nature Structural Biology* **7** (2000).
454 Lewis, M. *et al.* Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science* **271** (1996).
455 Chauvaux, S. CcpA and HPr(ser-P): mediators of catabolite repression in Bacillus subtilis. *Forum in Microbiology* **14** (1996).
456 Schumacher, M. A. *et al.* Structural basis for allosteric control of the transcription regulator CcpA by the phosphoprotein HPr-Ser46-P. *Cell* **118**, 731-741 (2004). https://doi.org/10.1016/j.cell.2004.08.027
457 Schumacher, M. A., Seidel, G., Hillen, W. & Brennan, R. G. Structural mechanism for the fine-tuning of CcpA function by the small molecule effectors glucose 6-phosphate

and fructose 1,6-bisphosphate. *J Mol Biol* **368**, 1042-1050 (2007). https://doi.org/10.1016/j.jmb.2007.02.054
458 Daber, R., Sharp, K. & Lewis, M. One is not enough. *J Mol Biol* **392**, 1133-1144 (2009). https://doi.org/10.1016/j.jmb.2009.07.050
459 Tsai, C. J. & Nussinov, R. A unified view of "how allostery works". *PLoS Comput Biol* **10**, e1003394 (2014). https://doi.org/10.1371/journal.pcbi.1003394
460 Liu, J. & Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput Biol* **12**, e1004966 (2016). https://doi.org/10.1371/journal.pcbi.1004966
461 Segers, M., Voorspoels, A., Sakaue, T. & Carlon, E. Mechanisms of DNA-Mediated Allostery. *Phys Rev Lett* **131**, 238402 (2023). https://doi.org/10.1103/PhysRevLett.131.238402
462 Nishikawa, K. K., Hoppe, N., Smith, R., Bingman, C. & Raman, S. Epistasis shapes the fitness landscape of an allosteric specificity switch. *Nat Commun* **12**, 5562 (2021). https://doi.org/10.1038/s41467-021-25826-7
463 Wang, J. *et al.* Mapping allosteric communications within individual proteins. *Nat Commun* **11**, 3862 (2020). https://doi.org/10.1038/s41467-020-17618-2
464 Glasgow, A. *et al.* Ligand-specific changes in conformational flexibility mediate long-range allostery in the lac repressor. *Nat Commun* **14**, 1179 (2023). https://doi.org/10.1038/s41467-023-36798-1
465 Kariyawasam, N. L., Ploetz, E. A., Swint-Kruse, L. & Smith, P. E. Simulated pressure changes in LacI suggest a link between hydration and functional conformational changes. *Biophys Chem* **304**, 107126 (2024). https://doi.org/10.1016/j.bpc.2023.107126
466 Tungtur, S., Schwingen, K. M., Riepe, J. J., Weeramange, C. J. & Swint-Kruse, L. Homolog comparisons further reconcile in vitro and in vivo correlations of protein activities by revealing over-looked physiological factors. *Protein Sci* **28**, 1806-1818 (2019). https://doi.org/10.1002/pro.3695
467 Barkley, M. D., Riggs, A. D., Jobe, A. & Bourgeois, S. Interaction of Effecting Ligands with Lac Repressor and Repressor-Operator Complex. *Biochem J* **14**, 1700-1712 (1975).
468 Scheler, A. & Hillen, W. Glucose is an anti-inducer for the *Bacillus licheniformis* encoded Xyl repressor. *FEMS Microbiology Letters* **107**, 299-302 (1993).
469 Anderson, B. W. *et al.* The nucleotide messenger (p)ppGpp is an anti-inducer of the purine synthesis transcription regulator PurR in Bacillus. *Nucleic Acids Res* **50**, 847-866 (2022). https://doi.org/10.1093/nar/gkab1281
470 Burklen, L., Schock, F. & Dahl, M. K. Molecular analysis of the interaction between the *Bacillus subtilis* trehalose repressor TreR and the *tre* operator. *Mol Gen Genet* **260**, 48-55 (1998).
471 Riggs, A. D., Newby, R. F. & Bourgeois, S. *lac* Repressor-Operator Interaction: II Effect of Galactosides and Other Ligands. *J Mol Biol* **51**, 303-314 (1970).
472 Tack, D. S. *et al.* The genotype-phenotype landscape of an allosteric protein. *Mol Syst Biol* **17**, e10179 (2021). https://doi.org/10.15252/msb.202010179
473 Taylor, N. D. *et al.* Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods* **13**, 177-183 (2016). https://doi.org/10.1038/nmeth.3696
474 Noor, S. *et al.* Intramolecular epistasis and the evolution of a new enzymatic function. *PLoS One* **7**, e39822 (2012). https://doi.org/10.1371/journal.pone.0039822
475 Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci* **25**, 1204-1218 (2016). https://doi.org/10.1002/pro.2897

476 Canale, A. S., Cote-Hammarlof, P. A., Flynn, J. M. & Bolon, D. N. Evolutionary mechanisms studied through protein fitness landscapes. *Curr Opin Struct Biol* **48**, 141-148 (2018). https://doi.org/10.1016/j.sbi.2018.01.001
477 Hartman, E. C. & Tullman-Ercek, D. Learning from protein fitness landscapes: a review of mutability, epistasis, and evolution. *Current Opinion in Systems Biology* **14**, 25-31 (2019). https://doi.org/10.1016/j.coisb.2019.02.006
478 Suckow, J. *et al.* Genetic Studies of the Lac Repressor XV: 4000 Single Amino Acid Substitutions and Analysis of the Resulting Phenotypes on the Basis of the Protein Structure. *J. Mol. Biol.* **261**, 509-523 (1996).
479 Meger, A. T. *et al.* Rugged fitness landscapes minimize promiscuity in the evolution of transcriptional repressors. *Cell Syst* **15**, 374-387 e376 (2024). https://doi.org/10.1016/j.cels.2024.03.002
480 Parente, D. J. & Swint-Kruse, L. Multiple co-evolutionary networks are supported by the common tertiary scaffold of the LacI/GalR proteins. *PLoS One* **8**, e84398 (2013). https://doi.org/10.1371/journal.pone.0084398
481 Swint-Kruse, L. *et al.* Rheostat functional outcomes occur when substitutions are introduced at nonconserved positions that diverge with speciation. *Protein Sci* **30**, 1833-1853 (2021). https://doi.org/10.1002/pro.4136
482 Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455-463 (2022). https://doi.org/10.1038/s41586-022-04506-6
483 Lewis, M. A tale of two repressors. *J Mol Biol* **409**, 14-27 (2011). https://doi.org/10.1016/j.jmb.2011.02.023
484 Wilson, C. J., Zhan, H., Swint-Kruse, L. & Matthews, K. S. The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. *Cell Mol Life Sci* **64**, 3-16 (2007). https://doi.org/10.1007/s00018-006-6296-z
485 Xu, S. *et al.* Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *Imeta* **1**, e56 (2022). https://doi.org/10.1002/imt2.56
486 Schumacher, M. A., Seidel, G., Hillen, W. & Brennan, R. G. Phosphoprotein Crh-Ser46-P displays altered binding to CcpA to effect carbon catabolite regulation. *J Biol Chem* **281**, 6793-6800 (2006). https://doi.org/10.1074/jbc.M509977200
487 Glasfeld, A., Schumacher, M. A., Choi, K.-Y., Zalkin, H. & Brennan, R. G. A Positively Charged Residue Bound in the Minor Groove Does Not Alter the Bending of a DNA Duplex. *J Am Chem Soc* **118**, 13073-13074 (1996).
488 Greenfield, N. & Fasman, G. D. Computed Circular Dichroism Spectra for the Evaluation of Protein Conformation. *Biochemistry* **8**, 4108-4116 (1969).
489 Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **1**, 2876-2890 (2006). https://doi.org/10.1038/nprot.2006.202
490 Young, G. M. & Williams, G. E. in *Encyclopedia of Geology* 557-570 (2021).
491 BAROSS, J. A. & HOFFMAN, S. E. SUBMARINE HYDROTHERMAL VENTS AND ASSOCIATED GRADIENT ENVIRONMENTS AS SITES FOR THE ORIGIN AND EVOLUTION OF LIFE. *Origins of Life* **15**, 327-345 (1985).
492 Sleep, N. H. Geological and Geochemical Constraints on the Origin and Evolution of Life. *Astrobiology* **18**, 1199-1219 (2018). https://doi.org/10.1089/ast.2017.1778
493 Georgieva, M. N. *et al.* The history of life at hydrothermal vents. *Earth-Science Reviews* **217** (2021). https://doi.org/10.1016/j.earscirev.2021.103602
494 Yang, Z., Zeng, X., Zhao, Y. & Chen, R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct Target Ther* **8**, 115 (2023). https://doi.org/10.1038/s41392-023-01381-z
495 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022). https://doi.org/10.1038/s41592-022-01488-1

496 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
497 Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **10** (2003).
498 Luking, M., Elf, J. & Levy, Y. Conformational Change of Transcription Factors from Search to Specific Binding: A lac Repressor Case Study. *J Phys Chem B* **126**, 9971-9984 (2022). https://doi.org/10.1021/acs.jpcb.2c05006
499 Chen, J. & Matthews, K. S. Subunit dissociation affects DNA binding in a dimeric lac repressor produced by C-terminal deletion. *Biochemistry* **33**, 8728-8735 (1994). https://doi.org/10.1021/bi00195a014
500 Georgelin, R. *Molecular evolution of the LacI/GalR transcription factor family* Honours thesis, Australian National University, (2021).
501 WEICKERT, M. J. & ADHYA, S. Control of Transcription of Gal Repressor and Isorepressor Genes in *Escherichia coli*. *Journal of Bacteriology*, 251-258 (1993).
502 Alfano, C. *et al.* Molecular Crowding: The History and Development of a Scientific Paradigm. *Chem Rev* **124**, 3186-3219 (2024). https://doi.org/10.1021/acs.chemrev.3c00615
503 Raicu, V. & Singh, D. R. FRET spectrometry: a new tool for the determination of protein quaternary structure in living cells. *Biophys J* **105**, 1937-1945 (2013). https://doi.org/10.1016/j.bpj.2013.09.015
504 Mirdha, L. & Chakraborty, H. Fluorescence-based techniques for the detection of the oligomeric status of proteins: implication in amyloidogenic diseases. *Eur Biophys J* **50**, 671-685 (2021). https://doi.org/10.1007/s00249-021-01505-9
505 Fukami-Kobayashi, K., Tateno, Y. & Nishikawa, K. Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Mol Biol Evol* **20**, 267-277 (2003). https://doi.org/10.1093/molbev/msg038
506 Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat Protoc* **2**, 2212-2221 (2007). https://doi.org/10.1038/nprot.2007.321
507 Shaw, J. & Stubbs, C. in *Methods in Molecular Biology* Vol. 2263 201-215 (2021).
508 Brandts, J. F. & Lin, L.-N. Study of Strong to Ultralight Protein Interactions Using Differential Scanning Calorimetry. *Biochemistry* **29**, 6927-6940 (1990).
509 Daumantas Matulis, Kranz, J. K., Salemme, F. R. & Todd, M. J. Thermodynamic Stability of Carbonic Anhydrase: Measurements of Binding Affinity and Stoichiometry Using ThermoFluor. *Biochemistry* **44**, 5258-5266 (2005).
510 Wu, T. *et al.* Protocol for performing and optimizing differential scanning fluorimetry experiments. *STAR Protoc* **4**, 102688 (2023). https://doi.org/10.1016/j.xpro.2023.102688
511 Gao, K., Oerlemans, R. & Groves, M. R. Theory and applications of differential scanning fluorimetry in early-stage drug discovery. *Biophys Rev* **12**, 85-104 (2020). https://doi.org/10.1007/s12551-020-00619-2
512 Vivoli, M., Novak, H. R., Littlechild, J. A. & Harmer, N. J. Determination of protein-ligand interactions using differential scanning fluorimetry. *J Vis Exp*, 51809 (2014). https://doi.org/10.3791/51809
513 Englebienne, P., Hoonacker, A. V. & Verhas, M. Surface plasmon resonance: principles, methods and applications in biomedical sciences. *Journal of Spectroscopy* **17**, 255-273 (2003). https://doi.org/10.1155/2003/372913
514 Maynard, J. A. *et al.* Surface plasmon resonance for high-throughput ligand screening of membrane-bound proteins. *Biotechnol J* **4**, 1542-1558 (2009). https://doi.org/10.1002/biot.200900195

515 Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein-ligand interactions and its potential for drug discovery. *Biochim Biophys Acta* **1838**, 43-55 (2014). https://doi.org/10.1016/j.bbamem.2013.04.028
516 Du, X. *et al.* Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *Int J Mol Sci* **17** (2016). https://doi.org/10.3390/ijms17020144
517 Pollack, S. J. in *Biophysical Techniques in Drug Discovery* (ed Angeles Canales) 0 (The Royal Society of Chemistry, 2017).
518 Campitelli, P., Ross, D., Swint-Kruse, L. & Ozkan, S. B. Dynamics-based protein network features accurately discriminate neutral and rheostat positions. *Biophys J* **123**, 3612-3626 (2024). https://doi.org/10.1016/j.bpj.2024.09.013
519 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). https://doi.org/10.1146/annurev-biophys-070816-033743
520 Lumry, R. & Rajender, S. Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: a ubiquitous property of water. *Biopolymers* **9**, 1125-1227 (1970). https://doi.org/10.1002/bip.1970.360091002
521 Sharp, K. Entropy-enthalpy compensation: fact or artifact? *Protein Sci* **10**, 661-667 (2001). https://doi.org/10.1110/ps.37801
522 Olsson, T. S., Ladbury, J. E., Pitt, W. R. & Williams, M. A. Extent of enthalpy-entropy compensation in protein-ligand interactions. *Protein Sci* **20**, 1607-1618 (2011). https://doi.org/10.1002/pro.692
523 Dragan, A. I., Read, C. M. & Crane-Robinson, C. Enthalpy-entropy compensation: the role of solvation. *Eur Biophys J* **46**, 301-308 (2017). https://doi.org/10.1007/s00249-016-1182-6
524 Jiménez, J. S. & Benítez, M. J. Gibbs Free Energy and Enthalpy–Entropy Compensation in Protein–Ligand Interactions. *Biophysica* **4**, 298-309 (2024). https://doi.org/10.3390/biophysica4020021
525 Camps, M., Herman, A., Loh, E. & Loeb, L. A. Genetic constraints on protein evolution. *Crit Rev Biochem Mol Biol* **42**, 313-326 (2007). https://doi.org/10.1080/10409230701597642
526 Jayaraman, V., Toledo-Patino, S., Noda-Garcia, L. & Laurino, P. Mechanisms of protein evolution. *Protein Sci* **31**, e4362 (2022). https://doi.org/10.1002/pro.4362
527 Yammine, A., Gao, J. & Kwan, A. H. Tryptophan Fluorescence Quenching Assays for Measuring Protein-ligand Binding Affinities: Principles and a Practical Guide. *Bio Protoc* **9**, e3253 (2019). https://doi.org/10.21769/BioProtoc.3253
528 Maliwal, B. P. & Lakowicz, J. R. Effect of ligand binding and conformational changes in proteins on oxygen quenching and fluorescence depolarization of tryptophan residues. *Biophys Chem* **19**, 337-344 (1984).
529 Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories. *PLoS Comput Biol* **13**, e1005541 (2017). https://doi.org/10.1371/journal.pcbi.1005541
530 Diaz-Colunga, J. *et al.* Global epistasis on fitness landscapes. *Philos Trans R Soc Lond B Biol Sci* **378**, 20220053 (2023). https://doi.org/10.1098/rstb.2022.0053
531 Johnson, M. S., Reddy, G. & Desai, M. M. Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biol* **21**, 120 (2023). https://doi.org/10.1186/s12915-023-01585-3
532 Tungtur, S., Parente, D. J. & Swint-Kruse, L. Functionally important positions can comprise the majority of a protein's architecture. *Proteins* **79**, 1589-1608 (2011). https://doi.org/10.1002/prot.22985
533 Campitelli, P., Swint-Kruse, L. & Ozkan, S. B. Substitutions at Nonconserved Rheostat Positions Modulate Function by Rewiring Long-Range, Dynamic

Interactions. *Mol Biol Evol* **38**, 201-214 (2021). https://doi.org/10.1093/molbev/msaa202
534 Bigman, L. S. & Levy, Y. Proteins: molecules defined by their trade-offs. *Curr Opin Struct Biol* **60**, 50-56 (2020). https://doi.org/10.1016/j.sbi.2019.11.005
535 Bar-Rogovsky, H. *et al.* Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein Eng Des Sel* **28**, 507-518 (2015). https://doi.org/10.1093/protein/gzv038
536 Semsey, S., Virnik, K. & Adhya, S. Three-stage regulation of the amphibolic gal operon: from repressosome to GalR-free DNA. *J Mol Biol* **358**, 355-363 (2006). https://doi.org/10.1016/j.jmb.2006.02.022
537 Stenberg, K. A. & Vihinen, M. Crystal structure of a 1.6-hexanediol bound tetrameric form of Escherichia coli lac-repressor refined to 2.1 A resolution. *Proteins* **75**, 748-759 (2009). https://doi.org/10.1002/prot.22284
538 Borkowski, O., Ceroni, F., Stan, G. B. & Ellis, T. Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. *Curr Opin Microbiol* **33**, 123-130 (2016). https://doi.org/10.1016/j.mib.2016.07.009
539 Boos, W., Schaedel, P. & Wallenfels, K. [Investigation on induction of lac-enzyme. 1. Induction process and permeation]. *Eur J Biochem* **1**, 382-394 (1967). https://doi.org/10.1111/j.1432-1033.1967.tb00086.x
540 Gorke, B. & Stulke, J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat Rev Microbiol* **6**, 613-624 (2008). https://doi.org/10.1038/nrmicro1932
541 Wagner, G. P. & Lynch, V. J. The gene regulatory logic of transcription factor evolution. *Trends Ecol Evol* **23**, 377-385 (2008). https://doi.org/10.1016/j.tree.2008.03.006
542 Erwin, D. H. in *Curr Top Dev Biol* Vol. 139 Ch. Evolutionary dynamics of gene regulation, 407-431 (2020).
543 Mitsis, T. *et al.* Transcription factors and evolution: An integral part of gene expression (Review). *World Academy of Sciences Journal* (2020). https://doi.org/10.3892/wasj.2020.32
544 Perica, T. *et al.* Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* **346**, 1254346 (2014). https://doi.org/10.1126/science.1254346
545 Liu, A. K. *et al.* Structural plasticity enables evolution and innovation of RuBisCO assemblies. *Science Advances* **8** (2022).
546 Klose, D. *et al.* Resolving distance variations by single-molecule FRET and EPR spectroscopy using rotamer libraries. *Biophys J* **120**, 4842-4858 (2021). https://doi.org/10.1016/j.bpj.2021.09.021
547 Wyborski, D. L. & Short, J. M. Analysis of inducers of the E.coli lac repressor system in mammalian cells and whole animals. *Nucleic Acids Res* **19**, 4647-4653 (1991).
548 Breslow, R. On the mechanism of the formose reaction. *Tetrahedron Letters* **21**, 22-26 (1959).
549 Jalbout, A. F. Prebiotic synthesis of simple sugars by an interstellar formose reaction. *Orig Life Evol Biosph* **38**, 489-497 (2008). https://doi.org/10.1007/s11084-008-9151-4
550 Cantillo, D. *et al.* On the prebiotic synthesis of D-sugars catalyzed by L-peptides: assessments from first-principles calculations. *Chemistry* **18**, 8795-8799 (2012). https://doi.org/10.1002/chem.201200466
551 Ritson, D. & Sutherland, J. D. Prebiotic synthesis of simple sugars by photoredox systems chemistry. *Nat Chem* **4**, 895-899 (2012). https://doi.org/10.1038/nchem.1467

552 Haas, M., Lamour, S., Christ, S. B. & Trapp, O. Mineral-mediated carbohydrate synthesis by mechanical forces in a primordial geochemical setting. *Commun Chem* **3**, 140 (2020). https://doi.org/10.1038/s42004-020-00387-w
553 Vinogradoff, V. *et al.* Olivine-catalyzed glycolaldehyde and sugar synthesis under aqueous conditions: Application to prebiotic chemistry. *Earth and Planetary Science Letters* **626** (2024). https://doi.org/10.1016/j.epsl.2023.118558
554 Buttin, G. Mecanismes Regulateurs dans la Biosynthese des Enzymes du Metabolisme du Galactose chez Escherichia coli K12. *J. Mol. Biol.* **7**, 164-182 (1963).
555 POSTMA, P. W. Galactose Transport in Salmonella typhimurium. *Journal of Bacteriology* **129**, 630-639 (1977).
556 SHEN, B. H. P. & BOOS, W. Regulation of the B-Methylgalactoside Transport System and the Galactose-Binding Protein by the Cell Cycle of Escherichia coli. *PNAS* **70**, 1481-1485 (1973).
557 Wilcox, G. The Interaction of l-Arabinose and d-Fucose with AraC Protein. *Journal of Biological Chemistry* **249**, 6892-6894 (1974). https://doi.org/10.1016/s0021-9258(19)42141-8
558 ENGLESBERG, E., IRR, J., POWER, J. & LEE, N. Positive Control of Enzyme Synthesis by Gene C in the L-Arabinose System. *Journal of Bacteriology* **90**, 946-957 (1965).
559 Hersey, A. N., Kay, V. E., Lee, S., Realff, M. J. & Wilson, C. J. Engineering allosteric transcription factors guided by the LacI topology. *Cell Syst* **14**, 645-655 (2023). https://doi.org/10.1016/j.cels.2023.04.008
560 Clifton, B. E. & Jackson, C. J. Ancestral Protein Reconstruction Yields Insights into Adaptive Evolution of Binding Specificity in Solute-Binding Proteins. *Cell Chem Biol* **23**, 236-245 (2016). https://doi.org/10.1016/j.chembiol.2015.12.010
561 Palma, M. *et al.* The Zygosaccharomyces bailii transcription factor Haa1 is required for acetic acid and copper stress responses suggesting subfunctionalization of the ancestral bifunctional protein Haa1/Cup2. *BMC Genomics* **18**, 75 (2017). https://doi.org/10.1186/s12864-016-3443-2
562 Colinas, M. *et al.* Subfunctionalization of Paralog Transcription Factors Contributes to Regulation of Alkaloid Pathway Branch Choice in Catharanthus roseus. *Front Plant Sci* **12**, 687406 (2021). https://doi.org/10.3389/fpls.2021.687406
563 Rojo-Bartolome, I., Santana de Souza, J. E., Diaz de Cerio, O. & Cancio, I. Duplication and subfunctionalisation of the general transcription factor IIIA (gtf3a) gene in teleost genomes, with ovarian specific transcription of gtf3ab. *PLoS One* **15**, e0227690 (2020). https://doi.org/10.1371/journal.pone.0227690
564 Klingel, S. *et al.* Subfunctionalization and neofunctionalization of vertebrate Lef/Tcf transcription factors. *Dev Biol* **368**, 44-53 (2012). https://doi.org/10.1016/j.ydbio.2012.05.012
565 Ueda, M. *et al.* Subfunctionalization of sigma factors during the evolution of land plants based on mutant analysis of liverwort (Marchantia polymorpha L.) MpSIG1. *Genome Biol Evol* **5**, 1836-1848 (2013). https://doi.org/10.1093/gbe/evt137
566 Moiseenko, K. V. *et al.* Orchestration of the expression of the laccase multigene family in white-rot basidiomycete Trametes hirsuta 072: Evidences of transcription level subfunctionalization. *Fungal Biol* **122**, 353-362 (2018). https://doi.org/10.1016/j.funbio.2018.02.006
567 Kluver, N., Kondo, M., Herpin, A., Mitani, H. & Schartl, M. Divergent expression patterns of Sox9 duplicates in teleosts indicate a lineage specific subfunctionalization. *Dev Genes Evol* **215**, 297-305 (2005). https://doi.org/10.1007/s00427-005-0477-x

568 Morcos, F. Characterizing the landscape of evolvability. *Nat Ecol Evol* **6**, 500-501 (2022). https://doi.org/10.1038/s41559-022-01731-0
569 Visweswariah, S. S. & Busby, S. J. Evolution of bacterial transcription factors: how proteins take on new tasks, but do not always stop doing the old ones. *Trends Microbiol* **23**, 463-467 (2015). https://doi.org/10.1016/j.tim.2015.04.009
570 MONOD, J., COHEN-BAZIRE, G. & COHN, M. SUR LA BIOSYNTHESE DE LA p-GALACTOSIDASE (LACTASE) CHEZ ESCHERICHIA COL LA SPECIFICITE DE L'INDUCTION. *Biochim Biophys Acta* **7**, 585-599 (1951).
571 MULLER-HILL, B., RICKENBERG, H. V. & WALLENFELS, K. Specificity of the Induction ofthe Enzymes ofthe Lac Operon in Escherichia coli. *J. Mol. Biol.* **10**, 303-318 (1964).
572 Urashima, T., Fukuda, K. & Messer, M. Evolution of milk oligosaccharides and lactose: a hypothesis. *Animal* **6**, 369-374 (2012). https://doi.org/10.1017/S1751731111001248
573 Urashima, T., Katayama, T., Sakanaka, M., Fukuda, K. & Messer, M. Evolution of milk oligosaccharides: Origin and selectivity of the ratio of milk oligosaccharides to lactose among mammals. *Biochim Biophys Acta Gen Subj* **1866**, 130012 (2022). https://doi.org/10.1016/j.bbagen.2021.130012
574 Grossnickle, D. M., Smith, S. M. & Wilson, G. P. Untangling the Multiple Ecological Radiations of Early Mammals. *Trends Ecol Evol* **34**, 936-949 (2019). https://doi.org/10.1016/j.tree.2019.05.008
575 Foley, N. M. *et al.* A genomic timescale for placental mammal evolution. *Science* **380**, eabl8189 (2023). https://doi.org/10.1126/science.abl8189
576 Yang, G. *et al.* Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat Chem Biol* **15**, 1120-1128 (2019). https://doi.org/10.1038/s41589-019-0386-3
577 Ju, K. S., Parales, J. V. & Parales, R. E. Reconstructing the evolutionary history of nitrotoluene detection in the transcriptional regulator NtdR. *Mol Microbiol* **74**, 826-843 (2009). https://doi.org/10.1111/j.1365-2958.2009.06904.x
578 Jiang, P. P., Corbett-Detig, R. B., Hartl, D. L. & Lozovsky, E. R. Accessible mutational trajectories for the evolution of pyrimethamine resistance in the malaria parasite Plasmodium vivax. *J Mol Evol* **77**, 81-91 (2013). https://doi.org/10.1007/s00239-013-9582-z
579 Gibbs, J. A METHOD OF GEOMETRICAL REPRESENTATION OF THE THERMODYNAMIC PROPERTIES OF SUBSTANCES BY MEANS OF SURFACES. *Transactions of the Connecticut Academy of Arts and Sciences* **2**, 382-404 (1873).
580 *Paleoclimatology*. (Springer Nature, 2020).
581 Isson, T. & Rauzi, S. Oxygen isotope ensemble reveals Earth's seawater, temperature, and carbon cycle history. *Science* **383**, 666-670 (2024).
582 Gedde, U. W. *Essential Classical Thermodynamics*. (Springer Cham, 2020).
583 Komarov, I. V., Bugrov, V. A., Cherednychenko, A. & Grygorenko, O. O. Insights into Modeling Approaches in Chemistry: Assessing Ligand-Protein Binding Thermodynamics Based on Rigid-Flexible Model Molecules. *Chem Rec* **24**, e202300276 (2024). https://doi.org/10.1002/tcr.202300276
584 Kuroki, R., Nitta, K. & Yutani, K. Thermodynamic changes in the binding of Ca2+ to a mutant human lysozyme (D86/92). Enthalpy-entropy compensation observed upon Ca2+ binding to proteins. *Journal of Biological Chemistry* **267**, 24297-24301 (1992). https://doi.org/10.1016/s0021-9258(18)35764-8

585 Joynt, S., Morillo, V. & Leng, F. Binding the mammalian high mobility group protein AT-hook 2 to AT-rich deoxyoligonucleotides: enthalpy-entropy compensation. *Biophys J* **96**, 4144-4152 (2009). https://doi.org/10.1016/j.bpj.2009.02.015
586 Alcaraz, A., Queralt-Martin, M., Verdia-Baguena, C., Aguilella, V. M. & Mafe, S. Entropy-enthalpy compensation at the single protein level: pH sensing in the bacterial channel OmpF. *Nanoscale* **6**, 15210-15215 (2014). https://doi.org/10.1039/c4nr03811h
587 Kragelj, J. *et al.* Enthalpy-Entropy Compensation in the Promiscuous Interaction of an Intrinsically Disordered Protein with Homologous Protein Partners. *Biomolecules* **11** (2021). https://doi.org/10.3390/biom11081204
588 Copp, J. N., Anderson, D. W., Akiva, E., Babbitt, P. C. & Tokuriki, N. in *Methods Enzymol* Vol. 620 Ch. 12, 315-347 (2019).
589 Tungtur, S., Meinhardt, S. & Swint-Kruse, L. Comparing the functional roles of nonconserved sequence positions in homologous transcription repressors: implications for sequence/function analyses. *J Mol Biol* **395**, 785-802 (2010). https://doi.org/10.1016/j.jmb.2009.10.001
590 Jiang, X. L., Dimas, R. P., Chan, C. T. Y. & Morcos, F. Coevolutionary methods enable robust design of modular repressors by reestablishing intra-protein interactions. *Nat Commun* **12**, 5592 (2021). https://doi.org/10.1038/s41467-021-25851-6
591 Swint-Kruse, L. Using Evolution to Guide Protein Engineering: The Devil IS in the Details. *Biophys J* **111**, 10-18 (2016). https://doi.org/10.1016/j.bpj.2016.05.030
592 Garruss, A. S., Collins, K. M. & Church, G. M. Deep representation learning improves prediction of LacI-mediated transcriptional repression. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org/10.1073/pnas.2022838118
593 Meger, A. T. *et al.* Rugged fitness landscapes minimize promiscuity in the evolution of transcriptional repressors. *Cell Systems* **15**, 374-387.e376 (2024). https://doi.org/https://doi.org/10.1016/j.cels.2024.03.002
594 Gasteiger, E. *et al. Protein Identification and Analysis Tools on the Expasy Server*. 571-607 (Humana Press, 2005).
595 Keller, S. *et al.* High-precision isothermal titration calorimetry with automated peak-shape analysis. *Anal Chem* **84**, 5066-5073 (2012). https://doi.org/10.1021/ac3007522
596 Scheuermann, T. H. & Brautigam, C. A. High-precision, automated integration of multiple isothermal titration calorimetric thermograms: new features of NITPIC. *Methods* **76**, 87-98 (2015). https://doi.org/10.1016/j.ymeth.2014.11.024
597 Zhao, H., Piszczek, G. & Schuck, P. SEDPHAT--a platform for global ITC analysis and global multi-method analysis of molecular interactions. *Methods* **76**, 137-148 (2015). https://doi.org/10.1016/j.ymeth.2014.11.012
598 Tellinghuisen, J. Isothermal titration calorimetry at very low c. *Anal Biochem* **373**, 395-397 (2008). https://doi.org/10.1016/j.ab.2007.08.039
599 Brautigam, C. A. in *Methods Enzymol* Vol. 562 Ch. 5, 109-133 (2015).
600 Planche, L. A. Note sur la sophistication de la resine de jalap et sur les moyens de la reconnaitre, etc. *Bull. Pharmacie* **2**, 578-580 (1810).
601 Heckmann, C. M. & Paradisi, F. Looking Back: A Short History of the Discovery of Enzymes and How They Became Powerful Chemical Tools. *ChemCatChem* **12**, 6082-6102 (2020). https://doi.org/10.1002/cctc.202001107
602 Lopes, G. R., Pinto, D. C. G. A. & Silva, A. M. S. Horseradish peroxidase (HRP) as a tool in green chemistry. *RSC Adv.* **4**, 37244-37265 (2014). https://doi.org/10.1039/c4ra06094f

603 Veitch, N. C. Horseradish peroxidase: a modern view of a classic enzyme. *Phytochemistry* **65**, 249-259 (2004). https://doi.org/10.1016/j.phytochem.2003.10.022
604 Grigorenko, V. G., Andreeva, I. P., Rubtsova, M. Y. & Egorov, A. M. Recombinant horseradish peroxidase: production and analytical applications. *Biochemistry (Mosc)* **80**, 408-416 (2015). https://doi.org/10.1134/S0006297915040033
605 Wei, X. *et al.* Microfluidic Distance Readout Sweet Hydrogel Integrated Paper-Based Analytical Device (muDiSH-PAD) for Visual Quantitative Point-of-Care Testing. *Anal Chem* **88**, 2345-2352 (2016). https://doi.org/10.1021/acs.analchem.5b04294
606 Yu, S. *et al.* Which one of the two common reporter systems is more suitable for chemiluminescent enzyme immunoassay: alkaline phosphatase or horseradish peroxidase? *Luminescence* **31**, 888-892 (2016). https://doi.org/10.1002/bio.3047
607 Zhao, L. *et al.* Novel dynamic light scattering immunosensor for prostate specific antigens based upon dual-tyramine signal amplification strategy. *Sensors and Actuators B: Chemical* **418** (2024). https://doi.org/10.1016/j.snb.2024.136242
608 Yang, X., Jin, C., Yang, S. & Tian, M. Paper-based enzyme-linked biosensor combined with smartphone for simultaneous colorimetric sensing of xanthines and sarcosine. *Sensors and Actuators B: Chemical* **412** (2024). https://doi.org/10.1016/j.snb.2024.135849
609 Koposova, E. *et al.* Bioelectrochemical systems with oleylamine-stabilized gold nanostructures and horseradish peroxidase for hydrogen peroxide sensor. *Biosens Bioelectron* **57**, 54-58 (2014). https://doi.org/10.1016/j.bios.2014.01.034
610 Li, F. *et al.* Luminol, horseradish peroxidase, and glucose oxidase ternary functionalized graphene oxide for ultrasensitive glucose sensing. *Anal Bioanal Chem* **410**, 543-552 (2018). https://doi.org/10.1007/s00216-017-0752-5
611 Bilal, M. *et al.* Horseradish peroxidase immobilization by copolymerization into cross-linked polyacrylamide gel and its dye degradation and detoxification potential. *Int J Biol Macromol* **113**, 983-990 (2018). https://doi.org/10.1016/j.ijbiomac.2018.02.062
612 Mirdamadian, S. H., Moghimi, H., Asad, S., Dastgheib, S. M. M. & Karimian, F. Horseradish peroxidase-calcium peroxide core–shell microcapsules as a novel permeable reactive barrier for bioremediation of phenol-contaminated waters. *International Journal of Environmental Science and Technology* **19**, 3165-3176 (2021). https://doi.org/10.1007/s13762-021-03458-0
613 Mohamed, S. A. *et al.* Development of chia gum/alginate-polymer support for horseradish peroxidase immobilization and its application in phenolic removal. *Sci Rep* **14**, 1362 (2024). https://doi.org/10.1038/s41598-024-51566-x
614 Danielson, A. P. *et al.* Well-Defined Macromolecules Using Horseradish Peroxidase as a RAFT Initiase. *Macromol Rapid Commun* **37**, 362-367 (2016). https://doi.org/10.1002/marc.201500633
615 Sahoo, J. K. *et al.* Horseradish Peroxidase Catalyzed Silk-Prefoldin Composite Hydrogel Networks. *ACS Appl Bio Mater* **6**, 203-208 (2023). https://doi.org/10.1021/acsabm.2c00836
616 Gokturk, E., Ocsoy, I., Turac, E. & Sahmetlioglu, E. Horseradish peroxidase-based hybrid nanoflowers with enhanced catalytical activities for polymerization reactions of phenol derivatives. *Polymers for Advanced Technologies* **31**, 2371-2377 (2020). https://doi.org/10.1002/pat.4956
617 Wei, Q. *et al.* One-pot preparation of double network hydrogels via enzyme-mediated polymerization and post-self-assembly for wound healing. *J Mater Chem B* **7**, 6195-6201 (2019). https://doi.org/10.1039/c9tb01667h

618 Guo, Z. *et al.* Biocompatibility and cellular uptake mechanisms of poly(N-isopropylacrylamide) in different cells. *Journal of Bioactive and Compatible Polymers* **32**, 17-31 (2016). https://doi.org/10.1177/0883911516648969
619 Lima, L. H., Morales, Y. & Cabral, T. Ocular Biocompatibility of Poly-N-Isopropylacrylamide (pNIPAM). *J Ophthalmol* **2016**, 5356371 (2016). https://doi.org/10.1155/2016/5356371
620 Heskins, M. & Guillet, J. E. Solution Properties of Poly(N-isopropylacrylamide). *Journal of Macromolecular Science: Part A - Chemistry* **2**, 1441-1455 (1968). https://doi.org/10.1080/10601326808051910
621 Halperin, A., Kroger, M. & Winnik, F. M. Poly(N-isopropylacrylamide) Phase Diagrams: Fifty Years of Research. *Angew Chem Int Ed Engl* **54**, 15342-15367 (2015). https://doi.org/10.1002/anie.201506663
622 Din, M. I. *et al.* Recent progress of poly (N-isopropylacrylamide) hybrid hydrogels: synthesis, fundamentals and applications – review. *Soft Materials* **16**, 228-247 (2018). https://doi.org/10.1080/1539445x.2018.1461650
623 Shaibie, N. A., Ramli, N. A., Mohammad Faizal, N. D. F., Srichana, T. & Mohd Amin, M. C. I. Poly(N-isopropylacrylamide)-Based Polymers: Recent Overview for the Development of Temperature-Responsive Drug Delivery and Biomedical Applications. *Macromolecular Chemistry and Physics* **224** (2023). https://doi.org/10.1002/macp.202300157
624 Tang, L. *et al.* Poly(N-isopropylacrylamide)-based smart hydrogels: Design, properties and applications. *Progress in Materials Science* **115** (2021). https://doi.org/10.1016/j.pmatsci.2020.100702
625 Ray, B. *et al.* Synthesis of Isotactic Poly(N-isopropylacrylamide) by RAFT Polymerization in the Presence of Lewis Acid. *Macromolecules* **36**, 543-545 (2003).
626 Behere, D. V., Gonzalez-Vergara, E. & Goff, H. COMPARISON OF HEME ENVIRONMENTS AND PROXIMAL LIGANDS IN PEROXIDASES AND OTHER HEMOPROTEINS THROUGH CARBON-13 NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY OF CARBON MONOXIDE COMPLEXES. *Biochemical and Biophysical Research Communications* **131**, 607-613 (1985).
627 Mathe, C., Barre, A., Jourda, C. & Dunand, C. Evolution and expression of class III peroxidases. *Arch Biochem Biophys* **500**, 58-65 (2010). https://doi.org/10.1016/j.abb.2010.04.007
628 Savelli, B. *et al.* RedoxiBase: A database for ROS homeostasis regulated proteins. *Redox Biol* **26**, 101247 (2019). https://doi.org/10.1016/j.redox.2019.101247
629 Mbadinga Mbadinga, D. L., Li, Q., Ranocha, P., Martinez, Y. & Dunand, C. Global analysis of non-animal peroxidases provides insights into the evolution of this gene family in the green lineage. *J Exp Bot* **71**, 3350-3360 (2020). https://doi.org/10.1093/jxb/eraa141
630 Koua, D. *et al.* PeroxiBase: a database with new tools for peroxidase family classification. *Nucleic Acids Res* **37**, D261-266 (2009). https://doi.org/10.1093/nar/gkn680
631 Zamocky, M., Furtmuller, P. G. & Obinger, C. Evolution of structure and function of Class I peroxidases. *Arch Biochem Biophys* **500**, 45-57 (2010). https://doi.org/10.1016/j.abb.2010.03.024
632 Mathe, C., Fawal, N., Roux, C. & Dunand, C. In silico definition of new ligninolytic peroxidase sub-classes in fungi and putative relation to fungal life style. *Sci Rep* **9**, 20373 (2019). https://doi.org/10.1038/s41598-019-56774-4
633 Bakalovic, N. *et al.* PeroxiBase: a class III plant peroxidase database. *Phytochemistry* **67**, 534-539 (2006). https://doi.org/10.1016/j.phytochem.2005.12.020

634 Freitas, C. D. T. *et al.* Class III plant peroxidases: From classification to physiological functions. *Int J Biol Macromol* **263**, 130306 (2024). https://doi.org/10.1016/j.ijbiomac.2024.130306
635 Maehly, A. C. & Chance, B. The assay of catalases and peroxidases. *Methods Biochem Anal* **1**, 357-424 (1954). https://doi.org/10.1002/9780470110171.ch14
636 Lebedeva, O. V. & Ugarova, N. N. Mechanism of peroxidase-catalyzed oxidation. Substrate-substrate activation in horseradish peroxidase-catalyzed reactions. *Russian Chemical Bulletin* **45**, 18-25 (1996).
637 Rodrigues, A. P., da Fonseca, L. M., de Faria Oliveira, O. M., Brunetti, I. L. & Ximenes, V. F. Oxidation of acetylacetone catalyzed by horseradish peroxidase in the absence of hydrogen peroxide. *Biochim Biophys Acta* **1760**, 1755-1761 (2006). https://doi.org/10.1016/j.bbagen.2006.09.008
638 Schubert, F. Mediated amperometric enzyme electrode incorporating peroxidase for the determination of hydrogen peroxide in organic solvents. *Analytica Chimica Acta* **245**, 133-138 (1991).
639 Henriksen, A., Smith, A. T. & Gajhede, M. The structures of the horseradish peroxidase C-ferulic acid complex and the ternary complex with cyanide suggest how peroxidases oxidize small phenolic substrates. *J Biol Chem* **274**, 35005-35011 (1999). https://doi.org/10.1074/jbc.274.49.35005
640 Welinder, K. G. Covalent structure of the glycoprotein horseradish peroxidase (EC 1.11.1.7). *FEBS Lett* **72**, 19-23 (1976). https://doi.org/10.1016/0014-5793(76)80804-6
641 Aibara, S., Kobayashi, T. & Morita, Y. Isolation and properties of basic isoenzymes of horseradish peroxidase. *J. Biochem.* **90**, 489-496 (1981).
642 Ye, Y. *et al.* Subcellular location of horseradish peroxidase in horseradish leaves treated with La(III), Ce(III) and Tb(III). *Ecotoxicol Environ Saf* **71**, 677-684 (2008). https://doi.org/10.1016/j.ecoenv.2007.11.020
643 Matsui, T., Nakayama, H., Yoshida, K. & Shinmyo, A. Vesicular transport route of horseradish C1a peroxidase is regulated by N- and C-terminal propeptides in tobacco cells. *Appl Microbiol Biotechnol* **62**, 517-522 (2003). https://doi.org/10.1007/s00253-003-1273-z
644 Humer, D., Ebner, J. & Spadiut, O. Scalable High-Performance Production of Recombinant Horseradish Peroxidase from E. coli Inclusion Bodies. *Int J Mol Sci* **21** (2020). https://doi.org/10.3390/ijms21134625
645 Lavery, C. B. *et al.* Purification of peroxidase from Horseradish (Armoracia rusticana) roots. *J Agric Food Chem* **58**, 8471-8476 (2010). https://doi.org/10.1021/jf100786h
646 Naatsaari, L., Krainer, F., Schubert, M., Glieder, A. & Thallinger, G. Peroxidase gene discovery from the horseradish transcriptome. *BMC Genomics* **15** (2014). https://doi.org/https://doi.org/10.1186/1471-2164-15-227
647 Jermyn, M. A. & Thomas, R. Multiple Components in Horse-Radish Peroxidase. *Biochem. J.* **56**, 631-639 (1954). https://doi.org/https://doi.org/10.1042/bj0560631
648 Krainer, F. W. *et al.* Purification and basic biochemical characterization of 19 recombinant plant peroxidase isoenzymes produced in Pichia pastoris. *Protein Expr Purif* **95**, 104-112 (2014). https://doi.org/10.1016/j.pep.2013.12.003
649 Hoyle, M. High Resolution of Peroxidase-Indoleacetic Acid Oxidase Isoenzymes from Horseradish by Isoelectric Focusing. *Plant Physiology* **60**, 787-793 (1977).
650 Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Mol Biol Rep* **49**, 539-565 (2022). https://doi.org/10.1007/s11033-021-06749-4
651 Liu, Y., Yang, Q. & Zhao, F. Synonymous but Not Silent: The Codon Usage Code for Gene Expression and Protein Folding. *Annu Rev Biochem* **90**, 375-401 (2021). https://doi.org/10.1146/annurev-biochem-071320-112701

652 Kaur, J., Kumar, A. & Kaur, J. Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements. *Int J Biol Macromol* **106**, 803-822 (2018). https://doi.org/10.1016/j.ijbiomac.2017.08.080
653 Lobstein, J. *et al.* SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories* **11** (2012).
654 Eichler, J. Protein glycosylation. *Curr Biol* **29**, R229-R231 (2019). https://doi.org/10.1016/j.cub.2019.01.003
655 Krainer, F. W. & Glieder, A. An updated view on horseradish peroxidases: recombinant production and biotechnological applications. *Appl Microbiol Biotechnol* **99**, 1611-1625 (2015). https://doi.org/10.1007/s00253-014-6346-7
656 Rutherford, S. L. Between genotype and phenotype: protein chaperones and evolvability. *Nat Rev Genet* **4**, 263-274 (2003). https://doi.org/10.1038/nrg1041
657 Bhatia, S. & Udgaonkar, J. B. Heterogeneity in Protein Folding and Unfolding Reactions. *Chem Rev* **122**, 8911-8935 (2022). https://doi.org/10.1021/acs.chemrev.1c00704
658 Kondo, A. *et al.* Improvement of Productivity of Active Horseradish Peroxidase in *Escherichia coli* by Coexpression of Dsb Proteins. *Journal of Bioscience and Bioengineering* **90**, 600-606 (2000).
659 Chauhan, S. & Kang, T. J. Soluble expression of horseradish peroxidase in Escherichia coli and its facile activation. *J Biosci Bioeng* **126**, 431-435 (2018). https://doi.org/10.1016/j.jbiosc.2018.04.004
660 Gundinger, T. & Spadiut, O. A comparative approach to recombinantly produce the plant enzyme horseradish peroxidase in Escherichia coli. *J Biotechnol* **248**, 15-24 (2017). https://doi.org/10.1016/j.jbiotec.2017.03.003
661 Park, Y. J. & Kim, D. M. Production of Recombinant Horseradish Peroxidase in an Engineered Cell-free Protein Synthesis System. *Front Bioeng Biotechnol* **9**, 778496 (2021). https://doi.org/10.3389/fbioe.2021.778496
662 Segura, M. a. d. l. M. *et al.* High-level expression and purification of recombinant horseradish peroxidase isozyme C in SF-9 insect cell culture. *Process Biochemistry* **40**, 795-800 (2005). https://doi.org/10.1016/j.procbio.2004.02.009
663 Utashima, Y., Matsumoto, H., Masaki, K. & Iefuji, H. Heterologous production of horseradish peroxidase C1a by the basidiomycete yeast Cryptococcus sp. S-2 using codon and signal optimizations. *Appl Microbiol Biotechnol* **98**, 7893-7900 (2014). https://doi.org/10.1007/s00253-014-5856-7
664 Krainer, F. W., Gerstmann, M. A., Darnhofer, B., Birner-Gruenberger, R. & Glieder, A. Biotechnological advances towards an enhanced peroxidase production in Pichia pastoris. *J Biotechnol* **233**, 181-189 (2016). https://doi.org/10.1016/j.jbiotec.2016.07.012
665 Lin, Z., Thorsen, T. & Arnold, F. H. Functional Expression of Horseradish Peroxidase in *E. coli* by Directed Evolution. *Biotechnol. Prog.* **15**, 467-471 (1999).
666 *Pichia protocols*. 2 edn, (Humana Press, 2007).
667 Alseekh, S., Kostova, D., Bulut, M. & Fernie, A. R. Genome-wide association studies: assessing trait characteristics in model and crop plants. *Cell Mol Life Sci* **78**, 5743-5754 (2021). https://doi.org/10.1007/s00018-021-03868-w
668 Tseng, W.-C., Hsu, C.-T., Chang, H.-C., Wang, M.-J. & Fang, T.-Y. Fusion of the peptide derived from the acidic tail of alpha-synuclein improves the thermostability and soluble expression of recombinant Agrobacterium sp. d-allulose 3-epimerase. *Biochemical Engineering Journal* **165** (2021). https://doi.org/10.1016/j.bej.2020.107828

669 Zhao, W., Liu, L., Du, G. & Liu, S. A multifunctional tag with the ability to benefit the expression, purification, thermostability and activity of recombinant proteins. *J Biotechnol* **283**, 1-10 (2018). https://doi.org/10.1016/j.jbiotec.2018.07.005
670 Yainoy, S. *et al.* Production and Characterization of Recombinant Wild Type Uricase from Indonesian Coelacanth (L. menadoensis) and Improvement of Its Thermostability by In Silico Rational Design and Disulphide Bridges Engineering. *Int J Mol Sci* **20** (2019). https://doi.org/10.3390/ijms20061269
671 Ayuso-Fernandez, I., Rencoret, J., Gutierrez, A., Ruiz-Duenas, F. J. & Martinez, A. T. Peroxidase evolution in white-rot fungi follows wood lignin evolution in plants. *Proc Natl Acad Sci U S A* **116**, 17900-17905 (2019). https://doi.org/10.1073/pnas.1905040116
672 Zitare, U. A. *et al.* Mutational and structural analysis of an ancestral fungal dye-decolorizing peroxidase. *FEBS J* **288**, 3602-3618 (2021). https://doi.org/10.1111/febs.15687
673 Barber-Zucker, S. *et al.* Stable and Functionally Diverse Versatile Peroxidases Designed Directly from Sequences. *J Am Chem Soc* **144**, 3564-3571 (2022). https://doi.org/10.1021/jacs.1c12433
674 Weinstein, J. J., Goldenzweig, A., Hoch, S. & Fleishman, S. J. PROSS 2: a new server for the design of stable and highly expressed protein variants. *Bioinformatics* **37**, 123-125 (2021). https://doi.org/10.1093/bioinformatics/btaa1071
675 Duroux, L. & Welinder, K. G. The peroxidase gene family in plants: a phylogenetic overview. *J Mol Evol* **57**, 397-407 (2003). https://doi.org/10.1007/s00239-003-2489-3
676 Ryan, B. J., O'Connell, M. J. & O'Fagain, C. Consensus mutagenesis reveals that non-helical regions influence thermal stability of horseradish peroxidase. *Biochimie* **90**, 1389-1396 (2008). https://doi.org/10.1016/j.biochi.2008.04.009
677 Loughran, N. B., O'Connell, M. J., O'Connor, B. & Ó'Fágáin, C. Stability properties of an ancient plant peroxidase. *Biochimie* **104**, 156-159 (2014). https://doi.org/10.1016/j.biochi.2014.05.012
678 JEANMOUGIN, F., THOMPSON, J. D., GOUY, M., HIGGINS, D. G. & GIBSON, T. J. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403-405 (1998).
679 Ryan, B. J. & O'Fagain, C. Effects of mutations in the helix G region of horseradish peroxidase. *Biochimie* **90**, 1414-1421 (2008). https://doi.org/10.1016/j.biochi.2008.05.008
680 Luthje, S. & Martinez-Cortes, T. Membrane-Bound Class III Peroxidases: Unexpected Enzymes with Exciting Functions. *Int J Mol Sci* **19** (2018). https://doi.org/10.3390/ijms19102876
681 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012). https://doi.org/10.1093/bioinformatics/bts565
682 One Thousand Plant Transcriptomes, I. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679-685 (2019). https://doi.org/10.1038/s41586-019-1693-2
683 Hinchliff, C. E. *et al.* Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci U S A* **112**, 12764-12769 (2015). https://doi.org/10.1073/pnas.1423041112
684 Wong, Y. & Rosindell, J. Dynamic visualisation of million-tip trees: The OneZoom project. *Methods in Ecology and Evolution* **13**, 303-313 (2021). https://doi.org/10.1111/2041-210x.13766

685 UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523-D531 (2023). https://doi.org/10.1093/nar/gkac1052
686 Jiao, Y. *et al.* A genome triplication associated with early diversification of the core eudicots. *Genome Biology* **13** (2012).
687 Bremer, B. in *The Timetree of Life* (eds S. Blair Hedges & Sudhir Kumar) 0 (Oxford University Press, 2009).
688 Zallot, R., Oberg, N. & Gerlt, J. A. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **58**, 4169-4182 (2019). https://doi.org/10.1021/acs.biochem.9b00735
689 Copp, J. N., Akiva, E., Babbitt, P. C. & Tokuriki, N. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry* **57**, 4651-4662 (2018). https://doi.org/10.1021/acs.biochem.8b00473
690 Subramanian, A. R., Kaufmann, M. & Morgenstern, B. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol Biol* **3**, 6 (2008). https://doi.org/10.1186/1748-7188-3-6
691 Suchard, M. A. & Redelings, B. D. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22**, 2047-2048 (2006). https://doi.org/10.1093/bioinformatics/btl175
692 Zhou, X., Shen, X. X., Hittinger, C. T. & Rokas, A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Mol Biol Evol* **35**, 486-503 (2018). https://doi.org/10.1093/molbev/msx302
693 Yang, Z. A Space-Time Process Model for the Evolution of DNA Sequences. *Genetics* **139**, 993-1005 (1995).
694 Goluch, T., Bogdanowicz, D., Giaro, K. & Price, S. Visual TreeCmp: Comprehensive Comparison of Phylogenetic Trees on the Web. *Methods in Ecology and Evolution* **11**, 494-499 (2020). https://doi.org/10.1111/2041-210x.13358
695 Yang, G. *et al.* Evolution of the catalytic mechanism at the dawn of the Baeyer-Villiger monooxygenases. *Cell Rep* **43**, 114130 (2024). https://doi.org/10.1016/j.celrep.2024.114130
696 Boverio, A. *et al.* Structure, mechanism, and evolution of the last step in vitamin C biosynthesis. *Nat Commun* **15**, 4158 (2024). https://doi.org/10.1038/s41467-024-48410-1
697 Capone, S. *et al.* Glyco-variant library of the versatile enzyme horseradish peroxidase. *Glycobiology* **24**, 852-863 (2014). https://doi.org/10.1093/glycob/cwu047
698 Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722-2728 (2013). https://doi.org/10.1093/bioinformatics/btt473
699 Guo, H. B. *et al.* AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* **12**, 10696 (2022). https://doi.org/10.1038/s41598-022-14382-9
700 Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* **40**, 1023-1025 (2022). https://doi.org/10.1038/s41587-021-01156-3
701 Li, Y. D. *et al.* The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* **436**, 8-11 (2009). https://doi.org/10.1016/j.gene.2009.01.015
702 Humer, D. & Spadiut, O. Improving the Performance of Horseradish Peroxidase by Site-Directed Mutagenesis. *Int J Mol Sci* **20** (2019). https://doi.org/10.3390/ijms20040916

703 Theorell, H. & Maehly, A. C. Untersuchungen an kunstlichen Peroxydasen. *Acta Chemica Scandinavica* **4**, 422-434 (1950).
704 Grigorenko, V. *et al.* New Approaches for Functional Expression of Recombinant Horseradish Peroxidase C InEscherichia Coli. *Biocatalysis and Biotransformation* **17**, 359-379 (1999). https://doi.org/10.3109/10242429909015236
705 Strickland, E. H. Circular dichroism of horseradish peroxidase and tis enzyme-substrate compounds. *Biochim Biophys Acta* **151**, 70-75 (1968).
706 Asad, S., Dabirmanesh, B., Ghaemi, N., Etezad, S. M. & Khajeh, K. Studies on the refolding process of recombinant horseradish peroxidase. *Mol Biotechnol* **54**, 484-492 (2013). https://doi.org/10.1007/s12033-012-9588-6
707 Pappa, H. S. & Cass, A. E. A step towards understanding the folding mechanism of horseradish peroxidase. Tryptophan fluorescence and circular dichroism equilibrium studies. *Eur J Biochem* **212**, 227-235 (1993). https://doi.org/10.1111/j.1432-1033.1993.tb17654.x
708 Ge, F. *et al.* Recent advances in enhanced enzyme activity, thermostability and secretion by N-glycosylation regulation in yeast. *Biotechnol Lett* **40**, 847-854 (2018). https://doi.org/10.1007/s10529-018-2526-3
709 Xu, Z., Cen, Y. K., Zou, S. P., Xue, Y. P. & Zheng, Y. G. Recent advances in the improvement of enzyme thermostability by structure modification. *Crit Rev Biotechnol* **40**, 83-98 (2020). https://doi.org/10.1080/07388551.2019.1682963
710 Nguyen, V. *et al.* Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* **355**, 289-294 (2017).
711 Perez-Jimenez, R. *et al.* Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol* **18**, 592-596 (2011). https://doi.org/10.1038/nsmb.2020
712 Rana, M. M., Natale, G. & De la Hoz Siegler, H. A greener route for smart PNIPAm microgel synthesis using a bio-based synthesis-solvent. *European Polymer Journal* **174** (2022). https://doi.org/10.1016/j.eurpolymj.2022.111311
713 Solis, D. M. & Czekanski, A. The effect of the printing temperature on 4D DLP printed pNIPAM hydrogels. *Soft Matter* **18**, 3422-3429 (2022). https://doi.org/10.1039/d2sm00201a
714 Islam, M. R. & Oyen, M. L. in *The Mechanics of Hydrogels* (eds Hua Li & Vadim Silberschmidt) 1-24 (Woodhead Publishing, 2022).
715 Drozdov, A. D. & Christiansen, J. D. Tension-compression asymmetry in the mechanical response of hydrogels. *J Mech Behav Biomed Mater* **110**, 103851 (2020). https://doi.org/10.1016/j.jmbbm.2020.103851
716 Zhang, Y. R. *et al.* Features of the volume change and a new constitutive equation of hydrogels under uniaxial compression. *J Mech Behav Biomed Mater* **85**, 181-187 (2018). https://doi.org/10.1016/j.jmbbm.2018.06.004
717 Nafar Dastgerdi, J. *et al.* Comprehensive characterisation of the compressive behaviour of hydrogels using a new modelling procedure and redefining compression testing. *Materials Today Communications* **28** (2021). https://doi.org/10.1016/j.mtcomm.2021.102518
718 Mooney, M. A Theory of Large Elastic Deformation. *Journal of Applied Physics* **11**, 582-592 (1940). https://doi.org/10.1063/1.1712836
719 Rivlin, R. S. LARGE ELASTIC DEFORMATIONS OF ISOTROPIC MATERIALS IV. FURTHER DEVELOPMENTS OF THE GENERAL THEORY. **241**, 380-397 (1948).

720 Liu, W. *et al.* Enhancing Temperature Responsiveness of PNIPAM Through 3D-Printed Hierarchical Porosity. *Advanced Functional Materials* (2024). https://doi.org/10.1002/adfm.202403794
721 Kouba, P. *et al.* Machine Learning-Guided Protein Engineering. *ACS Catal* **13**, 13863-13895 (2023). https://doi.org/10.1021/acscatal.3c02743
722 Weinstein, J. Y. *et al.* Designed active-site library reveals thousands of functional GFP variants. *Nat Commun* **14**, 2890 (2023). https://doi.org/10.1038/s41467-023-38099-z
723 Liechty, E. T. *et al.* Analysis of neutral mutational drift in an allosteric enzyme. *Protein Sci* **32**, e4719 (2023). https://doi.org/10.1002/pro.4719
724 Toft, C. & Fares, M. A. Structural calibration of the rates of amino acid evolution in a search for Darwin in drifting biological systems. *Mol Biol Evol* **27**, 2375-2385 (2010). https://doi.org/10.1093/molbev/msq123
725 Krainer, F. W. *et al.* Optimizing cofactor availability for the production of recombinant heme peroxidase in Pichia pastoris. *Microb Cell Fact* **14**, 4 (2015). https://doi.org/10.1186/s12934-014-0187-z
726 Hutchins, G. H. *et al.* An expandable, modular de novo protein platform for precision redox engineering. *Proc Natl Acad Sci U S A* **120**, e2306046120 (2023). https://doi.org/10.1073/pnas.2306046120
727 Fiege, K., Querebillo, C. J., Hildebrandt, P. & Frankenberg-Dinkel, N. Improved Method for the Incorporation of Heme Cofactors into Recombinant Proteins Using Escherichia coli Nissle 1917. *Biochemistry* **57**, 2747-2755 (2018). https://doi.org/10.1021/acs.biochem.8b00242
728 Scotese, C. R., Song, H., Mills, B. J. W. & van der Meer, D. G. Phanerozoic paleotemperatures: The earth's changing climate during the last 540 million years. *Earth-Science Reviews* **215** (2021). https://doi.org/10.1016/j.earscirev.2021.103503
729 Lubna *et al.* The dynamic history of gymnosperm plastomes: Insights from structural characterization, comparative analysis, phylogenomics, and time divergence. *Plant Genome* **14**, e20130 (2021). https://doi.org/10.1002/tpg2.20130
730 Yang, L. *et al.* Phylogenomic Insights into Deep Phylogeny of Angiosperms Based on Broad Nuclear Gene Sampling. *Plant Commun* **1**, 100027 (2020). https://doi.org/10.1016/j.xplc.2020.100027
731 Murat, F., Armero, A., Pont, C., Klopp, C. & Salse, J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet* **49**, 490-496 (2017). https://doi.org/10.1038/ng.3813
732 Magallon, S., Gomez-Acevedo, S., Sanchez-Reyes, L. L. & Hernandez-Hernandez, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol* **207**, 437-453 (2015). https://doi.org/10.1111/nph.13264
733 Walters, S. Horseradish: A Neglected and Underutilized Plant Species for Improving Human Health. *Horticulturae* **7** (2021). https://doi.org/10.3390/horticulturae7070167
734 Shen, F., Xu, S., Shen, Q., Bi, C. & Lysak, M. A. The allotetraploid horseradish genome provides insights into subgenome diversification and formation of critical traits. *Nat Commun* **14**, 4102 (2023). https://doi.org/10.1038/s41467-023-39800-y
735 Wheeler, L. C., Anderson, J. A., Morrison, A. J., Wong, C. E. & Harms, M. J. Conservation of Specificity in Two Low-Specificity Proteins. *Biochemistry* **57**, 684-695 (2018). https://doi.org/10.1021/acs.biochem.7b01086
736 Kohri, M. Development of HRP-mediated enzymatic polymerization under heterogeneous conditions for the preparation of functional particles. *Polymer Journal* **46**, 373-380 (2014). https://doi.org/10.1038/pj.2014.13

737 Belluati, A. *et al.* Self-decorating cells via surface-initiated enzymatic controlled radical polymerization. *Nanoscale* **15**, 19486-19492 (2023). https://doi.org/10.1039/d3nr04008a
738 Liu, A. P. *et al.* The living interface between synthetic biology and biomaterial design. *Nat Mater* **21**, 390-397 (2022). https://doi.org/10.1038/s41563-022-01231-3
739 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003). https://doi.org/10.1101/gr.1239303
740 Miles, A. J., Ramalli, S. G. & Wallace, B. A. DichroWeb, a website for calculating protein secondary structure from circular dichroism spectroscopic data. *Protein Sci* **31**, 37-46 (2022). https://doi.org/10.1002/pro.4153
741 Childs, R. & Bardsley, W. The Steady-State Kinetics of Peroxidase with 2,2'-Azino-di-(3-ethylbenzthiazoline-6-sulphonic acid) as Chromogen. *Biochem. J.* **145**, 93-103 (1975).
742 Tamura, M., Asakura, T. & Yonetani, T. Heme-modification studies on horseradish peroxidase. *Biochim Biophys Acta* **268**, 292-304 (1972).
743 Marshall, R. D. The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins.
744 Mellquist, J. L., Kasturi, L., Spitalnik, S. L. & Shakin-Eshleman, S. H. The Amino Acid Following an Asn-X-Ser/Thr Sequon Is an Important Determinant of N-Linked Core Glycosylation Efficiency. *Biochemistry* **37**, 6833-6837 (1998).
745 Siniscalchi, C. M. *et al.* Fagalean phylogeny in a nutshell: Chronicling the 1 diversification history of Fagales. (2023). https://doi.org/10.1101/2023.03.06.531381
746 Muller, D., Hauer, J., Schottner, K., Fritzsche, P. & Weinert, D. Seasonal adaptation of dwarf hamsters (Genus Phodopus): differences between species and their geographic origin. *J Comp Physiol B* **185**, 917-930 (2015). https://doi.org/10.1007/s00360-015-0926-4
747 Wang, X. *et al.* Smart drug delivery systems for precise cancer therapy. *Acta Pharm Sin B* **12**, 4098-4121 (2022). https://doi.org/10.1016/j.apsb.2022.08.013
748 Khadka, B., Lee, B. & Kim, K. T. Drug Delivery Systems for Personal Healthcare by Smart Wearable Patch System. *Biomolecules* **13** (2023). https://doi.org/10.3390/biom13060929
749 Frijlink, H., Lagarce, F., Touw, D. & Woerdenbag, H. in *Practical Pharmaceutics: An International Guideline for the Preparation, Care and Use of Medicinal Products* (eds Paul Le Brun *et al.*) 67-91 (Springer International Publishing, 2023).
750 Urs, D., Madesh, A., Mahmood, K., Sreeharsha, N. & Dharmappa, K. K. in *Novel Nanostructured Materials for Electrochemical Bio-Sensing Applications* (ed Jamballi G. Manjunatha) 367-378 (Elsevier, 2024).
751 Tang, X. *et al.* Current trends in biosensors for biotoxins (mycotoxins, marine toxins, and bacterial food toxins):principles, application, and perspective. *TrAC Trends in Analytical Chemistry* **165** (2023). https://doi.org/10.1016/j.trac.2023.117144
752 Geleta, G. S., Zhao, Z. & Wang, Z. Electrochemical Biosensors for Detecting Microbial Toxins by Graphene-Based Nanocomposites. *Journal of Analysis and Testing* **2**, 20-25 (2018). https://doi.org/10.1007/s41664-018-0051-y
753 Tokuyama, H., Aoyagi, R., Fujita, K., Maekawa, Y. & Riya, S. Ethanol fermentation using macroporous monolithic hydrogels as yeast cell scaffolds. *Reactive and Functional Polymers* **169** (2021). https://doi.org/10.1016/j.reactfunctpolym.2021.105075
754 Altin-Yavuzarslan, G., Sadaba, N., Brooks, S. M., Alper, H. S. & Nelson, A. Engineered Living Material Bioreactors with Tunable Mechanical Properties using

Vat Photopolymerization. *Small* **20**, e2306564 (2024). https://doi.org/10.1002/smll.202306564
755 Schwab, A. *et al.* Printability and Shape Fidelity of Bioinks in 3D Bioprinting. *Chem Rev* **120**, 11028-11055 (2020). https://doi.org/10.1021/acs.chemrev.0c00084
756 Smith, R. S. H. *et al.* Hybrid Living Materials: Digital Design and Fabrication of 3D Multimaterial Structures with Programmable Biohybrid Surfaces. *Advanced Functional Materials* **30** (2019). https://doi.org/10.1002/adfm.201907401
757 Butelmann, T. *et al.* Metabolism Control in 3D-Printed Living Materials Improves Fermentation. *ACS Appl Bio Mater* **4**, 7195-7203 (2021). https://doi.org/10.1021/acsabm.1c00754
758 Cui, Z., Feng, Y., Liu, F., Jiang, L. & Yue, J. 3D Bioprinting of Living Materials for Structure-Dependent Production of Hyaluronic Acid. *ACS Macro Lett* **11**, 452-459 (2022). https://doi.org/10.1021/acsmacrolett.2c00037
759 Saha, A. *et al.* Additive Manufacturing of Catalytically Active Living Materials. *ACS Appl Mater Interfaces* **10**, 13373-13380 (2018). https://doi.org/10.1021/acsami.8b02719
760 Li, Y. *et al.* Biocatalytic living materials built by compartmentalized microorganisms in annealable granular hydrogels. *Chemical Engineering Journal* **445** (2022). https://doi.org/10.1016/j.cej.2022.136822
761 Park, T. G. & Hoffman, A. S. Preparation of large, uniform size temperature-sensitive hydrogel beads. *Journal of Polymer Science Part A: Polymer Chemistry* **30**, 505-507 (1992). https://doi.org/10.1002/pola.1992.080300318
762 Dumitriu, R. P. *et al.* Biocompatible and biodegradable alginate/poly(N-isopropylacrylamide) hydrogels for sustained theophylline release. *Journal of Applied Polymer Science* **131** (2014). https://doi.org/10.1002/app.40733
763 Zheng, W. J., An, N., Yang, J. H., Zhou, J. & Chen, Y. M. Tough Al-alginate/poly(N-isopropylacrylamide) hydrogel with tunable LCST for soft robotics. *ACS Appl Mater Interfaces* **7**, 1758-1764 (2015). https://doi.org/10.1021/am507339r
764 de Moura, M. R. *et al.* Porous alginate-Ca2+ hydrogels interpenetrated with PNIPAAm networks: Interrelationship between compressive stress and pore morphology. *European Polymer Journal* **41**, 2845-2852 (2005). https://doi.org/10.1016/j.eurpolymj.2005.06.007
765 Lanzalaco, S., Mingot, J., Torras, J., Alemán, C. & Armelin, E. Recent Advances in Poly(N-isopropylacrylamide) Hydrogels and Derivatives as Promising Materials for Biomedical and Engineering Emerging Applications. *Advanced Engineering Materials* **25** (2023). https://doi.org/10.1002/adem.202201303
766 Haq, M. A., Su, Y. & Wang, D. Mechanical properties of PNIPAM based hydrogels: A review. *Mater Sci Eng C Mater Biol Appl* **70**, 842-855 (2017). https://doi.org/10.1016/j.msec.2016.09.081
767 Fischer, F. G. & Dorfel, H. Die Polyuronsäuren der Braunalgen (Kohlenhydrate der Algen I). *Hoppe Seylers Z Physiol Chem* **302**, 186-203 (1955).
768 Schweiger, R. G. Acetylation of Alginic Acid. II. Reaction of Algin Acetates with Calcium and Other Divalent Ions. *The Journal of Organic Chemistry* **27**, 1789-1791 (1962).
769 Petrusic, S. *et al.* Development and characterization of thermosensitive hydrogels based on poly(N-isopropylacrylamide) and calcium alginate. *Journal of Applied Polymer Science* **124**, 890-903 (2011). https://doi.org/10.1002/app.35122
770 Yang, C. H. *et al.* Strengthening alginate/polyacrylamide hydrogels using various multivalent cations. *ACS Appl Mater Interfaces* **5**, 10418-10422 (2013). https://doi.org/10.1021/am403966x

771 Park, T. G. & Choi, H. K. Thermally induced core-shell type hydrogel beads having interpenetrating polymer network (IPN) structure. *Macromolecular Rapid Communications* **19**, 167-172 (1998). https://doi.org/10.1002/(sici)1521-3927(19980401)19:4<167::Aid-marc167>3.0.Co;2-g

772 Lee, S. B. *et al.* Preparation of alginate/poly(N-isopropylacrylamide) semi-interpenetrating and fully interpenetrating polymer network hydrogels with γ-ray irradiation and their swelling behaviors. *Journal of Applied Polymer Science* **100**, 4439-4446 (2006). https://doi.org/10.1002/app.23726

773 Zhang, H. *et al.* Thermal-responsive poly(N-isopropyl acrylamide)/sodium alginate hydrogels: preparation, swelling behaviors, and mechanical properties. *Colloid and Polymer Science* **294**, 1959-1967 (2016). https://doi.org/10.1007/s00396-016-3951-2

774 Cui, Y. *et al.* Preparation of multifunctional composite materials based on PEDOT via enzymatic cascade polymerization. *Surfaces and Interfaces* **39** (2023). https://doi.org/10.1016/j.surfin.2023.102961

775 Kim, B. Y., Lee, Y., Son, J. Y., Park, K. M. & Park, K. D. Dual Enzyme-Triggered In Situ Crosslinkable Gelatin Hydrogels for Artificial Cellular Microenvironments. *Macromol Biosci* **16**, 1570-1576 (2016). https://doi.org/10.1002/mabi.201600312

776 Hofrichter, M. & Ullrich, R. Heme-thiolate haloperoxidases: versatile biocatalysts with biotechnological and environmental significance. *Appl Microbiol Biotechnol* **71**, 276-288 (2006). https://doi.org/10.1007/s00253-006-0417-3

777 Hofrichter, M. *et al.* Peroxide-Mediated Oxygenation of Organic Compounds by Fungal Peroxygenases. *Antioxidants (Basel)* **11** (2022). https://doi.org/10.3390/antiox11010163

778 Struwe, H. *et al.* Expanding the "Terpenome": Applications of Unspecific Peroxygenases (UPOs) in Oxidations of Unnatural Terpenoids. *ChemCatChem* (2024). https://doi.org/10.1002/cctc.202401414

779 Martin-Diaz, J., Paret, C., Garcia-Ruiz, E., Molina-Espeja, P. & Alcalde, M. Shuffling the Neutral Drift of Unspecific Peroxygenase in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* **84** (2018). https://doi.org/https://doi.org/10.1128/AEM.00808-18.

780 Pullmann, P. *et al.* A modular two yeast species secretion system for the production and preparative application of unspecific peroxygenases. *Commun Biol* **4**, 562 (2021). https://doi.org/10.1038/s42003-021-02076-3

781 Zamocky, M. *et al.* Independent evolution of four heme peroxidase superfamilies. *Arch Biochem Biophys* **574**, 108-119 (2015). https://doi.org/10.1016/j.abb.2014.12.025

782 Hofrichter, M. & Ullrich, R. Oxidations catalyzed by fungal peroxygenases. *Curr Opin Chem Biol* **19**, 116-125 (2014). https://doi.org/10.1016/j.cbpa.2014.01.015

783 Kellner, H. *et al.* Widespread occurrence of expressed fungal secretory peroxidases in forest soils. *PLoS One* **9**, e95557 (2014). https://doi.org/10.1371/journal.pone.0095557

784 Ramirez-Escudero, M. *et al.* Structural Insights into the Substrate Promiscuity of a Laboratory-Evolved Peroxygenase. *ACS Chem Biol* **13**, 3259-3268 (2018). https://doi.org/10.1021/acschembio.8b00500

785 Sundaramoorthy, M., Terner, J. & Poulos, T. L. The crystal structure of chloroperoxidase: a heme peroxidase-cytochrome P450 functional hybrid. *Curr Biol* **3**, 1367-1377 (1995).

786 Piontek, K. *et al.* Structural basis of substrate conversion in a new aromatic peroxygenase: cytochrome P450 functionality with benefits. *J Biol Chem* **288**, 34767-34776 (2013). https://doi.org/10.1074/jbc.M113.514521

787 Molina-Espeja, P., Beltran-Nogal, A., Alfuzzi, M. A., Guallar, V. & Alcalde, M. Mapping Potential Determinants of Peroxidative Activity in an Evolved Fungal Peroxygenase from Agrocybe aegerita. *Front Bioeng Biotechnol* **9**, 741282 (2021). https://doi.org/10.3389/fbioe.2021.741282
788 Stenner, A. *et al.* Chemo-Enzymatic One-Pot Oxidation of Cyclohexane via in-situ H2O2 Production over Supported AuPdPt Catalysts. *ChemCatChem* **15** (2023). https://doi.org/10.1002/cctc.202300162
789 Wood, T. K. Molecular approaches in bioremediation. *Curr Opin Biotechnol* **19**, 572-578 (2008). https://doi.org/10.1016/j.copbio.2008.10.003
790 Rahman, M.-u. *et al.* in *Handbook of Bioremediation* (eds Mirza Hasanuzzaman & Majeti Narasimha Vara Prasad) 437-444 (Academic Press, 2021).
791 Kleiner-Grote, G. R. M., Risse, J. M. & Friehs, K. Secretion of recombinant proteins from E. coli. *Eng Life Sci* **18**, 532-550 (2018). https://doi.org/10.1002/elsc.201700200
792 Incir, I. & Kaplan, O. Escherichia coli as a versatile cell factory: Advances and challenges in recombinant protein production. *Protein Expr Purif* **219**, 106463 (2024). https://doi.org/10.1016/j.pep.2024.106463
793 Schutz, A. *et al.* A concise guide to choosing suitable gene expression systems for recombinant protein production. *STAR Protoc* **4**, 102572 (2023). https://doi.org/10.1016/j.xpro.2023.102572
794 Vogl, T., Hartner, F. S. & Glieder, A. New opportunities by synthetic biology for biopharmaceutical production in Pichia pastoris. *Curr Opin Biotechnol* **24**, 1094-1101 (2013). https://doi.org/10.1016/j.copbio.2013.02.024
795 Ha, S., Wang, Y. & Rustandi, R. R. Biochemical and biophysical characterization of humanized IgG1 produced in Pichia pastoris. *MAbs* **3**, 453-460 (2011). https://doi.org/10.4161/mabs.3.5.16891
796 *Yeast*. 2 edn, (2012).
797 Guilliermond, A. *Zygosaccharomyces pastori, nouvelle espèce de levures à copulation hétérogamique*. (Société mycologique de France, 1920).
798 PHAFF, H. J., MILLER, M. W. & SHIFRINE, M. The taxonomy of yeasts isolated from *Drosophila* in the Yosemite region in California. *Antonie van Leeuwenhoek* **22**, 144-161 (1956).
799 Yamada, Y., Matsuda, M., Maeda, K. & Mikata, K. The phylogenetic relationships of methanol-assimilating yeasts based on the partial sequences of 18S and 26S ribosomal RNAs: the proposal of Komagataella gen. nov. (Saccharomycetaceae). *Biosci Biotechnol Biochem* **59**, 439-444 (1995).
800 Naumov, G. I., Naumova, E. S. & Boundy-Mills, K. L. Description of Komagataella mondaviorum sp. nov., a new sibling species of Komagataella (Pichia) pastoris. *Antonie Van Leeuwenhoek* **111**, 1197-1207 (2018). https://doi.org/10.1007/s10482-018-1028-6
801 Ahmad, M., Hirz, M., Pichler, H. & Schwab, H. Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production. *Appl Microbiol Biotechnol* **98**, 5301-5317 (2014). https://doi.org/10.1007/s00253-014-5732-5
802 Young, C. L. & Robinson, A. S. Protein folding and secretion: mechanistic insights advancing recombinant protein production in S. cerevisiae. *Curr Opin Biotechnol* **30**, 168-177 (2014). https://doi.org/10.1016/j.copbio.2014.06.018
803 Ou, Y. *et al.* Bioprinting microporous functional living materials from protein-based core-shell microgels. *Nat Commun* **14**, 322 (2023). https://doi.org/10.1038/s41467-022-35140-5

804 Karbalaei, M., Rezaee, S. A. & Farsiani, H. Pichia pastoris: A highly successful expression system for optimal synthesis of heterologous proteins. *J Cell Physiol* **235**, 5867-5881 (2020). https://doi.org/10.1002/jcp.29583
805 Juturu, V. & Wu, J. C. Heterologous Protein Expression in Pichia pastoris: Latest Research Progress and Applications. *Chembiochem* **19**, 7-21 (2018). https://doi.org/10.1002/cbic.201700460
806 Fischer, J. E. & Glieder, A. Current advances in engineering tools for Pichia pastoris. *Curr Opin Biotechnol* **59**, 175-181 (2019). https://doi.org/10.1016/j.copbio.2019.06.002
807 Krainer, F. W. *et al.* Recombinant protein expression in Pichia pastoris strains with an engineered methanol utilization pathway. *Microbial Cell Factories* **11** (2012).
808 *Yeast Metabolic Engineering*. (Springer Protocols, 2014).
809 Monterrey, D. T., Menés-Rubio, A., Keser, M., Gonzalez-Perez, D. & Alcalde, M. Unspecific peroxygenases: The pot of gold at the end of the oxyfunctionalization rainbow? *Current Opinion in Green and Sustainable Chemistry* **41** (2023). https://doi.org/10.1016/j.cogsc.2023.100786
810 PORSTMANN, B., PORSTMANN, T., GAEDE, D., NUGEL, E. & EGGER, E. TEMPERATURE DEPENDENT RISE IN ACTIVITY OF HORSERADISH PEROXIDASE CAUSED BY NON-IONIC DETERGENTS AND ITS USE IN ENZYME-IMMUNOASSAY. *Clinica Chimica Acta* **109**, 175-181 (1981).
811 Gallati, H. Peroxidase aus Meerrettich: Kinetische Studien sowie Optimierung der Aktivitätsbestimmung mit den Substraten H2O2 und 2,2'-Azino-di-(3-ethyl-benzthiazolinsulfonsäure-(6)) (ABTS). *Clinical Chemistry and Laboratory Medicine* **17**, 1-7 (1979).
812 Gnugge, R. & Rudolf, F. Saccharomyces cerevisiae Shuttle vectors. *Yeast* **34**, 205-221 (2017). https://doi.org/10.1002/yea.3228
813 Bretthauer, R. K. & Castellino, F. J. Glycosylation of Pichia pastoris -derived proteins. *Biotechnology and Applied Biochemistry* **30**, 193-200 (2010). https://doi.org/10.1111/j.1470-8744.1999.tb00770.x
814 Sickmann, A., Mreyen, M. & Meyer, H. E. Identification of modified proteins by mass spectrometry. *IUBMB Life* **54**, 51-57 (2002). https://doi.org/10.1080/15216540214314
815 Nguyen, D. N., Becker, G. W. & Riggin, R. M. Protein mass spectrometry: applications to analytical biotechnology. *Journal of Chromatography A* **705**, 21-45 (1995).
816 Jenkins, N. & Curling, E. M. A. Glycosylation of recombinant proteins: Problems and prospects. *Enzyme Microb. Technol.* **16**, 354-364 (1994).
817 Olucha, F., Martinez-Garcia, F. & Lopez-Garcia, C. A new stabilizing agent for the tetramethyl benzidine (TMB) reaction product in the histochemical detection of horseradish peroxidase (HRP). *Journal of Neuroscience Methods* **13**, 131-138 (1985).
818 Stanciuc, N., Aprodu, I., Ionita, E., Bahrim, G. & Rapeanu, G. Exploring the process-structure-function relationship of horseradish peroxidase through investigation of pH- and heat induced conformational changes. *Spectrochim Acta A Mol Biomol Spectrosc* **147**, 43-50 (2015). https://doi.org/10.1016/j.saa.2015.03.023
819 Chattopadhyay, K. & Mazumdar, S. Structural and Conformational Stability of Horseradish Peroxidase: Effect of Temperature and pH. *Biochemistry* **39**, 263-270 (2000).
820 Adediran, S. A. Effect of pH on the Formation of Compounds II and Ill of Horseradish Peroxidase. *Journal of Protein Chemistry* **3**, 437-444 (1984).

821 Chang, C. S., Yamazaki, I., Sinclair, R., Khalid, S. & Powers, L. pH Dependence of the Active Site of Horseradish Peroxidase Compound II. *Biochemistry* **32**, 923-928 (1993).
822 Dumitriu, R. P., Mitchell, G. R. & Vasile, C. Multi-responsive hydrogels based on N-isopropylacrylamide and sodium alginate. *Polymer International* **60**, 222-233 (2010). https://doi.org/10.1002/pi.2929
823 Hurtado, A., Aljabali, A. A. A., Mishra, V., Tambuwala, M. M. & Serrano-Aroca, A. Alginate: Enhancement Strategies for Advanced Applications. *Int J Mol Sci* **23** (2022). https://doi.org/10.3390/ijms23094486
824 Gallati, H. & Brodbeck, H. Peroxidase aus Meerrettich: Kinetische Studien und Optimierung der Aktivitätsbestimmung mit den Substraten H2O2 und o-Phenylendiamin. *Clinical Chemistry and Laboratory Medicine* **20**, 221-225 (1982).
825 Ribeiro, A. *et al.* Assessing bioink shape fidelity to aid material development in 3D bioprinting. *Biofabrication* **10**, 014102 (2017). https://doi.org/10.1088/1758-5090/aa90e2
826 Choi, E. J., Ha, S., Lee, J., Premkumar, T. & Song, C. UV-mediated synthesis of pNIPAM-crosslinked double-network alginate hydrogels: Enhanced mechanical and shape-memory properties by metal ions and temperature. *Polymer* **149**, 206-212 (2018). https://doi.org/10.1016/j.polymer.2018.06.080
827 Kim, J. H., Lee, S. B., Kim, S. J. & Lee, Y. M. Rapid temperature/pH response of porous alginate-g-poly(N-isopropylacrylamide) hydrogels. *Polymer* **43**, 7549–7558 (2002).
828 Owh, C. *et al.* Bottom-up design of hydrogels for programmable drug release. *Biomater Adv* **141**, 213100 (2022). https://doi.org/10.1016/j.bioadv.2022.213100
829 Matsubara, S. *et al.* in *Creep in Structures VI.* (eds Holm Altenbach & Konstantin Naumenko) 175-204 (Springer Nature Switzerland).
830 Klemperer, R. G. *3D Printing Enzyme Mediated Interpenetrating-Network Biohybrid Materials with Shape Changing Properties* PhD thesis, University of Bristol, (2021).
831 Liu, Y. *et al.* Comparison of Bulk- vs Layer-by-Layer-Cured Stimuli-Responsive PNIPAM–Alginate Hydrogel Dynamic Viscoelastic Property Response via Embedded Sensors. *ACS Applied Polymer Materials* **4**, 5596-5607 (2022). https://doi.org/10.1021/acsapm.2c00634
832 Guilherme, M. R. *et al.* Novel thermo-responsive membranes composed of interpenetrated polymer networks of alginate-Ca2+ and poly(N-isopropylacrylamide). *Polymer* **46**, 2668-2674 (2005). https://doi.org/10.1016/j.polymer.2005.01.082
833 Drury, J. L., Dennis, R. G. & Mooney, D. J. The tensile properties of alginate hydrogels. *Biomaterials* **25**, 3187-3199 (2004). https://doi.org/10.1016/j.biomaterials.2003.10.002
834 Anseth, K. S., Bowman, C. N. & Brannon-Peppas, L. Mechanical properties of hydrogels and their experimental determination. *Biomaterials* **17**, 1647-1657 (1996).
835 Sasaki, S. & Koga, S. Slow Relaxation of the Elastic N-Isopropylacrylamide Gel. *Macromolecules* **35**, 857-860 (2002).
836 Martínez, D. *et al.* Complete sucrose hydrolysis by heat-killed recombinant Pichia pastoris cells entrapped in calcium alginate. *Microbial Cell Factories* **13** (2014).
837 Sheppard, C. J. R. The Development of Microscopy for Super-Resolution: Confocal Microscopy, and Image Scanning Microscopy. *Applied Sciences* **11** (2021). https://doi.org/10.3390/app11198981
838 Mill, P. J. The Nature of the Interactions between Flocculent Cells in the Flocculation of Saccharomyces cerevisiae. *J. gen. Microbiol.* **35**, 61-68 (1964).

839 Dungrawala, H. *et al.* Identification of new cell size control genes in S. cerevisiae. *Cell Division* **7** (2012).
840 Pekarsky, A. *et al.* Production of a recombinant peroxidase in different glyco-engineered Pichia pastoris strains: a morphological and physiological comparison. *Microb Cell Fact* **17**, 183 (2018). https://doi.org/10.1186/s12934-018-1032-6
841 Barelle, C. J. *et al.* GFP as a quantitative reporter of gene regulation in Candida albicans. *Yeast* **21**, 333-340 (2004). https://doi.org/10.1002/yea.1099
842 Portle, S. *et al.* Cell population heterogeneity in expression of a gene-switching network with fluorescent markers of different half-lives. *J Biotechnol* **128**, 362-375 (2007). https://doi.org/10.1016/j.jbiotec.2006.09.026
843 Corish, P. & Tyler-Smith, C. Attenuation of green fluorescent protein half-life in mammalian cells. *Protein Engineering* **12**, 1035-1040 (1999).
844 Carmona-Gutierrez, D. *et al.* Guidelines and recommendations on yeast cell death nomenclature. *Microb Cell* **5**, 4-31 (2018). https://doi.org/10.15698/mic2018.01.607
845 Falcone, C. & Mazzoni, C. External and internal triggers of cell death in yeast. *Cell Mol Life Sci* **73**, 2237-2250 (2016). https://doi.org/10.1007/s00018-016-2197-y
846 Munoz, A. J., Wanichthanarak, K., Meza, E. & Petranovic, D. Systems biology of yeast cell death. *FEMS Yeast Res* **12**, 249-265 (2012). https://doi.org/10.1111/j.1567-1364.2011.00781.x
847 Grosfeld, E. V. *et al.* A Systematic Survey of Characteristic Features of Yeast Cell Death Triggered by External Factors. *J Fungi (Basel)* **7** (2021). https://doi.org/10.3390/jof7110886
848 Ruckenstuhl, C. *et al.* The Warburg effect suppresses oxidative stress induced apoptosis in a yeast model for cancer. *PLoS One* **4**, e4592 (2009). https://doi.org/10.1371/journal.pone.0004592
849 Kuo, S. C. & Yamamoto, S. in *Methods in Cell Biology* Vol. 11 (ed David M. Prescott) Ch. 8, 169-183 (1975).
850 George, J. D. *et al.* Evaluation of the Developmental Toxicity of Methacrylamide and N,N-Methylenebisacrylamide in Swiss Mice. *Toxicological Sciences* **46**, 124-133 (1998).
851 Chapin, R. E., Fail, P. A., George, J. D. & Grizzle, T. B. The Reproductive and Neural Toxicities of Acrylamide and Three Analogues in Swiss Mice, Evaluated Using the Continuous Breeding Protocol. *Toxicological Sciences* **27**, 9-24 (1995). https://doi.org/10.1093/toxsci/27.1.9
852 Gouzy-Olmos, M. *et al.* Statistical Improvement of Batch Culture with Immobilized Pichia pastoris Cells for rPOXA 1B Laccase Production. *American Journal of Biochemistry and Biotechnology* **14**, 88-107 (2018). https://doi.org/10.3844/ajbbsp.2018.88.107
853 Perez, E. R. *et al.* Fructooligosaccharides production by immobilized Pichia pastoris cells expressing Schedonorus arundinaceus sucrose:sucrose 1-fructosyltransferase. *J Ind Microbiol Biotechnol* **48** (2021). https://doi.org/10.1093/jimb/kuab036
854 Martínez, D. *et al.* Scaling-up batch conditions for efficient sucrose hydrolysis catalyzed by an immobilized recombinant Pichia pastoris cells in a stirrer tank reactor. *Electronic Journal of Biotechnology* **25**, 39-42 (2017). https://doi.org/10.1016/j.ejbt.2016.11.003
855 Wloch-Salamon, D. M. & Bem, A. E. Types of cell death and methods of their detection in yeast Saccharomyces cerevisiae. *J Appl Microbiol* **114**, 287-298 (2013). https://doi.org/10.1111/jam.12024

856 JOHNSTON, G. C., EHRHARDT, C. W., LORINCZ, A. & CARTER, B. L. A. Regulation of Cell Size in the Yeast Saccharomyces cerevisiae. *Journal of Bacteriology* **137**, 1-5 (1979).
857 Lorincz, A. & Carter, B. Control of Cell Size at Bud Initiation in Saccharomyces cerevisiae. *Journal of General Microbiology* **113**, 287-295 (1979).
858 Diaz Arias, C. A. *et al.* Influence of carbon source on cell size and production of anti LDL (-) single-chain variable fragment by a recombinant Pichia pastoris strain. *Mol Biol Rep* **46**, 3257-3264 (2019). https://doi.org/10.1007/s11033-019-04785-9
859 Tao, Z. *et al.* Yeast Extract: Characteristics, Production, Applications and Future Perspectives. *J Microbiol Biotechnol* **33**, 151-166 (2023). https://doi.org/10.4014/jmb.2207.07057
860 Kapteyn, J. C., Ende, H. V. D. & Klis, F. M. The contribution of cell wall proteins to the organization of the yeast cell wall. *Biochimica et Biophysica Acta* **1426**, 373-383 (1999).
861 MILLARD, P. J., ROTH, B. L., THI, H.-P. T., YUE, S. T. & HAUGLAND, R. P. Development of the FUN-1 Family of Fluorescent Probes for Vacuole Labeling and Viability Testing of Yeasts. *Applied and Environmental Microbiology* **63**, 2897–2905 (1997).
862 Yamazaki, Y. & Kono, K. Clathrin-mediated trafficking of phospholipid flippases is required for local plasma membrane/cell wall damage repair in budding yeast. *Biochem Biophys Res Commun* **606**, 156-162 (2022). https://doi.org/10.1016/j.bbrc.2022.03.129
863 de Nadal, E. & Posas, F. The HOG pathway and the regulation of osmoadaptive responses in yeast. *FEMS Yeast Res* **22** (2022). https://doi.org/10.1093/femsyr/foac013
864 Jin, X., Chen, M., Coldea, T. E., Yang, H. & Zhao, H. Protective effects of peptides on the cell wall structure of yeast under osmotic stress. *Appl Microbiol Biotechnol* **106**, 7051-7061 (2022). https://doi.org/10.1007/s00253-022-12207-3
865 Pogrányi, B., Mielke, T., Cartwright, J., Unsworth, W. P. & Grogan, G. Selective Oxidations of Toluenes and Benzyl Alcohols by an Unspecific Peroxygenase (UPO). *ChemCatChem* (2024). https://doi.org/10.1002/cctc.202400702
866 Xin, L. *et al.* Immobilization of yeast cells on hydrogel carriers obtained by radiation-induced polymerization. *Radiat. Phys. Chem.* **40**, 579-584 (1992).
867 Maslanka, R., Kwolek-Mirek, M. & Zadrag-Tecza, R. Autofluorescence of yeast Saccharomyces cerevisiae cells caused by glucose metabolism products and its methodological implications. *J Microbiol Methods* **146**, 55-60 (2018). https://doi.org/10.1016/j.mimet.2018.01.017
868 Huang, P. *et al.* Evaluating Protein Engineering Thermostability Prediction Tools Using an Independently Generated Dataset. *ACS Omega* **5**, 6487-6493 (2020). https://doi.org/10.1021/acsomega.9b04105
869 Hei, D. J. & Clark, D. S. Estimation of melting curves from enzymatic activity-temperature profiles. *Biotechnol Bioeng* **42**, 1245-1251 (1993). https://doi.org/10.1002/bit.260421015
870 Adrus, N. & Ulbricht, M. Rheological studies on PNIPAAm hydrogel synthesis via in situ polymerization and on resulting viscoelastic properties. *Reactive and Functional Polymers* **73**, 141-148 (2013). https://doi.org/10.1016/j.reactfunctpolym.2012.08.015
871 Kumar, K. *et al.* Manipulation of Thermoresponsive Polymers Using Biomolecules. *ACS Applied Polymer Materials* **5**, 3181-3200 (2023). https://doi.org/10.1021/acsapm.2c02162

872 Johnston, T. G. *et al.* in *Micro- and Nanotechnology Sensors, Systems, and Applications XI* (2019).
873 Ludovico, P., Madeo, F. & Silva, M. Yeast programmed cell death: an intricate puzzle. *IUBMB Life* **57**, 129-135 (2005). https://doi.org/10.1080/15216540500090553
874 Madeo, F., Fröhlich, E. & Fröhlich, K.-U. A Yeast Mutant Showing Diagnostic Markers of Early and Late Apoptosis. *Journal of Cell Biology* **139**, 729-734 (1997).
875 Tang, T. C. *et al.* Hydrogel-based biocontainment of bacteria for continuous sensing and computation. *Nat Chem Biol* **17**, 724-731 (2021). https://doi.org/10.1038/s41589-021-00779-6
876 Calinescu, I., Chipurici, P., Alexandrescu, E. & Trifan, A. Saccharomyces cerevisiae yeast immobilized on marrow stem sunflower and polyacrylamide hydrogels. *Open Chemistry* **12**, 851-857 (2014). https://doi.org/10.2478/s11532-014-0508-4
877 TYSON, C. B., LORD, P. G. & WHEALS, A. E. Dependency of Size of Saccharomyces cerevisiae Cells on Growth Rate. *J. Bacteriology* **138**, 92-98 (1979).
878 Lin, N.-X., Xu, Y. & Yu, X.-W. Overview of yeast environmental stress response pathways and the development of tolerant yeasts. *Systems Microbiology and Biomanufacturing* **2**, 232-245 (2021). https://doi.org/10.1007/s43393-021-00058-4
879 Priks, H. *et al.* Physical Confinement Impacts Cellular Phenotypes within Living Materials. *ACS Appl Bio Mater* **3**, 4273-4281 (2020). https://doi.org/10.1021/acsabm.0c00335
880 Davey, H. M. & Hexley, P. Red but not dead? Membranes of stressed Saccharomyces cerevisiae are permeable to propidium iodide. *Environ Microbiol* **13**, 163-171 (2011). https://doi.org/10.1111/j.1462-2920.2010.02317.x
881 Dirmeier, R. *et al.* Exposure of yeast cells to anoxia induces transient oxidative stress. Implications for the induction of hypoxic genes. *J Biol Chem* **277**, 34773-34784 (2002). https://doi.org/10.1074/jbc.M203902200
882 Shen, L., Li, Y., Jiang, L. & Wang, X. Response of Saccharomyces cerevisiae to the stimulation of lipopolysaccharide. *PLoS One* **9**, e104428 (2014). https://doi.org/10.1371/journal.pone.0104428
883 Rieger, A. M., Hall, B. E., Luong le, T., Schang, L. M. & Barreda, D. R. Conventional apoptosis assays using propidium iodide generate a significant number of false positives that prevent accurate assessment of cell death. *J Immunol Methods* **358**, 81-92 (2010). https://doi.org/10.1016/j.jim.2010.03.019
884 Rego, A., Ribeiro, A., Corte-Real, M. & Chaves, S. R. Monitoring yeast regulated cell death: trespassing the point of no return to loss of plasma membrane integrity. *Apoptosis* **27**, 778-786 (2022). https://doi.org/10.1007/s10495-022-01748-7
885 Lopez, P. C., Peng, C., Arneborg, N., Junicke, H. & Gernaey, K. V. Analysis of the response of the cell membrane of Saccharomyces cerevisiae during the detoxification of common lignocellulosic inhibitors. *Sci Rep* **11**, 6853 (2021). https://doi.org/10.1038/s41598-021-86135-z
886 Rocchi, R., Wolkers-Rooijackers, J. C. M., Liao, Z., Tempelaars, M. H. & Smid, E. J. Strain diversity in Saccharomyces cerevisiae thiamine production capacity. *Yeast* **40**, 628-639 (2023). https://doi.org/10.1002/yea.3906
887 Junter, G. A. & Vinet, F. Compressive properties of yeast cell-loaded Ca-alginate hydrogel layers: Comparison with alginate–CaCO3 microparticle composite gel structures. *Chemical Engineering Journal* **145**, 514-521 (2009). https://doi.org/10.1016/j.cej.2008.09.034
888 Ebbesen, M., Korvink, J. G., Islam, M. & Lantada, A. D. The ethics of engineered living materials. *Trends Biotechnol* **42**, 5-9 (2024). https://doi.org/10.1016/j.tibtech.2023.09.004

889 Gietz, R. D. & Schiestl, R. H. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* **2**, 35-37 (2007). https://doi.org/10.1038/nprot.2007.14
890 Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345 (2009). https://doi.org/10.1038/nmeth.1318
891 AmplifX v. 2.1.1 (Aix-Marseille Univ, CNRS, INP, Inst Neurophysiopathol, Marseille, France).
892 Cormack, B., Valdivia, R. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33-38 (1996).
893 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012). https://doi.org/10.1038/nmeth.2019
894 Rueden, C. T. *et al.* ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529 (2017). https://doi.org/10.1186/s12859-017-1934-z
895 (Invitrogen, 2010).
896 Day, A., Schneider, C. & Schneider, B. L. in *Cell Cycle Checkpoint Control Protocols* (ed Howard B. Lieberman) 55-76 (Humana Press, 2004).
897 Savory, F. R., Milner, D. S., Miles, D. C. & Richards, T. A. Ancestral Function and Diversification of a Horizontally Acquired Oomycete Carboxylic Acid Transporter. *Mol Biol Evol* **35**, 1887-1900 (2018). https://doi.org/10.1093/molbev/msy082
898 Vazquez Torres, S. *et al.* De novo designed proteins neutralize lethal snake venom toxins. *Nature* (2025). https://doi.org/10.1038/s41586-024-08393-x
899 Xie, H. *et al.* Machine-Learning-Aided Engineering Hemoglobin as Carbene Transferase for Catalyzing Enantioselective Olefin Cyclopropanation. *JACS Au* **4**, 4957-4967 (2024). https://doi.org/10.1021/jacsau.4c01045
900 d'Oelsnitz, S. *et al.* Biosensor and machine learning-aided engineering of an amaryllidaceae enzyme. *Nat Commun* **15**, 2084 (2024). https://doi.org/10.1038/s41467-024-46356-y
901 Robinson, S. L. *et al.* Global analysis of adenylate-forming enzymes reveals beta-lactone biosynthesis pathway in pathogenic Nocardia. *J Biol Chem* **295**, 14826-14839 (2020). https://doi.org/10.1074/jbc.RA120.013528
902 Gu, J., Xu, Y. & Nie, Y. Role of distal sites in enzyme engineering. *Biotechnol Adv* **63**, 108094 (2023). https://doi.org/10.1016/j.biotechadv.2023.108094
903 Anderson, D. W., McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* **4**, e07864 (2015). https://doi.org/10.7554/eLife.07864
904 Ott, J. B. B. & Boerio-Goates, J. *Chemical Thermodynamics: Principles and Applications*. (Elsevier, 2000).
905 Roodveldt, C. & Tawfik, D. S. Directed evolution of phosphotriesterase from Pseudomonas diminuta for heterologous expression in Escherichia coli results in stabilization of the metal-free state. *Protein Eng Des Sel* **18**, 51-58 (2005). https://doi.org/10.1093/protein/gzi005
906 Vee Aune, T. E. *et al.* Directed evolution of the transcription factor XylS for development of improved expression systems. *Microb Biotechnol* **3**, 38-47 (2010). https://doi.org/10.1111/j.1751-7915.2009.00126.x
907 Reetz, M. T. & Zheng, H. Manipulating the expression rate and enantioselectivity of an epoxide hydrolase by using directed evolution. *Chembiochem* **12**, 1529-1535 (2011). https://doi.org/10.1002/cbic.201100078
908 Martiny, H. M., Armenteros, J. J. A., Johansen, A. R., Salomon, J. & Nielsen, H. Deep protein representations enable recombinant protein expression prediction. *Comput Biol Chem* **95**, 107596 (2021). https://doi.org/10.1016/j.compbiolchem.2021.107596

909 Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022). https://doi.org/doi:10.1126/science.add2187
910 Information, N. C. f. B. *PubChem Compound Summary for CID 6547, 2,2'-Azobis(2-methylpropionitrile)*, < https://pubchem.ncbi.nlm.nih.gov/compound/2_2_-Azobis_2-methylpropionitrile> (2025).
911 Information, N. C. f. B. *PubChem Compound Summary for CID 7187, Benzoyl Peroxide*, < https://pubchem.ncbi.nlm.nih.gov/compound/Benzoyl-Peroxide> (2025).
912 Wan, J., Fan, B. & Thang, S. H. RAFT-mediated polymerization-induced self-assembly (RAFT-PISA): current status and future directions. *Chem Sci* **13**, 4192-4224 (2022). https://doi.org/10.1039/d2sc00762b
913 Junker, K. *et al.* The use of Trametes versicolor laccase for the polymerization of aniline in the presence of vesicles as templates. *Enzyme Microb Technol* **55**, 72-84 (2014). https://doi.org/10.1016/j.enzmictec.2013.12.008
914 Junker, K., Zandomeneghi, G., Schuler, L. D., Kissner, R. & Walde, P. Enzymatic polymerization of pyrrole with Trametes versicolor laccase and dioxygen in the presence of vesicles formed from AOT (sodium bis-(2-ethylhexyl) sulfosuccinate) as templates. *Synthetic Metals* **200**, 123-134 (2015). https://doi.org/10.1016/j.synthmet.2015.01.004
915 Junker, K. *et al.* Efficient Polymerization of the Aniline Dimer p-Aminodiphenylamine (PADPA) with Trametes versicolor Laccase/O2 as Catalyst and Oxidant and AOT Vesicles as Templates. *ACS Catalysis* **4**, 3421-3434 (2014). https://doi.org/10.1021/cs500769d
916 Sheng, Q. & Zheng, J. Bienzyme system for the biocatalyzed deposition of polyaniline templated by multiwalled carbon nanotubes: a biosensor design. *Biosens Bioelectron* **24**, 1621-1628 (2009). https://doi.org/10.1016/j.bios.2008.08.029
917 Wu, Q., Wang, X., Liao, C., Wei, Q. & Wang, Q. Microgel coating of magnetic nanoparticles via bienzyme-mediated free-radical polymerization for colorimetric detection of glucose. *Nanoscale* **7**, 16578-16582 (2015). https://doi.org/10.1039/c5nr05716g
918 Dwivedi, U. N., Singh, P., Pandey, V. P. & Kumar, A. Structure–function relationship among bacterial, fungal and plant laccases. *Journal of Molecular Catalysis B: Enzymatic* **68**, 117-128 (2011). https://doi.org/10.1016/j.molcatb.2010.11.002
919 Sharma, A., Jain, K. K., Jain, A., Kidwai, M. & Kuhad, R. C. Bifunctional in vivo role of laccase exploited in multiple biotechnological applications. *Appl Microbiol Biotechnol* **102**, 10327-10343 (2018). https://doi.org/10.1007/s00253-018-9404-8
920 Kawata, T. & Ogino, H. Enhancement of the organic solvent-stability of the LST-03 lipase by directed evolution. *Biotechnol Prog* **25**, 1605-1611 (2009). https://doi.org/10.1002/btpr.264
921 Song, J. K. & Rhee, J. S. Enhancement of stability and activity of phospholipase A1 in organic solvents by directed evolution. *Biochim Biophys Acta* **1547**, 370-378 (2001).
922 Hao, J. & Berry, A. A thermostable variant of fructose bisphosphate aldolase constructed by directed evolution also shows increased stability in organic solvents. *Protein Eng Des Sel* **17**, 689-697 (2004). https://doi.org/10.1093/protein/gzh081
923 Piotukh, K. *et al.* Directed evolution of sortase A mutants with altered substrate selectivity profiles. *J Am Chem Soc* **133**, 17536-17539 (2011). https://doi.org/10.1021/ja205630g
924 Huang, B., Zhao, Q., Zhou, J. H. & Xu, G. Enhanced activity and substrate tolerance of 7alpha-hydroxysteroid dehydrogenase by directed evolution for 7-ketolithocholic

acid production. *Appl Microbiol Biotechnol* **103**, 2665-2674 (2019). https://doi.org/10.1007/s00253-019-09668-4

925 Hibbert, E. G. *et al.* Directed evolution of transketolase substrate specificity towards an aliphatic aldehyde. *J Biotechnol* **134**, 240-245 (2008). https://doi.org/10.1016/j.jbiotec.2008.01.018

926 Leemhuis, H., Nightingale, K. P. & Hollfelder, F. Directed evolution of a histone acetyltransferase--enhancing thermostability, whilst maintaining catalytic activity and substrate specificity. *FEBS J* **275**, 5635-5647 (2008). https://doi.org/10.1111/j.1742-4658.2008.06689.x

927 Tang, Z. *et al.* Improved thermostability and enzyme activity of a recombinant phyA mutant phytase from Aspergillus niger N25 by directed evolution and site-directed mutagenesis. *Enzyme Microb Technol* **108**, 74-81 (2018). https://doi.org/10.1016/j.enzmictec.2017.09.010

928 Li, G., Zhang, H., Sun, Z., Liu, X. & Reetz, M. T. Multiparameter Optimization in Directed Evolution: Engineering Thermostability, Enantioselectivity, and Activity of an Epoxide Hydrolase. *ACS Catalysis* **6**, 3679-3687 (2016). https://doi.org/10.1021/acscatal.6b01113

929 Liu, F. *et al.* Preparation and Characterization of an Ancient Aminopeptidase Obtained from Ancestral Sequence Reconstruction for L-Carnosine Synthesis. *Molecules* **27** (2022). https://doi.org/10.3390/molecules27196620

930 Rozi, M., Rahman, R., Leow, A. T. C. & Ali, M. S. M. Ancestral sequence reconstruction of ancient lipase from family I.3 bacterial lipolytic enzymes. *Mol Phylogenet Evol* **168**, 107381 (2022). https://doi.org/10.1016/j.ympev.2021.107381

931 Frey, R., Hayashi, T. & Hilvert, D. Enzyme-mediated polymerization inside engineered protein cages. *Chem Commun (Camb)* **52**, 10423-10426 (2016). https://doi.org/10.1039/c6cc05301g

932 Rhee, H.-W. *et al.* Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging. *Science* **339**, 1328-1331 (2013). https://doi.org/doi:10.1126/science.1230593

933 Lam, S. S. *et al.* Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat Methods* **12**, 51-54 (2015). https://doi.org/10.1038/nmeth.3179

934 Gomez-Fernandez, B. J., Risso, V. A., Rueda, A., Sanchez-Ruiz, J. M. & Alcalde, M. Ancestral Resurrection and Directed Evolution of Fungal Mesozoic Laccases. *Applied and Environmental Microbiology* **86** (2020). https://doi.org/10.1128/AEM

935 Keser, M. *et al.* Stable and Promiscuous Galactose Oxidases Engineered by Directed Evolution, Atomistic Design, and Ancestral Sequence Reconstruction. *ACS Synth Biol* **14**, 239-246 (2025). https://doi.org/10.1021/acssynbio.4c00653

936 Liu, J. & Liu, B. Living cell-mediated in-situ polymerization for biomedical applications. *Progress in Polymer Science* **129** (2022). https://doi.org/10.1016/j.progpolymsci.2022.101545

937 Perrino, G., Hadjimitsis, A., Ledesma-Amaro, R. & Stan, G. B. Control engineering and synthetic biology: working in synergy for the analysis and control of microbial systems. *Curr Opin Microbiol* **62**, 68-75 (2021). https://doi.org/10.1016/j.mib.2021.05.004

938 Lee, H.-M., Vo, P. N. L. & Na, D. Advancement of Metabolic Engineering Assisted by Synthetic Biology. *Catalysts* **8** (2018). https://doi.org/10.3390/catal8120619

939 Duraj-Thatte, A. M. *et al.* Water-processable, biodegradable and coatable aquaplastic from engineered biofilms. *Nat Chem Biol* **17**, 732-738 (2021). https://doi.org/10.1038/s41589-021-00773-y

940 Binelli, M. R. *et al.* Complex Living Materials Made by Light-Based Printing of Genetically Programmed Bacteria. *Adv Mater* **35**, e2207483 (2023). https://doi.org/10.1002/adma.202207483

941 Wang, Y. *et al.* Advancing Engineered Plant Living Materials through Tobacco BY-2 Cell Growth and Transfection within Tailored Granular Hydrogel Scaffolds. *ACS Cent Sci* **10**, 1094-1104 (2024). https://doi.org/10.1021/acscentsci.4c00338

942 McBee, R. M. *et al.* Engineering living and regenerative fungal-bacterial biocomposite structures. *Nat Mater* **21**, 471-478 (2022). https://doi.org/10.1038/s41563-021-01123-y

943 Parra-Torrejon, B. *et al.* Bioinspired mineralization of engineered living materials to promote osteogenic differentiation. *Biomater Adv* **154**, 213587 (2023). https://doi.org/10.1016/j.bioadv.2023.213587

944 Sun, T. *et al.* Engineered Cyanobacteria-Based Living Materials for Bioremediation of Heavy Metals Both In Vitro and In Vivo. *ACS Nano* **18**, 17694-17706 (2024). https://doi.org/10.1021/acsnano.4c02493