The structure, dynamics, function and evolution of binding proteins

Joe Alexander Kaczmarski

July 2020

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

© Copyright by Joe Alexander Kaczmarski 2020

All Rights Reserved
Statement of authorship

Unless otherwise stated, the work described in this thesis was carried out by the author between February 2016 and July 2020 under the supervision of Professor Colin Jackson at the Research School of Chemistry, Australian National University. The work presented in this thesis has not been submitted as part of any other degrees.

However, this thesis does include several co-authored, original papers published in peer-reviewed journals, as well as manuscripts that are in preparation for submission to peer-reviewed journals. The following table outlines the status (as of 5th February 2020) and the contribution of Joe Alexander Kaczmarski to the research and authorship of each paper. Detailed statements outlining the contributions of the authors are also provided before each manuscript or publication.

Signed,

_______________________________
Joe A. Kaczmarski
21 July 2020

_______________________________
Professor Colin J. Jackson (on behalf of all collaborating authors)
21 July 2020
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Authorship</th>
<th>Publication Outlet &amp; Status</th>
<th>Overall contribution of JAK</th>
</tr>
</thead>
</table>
| 2.2     | Structural basis for the allosteric regulation of the SbtA bicarbonate transporter by the $P_{II}$-like protein, SbtB, from *Cyanobium* sp. PCC7001 | **Joe A. Kaczmarski**, Nan-Sook Hong, Britati Mukherjee, Laura T. Wey, Loraine Rourke, Britta Förster, Thomas S. Peat, G. Dean Price, Colin J. Jackson. | *Biochemistry*  
  In press.  
  https://doi.org/10.1021/acs.biochem.9b00880 | ~85 % |
  In press.  
  https://doi.org/10.1038/s41589-018-0043-2 | ~40 % |
  Available on BioRxiv at https://doi.org/10.1101/2020.02.03.932491 | ~30 % |
  In press.  
  https://doi.org/10.1371/journal.ppat.1006469 | ~15% |
  Available on BioRxiv at https://doi.org/10.1101/2020.01.12.903682 | ~5 % |
| 5.2     | *Drosophila melanogaster* nonribosomal peptide synthetase Ebony encodes an atypical condensation domain | Thierry Izoré, Julien Tailhades, Mathias Henning Hansen, **Joe A. Kaczmarski**, Colin J. Jackson, and Max J. Cryle. | *PNAS*  
  In press.  
  https://doi.org/10.1073/pnas.1811194116 | ~10% |
| 5.4     | Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics | Anja Greule, Thierry Izoré, Dumitrīta Iftime, Julien Tailhades, Melanie Schoppet, Yongwei Zhao, Madeleine Peschke, Iftekhar Ahmed, Andreas Kulik, Martina Adamek, Robert J.A. Goode, Ralf B. Schittenhelm, **Joe A. Kaczmarski**, Colin J. Jackson, Nadine Ziemert, Elizabeth H. Krenske, James J. De Voss, Evi Stegmann, Max J. Cryle | *Nature Communications*  
  In press.  
  https://doi.org/10.1038/s41467-019-10384-w | ~5% |
6.2 Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors


Current Opinion in Structural Biology
In press
https://doi.org/10.1016/j.sbi.2019.01.013

* These authors contributed equally
Acknowledgements

I would like to acknowledge the financial support I have received from the Research School of Chemistry, as well as from the Australian Government through the Research Training Program Stipend Scholarship.

Thanks to all my friends in the Jackson Lab (past and present) and the neighbouring biochemistry groups at the Research School of Chemistry. You have put a smile on my face every day of my PhD, and I have learnt a lot from each of you. I have been especially honoured to work directly with Ben Clifton on the cyclohexadienyl dehydratase project (Chapter 3). I also thank Dr Paul Carr and Lynn Tan for their generous help with protein crystallography and data collection.

I have had the pleasure of working with several great collaborators during my PhD. I thank Ian Cockburn for welcoming me into the world of structural immunology (Chapter 4), Dean Price for opening my eyes to the wonderful cyanobacterial CO₂-concentrating mechanism (Chapter 2), and Max Cryle for inviting me to work on his projects looking at the nonribosomal peptide synthesis machinery (Chapter 5). I thank Mithun Mahawaththage and Gottfried Otting for their patience and discussions regarding the study on the dynamics of solute-binding proteins (Chapter 3). I would also like to thank Nobu Tokuriki and his group for hosting me during my time at the University of British Columbia and for insightful discussions on science and life.

I would also like to thank my PhD supervisory panel for helpful feedback, and my academic mentor Graham Farquhar for the discussions we had while waiting in line for coffee. A special thanks to Mark Ellison for his support throughout my teaching endeavors. I would also like to thank all the students that I have had the pleasure of teaching and learning from.

A special thanks to my Canberra family, especially Samara, Oliver, Rosa, Owen, Galen and Jess for making the last few years among the best years of my life. A huge thank you to my mum, for teaching me to appreciate everything that is beautiful in life and for always being no further than phone call away.
Finally, I would like to say a massive thank you to my PhD supervisor, Professor Colin Jackson. I am very grateful for his consistent enthusiasm, optimism, patience, advice, encouragement and support during my time at ANU. I am thankful for the wonderful opportunities that Colin has given me to develop, both as a scientist and as a person, and I thank him for opening my eyes to the wonderful world of proteins, structural biology and protein engineering.
“Few scientists acquainted with the chemistry of biological systems at the molecular level can avoid being inspired. Evolution has produced chemical compounds exquisitely organized to accomplish the most complicated and delicate of tasks.”

— Donald J. Cram

Abstract

Macromolecular recognition – the ability of biological molecules (especially proteins) to produce physiological responses by forming specific, high-affinity, non-covalent complexes with other molecules – underlies every biological process, including cellular signalling, immune recognition and enzyme catalysis. Determining the chemical and physical factors (including the structure and dynamics of the interacting molecules) that govern the properties of macromolecular recognition events is central to appreciating how biological molecules work and why disease occurs, and drives developments in protein engineering, synthetic biology and rational drug and vaccine design. As demonstrated in this thesis, a detailed understanding of protein-ligand binding can be achieved by combining biophysical, structural, computational and evolution-based methods to elucidate the relationship between the structure and dynamics of the interacting molecules and the thermodynamics, properties (e.g. affinity, kinetics, allostery) and biological consequences of binding events.

In particular, this thesis centers around using X-ray crystallography and other biophysical methods, such as isothermal titration calorimetry, to understand the structural factors that govern several diverse molecular recognition events. In each section, I combine an understanding of molecular evolution with experimental methods to characterise how the function of these proteins is influenced by changes in their sequence, structure and structural flexibility. We discuss the implications that these findings have on protein engineering, medicine, and synthetic biology. In doing so, we not only expand our understanding of how these particular proteins function, but also provide insight into how their activity could be modified, engineered or targeted for useful applications.

Chapter 1 provides an introduction to the concepts and ideas that are central to this thesis and explores how this work fits into our current understanding of the process of molecular recognition. Chapter 2 explores the structural and biophysical properties that govern the function of a P1-like protein that is involved in regulating bicarbonate transport in cyanobacteria. Chapter 3 presents work in which we discuss the evolution of an enzyme from an ancestral solute-binding protein that was specialised for a non-catalytic role, with a focus on the historical changes in structure and dynamics that were required for the emergence and enhancement of the new catalytic activity. In Chapter 4, we characterise an unusual
mode of binding between neutralising antibodies and the *Plasmodium falciparum* circumsporozoite protein and discuss the implications that this has on efforts to design an improved malaria vaccine. In Chapter 5, we explore interactions between molecules involved in nonribosomal peptide synthesis, highlighting the structural basis for specificity and affinity. In Chapter 6, I present a short review that outlines current evolutionary and structural approaches for designing genetically-encoded biosensors for small molecules. This review brings together concepts discussed in the earlier chapters and recapitulates how an understanding of the evolutionary drivers and structural determinants of molecular recognition and protein function can be applied to guide the design of useful molecular tools. Finally, I conclude this thesis with a discussion on recent developments in areas related to the projects and aspects that could to be addressed in future work.
Table of Contents

1 Introduction ................................................................................................................................. 18
1.1 Introduction to macromolecular recognition ........................................................................... 18
1.2 The protein sequence–structure–function paradigm .............................................................. 20
1.3 The thermodynamics of binding: a simple bimolecular model ............................................... 20
1.4 Properties of binding interactions ........................................................................................... 23
1.5 Beyond a simple binding model ............................................................................................. 25
1.6 The role of protein dynamics in protein–ligand binding .............................................................. 27
1.7 The evolution of protein-ligand binding interactions ................................................................ 32
1.8 Techniques and approaches for investigating binding interactions ........................................... 33
1.9 Thesis motivation & research objectives ................................................................................ 36

2 Structural basis for the regulation of bicarbonate transport by the cyanobacterial P_{II}-like signal transduction protein, SbtB ........................................................................................................... 40
2.1 Preface .................................................................................................................................... 40
2.2 Paper 1: Structural basis for the regulation of bicarbonate transport by the P_{II}-like protein, SbtB, from Cyanobium sp. PCC 7001 ................................................................................... 42
2.3 Supplementary Information for “Structural basis for the regulation of bicarbonate transport by the P_{II}-like protein, SbtB, from Cyanobium sp. PCC 7001” ................................................ 54

3 The evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein...........69
3.1 Preface .................................................................................................................................... 69
3.2 Paper 1: Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein .................................................................................................................. 71
3.3 Supplementary Information for “Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein” ........................................................................................................ 83
3.4 Paper 2: Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme ..............................................................................................117
3.5 Supplementary Information for “Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme” .................................................... 154
4 Interactions between neutralising antibodies and the circumsporozoite protein from *Plasmodium falciparum* .................................................................................................................................. 175

4.1 Preface.................................................................................................................................. 175

4.2 Paper 1: T-dependent B cell responses to *Plasmodium* induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein ......................................................... 177

4.3 Supplementary Information for “T-dependent B cell responses to *Plasmodium* induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein” .............................................................................................................. 201

4.4 Paper 2: Avid binding by B cells to the *Plasmodium* circumsporozoite protein repeat supresses responses to protective subdominant epitopes .................................................. 218

4.5 Supplementary Information for “Avid binding by B cells to the *Plasmodium* circumsporozoite protein repeat supresses responses to protective subdominant epitopes” .......................... 252

5 Using ITC to characterise molecular interactions involving novel components of the nonribosomal synthetase machinery. .............................................................................................................. 272

5.1 Preface.................................................................................................................................. 272

5.2 Paper 1: *Drosophila melanogaster* nonribosomal peptide synthetase Ebony encodes an atypical condensation domain .............................................................................................................. 274

5.3 Supplementary Information for “*Drosophila melanogaster* nonribosomal peptide synthetase Ebony encodes an atypical condensation domain” ............................................................. 281

5.4 Paper 2: Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics .......................................................................... 305

5.5 Supplementary Information for “Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics” .............................. 321

6 Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors ........................................................................................................ 396

6.1 Preface.................................................................................................................................. 396

6.2 Review paper: Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors ................................................................. 398

7 Conclusions, implications and future directions ........................................................................ 408

7.1 New insight into SbtB and the PII protein fold .................................................................... 409

7.2 The role of remote mutations and dynamics in the evolution of a new enzyme activity ..... 411
7.3 Investigating the properties of antibodies that bind PfCSP\textsubscript{Repeat} and implications for vaccine design .............................................................. 414
7.4 Nonribosomal peptide synthesis machinery: applications and evolution .......................... 423
7.5 Concluding remarks ...................................................................................................... 425
8 References ........................................................................................................................................... 426
List of Figures (that are not included in publications)

Figure 1-1. A living cell is just a collection of interacting molecules.................................19

Figure 1-2. Simple binding curves can be modelled using the Langmuir model. (A) ........24

Figure 1-3. Protein dynamics and the conformational landscape....................................29

Figure 1-4. A typical isothermal titration calorimetry experiment....................................35

Figure 2-1. Binding of ATP stabilises the flexible T-loop of SbtB7001..............................41

Figure 3-1. Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein.................................................................70

Figure 6-1. Graphical Abstract from "Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors"..................397

Figure 7-1. Structural polarity and the evolvability of protein folds...............................411

Figure 7-2. Multi-valent binding of Fabs to PfCSP_{Repeat} confirmed by Cryo-EM.........416

Figure 7-3. Diversity amongst anti-PfCSP_{Repeat} Fabs............................................418

Figure 7-4. NVDP motifs act as anchors in PfCSP.......................................................422
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta G$</td>
<td>Change in Gibbs free energy</td>
</tr>
<tr>
<td>$\Delta H$</td>
<td>Change in enthalpy</td>
</tr>
<tr>
<td>$\Delta S$</td>
<td>Change in entropy</td>
</tr>
<tr>
<td>Ab</td>
<td>Antibody</td>
</tr>
<tr>
<td>Ag</td>
<td>Antigen</td>
</tr>
<tr>
<td>ASR</td>
<td>Ancestral sequence reconstruction</td>
</tr>
<tr>
<td>CCM</td>
<td>CO$_2$-concentrating mechanism</td>
</tr>
<tr>
<td>CDR</td>
<td>Complementary-determining region</td>
</tr>
<tr>
<td>CDT</td>
<td>Cyclohexadienyl dehydratase</td>
</tr>
<tr>
<td>Cryo-EM</td>
<td>Cryogenic electron microscopy</td>
</tr>
<tr>
<td>CSP</td>
<td>Circumsporozoite protein</td>
</tr>
<tr>
<td>DEER</td>
<td>Double electron electron resonance</td>
</tr>
<tr>
<td>DSF</td>
<td>Differential scanning fluorimetry</td>
</tr>
<tr>
<td>EPR</td>
<td>Electron paramagnetic resonance</td>
</tr>
<tr>
<td>$f$</td>
<td>Fractional occupancy</td>
</tr>
<tr>
<td>Fab</td>
<td>Antigen-binding fragment</td>
</tr>
<tr>
<td>gAb</td>
<td>Germline antibody</td>
</tr>
<tr>
<td>Ig</td>
<td>Immunoglobulin</td>
</tr>
<tr>
<td>ITC</td>
<td>Isothermal titration calorimetry</td>
</tr>
<tr>
<td>$K_a/K_d$</td>
<td>Equilibrium association/dissociation constant</td>
</tr>
<tr>
<td>$k_{cat}$</td>
<td>Turnover number</td>
</tr>
<tr>
<td>$K_M$</td>
<td>Michaelis constant</td>
</tr>
<tr>
<td>$k_{cat}/k_{off}$</td>
<td>First order rate constant for association/dissociation</td>
</tr>
<tr>
<td>LDE</td>
<td>Lab directed evolution</td>
</tr>
<tr>
<td>mAb</td>
<td>Monoclonal antibody</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular dynamics</td>
</tr>
<tr>
<td>NRP(S)</td>
<td>Nonribosomal peptide (synthetase)</td>
</tr>
<tr>
<td>nsEM</td>
<td>Negative stain electron microscopy</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>PaCDT</td>
<td>CDT from <em>Pseudomonas aeruginosa</em></td>
</tr>
<tr>
<td>PCP</td>
<td>Peptidyl carrier protein</td>
</tr>
<tr>
<td>PfCSP</td>
<td>CSP from <em>Plasmodium falciparum</em> (repeat region)</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>rCSP</td>
<td>Recombinant CSP</td>
</tr>
<tr>
<td>rsCSP</td>
<td>Recombinant, short CSP</td>
</tr>
<tr>
<td>SBP</td>
<td>Solute binding protein</td>
</tr>
<tr>
<td>SbtA/B</td>
<td>Sodium-dependent bicarbonate transporter A/B</td>
</tr>
<tr>
<td>SPR</td>
<td>Surface plasmon resonance</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction.
“...if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jiggings and wigglings of atoms.”

— Richard Feynman, The Feynman Lectures on Physics
1 Introduction

1.1 Introduction to macromolecular recognition

Proteins realise their biological function by forming specific, noncovalent complexes with other molecules (including copies of themselves, other proteins, nucleotides, lipids, metabolites, ions and solvent molecules). Through a process of genetic diversification and natural selection, proteins adopt diverse physical and chemical characteristics (including a wide array of structures and dynamics) that allow them to coordinate and elicit specific responses to certain environmental signals (such as an increase in the local concentration of a ligand or substrate). These protein-mediated molecular recognition events underlie every biological process, including cellular signalling (e.g. Chapter 2), enzyme catalysis (e.g. Chapter 3) and immune recognition (e.g. Chapter 4). Together, molecular recognition events allow the host organism to respond and adapt to changing environmental and cellular conditions. Recognising the physiochemical drivers, energetics and structural basis of molecular recognition events is central to appreciating how and why biological processes occur, understanding the molecular causes of disease, and engineering novel protein function.

The work presented in this thesis explores the structural factors that govern the properties of several unique molecular recognition events, drawing on computational and evolution-based approaches, X-ray crystallography and biophysical techniques (such as isothermal titration calorimetry, ITC) to elucidate the connection between the physical and chemical properties of proteins and the thermodynamic parameters, binding properties and biological consequences of the molecular interactions that they participate in. I focus on binding interactions that involve proteins from a diverse set of protein families and folds: a cyanobacterial PII-like signal transduction protein (Chapter 2), bacterial amino acid solute-binding proteins (Chapter 3), anti-malarial antibodies (Chapter 4), protein domains involved in nonribosomal peptide synthesis (Chapter 5), and engineered genetically-encoded fluorescence-based small-molecule biosensors (Chapter 6). In each piece of work, we discuss how changes to the sequence, structure and dynamics of the proteins affect the properties of the protein-ligand binding process, propose models that describe the overall
binding process and explore the implications of these findings to specific problems and challenges within the fields of protein evolution, medicine, protein engineering and synthetic biology.

In the following sections, I introduce some basic concepts that are central to this work: the relationship between the sequence, structure, dynamics and function of proteins; the thermodynamic drivers and important properties of binding events; theoretical models that describe binding processes and that can be used to calculate binding parameters from experimental data; the evolution of molecular recognition events; and the theory behind experimental techniques that I have used in this work.

**Figure 1-1. A living cell is just a collection of interacting molecules.** Proteins do not exist in isolation. In biological environments, such as the inside of this crowded bacterial cell, proteins are constantly interacting with other molecules. Indeed, the biological function of a protein is realised through direct interactions with other molecules, and the properties of these interactions are primarily determined by the physical and chemical properties of the interacting molecules. Illustration by David S. Goodsell, Scripps Research Institute.
1.2 The protein sequence–structure–function paradigm

A core theme explored in this thesis is the relationship between protein sequence, protein structure and the properties of molecular recognition events. It has long been recognised that the structure of a protein determines the types of molecules that it will bind and the properties of the resulting binding event (including the stoichiometry, affinity, kinetics and associated biological consequences). For example, the Lock-and-Key model, first proposed by Emil Fischer in 1894 (Fischer, 1894), recognises that high-affinity binding will occur when the shape and chemistry of the ligand and protein binding-site are complementary. According to this classic model of molecular recognition, ligands that are the wrong shape will not be able to bind the protein, and so the shape of the protein’s binding site becomes the primary factor that determines the binding specificity of a protein. Indeed, thousands of structures of protein–ligand complexes published on the Protein Data Bank (PDB) confirm that complementarity between interaction surfaces in protein-ligand complexes is central to determining the affinity and selectivity of molecular recognition events (Kuroda and Gray, 2016; Tsuchiya et al., 2006). When combined with fundamental laws of thermodynamics, even simplistic models such as the Lock-and-key model, which treat the protein and ligands as rigid entities and binding as a simple two-state process (unbound → bound), can help to explain some characteristics of binding interactions and be used to rationalise data from structural and biophysical experiments. However, as I will discuss below, more comprehensive models that take into account competing ligands, multiple binding interfaces, multi-step kinetics, and the intrinsic flexibility of molecules are often required to accurately describe, understand, and engineer the more subtle physiochemical characteristics and mechanisms that control the outcome of molecular recognition events.

1.3 The thermodynamics of binding: a simple bimolecular model

Determining the thermodynamic drivers of binding events has been central to the work presented in this thesis. To introduce the basics of binding thermodynamics, let us consider the simple two-state (unbound → bound), noncovalent interaction between a protein with a single binding site (P) and a single specific ligand (L). The ligand could be another copy of the same protein (in the case of oligomerization), another macromolecule (protein, DNA, lipid
etc.) or a small molecule. When the protein and ligand are mixed together in an aqueous solution (i.e. in a simple protein–ligand–solvent system) an equilibrium will develop between the free species and the protein–ligand complex (PL):

\[
P + L \rightleftharpoons PL, \quad (1.1)
\]

where \( k_{\text{on}} \) and \( k_{\text{off}} \) are the kinetic rate constants for the binding (in \( \text{M}^{-1} \cdot \text{s}^{-1} \)) and dissociation (in \( \text{s}^{-1} \)) reactions, respectively.

The protein–ligand complex will form spontaneously when the change in Gibbs free energy of the system (\( \Delta G \)) is negative under equilibrium conditions at constant pressure and temperature. Additionally, the magnitude of the standard Gibbs free energy of binding, \( \Delta G^\circ \), (at 1 atm, 298 K, 1 M concentrations of reactants) is related to the affinity and kinetics of the binding interaction under equilibrium conditions (Kuriyan et al., 2013):

\[
\Delta G^\circ = \frac{RT}{K_d} = -RT \ln \left( \frac{k_{\text{on}}}{k_{\text{off}}} \right), \quad (1.2)
\]

where \( R \) is the universal gas constant (1.987 cal.K\(^{-1}\).mol\(^{-1}\) or 8.314 J.mol\(^{-1}\).K\(^{-1}\)), \( T \) is the temperature of the system in Kelvin, and \( K_d \) is the equilibrium dissociation constant.

Importantly, \( \Delta G \) can be parsed into terms for the entropic and enthalpic contributions to binding:

\[
\Delta G = \Delta H - T \Delta S \quad (1.3)
\]

where \( \Delta H \) is the total change in enthalpy of the system upon binding, \( \Delta S \) is the total change in entropy of the system upon binding, and \( T \) is the temperature of the system (in Kelvin). Therefore, at a given temperature, \( \Delta H \) and \( \Delta S \) can be considered as the drivers of molecular association. \( \Delta H \) and \( \Delta S \) are global properties of the protein-ligand-solvent system, arising from differences in the arrangement of atoms between the unbound and bound states. Therefore, \( \Delta H \) and \( \Delta S \) of binding are determined by the inherent physical and chemical properties of all the individual molecules within the system (the protein, ligand, and solvent),
and by the intramolecular and intermolecular interactions between the individual atoms in the system.

The $\Delta H$ of binding, for example, arises from the formation and breaking of many individual noncovalent interactions, including those between the protein and the ligand, between the protein and the solvent, between the ligand and the solvent, between solvent molecules, as well as intramolecular interactions within the protein molecule and ligand (Du et al., 2016). This includes the breaking and forming of hydrogen bonds, halogen bonds, electrostatic interactions, aromatic interactions (including $\pi-\pi$ interactions) and van der Waals interactions.

The $\Delta S$ of binding, a measure of the change in disorder of the system, is made up of the change in entropy of the solvent ($\Delta S_{\text{solv}}$), conformational entropy ($\Delta S_{\text{conf}}$) and rotational/translational entropy ($\Delta S_{\text{r/t}}$):

$$
\Delta S = \Delta S_{\text{solv}} + \Delta S_{\text{conf}} + \Delta S_{\text{r/t}} \quad (1.4)
$$

The total $\Delta S$ during protein–ligand binding tends to be dominated by large, positive contributions from $\Delta S_{\text{solv}}$, caused by the release of ordered solvent molecules associated with the burial of solvent-accessible surface area upon binding (Du et al., 2016). $\Delta S_{\text{conf}}$, the change in conformational freedom of both the protein and the ligand upon binding (which can be positive or negative), and $\Delta S_{\text{r/t}}$, the loss of rotational and translational degrees of freedom that occurs upon binding, also contribute to the total change in entropy of the system upon binding, but normally to a smaller extent. Since protein–ligand binding is inevitably associated with some unfavourable entropic costs (such as negative $\Delta S_{\text{r/t}}$ upon the formation of a protein–ligand complex), binding will only occur if these entropic penalties are compensated for by other factors, such as a large positive $\Delta S_{\text{solv}}$ and/or negative $\Delta H$.

Identifying the structural changes and molecular rearrangements that occur during binding and contribute to $\Delta H$ and $\Delta S$ (and how these change during the evolution of new protein function) is central to appreciating how proteins achieve their function and understanding how molecular recognition events can be manipulated or engineered; this is a theme that is central to this thesis.
1.4 Properties of binding interactions

Proteins adopt many different ligand-binding properties, including affinity and kinetics, to carry out their function in the cell. These attributes are optimised through the process of molecular evolution.

1.4.1 Binding affinity & fractional occupancy

The strength, or affinity, of binding is a key property of binding interactions and determines the proportion of bound protein under equilibrium conditions (Kuriyan et al., 2013). In each of the chapters in this thesis, assessing the affinity of the binding interactions by ITC has been important for understanding the molecular recognition events being studied. The affinity of a molecular interaction is normally reported as the equilibrium association constant \( K_a \) or the equilibrium dissociation constant \( K_d \), which are defined as the ratio of the \( k_{on} \) and \( k_{off} \):

\[
K_a = \frac{k_{on}}{k_{off}} = \frac{[PL]}{[P][L]} = \frac{1}{K_d}. \quad (1.5)
\]

The \( K_d \) reflects the proportion of protein that will be bound by a ligand at a given ligand concentration. For simple bimolecular interactions at equilibrium, the fractional occupancy \( (f) \) of the protein (at equilibrium) can be approximated by the Langmuir equation:

\[
f = \frac{[PL]}{[P]_{total}} = \frac{[L]}{K_d + [L]} \quad (1.6)
\]

Importantly, \( K_d \) values of simple interactions can be estimated by fitting this model (or similar derivatives of this model) to data from \textit{in vitro} binding assays that measure the proportion of ligand-bound protein as a function of the ligand concentration (Latour, 2015). Figure 1-2a shows \( f \) plotted as a function of [L] for three different \( K_d \) values, each of which displays the shape of a typical hyperbolic binding isotherm. Plotting \( f \) as a function of \([L]/K_d\) on a logarithmic scale yields the universal binding isotherm for bimolecular interactions (Figure 1-2b), from which it is clear that the \( K_d \) of a protein–ligand interaction defines the ligand concentration range over which the protein switches from an unbound to a bound state. For simple bimolecular binding interactions, this switch occurs over about a 100-fold increase in [L], centered around the \( K_d \). That is, the protein goes from being mostly unoccupied at 0.1 \( \times K_d \) (9 % occupied) to mostly occupied at 10 \( \times K_d \) (91 % occupied). In each of the chapters in...
this thesis, assessing the affinity of the binding interactions by ITC has been key to understanding the molecular recognition events being studied.

Figure 1-2. Simple binding curves can be modelled using the Langmuir model. (A) A plot of the Langmuir model showing the fractional occupancy of three binding interactions with different $K_d$ values. The protein will be at 50% occupancy when $[L]=K_d$. (B) The universal binding isotherm of simple protein–ligand interactions highlights that a protein switches from mostly unoccupied (9%) to mostly occupied (~90%) between $[L]=0.1K_d$ and $[L]=10K_d$. Figures adapted from (Kuriyan et al., 2013).

1.4.2 Binding kinetics

While the affinity of a binding interaction is traditionally viewed as the most important parameter of binding, the rates at which molecules bind and unbind are also critical to all biological processes, especially if there are competing reactions. Indeed, many biological processes are kinetically controlled rather than thermodynamically controlled and may never reach steady state-conditions. For example, protein-ligand complexes that are short-lived might not be able to activate slower, downstream processes in signalling cascades. Studies on protein–drug interactions have indicated that the magnitudes of the mean lifespan ($t_{mean} = \frac{1}{k_{off}}$) and mean half-life ($t_{0.5} = \frac{\ln(2)}{k_{off}}$) of a protein–ligand complex can be more important factors for determining the efficacy of drugs than the binding affinity of the interaction (Corzo and Santamaria, 2006; Pang and Zhou, 2017). Drugs that have a longer off-rate may occupy the receptor in favour of others even if the $K_d$ values are similar, for example. In other cases, a fast association rate may be needed to target short-lived receptors. Similarly, the rate of substrate/product binding and release can limit the catalytic efficiently of enzymes (Bar-Even
Introduction

et al., 2011); the rates of these processes are often controlled by the intrinsic dynamics of the protein (See Chapter 3). While explicitly quantifying binding kinetics has not been a primary goal of this thesis, the rates of ligand binding and dissociation and the effect that these could have on the outcome of molecular recognition events have been taken into consideration and are discussed in the work presented in this thesis.

1.5 Beyond a simple binding model

Biological molecular recognition events are rarely as simple as the bimolecular model discussed above, and are typically complicated by effects arising from molecular crowding (which can limit protein-ligand collisions), competition between multiple ligands, the presence of multiple binding sites, allosteric effects, irreversible covalent binding, multistep kinetics and the intrinsic dynamics of proteins and ligands.

1.5.1 Competitive binding at a single binding site

When more than one ligand has access to a given site on a protein (or a neighbouring, overlapping binding site), the specificity of a protein’s binding interface – i.e. its ability to discriminate between different (but often very similar) ligands – becomes a key determinant of a protein’s physiological output and function (Du et al., 2016). Some proteins have evolved specific binding sites that only bind one physiological ligand with high affinity, consistent with the traditional “one protein-one function” view of protein function. However, a large proportion of proteins are multi-specific, participating in multiple high-affinity, biologically-important interactions, each of which is maintained by selection (Schreiber and Keating, 2011). For example, about 37 % of metabolic enzymes in E. coli are multifunctional (Nam et al., 2012). A large proportion of proteins also participate in promiscuous binding or have promiscuous functions that, by definition, occur at low levels and are neutral to selection pressures (James and Tawfik, 2003a). The widespread presence of multi-specificity, promiscuous functions and promiscuous binding is likely to be due to the unselective nature of structural features such reactive residues and hydrophobic binding pockets (Khersonsky and Tawfik, 2010).

In a given environment (i.e. at fixed temperature, pressure, and protein/ligand concentrations), the specificity of an interaction can be enhanced by increasing the affinity
for the target molecule and/or by decreasing the affinity for off-target molecules. However, mutations in the protein receptor that increase the affinity for one ligand (through hydrophobic effects, for example) tend to increase the affinity of similar ligands – giving rise to a commonly observed affinity/specificity trade-off (Copley, 2015). In some cases, specificity can be achieved by the incorporation of residues that facilitate the formation of ligand-specific salt-bridges or hydrogen bonds; unlike hydrophobic effects, hydrogen bonds and ionic interactions are highly sensitive to interatomic distances and residue orientation (Chen et al., 2016). As an example, specific polar interactions are important for determining the selectivity between adeny l nucleotides in the P$_{II}$-like protein, SbtB (Chapter 2).

In a biological context, problems due to poor specificity can also be overcome by controlling the relative local concentrations of ligands near the target protein, through cellular compartmentalisation of metabolic pathways or by forming closely connected, multi-protein complexes. For example, cyanobacteria overcome the problematic dual specificity of the CO$_2$-fixing enzyme, Rubisco, by maintaining near saturating concentrations of CO$_2$ around the enzyme inside specialized proteinaceous compartments called carboxysomes as part of their CO$_2$-concentrating mechanism (discussed in Chapter 2) (Rae et al., 2013). The formation of large, modular complexes, such as those involved in nonribosomal peptide synthesis (discussed in Chapter 5) facilitate direct shuttling of substrates between individual subunits (McErlean et al., 2019), and thereby increase the overall efficiency of the biosynthetic pathway by minimising side reactions.

1.5.2 Multiple binding sites, allostery & cooperativity

Many proteins form specific interactions with small molecules and/or other proteins via more than one surface or binding site. Distinct binding sites on a protein can be independent – that is, binding at one site does not affect the properties at the second binding site. However, the presence of multiple binding sites in a protein tends to give rise to allostery, where the binding of a molecule to one binding-site alters properties at the other binding site (Liu and Nussinov, 2016). Importantly, allosteric proteins display cooperativity, in that the overall free energy of binding associated with the cooperative binding events is different than the sum of the individual binding event contributions to the total binding free energy. Binding cooperativity can be positive (synergistic) or negative (interfering) and allows cellular signals to be
enhanced and attenuated in a non-linear manner. While proteins with a single binding site require >100-fold change in ligand concentration to go from mostly unbound to bound (Figure 1-2), cooperativity between binding sites can lead to ultrasensitive proteins that respond to much more subtle (and physiologically-relevant) changes in ligand concentrations.

Allostery and cooperativity plays a key role in many biological processes. For example, signal transduction proteins are typically allosteric proteins (e.g. Chapter 2) in which binding of one ligand causes a conformational change in the protein that controls how readily it forms complexes with other ligands at a remote interaction interface. Metabolic enzymes and transporters are often allosterically regulated (often by the product of the reaction) to minimise futile cycling (Wegner et al., 2015). Binding cooperativity is also a common property of multi-or poly-valent interactions, including protein-carbohydrate interactions and interactions involving proteins with tandem repeats (see Chapter 4) (Klein et al., 2003).

Unfortunately (for protein engineers), the general rules that define the sequence and structural basis of allostery and cooperativity are not completely understood and attempts to introduce allostery into protein scaffolds still rely on screening of large libraries or on extensive optimisation by techniques such as directed evolution (Greener and Sternberg, 2018). However, it is now widely accepted that the conformational flexibility of proteins plays an important role in facilitating many properties of binding events, including allosteric regulation and binding promiscuity (Motlagh et al., 2014; Wrabl et al., 2011).

1.6 The role of protein dynamics in protein–ligand binding.

Proteins and ligands are not rigid species but exist in solution as an ensemble of conformations, and the intrinsic structural dynamics of proteins play an important role in determining protein function and the properties of molecular recognition events (Nussinov et al., 2019). Protein motions occur on a wide range of timescales (Figure 1-3a), from fast bond vibrations and side chain rotations (fs – ns), to slower loop rearrangements and rigid body motions (ns – s), and these motions are fine-tuned by natural selection to facilitate protein function (Villali and Kern, 2010). Protein dynamics may be required to reach functionally-relevant conformational sub-states (Henzler-Wildman et al., 2007), to allow ligand binding and release (Jackson et al., 2009), or to facilitate allosteric regulation of protein function. Studying
intrinsic protein dynamics and ligand-induced conformational changes has been central to the work presented throughout this thesis.

Protein dynamics are commonly discussed using the metaphor of a free energy landscape, where each conformational sub-state (i.e. each possible arrangement of all the atoms in the given protein–solvent system) is represented as a point within the space of all conformational possibilities and the height at each point reflects the free energy of that sub-state (Figure 1-3) (Frauenfelder et al., 1991). With this free energy landscape in mind, the dynamics of a protein can be discussed in terms of statistical thermodynamics; the most thermodynamically stable conformational sub-states will be the most populated in a given population of molecules (at equilibrium), and the energy barriers between local minima determine the rate at which these conformations interconvert. Each protein-solvent system has a unique conformational free-energy landscape that is determined by the amino acid sequence of the protein and the conditions of the system (pH, temperature, pressure etc.). The intrinsic large-scale motions of a protein (e.g. rigid-body domain movement and loop motions) are largely dictated by the protein’s topology and overall fold; tightly-packed, folded domains tend to be rigid, while regions without secondary structure (such as loops and linkers) are intrinsically flexible (Papaleo et al., 2016). Consequently, large scale protein motions tend to be conserved amongst proteins that share the same fold. For example, almost all members of the solute-binding protein superfamily undergo intrinsic rigid-body, open-closed transitions around a flexible hinge region (see Chapter 3); their free energy landscape is typically characterised by two energy minima (Scheepers et al., 2016). On the other hand, it is now clear that seemingly subtle changes in protein dynamics can dramatically affect protein function and that the distribution of conformational sub-states of a protein–solvent system is readily shifted by changing environmental conditions (e.g. pH, temperature), mutations (as a consequence of protein evolution or protein engineering) or by the presence of other interacting molecules (Campbell et al., 2016) (Figure 1-3b).

While protein folding, intrinsic protein dynamics and protein-ligand binding are often discussed as being separate processes, they are far from being independent. Each of these processes is driven by the change in the total Gibbs free energy of the whole thermodynamic system and the only difference is the lack of covalent connectivity in the case of protein-ligand binding (Tsai et al., 1998). Because of this, the intrinsic protein-folding/conformational...
landscape influences the way that a protein interacts with ligands, and the presence of ligands can alter the way that proteins fold and move by causing a redistribution of conformational sub-states. Like protein folding, protein-ligand binding is a dynamic process consisting of many individual steps involving coordinated motions of the protein, ligand and solvent molecules, and can be represented as a free energy landscape. In some cases, ligand binding induces a protein conformation that is not normally thermodynamically accessible in the absence of the ligand (this process can, therefore, be approximated by the induced-fit model (Koshland, 1959)). In other cases, ligand binding stabilises conformations that were sampled before the addition of the ligand (i.e. the conformational-selection model (Ma et al., 1999)). While discussions on whether ligand binding occurs via an induced fit process or by conformational selection are still prevalent in the literature (Paul and Weikl, 2016), it should be recognised that these two models are not necessarily mutually exclusive; the induced fit model can be explained in terms of the conformational selection model in which high-energy conformations are only rarely sampled in the absence of ligands. Depending on the system, both models can be used, separately or in combination, to describe the characteristics of protein-ligand free energy landscapes.

Figure 1-3. Protein dynamics and the conformational landscape. Protein dynamics occur on a range of timescales (from fs to s) and these dynamics are determined by the free energy landscape of the system. (A) A one-dimensional cross-section of the free energy landscape of a protein reveals how the energy barriers that separate distinct states (e.g. States A and B) define the rate of motions between these two states. States separated by large energy barriers interconvert over larger timescales. A change in the system (mutation, addition of ligand, change in external conditions etc.) alters the free energy landscape and dynamics of the system (dark blue to light blue line). Figure from (Henzler-Wildman and Kern, 2007). (B) Changes to the system, such as mutations, can alter the conformational landscape such that
pre-existing promiscuous conformations or functions are enriched within the population of protein molecules. Figure from (Maria-Solano et al., 2018).

Recognising that proteins and ligands exist as an ensemble of alternate conformations and representing the protein-ligand binding process as a free-energy landscape helps to explain properties of ligand binding that static models of binding (such as the Lock-and-Key model) cannot explain. For example, allostery can be explained as arising from changes in the conformational landscape, where binding of a ligand or mutation at a remote site shifts the distribution of conformational sub-states and stabilises a certain conformation at the allosterically coupled binding site. Transient sampling of distinct protein conformations also contributes to the ability of proteins to recognise multiple ligands with high affinity, and may also facilitate non-physiological promiscuous binding or side-reactions (Tokuriki and Tawfik, 2009). For example, cytochrome P450 3A4 can bind structurally distinct compounds by adopting two different active site conformations (Ekroos and Sjögren, 2006), and the IgE antibody SPE7 binds 11 distinct ligands by adopting slightly different conformations (James and Tawfik, 2003b). In cases where protein flexibility facilitates promiscuous binding, small changes in the protein dynamics can improve these functions to near-native levels when selection pressures change (Aharoni et al., 2005; Bloom and Arnold, 2009; Campbell et al., 2018). The role of protein dynamics during the evolution of a new enzyme function is explored in detail in Chapter 3.

Appreciating and understanding protein dynamics is particularly important for the field of protein engineering and computational protein design. The connection between protein sequence and structure is now reasonably well understood and small, stable rigid protein folds can be accurately designed de novo (Huang et al., 2016). This is normally achieved using computational techniques that optimise a single target fold by maximising favourable non-bonded interactions made by side-chain rotamers in the desired folded state. However, the importance of protein flexibility and the often-subtle role of protein dynamics in determining protein function can complicate protein engineering efforts; proteins designed using standard computational approaches often lack appropriate dynamics and flexibility and, as a consequence, tend to have sub-optimal activities. For example, MD simulations revealed that the limited activity of a designed Diels-Alderase was due to non-productive sampling of a catalytic tyrosine; substitution of a neighbouring alanine changed the sampling
of this tyrosine and improved catalytic activity (Siegel et al., 2010). Engineering dynamics and allostery into proteins is difficult since it requires the design of protein sequences that adopt multiple conformational states with comparable (and interconvertible) free energies. The ensemble view of proteins has given rise to several approaches for designing and understanding dynamics and allostery, including using flexible linkers to create pH or temperature dependent switches (Choi et al., 2015) or generalisable systems-level approaches that dissect the role of individual mutations on dynamics and allostery using large-scale functional screens (Raman, 2018). As we begin to gain a better understanding of the role of dynamics in protein function and the connection between sequence and dynamics, the modification of protein dynamics will become a powerful avenue for engineering and optimising new protein function (see Chapter 6, for example) and lead to the emergence of computational tools that can rationally design dynamics. In a recent example, a multistate design approach was used to introduce two distinct, exchanging conformations in a protein G domain (Davey et al., 2017).

Ideally, investigations into protein-ligand binding processes (and the effect of sequence changes on this process) would involve obtaining a detailed description of the whole protein-ligand energy landscape, allowing all possible binding pathways, microscopic energy states and intermediates to be characterised and manipulated. However, the thermodynamic profiles of the individual steps within binding pathways are difficult to obtain through experiments, and the conformational space of even the simplest protein-ligand-solvent system has too many dimensions to be exhaustively characterised by experiments or computational approaches. When an initial structure is available, techniques such as molecular docking, MD and free energy calculations (Maria-Solano et al., 2018) can be used to elucidate the local free-energy landscape that surround this ground state and can provide some useful information about the dynamic binding process. However, this process typically involves compressing the multidimensional conformational space into fewer dimensions (e.g. by principal-component analysis), which can cause important properties of the free-energy landscape to be hidden or missed.

Since the protein-ligand energy landscape is too large to be completely sampled and characterised, most structure-function studies continue to take the more traditional approach of comparing the structures and energies of the thermodynamically stable ground states of
proteins under different conditions (e.g. the same protein with and without ligands, or closely related proteins) to elucidate the structural determinants of binding properties. These structures represent minima in the overall free energy landscape and can be characterised more readily by techniques such as X-ray crystallography. While such approaches do not uncover all aspects of the binding process, there is still great value in this approach (as exemplified by the work presented in this thesis). For example, comparing apo and ligand-bound structures can help identify residues that are involved in ligand binding, and highlights conformational changes that occur during ligand binding, thereby highlighting the origins of $\Delta H$ and $\Delta S$ (which can be quantified by techniques such as ITC). Further, comparison of the function and structures of homologous proteins, intermediates along evolutionary trajectories, or artificially modified proteins (e.g. by site directed mutagenesis) can be used to identify how changes in the sequence affect the structure and function of proteins, thereby providing valuable information about the connection between protein sequence, structure, dynamics and function.

1.7 The evolution of protein-ligand binding interactions

As mentioned above, combining evolutionary analysis with biophysical and biochemical studies can be particularly useful when trying to understand how biological molecules work and how changes to protein sequence affect protein folding and function (Harms and Thornton, 2013). Rather than asking the broad question of how protein sequence determines structure and function (answering this is an impossible task given the vastness of sequence-structure-function space), evolutionary biochemistry focuses on identifying how historical changes in sequence during evolution have contributed to changes in protein structure, dynamics and function (Harms and Thornton, 2013). One such approach is to compare the sequence, structures and functions of homologous proteins using multiple sequence alignments or sequence similarity networks. These “horizontal” approaches are often sufficient for identifying residues essential for a function since such residues will be highly conserved amongst related proteins that share a primary function. However, horizontal approaches tend to miss subtle details about the sequence-function relationship. For example, it can be difficult to identify function-shifting mutations by comparing the sequences of homologs, which often differ by up to $\sim$80%. In contrast, more “vertical” evolution-based
approaches, such as ancestral sequence reconstruction (ASR) (Merkl and Sterner, 2016) and laboratory-based directed evolution (LDE) (Campbell et al., 2018) can be used to structurally and functionally characterise intermediates along evolutionary pathways, which can help to identify a smaller subset of function-shifting substitutions. For example, ASR was used to identify two historical substitutions that caused a 70,000-fold shift in the ligand specificity during the divergence of two clades of vertebrate hormone receptors (Harms and Thornton, 2013). Structural and biochemical analysis of these evolutionary intermediates revealed how these historical substitutions led to a rearrangement of hydrogen-bond networks, which caused the shift in specificity. Identifying these substitutions using alignments of present-day sequences alone would have been challenging since the extant proteins differ at ~70% of their residues. Throughout this thesis, we have used horizontal and vertical evolution-based approaches (i) to understand how specific protein functions have arisen within protein families or along evolutionary trajectories by identifying conserved residues or substitutions that have caused shifts in protein function, and (ii) to illuminate fundamental characteristics of molecular evolution and the protein-ligand binding process.

1.8 Techniques and approaches for investigating binding interactions

A detailed understanding of any macromolecular binding event requires data obtained from multiple techniques, including biophysical, structural and evolution-based approaches. This thesis is centered around using ITC and X-ray crystallography, in combination with other computational and experimental approaches, to understand protein-ligand binding processes.
1.8.1 Isothermal titration calorimetry for determining binding parameters

Isothermal titration calorimetry (ITC) is the gold standard for elucidating the thermodynamic signatures of macromolecular binding interactions. For simple bimolecular systems, a single, well-planned ITC experiment can yield the binding association constant ($K_a$), binding stoichiometry ($n$), as well as the enthalpic ($\Delta H$) and entropic ($\Delta S$) contributions to binding. In a typical ITC experiment, the ligand is titrated into a sample cell that contains a solution of the protein of interest. In modern power compensation ITC instruments, power is constantly applied to keep the sample cell and an identical water-filled reference cell at the same temperature. Following each ligand injection, any heat evolved from a binding event is compensated by altering the amount of power applied to the sample cell so that both cells stay at the same temperature. Each injection is therefore normally associated with a peak in the power that is applied to the sample cell, and integration of these peaks yields a binding isotherm that reflects the total amount of heat produced or absorbed during each injection of ligand as a function of the ligand:protein ratio. Fitting of this binding isotherm to an appropriate binding model yields the parameters of binding (Figure 1-4a).

Experimental design plays a key role in obtaining accurate binding parameters from ITC experiments. For the standard ITC experiment, the main variables that can be altered are the concentration of the protein in the cell, the ligand:protein ratio at the end of the experiment, and the number and volume of the injections. A typical ITC experiments has around 20 injections (which yields 20 data points in the binding isotherm), and the volume of each injection is limited by the volume of the injection syringe. The informative fitting of a simple independent model to a simple binding isotherm requires that the protein is sufficiently bound to ligand at the end of the titration; for high affinity, bimolecular interactions this is normally achieved with a 2-3 molar excess of ligand, but lower affinity interaction may require the ligand to be in a higher excess (Tellinghuisen, 2008).

Importantly, the concentration of protein in the cell ($[P]$) influences the shape of the binding isotherm. The shape of the binding isotherm is determined by the value of the parameter $c$:

$$c = [P] \times n \times K_a$$

(1.9)
Typically, experiments in which the c value is between 1-1000 yield a binding isotherm that can be used to determine accurate values of $K_a$, $n$ and $\Delta H$ (Velázquez-Campoy et al., 2004). At very large c (>1000), the binding isotherm becomes too steep to accurately determine $K_a$, and at low c (<10) the binding isotherm is shallow and prone to errors in the fitting parameters such as $\Delta H$ (Velázquez-Campoy et al., 2004) (Figure 1-4). This has been taken into account throughout this thesis.

One of the biggest limitations to using ITC for studying the binding parameters of protein-ligand interactions is that it requires a lot of material and is typically low-throughput. In cases where material is precious, other techniques, such as equilibrium dialysis, analytical ultracentrifugation, NMR spectroscopy, microscale thermophoresis or surface plasmon resonance can be used for investigating parameters of binding.

**Figure 1-4. A typical isothermal titration calorimetry experiment.** A power compensation isothermal titration calorimeter measures the power applied to the cell following consecutive injections of the analyte into the reaction well (top panel). Integration of the area under each of the peaks produces a binding isotherm that can be fit using an appropriate model to yield the $\Delta H$, $\Delta S$, $K_d$, and stoichiometry ($n$) of a binding interaction (bottom panel).
1.8.2 X-ray crystallography provides information about how three-dimensional structure relates to binding properties

X-ray crystallographic structures of proteins in complex with target molecules can provide details about the static interactions that define the interaction, providing information about enthalpic contributions to the binding free energy. Entropic contributions can be inferred by comparing structures of the protein with and without its targets; comparisons of ligand-free and ligand-bound states can highlight ligand-induced conformational changes in the protein. X-ray crystallography experiments can also highlight the intrinsic dynamics of proteins; regions with poor electron density are often associated with high levels of disorder, and B-factors reflect the mean square displacements of non-hydrogen atoms in different regions of the protein. However, in some cases the dynamics of the protein may be constrained by the crystal lattice, and other techniques, such as EPR or MD simulations may be required to observe and quantify these dynamics (e.g. see Chapter 3).

1.9 Thesis motivation & research objectives

1.9.1 Research Objectives

The work presented in this thesis is centred around using X-ray crystallography and ITC, combined with evolutionary information, to determine the structural and thermodynamic drivers of molecular recognition events. I aimed to explore, using distinct protein-ligand systems from a range of organisms, how changes in enthalpy and entropy during binding contribute to the properties of binding (affinity, kinetics, specificity) and discuss the physiological consequences and molecular evolution of these interactions. Throughout, the combination of structural and biophysical data facilitates discussion on the relationship between the structure and dynamics of proteins and the properties of binding interactions. In each case, we discuss how changes to protein structure alters the properties of any binding interactions, and how these changes affect the physiological consequences of the interaction. Comparisons of the structure of homologous or ancestrally related proteins highlight how differences in the sequence and structure of proteins lead to differences in the thermodynamics and binding properties. Finally, as well as providing insight into the
fundamental aspects of molecular recognition, each project is connected with a ‘big-picture’ or practical application, and we discuss how understanding the underlying determinants of protein function can guide protein engineering, synthetic biology, biotechnology and medicine.

1.9.2 Research Outline

In the following chapter (Chapter 2), I present a publication that describes the structural basis for the binding of effector molecules, including adenylnucleotides, to a cyanobacterial P\textsubscript{II}-like protein, SbtB. Using ITC and X-ray crystallography, we provide a structural explanation for differences in the affinities between SbtB and AMP, ADP, cAMP and ATP. Further, we show, for the first time, that ATP and Ca\textsuperscript{2+}/ATP stabilise the otherwise flexible T-loop, and present a testable model in which ATP binding regulates the formation of an inhibitory complex with the bicarbonate transporter, SbtA. This work provides new insight into this family of P\textsubscript{II}-like proteins, highlights differences between SbtB homologs, and we discuss how this information could be useful for efforts that aim to improve photosynthetic yield of C3 crop by incorporating components of the cyanobacterial CO\textsubscript{2}-concentrating mechanism into the chloroplasts of these plants.

In Chapter 3, I present two pieces of work that explore the evolution of cyclohexadienyl dehydratase (CDT) from an ancestral solute-binding protein. In the first paper, we use ancestral protein reconstruction to infer the sequences of ancestors of CDT. Using ITC and differential scanning fluorimetry (DSF), we show that CDT (which has a type II periplasmic binding protein fold) evolved from a cationic amino acid binding protein. Using structural, computational and phylogenetical analyses, as well as directed evolution, we highlight the steps that were required for the emergence and evolution of CDT-like activity: the functionalisation of the active site residues and restructuring of the binding cavity, and the accumulation of mutations remote from the binding site that likely stabilise the catalytically-relevant closed state of the protein.

In the second manuscript, we use double electron-electron resonance (DEER) distance measurements on CDT variants tagged on their two domains using site-specifically incorporated Gd(III) tags to confirm that CDT is stabilised in a closed state in solution. Combined with results from a colorimetric prephenate dehydratase assay and molecular
dynamic simulations, we discuss how substitutions between the more flexible ancestral binding proteins and CDT could have caused the decreased sampling of a catalytically unproductive wide open state. Together, these two pieces of work highlight the importance of conformational dynamics in molecular recognition and we discuss the implication of these findings to the field of enzyme design.

In Chapter 4, I present work on the binding of monoclonal antibodies to the *Plasmodium falciparum* circumsporozoite surface protein (CSP). First, I performed ITC experiments that highlighted a novel mode of binding of these anti-CSP antibodies, in which they radiate out from the repeating epitopes on the CSP molecule. Second, we present a paper that looks at the effect of truncating the CSP repeat region on the antibody response. I discuss implications of these results for the structure-based design of new molecules to be used as part of malaria vaccines.

In Chapter 5, I present two publications that explore the structure and function of two novel domains involved in nonribosomal peptide synthesis (NRPS). Understanding how these domains interact with other modules within the NRPS complex and with substrates and products is critical to understanding the molecular mechanisms that underly their biological function and NRP biosynthesis. For this work, I performed ITC experiments that supported X-ray crystallographic data, and helped elucidate the properties of molecular interactions that are important for NRPS.

In Chapter 6, I present a review paper that discusses the structural and evolutionary approaches for designing genetically encodable fluorescent-based small molecule biosensors. This work ties in nicely with the previous chapters, demonstrating how a structural and evolutionary understanding of naturally occurring molecular interactions is important for the engineering of proteins with novel functions that can have far-reaching applications.

Finally, in Chapter 7 I summarise the work presented in this thesis, mention recent studies that have built upon our work, and discuss future directions and challenges for the field.
Chapter 2. Structural basis for the regulation of bicarbonate transport by the cyanobacterial $P_{II}$-like signal transduction protein, SbtB
2 Structural basis for the regulation of bicarbonate transport by the cyanobacterial P$_{II}$-like signal transduction protein, SbtB

Key Ideas: P$_{II}$-like signal transduction proteins, competitive binding, ligand-induced conformational change, allosteric regulation.

Techniques: Evolution-based approaches, X-ray crystallography, isothermal titration calorimetry.

2.1 Preface

2.1.1 Introduction

The P$_{II}$ signal-transduction proteins make up one of the largest structurally-conserved superfamilies of regulatory proteins and are involved in the control of diverse metabolic pathways in all domains of life (Forchhammer and Lüddecke, 2016). P$_{II}$ proteins share a core trimeric structure, with ligand-binding sites at the clefts between the monomers. While the core structure is conserved amongst P$_{II}$ proteins, the sequences and lengths of flexible loop regions vary, allowing them to respond to a range of cellular signals, form complexes with numerous protein targets, and regulate diverse metabolic systems.

In this chapter, I present work on the structure and function of a P$_{II}$-like signal transduction protein found in cyanobacteria that is involved in regulating bicarbonate uptake (Figure 2-1). This work extends our general understanding of the structural basis for metabolic regulation by P$_{II}$-like proteins, highlighting a mode of ATP binding and recognition that is distinct from what is observed in canonical P$_{II}$ proteins and other P$_{II}$-like proteins. Further, X-ray crystallography and ITC data suggest that Ca$^{2+}$ plays a key role in the function of SbtB. Most importantly, it provides a structural explanation of how SbtB could regulate bicarbonate transport in response to changing cellular conditions via the ligand-dependent formation and dissociation of a complex with the sodium-dependent bicarbonate transporter, SbtA.
Understanding how SbtB senses cellular conditions and regulates components of the cyanobacterial CO₂-concentrating mechanism (CCM), is important for on-going synthetic biology-based efforts that aim to improve the photosynthetic output of crop plants through the incorporation of cyanobacterial CCM proteins into the chloroplasts of C3 plants (Rae et al., 2017). Finally, this work demonstrates how valuable information can be obtained by combining structural and biophysical experiments and highlights key concepts (such as cooperativity between ligands, and the importance of conformational flexibility for protein function) that will be revisited throughout this thesis.

Figure 2-1. Binding of ATP stabilises the flexible T-loop of SbtB7001. Graphical abstract from the manuscript “Structural basis for the regulation of bicarbonate transport by the PII-like protein, SbtB, from Cyanobium sp. PCC 7001.”
2.2 Paper 1: Structural basis for the regulation of bicarbonate transport by the PII-like protein, SbtB, from *Cyanobium* sp. PCC 7001.

2.2.1 Publication status

This manuscript presented in this section has been published in *Biochemistry* ([https://doi.org/10.1021/acs.biochem.9b00880](https://doi.org/10.1021/acs.biochem.9b00880)). Supporting information is provided in Section 2.3.

2.2.2 Author’s contribution

I was the lead author on this project and conducted most of the protein expression, protein purification, ITC and X-ray crystallography work. I helped with the initial conceptualization of the project and led the direction of this project. I designed the experiments (with advice from co-authors), and collected, analysed and interpreted all the data. I solved and validated all of the X-ray crystal structures presented in this manuscript. I wrote most of the manuscript and produced all the figures.

2.2.3 Contributions from others

This work was a collaboration with Professor G. Dean Price (Research School of Biology, Australian National University) and his group. Professor Price and his group proposed the original concept for this project (that is, to identify ligands of SbtB7001 and to get X-ray crystal structures of the protein) and helped with editing the final manuscript. Cloning of the gene encoding SbtB7001 was done in Professor Price’s group. Members of the Price group also provided some Ni-NTA-purified SbtB7001 that was used to obtain crystals of apo-SbtB7001. Dr Nansook Hong (Jackson Lab, ANU) performed size-exclusion chromatography on this sample. Initial crystal screens were prepared from this sample at the Collaborative Crystallisation Centre (C3, CSIRO). We thank staff at C3 for their help with setting up these crystallization screens.
Lynn Tan (ANU), Dr. Tom Peat (CSIRO) and colleagues, and Dr. Paul Carr (ANU) aided in freezing of protein crystals and the collection of X-ray crystallography data. Lynn Tan also expressed and purified the USP2-catalytic core protease that was used for this work. Professor Colin Jackson helped with experimental design, interpretation of the data, and writing of the manuscript. All co-authors contributed to the editing of this manuscript.
Structural Basis for the Allosteric Regulation of the SbtA Bicarbonate Transporter by the P$_{II}$-like Protein, SbtB, from Cyanobium sp. PCC7001

Joe A. Kaczmarski,$^{, }$ Nan-Sook Hong,$^+$ Brattti Mukherjee,$^+$ Laura T. Wey,$^+$ Loraine Rourke,$^+$ Britta Förster,$^{, }$ Thomas S. Peat,$^+$ G. Dean Price,$^{, +}$ and Colin J. Jackson$^{*, +}$

$^1$Research School of Chemistry, The Australian National University, 137 Sullivans Creek Road, Canberra, ACT 0200, Australia
$^2$ARC Centre of Excellence for Translational Photosynthesis, Division of Plant Sciences, Research School of Biology, The Australian National University, 134 Linnaeus Way, Canberra, ACT 0200, Australia
$^3$CSIRO Biomedical Program, 343 Royal Parade, Parkville, VIC 3052, Australia

Supporting Information

ABSTRACT: Cyanobacteria have evolved a suite of enzymes and inorganic carbon (C$_i$) transporters that improve photosynthetic performance by increasing the localized concentration of CO$_2$ around the primary CO$_2$-fixing enzyme, Rubisco. This CO$_2$-concentrating mechanism (CCM) is highly regulated, responds to illumination/darkness cycles, and allows cyanobacteria to thrive under limiting C$_i$ conditions. While the transcriptional control of CCM activity is well understood, less is known about how regulatory proteins might allosterically regulate C$_i$ transporters in response to changing conditions. Cyanobacterial sodium-dependent bicarbonate transporters (SbtAs) are inhibited by P$_{II}$-like regulatory proteins (SbtBs), with the inhibitory effect being modulated by adenylnucleotides. Here, we used isothermal titration calorimetry to show that SbtB from Cyanobium sp. PCC7001 (SbtB7001) binds AMP, ADP, cAMP, and ATP with micromolar-range affinities. X-ray crystal structures of apo and nucleotide-bound SbtB7001 revealed that while AMP, ADP, and cAMP have little effect on the SbtB7001 structure, binding of ATP stabilizes the otherwise flexible T-loop, and that the flexible C-terminal C-loop adopts several distinct conformations. We also show that ATP binding affinity is increased 10-fold in the presence of Ca$^{2+}$, and we present an X-ray crystal structure of Ca$^{2+}$-ATP:SbtB7001 that shows how this metal ion facilitates additional stabilizing interactions with the apex of the T-loop. We propose that the Ca$^{2+}$-ATP-induced conformational change observed in SbtB7001 is important for allosteric regulation of SbtA activity by SbtB and is consistent with changing adenylnucleotide levels in illumination/darkness cycles.

Cyanobacteria possess a highly regulated network of enzymes, inorganic carbon (C$_i$) transporters, and specialized compartments that enhance photosynthetic performance by increasing the localized concentration of CO$_2$ around the primary CO$_2$-fixing enzyme, Rubisco. This CO$_2$-concentrating mechanism (CCM) responds to illumination/darkness cycles and allows cyanobacteria to thrive under limiting C$_i$ conditions. Bicarbonate (HCO$_3^−$) transporters, which maintain high concentrations of HCO$_3^−$ in the cell (≤40 mM), are vitally important to the CCM. While the transcriptional regulation of these HCO$_3^−$ transporters has been characterized, the molecular signals and post-translational mechanisms that transiently control the activity of cyanobacterial HCO$_3^−$ transporters are only beginning to be elucidated.5,6

High-affinity, low-flux, sodium-dependent bicarbonate transporter A (SbtA), which is found in the plasma membrane of many cyanobacterial species, is a key component of the cyanobacterial CCM. The dicistronic operon that typically encodes both SbtA and a cytoplasmic P$_{II}$-like regulatory protein (SbtB) is upregulated under C$_i$-limiting conditions in Synechocystis and Synechococcus species.7–10 We have previously shown that SbtB from Synechococcus elongatus PCC7942 (SbtB7942) forms a stable complex with SbtA7942 and that several SbtB homologues from Synechococcus, Synechocystis, and Cyanobium sp. inhibit their cognate SbtA when co-expressed in Escherichia coli.11 Because SbtA-mediated HCO$_3^−$ transport must be coupled with the concurrent activity of ATP-dependent sodium exporters, we proposed that SbtB limits futile cycling by directly regulating SbtA-mediated HCO$_3^−$ transport in cyanobacteria through the formation of an inhibitory SbtA–SbtB complex.11 This general

Received: September 25, 2019
Revised: November 5, 2019
Published: November 20, 2019
model was recently confirmed by Selim et al., who showed that SbtB from \textit{Synechocystis} sp. PCC6803 (SbtB6803) forms a stable complex with membrane-bound SbtA6803 under limiting C\textsubscript{4} conditions \textit{in vivo} and responds to changes in the relative levels of adenylnucleotides in the cell.

SbtB is a noncanonical member of the P\textsubscript{II} signal transduction superfamily.\textsuperscript{5,6,11} P\textsubscript{II} proteins are homotrimERIC proteins that typically sense and tune the metabolic state of cells by binding adenylnucleotides and small molecules and forming regulatory complexes with diverse protein targets, including transcription factors, enzymes, and membrane transporters.\textsuperscript{12} The functions of P\textsubscript{II} proteins are primarily driven by large, ligand-induced conformational changes in the three flexible, solvent-exposed T-loops that mediate interactions with target proteins. Competitive binding of adenylnucleotides (typically ADP/Mg\textsuperscript{2+}/ATP) and/or other effector molecules at the clefts between the subunits can alter the conformation of the T-loops and modulate the affinity of the P\textsubscript{II} protein for its target.\textsuperscript{13}

Selim et al.\textsuperscript{6} recently reported that SbtB6803 binds AMP, ADP, cyclic AMP (cAMP), and ATP with micromolar-range affinities, although the reported range of \textit{K}_{D} values was relatively narrow (11–46 \textmu M from microscale thermophoresis experiments and 75–250 \textmu M for \textit{K}_{D} from ITC experiments). AMP and ADP were observed to stabilize the SbtA–SbtB6803 complex \textit{in vivo}, and the authors suggested that the cAMP:AMP ratio may act as a signal to control SbtA6803-mediated HCO\textsubscript{3}−/HCO\textsubscript{3} transport \textit{in vivo}. However, the structures of SbtB6803 in complex with AMP and cAMP did not show any ligand-induced T-loop rearrangements that are typically required for the differential binding of P\textsubscript{II} proteins to their target proteins.\textsuperscript{6} Therefore, the structural basis for SbtA inhibition by the formation of the SbtA–SbtB complex remains unclear. Furthermore, the functional consequences of sequence differences among SbtB homologues have not been explored. For example, many SbtB homologues, such as SbtB from \textit{Cyanobium} sp. PCC7001 (SbtB7001), lack the C-terminal extension that forms a putative redox-sensing domain in SbtB6803.\textsuperscript{6} Additional structural and functional data on phylogenetically distinct SbtB homologues are required to achieve a deeper understanding of the molecular basis for the regulation of CCM activity, via regulation of SbtA-mediated HCO\textsubscript{3}−/HCO\textsubscript{3} transport, by SbtB.

Here we show that although SbtB7001 binds AMP, ADP, ATP, and cAMP with micromolar-range affinities (like SbtB6803\textsuperscript{3}), the presence of Ca\textsuperscript{2+} dramatically increases the affinity of SbtB7001 for ATP affinity such that it is \textasciitilde 50- and 100-fold greater than the affinity for ADP/cAMP and AMP (in the absence of Ca\textsuperscript{2+}), respectively. High-resolution crystal structures reveal that AMP, ADP, and cAMP have little effect on the structure of SbtB7001, while ATP/Ca\textsuperscript{2+}/ATP binding causes dramatic rearrangements in the structure of SbtB7001, providing a molecular mechanism for the allosteric modulation of SbtB7001 function by ligand binding.

\section*{Materials and Methods}

\textbf{Cloning, Expression, and Purification of SbtB7001.} The gene encoding SbtB7001 (PCC7001_1671) was codon-optimized for expression in \textit{E. coli}, synthesized and cloned into pUC57 by GenScript, and then subcloned into the pHUE vector\textsuperscript{14} between BamHI and HindIII sites to yield pHUE-SbtB7001. \textit{E. coli} strain DH5α cells were used for cloning. pHUE adds polyhistidine and ubiquitin tags (His\textsubscript{6}-Ub) to the N-terminus of SbtB7001. Successful cloning was confirmed by Sanger sequencing at Garvan Molecular Genetics.

Protein expression was carried out in BL21(DE3) cells. Briefly, cells were grown in autoinduction media (20 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl, 6 g/L Na\textsubscript{2}HPO\textsubscript{4}, 3 g/L KH\textsubscript{2}PO\textsubscript{4}, 6 mL/L glycerol, 2 g/L lactose, 0.5 g/L glucose, and 100 mg/L ampicillin), and protein expression was carried out overnight at 25 °C with shaking. His\textsubscript{6}-Ub-tagged SbtB7001 was purified from the cell lysate using a 5 mL HiTrap FF column (GE Healthcare). The pre-equilibrated and lysate-loaded column was washed with binding buffer [20 mM NaH\textsubscript{2}PO\textsubscript{4}, 500 mM NaCl, and 20 mM imidazole (pH 8)]. nonspecific binders were removed using a step of 7% elution buffer, and tagged SbtB7001 was eluted using 100% elution buffer [20 mM NaH\textsubscript{2}PO\textsubscript{4}, 500 mM NaCl, and 500 mM imidazole (pH 8.0)]. Fractions containing tagged SbtB7001 were pooled and transferred into binding buffer using a HiPrep 26/10 desalting column (GE Healthcare). The His\textsubscript{6}-Ub tags were removed using a ubiquitin carboxyl-terminal hydrolase 2 catalytic core domain (USP2cc):SbtB7001 molar ratio of 1:25 at 4 °C overnight as previously described.\textsuperscript{14} Untagged SbtB7001 was recovered using a 5 mL HisTrap FF column (GE Healthcare) using binding buffer and concentrated using a 3 Kd high-molecular-weight cutoff Amicon Ultra Centrifugal Filter (Merck). Sodium dodecyl sulfate–polyacrylamide gel electrophoresis was used to assess the purity of untagged SbtB7001. For preparative purification, untagged SbtB7001 was loaded on a HiLoad Superdex 75 16/600 size-exclusion chromatography (SEC) column (GE Healthcare) and eluted in SEC buffer [20 mM HEPES and 150 mM NaCl (pH 8.0)]. For analytical size exclusion, untagged SbtB7001 was loaded on a Superdex 75 10/300 GL column (GE Healthcare) and eluted in SEC buffer. The size-exclusion column was calibrated using a set of standard proteins (Gel Filtration HMW Calibration Kit, GE Healthcare) in SEC buffer. The protein concentration was measured spectrophotometrically using the molar absorption coefficient for USP2-cleaved SbtB7001 (12950 M\textsuperscript{-1} cm\textsuperscript{-1}) as calculated using ProtParam (http://expasy.org/tools/protparam.html). SEC-purified, untagged SbtB7001 was used for isothermal titration calorimetry and crystallization.

\textbf{Isothermal Titration Calorimetry.} All isothermal titration calorimetry (ITC) experiments were performed on a Nano-ITC Low Volume calorimeter (TA Instruments). ITC experiments were performed at 25 °C with stirring at 350 rpm. Protein and ligand solutions were prepared in matched SEC buffer and degassed before use. Solutions of HCO\textsubscript{3}−, AMP, ADP, ATP, and cAMP (all sodium salts, Sigma) were prepared volumetrically. Titrations were performed using 100–400 \mu M (monomer concentration) USP2-cleaved SbtB7001 and typically involved 1 × 1 \mu L followed by 22 × 2 \mu L injections of 2–25 mM ligand with 250–300 s intervals between injections. Data from each titration were analyzed using NITPIC\textsuperscript{15} and SEDPHAT;\textsuperscript{16} the baseline-subtracted power fitting was achieved by iteratively fitting of algorithms in SEDPHAT until modeling parameters converged. Model parameters were obtained through the global fitting of...
single titrations or replicates; 68.3% confidence intervals were calculated using the automatic confidence interval search with the projection method using F-statistics in SEDPHAT. Other binding parameters and associated confidence intervals were obtained by propagation.

**Protein Crystallization.** High-throughput, sitting drop vapor diffusion crystallization screens and subsequent hanging drop optimization screens were set up at 18 or 4 °C in house or at the Collaborative Crystallisation Centre (C3, CSIRO), and crystals formed under a number of conditions. For apo-SbtB7001 crystals, SEC-purified, untagged SbtB7001 was transferred to MES crystallization buffer [50 mM MES (pH 6.0) and 100 mM NaCl]. Crystals in sitting drops at 18 °C containing 0.15 μL of a 10 mg/mL protein solution and 0.15 μL of a reservoir solution (0.025 M disodium hydrogen phosphate and 20% PEG 3350) were cryoprotected with a 30% (v/v) 1:1 glycerol/ethylene glycol mixture. Crystals of AMP:SbtB7001 formed at 18 °C in hanging drops containing 1 μL of a protein solution [22 mg/mL untagged SbtB7001 in 20 mM HEPES, 100 mM NaCl (pH 8.0), 10 mM MgCl₂, 12 mM AMP, and 50 mM NaHCO₃] and 1 μL of a reservoir solution (0.2 M sodium acetate and 18% PEG 3350). These crystals were cryoprotected using 30% (v/v) 2-methyl-2,4-pentanediol (MPD) and 100 mM NaHCO₃. Crystals of ADP:SbtB7001 formed at 18 °C in sitting drops containing 0.15 μL of a protein solution [22 mg/mL untagged SbtB7001 in 20 mM HEPES, 100 mM NaCl (pH 8.0), 10 mM MgCl₂, 12 mM ADP, and 50 mM NaHCO₃] and 0.15 μL of a reservoir solution (2.5 M ammonium sulfate). Crystals were cryoprotected with 30% (v/v) MPD and 100 mM NaHCO₃. Crystals of Ca²⁺:AMP:SbtB7001 formed at 18 °C in hanging drops containing 1 μL of a protein solution [22 mg/mL untagged SbtB7001 in 20 mM HEPES, 100 mM NaCl (pH 8.0), 10 mM MgCl₂, 12 mM AMP, and 50 mM NaHCO₃] and 1 μL of a reservoir solution (0.2 M potassium acetate and 18% PEG 3350). These crystals were cryoprotected using 30% (v/v) MPD and 100 mM NaHCO₃. Crystals of Ca²⁺:ADP:SbtB7001 formed at 18 °C in hanging drops containing 1 μL of a protein solution [22 mg/mL untagged SbtB7001 in 20 mM HEPES, 100 mM NaCl (pH 8.0), 10 mM MgCl₂, 12 mM ADP, and 50 mM NaHCO₃] and 1 μL of a reservoir solution (0.2 M potassium acetate and 16% PEG 3350). These crystals were cryoprotected with 30% (v/v) MPD and 100 mM NaHCO₃. Crystals of Ca²⁺:AMP:SbtB7001 formed at 18 °C in sitting drops containing 0.15 μL of a protein solution [22 mg/mL untagged SbtB7001 in 20 mM Tris, 100 mM NaCl (pH 8.0), 10 mM MgCl₂, 12 mM AMP, and 20 mM sodium bicarbonate] and 0.15 μL of a reservoir solution [0.2 M lithium chloride, 20% PEG 6000, and 100 mM sodium acetate-acetic acid (pH 5)]. Crystals were cryoprotected with 30% (v/v) glycerol. Crystals of Ca²⁺:ATP:SbtB7001 formed at 18 °C in sitting drops containing 0.15 μL of a protein solution [20 mg/mL untagged SbtB7001 in 20 mM Tris, 100 mM NaCl (pH 8.0), 10 mM MgCl₂, 12 mM ATP, and 20 mM NaHCO₃] and 0.15 μL of a reservoir solution [30% MPD, 0.02 M CaCl₂, and 0.1 M sodium acetate-acetic acid (pH 4.6)]. Crystals were cryoprotected with 30% (v/v) MPD.

All crystals were flash-cooled in liquid nitrogen. X-ray diffraction data were collected at beamline MX2 at The Australian Synchrotron. Data were processed using XDS17 and Aimless,18 and molecular replacement was performed using either MolRep19 or Phaser.20 Chain A of apo-SbtB6803 (PDB entry 5O3P, 1.7 Å) was used for molecular replacement of Ca²⁺:ATP:SbtB7001 structure. Iterative cycles of manual model building and refinement were performed using Coot 0.8.2,21 reﬁmac,22 and/or phenix.refine.23 TLS parameter refinement was also used, using TLS groups automatically selected by phenix.refine. Anisotropic B-factor refinement (all atoms except water) was used for refinement of Ca²⁺:ATP:SbtB7001 (1.04 Å) and ADP:SbtB7001 (1.42 Å). Models were optimized using the PDB_REDO server.24 Data collection and refinement statistics are provided in Table 2.

Table 1. Parameters for the Binding of Effector Molecules to SbtB7001, Measured by Isothermal Titration Calorimetry (ITC)²

<table>
<thead>
<tr>
<th>Effector</th>
<th>Kᵢ (μM)</th>
<th>ΔH (kJ/mol)</th>
<th>TΔS (kJ/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMP (n = 3) with 500 μM CaCl₂ (n = 2)</td>
<td>447 (398, 502)</td>
<td>-31.8 (-36.2, -29.5)</td>
<td>-12.7 (-17.4, -10.14)</td>
</tr>
<tr>
<td>ADP (n = 4) with 500 μM CaCl₂ (n = 2)</td>
<td>270 (221, 301)</td>
<td>-49.3 (-54.1, -37.7)</td>
<td>-29.0 (-34.0, -16.8)</td>
</tr>
<tr>
<td>cAMP (n = 4) with 500 μM CaCl₂ (n = 2)</td>
<td>238 (213, 266)</td>
<td>-60.8 (-65.6, -57.6)</td>
<td>-40.1 (-45.2, -36.6)</td>
</tr>
<tr>
<td>ATP (n = 5) with 10 mM MgCl₂ (n = 2)</td>
<td>144 (140, 147)</td>
<td>-67.6 (-68.1, -65.8)</td>
<td>-45.7 (-46.3, -43.8)</td>
</tr>
<tr>
<td>CaCl₂ (n = 2)</td>
<td>37.7 (33.2, 43.5)</td>
<td>-81.1 (-86.3, -76.3)</td>
<td>-55.9 (-61.4, -50.8)</td>
</tr>
<tr>
<td>with 400 μM ATP (n = 4)</td>
<td>58.9 (49.4, 66.5)</td>
<td>-11.3 (-11.9, -10.1)</td>
<td>12.6 (11.9, 14.5)</td>
</tr>
<tr>
<td>NaHCO₃ (n = 2)</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

*ITC thermograms are provided in Figures S1–S15. Binding parameters were determined by ITC at 25 °C in 20 mM HEPES and 150 mM NaCl (pH 8.0) using a single-site binding model. Global fitting was performed using data from replicate titrations (number of replicates indicated in table) to obtain best-fit values. The 68.3% confidence intervals for the fit of the model to the data are given in parentheses. ND, specific binding not detected.

## RESULTS

**Nucleotide Binding to SbtB7001.** We used isothermal titration calorimetry (ITC) to investigate the affinity of SbtB7001 for the proposed P2P protein ligands AMP, ADP, cAMP, and ATP (Table 1 and Figures S1–S15). Titrations with AMP (Figure S1) gave a macroscopic dissociation constant (Kᵢ) averaged across the three binding sites in the homotramer of approximately 450 μM, with an enthalpic contribution to binding (ΔH) of -32 kJ/mol and an entropic cost (TΔS) of -12.7 kJ/mol. Titrations with ADP (Figure S3) and cAMP (Figure S5) gave intermediate Kᵢ values of 224 and 238 μM, respectively, with similar enthalpic (ΔH) values of -44 kJ/mol for ADP and -61 kJ/mol for cAMP and entropic contributions (TΔS values of -23.3 kJ/mol for ADP and 24.3 kJ/mol for cAMP).
---40.1 kJ/mol for cAMP) to binding. In contrast, ATP showed the highest binding affinity with a K_D of approximately 37 μM (Figure S7). The binding of ATP was associated with a much larger change in enthalpy (ΔH of −81.1 kJ/mol), indicative of the formation of numerous additional electrostatic interactions between ATP and SbtB7001. In addition, ATP binding resulted in a substantial loss of entropy (−ΔS of −55.9 kJ/mol), which is consistent with the stabilization of a mobile region of the protein upon binding of ATP.

Because Mg^{2+} is often required for the binding of ATP to PII proteins, we also tested the affinity of SbtB7001 for ATP in the presence of 10 mM Mg^{2+} (Figure S8) and in the presence of the Mg^{2+} chelator EDTA (Figure S9). The resulting values for the K_D, ΔH, and ΔS of binding were not significantly different from those for the case in which ATP was tested alone. We then tested the effect of Ca^{2+}, which can also coordinate nucleotides. 25,26 ITC measurements revealed that the affinity for ATP was increased 10-fold, to ~4 μM (approximately 100-fold higher than that for AMP in the absence of Ca^{2+}) when 7 mM Ca^{2+} was present in the reaction mixture (Figure S10). In the presence of a lower concentration of Ca^{2+} (500 μM), ATP binding was still enhanced, with a K_D of ~7 μM (Table 1 and Figure S11). Knowledge of the upper (in the presence of saturating Ca^{2+}) and lower (in the absence of Ca^{2+}) limits for the K_D for ATP allows the affinity for ATP at intermediate Ca^{2+} concentrations to be estimated. In the absence of nucleotides, binding of Ca^{2+} could not be detected by ITC (Figure S12), suggesting that binding of Ca^{2+} to SbtB7001 is dependent on prebound ATP or that Ca^{2+} and ATP bind cooperatively. Indeed, when Ca^{2+} was titrated into a solution of SbtB7001 in the presence of saturating concentrations of ATP (400 μM, >10 K_D), isotherms indicated that Ca^{2+} bound with a K_D of approximately 60 μM (Table 1 and Figure S13). The effect of Ca^{2+} on the binding of nucleotides was most pronounced for ATP; K_D values for cAMP (Figure S6) and AMP (Figure S2) were reduced by factors of <2 and <4, respectively, in the presence of 500 μM Ca^{2+}, and the K_D of ADP was unaffected by Ca^{2+} (Table 1 and Figure S4). Because of their known common interaction with nucleotides, Ca^{2+} and Mg^{2+} were tested; no other metals were tested at this time.

Despite the role of SbtB in regulating SbtA-mediated HCO_3^- transport, specific binding of HCO_3^- to SbtB7001 could not be detected using ITC under the conditions tested [i.e., K_D for HCO_3^- > 10 mM (Figure S14)]. At the concentration of HCO_3^- used in these titrations, the minor changes in heat observed in the control-subtracted thermograms are most likely due to nonspecific binding events or artifacts. Indeed, when the same concentration of HCO_3^- was titrated into the lac repressor (LacI) from E. coli, similar heats were observed (Figure S15).

### Table 2: Data Collection and Refinement Statistics for Structures Described in This Work

<table>
<thead>
<tr>
<th>PDB entry</th>
<th>apo-SbtB7001</th>
<th>AMP:SbtB7001</th>
<th>ADP:SbtB7001</th>
<th>cAMP:SbtB7001</th>
<th>ATP:SbtB7001</th>
<th>Ca^{2+}:ATP:SbtB7001</th>
</tr>
</thead>
<tbody>
<tr>
<td>space group</td>
<td>6N4A</td>
<td>6MMO</td>
<td>6MMIC</td>
<td>6MMQQ</td>
<td>6NTB</td>
<td>6MM2</td>
</tr>
<tr>
<td>cell dimensions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>46.6, 51.0, 66.6</td>
<td>46.7, 51.3, 66.8</td>
<td>46.8, 68.6, 89.1</td>
<td>46.6, 50.2, 66.6</td>
<td>132.9, 132.9, 33.8</td>
<td>52.0, 52.1, 95.2</td>
</tr>
<tr>
<td>α, β, γ (deg)</td>
<td>94.3, 106.1, 103.6</td>
<td>97.9, 102.5, 103.1</td>
<td>90, 90, 120</td>
<td>97.6, 102.9, 102.5</td>
<td>90, 90, 120</td>
<td>90, 90, 120</td>
</tr>
<tr>
<td>resolution (Å)</td>
<td>49.03–2.30</td>
<td>43.22–1.86</td>
<td>49.44–1.42</td>
<td>35.32–2.02</td>
<td>45.32–1.90</td>
<td>15.03–1.04</td>
</tr>
<tr>
<td>Rmerge (overall/inner/outer)</td>
<td>0.040/0.015/1.05</td>
<td>0.046/0.033/1.922</td>
<td>0.105/0.037/6.445</td>
<td>0.05/0.029/1.270</td>
<td>0.107/0.054/2.680</td>
<td>0.039/0.035/1.337</td>
</tr>
<tr>
<td>CC12</td>
<td>0.999/0.999/0.574</td>
<td>0.998/0.996/0.335</td>
<td>1.000/1.000/0.380</td>
<td>0.998/0.998/0.307</td>
<td>0.999/0.999/0.560</td>
<td>0.999/0.997/0.405</td>
</tr>
<tr>
<td>completeness (%)</td>
<td>98.0/96.6/95.2</td>
<td>97.1/86.5/96.3</td>
<td>100/99.7/100.0</td>
<td>96.4/93.5/96.6</td>
<td>100/99.3/100.0</td>
<td>95.2/53.4/50.9</td>
</tr>
<tr>
<td>multiplicity</td>
<td>3.8/3.8/3.8</td>
<td>3.6/3.2/3.8</td>
<td>37.8/26.9/36.3</td>
<td>2.4/2.4/2.5</td>
<td>20/17.7/18.1</td>
<td>9.0/13.4/5.5</td>
</tr>
<tr>
<td>resolution (Å)</td>
<td>49.03–2.30</td>
<td>41.58–1.86</td>
<td>49.44–1.42</td>
<td>35.32–2.02</td>
<td>45.32–1.90</td>
<td>15.03–1.04</td>
</tr>
<tr>
<td>no. of reflections</td>
<td>24661 (2413)</td>
<td>47132 (4612)</td>
<td>24048 (2343)</td>
<td>35629 (3583)</td>
<td>27442 (2714)</td>
<td>44104 (2995)</td>
</tr>
<tr>
<td>Rwork/Rfree</td>
<td>0.2170/0.2767</td>
<td>0.2234/0.2579</td>
<td>0.2029/0.2202</td>
<td>0.2208/0.2794</td>
<td>0.210/0.2475</td>
<td>0.1452/1.632</td>
</tr>
<tr>
<td>components in the asymmetric unit</td>
<td>two trimers (six chains)</td>
<td>two trimers (six chains)</td>
<td>one chain</td>
<td>two trimers (six chains)</td>
<td>one chain</td>
<td>one chain</td>
</tr>
<tr>
<td>no. of atoms</td>
<td>3989</td>
<td>4189</td>
<td>788</td>
<td>4142</td>
<td>2541</td>
<td>997</td>
</tr>
<tr>
<td>protein</td>
<td>3961</td>
<td>4002</td>
<td>722</td>
<td>3985</td>
<td>2350</td>
<td>845</td>
</tr>
<tr>
<td>ligands/ions</td>
<td>10</td>
<td>138</td>
<td>1</td>
<td>132</td>
<td>95</td>
<td>33</td>
</tr>
<tr>
<td>solvent</td>
<td>18</td>
<td>49</td>
<td>65</td>
<td>34</td>
<td>96</td>
<td>119</td>
</tr>
<tr>
<td>average β-factor</td>
<td>75.0</td>
<td>61.25</td>
<td>27.31</td>
<td>66.26</td>
<td>45.8</td>
<td>24.1</td>
</tr>
<tr>
<td>protein</td>
<td>75.1</td>
<td>61.3</td>
<td>26.44</td>
<td>66.28</td>
<td>45.9</td>
<td>20.9</td>
</tr>
<tr>
<td>ligands/ions</td>
<td>84.12</td>
<td>59.9</td>
<td>28.82</td>
<td>60.20</td>
<td>38.8</td>
<td>15.5</td>
</tr>
<tr>
<td>solvent</td>
<td>56.5</td>
<td>57.0</td>
<td>36.94</td>
<td>64.02</td>
<td>49.9</td>
<td>48.7</td>
</tr>
<tr>
<td>RMSD</td>
<td>0.011</td>
<td>0.007</td>
<td>0.005</td>
<td>0.008</td>
<td>0.007</td>
<td>0.013</td>
</tr>
<tr>
<td>bond angles (deg)</td>
<td>1.46</td>
<td>0.91</td>
<td>0.84</td>
<td>1</td>
<td>1.17</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Statistics for the highest-resolution shell are shown in parentheses.
Figure 1. Overall fold of SbtB7001. (A) Size-exclusion chromatogram of SbtB7001 (left) and calibration curve for analytical size-exclusion chromatography (right). Empty circles represent molecular weight standards, and the closed circle represents untagged SbtB7001. The calculated molecular weight of SbtB7001 is consistent with a trimeric structure (calcd, 33.7 kDa; theoretical, 34.5 kDa; for the trimer). (B) Crystal structure of apo-SbtB7001 (chains A−C, colored by chain) showing the conserved trimeric PII-like fold. The approximate positions of the flexible T-loop regions are indicated by the dashed lines, and the locations of the B-loop and C-loop are labeled. (C) Structural alignment of apo-SbtB7001 (green) and apo-SbtB6803 (PDB entry 5O3P, white) showing the conserved PII-like fold. (D) Nucleotides (e.g., ATP, shown as a ball-and-stick model) bind at the interface between subunits. The surface representation of SbtB7001 with T-loop and C-loop residues removed highlights the accessibility of the nucleotide binding site. Lines indicate the approximate position of “open” (dashed line) and “closed” (solid line) states of the T-loop and C-loop. (E) Sequence alignment of SbtB7001, SbtB homologues from other cyanobacterial strains, and related PII and PII-like proteins. Residues important for interactions with the phosphates of ATP are indicated by black circles. The primary and secondary structure of SbtB7001 is shown above the alignment. Sequences were aligned using Clustal Omega.27 and the figure was produced using ESPtcript 3.0.28 Abbreviations: SbtB7001, *Cyanobium* sp. PCC7001 (WP_011244772.1); SbtB6307, *Cyanobium* sp. PCC6307 (WP_015110057.1); SbtB5701, *Synechococcus* WH5701 (WP_037981031.1); SbtB7002, *Synechococcus* sp. PCC77002 (WP_012306104.1); AvSbtB, *Anaabaena* variabilis (yp_323533.1); TiCPII, carboxysome-associated PII protein (CPII) from *Thiimonas intermedia* K12 (WP_013104276.1); EcNPR, PII protein from *E. coli* (CCK45606.1).

Binding of AMP, ADP, and cAMP at the Intersubunit Clefts of SbtB7001. When SbtB7001 was co-crystallized with AMP (Figure 2A), ADP (Figure 2B), and cAMP (Figure 2C), electron density was observed for the respective ligands within each of the clefts between the monomers. The ADP:SbtB7001 co-crystal displayed ambiguous electron density around the β-phosphate of ADP, consistent with either a mobile β-phosphate or hydrolysis of ADP and occupancy by AMP (as has been observed in other studies of related PII proteins5,6,29).

Each nucleotide binds with the shared adenine motif sitting within a mostly hydrophobic cleft (lined by the side chains of Ile11, Leu65, and Ile90), while the backbone carbonyl and amides of Gly31 and Leu65 and the hydroxyl groups of Ser30 and Thr33 form specific polar interactions with the base. The hydroxyl groups of the ribose sugar form hydrogen bonds with the side chains of Thr33 and Asn59 in each case, and the C-loop is colored yellow. A dashed line indicates the approximate location of the unmodeled, flexible T-loop. Coordinated water molecules are shown as cyan spheres.

Figure 2. Ligand binding sites of SbtB7001. Representative binding sites from structures of SbtB7001 co-crystallized with (A) AMP, (B) ADP, and (C) cAMP. Ligands (yellow ball-and-stick models) bind in the clefts between subunits (neighboring subunits are colored white and purple). Key residues for binding are shown as sticks. F3−F5. Polder omit maps are shown as a green mesh contoured at 3σ for the ADP:SbtB7001 structure.

trimeric in solution (Figure 1A), and all crystal structures showed a trimeric PII-like core structure (Figure 1B) that was essentially identical to that of SbtB6803, with a root-mean-square deviation (RMSD) of 0.45 Å between monomer α-carbons of apo-SbtB6803 (PDB entry 5O3P) and apo-SbtB7001 (Figure 1C). The canonical nucleotide binding sites, which sit between the subunits of the trimer, can be easily accessed from the solvent in apo-SbtB7001 (Figure 1D). SbtB7001 shares 56% sequence identity with SbtB6803 (Figure 1E), which is consistent with the low RMSD value. However, SbtB6803 contains a C-terminal extension that is present in only a small subset of SbtB homologues, while SbtB7001 does not contain this extension.

(SbtB7001, 625.4x818.4)
Å for AMP:SbtB7001 and ∼3.3 Å for ADP:SbtB7001), as well as with the backbone of Cys89 (∼3.5 and 2.8 Å, respectively) and Ser42 (2.7 and 2.8 Å, respectively). The interaction between the α-phosphate and the backbone of Ser42 and Gly41 seems to be important for stabilizing residues 40–42; in apo-SbtB7001 and cAMP:SbtB7001, these interactions are not observed, and this region becomes unstructured. The additional interactions between the β-phosphate of ADP and Ser88 (distance of 2.3 Å) and Tyr57 (2.6 Å) are consistent with the ITC experiments that reveal a higher affinity [Kₐ = 225 μM vs 387 μM (Table 1)] and a larger change in enthalpy upon binding of ADP compared with AMP (Table 1). Any interactions between AMP/ADP and T-loop residues 40–56 are likely to be only transient, at most, because additional interactions with T-loop residues were not observed in the crystal structures and the electron density around the β-phosphate of ADP is consistent with a mobile β-phosphate that is not restrained by such interactions.

Due to the cyclic nature of cAMP, its binding mode is somewhat different from that of AMP/ADP; the interaction with Ser42 is lost, and the phosphate forms an additional interaction with side chain of Asn59 (distance of 3.1 Å) (Figure 2C). Asn59 is conserved among SbtB homologues (Figure 1E); the Asn59–cAMP interaction is likely important for conferring a higher affinity for cAMP over AMP in SbtB7001 [Kₐ = 238 μM for cAMP vs 447 μM for AMP (Table 1)] and SbtB6803.⁶

**Ca²⁺ and ATP Stabilize the T-Loop.** In contrast to AMP, ADP, and cAMP, the γ-phosphate of ATP supports additional interactions with the residues at the base of the T-loop (residues 40–46) (Figure 3A,B), explaining why SbtB7001 binds ATP with a much higher affinity and has a more favorable enthalpy of binding compared with AMP, ADP, or cAMP. The salt bridge formed between the γ-phosphate of ATP and the side chain of Arg46 appears to be particularly important, given the high degree of conservation of this residue among SbtB homologues (Figure 1E). This ATP-specific interaction appears to stabilize an interaction between Arg46 and Asn59, which pulls the base of the T-loop toward the core of the protein, where it tightly encircles the phosphate groups of the ligand. In canonical PII proteins, the residue at the position analogous to Asn59 in SbtB7001 is involved in forming a key salt bridge with T-loop residues, and the formation and breaking of this salt bridge has been shown to play an important role in conferring functionally relevant T-loop conformations.³⁰,³¹ In SbtB7001, when the T-loop is held in position by the Arg46–Asn59 interaction, additional interactions are also formed between the phosphates of ATP, Ser88, Arg43, and multiple points on the backbone of the T-loop, which further stabilize this region; the average B-factors...
for Ca atoms of residues 40−46 and 47−55 are 19.2 and 25.0 Å², respectively. While the interaction between the α-phosphate and Arg43 could be possible when AMP or ADP is bound, it seems that the additional ATP-specific interaction with Arg46 is required to stabilize this basal T-loop region and promote this interaction between the α-phosphate and Arg43. Thus, in comparison to the largely disordered T-loop residues 43−55 in apo-SbtB7001, AMP:SbtB7001, ADP:SbtB7001, and cAMP:SbtB7001 (as was also observed in structures of SbtB6803 bound to AMP and cAMP3), the binding of ATP to SbtB7001 resulted in stabilization of the T-loop. These structural observations are consistent with our ITC experiments for ATP, which revealed a substantially greater entropic cost associated with ATP binding, compared with the binding of AMP, ADP, and cAMP (Table 1).

Although both ATP:SbtB7001 and Ca²⁺ATP:SbtB7001 co-crystals were grown in the presence of 10 mM Mg²⁺, electron density consistent with an octahedrally coordinated metal ion near the γ-phosphate of ATP was observed only in the crystal grown in the presence of 20 mM Ca²⁺ (Figure 3B). Metal−oxygen distances (~2.4 Å) and environmental B-factor analysis (B-factor of 20.38 for Ca²⁺, averaged B-factor of 21.94 for coordinated atoms) are consistent with a Ca²⁺ binding site. This was confirmed using the "place elemental ions" tool in phenix.refine (valence sum for Ca²⁺ of 1.989, expected 2; valence sum for Mg²⁺ of 0.948, expected 2).35 The Ca²⁺ ion is coordinated by the side chains of Glu50 and Asn44, the backbone carboxyl of Gly41, the β- and γ-phosphates of ATP, and a single water molecule that is held in place by Ser54 (Figure 3C). The interactions with Glu50 and Asn44 appear to be important for stabilizing the apex of the T-loop; when Ca²⁺ was absent from the crystallization conditions, the complete T-loop could be completely modeled in only one of the three chains (ATP:SbtB7001, chain C). In chain A, the T-loop residues that could be modeled adopt similar conformations as in chain C, except for residues 50−57, which have become disordered, and Asn44, which adopts an alternate rotamer. In chain B, all residues except unmodeled residues 49−54 adopt similar conformations as in chain C. These results are consistent with the apex region of the T-loop (between residues 49 and 57) being less ordered in the absence of Ca²⁺. Even when this region could be completely modeled, the T-loop of ATP:SbtB7001 exhibits high B-factors; the B-factors for the Ca atoms of residues 40−46 and 47−55 of chain C were 47.7 and 61.0 Å², respectively. Moreover, this loop adopts a conformation distinct from that seen in Ca²⁺ATP:SbtB7001 (Figure 3D). Asn44 and Tyr57 flip away from the Ca²⁺ binding site; Glu50 moves away from the ATP binding site, and Pro51 adopts a cis conformation (Figure S17). The trans−cis interconversion of Pro51 is notable, given that proline isomerization has been noted as a regulatory mechanism in other signaling proteins.35 Together, it seems that the Ca²⁺ specific interactions further stabilize the "closed" conformation of the T-loop, which is consistent with our observation from ITC that the affinity for ATP was increased in the presence of Ca²⁺, but not Mg²⁺. This "closed" Ca²⁺ATP:SbtB7001 T-loop conformation almost completely seals the nucleotide binding site from one side, leaving only a narrow channel near the β- and γ-phosphates that could allow for the passage of an ion to and from the metal binding site, as well a thin access tunnel near the C-terminus of the protein (Figure 3E). In contrast, the flexibility observed in the ATP:SbtB7001 T-loop in the absence of Ca²⁺ means that the phosphate tail of ATP is not completely occluded by the T-loop.

The Dynamic C-loop Forms Part of the Nucleotide Binding Site. The C-terminal C-loop forms part of the nucleotide binding site in canonical PII proteins.34 In SbtB7001, C-terminal residues Leu99−Phe104 could be completely modeled only in AMP:SbtB7001 (chain A only), cAMP:SbtB7001 (chain B only), ADP:SbtB7001, and ATP:SbtB7001 (Figure 4). All other chains showed poor electron density in this region, consistent with conformational heterogeneity. In crystals where the C-terminal residues could be modeled, this region occupies part of the intersubunit cleft, where it is within hydrogen bonding distance (~3 Å) of the α-phosphate of the ligand (Figure 4B). While this interaction could aid in binding the α-phosphate of ADP/AMP, it is sterically incompatible with the "closed" position of the T-loop observed in Ca²⁺ATP:SbtB7001 and ATP:SbtB7001, clashing
with the position of Arg43 (Figure 4C). The rigid C-terminal helix observed in SbtB6803 holds His102 in a similar location and would cause a similar steric clash with the position of Arg43 that appears to be important for ATP binding (Figure 4D).

In AMP:SbtB7001 [chain A (Figure 4E)] and cAMP:SbtB7001 [chain B (Figure 4F)], β4 continues to run antiparallel with β1 until Ala101 before wrapping back around toward the ligand binding site. In this more compact conformation, His102 is no longer within bonding distance of the ligand. The C-loop of ATP:SbtB7001 traces a similar route, although His102 adopts an alternate rotamer that positions it within bonding distance of the α-phosphate of the ligand (Figure 4G). Unlike the extended conformation observed in apo-SbtB7001 and ADP:SbtB7001, the more compact conformation observed in AMP:SbtB7001, cAMP:SbtB7001, and ATP:SbtB7001 is compatible with the closed conformation of the T-loop (e.g., as observed in ATP:SbtB7001). When both the T-loop and the C-loop adopt a closed conformation (as in ATP:SbtB7001), the nucleotide binding site access is almost completely occluded.

In the related, carboxysome-associated Pii protein from Thiomonas intermedia (TcPii), HCO₃⁻ binds near the C-loop and alters the affinity of the protein for nucleotides. While crystals that yielded structures of AMP:SbtB7001, ADP:SbtB7001, cAMP:SbtB7001, ATP:SbtB7001, and Ca²⁺ATP:SbtB7001 were grown in the presence of 20–100 mM NaHCO₃, electron density that unambiguously corresponded to HCO₃⁻ could not be identified in any of the structures. This is consistent with our ITC results that showed only low-affinity binding of HCO₃⁻ to SbtB7001, consistent with nonspecific protein stabilization.

**DISCUSSION**

In this work, we have shown that, like SbtB6803, SbtB7001 binds AMP, ADP, cAMP, and ATP with micromolar-range affinities. However, in contrast to SbtB6803, we found that SbtB7001 binds ATP with an affinity that is 5–10-fold greater than those of the other nucleotides tested and that the affinity for ATP increases a further 10-fold (from a $K_D$ of ~4 μM in the presence of Ca²⁺). In contrast, the presence of Ca²⁺ had a weaker effect on the binding of AMP, ADP, or cAMP. ITC experiments revealed that binding of ATP and Ca²⁺ATP to SbtB7001 is associated with a large change in enthalpy and entropy, consistent with the formation of a number of charged interactions between the ligand and SbtB7001 that cause a significant stabilization of a mobile region of the protein, respectively. Structures of SbtB7001 co-crystallized with ligands support the ITC results and (i) provide a structural explanation for the preferential binding of ATP, (ii) account for differences in the affinities of SbtB7001 and those reported for SbtB6803, and (iii) reveal how binding of ATP stabilizes the otherwise flexible T-loop. Together, these results provide a structural explanation for how SbtB could regulate SbtA, via changes in the adenylate charge ratio ([ATP]/[ATP + ADP + AMP]) observed in a transition from light to darkness in cyanobacteria.

The mode of binding of ATP to SbtB7001 is distinct from the mode of ATP binding in the canonical Pii proteins, in which a conserved RXR motif in the C-terminal C-loop is responsible for forming interactions with the terminal phosphates of the ligand. In SbtB7001, these interactions are instead provided by arginine residues from the T-loop.

Indeed, in SbtB7001 [and the majority of the cyanobacterial SbtBs (Figure 1E)], the C-terminal extension is largely truncated, suggesting that it is less essential for ligand binding in SbtBs from most cyanobacteria. The reduced C-loop in SbtB7001 displayed conformational heterogeneity, adopting extended and compact states, with the extended conformation being incompatible with Ca²⁺ATP binding. It is unlikely that the C-loop substantially contributes to the differentiation among AMP, ADP, and ATP, because C-loop residues (e.g., His102) are unlikely to be able to form interactions with the β- or γ-phosphates of ADP or ATP. It is also unclear why the conformation of the C-loop is different between AMP- and ADP-bound structures, for example. In this case, we cannot exclude the possibility that it is a crystallization artifact.

The T-loop residues involved in binding ATP are highly conserved among most SbtB homologues (Figure 1E), suggesting that all SbtB homologues should bind ATP more tightly than AMP, ADP, or cAMP. However, this was not observed for Sbtb6803. The C-terminal extension found in SbtB6803, which is present in only a small subset of SbtB homologues, may contribute to the lower affinity between SbtB6803 and ATP. Indeed, on the basis of the structures presented here, the rigid, disulfide-linked C-terminal region of SbtB6803 would clash with the position that Arg43 is required to adopt for the T-loop to accommodate ATP. Interestingly, the C-terminal region is unresolved in the structure of a related SbtB from Anabaena variabilis ATCC 29413 (AvSbtB, PDB entry 3DFE), which also has the putative redox-sensing C-terminal extension, suggesting that this region can switch between open and closed states. It is therefore possible that the C-terminal extension found in SbtB6803 and a small subset of SbtBs could alter the affinity of SbtB for ATP based on the cellular redox conditions of the cell by preventing conformations of the T-loop that are required for ATP binding. The ATP binding activity of SbtB6803 was not tested in different reducing or oxidizing conditions, and it is possible that SbtB6803 has a higher affinity for ATP under certain redox conditions. In contrast, the residues involved in binding ATP and Ca²⁺ATP in SbtB7001 are not found in the structurally related TcPii (which undergoes ligand-induced T-loop conformational changes) (Figure 1E), consistent with these proteins having different functions in cyanobacteria.

In the cyanobacterial cell, the ligand-bound state of SbtB will depend on both the affinity of the ligands and the local concentrations of each nucleotide and/or effector molecules. Our results suggest that the apo form of SbtB7001 is unlikely to be physiologically relevant because there will always be AMP, ADP, cAMP, or ATP present at concentrations in excess of the $K_D$. The [ATP]/[ATP + ADP] ratio has been reported to be around 0.8 in Synechococcus under illumination but varies significantly in cyanobacteria in response to changing light and C₄ availability. For example, the [ATP]/[ADP + ATP] ratio decreases from 0.9 to ~0.4 within 5–6 h of S. elongatus PCC7942 being moved to darkness and returns to ~0.85 within 1 h of being returned to the light. Thus, when ATP levels decrease (such as during darkness), the binding sites of SbtB7001 are likely to become occupied by ADP or cAMP rather than ATP. Because the SbtA–SbtB6803 complex is stabilized by AMP and ADP, we expect that binding of these ligands would also stabilize the inactive SbtA–SbtB7001 complex (with open T-loops) to prevent futile metabolic cycling in the dark. We envision that when ATP levels increase again (upon exposure to light), ATP...
could displace AMP/ADP and cause the T-loops to adopt the closed conformation, which we envisage would be the allosteric signal to cause the complex to dissociate. Further in vivo studies will be required to test this model.

The finding that Ca\(^{2+}\) enhances the binding of ATP to SbtB7001 raises the possibility that Ca\(^{2+}\) might act as a second messenger to influence how SbtB regulates SbtA-mediated transport. Interestingly, transient increases in the Ca\(^{2+}\) concentration have been shown to upregulate expression of CCM-related genes, including the sbt operon, in *Anabaena* sp. PCC7120.\(^{38}\) Ca\(^{2+}\) is known to play an important role in the regulation of other cyanobacterial systems, as well. For example, Ca\(^{2+}\) acts as a signal for heterocyst formation\(^{39}\) and other regulatory proteins involved in sensing C\(_{\text{p}}\) such as soluble adenylyl cyclases, can simultaneously sense ATP, Ca\(^{2+}\), and HCO\(_3^-\)/CO\(_2\)/pH.\(^{40}\) Further evidence that Ca\(^{2+}\) plays an important role in the regulation of cyanobacterial HCO\(_3^-\) transport includes a structure of the periplasmic domain of another cyanobacterial HCO\(_3^-\) transporter, CmpA, which revealed that HCO\(_3^-\) bound strongly only if Ca\(^{2+}\) was present.\(^{41}\) Future experiments could focus on testing whether physiological levels of Ca\(^{2+}\) could act as a secondary effector for SbtB in vivo because free calcium levels in cyanobacteria can vary between 100 and 200 nM under normal conditions\(^{42}\) and can reach low micromolar-range concentrations following light to dark transitions\(^{36}\) or heat, cold, or ammonia shock.\(^{37}\) If ATP binding does cause the complex to dissociate, then variations in the level of Ca\(^{2+}\) in the cell could further control this equilibrium by altering the affinity of SbtB for ATP. For example, transient increases in the concentration of free Ca\(^{2+}\) following light-to-dark transitions could conceivably help delay the formation of the SbtA–SbtB complex following the onset of darkness.

In summary, this study has provided a molecular basis for the design of future in vivo studies to test a number of hypotheses relating to the regulation of the SbtA–SbtB complex that will be required to propose a comprehensive model describing the regulation of SbtA-mediated HCO\(_3^-\) transport by SbtB. First, increasing ATP concentrations (due to exposure to light) are likely to be an allosteric signal to release the inhibitory SbtB protein from the SbtA HCO\(_3^-\) transporter. Second, fluctuations in Ca\(^{2+}\) levels are likely to modulate the effect of changing nucleotide concentrations. Third, the C-terminal extension, which is present in a subset of SbtBs, may play an additional role in modulating allosteric signaling by controlling ATP binding in a redox-sensitive fashion. These findings will guide future experiments that will ultimately contribute to attempts to use transgenic SbtA to improve the photosynthetic performance and yield of C\(_3\) crop plants by increasing the steady state concentration of C\(_{\text{p}}\) in the chloroplast stroma.\(^{5}\)

### ASSOCIATED CONTENT

**Supporting Information**

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.9b00880.

ITC thermograms (Figures S1–S15), structural heterogeneity of individual chains of SbtB7001 in apo and ligand-bound states (Figure S16), and difference map of the region near Pro51 in ATP:SbtB7001 (Figure S17) (PDF)

### Accession Codes

The atomic coordinates and structure factors for apo-SbtB7001 (6N4A), AMP:SbtB7001 (6MMO), ADP:SbtB7001 (6MMC), cAMP:SbtB7001 (6MMQ), ATP:SbtB7001 (6NTB), and Ca\(^{2+}\)ATP (6MM2) have been deposited in the Protein Data Bank.

### AUTHOR INFORMATION

**Corresponding Authors**

*E-mail: colin.jackson@anu.edu.au.*

*E-mail: dean.price@anu.edu.au.*

**ORCID**

Joe A. Kaczmarski: 0000-0001-8549-0122

Colin J. Jackson: 0000-0001-6150-3822

**Funding**

This work was supported by the Australian Government through the Australian Research Council Centre of Excellence for Translational Photosynthesis (CE140100001S). J.A.K. acknowledges support from an Australian Government Research Training Program Scholarship.

**Notes**

The authors declare no competing financial interest.

### ACKNOWLEDGMENTS

The authors thank the staff of the Collaborative Crystallisation Centre (C3, CSIRO) for assistance in crystallizing these proteins and Li Lynn Tan (ANU) and Australian Synchrotron beamline scientists for help with freezing crystals and the collection of diffraction data.

### ABBREVIATIONS

C\(_{\text{p}}\), inorganic carbon; CCM, CO\(_2\)-concentrating mechanism; CPII, carboxysome-associated protein II; SbtA, sodium-dependent bicarbonate transporter A; SbtB, sodium-dependent bicarbonate transport protein B.

### REFERENCES


2.3 Supplementary Information for “Structural basis for the regulation of bicarbonate transport by the PII-like protein, SbtB, from *Cyanobium* sp. PCC 7001”
Supporting Information for Structural basis of SbtA-regulation by SbtB7001

SUPPORTING INFORMATION

Structural basis for the allosteric regulation of the SbtA bicarbonate transporter by the PII-like protein, SbtB, from Cyanobium sp. PCC7001

Joe A. Kaczmarski, Nan-Sook Hong, Bratati Mukherjee, Laura T. Wey, Loraine Rourke, Britta Förster, Thomas S. Peat, G. Dean Price and Colin J. Jackson.

Research School of Chemistry, 137 Sullivans Creek Road, The Australian National University, Canberra, ACT 0200, Australia

ARC Centre of Excellence for Translational Photosynthesis, Division of Plant Sciences, Research School of Biology, 134 Linnaeus Way, The Australian National University, Canberra, ACT 0200, Australia

CSIRO Biomedical Program, 343 Royal Parade, Parkville, VIC 3052, Australia
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

**Figure S1. ITC isotherms of titration of AMP into SbtB7001. (L-R) 1.** Cell = 102 μM SbtB7001 (monomer concentration); syringe = 20 mM AMP 2. Cell = 400 μM SbtB7001 (monomer concentration); syringe = 23 mM AMP 3. Cell = 143 μM SbtB7001 (monomer concentration); syringe = 20 mM AMP. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).

**Figure S2. ITC isotherms of titration of AMP into SbtB7001 in presence of CaCl₂. (L-R) 1.** Cell = 160 μM SbtB (monomer concentration) + 500 μM CaCl₂; syringe = 4 mM ATP + 500 μM CaCl₂ 2. Cell = 160 μM SbtB (monomer concentration) + 500 μM CaCl₂; syringe = 4 mM ATP + 500 μM CaCl₂. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

**ADP (n=4)**

Figure S3. ITC isotherms of titration of ADP into SbtB7001. (L-R) 1. Cell = 102 μM SbtB7001 (monomer concentration); syringe = 10 mM ADP 2. Cell = 110 μM SbtB7001 (monomer concentration); syringe = 5 mM ADP 3. Cell = 143 μM SbtB7001 (monomer concentration); syringe = 5 mM ADP 4. Cell = 160 μM SbtB7001 (monomer concentration); syringe = 4 mM ADP. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).

**ADP + 500 μM CaCl2 (n=2)**

Figure S4. ITC isotherms of titration of ADP into SbtB7001 in presence of CaCl2. (L-R) 1. Cell = 143 μM SbtB7001 (monomer concentration) + 500 μM CaCl2; syringe = 5 mM ADP + 500 μM CaCl2 2. Cell = 160 μM SbtB7001 (monomer concentration) + 500 μM CaCl2; syringe = 4 mM ADP + 500 μM CaCl2. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

**Figure S5.** ITC isotherms of titration of cAMP into SbtB7001. (L-R) 1. Cell = 110 μM SbtB7001 (monomer concentration); syringe = 2 mM cAMP 2. Cell = 143 μM SbtB7001 (monomer concentration); syringe = 5 mM cAMP 3. Cell = 160 μM SbtB7001 (monomer concentration); syringe = 5 mM cAMP 4. Cell = 160 μM SbtB7001 (monomer concentration); syringe = 4 mM cAMP. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).

**Figure S6.** ITC isotherms of titration of cAMP into SbtB7001 in presence of CaCl₂. (L-R) 1. Cell = 160 μM SbtB7001 (monomer concentration) + 500 μM CaCl₂; syringe = 4 mM cAMP + 500 μM CaCl₂ 2. Cell = 160 μM SbtB7001 (monomer concentration) + 500 μM CaCl₂; syringe = 4 mM cAMP + 500 μM CaCl₂. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).
Figure S7. ITC isotherms of titration of ATP into SbtB7001. (L-R) 1. Cell = 143 µM SbtB7001 (monomer concentration); syringe = 4 mM ATP. 2. Cell = 400 µM SbtB7001 (monomer concentration); syringe = 7 mM ATP. 3. Cell = 96 µM SbtB7001 (monomer concentration); syringe = 1.2 mM ATP. 4. Cell = 100 µM SbtB7001 (monomer concentration); syringe = 2 mM ATP. 5. Cell = 102 µM SbtB7001 (monomer concentration); syringe = 10 mM ATP. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

**Figure S8.** ITC isotherms of titration of ATP into SbtB7001 in the presence of MgCl₂. (L-R) 1. Cell = 400 μM SbtB7001 (monomer concentration) + 10 mM MgCl₂; syringe = 4 mM ATP + 10 mM MgCl₂ 2. Cell = 143 μM SbtB7001 (monomer concentration) + 10 mM MgCl₂; syringe = 4 mM ATP + 10 mM MgCl₂. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).

**Figure S9.** ITC isotherm of titration of ATP into SbtB7001 in the presence of EDTA. Cell = 400 μM SbtB7001 + 0.5 mM EDTA; syringe = 7 mM ATP + 0.5 mM EDTA. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

**Figure S10.** ITC isotherms of titration of ATP into SbtB7001 in the presence of 7 mM CaCl$_2$. (L-R) 1. Cell = 400 μM SbtB7001 (monomer concentration) + 7 mM CaCl$_2$; syringe = 7 mM ATP + 7 mM CaCl$_2$. 2. Cell = 143 μM SbtB7001 (monomer concentration) + 7 mM CaCl$_2$; syringe = 1.5 mM ATP + 7 mM CaCl$_2$. 3. Cell = 143 μM SbtB7001 (monomer concentration) + 7 mM CaCl$_2$; syringe = 4 mM ATP + 7 mM CaCl$_2$. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).

**Figure S11.** ITC isotherms of titration of ATP into SbtB7001 in the presence of 500 μM CaCl$_2$. (L-R) 1. Cell = 160 μM SbtB7001 (monomer concentration) + 500 μM CaCl$_2$; syringe = 4 mM ATP + 500 μM CaCl$_2$. 2. Cell = 160 μM SbtB7001 (monomer concentration) + 500 μM CaCl$_2$; syringe = 4 mM ATP + 500 μM CaCl$_2$. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

**Figure S12.** ITC isotherm of titration of CaCl$_2$ into SbtB7001. Cell = 143 μM SbtB7001 (monomer concentration); syringe = 50 mM CaCl$_2$. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).

**Figure S13.** ITC isotherms of titration of CaCl$_2$ into SbtB7001 in the presence of ATP. (L-R) 1. Cell = 160 μM SbtB7001 (monomer concentration) + 400 μM ATP; syringe = 4 mM CaCl$_2$ + 400 μM ATP. 2. Cell = 160 μM SbtB7001 (monomer concentration) + 400 μM ATP; syringe = 7 mM CaCl$_2$ + 400 μM ATP. 3. Cell = 160 μM SbtB7001 (monomer concentration) + 400 μM ATP; syringe = 10 mM CaCl$_2$ + 400 μM ATP. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT (16). Figures were produced using GUSSI (27). The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC (15).
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

**Figure S14. ITC isotherms of titration of NaHCO₃ into SbtB7001.** (L-R) 1. Cell = 102 µM SbtB7001 (monomer concentration); syringe = 20 mM NaHCO₃ 2. Cell = 143 µM SbtB7001 (monomer concentration); syringe = 200 mM NaHCO₃. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT. Figures were produced using GUSSI. The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC.

**Figure S15. ITC isotherms of titration of NaHCO₃ into *E. coli* lacI.** (L-R) 1. Cell = 100 µM *E. coli* lacI; syringe = 200 mM NaHCO₃ 2. Cell = 100 µM *E. coli* lacI; syringe = 200 mM NaHCO₃. The upper panels represent baseline-corrected power traces. The middle panels represent the integrated heat data and best fit (globally fitted data from replicate experiments where possible) of the independent binding site model in SEDPHAT. Figures were produced using GUSSI. The bottom panels show the residuals of the fit. Error bars represent the standard error in the integration of the peaks as calculated by NITPIC.
Supporting Information for *Structural basis of SbtA-regulation by SbtB7001*

Figure S16. Structural heterogeneity of individual chains of SbtB7001 in apo- and ligand-bound states. Structural alignment of individually refined chains (white – purple) highlighting the subtle conformational differences in the T-loop (cyan/teal), C-loop (yellow), N-terminal and other loops (e.g. residues 28-32) between chains within the same asymmetric unit (when more than one). (A) apo-SbtB7001, (B) AMP:SbtB7001, (C) ADP:SbtB7001, (D) cAMP:SbtB7001, (E) ATP:SbtB7001, (F) Ca²⁺ATP:SbtB7001.
Figure S17. Trans-cis isomerization of Pro51 in ATP:SbtB7001. $2F_o-F_c$ (1.5 $\sigma$) and $F_o-F_c$ maps (+3 $\sigma$ [green] and -3 $\sigma$ [red]) following refinement of ATP:SbtB7001 with: (A) Pro51 (chain C) omitted; (B) Pro51 modelled in cis conformation and; (C) Pro51 modelled in trans conformation.
Chapter 3. The evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein.
3 The evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein

Key Ideas: Enzyme evolution, protein dynamics.

Techniques: Evolution-based approaches, ancestral protein reconstruction, colorimetric enzyme assays, auxotroph complementation assays, X-ray crystallography, molecular dynamics simulations.

3.1 Preface

Molecular recognition and protein-ligand binding are not only important for proteins specialised for binding functions (e.g. Chapter 2), but also underlie all enzyme activity. Indeed, substrate recognition and binding precede the chemical step during enzyme catalysis. Like proteins specialised for non-catalytic binding functions, enzymes must also be able to differentiate between different substrates and bind them with affinities that are physiologically relevant (Bar-Even et al., 2011). Further, enzymes must sample conformations that are important for catalysis, and cycle between states that facilitate for substrate binding, release and catalysis (Campbell et al., 2016). In this chapter, I present work that explores the evolution of the enzyme cyclohexadienyl dehydratase (CDT) from an ancestral solute-binding protein.

In the first paper, we describe how we used ancestral protein reconstruction to infer and characterise the sequences of ancestral proteins along the evolutionary trajectory leading to the present-day CDT from Pseudomonas aerogenosa (PaCDT). While other evolutionary biochemistry studies have looked at the evolution of new enzyme activity from proteins that had other catalytic activities, this was one of two original studies in which the natural evolution of an enzyme from a non-catalytic progenitor was described (the other being (Kaltenbach et al., 2018)). Through structural, biophysical analysis, and complementation assays, we show that PaCDT evolved from a cationic amino acid binding protein, and demonstrated that the emergence and improvement of CDT activity occurred via a number of distinct steps (Figure 3-1): (i) the co-adaption of ancestral binding site residues; (ii) the functionalisation of the
active site for the new function; (iii) conformational changes in the active site that improved complementary between the protein and the substrates; and (iv) optimisation of activity through substitutions remote from the active site. We propose that some of these remote mutations altered the relative sampling of the open and closed states of the protein to stabilise the more closed, catalytically competent state.

In the second manuscript, we expand on our previous work, using double electron-electron resonance (DEER) distance measurements in combination with MD simulations to confirm that PaCDT is stable in a more closed conformation compared with the ancestral proteins and the related amino acid binding proteins. We provide a more detailed structural explanation for the optimisation of CDT activity along this evolutionary pathway.

Both pieces of work explore and provide new insight into the process of molecular evolution and the role of protein conformational sampling in catalysis. Similar to the work presented in Chapter 2, this work highlights how using a combination of evolution-based, computational and experimental techniques can help us to better understand the physical and chemical determinants of protein function.

Figure 3-1. Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. Graphical abstract from the paper “Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein”.
3.2 Paper 1: Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein

3.2.1 Publication status

This manuscript presented in this section has been published by *Nature Chemical Biology*. Supporting information is provided in Section 3.3.

3.2.2 Author's contribution

The work presented in this publication was a collaboration between me and the co-authors of this paper. Some of the preliminary work presented in this paper (such as the initial complementation growth assay experiments) was performed by me during my Honours year under the supervision of Colin Jackson at the Research School of Chemistry, Australian National University. During my PhD studies, I performed X-ray crystallography to obtain and validate the structure of apo-AncCDT-1. I also completed ITC experiments, designed and performed colorimetric enzyme assays to measure the prephenate dehydratase activity of PaCDT and several of the ancestral proteins, and performed differential scanning fluorimetry experiments. In addition, I performed analytical size-exclusion chromatography and circular dichroism experiments. I also performed several of the molecular dynamics simulations and aided in the analysis of crystal structures and these simulations. Further, I performed analysis of the genomic context of PaCDT and its homologs. I helped write and edit the manuscript and produced a number of the figures for this publication.

3.2.3 Contributions from others

Ben Clifton and Colin Jackson developed the idea for this project. Ben Clifton performed most of the ancestral sequence reconstruction, protein crystallography, directed evolution experiments, structural analysis and molecular dynamics simulations. Ben Clifton wrote the majority of the draft of this manuscript. Colin Jackson helped with the writing and editing of the manuscript and the design of the experiments.
Paul D. Carr helped with the freezing of protein crystals and the collection of X-ray diffraction data. Monica L. Gerth (University of Otago) performed differential scanning fluorimetry on Pu1068 and AncCDT-2 using the Biolog Plates. Nobuhiko Tokuriki (University of British Columbia) helped design and perform the directed evolution experiments. Patrick Yates and Abera Saeed were undergraduate project students that helped me with enzyme assays and differential scanning fluorimetry experiments.
Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein

Ben E. Clifton1, Joe A. Kaczmarski1, Paul D. Carr1, Monica L. Gerth2,4, Nobuhiko Tokuriki3 and Colin J. Jackson4*  

The emergence of enzymes through the neofunctionalization of noncatalytic proteins is ultimately responsible for the extraordinary range of biological catalysts observed in nature. Although the evolution of some enzymes from binding proteins can be inferred by homology, we have a limited understanding of the nature of the biochemical and biophysical adaptations along these evolutionary trajectories and the sequence in which they occurred. Here we reconstructed and characterized evolutionary intermediate states linking an ancestral solute-binding protein to the extant enzyme cyclohexadienyl dehydratase. We show how the intrinsic reactivity of a desolvated general acid was harnessed by a series of mutations radiating from the active site, which optimized enzyme–substrate complementarity and transition-state stabilization and minimized sampling of noncatalytic conformations. Our work reveals the molecular evolutionary processes that underlie the emergence of enzymes de novo, which are notably mirrored by recent examples of computational enzyme design and directed evolution.

In this work, we used ancestral protein reconstruction10 to investigate the biophysical and biochemical mechanisms underlying the evolutionary transition between SBPs and CDTs. By analyzing the evolutionary trajectory between reconstructed ancestors and extant proteins, we show that the emergence and optimization of catalytic activity involves several distinct processes. The emergence of CDT activity was potentiated by the incorporation of a desolvated general acid into the ancestral binding site, which provided an intrinsically reactive catalytic motif, and reshaping of the ancestral binding site, which facilitated enzyme–substrate complementarity. Catalytic activity was subsequently gained via the introduction of hydrogen bonding networks that positioned the catalytic residue precisely and contributed to transition-state stabilization. Finally, catalytic activity was enhanced by remote substitutions that refined the active site structure and reduced sampling of noncatalytic states.

Results

Evolutionary history of CDT. Ws0279, the SBP with the highest sequence identity to CDT that has been functionally or structurally characterized, has been annotated as a lysine-binding protein on the basis of sequence similarity. CDT activity was subsequently gained via the introduction of hydrogen bonding networks that positioned the catalytic residue precisely and contributed to transition-state stabilization. Finally, catalytic activity was enhanced by remote substitutions that refined the active site structure and reduced sampling of noncatalytic states.

To reconstruct the evolutionary history of CDT, we inferred the maximum-likelihood phylogeny of 113 homologs of Ws0279 and PaCDT and used ancestral protein reconstruction to infer the most likely amino acid sequence for each ancestral node in the phylogeny. We evaluated the binding specificity of Ws0279 using differential scanning fluorimetry (DSF), thereby confirming that the protein is an AABP that is specific for the cationic amino acids L-lysine and, to a lesser extent, L-arginine (Supplementary Fig. 1a).

To reconstruct the evolutionary history of CDT, we inferred the maximum-likelihood phylogeny of 113 homologs of Ws0279 and PaCDT and used ancestral protein reconstruction to infer the most likely amino acid sequence for each ancestral node in the phylogeny. We evaluated the binding specificity of Ws0279 using differential scanning fluorimetry (DSF), thereby confirming that the protein is an AABP that is specific for the cationic amino acids L-lysine and, to a lesser extent, L-arginine (Supplementary Fig. 1a).

To reconstruct the evolutionary history of CDT, we inferred the maximum-likelihood phylogeny of 113 homologs of Ws0279 and PaCDT and used ancestral protein reconstruction to infer the most likely amino acid sequence for each ancestral node in the phylogeny. We evaluated the binding specificity of Ws0279 using differential scanning fluorimetry (DSF), thereby confirming that the protein is an AABP that is specific for the cationic amino acids L-lysine and, to a lesser extent, L-arginine (Supplementary Fig. 1a).

To reconstruct the evolutionary history of CDT, we inferred the maximum-likelihood phylogeny of 113 homologs of Ws0279 and PaCDT and used ancestral protein reconstruction to infer the most likely amino acid sequence for each ancestral node in the phylogeny. We evaluated the binding specificity of Ws0279 using differential scanning fluorimetry (DSF), thereby confirming that the protein is an AABP that is specific for the cationic amino acids L-lysine and, to a lesser extent, L-arginine (Supplementary Fig. 1a).

To reconstruct the evolutionary history of CDT, we inferred the maximum-likelihood phylogeny of 113 homologs of Ws0279 and PaCDT and used ancestral protein reconstruction to infer the most likely amino acid sequence for each ancestral node in the phylogeny. We evaluated the binding specificity of Ws0279 using differential scanning fluorimetry (DSF), thereby confirming that the protein is an AABP that is specific for the cationic amino acids L-lysine and, to a lesser extent, L-arginine (Supplementary Fig. 1a).
We experimentally characterized the five ancestral proteins, using isothermal titration calorimetry (ITC) to test for amino acid binding and genetic complementation to test for enzymatic activity; in the genetic complementation assay, expression of CDT recapitulated the higher growth rate associated with AncCDT-3W (Supplementary Fig. 4b). Spectrophotometric kinetic assays in vitro confirmed that AncCDT-3 and AncCDT-3(P188L), but not AncCDT-2, have prephenate dehydratase activity (Supplementary Fig. 4d–h). Altogether, characterization of both sets of ancestral proteins supported the conclusions that ancestral amino acid-binding activity was lost between AncCDT-1 and AncCDT-2, that CDT activity was gained between AncCDT-2 and AncCDT-3, and that AncCDT-2 apparently had neither CDT activity nor binding affinity toward amino acids.

An intermediate function between AABPs and CDTs. To test whether AncCDT-2 was rendered nonfunctional by an error in its reconstructed sequence or if it had a function distinct from AncCDT-1 and AncCDT-3, we examined representatives of the previously uncharacterized evolutionary clades consisting of extant descendants of AncCDT-2 and AncCDT-3: Pu1068 from “Candidatus Pelagibacter ubique” and Ea1174 from Euxiobacterium antarcticum (Fig. 1d). Genetic complementation experiments showed that Ea1174, but not Pu1068, has CDT activity (Supplementary Fig. 4c), and DSF experiments showed that Pu1068 is not an AABP (Supplementary Fig. 1c). Analysis of the genomic context of Pu1068 and several of its orthologs revealed that these genes, like the SBP gene Ws0279, are adjacent to genes encoding transmembrane components of ABC importers, suggesting that Pu1068 encodes an SBP rather than an enzyme (Supplementary Table 3).

We used two strategies that have previously been used for functional annotation of SBPs\(^ {12} \) to identify the physiological ligands of Pu1068 and AncCDT-2: crystallization of Pu1068 with co-purified ligands bound during heterologous expression of the protein...
in *E. coli* and DSF screening of Pu1068 and AncCDT-2 against metabolite libraries. Although crystalization and structure determination of Pu1068 were successful, the protein crystallized in the unliganded state, suggesting the absence of high-affinity ligands of the protein in the *E. coli* metabolome. DSF experiments to identify the ligands of Pu1068 and AncCDT-2 were performed using several hundred potential metabolites from libraries and rationally selected metabolites with plausible physiological importance for oceanic bacteria such as *C. Pelagibacter* ubique (Supplementary Fig. 5 and Supplementary Dataset 1). Although the exact physiological ligands of AncCDT-2 and Pu1068 could not be identified, we found that these proteins have some affinity for a variety of carboxylates (Supplementary Fig. 5) and for some sultones, such as the sulfohexitol NDSB-221, which binds Pu1068 with a *K*_D of 0.53 mM (Supplementary Fig. 6). Regardless of the specific physiological ligands of AncCDT-2 and Pu1068, the functional properties of the various extant clades (W10279: cationic AABP; Pu1068: SBP of unknown function; Ea174 and PaCDT: CDTs) agreed with those expected based on functional characterization of the ancestral proteins, supporting a likely evolutionary trajectory from a cationic AABP to a carboxylic-acid-binding protein, and finally to CDT, an enzyme with carboxylic acid substrates (Fig. 1d).

**Emergence of catalytic activity.** To establish the molecular basis for the functional transition from binding protein to enzyme, we first attempted to rationalize the catalytic activity of the extant enzyme PaCDT in structural terms. PaCDT has previously been crystallized (PDB ID 3KBR) in complex with the nonphysiological ligand HEPES, which shares some fortuitous similarities with the cyclohexadienol substrates of the enzyme. We therefore attempted to solve the crystal structure of the native, unliganded enzyme, and obtained a structure in which the active site cavity was occupied by only an acetate molecule from the crystallization buffer and four ordered water molecules (Fig. 2a). Unlike that in the structure of the PaCDT–HEPES complex, the active site of the PaCDT–acetate complex was fully occluded from solvent and highly complementary to its cyclohexadienol substrates (Fig. 2a and Supplementary Fig. 7). Docking of prephenate and *L*-arogenate into the PaCDT–acetate structure implied a binding mode in which Glu173 is positioned adjacent to the departing hydroxyl group of the substrate, suggesting that the enzyme mechanism involves general acid catalysis by Glu173 (Fig. 2b and Supplementary Fig. 8a). Consistent with its proposed role as a general acid, Glu173 is partially deoslated and predicted by PROCHECK to be protonated at neutral pH (p*K*_A = 7.75), and the substitution E173Q abolishes prephenate dehydratase activity with minimal impact on secondary structure and thermostability (Supplementary Fig. 8b–d). The active site of PaCDT is pre-organized for protonation and elimination of the departing hydroxyl group of the substrate by an intricate hydrogen bonding network extending from Glu173 (Fig. 2c). Other active site residues most likely contribute to stabilization of the departing carboxylate group and delocalized electrons in the developing π system in the transition state (Fig. 2d).

We next solved the crystal structures of AncCDT-1 and AncCDT-3 (P188L). Comparison of these structures with the structures of the extant proteins PaCDT and Pu1068 revealed the contribution of historical amino acid substitutions to remodeling, functionalization, and refinement of the ancestral amino acid binding site (Fig. 3a–e). First, mutations that occurred between AncCDT-1 and AncCDT-2 caused two important structural changes that potentiated the emergence of catalytic activity: the substitution V173E introduced a general acid that is positioned appropriately for general acid catalysis (Fig. 3b), whereas substitutions D197T and A20G allowed a conformational change for Trp60, reshaping the ancestral binding site and facilitating steric complementarity between CDT and its substrates (Fig. 3c). These substitutions can be considered potentiating because the structural features associated with them are also observed in Pu1068, and they initially enabled binding of a different ligand rather than CDT activity (Fig. 3d). Indeed, each residue associated with these structural changes was reconstructed with high statistical confidence in the noncatalytic protein AncCDT-2 (Supplementary Fig. 2c). Thus, the evolution of CDT from AABPs required the acquisition of a new binding function, resulting in the introduction of amino acids that potentiated the structure for subsequent evolution of catalytic function.

Structural analysis indicated that functionalization of the ancestral binding site for catalytic activity occurred by subsequent mutations that fixed either between AncCDT-1 and AncCDT-2 or between AncCDT-2 and AncCDT-3. The substitutions Q100K, Q128N, and S133N introduced the hydrogen bonding network that positions the catalytic group precisely and contributes to transition-state stabilization through interactions with the departing hydroxyl and carboxylate groups of the substrate (Fig. 3b). Additionally, the substitutions Q100K and L198K likely contributed to dual specificity for α-amino and α-keto acid substrates (i.e., *L*-arogenate and prephenate) via electrostatic shielding of Asp170 (Fig. 3e). However, AncCDT-2 contains each of these four active site substitutions (except L198K, which is not itself sufficient to introduce catalytic activity (Fig. 3a), implying that additional substitutions between AncCDT-2 and AncCDT-3 were required for the emergence of CDT activity. To identify these substitutions, we performed site-directed mutagenesis and three rounds of directed evolution, which resulted in the isolation of an AncCDT-2 variant with only six substitutions (CDT-M5) that allowed slow growth of *E. coli* *L*-phenylalanine auxotrophs and exhibited prephenate dehydratase activity in vitro (Fig. 3f and Supplementary Figs. 4h and 9). Although three of these substitutions (T131G, A155I, and L198K) are present in AncCDT-3, the other three substitutions (F25L, G99S, and P102L) represent an
While the T131G substitution removes a steric clash between the alternative evolutionary trajectory toward higher catalytic activity. The introduction of additional mutations in various combinations supported faster growth of L-phenylalanine auxotrophs (Fig. 3f and Supplementary Fig. 9d). These results show that there are multiple mutational pathways to higher CDT activity via remote substitutions following the introduction of key active site residues.

**Evolution of an efficient enzyme.** Although AncCDT-3(P188L) has CDT activity, its second order rate constant ($k_{cat}/K_m$) is ~6,000-fold lower than that of PaCDT despite their active sites being virtually identical (Supplementary Figs. 4h and 10). We therefore investigated the role of structural dynamics in the evolutionary process. Upon ligand binding, SBPs undergo domain-scale open–closed conformational changes that are essential for function, and these are exemplified by the unliganded and arginine-bound crystal structures of AncCDT-1 (Fig. 4a). The open–closed conformational equilibrium of an SBP controls binding affinity and the rate of solute transport, suggesting that the position of this equilibrium must be adjusted for optimization of solute transport in a cellular context. On the other hand, efficient enzyme catalysis depends on pre-organization of the active site; unproductive conformational sampling has been shown to constrain the catalytic efficiency of recently evolved enzymes. The closed conformation of CDT is the catalytically competent conformation; the open–closed conformational change would be necessary only to the extent needed to enable substrate binding and product release from the occluded active site. These considerations suggest a possible point of adaptive conflict between solute binding and catalytic activity that may have necessitated changes in structural dynamics during the evolution of CDT.

The unliganded SBPs AncCDT-1 and Pu1068 and the inefficient ancestral enzyme AncCDT-3(P188L), whose structures were solved in this work, crystallized in an open conformation (Fig. 4a). This is consistent with previous studies showing that unliganded AABPs sample closed or semi-closed conformations only transiently, as well as with previously reported crystal structures of unliganded AABPs, of which only 1/14 crystallized in a closed conformation (Supplementary Table 4). By contrast, PaCDT crystallized in a closed conformation in the absence of substrate or substrate analogs in multiple, differently packed crystals, suggesting that the closed conformation of the enzyme is unusually stable for this protein fold (Fig. 4a and Supplementary Fig. 7a). It should be noted that the adventitiously bound acetate molecule observed in the active site of the PaCDT–acetate structure makes several polar interactions with the large domain of the enzyme (via Arg85 and Ser80), but no polar interactions with the small domain, with Thr132 being the only residue within potential hydrogen bonding distance (Supplementary Fig. 11). The acetate molecule is therefore unlikely to make a substantial contribution to stabilization of the closed conformation of the PaCDT–acetate complex.

To further investigate the structural dynamics of PaCDT, we simulated molecular dynamics (MD) simulations of the PaCDT trimer from the PaCDT–acetate structure with the acetate molecule removed. These simulations indicated that the open conformation is accessible in PaCDT, although most of the subunits remained closed throughout each 170 ns trajectory (Fig. 4b,c and Supplementary Fig. 12g). Additional simulations using a different initial structure (PaCDT–HEPES structure, with the HEPES molecule removed) or a different force field gave similar results (Supplementary Fig. 12a,b). In contrast, MD simulations of AncCDT-1 initialized from unliganded closed and unliganded open structures showed that the ancestral protein had dynamical properties typical of SBPs; in each simulation, the unliganded closed structure transitioned to an open conformation within 150 ns (Supplementary Fig. 12d,f), whereas the unliganded open structure remained in an open conformation for the duration of the simulations (Supplementary Fig. 12c,e,h).
The domain-scale conformational fluctuations that did occur in the PaCDT MD simulations were characteristic of SBPs; principal component analysis showed that hinge-bending and hinge-twisting motions typical of AABPs\(^{17,20}\) accounted for >85% of conformational variance (Fig. 4b). Indeed, these domain-scale motions were similar to those observed in the simulations initialized from the unliganded, closed AncCDT-1 structure. Furthermore, the open structure of AncCDT-3(P188L), which provided experimental evidence for sampling of the open conformation in CDTs, resembled the simulated open conformation of PaCDT (Supplementary Fig. 12g). Thus, the characteristic domain-scale dynamics of the SBP fold are retained in CDTs and are indeed necessary for substrate or product diffusion from the occluded active site. However, the unusual stability of the closed conformation of PaCDT suggests that the conformational landscape of the enzyme has evolved between AncCDT-3(P188L) and PaCDT to minimize unproductive sampling of the noncatalytic open conformation, contributing to improvements in catalytic efficiency toward the end of the evolutionary trajectory.

**Discussion**

In this work, we have outlined the steps required for a noncatalytic protein to not only evolve some initial catalytic activity, but also to become a proficient, bona fide enzyme (for example, PaCDT, \(k_{cat}/K_m \approx 10^5 \text{M}^{-1} \text{s}^{-1}\)). One of the most notable aspects of the evolutionary trajectory of CDT is the extent to which the emergence of catalysis depended on the repurposing of structural features of the ancestral SBPs AncCDT-1 and AncCDT-2. As an obvious example, the ancestral AABP AncCDT-1 provided an amino acid binding motif that was exploited for substrate binding in CDT. Additionally, some residues that are important for CDT activity, but not the ancestral amino acid binding activity, are observed in AncCDT-1 (Asp21, Asn152, and Thr169). The presence of these residues in AncCDT-1 shows that they were compatible with the ancestral function and could have evolved neutrally. Moreover, our phylogenetic and functional analyses showed that the evolution of CDT, which has enzymatic activity on the carboxylic acid prephenate, from an AABP required the prior acquisition of a new binding function, represented by AncCDT-2 and the extant protein Pu1068, which have affinity for carboxylic acids. The evolution of this new binding function resulted in fixation of two structural changes that were required for the evolution of CDT: the conformational change of Trp60 and the incorporation of Glu173, which later became the catalytic acid in CDT, into the binding site. Altogether, these results suggest that the major change in function between AABPs and CDT was contingent on various preexisting structural features, both essential and nonessential with respect to the ancestral functions.

Directed evolution of ancestral CDT variants highlighted the existence of multiple mutational pathways to higher catalytic activity following the introduction of key catalytic residues, raising the possibility that the evolution of CDT activity could have occurred gradually and nondeterministically. Previous studies of enzyme evolution have provided examples of evolutionary trajectories that are deterministic and constrained, in which the occurrence of a particular mutation determines the nature and occurrence of subsequent mutations\(^{19,20}\). In contrast, our directed evolution experiments demonstrated that there were alternative mutational pathways to higher catalytic activity in ancestral CDT variants that were distinct from the historical mutational pathway between AncCDT-2 and AncCDT-3, indicating either that the evolutionary trajectory of CDT was nondeterministic or that the historical mutations were fixed instead of the alternative mutations because of pleiotropy and/or adaptive conflict (i.e., because the alternative mutations had negative effects on fitness independent of their effect on enzymatic activity). One potential reason for the existence of numerous mutations that increase the enzymatic activity of ancestral CDT variants is the relatively simple chemistry associated with the enzyme mechanism; CDT has a simple one-step (or possibly two-step) mechanism that depends on a single catalytic residue, and the substrate of CDT is a high-energy metabolic intermediate that is predisposed to the reaction catalyzed by the enzyme. Therefore, higher catalytic activity might be accessed by a larger number of mutational pathways in CDT than in enzymes that catalyze more complex reactions with higher activation energy barriers, such as hydrolysis of unactivated substrates, for which more precise positioning of active site residues is likely necessary.

Our structural analysis of evolutionary intermediates and extant proteins demonstrates in molecular detail how the evolution of highly specialized and efficient CDTs from noncatalytic ancestors occurred in several distinct stages. Incorporation of the desolvated general acid Glu173 into the binding pocket of an ancestral SBP most likely provided sufficient chemical reactivity for initial, promiscuous CDT activity. Indeed, the intrinsic reactivity of desolvated acidic and basic residues has been exploited similarly in enzymes that have evolved recently in response to anthropogenic substrates\(^21\) and in enzymes engineered via single substitutions in noncatalytic proteins\(^22\). Following the introduction of a reactive general acid, optimization of enzyme–substrate complementarity and introduction of hydrogen
bonding networks to position the catalytic residue precisely and stabilize the departing carboxylate group of the substrate appear to have occurred. Further improvements in catalytic efficiency could have been gained by second- and third-shell substitutions that refine the structure of the active site and optimize conformational sampling to favor catalytically relevant conformations. Similar mutational patterns have been documented in directed evolution experiments\(^{(1,2)}\). Additionally, adaptation of protein dynamics has been shown to occur analogously in the evolution of a binding protein from an enzyme, in which the catalytically relevant conformation was disfavored by the function-switching mutation\(^{(3)}\).

Although some computationally designed protein structures have been made with atomic-level accuracy\(^{(4)}\), and various strategies have been developed to introduce catalytic activity into arbitrary protein scaffolds\(^{(5,6,7)}\), replicating the catalytic proficiency of natural enzymes using computational design remains a major challenge\(^{(8,9)}\). The evolutionary trajectory of CDT has striking similarities with the optimization of rationally designed enzymes by directed evolution; catalytic activity can be initialized by computationally guided grafting of a reactive catalytic motif (for example, a desolvated carboxylate) into a protein scaffold that can accommodate the transition state for a given reaction, and directed evolution can be used to introduce additional stabilizing interactions, optimize positioning of catalytic groups, improve enzyme-transition state complementarity, and optimize conformational sampling, frequently via remote substitutions\(^{(10,11)}\). Thus, the strategies that have been used to improve catalytic activity in computational design and directed evolution experiments appear to mirror those that drove the emergence of an enzyme from a non-catalytic protein by natural evolution.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41589-018-0043-2.

Received: 31 July 2017; Accepted: 1 March 2018; Published online: 23 April 2018

References


Acknowledgements

B.E.C. and J.A.J. were supported by Australian Postgraduate Awards. B.E.C. was also supported by a Rod Rickards PhD scholarship and an Alan Sargenson scholarship. This research was undertaken with the assistance of resources, services, and staff from the Australian National Computational Infrastructure (NCI), the Australian Synchrotron, and the CSIRO Collaborative Crystallisation Centre, and funding from the Australian Research Council Discovery Project scheme (C.J.J.). We thank A. Saeed, P. Yates, L. Tan and S. Waring for additional technical contributions. We thank H. Janošik (IST Austria) for gifting us the pDOTS7 plasmid.

Author contributions

B.E.C. and C.J.J. conceived the study; B.E.C. and J.A.K. performed computational analysis; J.A.K., B.E.C., and M.L.G. performed experimental characterization of proteins; B.E.C., J.A.K., P.D.C., and C.J.J. solved the crystal structures; N. T. and C.J.J. supervised students; B.E.C., J.A.K., and C.J.J. wrote the paper. All authors contributed to experimental design, editing of the paper, and interpretation of results.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-018-0043-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.J.J.

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

Materials. pDOTS7 is a derivative of pQE-82L (Qiagen) modified to enable Golden Gate cloning and was created by removal of the SapI site from pQE-82L and introduction of two reciprocal SapI sites following the His6 tag, with the SapI and large insertions and combined with a subset of a previous alignment of refolding during the affinity chromatography step to remove endogenously K partition coefficient (\(K_p\)).

Phylogenetic analysis and ancestral sequence reconstruction. The protein sequences of 113 homologs of W0279 and PaCDT were collected from the NCBI reference sequence database using the BLAST server. The sequences were aligned in MUSCLE. The alignment was edited to remove N-terminal signal peptides and large insertions and combined with a subset of a previous alignment of the empirical Bayes method implemented in PAML. Evaluation of BIONJ trees reconstructed using different amino acid substitution models, using the Akaike information criterion as implemented in ProtTest, supported the use of the WAG substitution matrix with gamma-distributed rate heterogeneity, a fixed proportion of invariant sites, and equilibrium amino acid frequencies estimated from the data (WAG-I + I + F + M model). Phylogenies were reconstructed in PhyML by optimization of an initial BIONJ tree using the nearest-neighbor interchange as well as subtree pruning and regrafting algorithms. Robustness of the resulting tree topology to the substitution model was assessed by repeating the analysis using the LG and JTT substitution matrices (LG/TT/I + F + M models), and convergence to the ML tree was checked by repeating the analyses with ten randomized initial trees. Although the resulting trees had essentially identical topologies, the tree inferred using the LG/TT/I + F + M model had the highest likelihood and was therefore taken as the ML tree. Ancestral protein sequences were reconstructed using the empirical Bayes method implemented in PAML. The ancestral sequence AncCDT-I to AncCDT-5 were inferred using the LG substitution matrix together with the ML tree inferred using the LG/TT/I + F + M model, and the ancestral sequences AncCDT-1W to AncCDT-5W were reconstructed using the WAG substitution matrix together with the tree inferred using the WAG-I + F + M model (Supplementary Fig. 2).

Cloning and mutagenesis. Codon-optimized synthetic genes encoding the ancestral proteins, W0279 (UniProt: Q7MAG0; residues 24–258), Pu1068 (UniProt: Q4HFRS; residues 19–255), Ea1714 (UniProt: K0ABP5; residues 31–268), and PaCDT (UniProt: Q01209; residues 26–268) were cloned into the pDOTS7 vector using the Golden Gate method. Site-directed mutagenesis was achieved using Gibson assembly gene fragments with ~30 bp overlap were synthesized by PCR using complementary primers encoding the desired mutation and assembled together with the linearized pDOTS7 vector using Gibson assembly. Successful cloning and mutagenesis was confirmed by Sanger sequencing of the vector insert.

Protein expression and purification. Proteins were generally expressed in E. coli (BL21)DE3 cells, except for enzyme assays, in which case they were expressed in ApheA cells to exclude the possibility of contamination with endogenous prephenate dehydratase. Cells were typically grown in Luria-Bertani (LB) or Terrific Broth (TB) media at 37 °C to OD\(_{600}\) \(0.8\) with 0.5 mM β-1,1,1-2-isopropylglycidoxy propane and incubated for a further 20 h at 37 °C. Cells were pelleted and stored at −80 °C before protein purification.

For most applications, proteins were purified under native conditions by nickel-nitrilotriacetic acid (Ni-NTA) affinity chromatography and size-exclusion chromatography (SEC). Cells were thawed, resuspended in equilibration buffer (50 mM Na\(_2\)HPO\(_4\), 500 mM NaCl, 20 mM imidazole, pH 7.4), lysed by sonication, and fractionated by ultracentrifugation (24,200 x g for 1 h at 4 °C). The supernatant was filtered through a 0.45 μm filter and loaded onto a 5 mL HisTrap HP column (GE Healthcare) equilibrated with equilibration buffer. The column was washed with 50 μL equilibration buffer and 25 μL wash buffer (50 mM Na\(_2\)HPO\(_4\), 500 mM NaCl, 44 mM imidazole, pH 7.4), and the target protein was eluted in 25 μL elution buffer (50 mM Na\(_2\)HPO\(_4\), 500 mM NaCl, 500 mM imidazole, pH 7.4). For ITC experiments, proteins were subjected to on-column refolding during the affinity chromatography step to remove endogenously bound ligands, as described previously. Proteins were concentrated using a centrifuge filter (Amicon Ultra 15 filter unit with 10 kDa cut-off) and purified by SEC on a HiLoad 26/600 Superdex 200 column (GE Healthcare), typically eluting in SEC buffer (20 mM Na\(_2\)HPO\(_4\), 150 mM NaCl, pH 7.6), 5x SYPRO orange dye (Sigma-Aldrich) and 1 (1 mM or 10 mM for amino acids, ≥10 mM for other metabolites) in a total volume of 20 μL and dialyzed on a 384-well PCR plate in at least triplicate. At least eight replicates of ligand-free control were also included on each plate.

Fluorescence intensities were monitored continuously as the samples were heated from 20 °C to 99 °C at a rate of 0.05 °C/s, with excitation at 580 nm and emission monitored at 595 nm. Time-averaged fluorescence (\(T_{avg}\)) was determined by fitting the data to a Boltzmann function, \(F = A + B (C + D)/(1 + e^{(T_{avg} - T)/E})\), where \(F\) is fluorescence and \(T\) is temperature. The parameters \(A\) and \(C\), accounting for the slopes of the pre- and post-transition baselines, were fixed at zero if possible.

Pu1068, AncCDT-I, and AncCDT-2 were also screened against a subset of Biolog Phenotype Microarray (PM) plates (Biolog, Hayward, CA, USA), as described previously. Libraries of biologically relevant potential ligands were generated by dissolving each compound in 50 μL water, resulting in concentrations of approximately 10–20 μM in the assay (the exact concentrations vary from well to well, and are not released by the manufacturer). Plates PM1–PM5 contain six concentrations of each compound, whereas plate PM9 contains a series of concentrations of each compound. Fluorescence intensities were measured on a Lightcycler 480 real-time PCR instrument (Roche Diagnostics). Initial hits were further tested using known concentrations (0–600 μM) of each potential ligand to confirm binding. An additional in-house screen consisted of a subset of the Solubility and Stability Screen (Hampton Research), which was tested by the Crick Centre for Crystallisation (http://www.mrc-lmb.cam.ac.uk), Cambridge, UK. For this screen, the reaction mixtures contained 0.3 μg Pu1068, 3.75x SYPRO orange and 5.3 ligand in a total volume of 20 μL, in a 96-well plate format; each ligand was tested at three concentrations and three replicates of a ligand-free control were also included. Fluorescence intensities were measured on a Bio-Rad CFX384 real-time PCR instrument with excitation at 490 nm and emission at 570 nm. The temperature was ramped from 20 °C to 100 °C at a rate of 0.05 °C/s, and the fluorescence intensity was measured at 0.5°C intervals. Melting temperatures were taken as the temperature at the minimum of the first derivative of the melt curve, which was determined by fitting the data to a quadratic function in the vicinity of the melting temperature using GraphPad Prism 7 software.

Isothermal titration calorimetry. ITC experiments were performed using a Nano-ITC low-volume calorimeter (TA Instruments); details of instrument calibration have been described previously. ITC experiments were performed at 25°C with stirring at 200 r.p.m. Protein and ligand solutions were prepared in matched SEC buffer and degassed before use. Amino acid solutions were prepared volumetrically from commercial samples (Sigma-Aldrich, Alfa Aesar) with stated purity ≥98%. Ancestral proteins were tested for binding of proteinogenic amino acids via screening experiments in which 45 μL of 0.844 mM ligand was injected continuously into 164 μL of 50 mM protein over 300 s. In some cases, ligands were tested in mixtures of structurally related amino acids. Titrations with 100 μM protein was generally titrated with 1 x 1 μL and then 28 x 1.6 μL injections of 0.69 mM ligand at 300 s intervals. The background heat was estimated as the average heat associated with each injection in a control titration of ligand into buffer and subtracted from each protein–ligand titration. Association constants (\(K_a\)) were determined by fitting the integrated heat data to the independent binding sites model in NanoAnalyze software (TA Instruments).

Genetic complementation. E. coli strain JW2580-1 (ΔpheA) cells were transformed with the appropriate plasmid by electroporation, plated on LB agar supplemented with 100 μg/mL ampicillin (LBA agar), and incubated at 37 °C overnight. Single colonies were used to inoculate 20 mL M9 minimal media supplemented with 1-tyrosine, ampicillin and IPTG (M9 – F: per L: 6 g Na\(_2\)HPO\(_4\), 3 g KH\(_2\)PO\(_4\), 0.5 g NaCl, 1 g NH\(_4\)Cl, 20 mL 20% (w/v) glucose, 2 mL 1 M MgCl\(_2\), 0.1 mL 1 M CaCl\(_2\), 2 mL 2.5 mg/mL l-tyrosine, 1 mL 100 mg/mL ampicillin, 0.2 mL 1 M IPTG). The cultures were incubated at 37 °C with shaking at 180 r.p.m., and OD\(_{600}\) was measured periodically. We confirmed that the observed differences in growth rates could not be explained by differences in protein expression levels by culturing the same media supplemented with 20 μg/mL l-phenylalanine (M9 + F media) and assessing protein expression by SDS–PAGE of the soluble fraction of the crude cell lysate from each culture.

Preparation of sodium prephene. Sodium prephene was prepared from barium chorismate (Sigma, 60–80% purity). Barium chorismate (40 mM in \(\text{H}_2\)O) was mixed with an equimolar amount of 1 M Na\(_2\)SO\(_4\). An equal volume of 100 mM Na\(_2\)HPO\(_4\) (pH 8.0) was added to the mixture and the BaSO\(_4\) precipitate was removed by centrifugation. Sodium prephene was obtained by heating the resulting sodium chorismate solution at 70 °C for 1 h. Aliquots were stored
at −80°C. The concentration of prephenate was measured by quantitative conversion of prephenate to phenylpyruvate under acidic conditions (0.5 M HCl, 15 min, 25°C) and spectrophotometric determination of phenylpyruvate concentration, as described previously17.

**Prephenate dehydratase assay.** Prephenate dehydratase activity was determined by spectrophotometric measurement of phenylpyruvate formation, as described previously17. Protein solutions were prepared in 20 mM Na2HPO4, 150 mM NaCl (pH 8.0), and prephenate dehydratase solutions were prepared in 50 mM Na2HPO4 (pH 8.0). After equilibration at room temperature (20–25°C) for 5 min, the reaction was initiated by mixing equal volumes of protein and substrate solutions. Aliquots (50 µL or 100 µL) were regularly removed from the reaction mixture and quenched by addition of an equal volume of 2 M NaOH. Absorbance at 320 nm was measured using an Epoch Microplate Spectrophotometer (BioTek), and phenylpyruvate concentrations were determined assuming a molar extinction coefficient of 17,500 M−1 cm−1. Reaction times and enzyme concentrations were adjusted to ensure <20% conversion of prephenate to phenylpyruvate. The rate of nonenzymic turnover was subtracted from the observed rate of enzyme-catalyzed turnover.

**Circular dichroism spectroscopy.** Circular dichroism (CD) experiments were performed using a Chirascan spectropolarimeter (Applied Photophysics) with a 1-mm path length quartz cuvette. Proteins were diluted to 0.3 mg/mL in water (for recording CD spectra) or SEC buffer (for thermal denaturation experiments) and degassed before measurements. CD spectra were recorded at 20°C between 190 nm and 260 nm, with a bandwidth of 0.5 nm and a scan rate of 3° per point, with adaptive sampling. For thermal denaturation experiments, CD was monitored at 222 nm over a temperature range of 20°C to 90°C, heating at 1°C min−1. Tm values were determined by fitting the data to a two-state model:

\[ y_{\text{obs}} = (y + m_{T}) \times \exp \left( \frac{\Delta H_{\text{m}}}{R} \times \left( \frac{1}{T} - \frac{1}{T_{\text{m}}} \right) \right) + \left(1 + \exp \left( \frac{\Delta H_{\text{m}}}{R} \times (1 - 1/T_{\text{m}}) \right) \right) \]

where \( y_{\text{obs}} \) is ellipticity at 222 nm, \( y \), \( m_{T} \), \( y_{\text{ref}} \), and \( m_{\text{ref}} \) describe the pre-transition and post-transition baselines, \( T \) is temperature, \( R \) is the gas constant, and \( \Delta H_{\text{m}} \) is the apparent van’t Hoff enthalpy of unfolding.

**Crystallization and structure determination.** Crystal structures of AncCDT-1 (complexed with l-arгинine), Pu1068 (unliganded), AncCDT-3 (P188L), and PaCDT (complexed with acetate) were solved and refined at resolutions between 1.6 Å and 2.6 Å. An additional low-resolution structure of the PaCDT–acetate complex (3.2 Å) shows an alternate crystal packing arrangement; except where explicitly noted, the high-resolution PaCDT–acetate structure was used for structural analysis. A low-resolution (3.4 Å) structure of unliganded AncCDT-1, which illustrates the domain-scale conformational change resulting from ligand binding, was also solved. Finally, Pu1068 was also co-crystallized with NDSB-221 which illustrates the domain-scale conformational change resulting from ligand binding. When explicitly noted, the high-resolution PaCDT–acetate structure was used for computational docking. The resulting complexes were energy minimized using the OPLS3 force field, with heavy atoms restrained within 0.3 Å of their initial position. Water and missing side chains and a missing residue (Gln190) in the PaCDT–HEPES strand of the PaCDT complex were rebuilt. Glu173 was protonated, and other residues were assigned the appropriate protonation states.

**Library creation and selection.** Purified PCR products (0.5 µg) from STEP or ISOR reactions were digested with 2.5 µL each of HindIII FD and EcoRI FD (Thermo Scientific) in a 50 µL reaction at 37°C for 30 min. The reaction mixture was purified immediately using a PCR purification kit. The PaCDT–90 ng gene fragments, 2.2 µL 10X buffer, 0.2 mM dNTPs, 1.25 U Taq polymerase, and 10 ng template DNA were amplified using various concentrations of equimolar mutagenic oligonucleotides (500 nM total concentration) in a volume of 20 µL (see Supplementary Table 13 for a list of oligonucleotides included in each round). The thermocycling program consisted of (i) an initial denaturation step at 95°C for 2 min; (ii) 40 cycles of a denaturation step at 95°C for 30 s, then 13 hybridization steps from 65°C to 41°C in 2°C steps, each for 90 s (total 13.5 min), then an extension step at 72°C for 1 min; and (iii) a final extension step at 72°C for 7 min. 0.5 µL of the unpurified assembly reaction mixture was amplified in a 50 µL nested PCR reaction using Taq polymerase and the primers P7SF and P7NR (Supplementary Table 13). The nested PCR product was run on a 1% agarose gel and purified by gel extraction.

**Articles**

Staggered extension process. AncCDT-3 and AncCDT-3W were recombined using the staggered extension process (STEP) following a literature protocol18. The STEP reaction mixture contained 5 µL 10X Taq buffer, 1.5 mM MgCl2, 0.2 mM of each dNTP, 75 fmol of each template plasmid, 30 pmol of each primer, and 2.5 U Taq polymerase (New England Biolabs) in a total volume of 50 µL. The primers used in the reaction were the 5’ flanking primer P7XF and the 3’ flanking primer P7XR (Supplementary Table 13), which amplify ~100 bp on either side of the SapI site of the pDOTS7 vector. The thermocycling program consisted of 80 cycles of (i) a denaturation step at 95°C for 30 s and (ii) an annealing/extension step for 5 s at 52°C. 2 µL of the resulting PCR product was incubated with 10 U DpnI (Thermo Scientific) in a reaction volume of 10 µL at 37°C for 1 h to digest the parental plasmid DNA. 5 µL of the DpnI-digested STEP product was then amplified in a nested PCR reaction using Taq polymerase, in a total volume of 100 µL. The primers used for the nested PCR reaction, P7NF and P7NR (Supplementary Table 13), which amplify ~100 bp on either side of the SapI site of the pDOTS7 strand, respectively. The nested PCR product was run on a 1% agarose gel and purified by gel extraction.

**Molecular dynamics simulations.** MD simulations of PaCDT were initialized from the PaCDT–HEPES and PaCDT–acetate structures. The structure of the PaCDT trimer was generated from the monomer structure by application of the crystallographic three-fold rotation operation. Small molecules, including the active site HEPES and acetate molecules, were removed from the structures, and missing side chains and a missing residue (Gln190) in the PaCDT–HEPES structure were modeled in MODELLER19. N-terminal acetyl caps and C-terminal
amide caps were added using MODELLER and Coot\(^9\). MD simulations of AncCDT-1 were initialized from the \(-\)arginine-bound and unliganded structures. Small molecules were removed from the structures, and missing side chains and residues were modeled in Desmond (Schrodinger) and Coot\(^9\). Terminal amide caps and C-terminal amide caps were added using Desmond (Schrodinger). GROMOS MD simulations were performed using GROMACS version 4.5.5 (ref. \(^9\)) for the PaCDT–HEPES structure and GROMACS version 4.6.5 for the PaCDT–acetate structure and AncCDT-1 structures, using the GROMOS 54a6 force field\(^9\) in all cases. Each protein was solvated in a rhombic dodecahedron with SPC water molecules, such that the minimal distance of the protein to the periodic boundary was 15 Å, and an appropriate number of ions were added to neutralize the system (15 Na\(^+\) ions for PaCDT simulations, 1 Cl\(^-\) ion for AncCDT-1 simulations). Energy minimization was achieved using the steepest descent algorithm. A 100 ps isothermal (NVT) MD simulation with position restraints on the protein was used to equilibrate the system at 300 K. For production MD simulations of the NPT ensemble, the temperature was maintained at 300 K using Berendsen's thermostat (\(\tau = 0.1\) ps), and the pressure was maintained at 1 bar using Berendsen's barostat (\(\tau = 0.5\) ps, compressibility = \(4.5 \times 10^{-5}\) bar\(^{-1}\)). All protein bonds were constrained with the LINCS algorithm; water molecules were constrained using the SETTLE algorithm; the time step for numerical integration was 2 fs; the cut-offs for short-range electrostatics and van der Waals forces were 9 Å and 14 Å, respectively; the Particle-Mesh Ewald method was used to evaluate long-range electrostatics; neighbor lists were updated every 10 steps. Following a 1 ns equilibration phase, which was not considered in the analysis, the four simulations of the PaCDT–HEPES structure were continued for 100 ns, and the four simulations of the PaCDT–acetate structure were continued for 170 ns. The three AncCDT-1 simulations were continued for 150 ns. Additional 150 ns simulations were performed in Desmond version 4.8 (Schrodinger 2016–4\(^9\)) using the OPLS3 force field\(^9\). These simulations were initiated from the PaCDT–acetate trimer, unliganded AncCDT-1, and \(-\)arginine-bound AncCDT-1 structures, with all small molecules removed. Desmond was used to add the N-terminal acetyl caps and C-terminal amide caps in each case, and for energy minimization of the protein structure. The protein was solvated in an orthorhombic box (15 Å periodic boundary) with SPC water molecules, and the system was neutralized as described above. Energy minimization was achieved using a hybrid method of the steepest descent algorithm and the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (maximum of 2,000 iterations and a convergence threshold of 1 kcal/mol/Å). The system was relaxed using the default relaxation procedure in Desmond. For production MD simulations of the NPT ensemble, the temperature was maintained at 300 K using a Nosé–Hoover thermostat (\(\tau = 1.0\)), and the pressure was maintained at 1.0 bar (\(\tau = 2.0\)). A Martyna–Tobias–Klein barostat was used. Otherwise, default Desmond options were used. Following relaxation of the system, each simulation was run for 150 ns.

**Structure analysis.** Residues in extant CDT homologs (Wu0279, Pu1068, Ea1174, PaCDT) are numbered according to the equivalent position in the ancestral proteins. Bio3D\(^9\) was used for r.m.s. deviation, radius of gyration, and interdomain angle calculations, and principal component analysis. These analyses were performed on the PaCDT–HEPES and PaCDT–acetate–GROMOS simulations using protein backbone atoms (N, C\(_\alpha\), and C\(_\beta\)) of individual protein subunits at 0.1 ns intervals. The PaCDT–acetate–OPLS simulations were analyzed separately and projected onto the principal component derived from the PaCDT–HEPES and PaCDT–acetate–GROMOS simulations. The AncCDT-1 simulations were also analyzed separately and projected onto the principal components derived from simulations initialized from closed AncCDT-1. The interdomain angle was calculated as the angle between the centers of mass of three groups of backbone atoms: the large domain (residues 2–97 and 196–234), the hinge region (residues 98–99 and 196–198) and the small domain (residues 98–195). Hinge axes for rigid-body domain displacements were determined using DynDom\(^9\) (Supplementary Fig. 7d). PROPKA version 3.0\(^9\) was used for pK\(_a\) prediction.

**Intrinsic tryptophan fluorescence spectroscopy.** Intrinsic tryptophan fluorescence spectroscopy was recorded using a Pu1068 prepared at a concentration of 5 \(\mu\)M in DSP buffer. The excitation wavelength was 280 nm, and emission was measured between 300 nm and 400 nm. Following addition of each aliquot of NDSB-221, the sample was incubated at ambient temperature for 1 min before the fluorescence spectrum was recorded. The \(K_0\) for the Pu1068/NDSB-221 interaction was calculated by fitting the fluorescence data to a hyperbolic binding curve \((F - F_m)/(F_m - F) = [L]/(K + [L])\), where \(F\) is fluorescence, \(F_m\) and \(F_0\) are initial and final fluorescence, and \(K\) is ligand concentration.

**Statistics.** For DSF experiments, one-way analysis of variance (ANOVA) with Dunnett's test for multiple comparisons was used to assess the statistical significance of differences in T\(_{m}\) between untreated and ligand-treated samples. \(\Delta T_{m}\) values were taken to be indicative of binding if \(\Delta T_{m} > 2 \mp 2°C\) and \(P < 0.05\).


3.3 Supplementary Information for “Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein”
Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein

Ben E. Clifton¹, Joe A. Kaczmarski¹, Paul D. Carr¹, Monica L. Gerth²,⁴, Nobuhiko Tokuriki³ and Colin J. Jackson¹

¹Research School of Chemistry, Australian National University, Canberra, ACT, Australia. ²Department of Biochemistry, University of Otago, Dunedin, New Zealand. ³Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada. ⁴Present address: School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand. *e-mail: colin.jackson@anu.edu.au
### Supplementary Table 1. Enzymes with the SBP fold.

<table>
<thead>
<tr>
<th>Name</th>
<th>PDB</th>
<th>EC</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP phosphoribosyltransferase</td>
<td>1H3D</td>
<td>2.4.2.17</td>
<td>Histidine biosynthesis</td>
</tr>
<tr>
<td>Thiaminase I</td>
<td>3THI</td>
<td>2.5.1.2</td>
<td>Thiamine catabolism</td>
</tr>
<tr>
<td>Porphobilinogen deaminase</td>
<td>1PDA</td>
<td>2.5.1.61</td>
<td>Porphyrin biosynthesis</td>
</tr>
<tr>
<td>1,4-dihydroxy-6-naphthoate synthase</td>
<td>3A3U</td>
<td>4.1.-.-</td>
<td>Menaquinone biosynthesis</td>
</tr>
<tr>
<td>Prephenate dehydratase</td>
<td>3MWB</td>
<td>4.2.1.51</td>
<td>Phenylalanine biosynthesis</td>
</tr>
<tr>
<td>Cyclohexadienyl dehydratase</td>
<td>3KBR</td>
<td>4.2.1.51,4.2.1.91</td>
<td>Phenylalanine biosynthesis</td>
</tr>
</tbody>
</table>

These enzymes were identified by searching the PDB for proteins belonging to CATH superfamily 3.40.190.10 (periplasmic binding protein-like II)\(^1\). The PDB code of a representative structure, the Enzyme Commission (EC) number, and the gene ontology (GO) annotation are given for each enzyme.
### Supplementary Table 2. Ancestral protein reconstruction statistics.

<table>
<thead>
<tr>
<th>Ancestral protein</th>
<th>Mean posterior probability&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Uncertain sites&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Sequence differences between alternative versions&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Sequence differences between alternative versions at uncertain sites&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>AncCDT-1</td>
<td>0.80</td>
<td>88 (37%)</td>
<td>22 (9%)</td>
<td>22</td>
</tr>
<tr>
<td>AncCDT-2</td>
<td>0.79</td>
<td>95 (40%)</td>
<td>21 (9%)</td>
<td>21</td>
</tr>
<tr>
<td>AncCDT-3</td>
<td>0.85</td>
<td>67 (29%)</td>
<td>10 (4%)</td>
<td>10</td>
</tr>
<tr>
<td>AncCDT-4</td>
<td>0.85</td>
<td>71 (30%)</td>
<td>13 (6%)</td>
<td>13</td>
</tr>
<tr>
<td>AncCDT-5</td>
<td>0.89</td>
<td>54 (23%)</td>
<td>11 (5%)</td>
<td>9</td>
</tr>
</tbody>
</table>

<sup>a</sup>Mean posterior probability of the maximum-likelihood version of the ancestral sequence inferred using the LG model.

<sup>b</sup>Number (percentage) of sites at which the maximum-likelihood reconstruction has a posterior probability < 0.8.

<sup>c</sup>Number (percentage) of sites at which the alternative versions of the ancestral protein differ in sequence.

<sup>d</sup>Number of sites where a sequence difference between the alternative versions of the ancestral protein occurs at a site where the maximum-likelihood reconstruction has a posterior probability < 0.8. This metric indicates that the alternative ancestral proteins characterized in this work had sequence differences at 15–25% of sites where the maximum-likelihood reconstruction was ambiguous.
## Supplementary Table 3. Genomic context of CDT homologs.

<table>
<thead>
<tr>
<th>Clade</th>
<th>CDT homolog STRING ID</th>
<th>Organism</th>
<th>Top scoring interaction partner STRING identifier</th>
<th>Top scoring interaction partner annotation</th>
<th>STRING interaction score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>WS0279</td>
<td>Wolinella succinogenes DSM 1740</td>
<td>WS0280</td>
<td>ABC transporter permease</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>SAR11_1069</td>
<td>Pelagibacter ubique HTCC1062</td>
<td>SAR11_1069</td>
<td>ABC transporter permease, glnP</td>
<td>0.906</td>
</tr>
<tr>
<td></td>
<td>OA307_5441</td>
<td>Octadecabacter antarcticus 307</td>
<td>OA307_4070</td>
<td>Putative inner membrane amino-acid ABC transporter permease protein</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>VP3065</td>
<td>Vibrio parahaemolyticus RIMD 2210633</td>
<td>CP3066</td>
<td>Amino acid ABC transporter permease</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td>JS86_22305</td>
<td>Vibrio vulnificus CMCP6</td>
<td>JS86_22310</td>
<td>Amino acid ABC transporter permease</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>PSE_1115</td>
<td>Pseudovibrio sp. JE062</td>
<td>PSE_1116</td>
<td>Putative glutamine transport system permease protein GlnP</td>
<td>0.888</td>
</tr>
<tr>
<td>II</td>
<td>SAR11_1068</td>
<td>Pelagibacter ubique HTCC1062</td>
<td>SAR11_1069</td>
<td>ABC transporter permease, glnP</td>
<td>0.906</td>
</tr>
<tr>
<td></td>
<td>OA307_5441</td>
<td>Octadecabacter antarcticus 307</td>
<td>OA307_4070</td>
<td>Putative inner membrane amino-acid ABC transporter permease protein</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>VP3065</td>
<td>Vibrio parahaemolyticus RIMD 2210633</td>
<td>CP3066</td>
<td>Amino acid ABC transporter permease</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td>JS86_22305</td>
<td>Vibrio vulnificus CMCP6</td>
<td>JS86_22310</td>
<td>Amino acid ABC transporter permease</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>PSE_1115</td>
<td>Pseudovibrio sp. JE062</td>
<td>PSE_1116</td>
<td>Putative glutamine transport system permease protein GlnP</td>
<td>0.888</td>
</tr>
<tr>
<td>Accession</td>
<td>Organism</td>
<td>Description</td>
<td>Similarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>-------------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RKLH11_3656</td>
<td>Rhodobacteraceae bacterium KLH11</td>
<td>ABC-type glutamine transport system, permease component</td>
<td>0.875</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSK20926_18177</td>
<td>Roseobacter sp. SK209-2-6</td>
<td>Putative amino acid ABC transporter, permease protein</td>
<td>0.883</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DND132_2219</td>
<td>Desulfovibrio desulfuricans ND132</td>
<td>Extracellular solute-binding protein family 3</td>
<td>0.746</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exig_1225</td>
<td>Exiguobacterium sibiricum</td>
<td>Polar amino acid ABC transporter inner membrane subunit</td>
<td>0.696</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GYMC10_0791</td>
<td>Paenibacillus sp. Y412MC10</td>
<td>Polar amino acid ABC transporter inner membrane subunit</td>
<td>0.591</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBR47_56430</td>
<td>Brevibacillus brevis</td>
<td>Hypothetical protein</td>
<td>0.731</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAE_01295</td>
<td>Enterobacter aerogenes</td>
<td>Putative transferase</td>
<td>0.840</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ilyop_1568</td>
<td>Ilyobacter polytropus</td>
<td>PAAT family amino acid ABC transporter membrane protein</td>
<td>0.780</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT01EI_3690</td>
<td>Edwardsiella ictaluri 93-146</td>
<td>Hypothetical protein</td>
<td>0.815</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KPN_02982</td>
<td>Klebsiella pneumoniae</td>
<td>Amino acid ABC transporter membrane protein</td>
<td>0.669</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YPO3343</td>
<td>Yersinia pestis CO92</td>
<td>Amino-acid ABC transporter (permease)</td>
<td>0.718</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA3475</td>
<td>Pseudomonas aeruginosa PAO1</td>
<td>ABC transporter</td>
<td>0.601</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EbC_38320</td>
<td>Erwinia billingiae Eb661</td>
<td>Modulator of drug activity</td>
<td>0.675</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avi_0614</td>
<td>Agrobacterium vitis S4</td>
<td>Amino acid ABC transporter substrate-binding protein</td>
<td>0.677</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenBank Accession</td>
<td>Organism</td>
<td>Protein Description</td>
<td>STRING Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------</td>
<td>----------------------------------------------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dd586_3436</td>
<td><em>Dickeya dadantii</em> Ech586</td>
<td>GCN5-like N-acetyltransferase</td>
<td>0.584</td>
<td></td>
<td></td>
</tr>
<tr>
<td>blr4561</td>
<td><em>Bradyrhizobium japonicum</em> USDA110</td>
<td>Hypothetical protein (DUF1993)</td>
<td>0.665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV_0224</td>
<td><em>Chromobacterium violaceum</em></td>
<td>Hypothetical protein (Chorismate-lyase like)</td>
<td>0.859</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strvi_2438</td>
<td><em>Streptomyces violaceusniger</em></td>
<td>ABC transporter substrate-binding protein</td>
<td>0.628</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsero_4358</td>
<td><em>Herbaspirillum seropedicae</em></td>
<td>Chorismate mutase</td>
<td>0.529</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Clades I-V refer to the clades that branch from AncCDT-1 to AncCDT-5, respectively.

b The genomic context and protein interactions of the CDT homologs were analyzed using the STRING database (version 10.5). STRING interaction scores were calculated based on neighborhood, gene-fusion and co-occurrence scores.

c Although the top scoring interaction partners of some clade III genes are also associated with ABC transporters, the corresponding STRING interaction scores are lower than for clade I and II genes.
Supplementary Table 4. PDB codes of previously crystallized unliganded AABPs.

<table>
<thead>
<tr>
<th>Open conformation</th>
<th>Closed conformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4P0I</td>
<td>3KZG</td>
</tr>
<tr>
<td>4PRS</td>
<td></td>
</tr>
<tr>
<td>4OEN</td>
<td></td>
</tr>
<tr>
<td>4KPT</td>
<td></td>
</tr>
<tr>
<td>4KR5</td>
<td></td>
</tr>
<tr>
<td>4LA9</td>
<td></td>
</tr>
<tr>
<td>3ZSF</td>
<td></td>
</tr>
<tr>
<td>3HV1</td>
<td></td>
</tr>
<tr>
<td>3DEL</td>
<td></td>
</tr>
<tr>
<td>2O1M</td>
<td></td>
</tr>
<tr>
<td>2IEE</td>
<td></td>
</tr>
<tr>
<td>1GGG</td>
<td></td>
</tr>
<tr>
<td>2LAO</td>
<td></td>
</tr>
</tbody>
</table>

These structures were identified by searching the PDB for proteins belonging to PFAM family PF00497 (bacterial extracellular solute-binding proteins, family 3). Only ABC transporter-associated SBPs were considered; SBP-like domains such as the regulatory domains of lytic murein transglycosylases were excluded.
Supplementary Table 5. Experimental details for crystallization and structure determination of PaCDT.

<table>
<thead>
<tr>
<th>Structure</th>
<th>PaCDT-acetate (high-resolution, trigonal)</th>
<th>PaCDT-acetate (low-resolution, tetragonal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5HPQ</td>
<td>6BQE</td>
</tr>
<tr>
<td>Protein buffer</td>
<td>20 mM TRIS pH 8.0, 100 mM NaCl, 0.5 mM DTT, 10% glycerol</td>
<td>20 mM TRIS pH 8.0, 100 mM NaCl, 0.5 mM DTT, 10% glycerol</td>
</tr>
<tr>
<td>Crystallization</td>
<td>Hanging drop, 2 µL PaCDT (10 mg/mL) + 2 µL 18% (w/v) PEG 3350, 0.2 M ammonium acetate, 0.1 M TRIS pH 8.0</td>
<td>Hanging drop, 2 µL PaCDT (10 mg/mL) + 2 µL 18% (w/v) PEG 3350, 0.2 M ammonium acetate, 0.1 M TRIS pH 8.2</td>
</tr>
<tr>
<td>Cryoprotectant</td>
<td>30% (v/v) PEG 400, 70% (v/v) mother liquor</td>
<td>30% (w/v) PEG 3350, 0.2 M ammonium acetate, 0.1 M TRIS pH 8.0</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9537</td>
<td>0.9537</td>
</tr>
<tr>
<td>Beamline</td>
<td>Australian Synchrotron MX1</td>
<td>Australian Synchrotron MX1</td>
</tr>
<tr>
<td>Data processing</td>
<td>iMOSFLM*, Aimless*</td>
<td>XDS*, Aimless*</td>
</tr>
<tr>
<td>MR model</td>
<td>3KBR, using each domain as a separate search model.</td>
<td>5HPQ</td>
</tr>
<tr>
<td>Monomers in asymmetric unit</td>
<td>1</td>
<td>3 ($V_M = 2.43 \text{Å}^3$/Da, 49% solvent content)</td>
</tr>
<tr>
<td>Refinement</td>
<td>The model was rebuilt manually from a polyalanine model in Coot and refined by real-space refinement in Coot and reciprocal-space refinement in REFMAC5.</td>
<td>Minor adjustments to the model were made in Coot, and reciprocal-space refinement was done in REFMAC5 with non-crystallographic symmetry (NCS) restraints.</td>
</tr>
<tr>
<td>Model quality</td>
<td>Ramachandran statistics: 94.9% favored, 0.9% outliers. MolProbity score: 1.45 (98th percentile).</td>
<td>Ramachandran statistics: 94.1% favored, 0.4% outliers. MolProbity score: 2.02 (100th percentile).</td>
</tr>
</tbody>
</table>
Supplementary Table 6. Experimental details for crystallization and structure determination of Pu1068.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Pu1068 (unliganded)</th>
<th>Pu1068 (NDSB-221 complex)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PDB code</strong></td>
<td>5WJP</td>
<td>5KKW</td>
</tr>
<tr>
<td><strong>Protein buffer</strong></td>
<td>10 mM TRIS pH 8.0, 100 mM NaCl, 10% glycerol</td>
<td>10 mM TRIS pH 8.0, 50 mM NaCl, 10 mM NDSB-221</td>
</tr>
<tr>
<td><strong>Crystallization</strong></td>
<td>Hanging drop, 1.5 µL Pu1068 (18 mg/mL) + 1.5 µL 24% (w/v) PEG 8000, 5% (v/v) PEG 400, 0.1 M BIS-TRIS propane pH 9.0.</td>
<td>Hanging drop, 1 µL Pu1068 (24 mg/mL) + 1 µL 27% (w/v) PEG 3350, 0.1 M lithium sulfate, 0.1 M MES pH 6.5.</td>
</tr>
<tr>
<td><strong>Cryoprotectant</strong></td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td><strong>Beamline</strong></td>
<td>Australian Synchrotron MX2</td>
<td>Australian Synchrotron MX1</td>
</tr>
<tr>
<td><strong>Wavelength (Å)</strong></td>
<td>1.0332</td>
<td>0.9537</td>
</tr>
<tr>
<td><strong>Data processing</strong></td>
<td>iMOSFLM³, Aimless⁴</td>
<td>XDS⁵, Aimless⁴</td>
</tr>
<tr>
<td><strong>MR model</strong></td>
<td>3KBR, using each domain as a separate search model.</td>
<td>5WJP, using each domain as a separate search model.</td>
</tr>
<tr>
<td><strong>Monomers in asymmetric unit</strong></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td>Model building and real-space refinement were done in Coot⁶. Reciprocal space refinement was done in REFMAC5⁷.</td>
<td>Model building and real-space refinement were done in Coot⁶. Reciprocal space refinement was done in REFMAC5⁷. Geometric restraints for NDSB-221 were generated using eLBOW in the PHENIX⁸ package. Translation-libration-screw (TLS) parameters for one TLS group included in final refinement round in REFMAC5.</td>
</tr>
<tr>
<td><strong>Model quality</strong></td>
<td>Ramachandran statistics: 97.9% favored, 0% outliers. MolProbity score: 1.21 (98th percentile).</td>
<td>Ramachandran statistics: 97.6% favored, 0% outliers. MolProbity score: 1.10 (100th percentile).</td>
</tr>
</tbody>
</table>
**Supplementary Table 7. Experimental details for crystallization and structure determination of AncCDT-1.**

<table>
<thead>
<tr>
<th>Structure</th>
<th>AncCDT-1 (L-arginine complex)</th>
<th>AncCDT-1 (unliganded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5T0W</td>
<td>5TUJ</td>
</tr>
<tr>
<td>Protein buffer</td>
<td>20 mM HEPES pH 7.5, 50 mM NaCl, 1 mM L-arginine</td>
<td>20 mM HEPES pH 7.5, 150 mM NaCl</td>
</tr>
<tr>
<td>Crystallization</td>
<td>Initial crystals were obtained from a hanging drop containing 2 µL AncCDT-1 (18 mg/mL) and 2 µL 22% (w/v) PEG 3350, 0.2 M lithium sulfate, 0.1 M TRIS pH 8.2. These crystals were crushed and serially diluted in the precipitant, and a hanging drop was prepared by mixing 2 µL of the resulting microseed suspensions with 2 µL protein. Three iterations of microseeding using the resulting crystals yielded the final crystal used for structure determination.</td>
<td>Hanging drop, 1 µL AncCDT-1 (26 mg/mL) + 1 µL 22.5% (w/v) PEG 3350, 0.2 M lithium sulfate, 0.1 M TRIS pH 8.0.</td>
</tr>
<tr>
<td>Cryoprotectant</td>
<td>30% (w/v) PEG 3350, 0.2 M lithium sulfate, 0.1 M TRIS pH 8.0.</td>
<td>None</td>
</tr>
<tr>
<td>Beamline</td>
<td>Australian Synchrotron MX1</td>
<td>Australian Synchrotron MX1</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9501</td>
<td>0.9537</td>
</tr>
<tr>
<td>Data processing</td>
<td>iMOSFLM&lt;sup&gt;°&lt;/sup&gt;, Aimless&lt;sup&gt;°&lt;/sup&gt;</td>
<td>XDS&lt;sup&gt;°&lt;/sup&gt;, Aimless&lt;sup&gt;°&lt;/sup&gt;</td>
</tr>
<tr>
<td>MR model</td>
<td>4ZV2, processed using CHAINSAW&lt;sup&gt;°&lt;/sup&gt; to remove non-conserved side chains.</td>
<td>5T0W, using each domain as a separate search model.</td>
</tr>
<tr>
<td>Monomers in asymmetric unit</td>
<td>4 (Vₐ = 2.34 Å³/Da, 47% solvent content)</td>
<td>1</td>
</tr>
<tr>
<td>Refinement</td>
<td>Model building and real-space refinement were done in Coot&lt;sup&gt;°&lt;/sup&gt;. Reciprocal-space refinement with torsion-based NCS restraints was done by simulated annealing and restrained refinement in PHENIX&lt;sup&gt;°&lt;/sup&gt; and REFMAC5&lt;sup&gt;°&lt;/sup&gt;. TLS parameters for four groups of atoms per subunit, chosen using TLSMD&lt;sup&gt;°&lt;/sup&gt;, were included in the final refinement round in REFMAC5.</td>
<td>Model building and real-space refinement were done in Coot&lt;sup&gt;°&lt;/sup&gt;. Reciprocal space refinement was done in PHENIX&lt;sup&gt;°&lt;/sup&gt;, with secondary structure restraints and group B-factor refinement (2 groups per amino acid). TLS parameters for five TLS groups selected automatically with phenix.refine were included in the final refinement round.</td>
</tr>
<tr>
<td>Model quality</td>
<td>Ramachandran statistics: 95.0% favored, 0.5% outliers. MolProbity score: 2.34 (89&lt;sup&gt;th&lt;/sup&gt; percentile).</td>
<td>Ramachandran statistics: 93.3% favored, 0.4% outliers. MolProbity score: 2.26 (100&lt;sup&gt;th&lt;/sup&gt; percentile).</td>
</tr>
</tbody>
</table>
Supplementary Table 8. Experimental details for crystallization and structure determination of AncCDT-3(P188L).

<table>
<thead>
<tr>
<th>Structure</th>
<th>AncCDT-3(P188L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5JOS</td>
</tr>
<tr>
<td>Protein buffer</td>
<td>20 mM HEPES pH 7.5, 50 mM NaCl</td>
</tr>
<tr>
<td>Crystallization</td>
<td>Sitting drop, 1 µL AncCDT-3(P188L) (18 mg/mL) + 1 µL 1.2 M sodium citrate, 0.1 M TRIS pH 8.0.</td>
</tr>
<tr>
<td>Cryoprotectant</td>
<td>1.8 M sodium citrate, 0.05 M TRIS pH 8.0.</td>
</tr>
<tr>
<td>Beamline</td>
<td>Australian Synchrotron MX1</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9537</td>
</tr>
<tr>
<td>Data processing</td>
<td>IMOSFLM&lt;sup&gt;3&lt;/sup&gt;, Aimless&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td>MR model</td>
<td>3KBR, using each domain as a separate search model.</td>
</tr>
<tr>
<td>Monomers in asymmetric unit</td>
<td>1</td>
</tr>
<tr>
<td>Refinement</td>
<td>Model building and real-space refinement were done in Coot&lt;sup&gt;6&lt;/sup&gt;. Reciprocal space refinement was done in REFMAC5&lt;sup&gt;7&lt;/sup&gt;. TLS parameters for three groups of atoms, chosen using TLSMD&lt;sup&gt;8&lt;/sup&gt;, were included in the final refinement round in REFMAC5.</td>
</tr>
<tr>
<td>Model quality</td>
<td>Ramachandran statistics: 95.4% favored, 1.3% disallowed. MolProbity score: 1.74 (93&lt;sup&gt;rd&lt;/sup&gt; percentile).</td>
</tr>
</tbody>
</table>
Supplementary Table 9. Data collection and refinement statistics for PaCDT.

<table>
<thead>
<tr>
<th>Structure</th>
<th>PaCDT (H3)</th>
<th>PaCDT (P4₃22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5HPQ</td>
<td>6BQE</td>
</tr>
</tbody>
</table>

**Data collection**

<table>
<thead>
<tr>
<th></th>
<th>PaCDT (H3)</th>
<th>PaCDT (P4₃22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>H3</td>
<td>P4₃22</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>124.89, 124.89, 40.63</td>
<td>95.71, 95.71, 187.53</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.0, 90.0, 120.0</td>
<td>90.0, 90.0, 90.0</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>38.03-2.05 (2.11-2.05)*</td>
<td>46.88-3.20 (3.42-3.20)</td>
</tr>
<tr>
<td>Rmerge (%)</td>
<td>11.0 (57.9)</td>
<td>36.7 (84.4)</td>
</tr>
<tr>
<td>Rsym (%)</td>
<td>5.5 (46.2)</td>
<td>17.3 (39.9)</td>
</tr>
<tr>
<td>CC1/2 (%)</td>
<td>99.5 (71.4)</td>
<td>94.4 (59.0)</td>
</tr>
<tr>
<td>I / σ (I)</td>
<td>8.3 (1.6)</td>
<td>5.0 (1.9)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>97.9 (82.3)</td>
<td>99.7 (100.0)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>4.8 (2.5)</td>
<td>5.2 (5.3)</td>
</tr>
</tbody>
</table>

**Refinement**

<table>
<thead>
<tr>
<th></th>
<th>PaCDT (H3)</th>
<th>PaCDT (P4₃22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>36.05-2.05</td>
<td>46.88-3.20</td>
</tr>
<tr>
<td>No. unique reflections</td>
<td>13739</td>
<td>14976</td>
</tr>
<tr>
<td>Rwork / Rfree (%)</td>
<td>17.59 / 24.10</td>
<td>23.66 / 28.81</td>
</tr>
<tr>
<td>No. atoms (chain A / B / C)</td>
<td>1870 / 1899 / 1895 / 1883</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>1870</td>
<td>1899 / 1895 / 1883</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>4</td>
<td>4 / 4 / 4</td>
</tr>
<tr>
<td>Water</td>
<td>55</td>
<td>12</td>
</tr>
<tr>
<td>B-factors (Å²)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>30.88</td>
<td>36.00</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>24.62</td>
<td>14.90</td>
</tr>
<tr>
<td>Water</td>
<td>26.86</td>
<td>12.70</td>
</tr>
<tr>
<td>R.m.s. deviations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.016</td>
<td>0.011</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.78</td>
<td>1.41</td>
</tr>
</tbody>
</table>

*Values in parentheses are for highest-resolution shell.*
Supplementary Table 10. Data collection and refinement statistics for Pu1068.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Pu1068 (unliganded)</th>
<th>Pu1068 (NDSB-221)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5WJP</td>
<td>5KKW</td>
</tr>
</tbody>
</table>

**Data collection**

<table>
<thead>
<tr>
<th>Space group</th>
<th>$P2_1\overline{2}2_1$</th>
<th>$P4_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a, b, c$ (Å)</td>
<td>38.73, 65.87, 90.97</td>
<td>77.69, 77.69, 44.80</td>
</tr>
<tr>
<td>$\alpha, \beta, \gamma$ (°)</td>
<td>90.0, 90.0, 90.0</td>
<td>90.0, 90.0, 90.0</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>37.43-1.57 (1.59-1.57)*</td>
<td>38.84-1.88 (1.92-1.88)</td>
</tr>
<tr>
<td>$R_{merge}$ (%)</td>
<td>12.9 (167.8)</td>
<td>7.8 (58.1)</td>
</tr>
<tr>
<td>$CC_1/2$ (%)</td>
<td>99.8 (56.6)</td>
<td>99.8 (67.7)</td>
</tr>
<tr>
<td>$I / \sigma (I)$</td>
<td>12.7 (1.9)</td>
<td>17.2 (3.0)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>97.3 (96.6)</td>
<td>95.7 (67.7)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>11.5 (10.8)</td>
<td>7.4 (6.2)</td>
</tr>
</tbody>
</table>

**Refinement**

<table>
<thead>
<tr>
<th>Resolution (Å)</th>
<th>37.43-1.57</th>
<th>38.86-1.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. unique reflections</td>
<td>30761</td>
<td>19884</td>
</tr>
<tr>
<td>$R_{work}$ / $R_{free}$ (%)</td>
<td>18.25 / 21.30</td>
<td>20.50 / 23.76</td>
</tr>
<tr>
<td>No. atoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>1885</td>
<td>1858</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>-</td>
<td>19</td>
</tr>
<tr>
<td>Water</td>
<td>207</td>
<td>95</td>
</tr>
<tr>
<td>$B$-factors (Å$^2$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>14.96</td>
<td>28.84</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>-</td>
<td>51.69</td>
</tr>
<tr>
<td>Water</td>
<td>23.50</td>
<td>29.92</td>
</tr>
<tr>
<td>R.m.s. deviations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.0249</td>
<td>0.0196</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>2.20</td>
<td>1.92</td>
</tr>
</tbody>
</table>

*Values in parentheses are for highest-resolution shell.
### Supplementary Table 11. Data collection and refinement statistics for AncCDT-1.

<table>
<thead>
<tr>
<th>Structure</th>
<th>AncCDT-1 (L-arginine)</th>
<th>AncCDT-1 (unliganded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5T0W</td>
<td>5TUJ</td>
</tr>
</tbody>
</table>

#### Data collection

<table>
<thead>
<tr>
<th>Space group</th>
<th>P2(_1)2(_1)2(_1)</th>
<th>P3(_2)1</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, c (Å)</td>
<td>47.03, 68.88, 318.6</td>
<td>71.64, 71.64, 124.0</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.0, 90.0, 90.0</td>
<td>90.0, 90.0, 120.0</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>33.66-2.59 (2.71-2.59)*</td>
<td>27.74-3.35 (3.47-3.35)</td>
</tr>
<tr>
<td>R(_{merge}) (%)</td>
<td>15.8 (77.5)</td>
<td>21.5 (100)</td>
</tr>
<tr>
<td>CC(_{1/2}) (%)</td>
<td>99.3 (75.5)</td>
<td>99.4 (74.4)</td>
</tr>
<tr>
<td>I / σ (I)</td>
<td>8.2 (2.4)</td>
<td>8.0 (1.9)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.4 (99.3)</td>
<td>99.8 (99.8)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>6.8 (6.9)</td>
<td>9.5 (7.8)</td>
</tr>
</tbody>
</table>

#### Refinement

| Resolution (Å)     | 33.19-2.59              | 27.74-3.35            |
| No. unique reflections | 31387                  | 5601                  |
| R\(_{work}\) / R\(_{free}\) (%) | 25.35 / 28.73          | 30.49 / 33.75         |
| No. atoms (chain A / B / C / D) | 1769 / 1734 / 1718 / 1721 | 1738 |
| Protein            | 12 / 12 / 12 / 12       | -                     |
| Ligand/ion         | 21                      | -                     |
| Water              |                          |                        |
| B-factors (Å\(^2\)) | 45.04                   | 90.5                  |
| Protein            | 23.67                   | -                     |
| Ligand/ion         | 18.10                   | -                     |
| R.m.s. deviations  | Bond lengths (Å)        | 0.0108                |
|                    | Bond angles (°)         | 1.37                  |

*Values in parentheses are for highest-resolution shell.*
Supplementary Table 12. Data collection and refinement statistics for AncCDT-3(P188L).

<table>
<thead>
<tr>
<th>Structure</th>
<th>AncCDT-3(P188L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5JOS</td>
</tr>
</tbody>
</table>

**Data collection**
- Space group: $P4_212$
- Cell dimensions:
  - $a, b, c$ (Å): 90.37, 90.37, 101.81
  - $\alpha, \beta, \gamma$ (°): 90.0, 90.0, 90.0
- Resolution (Å): 45.19-2.10 (2.16-2.10)*
- $R_{merge}$ (%): 7.3 (76.0)
- $CC_{1/2}$ (%): 99.9 (61.9)
- $I / \sigma (I)$: 19.7 (2.0)
- Completeness (%): 99.8 (98.2)
- Redundancy: 11.8 (5.0)

**Refinement**
- Resolution (Å): 44.35-2.10
- No. unique reflections: 23894
- $R_{work} / R_{free}$ (%): 17.56 / 20.68
- No. atoms:
  - Protein: 1932
  - Ligand/ion: 23
  - Water: 70
- $B$-factors ($\AA^2$):
  - Protein: 48.82
  - Ligand/ion: 38.70
  - Water: 44.64
- R.m.s. deviations:
  - Bond lengths (Å): 0.0213
  - Bond angles (°): 2.05

*Values in parentheses are for highest-resolution shell.*
Supplementary Table 13. Sequences of oligonucleotides used for directed evolution.

<table>
<thead>
<tr>
<th>General primers</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>P7XF</td>
<td>cgctttcacctcgagaaatcc</td>
</tr>
<tr>
<td>P7XR</td>
<td>caaaccgcagctttctgaac</td>
</tr>
<tr>
<td>P7NF</td>
<td>caatttcacacagaattccataaag</td>
</tr>
<tr>
<td>P7NR</td>
<td>gctcagcttaattaaagtctttattag</td>
</tr>
<tr>
<td>P7SF</td>
<td>gcctttctcttttccac</td>
</tr>
<tr>
<td>P7SR</td>
<td>gcctttgatatttccacc</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ISOR-R1 Substitution</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>F27Y</td>
<td>aaaccgtttagctatataaagatccgaacgggtca</td>
</tr>
<tr>
<td>K28R</td>
<td>accgatctttctgatccgaacggtc</td>
</tr>
<tr>
<td>N31D</td>
<td>gatctctgcttcttcag</td>
</tr>
<tr>
<td>A44S</td>
<td>gcacgtgcaaagaagtctgctataaag</td>
</tr>
<tr>
<td>D48S</td>
<td>gatctgtgagctttcagttaaatccg</td>
</tr>
<tr>
<td>I64M</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>G70D</td>
<td>gatctcagcagaattttagatataaag</td>
</tr>
<tr>
<td>I81V</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>N108D</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>L118I</td>
<td>caaaccgcagctttcttcag</td>
</tr>
<tr>
<td>K120R</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>K124R</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>K138R</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>P142K</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>R147T</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>F149Y</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>A153V</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>S161A</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>V200Y</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>A202M</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>P207Q</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>N215Q</td>
<td>gctctgctttcttcag</td>
</tr>
<tr>
<td>F234L</td>
<td>gctctgctttcttcag</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ISOR-R2 Substitution</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>F25L</td>
<td>gcggtagttatatataaacccgcttagctttaaagatcc</td>
</tr>
<tr>
<td>I64M/G66D</td>
<td>gcgcaccctctgtgaccagaatctgcag</td>
</tr>
<tr>
<td>G66D</td>
<td>gcggtagttatatataaacccgcttagctttaaagatcc</td>
</tr>
<tr>
<td>G99S/P102L</td>
<td>gcggtagttatatataaacccgcttagctttaaagatcc</td>
</tr>
<tr>
<td>T101A/P102L</td>
<td>gcggtagttatatataaacccgcttagctttaaagatcc</td>
</tr>
<tr>
<td>P102L</td>
<td>gcggtagttatatataaacccgcttagctttaaagatcc</td>
</tr>
<tr>
<td>L129P/T131G</td>
<td>gcggtagttatatataaacccgcttagctttaaagatcc</td>
</tr>
<tr>
<td>T131G</td>
<td>gcggtagttatatataaacccgcttagctttaaagatcc</td>
</tr>
<tr>
<td>Substitution</td>
<td>Oligonucleotide</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>F25L</td>
<td>ccggtgattataaaccgccttagccttaaagatcc</td>
</tr>
<tr>
<td>G99S/P102L</td>
<td>ccgtatatgccttttagtaaaaaacactgtcg</td>
</tr>
<tr>
<td>P102L</td>
<td>ctttggtaaaaaacactgtcggttcgtaaag</td>
</tr>
<tr>
<td>T131G</td>
<td>ggcagtttaatctggccggcaccaatgaaaaatgg</td>
</tr>
<tr>
<td>A155I</td>
<td>gtttgaaaaataatgcccgaatatcttaaggtgttagcgg</td>
</tr>
<tr>
<td>A166V</td>
<td>gcgggtcgctgccgatgtgatgattaccgatacgcg</td>
</tr>
<tr>
<td>P188L</td>
<td>ctggcagcagttctggtgataaacc</td>
</tr>
<tr>
<td>P197E</td>
<td>gaaccggtttaccatagtggaactggttttcatgattcgg</td>
</tr>
<tr>
<td>L198K</td>
<td>ccggtttaccccatagtcggaaaggttttcatgattcggaaag</td>
</tr>
<tr>
<td>P197E/L198K</td>
<td>ggttgatgaaccggtttaccatagtgaaaaaggttttcatgattcggaaaggtgatc</td>
</tr>
</tbody>
</table>

**ISOR-R3**

<table>
<thead>
<tr>
<th>Substitution</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>F25L</td>
<td>ccggtgattataaaccgccttagccttaaagatcc</td>
</tr>
<tr>
<td>G99S/P102L</td>
<td>ccgtatatgccttttagtaaaaaacactgtcg</td>
</tr>
<tr>
<td>P102L</td>
<td>ctttggtaaaaaacactgtcggttcgtaaag</td>
</tr>
<tr>
<td>T131G</td>
<td>ggcagtttaatctggccggcaccaatgaaaaatgg</td>
</tr>
<tr>
<td>A155I</td>
<td>gtttgaaaaataatgcccgaatatcttaaggtgttagcgg</td>
</tr>
<tr>
<td>P197E</td>
<td>gaaccggtttaccatagtggaactggttttcatgattcgg</td>
</tr>
<tr>
<td>L198K</td>
<td>ccggtttaccccatagtcggaaaggttttcatgattcggaaag</td>
</tr>
<tr>
<td>P197E/L198K</td>
<td>ggttgatgaaccggtttaccatagtgaaaaaggttttcatgattcggaaaggtgatc</td>
</tr>
</tbody>
</table>
Supplementary Figure 1. Amino acid binding profiles of Ws0279, AncCDT-1, and Pu1068. a, Ws0279. b, AncCDT-1. c, Pu1068. Left panels: examples of fluorescence-monitored thermal denaturation data in the absence (grey) and presence (black) of an amino acid. Three replicate curves are shown for each condition. Right panels: melting temperature ($T_M$) of each protein in the presence of amino acids (10 mM, except for Trp, Tyr and Cyi at 1 mM), relative to a protein-only control. Columns represent the mean of the experimental replicates, shown as circles. Asterisks indicate $\Delta T_M > 2$ °C and significantly different from the control by one-way ANOVA with Dunnett’s test for multiple comparisons, using a significance threshold of $P < 0.05$ (ANOVA: Ws0279, $F(23, 62) = 17.83$, $P < 0.0001$; AncCDT-1, $F(23, 55) = 60.98$, $P < 0.0001$; Dunnett’s test: Ws0279 + Arg, $P = 0.0039$; Ws0279 + Lys, AncCDT-1 + His, AncCDT-1 + Lys, AncCDT-1 + Arg, AncCDT-1 + Orn, all $P < 0.0001$). The $\Delta T_M$ for Ws0279 was 7.2 °C with 10 mM Lys and 6.1 °C with 1 mM Lys, comparable with $\Delta T_M$ values observed for other AABPs in the presence of their physiological ligands. $^{10}$ Orn, L-ornithine; Cyi, L-cystine.
Supplementary Figure 2. Phylogenetic analysis of CDT homologs. a-b, Maximum-likelihood phylogenies inferred using the LG substitution matrix (a) and the WAG substitution matrix (b). Branches are labeled with bootstrap values from 100 replicates. For each protein sequence, the NCBI accession code and the genus of the source organism are given. Experimentally characterized extant proteins are highlighted, and experimentally characterized ancestral nodes are labeled. The scale bar represents the mean number of substitutions per site. The outgroup of 271 AABP sequences is not shown. c, Posterior probability distributions of ancestral protein sequences at positions important for amino acid binding or CDT activity, as indicated by structural analysis or directed evolution. The sequences of Ws0279, Pu1068, Ea1174, and PaCDT at the corresponding positions are shown. The mean posterior probability of each ancestral sequence is given in parentheses.
Supplementary Figure 3. Multiple sequence alignment of ancestral and extant proteins characterized in this work. N-terminal signal peptides and C-terminal extensions that were not considered in the phylogenetic analysis are not shown.
Supplementary Figure 4. Characterization of ancestral and extant CDT variants. 

**a-c**, Complementation of auxotrophic E. coli ΔpheA cells in selective M9–F media by ancestral and extant CDT variants. Results are mean ± s.e.m. of replicate cultures (a, n = 3; b, n = 5; c, n = 3). 

**a**, Alternative versions of the ancestral proteins inferred using the WAG substitution matrix (AncCDT-1W to AncCDT-5W). 

**b**, AncCDT-3(P188L). 

**c**, Pu1068 and Ea1174. 

**d-f**, Michaelis-Menten plots for AncCDT-3 (d), AncCDT-3(P188L) (e) and PaCDT (f). Results are mean ± s.d. of technical replicates (d, n = 4, e, n = 3, f, n = 3). 

**g**, Conversion of 1.6 mM prephenate to phenylpyruvate by 20 μM PaCDT and AncCDT-2. No activity was detected for AncCDT-2. Results are mean ± s.d., n = 3 technical replicates. 

**h**, Kinetic parameters for prephenate dehydratase activity of CDT variants characterized in this work. Values are derived from non-linear fitting of the data in **Supplementary Figure 4d-f** and **Supplementary Figure 9c** to the Michaelis-Menten equation. Errors indicate s.e. for $K_M$ and $k_{cat}$, and errors propagated from these quantities for $k_{cat}/K_M$. n.d., no detectable activity.

<table>
<thead>
<tr>
<th>Protein</th>
<th>$K_M$ (μM)</th>
<th>$k_{cat}$ (s⁻¹)</th>
<th>$k_{cat}/K_M$ (M⁻¹ s⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AncCDT-3</td>
<td>1830 ± 190</td>
<td>(1.04 ± 0.07) × 10⁻²</td>
<td>5.67 ± 0.70</td>
</tr>
<tr>
<td>AncCDT-3(P188L)</td>
<td>294 ± 27</td>
<td>(4.58 ± 0.30) × 10⁻²</td>
<td>155 ± 18</td>
</tr>
<tr>
<td>PaCDT</td>
<td>18.7 ± 2.9</td>
<td>18.4 ± 0.7</td>
<td>(9.83 ± 1.60) × 10⁵</td>
</tr>
<tr>
<td>PaCDT(E173Q)</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>CDT-5W</td>
<td>134 ± 49</td>
<td>(6.03 ± 0.60) × 10⁻⁴</td>
<td>4.49 ± 1.69</td>
</tr>
</tbody>
</table>
Supplementary Figure 5. Ligand screening of Pu1068 and AncCDT-2. DSF was used to screen Pu1068, AncCDT-2, and AncCDT-1 (as a positive control) against 650 different conditions from six proprietary screens from Biolog (PM1–5 and PM9) and an in-house screen comprised of various additional compounds. ΔT_M values are given relative to a protein-only control. Compounds that produced a ΔT_M greater than 2 °C are listed. No binding of prephenate, the substrate of CDT, was observed. Details of the screen compositions and ligand concentrations are provided in Supplementary Data Set 1.
Supplementary Figure 6. NDSB-221 is a low-affinity ligand of Pu1068. a, Fluorescence spectrum of Pu1068 in the presence and absence of 10 mM NDSB-221, with an excitation wavelength of 280 nm. The figure shows data representative of two independent experiments with similar results. b, Fluorescence titration of Pu1068 with NDSB-221; peak fluorescence is plotted against ligand concentration. Two replicate titrations are shown. Fitting the data to a Boltzmann function gives a $K_d$ of 530 µM and a maximum fluorescence change of 20%. The structure of NDSB-221 is inset.
Supplementary Figure 7. Comparison of PaCDT crystal structures. a, Crystallographic oligomers of PaCDT viewed down the three-fold symmetry axis. The PaCDT-HEPES structure (PDB: 3KBR) and the low-resolution tetragonal PaCDT-acetate structure show a hexameric assembly, whereas the high-resolution trigonal PaCDT-acetate structure shows a trimeric assembly. b, Size-exclusion chromatogram of PaCDT. The figure shows data representative of two independent experiments with similar results. c, Calibration curve for analytical size-exclusion chromatography. Open circles represent molecular weight standards and the closed circle represents PaCDT. The calculated molecular weight of PaCDT is consistent with a trimeric structure (calc. 94 kDa, theor. 88 kDa for trimer). d-e, Conformational differences between the PaCDT-acetate structure (grey) and the PaCDT-HEPES structure (PDB: 3KBR, green). d, Superimposition of the two structures using the two large domains shows a rigid-body displacement of the small domain, which corresponds to an 11° rotation about the axis indicated by the blue line. This conformational change accounts for occlusion of the active site in the PaCDT-acetate structure. e, HEPES disrupts the hydrogen bonding network between Asp21, Asn152, and the general acid Glu173.
**Supplementary Figure 8. Mechanism of PaCDT.**

**a** Proposed mechanism for CDT-catalyzed decarboxylative aromatization of cyclohexadienols, and basis for transition state stabilization. The general acid Glu173 donates a proton to the departing hydroxyl group of the substrate. The given mechanism shows a concerted elimination of CO₂ and H₂O, although stepwise elimination of H₂O and CO₂ via a divinyl carbocation intermediate is an alternative possibility. **b** Conversion of 1.6 mM prephenate to phenylpyruvate by 20 µM PaCDT and PaCDT(E173Q). The E173Q substitution abolishes prephenate dehydratase activity. Results are mean ± s.d., n = 3 technical replicates. Data for PaCDT are duplicated from **Supplementary Figure 4g**; these experiments were done concurrently. **c** Circular dichroism (CD) spectra of PaCDT (black) and PaCDT(E173Q) (blue). The E173Q substitution does not disrupt the secondary structure of PaCDT. The figure shows data representative of two independent experiments with similar results. **d** CD-monitored thermal denaturation of PaCDT (black) and PaCDT(E173Q) (blue). The E173Q substitution has minimal impact on the Tₘ of PaCDT. The figure shows data representative of two independent experiments with similar results. The Tₘ of WT PaCDT was 56.6 °C (mean of two technical replicates [56.6 °C, 56.6 °C]) and the Tₘ of PaCDT E173Q was 54.9 °C (mean of two technical replicates [55.1 °C, 54.7 °C]).
Supplementary Figure 9. Directed evolution of AncCDT-2. a, Overview of the strategy used for directed evolution of AncCDT-2. b, Complementation of auxotrophic *E. coli* ΔpheA cells in selective M9–F media by CDT-M5 (mean ± s.e.m., three replicate cultures). AncCDT-2 and empty vector transformants were used as negative controls. c, Michaelis-Menten plot for CDT-M5 (mean ± s.d. of technical replicates; \( n = 4 \) for [prephenate] = 0.05, 0.10, 0.20, 0.80, 1.60 mM; \( n = 8 \) for [prephenate] = 0.40 mM). d, Sequences of AncCDT-2 variants with CDT activity, isolated by genetic selection of ISOR libraries. Amino acid substitutions in blue originated from the template gene (via site-directed mutagenesis), substitutions in green were encoded in oligonucleotides, and substitutions in orange were acquired randomly.
**Supplementary Figure 10. Active site structures of PaCDT and AncCDT-3(P188L).** The closed conformation of AncCDT-3(P188L) was modeled by superimposing the two domains of AncCDT-3(P188L) (dark green) separately on the PaCDT-acetate structure (grey, with docked prephenate in light green). Excluding Asn152, which is involved in crystal packing in the AncCDT-3(P188L) structure, the active site structures of the two proteins are virtually identical, despite the difference in global conformation.
Supplementary Figure 11. Protein-ligand interactions in the PaCDT-acetate complex. The large and small domains are shown in green and gray, respectively. Bond distances (Å) are indicated. The acetate molecule interacts with PaCDT via polar interactions with Arg85 and Ser80. Thr132 is the only residue of the small domain within potential hydrogen bonding distance of the acetate molecule. A water molecule is represented as a cyan sphere.
Supplementary Figure 12. Molecular dynamics simulations of PaCDT and AncCDT-1.

a-b, Projections of the trajectories of individual PaCDT subunits onto the PC1 axis. Each color represents a subunit of the PaCDT homotrimer. The dotted line represents the crystallographic PaCDT-acetate conformation (5HPQ). a, 4 × 100 ns simulations initialized from the trimeric PaCDT-HEPES structure using the GROMOS 53a6 force field. b, 1 × 150 ns simulation initialized from the trimeric PaCDT-acetate structure using the OPLS3 force field.

c-f, Projections of the trajectories of individual AncCDT-1 subunits onto the PC1 axis derived from simulations of closed AncCDT-1. Each color represents the trajectory of a replicate simulation of the monomeric protein. The dotted lines represent the crystallographic conformations of unliganded AncCDT-1, AncCDT-1/l-arginine, and AncCDT-3(P188L). Triplicate 100 ns simulations initialized from the closed AncCDT-1 monomer (c) and open AncCDT-1 monomer (d) structures using the GROMOS 53a6 force field.
simulations initialized from closed AncCDT-1 monomer (e) and open AncCDT-1 monomer (f) structures using the OPLS3 force field. g-i, Frequency histograms of the interdomain angle (left panel), the radius of gyration (middle panel), and the projection onto the first principal component (PC1) (right panel) from simulations of PaCDT and AncCDT-1. Each quantity can be used as a descriptor of the conformational change between the open and closed states of the protein. g, Frequency histograms for individual PaCDT subunits in the eight simulations using the GROMOS 53a6 force field. The corresponding values for the crystal structures of PaCDT (PaCDT-acetate and PaCDT-HEPES; PDB: 5HPQ and 3KBR, respectively) and AncCDT-3(P188L) (PDB: 5JOS) are also shown. h-i, Frequency histograms for the open AncCDT-1 (h) and closed AncCDT-1 (i) simulations using the GROMOS 53a6 force field. The corresponding values for the crystal structures of AncCDT-L-arginine (PDB: 5T0W), unliganded AncCDT-1 (PDB: 5TUJ), and AncCDT-3(P188L) (PDB: 5JOS) are also shown.
Supplementary Figure 13. Representative electron density for structures solved in this work. Blue mesh shows $2mF_o - DF_c$ maps contoured at 1σ. Green mesh shows $mF_o - DF_c$ omit maps corresponding to ligands, contoured at +3σ. (a) PaCDT-acetate, 2.05 Å, (b) PaCDT-acetate, 3.20 Å, (c) Pu1068, unliganded, 1.57 Å, (d) Pu1068, NDSB-221 complex, 1.88 Å, (e) AncCDT-1, l-arginine complex, 2.59 Å, (f) AncCDT-1, unliganded, 3.35 Å, (g) AncCDT-3(P188L), 2.10 Å.
References


3.4 Paper 2: Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme

3.4.1 Publication status

This manuscript presented in this section has been submitted. Supporting information is provided in Section 3.5. A version of this manuscript is also available on BioRxiv at https://doi.org/10.1101/2020.02.03.932491.

3.4.2 Author’s contribution

I processed and solved the X-ray crystal structure of AncCDT-5. I performed structural analysis on this and performed comparisons with the X-ray structures of related proteins. I performed the additional molecular dynamics simulations presented in this work and analysed and interpreted the results. I expressed and purified proteins used for the prephenate dehydratase activity assays. I designed and performed the prephenate dehydratase activity assays and analysed the results. I performed analytical SEC-MALS on the proteins in this study. I contributed to the overall interpretation of the results and made a significant contribution to the writing and editing of this manuscript. I prepared most of the figures in this manuscript.

3.4.3 Contributions from others

This work was a collaboration with Professor Gottfried Otting’s group (ANU Research School of Chemistry) and Professor Daniella Goldfarb’s group (Weizmann Institute of Science, Israel). Mithun C. Mahawaththa (Otting Group) performed the cloning, mutagenesis, protein expression and sample preparation for DEER studies. Mithun expressed, purified and crystallised AncCDT-5 and modelled the positions of the Gd(III) tags onto the structures of the proteins. Li Lynn Tan (Jackson Group) helped with the crystallisation, freezing and collection of X-ray diffraction data of AncCDT-5. Akiva Feintuch (Goldfarb Group) performed
DEER experiments and analysis. All authors contributed to the interpretation of the data and the writing and editing of the manuscript.
Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme

Joe A. Kaczmarski, a,5 Mithun C. Mahawaththa, a,5 Akiva Feintuch, b,5 Ben E. Clifton, a,c Luke A. Adams, d Daniella Goldfarb, b* Gottfried Otting, a* Colin J. Jackson a*

a Research School of Chemistry, The Australian National University, Canberra, ACT 2601, Australia.
b Department of Chemical Physics, Weizmann Institute of Science, Rehovot 76100, Israel.
c Present address: Protein Engineering and Evolution Unit, Okinawa Institute of Science and Technology, 1919-1 Tancha, Onna-son, Okinawa 904-0412, Japan.
d Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, VIC 3052, Australia.

§ These authors contributed equally.
* Corresponding authors. E-mails: daniella.goldfarb@weizmann.ac.il, gottfried.otting@anu.edu.au, colin.jackson@anu.edu.au.

Abstract

Several enzymes are known to have evolved from non-catalytic proteins such as solute-binding proteins (SBPs). Although attention has been focused on how a binding site can evolve to become catalytic, an equally important question is: how do the structural dynamics of a binding protein change as it becomes an efficient enzyme? Here we performed a variety of experiments, including propargyl-DO3A-GD(III) tagging and double electron–electron resonance (DEER) to study the rigid body protein dynamics of reconstructed evolutionary intermediates to determine how the conformational sampling of a protein changes along an evolutionary trajectory linking an arginine SBP to a cyclohexadienyl dehydratase (CDT). We observed that primitive dehydratases predominantly populate catalytically unproductive conformations that are vestiges of their ancestral SBP function. Non-productive conformational states, including a wide-open state, are frozen out of the conformational landscape via remote mutations, eventually leading to extant CDT that exclusively samples catalytically relevant compact states. These results show that remote mutations can reshape the global conformational landscape of an enzyme as a mechanism for increasing catalytic activity.
Introduction

Solute-binding proteins (SBPs) comprise a large superfamily of extra-cytoplasmic receptors that are predominantly involved in sensing and the uptake of nutrients, including amino acids, carbohydrates, vitamins, metals and osmolytes\(^1\)-\(^4\). The SBP fold consists of two $\alpha/\beta$ domains linked by a flexible hinge region that mediates conformational change in solution, with hinge-bending and hinge-twisting motions moving the two domains together (closed state) and apart (open state; Figure 1a). Ligands bind at the cleft between the two domains and stabilize the closed state by forming bridging interactions between the two domains and the hinge region\(^1\). The ligand-induced open-to-closed conformational switch is important for the function of SBPs in ATP-binding cassette transporter systems\(^1,5,6\), tripartite ATP-independent periplasmic transporter sytems\(^7\) and signalling cascades\(^4\).

Although the open conformation is the ground state for most ligand-free SBPs\(^1,5,8\)-\(^19\), numerous SBPs also sample semi-closed and closed states in the absence of ligands. For example, X-ray crystallography, nuclear magnetic resonance (NMR), molecular dynamics (MD) simulation, double electron-electron resonance (DEER, a.k.a. PELDOR) and Förster resonance energy transfer (FRET) studies on the maltose-binding protein (MBP)\(^20\)-\(^22\), glucose-galactose-binding protein\(^23,24\), histidine-binding protein\(^25\), ferri-bacillibactin-binding protein\(^26\), glutamine binding protein\(^27,28\) and choline/acetylcholine binding protein\(^29\) demonstrate sampling of both open and closed conformations in the absence of ligands. The intrinsic open/closed equilibrium is fundamental to determining the binding affinity\(^21,23,30,31\) and binding promiscuity\(^5,32\) of SBPs, and controlling the transport activity of SBP-associated systems\(^5,28\). Indeed, the extent of the open/closed motion differs between SBPs\(^1\), and the function of an SBP can be changed by mutations that alter conformational sampling, without changing the architecture of the SBP-ligand interface\(^19,21,30\).

While most SBPs are non-catalytic binding proteins, a small fraction of proteins within the periplasmic binding protein-fold superfamily have evolved enzymatic activity\(^33\)-\(^37\). One example is cyclohexadienyl dehydratase (CDT), which is closely related to the polar and cationic amino acid binding proteins, such as the arginine binding protein (ArgBP)\(^37,38\). Like ArgBP, CDT adopts a periplasmic type-II SBP fold, but instead of binding amino acids it catalyzes the Grob-like fragmentation of prephenate and L-arogenate to form phenylpyruvate and L-phenylalanine, respectively\(^39\). We previously used ancestral sequence reconstruction to show that the trimeric CDT from Pseudomonas aeruginosa (PaCDT) plausibly evolved from a monomeric cationic amino acid binding protein ancestor (AncCDT-1) via a series of
intermediates (AncCDT-2 to AncCDT-5)\textsuperscript{40} (Figure 1b). While AncCDT-1 and AncCDT-2 did not display any enzymatic activity and appeared to be binding proteins, low-level CDT activity became detectable in two alternative reconstructions of AncCDT-3(P188/L188) ($k_{\text{cat}}/K_M \sim 10^1/10^2\, \text{s}^{-1}\, \text{M}^{-1}$) and increased along the trajectory towards the efficient extant enzyme PaCDT ($k_{\text{cat}}/K_M \sim 10^6\, \text{s}^{-1}\, \text{M}^{-1}$). Despite the large difference in catalytic efficiency, AncCDT-3 and PaCDT share 14/15 inner-shell residues, including all catalytic residues, but only about 50 % amino acid sequence identity over the rest of the protein (the outer shells). This suggests that these outer-shell substitutions must be substantially responsible for the $\sim 10^5$-fold increase in catalytic efficiency. For example, the remote P188L substitution in AncCDT-3 was shown to result in a 27-fold increase in $k_{\text{cat}}/K_M$\textsuperscript{40}.

In this study, we investigated two plausible explanations for how remote mutations could have led to an increase in catalytic activity along this evolutionary trajectory: they may have (i) altered the sampling of rotamers of active site residues, such as the general acid Glu173, thereby changing the structure and character of the active site and controlling the configuration of active site residues, and/or (ii) altered the equilibrium between open and closed states of the protein to minimize sampling of the catalytically unproductive open state. To test these hypotheses, we used a combination of protein crystallography, MD simulations and DEER distance measurements on protein variants into which we incorporated the unnatural amino acid $p$-azidophenylalanine (AzF) to allow biorthogonal conjugation, making it specific even in the presence of native cysteine residues, to a propargyl-DO3A-Gd(III) tag, which encapsulates the Gd(III) ion in a hydrophilic complex of neutral charge, while providing a relatively rigid linker between the Gd(III) ion and protein backbone\textsuperscript{41}. The Gd(III)–Gd(III) DEER measurements were carried out at W-band frequency (94.9 GHz), affording superior sensitivity and measurements free from orientation selection\textsuperscript{42,43} and multi-spin effects\textsuperscript{44–47}. The results from these experiments showed that the ancestral proteins, which all displayed either no catalytic activity (AncCDT-1) or very high $K_M$ values that were outside the physiologically relevant substrate concentration (AncCDT-3, AncCDT-5), significantly or even predominantly sampled open states, including a wide-open state that is not observed in the extant and efficient enzyme PaCDT. Finally, structural analysis revealed that multiple changes to intra- and intermolecular (via oligomerization) interaction networks have shifted the conformational equilibrium towards more compact states along the evolutionary trajectory.
Figure 1. The conformational change of SBPs and the evolution of PaCDT. (A) X-ray crystal structures of SBPs that are specialized for binding solutes, such as AncCDT-1 (shown), typically capture open ligand-free (left, PDB 5TUJ) and closed liganded (right, PDB 5T0W) states. (B) Schematic drawing (not to scale) of the phylogenetic tree used for ancestral sequence reconstruction in Clifton et al., 2018\textsuperscript{40}, which highlights the evolutionary relationship between the polar amino acid binding proteins (e.g. Thermotoga maritima L-arginine binding protein, TmArgBP), AncCDT-1, AncCDT-3, AncCDT-5 and PaCDT. Clades are collapsed. Figure adapted from Clifton et al., 2018\textsuperscript{40}.
Results

The evolution of prephenate dehydratase activity. We previously showed that while AncCDT-1 binds L-arginine with comparable affinity to extant L-arginine binding proteins, the P188 AncCDT-3 reconstruction (hereafter AncCDT-3/P188), AncCDT-5 and PaCDT have all lost the ancestral ability to bind proteogenic amino acids\textsuperscript{40}. Instead, along with the AncCDT-3/L188 AncCDT-3 reconstruction, they can rescue the growth of a phenylalanine auxotroph of \textit{E. coli} (ΔpheA) grown in the absence of L-phenylalanine, indicating their capacity to deliver the dehydratase activity required to synthesize L-phenylalanine from L-arogenate, or the precursor phenylpyruvate from prephenate\textsuperscript{40}. Furthermore, these variants exhibited prephenate dehydratase activity \textit{in vitro}, which increased about 10\textsuperscript{5}-fold ($k_{\text{cat}}/K_M$) between AncCDT-3/P188 and PaCDT. In this work, we have additionally assessed the dehydratase activity of AncCDT-5, which proved to be intermediate between the AncCDT-3 variants and the extant enzyme PaCDT (Table 1, Supplementary Figure 1). Although the $k_{\text{cat}}$ value of AncCDT-5 (4 s\textsuperscript{-1}) exceeds the level observed in the AncCDT-3 variants (~10\textsuperscript{-2} s\textsuperscript{-1}) almost as much as PaCDT (18 s\textsuperscript{-1}), the $K_M$ of AncCDT-5 remains ~15-fold higher than that of PaCDT (277 µM vs. 19 µM). This value is well above the likely physiological concentration of the substrate (the intracellular concentrations of molecules in this biosynthetic pathway are about 14–18 µM in Gram-negative bacteria\textsuperscript{48}).

Table 1. Kinetic parameters for prephenate dehydratase activity of CDT variants related to this study.*

<table>
<thead>
<tr>
<th></th>
<th>$K_M$ (µM)</th>
<th>$k_{\text{cat}}$ (s\textsuperscript{-1})</th>
<th>$k_{\text{cat}}/K_M$ (M\textsuperscript{-1} s\textsuperscript{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td>AncCDT-1**</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>AncCDT-3/P188**</td>
<td>1830 ± 190</td>
<td>(1.04 ± 0.07) × 10\textsuperscript{-2}</td>
<td>5.67 ± 0.70</td>
</tr>
<tr>
<td>AncCDT-3/L188**</td>
<td>294 ± 27</td>
<td>(4.58 ± 0.30) × 10\textsuperscript{-2}</td>
<td>155 ± 18</td>
</tr>
<tr>
<td>AncCDT-5</td>
<td>277 ± 4</td>
<td>4.1 ± 0.2</td>
<td>(1.5 ± 0.1) × 10\textsuperscript{4}</td>
</tr>
<tr>
<td>PaCDT**</td>
<td>18.7 ± 2.9</td>
<td>18.4 ± 0.7</td>
<td>(9.8 ± 1.6) × 10\textsuperscript{5}</td>
</tr>
</tbody>
</table>

ND, not detected.

* Results are reported as mean ± standard error for $K_M$ and $k_{\text{cat}}$. Errors for $k_{\text{cat}}/K_M$ were obtained by propagation. Source data are provided as a Source Data file.

** Sequence and structural analysis of AncCDT-3/L188 and PaCDT revealed that although their catalytic efficiency differ by ~10\textsuperscript{5} M\textsuperscript{-1} s\textsuperscript{-1}, they share 14 of 15 residues in the active/substrate binding site (the only difference being a conservative Thr80Ser substitution)\textsuperscript{40}. Comparison between the structures of AncCDT-3/L188 and PaCDT revealed that the catalytic general acid Glu173
adopts a rotameric state in AncCDT-3/L188 that differs from that observed in PaCDT (Figure 2a)\textsuperscript{40}. This conformational difference appears to be caused by the neighbouring Tyr177Gln mutation that occurs on the branch between AncCDT-3 and AncCDT-5; the presence and position of the side chain of Tyr177 in AncCDT-3 (held in place, in part, through pi-stacking with the neighbouring Phe156) sterically prevents Glu173 from adopting the catalytically competent conformation observed in PaCDT\textsuperscript{40}. Additionally, this mutation causes coupled conformational changes in the neighbouring residues Phe156 and Met167, which leads to further changes in the shape and electrostatic character of the active site. These findings are analogous to those from recent work on enzyme variants produced through directed evolution and computational design that show how second-shell mutations affect enzyme activity by constraining the conformational sampling of catalytic residues\textsuperscript{49,50}.

To better understand how the mutations along this evolutionary trajectory affect activity, we also solved a 1.49 Å X-ray crystal structure of AncCDT-5 (Supplementary Table 1), which shares 15/15 residues in the active/substrate binding site with PaCDT. In contrast to the comparison between AncCDT-3 and PaCDT, a structural alignment of AncCDT-5 and PaCDT (PDB 3KBR) reveals that these two proteins not only share identical binding sites in terms of amino acid composition, but that these residues also adopt identical conformations (Figure 2b). Since these active/binding sites are identical, the ~15-fold decrease in $K_M$ and ~5-fold increase in $k_{\text{cat}}$ of the dehydratase activity must be attributed to amino acid substitutions that are outside of the binding site. The sequences of AncCDT-5 and PaCDT differ at 98 positions. Excluding the 9 amino acid C-terminal extension of PaCDT, the remaining 89 positions are evenly distributed around the SBP fold (Figure 2c), making it difficult to rationalize how individual mutations contribute to changes in catalytic profiles. Instead, it is possible that some of the substitutions may be affecting catalysis by altering the open/closed dynamics of these proteins.
Figure 2. Crystal structures of AncCDT-3/P188, AncCDT-5 and PaCDT. (A) Structural overlay of the small domains of AncCDT-3/L188 (purple, PDB 5JOS), AncCDT-5 (green) and PaCDT (white, PDB 3KBR). The Tyr177Gln substitution that occurs between AncCDT-3 and AncCDT-5 contributes to the repositioning of Glu173. (B) Structural alignment of the active sites of PaCDT (PDB 3KBR, white/grey) and AncCDT-5 (green). The small domain (white, residues 99–194) and large domain (grey, residues 1–98 and 195–236) of AncCDT-5 were individually aligned with the corresponding domains of PaCDT. The side chains of the 15 inner-shell residues identified by Clifton et al., 2018 are shown as sticks (Gly131 not shown), highlighting identical active site residues and side-chain conformations between AncCDT-5 and PaCDT. The HEPES molecules present in the crystal structures are omitted for clarity. (C) The locations of the substitutions (green spheres) between AncCDT-5 and PaCDT are shown projected onto the structure of PaCDT (PDB 3KBR). Inner-shell residues are shown as orange sticks.
Solution dynamics of the amino acid binding protein, evolutionary intermediates and PaCDT. Next, we investigated whether the accumulation of substitutions between AncCDT-1 and PaCDT at positions remote from the active site could have affected catalytic activity by altering the open/closed dynamics of these proteins. To do so, we used DEER distance measurements and MD simulations to investigate whether the distribution of open and closed solution states differed between the ancestral amino acid binding protein (AncCDT-1), the intermediate low-activity CDTs (AncCDT-3, AncCDT-5) and the efficient extant enzyme, PaCDT.

AncCDT-1. First, we confirmed that AncCDT-1 was monomeric (>95%) in solution using size-exclusion chromatography-multi angle light scattering (SEC-MALS; (Supplementary Figure 2a). We then assessed the solution conformations of AncCDT-1 using DEER distance measurements with Gd(III)-tags introduced through unnatural amino acid (UAA) mutagenesis (Supplementary Figure 3). Amber stop codons were used to incorporate the UAA AzF into the small (Lys138) and large (Gln68) domains of AncCDT-1. These sites were chosen because they (i) are known to be located in regions of the protein that undergo substantial changes in their relative distance upon ligand binding (Supplementary Figure 4), (ii) are located on rigid α-helices, (iii) are surrounded by residues that will minimize unwanted tag motions via packing and (iv) are at the surface of the protein, which minimizes the risk that mutagenesis and tagging at these positions will significantly affect the open-closed protein conformations. The mutant protein heterologously expressed in soluble form from E. coli, yielding ~15 mg of purified protein per litre of cell culture. To remove any co-purified ligands, on-column protein denaturation and refolding was performed to obtain ligand-free samples. This is referred to as refolded AncCDT-1. The AzF residues were modified with propargyl-DO3A-Gd(III) tags as previously described (Supplementary Figure 3a) and mass spectrometry analysis indicated tag ligation yields of >80% at each site.

DEER measurements of refolded AncCDT-1-propargyl-DO3A-Gd(III) revealed a distance distribution with two dominant peaks with maxima corresponding to Gd(III)–Gd(III) distances of 3.2 nm and 4.4 nm (Figure 3a, Supplementary Table 2, Supplementary Figure 4a,b). When saturating concentrations of L-arginine were added to the refolded sample, the distance distribution revealed an almost complete shift to the shorter distance (Figure 3a), as expected for a ligand-induced open-to-closed transition. Together, these data indicate that the Gd(III)–Gd(III) distance varies by ~1.2 nm between the closed state and any open conformations and that the DEER approach employed here can be used to differentiate between
distinct conformational substates of a SBP. Experiments with natively purified protein rather than unfolded/refolded protein showed similar results (Supplementary Figure 4a,b), except that a greater proportion of the protein adopted the closed state (3.2 nm Gd(III)–Gd(III) distance) in the absence of exogenous L-arginine than in the sample subjected to unfolding and renaturation. This was expected, as the protein is known to co-purify from *E. coli* with some ligand bound, which stabilizes the closed conformation<sup>40</sup>. A ligand-induced shift towards a more closed state was also observed when L-arginine was added to a sample of a lysine-arginine-ornithine binding protein from *Salmonella enterica* (SeLAOBP, Supplementary Figure 5).

We also performed experiments on AncCDT-1 that had been tagged at two sites on either the large domain (positions 68 and 219, Supplementary Figure 4c) or small domain (positions 138 and 161; Supplementary Figure 4d). These each showed a peak in the distance distribution that were in agreement with Gd(III)-Gd(III) distance predictions based on crystal structures of AnCDT-1 (Supplementary Figure 4g,h). These results confirm that the two distances measured for refolded AncCDT-1 are due to rigid body motions of the two domains, rather than any intra-domain flexibility or tag dynamics.
Figure 3. Solution dynamics of AncCDT-1, AncCDT-3/P188, AncCDT-5 and PaCDT probed by DEER. Gd(III)–Gd(III) distance distributions for samples of propargyl-DO3A-Gd(III)-tagged (A) refolded AncCDT-1 (+/− L-Arg, tagged at positions 68 and 138), (B) natively purified AncCDT-3/P188 (tagged at positions 68 and 138), (C) natively purified AncCDT-5 (tagged at positions 68 and 138) and (D) natively purified PaCDT (tagged at positions 68 and 139 and diluted by a factor of 10 using unlabelled PaCDT). Distance distributions were obtained using DeerAnalysis. The original DEER data and DEER form factor traces are shown in Supplementary Figures 4, 6 & 7. Vertical dashed lines represent the Gd(III)–Gd(III) distances estimated by modelling propargyl-DO3A-Gd(III) tags onto the crystal structures of each protein. The shaded areas represent the range of calculated Gd(III)–Gd(III) distances when propargyl-DO3A-Gd(III) tags are modelled onto a number of individual MD snapshots representing the closed (red), open (grey) and wide-open (orange) states.

We next sought to correlate the distance changes observed through DEER to structural changes in the protein. We previously solved the crystal structure of AncCDT-1 bound to L-arginine (PDB 5T0W), which has a Cα–Cα distance between Gln68 and Lys138 of 2.6 nm (equivalent to a Gd(III)–Gd(III) distance of ~2.9 nm, Supplementary Figure 4e). We also solved a structure of the protein in an open state (PDB 5TUJ) in which the Cα–Cα distance between Gln68 and Lys138 is 3.4 nm (Gd(III)–Gd(III) distance of ~4.1 nm, Supplementary Figure 4f). This crystal structure presents a snapshot of an open conformation, but additional open
states may be accessible to the protein in solution. To probe these, we also explored the solution dynamics of AncCDT-1 using MD simulations initiated from the closed X-ray crystal structure of AncCDT-1 (PDB 5T0W) with the ligand, L-arginine, removed. These were conducted solely to identify conformational substates that can be accessed on the ns-µs timescale from the initial closed conformation (i.e. easily accessible, low-energy states), not to obtain quantitative predictions of the solution dynamics or the relative occupancy of each state. The simulations suggest that AncCDT-1 can easily access three dominant conformational substates (Figure 4a, Supplementary Table 2): the closed state corresponding to PDB 5T0W (68-138 Cα–Cα distance range of 2.5-2.9 nm; equivalent to a Gd(III)–Gd(III) distance range of ~2.9-3.4 nm), an intermediate open state corresponding to PDB 5TUJ (68-138 Cα–Cα distance 3.2-3.7 nm; equivalent to a Gd(III)–Gd(III) distance range of ~3.9-4.9 nm) and an additional wide-open state (68-138 Cα–Cα distance 3.6-4.2 nm; equivalent to a Gd(III)–Gd(III) distance range of ~4.1-5.2 nm).

The results obtained by DEER suggest that the dominant states are indeed the closed state and an open conformation closely related to the open state observed through protein crystallography, but which might also include the wide-open state, as there is significant overlap between the modelled Gd(III)–Gd(III) distances in the snapshots obtained from MD. Although we do not have crystallographic evidence that AncCDT-1 samples the wide-open state, analogous wide-open states have been observed in crystal structures of extant L-arginine binding proteins, such as that from Thermotoga maritima (TmArgBP, PDB 4PRS)38, which has a Cα–Cα distance of 4.0 nm between positions equivalent to 68-138 in AncCDT-1 (which is similar to the 3.6-4.2 nm range observed for the wide-open state in AncCDT-1 MD simulations). Indeed, it has been suggested that the specific open states of SBPs observed crystallographically can be dictated by crystal packing interactions8–11. The wide-open state thus is a part of the conformational landscape for L-arginine binding proteins, although we could not unambiguously detect it in our DEER experiments on AncCDT-1 owing to overlap with the open state. Altogether, the data from DEER, protein crystallography and MD demonstrate that AncCDT-1 can sample both open and closed states in the absence of ligands, and that ligand binding shifts the conformational equilibrium towards the closed state.
**Figure 4. Conformational substates captured by MD and crystallography.** Conformational substates of (A) AncCDT-1, (B) AncCDT-3, (C) AncCDT-5 and (D) PaCDT captured by MD and X-ray crystallography. The plot of Cα–Cα distances between tagged residues vs. radius of gyration obtained from MD simulation snapshots highlights populations (or lack thereof) of closed, open and wide-open conformational states. Each data point represents a single frame of the MD simulation (sampled every 0.1 ns). Crosses represent snapshots of closed (red), open (cyan) and wide-open (magenta) states that were used to model Gd(III)–Gd(III) distances. Representative structures of each of these states are shown on the right and highlight the extent of the open/closed transition, with distances between Cαs (red spheres) of tagged sites shown by yellow lines. Dark grey structures are crystal structures (corresponding to black dots on plot), while white structures are MD snapshots. The time evolution of Cα–Cα distances are shown in Supplementary Figure 9.

**AncCDT-3 and AncCDT-5.** Considering the significant difference in dehydratase activity between AncCDT-3, AncCDT-5 and PaCDT, as well as our inability to fully rationalize these differences through protein crystallography, we next used DEER and MD to investigate the solution conformational distributions of these proteins. Like AncCDT-1, SEC-MALS indicated
that AncCDT-3/P188 and AncCDT-5 were predominantly monomeric in solution (>95%), as reported previously40 (Supplementary Figure 2b,c). We tagged the small and large domains of AncCDT-3/P188 and AncCDT-5 at sites 68 (large domain) and 138 (small domain) in the same way as AncCDT-1 and performed DEER experiments (Figure 3b,c, Supplementary Figure 6a,b). For natively purified AncCDT-3/P188, the DEER distance distributions showed a clear maximum that corresponded to a Gd(III)–Gd(III) distance of ~5.2 nm (Figure 3b, Supplementary Figure 6a). Likewise, AncCDT-5 displayed the most prominent distance-distribution peak at ~5.4 nm (Figure 3c, Supplementary Figure 6b). Similar results were obtained for the refolded proteins (Supplementary Figure 6a,b), where the observed shifts are within experimental error.

We previously solved a crystal structure of AncCDT-3/L188 and found a conformation analogous to the wide-open conformation observed in the MD simulation of AncCDT-1 (PDB 5JOS)40. In this structure, the Q68-R138 Ca–Ca distance is 4.2 nm (Gd(III)–Gd(III) distance: ~5.3 nm). In triplicate 500 ns MD simulations initiated from the open AncCDT-3/L188 and AncCDT-3/P188 structures (after modelling the L188P mutation), the protein did not sample the closed state at all across the combined 1.5 µs of simulation time and the wide-open state was the major conformation (Figure 4b), exhibiting a Ca–Ca distance range of 4.0-4.8 nm (Gd(III)–Gd(III) distance range: 5.0-6.3 nm).

Unlike the wide-open crystal structure of AncCDT-3/L188, the 1.49 Å crystal structure of AncCDT-5 determined in the present work adopts the closed conformation (a HEPES molecule is bound in the active site), with a Q68-R138 Ca–Ca distance of 2.8 nm (Gd(III)–Gd(III) distance: 3.1 nm). Beginning from this closed state, we again performed triplicate 500 ns MD simulations after removing the HEPES molecule. The simulations predict that the protein rapidly adopts a wide-open conformation, with a Q68-R138 Ca–Ca distance range of 3.6-4.2 nm (Gd(III)–Gd(III) distance range: 4.5-5.4 nm) that is analogous to the wide-open conformation seen in the MD simulations of AncCDT-1, the crystal structure of the L-arginine binding protein from T. maritima, and the crystal structure of AncCDT-3 (Figure 4c). An intermediate state, analogous to the open crystal structure of unliganded AncCDT-1 was also briefly sampled, which corresponded to a Q68-R138 Ca–Ca distance range of 3.0–3.4 nm (Gd(III)–Gd(III) distance range: 3.4–4.2 nm). The combination of the DEER, crystallographic and MD data for the experiments performed on AncCDT-3 and AncCDT-5 suggest that both proteins predominantly access a wide-open state in solution. Smaller peaks in the distance
distribution could correspond to minor fractions of intermediate states but are too small to be unambiguous.

**PaCDT.** We next aimed to probe the open/closed dynamics of **PaCDT.** SEC-MALS experiments confirmed that **PaCDT** is primarily homotrimeric in solution\(^{40}\), with a small population (< 5 %) of higher order or aggregated species (Supplementary Figure 2d), which complicates the measurement of intra-monomer distances by DEER experiments. To confirm the existence of the tagged trimer, we first measured samples with propargyl-DO3A-Gd(III) tags at single AzF residues of each **PaCDT** monomer, at either the large domain (position 68) or the small domain (position 139). The maxima of the DEER distance distributions measured were consistent with those expected for the trimeric structure observed by X-ray crystallography (3.5 nm vs. 3.3 nm for proteins tagged at position 68, 5.4 nm vs. 4.6 nm for samples tagged at position 139; Supplementary Figure 7a,b). This indicated that the trimeric structure was preserved under the conditions of the DEER experiments, including sample preparation and freezing. When we attempted to measure the open/closed distribution via labelling at sites on both the large and small domain (68 and 139, respectively) the observed Gd(III)–Gd(III) distance distribution was very broad as was expected due to multiple inter-chain Gd(III)–Gd(III) distances in a trimer (Supplementary Figure 7c). In order to extract the inter-domain distance from a single **PaCDT** monomer, a protein sample labelled at sites 68 and 139 was diluted 10-fold with unlabelled protein and allowed to equilibrate at room temperature, with the aim of obtaining a mixture in which there would be a ~24 %, 3 % and 0.1 % chance of encountering a trimer consisting of 1, 2 or 3 tagged monomers, respectively. DEER experiments on samples prepared in this way yielded a significantly more narrow peak that indicated 3.5 nm was the predominant Gd(III)–Gd(III) distance (Figure 3d, Supplementary Figure 7d).

We again used protein crystallography and MD simulation to identify stable conformations of **PaCDT.** There are three crystal structures, one with HEPES bound at the active site (Structural Genomics Project, PDB 3KBR) and two with acetate bound at the active site\(^ {40}\). The highest resolution acetate-bound structure (PDB 5HPQ) and the HEPES-bound structure (PDB 3KBR) display Q68-A139 Cα–Cα distances of 2.5 nm (Gd(III)–Gd(III) distance: ~3.5 nm). A MD simulation initiated from the trimeric X-ray crystal structure of acetate-**PaCDT** (PDB 5HPQ), with acetate removed, indicated that the monomers can sample both closed (Cα–Cα distance range of 2.5–3.0 nm; equivalent to a Gd(III)–Gd(III) distance
range of ~3.5–3.9 nm) and open (Cα–Cα distance range of 3.2–3.8 nm; equivalent to a Gd(III)–Gd(III) distance range of ~4.0–4.5 nm) conformations, but there was no evidence of a wide-open conformation (Figure 4d). The Gd(III)–Gd(III) distances predicted for the closed and open states notably fall either side of the broad peak centred at 3.9 nm. This could be consistent with a broad distribution of rapidly interconverting open and closed states that have been snap-frozen, in contrast with the AncCDT-1 DEER analysis, which displayed two distinct peaks corresponding to similar distances.

All three crystal structures of PaCDT display closed conformations, which leave limited access to the active site. To allow exchange of substrate and product, the enzyme must also populate open conformations, which are likely related to those obtained in the MD simulations. Interestingly, however, the wide-open conformation, which appears in MD simulations of AncCDT-1, AncCDT-3 and AncCDT-5, was not significantly sampled. This is, again, consistent with the DEER measurements, where the prominent peaks at ~5.4 nm in AncCDT-3 and AncCDT-5 were minimal (or absent) for PaCDT. Altogether, the combined use of DEER, crystallography and MD simulations suggests that the wide-open conformation observed in the ancestral reconstructions of CDT (computationally observed in all variants and empirically observed in AncCDT-3, AncCDT-5 and orthologs of AncCDT-1) has been largely eliminated from the conformational landscape of the extant PaCDT. This provides a molecular explanation for the observed change in $K_M$ between AncCDT-5 and PaCDT; despite their identical substrate binding sites when closed, the open ligand-binding state of PaCDT likely has higher affinity for the substrate as it is much closer to the conformation of the Michaelis complex than the wide-open state that is predominantly populated in AncCDT-3 and AncCDT-5.

**Structural basis for the increased dehydratase activity of PaCDT.** Having established that the conformational sampling of PaCDT is different from AncCDT-5, we compared the structures of AncCDT-5 and PaCDT to ascertain the molecular basis for this change. Structural comparison between the two proteins revealed three regions with substantial differences (unlike the substrate binding site) that could collectively stabilize the closed state and/or destabilize the wide-open state. First, in comparison to AncCDT-5, which terminates at Lys235 (although it is not present in clear electron density), PaCDT has an additional nine amino acids at the C-terminus (RWPTAHGKL), of which the first 4 and 6 residues produce clear electron density in crystal structures of PaCDT (PDB 5HPQ and 6BQE, respectively). As shown in Figure 5a, these amino acids interact extensively with both the small and large domains,
substantially increasing the number of van der Waals and hydrogen bond interactions between the two domains in the closed and open, but not wide-open, conformation.

Second, one of the most substantial structural differences between AncCDT-5 and PaCDT is the oligomerization of PaCDT, which forms a trimer (Figure 5b). A number of key residues that contribute to the inter-subunit interfaces in the crystal structures of PaCDT are not found in AncCDT-5, explaining this difference in oligomeric states. For example, the substitutions T97R, M221Q and L217H introduce specific electrostatic interactions between chains in PaCDT, while H195F, H218I and P83L appear to improve packing between the subunits. However, oligomerization is a complex process and remote and unpredictable mutations can also have substantial influence on oligomeric states. Accordingly, we cannot exclude the possibility that additional mutations remote from the interface also play a role.

In the resulting trimeric form of PaCDT, the small domain remains free to move, allowing fluctuation between the closed and open states. However, there are inter-subunit contacts at the hinge region that appear to preclude some of the conformational changes expected in the wide-open state. Specifically, the crystal structure of the wide-open conformation of AncCDT-3 shows that the hinge region with His195 at the centre is fully solvent exposed and extended. The equivalent region of PaCDT, however, is in a retracted conformation, with Phe195 (chain A; the equivalent position to His195 in AncCDT-3) being stabilized in a buried hydrophobic pocket formed by Ile218 (chain A) and the non-polar region of Ser222 (chain A), as well as the non-polar regions of two hydrogen-bonded pairs, Arg97 (chain A):Gln221 (chain B) and Glu197 (chain A):His217 (chain B). The positioning of these residues and interactions with the neighbouring subunit would likely stabilize PaCDT in the closed and open conformations, while limiting the sampling of wide-open states.

Finally, for the proteins to adopt the wide-open state, the region between Val187 and Pro191 must undergo a conformational change, from the tightly kinked conformation seen in the closed structures of AncCDT-5 and PaCDT (PDB 5HPQ, 6BQE; it is in an altered conformation due to crystal packing interactions in 3KBR), to an extended conformation seen in the crystal structure of the wide-open conformation of AncCDT-3 (Figure 5c). In AncCDT-3/5, the region between Val187 and Pro191 is relatively flexible, consisting of polar and relatively small hydrophobic residues (Leu/Val/Glu in AncCDT-3; His/Val/Glu in AncCDT-5). In contrast, in PaCDT, the central residue in the kink is mutated to proline, and the neighbouring glutamine forms a hydrogen bond to Arg186, stabilizing the kinked structure. Further support for the hypothesis that this region is important in conformational sampling and dehydratase activity can be found in the observation that (i) the P188L mutation in AncCDT-3
causes a large decrease in $K_M^{40}$, (ii) it was observed to be a mutational hotspot during the directed evolution of ancestral CDTs for increased dehydratase activity$^{40}$ and (iii) this region has been recognized to be important for dynamics and affinity in other SBPs such as MBP$^{21,30}$.

The combination of these three structural effects is likely sufficient to substantially change the relative conformational free energy of the closed, open and wide-open states, consistent with the DEER and MD results. Interestingly, none of the structural changes observed between AncCDT-3 and PaCDT preclude PaCDT adopting the open conformation, but essentially prevent it adopting the wide-open conformation. Given the large number of changes between AncCDT-5 and PaCDT, it is impossible to ascribe the difference in populations of open and wide-open conformations to a single mutation or structural rearrangement. Indeed, recent work has highlighted the gradual nature by which protein conformational sampling can change, and the overlapping and often redundant effects of mutations$^{49}$. 
Figure 5. Structural basis for change in conformational sampling. (A) Comparison of the crystal structures of AncCDT-5 (left) and PaCDT (5HPQ, right) highlight how the C-terminal extension (residues beyond 232 are highlighted in red) in PaCDT facilitates additional interactions between the small (white) and large (grey) domains. (B) PaCDT is a trimer in solution, and analysis of the crystal structures of PaCDT (e.g. 5HPQ, shown here) reveals several interactions between the hinge region of chain A with the large domain of chain B that are likely to prevent the wide-open conformation being sampled. These include packing of Phe195 against the neighbouring subunit, and polar interactions between neighbouring residues. Overlay of the crystal structure of AncCDT-3/L188 (purple, PDB 5JOS) further highlights how the wide-open state is likely to be incompatible with the trimeric arrangement of PaCDT. (C) The flexibility of the hinge region is known to control the open/closed dynamics of SBPs. Sequence and structural comparison of these regions in AncCDT-3/L188 (purple,
PDB 5JOS), AncCDT-5 (green, PDB 6WUP) and PaCDT (white, PDB 5HPQ) highlight that this is an intrinsically flexible region, which is likely affected by mutations.

**Discussion**

SBPs have become a model system for understanding the role of rigid-body motions in determining ligand and substrate binding affinity, specificity and kinetics. Early crystallographic studies suggested that SBPs act as simple binary switches, adopting a closed conformation when in complex with ligands and an open conformation in the ligand-free state\(^4\). More recently, however, NMR, FRET and MD studies have probed the intrinsic dynamics of SBPs, showing that they can adopt closed and semi-closed states in the absence of ligands\(^{20-29}\), and have provided evidence that the extent of the open-closed dynamics evolves to support specific binding properties and function. Indeed, altering the open-closed dynamics of SBPs is now also being used as an additional parameter for engineering new binding functions into SBPs\(^{21}\) and the development of novel SBP-based biosensors\(^{54}\). Such efforts, however, are complicated by the fact that very subtle changes to protein motion can alter SBP function and the structural basis for changes in these open-closed dynamics are poorly understood. Here we have expanded on this work, demonstrating that SBPs of this protein fold can access three relatively well-defined conformational substates, including a wide-open state for which the possible physiological role must still be determined. In view of the ease with which it is sampled in MD simulations and unambiguously observed by DEER experimentally in two of the native proteins we studied (AncCDT-3 and AncCDT-5), as well as the apo-state of the L-arginine binding protein from *T. maritima*\(^{38}\), it is unlikely to be an artefact. The apparently facile sampling of both closed and wide-open states could explain why FRET experiments, which have been conducted with fluorescent tags attached to either domain, sometimes report smaller changes in signal upon ligand binding than would be expected if a single open conformation was fully populated in the apo-protein\(^{55,56}\).

We previously suggested that mutations accumulated during evolution most likely increased the dehydratase activity of PaCDT by shifting the conformational equilibrium away from the catalytically-unproductive open state to the catalytically-competent closed state\(^{40}\). Crystal structures and MD simulations provided some preliminary evidence that this is the case, but now we have strong quantitative evidence in solution. First, the crystal structure of AncCDT-5 revealed identical active/binding sites of AncCDT-5 and PaCDT, demonstrating that amino acid substitutions remote from the active site can have very large effects on activity via mechanisms other than controlling the structure of the active site. Second, the dramatic
shift in the open/closed equilibrium of the proteins is the main change along the evolutionary trajectory. These results show that the wide-open state, which is the lowest-energy state in AncCDT-3 and AncCDT-5 and is unlikely to play any catalytic role owing to low affinity for substrate owing to the spatial separation between the binding site/catalytic residues, appears to be depopulated in the final evolutionary step between AncCDT-5 and PaCDT. Third, the extant and efficient PaCDT appears to exist in a broad distribution between a closed state optimized for catalysis of the chemical step in the reaction and a moderately open state that retains the shape of the binding pocket but opens a pathway for substrate/product diffusion. Structural analysis provides a molecular explanation for this shift in conformational sampling, with a number of structural changes stabilizing the open/closed states at the expense of the wide-open state. Interestingly, this seems to have occurred at least in part through oligomerization, suggesting that evolution can favour oligomerization both for stability and activity33,57.

The role of the wide-open state along the evolutionary trajectory from AncCDT-1 to PaCDT is more difficult to interpret. We observe this state in crystal structures of extant relatives of AncCDT-1, it is stable for several hundred ns in MD simulations of AncCDT-1 and (because of the overlap of the DEER signals between the wide-open and open states) it is also consistent with the DEER data. However, it is clear that there was a shift towards the wide-open state after AncCDT-1 (as seen in AncCDT-3 and AncCDT-5). In our previous work40, we were unable to identify a natural ligand for the non-enzymatic binding protein AncCDT-2 (which predates AncCDT-3 and AncCDT-5); although it had lost its ancestral arginine binding function, it (and its decedents such as Pu1068 from the aquatic bacterium Pelagibacter ubique) retain the genomic organization and biophysical properties of solute binding proteins. Thus, we speculate that the shift towards the wide-open state in the intermediate ancestors could have resulted from the selective pressure for the (still unknown) solute-binding function that preceded the evolution of CDT activity.

The findings from this work add to growing evidence that the modification of protein dynamics often plays an important role in the evolution of new protein function. While the current work provides a unique example for pronounced, historical conformational changes between a non-catalytic ancestral protein and an efficient enzyme, our findings mirror those of other work that highlight how catalytic activity can be enhanced along evolutionary trajectories through the accumulation of remote substitutions that alter conformational dynamics49,58–62. Modified protein dynamics are not mutually exclusive with the well-established concept of evolution of protein function via specific changes in shape or electrostatic character of
active/binding sites, but serve as a complementary mechanism that may often overlap with these other more easily observable changes.

In addition, this work expands the scope of the DEER approach\textsuperscript{19,63–67} for use in this context by using Gd(III) spin labels; it is the first time that the propargyl-DO3A-Gd(III)/DEER approach has been used to study rigid-body protein dynamics and it demonstrates the excellent performance of the propargyl-DO3A-Gd(III) tag. The tag is independent of cysteine residues as it can be ligated to AzF residues and the small size, lack of net charge and hydrophilic character of the DO3A-Gd(III) complex disfavours specific interactions with the protein surface, thus increasing the reliability of computational predictions of the Gd(III)–Gd(III) distance distributions obtained with this tag\textsuperscript{41}. This study highlights its efficiency (the DEER measurements requiring only nanogram amounts of protein) in studying the protein dynamics of a series of states along an evolutionary trajectory and how it afforded distribution widths that were sufficiently narrow to observe the co-existence of open and closed protein conformations that differ by only about 1 nm, to show the presence of two conformations simultaneously sampled in solution and to detect ligand-induced conformational changes. This approach will therefore be a welcome addition to the toolbox of methods that can be used to investigate rigid-body motions, such as those typical of SBPs.

Considering that DEER data are recorded of frozen solutions, it cannot be ruled out that conformational equilibria established at room temperature change during snap-freezing of the samples. Notably, however, the freezing point of 20 % deuterated glycerol solutions is only a little below 0 °C,\textsuperscript{68} suggesting that any changes in conformational equilibria due to freezing would be minor. It is thus significant that the energy landscape of the extant enzyme \textit{Pa}CDT led to the freezing out of a continuous range of conformations, whereas that of the amino-acid binding protein AncCDT-1 funnelled the protein into two clearly distinct conformations. Importantly, the distances measured by DEER here are in good agreement with distances predicted from the crystal structures and states obtained from MD simulations.

**Conclusions**

In this work, we studied the structural dynamics of a series of states along an evolutionary trajectory from an ancestral L-arginine solute binding protein (AncCDT-1) to an extant dehydratase (\textit{Pa}CDT). The challenge of accurately defining the conformational landscape of a protein is substantial: virtually all approaches have serious limitations or are so demanding that they are not practical for studying more than a single protein. Here, by following an integrated structural biology approach involving the combined use of the three methods, the shortcomings
of each approach (crystal packing forces in single crystals, imperfect force fields in MD simulations, potential overlap in the signals from different states for DEER measurements, etc.) could be compensated for. We compared crystal structures of AncCDT-3, a new structure of AncCDT-5, and PaCDT, which revealed that second-shell residues alter the active site configuration between AncCDT-3 and AncCDT-5, but that AncCDT-5 and PaCDT display large differences in catalytic activity despite identical active site configurations. Using a propargyl-DO3A-Gd(III) tag and W-band DEER distance measurements, we were able to experimentally assess the distribution of different conformational substates in frozen solutions along this evolutionary trajectory, with reference to structures obtained through protein crystallography and MD simulations. These data provide new insight into how proteins can evolve new functions: while the correct active site configuration must be established, this study shows that the ability of an enzyme to adopt a range of conformational states that are tailored to its catalytic role is equally important. This result contributes to the growing body of literature that suggests molecular evolution is a relatively smooth process, whereby conformational substates that are not productive are gradually depopulated while beneficial substates are enriched.

Methods

Protein numbering convention

Residues of all proteins are referred to using the equivalent position in the reconstructed ancestral sequences (i.e. not including N-terminal tags), as described in previous work\textsuperscript{40} (see sequence alignment, Supplementary Figure 8). As a consequence, residue numbers mentioned in this work differ from the residue numbers in the published PDB files of 3KBR (shift of -25), 5HPQ (-11), 6BQE (-11), 5T0W (-11), 5TUJ (-11), 5JOS (-11) and 6WUP (-11).

Materials

The genes encoding SeLAOBP (UniProt: P09551), AncCDT-1, AncCDT-3/P188, AncCDT-3/L188, AncCDT-5 and PaCDT (UniProt: Q01269, residues 26–268) in pDOTS7 vectors (a derivative of pQE-28L, Qiagen) were obtained from previous work\textsuperscript{32,40}. The ΔpheA strain of \textit{E. coli} K-12 from the Keio collection (strain JW2580-1) was obtained from the Coli Genetic Stock Center (Yale University, CT).
Prephenate dehydratase assay

For prephenate dehydratase assays, protein expression and purification and spectrophotometric assays were performed as described previously. Proteins were expressed in ΔpheA cells to exclude the possibility of contamination with endogenous prephenate dehydratase. Proteins were purified by immobilized metal affinity chromatography (IMAC) and size-exclusion chromatography (SEC), and prephenate dehydratase activity was measured by end-point enzyme assays monitoring formation of phenylpyruvate by absorbance at 320 nm.

Measurement of oligomeric state

IMAC-purified, His6-tagged AncCDT-1, AncCDT-3/P188, AncCDT-5 and PaCDT (expressed in ΔpheA cells) were transferred into size-exclusion chromatography-multi angle light scattering (SEC-MALS) buffer (20 mM Na2HPO4, 150 mM NaCl, pH 7.4) and concentrated using an Amicon centrifuge filter (10 kDa molecular weight cut-off, MWCO). Samples (100 µL) of each protein were loaded at 10 mg mL\(^{-1}\) onto a pre-equilibrated Superdex 200 10/300 GL size-exclusion column (GE Healthcare) attached to multi-angle light scattering (DAWN HELEOS 8; Wyatt Technologies) and refractive index detection (Optilab rEX; Wyatt Technologies) units. A flow rate of 0.5 mL min\(^{-1}\) was used. The multi-angle detectors were normalized using monomeric bovine serum albumin (Sigma, A1900). A \(dn/dc\) value of 0.186 g\(^{-1}\) was used for each sample. The data were processed using ASTRA (Wyatt Technologies). Data were collected from a single experiment (n = 1).

Protein sample preparation for DEER experiments

For DEER experiments, genes encoding each protein were cloned into pETMCSIII and expressed with an N-terminal His6-tag followed by a TEV cleavage site. In order to minimize tag side-chain dynamics, the crystal structures of the native proteins were inspected in PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) to select mutation sites where the sidechains of AzF residues were predicted to populate single \(\chi_1\)-angle conformations. Amber stop codons were introduced at these positions (Supplementary Table 3) by a modified QuikChange protocol using mutant T4 DNA polymerase.

All proteins for DEER experiments were expressed in RF1-free E. coli strain B-95.ΔA cells co-transformed with a plasmid for the aminoacyl-tRNA synthetase/tRNA pair for incorporation of AzF (Chem-Impex, USA). To minimize the amount of AzF required (provided at 1 mM), 1 L of cell-culture grown in LB medium was concentrated to 300 mL (by
centrifugation and resuspension) before induction with 1 mM IPTG. Expression was conducted at 37 °C and limited to 3 hours after IPTG induction to minimize the chemical reduction of AzF. Cells were harvested by centrifugation at 5,000 × g for 15 min and lysed by passing two times through a French Press (SLM Aminco, USA) at 830 bars. The lysate was centrifuged at 13,000 × g for 30 min and the filtered supernatant was loaded onto a 5 mL Ni-NTA HisTrap HP column (GE Healthcare, USA) equilibrated with binding buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 5 % w/v glycerol). The protein was eluted using elution buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 5 % w/v glycerol, 300 mM imidazole) and fractions were analyzed by 12 % SDS-PAGE.

To liberate and remove ligand molecules that may have bound during protein expression, the preparation of a second set of protein samples included a denaturation and on-column refolding step. After cell lysis, guanidinium hydrochloride was added to the supernatant to a final concentration of 6 M. The filtered solution was then loaded onto a Ni-NTA column and washed with binding buffer containing 8 M urea to remove any bound ligand. Refolding was achieved using a gradient of binding buffer with decreasing amounts of urea overnight at a flow rate of 0.5 mL min⁻¹. Refolded protein was eluted as described above.

His₆-tags were removed by digestion with His₆-tagged TEV protease in 1:100 molar ratio overnight at 4 °C in TEV cleavage buffer (50 mM Tris-HCl, pH 8.0, 300 mM NaCl and 1 mM β-mercaptoethanol). Finally, the cleaved His₆-tag and TEV protease were removed by passing through the Ni-NTA column (pre-equilibrated with binding buffer) prior to ligation with the propargyl-DO3A tag loaded with Gd(III).

Click reactions with the propargyl-DO3A-Gd(III) tag were performed overnight at room temperature as described previously (Supplementary Figure 3a) and samples were exchanged into EPR buffer (20 mM Tris-HCl, pH 7.5 in D₂O) using a Amicon Ultra-15 filter unit (10 kDa MWCO). The tagging yields were assessed by mass spectrometry, using an Elite Hybrid Ion Trap-Orbitrap mass spectrometer coupled with an UltiMate S4 3000 UHPLC (Thermo Scientific, USA).

DEER measurements were done on samples containing 100 µM protein and 20 % w/v glycerol-d₈ in D₂O. A 1.5-fold molar excess of L-arginine was added to one sample of refolded AncCDT-1 and one sample of SeLAOBP prior to the DEER measurement. Samples were flash-frozen by quick insertion of the sample holder into the cold (10 K) magnet bore.
DEER measurements

DEER measurements were conducted at 10 K on a home-built pulse EPR spectrometer operating at W-band (94.9 GHz). For the doubly labelled AncCDT-1 samples, a variant of the standard four-pulse DEER sequence \((\pi/2(\nu_{\text{obs}}) - \tau_1 - \pi(\nu_{\text{obs}}) - (\tau_1 + t) - \pi(\Delta \nu_{\text{pump}}) - (\tau_2 - t) - \pi(\nu_{\text{obs}}) - \tau_2 - \text{echo})\) was used. The DEER echo was observed at 94.9 GHz with \(\pi/2\) and \(\pi\) pulses of 15 ns and 30 ns, respectively, and the field was positioned at the peak of the Gd(III) spectrum. The pump pulse was replaced by two consecutive chirp pulses produced by an arbitrary waveform generator which were positioned on both sides of the centre of the Gd(III) spectrum with frequency ranges of 94.5-94.8 GHz and 95-95.3 GHz, respectively. The length of each chirp pulse was 96 ns. The cavity was tuned at 94.9 GHz with a maximal microwave amplitude, \(\omega_1/2\pi\), of approximately 30 MHz. Other parameters used were \(\tau_1 = 350\) ns, a repetition rate of 0.8 ms, and \(\tau_2\) typically ranging between 4 and 7 \(\mu s\) for samples with shorter and longer distances, respectively. The delay \(t\) was varied during the experiments. For the rest of the samples, a reverse DEER (rDEER) sequence \((\pi/2(\nu_{\text{obs}}) - \tau_1 - \pi(\nu_{\text{obs}}) - (\tau_1 - t) - \pi(\nu_{\text{pump}}) - (\tau_2 + t) - \pi(\nu_{\text{obs}}) - \tau_2 - \text{echo})\) was used to reduce artefacts. The chirp pump pulses and the positioning of the observe and the pump pulses were unchanged. In this implementation, both \(\tau_2\) and \(\tau_1\) typically ranged between 4 and 7 \(\mu s\) for shorter and longer distances, respectively.

The data were analyzed using DeerAnalysis and distance distributions were obtained using Tikhonov regularization. The regularization parameter was chosen either by the L-curve criterion or by visual inspection in order to obtain reasonable fits. Estimation of uncertainties in distance distributions due to background correction were obtained using the validation option in DeerAnalysis.

Crystallization of AncCDT-5 and structure determination

A solution of TEV-cleaved AncCDT-5 (expressed in BL21(DE3) cells and purified by IMAC as described above) was further purified by SEC using a HiLoad 16/600 Superdex 75 pg column (GE Healthcare, USA) in 50 mM Tris-HCl, pH 7.5, 20 mM NaCl, and was concentrated to 30 mg mL\(^{-1}\) using an Amicon Ultra-15 filter unit (10 kDa MWCO). The protein was crystallized using the vapor diffusion method at 18 °C. Crystallization screens were set up using the PACT premier HT-96 C10 screen from Molecular Dimensions (Newmarket, USA) with drops containing 200 nL reservoir solution and 200 nL AncCDT-5 solution. AncCDT-5 crystals formed in 28 days in 20 % w/v PEG 6000, 0.2 M MgCl\(_2\), 0.1 M HEPES, pH 7.0.
Crystals were cryoprotected (30% w/v PEG 6000, 0.2 M MgCl₂, 0.1 M HEPES pH 7.0) and frozen in liquid nitrogen. Data were collected at 100 K on the MX2 beamline of the Australian Synchrotron. The data were indexed and integrated in XDS⁷⁹ and scaled in AIMLESS⁸⁰. The structure was solved by molecular replacement in Phaser⁸¹ using the two domains of PaCDT as search models (PDB 3KBR small domain = residues 122-223; large domain = residues 27-121 and 224-258). The structure was refined by real space refinement in Coot⁸² and using REFMAC⁸³ and phenix.refine⁸⁴. Data collection and refinement statistics are given in Supplementary Table 1. Residues in the AncCDT-5 structure were numbered according to the equivalent positions in AncCDT-1 (PDB 5TUJ)⁴⁰ and the coordinates deposited in the Worldwide Protein Data Bank (PDB 6WUP).

Molecular dynamics (MD) simulations

500 ns simulations were performed in Desmond⁸⁵ (in Schrödinger 2019-11) using the OPLS3e force field⁸⁶. These simulations were initiated from the PaCDT–acetate trimer (PDB 5HPQ, residues 13-250 in PDB, n=1), unliganded AncCDT-1 (PDB 5T0W, chain A, residues 14-246, n=3), AncCDT-3/L188 (PDB 5JOS, residues 12-247, n=3), AncCDT-3/P188 (n=3) and AncCDT-5 (PDB 6WUP, residues 13-245, n=3) structures, with all small molecules removed. The AncCDT-3/P188 structure was obtained by making the L188P mutation and local minimization in Maestro (Schrödinger Release 2019-1: Maestro, Schrödinger, LLC, New York, NY, 2019). Desmond was used to add N-terminal acetyl caps and C-terminal amide caps to each structure and for energy minimization of the protein structures. Each protein was solvated in an orthorhombic box (15 Å buffer periodic boundary) with SPC water molecules. The smallest number of counter ions (Na⁺ or Cl⁻) needed to neutralize the net charge of each system were added. Energy minimization was achieved using a hybrid method of the steepest descent algorithm and the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (maximum of 2,000 iterations and a convergence threshold of 1 kcal mol⁻¹ Å⁻¹). The system was relaxed using the default relaxation procedure in Desmond. For production MD simulations of the NPT ensemble, the temperature was maintained at 300 K using a Nosé–Hoover thermostat (relaxation time = 1.0 ps) and the pressure was maintained at 1.01 bar (relaxation time = 2.0 ps) using a Martyna–Tobias–Klein barostat. Otherwise, default Desmond options were used. Following relaxation of the system, each simulation was run for 500 ns. Distances between the α-carbons of the tagged residues and the radii of gyration were calculated using the ProDy package on snapshots sampled every 0.1 ns⁸⁷. In addition,
representative snapshots (with a range of Cα–Cα distances) corresponding to the closed, open and wide-open states (if sampled) were selected for modelling of Gd(III)–Gd(III) distances. Allocation of states (closed/open/wide-open) was done based on Cα–Cα measurements. For AncCDT-1, conformations with Cα–Cα distances between ~3.3 and 3.7 nm could be separated into the open and wide-open conformations by principle component analysis. Plots of Cα-Cα distances over the course of each simulation are provided in Supplementary Figure 9.

Modelling Gd(III)–Gd(III) distances

To calculate predicted Gd(III)–Gd(III) distances of different conformations, propargyl-DO3A-Gd(III) tags were modelled onto the crystal structures of AncCDT-1 (PDB 5T0W, 5TUJ), AncCDT-3/L188 (PDB 5JOS), AncCDT-5 (PDB 6WUP), PaCDT (PDB 3KBR, 5HPQ) and selected MD snapshots. To do this, the mutation tool in PyMOL was used to identify the side-chain dihedral angles $c_1$ and $c_2$ of a phenylalanine residue that produced the least amount of steric clashes with the rest of the protein; these angles were used when modelling the tag. The angles $c_9$ and $c_{10}$ (Supplementary Figure 3b) were fixed to -140° and 70°, respectively, to allow coordination of the metal ion by the nearest nitrogen of the triazole ring. The angle $c_6$ was set to 180°. As in previous work with a closely related cyclen tag$^{51}$, setting $c_6$ to 180° (Supplementary Figure 3c) yielded a better correlation between experimental and back-calculated Gd(III)–Gd(III) distances compared with setting $c_6$ to 0° (Supplementary Figure 3d). Adding additional flexibility to the model by allowing rotation around $c_9$ (any angle) and $c_{10}$ (-60°, 60° or 180°) increased the width of modelled distance distributions but did not significantly change the positions of the peaks within these distributions, which still correlated well with experimental measurements (Supplementary Figure 3e) and the distances predicted using the single conformation approach mentioned above.
Acknowledgements

JAK acknowledges financial support from an Australian Government Research Training Program Scholarship. BEC acknowledges financial support from a Rod Rickards PhD Scholarship. Funding by the Australian Research Council, including a Laureate Fellowship to GO, is gratefully acknowledged. DG acknowledges the support of the Minerva foundation and this research was made possible in part by the historic generosity of the Harold Perlman Family (DG). DG holds the Erich Klieger Professorial Chair in Chemical Physics.

Conflict of Interest

The authors declare that they have no conflicts of interest with the contents of this article.

Data Availability

The atomic coordinates and structure factors for AncCDT-5 (PDB 6WUP) have been deposited in the Worldwide Protein Data Bank (http://wwpdb.org/). The source data underlying Supplementary Figure 1 are provided as a Source Data file. Any other relevant data are available from the authors.
References


39. Zhao, G., Xia, T., Fischer, R. & Jensen, R. Cyclohexadienyl dehydratase from


3.5 Supplementary Information for “Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme”
Supplementary Information

Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme

Joe A. Kaczmarski,a,§ Mithun C. Mahawaththa,a,§ Akiva Feintuch,b,§ Ben E. Clifton,a,c Luke A. Adams,d Daniella Goldfarb,b* Gottfried Otting,a* Colin J. Jacksona*

a Research School of Chemistry, The Australian National University, Canberra, ACT 2601, Australia.
b Department of Chemical Physics, Weizmann Institute of Science, Rehovot 76100, Israel.
c Present address: Protein Engineering and Evolution Unit, Okinawa Institute of Science and Technology, 1919-1 Tancha, Onna-son, Okinawa 904-0412, Japan.
d Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, VIC 3052, Australia.

§ These authors contributed equally.
* Corresponding authors. E-mails: daniella.goldfarb@weizmann.ac.il, gottfried.otting@anu.edu.au, colin.jackson@anu.edu.au
Supplementary Table 1. Data collection and refinement statistics for the AncCDT-5 X-ray crystal structure.

<table>
<thead>
<tr>
<th>PDB</th>
<th>6WUP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AncCDT-5 (HEPES)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>P 6₅ 2 2</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>70.95, 70.95, 175.39</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90, 90, 120</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>35.69–1.49 (1.52–1.49)*</td>
</tr>
<tr>
<td>R_{merge}</td>
<td>0.083 (3.91)</td>
</tr>
<tr>
<td>R_{free}</td>
<td>0.014 (0.729)</td>
</tr>
<tr>
<td>I/σI</td>
<td>22.4 (1.40)</td>
</tr>
<tr>
<td>CC₁/₂</td>
<td>1.00 (0.501)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>35.3 (30.3)</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>43.85–1.49 (1.52–1.49)</td>
</tr>
<tr>
<td>No. reflections</td>
<td>43560 (4265)</td>
</tr>
<tr>
<td>R_{work} / R_{free}</td>
<td>0.1975 (0.2972) / 0.2217 (0.3129)</td>
</tr>
<tr>
<td>No. atoms</td>
<td>2194</td>
</tr>
<tr>
<td>Protein</td>
<td>1882</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>16</td>
</tr>
<tr>
<td>Water</td>
<td>296</td>
</tr>
<tr>
<td>B-factors</td>
<td>32.5</td>
</tr>
<tr>
<td>Protein</td>
<td>31.0</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>28.9</td>
</tr>
<tr>
<td>Water</td>
<td>42.1</td>
</tr>
<tr>
<td>R.m.s. deviations</td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.007</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>0.89</td>
</tr>
</tbody>
</table>

* Statistics for the highest-resolution shell are shown in parentheses.
Supplementary Table 2. Cα–Cα distances (tag positions) and calculated Gd(III)–Gd(III) distances for crystal structures and representative MD snapshots.a

<table>
<thead>
<tr>
<th>Name of Crystal Structure/MD Snapshot</th>
<th>Cα–Cα distance (nm)</th>
<th>Calculated Gd(III)–Gd(III) distance (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Arg AncCDT-1 (PDB 5T0W)</td>
<td>2.61</td>
<td>2.94</td>
</tr>
<tr>
<td>MD_Anc1_closed_01</td>
<td>2.52</td>
<td>2.89</td>
</tr>
<tr>
<td>MD_Anc1_closed_02</td>
<td>2.63</td>
<td>2.88</td>
</tr>
<tr>
<td>MD_Anc1_closed_03</td>
<td>2.73</td>
<td>3.28</td>
</tr>
<tr>
<td>MD_Anc1_closed_04</td>
<td>2.82</td>
<td>3.18</td>
</tr>
<tr>
<td>MD_Anc1_closed_05</td>
<td>2.91</td>
<td>3.39</td>
</tr>
<tr>
<td>apo-AncCDT-1 (PDB 5TUJ)</td>
<td>3.36</td>
<td>4.06</td>
</tr>
<tr>
<td>MD_Anc1_open_01</td>
<td>3.19</td>
<td>3.87</td>
</tr>
<tr>
<td>MD_Anc1_open_02</td>
<td>3.40</td>
<td>4.90</td>
</tr>
<tr>
<td>MD_Anc1_open_03</td>
<td>3.53</td>
<td>4.19</td>
</tr>
<tr>
<td>MD_Anc1_open_04*</td>
<td>3.63</td>
<td>4.10</td>
</tr>
<tr>
<td>MD_Anc1_open_05</td>
<td>3.65</td>
<td>4.30</td>
</tr>
<tr>
<td>MD_Anc1_wide_01</td>
<td>3.79</td>
<td>4.53</td>
</tr>
<tr>
<td>MD_Anc1_wide_02</td>
<td>3.95</td>
<td>5.23</td>
</tr>
<tr>
<td>MD_Anc1_wide_03</td>
<td>3.97</td>
<td>4.68</td>
</tr>
<tr>
<td>MD_Anc1_wide_04</td>
<td>4.07</td>
<td>5.20</td>
</tr>
<tr>
<td>MD_Anc1_wide_05</td>
<td>4.17</td>
<td>4.99</td>
</tr>
<tr>
<td>AncCDT-3/L188 (PDB 5JOS)</td>
<td>4.21</td>
<td>5.25</td>
</tr>
<tr>
<td>MD_Anc3_L188_wide_01</td>
<td>4.04</td>
<td>5.03</td>
</tr>
<tr>
<td>MD_Anc3_L188_wide_02</td>
<td>4.07</td>
<td>5.42</td>
</tr>
<tr>
<td>MD_Anc3_L188_wide_03</td>
<td>4.26</td>
<td>5.54</td>
</tr>
<tr>
<td>MD_Anc3_L188_wide_04</td>
<td>4.46</td>
<td>5.85</td>
</tr>
<tr>
<td>MD_Anc3_L188_wide_05</td>
<td>4.67</td>
<td>6.24</td>
</tr>
<tr>
<td>MD_Anc3_P188_wide_01</td>
<td>4.10</td>
<td>5.27</td>
</tr>
<tr>
<td>MD_Anc3_P188_wide_02</td>
<td>4.27</td>
<td>5.57</td>
</tr>
<tr>
<td>MD_Anc3_P188_wide_03</td>
<td>4.56</td>
<td>5.73</td>
</tr>
<tr>
<td>MD_Anc3_P188_wide_04</td>
<td>4.68</td>
<td>6.22</td>
</tr>
<tr>
<td>MD_Anc3_P188_wide_05</td>
<td>4.78</td>
<td>6.34</td>
</tr>
<tr>
<td>apo-TmArgBP (PDB 4PRS)</td>
<td></td>
<td>4.00</td>
</tr>
</tbody>
</table>

a Gd(III)–Gd(III) distances were calculated following modelling of the AzF-propargyl-DO3A-Gd(III) residue at each mutation site on structures obtained from crystallography (bold) or MD simulations. The models were built by identifying the most favourable values of the dihedral angles $\chi_1$ and $\chi_2$ of the tag (using the mutation tool of PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LCC.) and setting $\chi_6$ to 180°1. For the MD snapshots, structures with
varying Cα–Cα distances (“MD.....01-05”) were selected as representative structures from each conformational state (i.e. closed, open, wide-open). The data points representing these structures are shown as colored crosses in Figure 4.

*This snapshot was initially grouped as being open (based on the Cα–Cα distance), but was allocated into the wide-open group based on PCA analysis and radius of gyration measurements.
Supplementary Table 3. Sites mutated for the incorporation of AzF residues.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Mutation sitesa</th>
</tr>
</thead>
<tbody>
<tr>
<td>AncCDT-1</td>
<td>Q68/K138; Q68/Q219; K138/S161</td>
</tr>
<tr>
<td>AncCDT-3</td>
<td>Q68/R138</td>
</tr>
<tr>
<td>AncCDT-5</td>
<td>Q68/R138</td>
</tr>
<tr>
<td>PaCDTb</td>
<td>A68/A139</td>
</tr>
<tr>
<td>SeLAOBP</td>
<td>A61/T151</td>
</tr>
</tbody>
</table>

a Amber stop codons were introduced to incorporate pairs of AzF residues. The positions in the amino acid sequence are indicated together with the original amino acid types.

b The mutant A68/R138 was also prepared, but the tagging yields obtained were too low for DEER experiments.
Supplementary Figure 1. Michaelis-Menten plot for AncCDT-5 prephenate dehydratase activity. Data points from three technical replicates are shown (n = 3). Source data are provided as a Source Data file.
**Supplementary Figure 2. SEC-MALS of proteins.** Results from analytical size-exclusion chromatography coupled with multiangle light scattering (SEC-MALS) for (A) AncCDT-1, (B) AncCDT-3/P188, (C) AncCDT-5 and (D) PaCDT.
Supplementary Figure 3. The AzF-propargyl-DO3A-Gd(III) residue and correlations between modelled and experimental Gd(III)–Gd(III) distances. (A) Scheme of the Cu(I)-catalyzed cycloaddition reaction used to attach the propargyl-DO3A-Gd(III) tag to the AzF residue\(^4\). (B) Dihedral angles in the AzF-propargyl-DO3A-Gd(III) residue. \(\chi_6 = 0^\circ\) in the conformation shown. (C) Correlation between modelled and experimental Gd(III)–Gd(III) distances, using the crystal structures of AncCDT-1 (blue; PDB ID 5T0W: mutants 68/138, 68/219, 138/161; PDB ID 5TUJ: mutant 68/138), AncCDT-3/P188 (green; PDB ID 5JOS: mutant 68/138) and PaCDT (orange; PDB ID 3KBR: single mutants 68 and 139, and double mutant 68/139). The experimental distances correspond to the maxima of the respective DEER distance distributions. Distances were predicted using \(\chi_6 = 0^\circ\). The angles \(\chi_9\) and \(\chi_{10}\) were fixed to -140° and
70°, respectively, to allow coordination of the metal ion (cyan sphere in B) by the nearest nitrogen of the triazole ring. \( \chi_1 \) and \( \chi_2 \) angles were selected by using those conformations, which were identified by the mutation tool of the program PyMOL to generate minimal steric clashes. (D) Same as (C), but using \( \chi_6 = 180° \). This conformation was used to represent the modelled distances in all other figures of the present work. (E) Same as (D), but \( \chi_9 \) was allowed to be random while the \( \chi_{10} \) angle was randomly set to -60°, 60°, or 180° with equal likelihood. Although the variation of these dihedral angles broadened the modelled distance distributions, the peaks of the Gd(III)–Gd(III) distance distributions changed only a little.
Supplementary Figure 4. Additional and primary DEER data for AncCDT-1. Data from DEER measurements on natively purified (black) and refolded (red) AncCDT-1 samples (100 μM protein in D$_2$O and 20 % w/v deuterated glycerol) tagged at positions (A) 68 and 138 in the absence of exogenous L-Arg, (B) 68 and 138 in the presence of 1.5-fold molar excess L-Arg, (C) 68 and 219, or (D) 138 and 161. The left panel (i) shows the primary DEER data for the sample and background (gray) decay function. The central panel (ii) shows the DEER form factors after
background correction, with fits corresponding to the distance distributions shown in (iii). The right panel (iii) shows the corresponding distance distributions with color coding of the reliability regions as defined in DeerAnalysis. (Pale green: the shape of the distance distribution is reliable. Pale yellow: the mean distance and distribution width are reliable. Pale orange: the mean distance is reliable. Pale red: long-range distance contributions may be detectable, but cannot be quantified.) The striped regions, representing the uncertainty range of the data, show alternative distributions obtained by varying the parameters of the background correction as calculated by the validation tool in the DeerAnalysis software package[^5]. The parameter ranges used for the validations are: white noise 0–1.5, background start 0.2*t_{max}–0.6*t_{max}, and background dimension 3–3.6. The limited variation in background dimension was allowed, as a non-homogeneous background improved the fit for some of the traces. Vertical lines represent the predicted Gd(III)–Gd(III) distances when AzF-propargyl-DO3A-Gd(III) residues are modelled on crystal structures of (E) closed AncCDT-1 (PDB 5T0W) at positions 68 and 138, (F) open AncCDT-1 (PDB 5TUJ) at positions 68 and 138, (G) closed AncCDT-1 at positions 68 and 219, or (H) closed AncCDT-1 at positions 138 and 161.

[^5]: DeerAnalysis software package.
Supplementary Figure 5. DEER data for SeLAOBP samples. (A) Comparison of Gd(III)–Gd(III) distance distributions of refolded SeLAOBP tagged at positions 61 and 151, in the absence (red) and presence (black) of a 1.5 molar excess of L-arginine. (B-C) Raw and processed data from DEER measurements of natively purified (black) and refolded (red) SeLAOBP samples in the (B) absence and (C) presence of a 1.5 molar excess of L-arginine. Samples were measured using 100 μM protein in D$_2$O and 20 % w/v deuterated glycerol. The left panel shows the primary DEER data for the sample and background (gray) decay function. The central panel shows the DEER form factors after background correction, with fits corresponding to the distance distributions shown in the right-hand side panel. The right panel shows the corresponding distance distributions with color coding of the reliability regions as defined in DeerAnalysis and striped regions showing alternative distributions, as described in the caption of Supplementary Figure 4. The validation parameters were as described in Supplementary Figure 4.
Supplementary Figure 6. Primary DEER data for AncCDT-3/P188 and AncCDT-5 samples. Primary data from the DEER measurements (as shown in Figure 3) on natively purified (black) and refolded (red) of (A) AncCDT-3/P188 tagged at positions 68 and 138, and (B) AncCDT-5 tagged at positions 68 and 138. Samples were measured using 100 μM protein in D₂O and 20 % w/v deuterated glycerol. The left panel (i) shows the primary DEER data for the sample and background (gray) decay function. The central panel (ii) shows the DEER form factors after background correction, with fits corresponding to the distance distributions shown in (iii). The right panel (iii) shows the corresponding distance distributions with color coding of the reliability regions as defined in DeerAnalysis and striped regions showing alternative distributions as
described in the caption of Supplementary Figure 4. The validation parameters were as described in Supplementary Figure 4. In (B), the primary DEER traces were shortened in the analysis to avoid an instrumental artefact. Vertical lines represent the predicted Gd(III)–Gd(III) distances when AzF-propargyl-DO3A-Gd(III) residues are modelled onto crystal structures of (C) AncCDT-3/L188 (PDB 5JOS) at positions 68 and 138, or (D) AncCDT-5 (PDB 6WUP) at positions 68 and 138.
Supplementary Figure 7. Additional and primary DEER data for PaCDT samples. Data from the DEER measurements on natively purified samples of PaCDT tagged at (A) position 68, (B) position 139, (C) position 68 and 139, or (D) position 68 and 139, but following a 10-fold dilution with unlabelled PaCDT for selective observation of the intramolecular distance between the two domains (as shown in Figure 3). Samples were measured using 100 μM protein in D₂O and 20 % w/v deuterated glycerol. The left panel (i) shows the primary DEER data for the sample and background (gray) decay function. The central panel (ii) shows the DEER form factors after background correction, with fits corresponding to the distance distributions shown in (iii). The right panel (iii) shows the corresponding distance distributions with color coding of the reliability regions as defined in DeerAnalysis and striped regions showing alternative distributions as described in the caption of Supplementary Figure 4. The validation parameters were as described in Supplementary Figure 4 except for panel (D) (also shown in Figure 4D), in which the background dimension was varied in the range of 3-4 to take into account the D=4 we used to get a good fit. In (B), the primary DEER traces were shortened in the analysis to avoid an instrumental artefact. Vertical lines represent the predicted Gd(III)–Gd(III) distances when AzF-propargyl-
DO3A-Gd(III) residues are modelled onto the crystal of structure PaCDT (PDB 3KBR) at tagged positions, as shown in panel (iv).
Supplementary Figure 8. Multiple sequence alignment. Multiple sequence alignment of the proteins studied in this work. N-terminal residues that were introduced during cloning, including the His<sub>6</sub>-tag, have been omitted. The residue numbering corresponds to that used throughout this work. The sequences were aligned using Clustal Omega<sup>2</sup>. The figure was generated using ESPript<sup>3</sup>. Every 10<sup>th</sup> residue position in the alignment is indicated by a black dot. Shading is based on the percentage of equivalent residues at that position (i.e. black = 100 % conserved).
Supplementary Figure 9. Time evolution of Ca–Ca distances in MD simulations. The time evolution of distances between the alpha carbons of tagging sites in simulations of (A) AncCDT-1, (B) AncCDT-3/P188, (C) AncCDT-3/L188, (D) AncCDT-5 and (E) PaCDT. For AncCDT-1, AncCDT-3 and AncCDT-5, individual replicates (n=3) are shown as different colored traces. For the PaCDT plot, the Ca–Ca distances between the tagged sites of each monomer of the PaCDT trimer (n=1) are shown as different colored traces.
References


Chapter 4. Interactions between neutralising antibodies and the circumsporozoite protein from *Plasmodium falciparum*. 
Interactions between neutralising antibodies and the circumsporozoite protein from *Plasmodium falciparum*

**Key Ideas:** Antibody-antigen recognition, vaccine design, polyvalent binding modes, avidity.

**Techniques:** X-ray crystallography, ITC, SPR.

### 4.1 Preface

Macromolecular recognition is central to the function of the immune system. In particular, the adaptive immune system relies on the production of pathogen-targeting immunoglobulin (Ig) receptors and antibodies to detect, neutralise and protect animals against invading pathogens that evade the innate immune system. High-affinity, antigen-specific antibodies are generated by the immune system through an iterative process of genetic diversification and selection that is somewhat analogous to the production of high-affinity binding proteins via evolution by natural selection. However, rather than occurring over several generations of the host (as is the case for protein germline evolution), the somatic hypermutation (SHM) and affinity maturation of antibodies can (and must) generate high-affinity complementary binding surfaces in a matter of days. Because this process rapidly produces such specialised binders, studying antibody-antigen binding is an excellent way to understand the rules that govern macromolecular recognition. Further, understanding the molecular processes and structural mechanisms by which antibodies adapt and detect pathogens is central for understanding how some pathogens, such as the malaria-causing *Plasmodium falciparum*, are able to evade the adaptive immune system, and drives the rational design of new, effective vaccines against these prevailing diseases (Cockburn and Seder, 2018).

For the following work, I performed ITC experiments that helped to determine the unique mode of binding of a neutralising murine monoclonal antibody, mAb2A10, to the circumsporozoite surface protein (CSP) of *P. falciparum*. The immunodominant central repeat region of PfCSP (PfCSP\(_{\text{repeat}}\)) to which mAb2A10 binds is a non-classical antigen in the sense that it is thought to be mostly disordered in solution (Dyson et al., 1990; Fisher et al., 2017; Ghasparian et al., 2006; Gibson and Scheraga, 1986; Patra et al., 2017) and is made up of
4-amino acid repeat units; for example, the repeat region of CSP from the *P. falciparum* 3D7 strain is made up of one Asn-Pro-Asp-Pro (NPDP) unit (i.e. the “junction motif”), followed by three sets of alternating Asn-Ala-Asn-Pro (NANP) and Asn-Val-Asp-Pro (NVDP) units, followed by a region that contains 35 additional NANP units and a single, central NVDP. Our ITC data revealed that 2A10 bind PfCSP in a polyvalent manner, with several antibodies fitting around PfCSP_repeat. Comparison of the CSP-binding properties of 2A10 and 2A10 antigen-binding fragments (Fabs), as well as 2A10Fab binding to shorter NANP repeats, highlights how avidity effects contribute to the formation of a tight complex (nanomolar-range avidity) between this antibody and the CSP molecule. ITC data are supported by an X-ray crystal structure of the 2A10Fab and computational modelling studies. Further, we recognise that this unique mode of binding allows for cross-linking of B cell receptors and investigate the effect this has on the T cell and B cell response. Together, our results suggest that the unique structure of PfCSP_repeat may limit the protective properties of the immune response. We discuss how these findings could guide efforts to improve the malaria vaccine.

In the second piece of work, we explore the effect that truncating the PfCSP_repeat region has on the immunogenicity of CSP. I show, using surface plasmon resonance (SPR), that CSP9 has fewer binding sites than CSP27, which likely contributes to its increased ability to protect against malaria challenge in a mouse model. The implications for vaccine design are discussed further in Chapter 7.
4.2 Paper 1: T-dependent B cell responses to *Plasmodium*

...induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein

4.2.1 Publication status

This manuscript presented in this section has been published by *PLOS Pathogens*. Supporting information is provided in Section 4.3.

4.2.2 Author’s contribution

I designed and performed ITC experiments for this paper. I prepared protein samples used in these experiments and analyzed and interpreted the ITC data. I contributed to the writing and editing of the manuscript and produced Table 1 and Figure 1.

4.2.3 Contributions from others

The majority of the work presented in this paper was performed by my co-authors, primarily from the Jackson (ANU Research School of Chemistry) and Cockburn (ANU John Curtin School of Medical Research) research groups.
T-dependent B cell responses to *Plasmodium* induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein

Camilla R. Fisher¹, Henry J. Sutton², Joe A. Kaczmarski¹, Hayley A. McNamara², Ben Clifton¹, Joshua Mitchell¹, Yeping Cai², Johanna N. Dups², Nicholas J. D’Arcy², Mandeep Singh², Aaron Chuah², Thomas S. Peat³, Colin J. Jackson¹*, Ian A. Cockburn²*

¹ Research School of Chemistry, The Australian National University, Canberra, Australian Capital Territory, Australia, ² John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory, Australia, ³ CSIRO Biomedical Manufacturing Program, Parkville, Victoria, Australia

* These authors contributed equally to this work.

* colin.jackson@anu.edu.au (CJJ); ian.cockburn@anu.edu.au (IAC)

**OPEN ACCESS**


**Editor:** Matthew K. Higgins, University of Oxford, UNITED KINGDOM

**Received:** February 8, 2017

**Accepted:** June 13, 2017

**Published:** July 31, 2017

**Copyright:** © 2017 Fisher et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Sequence data generated in this paper is deposited at the NCBI BioProject database accession number PRJNA352758. Atomic coordinates and related experimental data for structural analyses are deposited in the Protein Data Bank (PDB) with PDB codes 5SZF and 5TOY.

**Funding:** This work was supported by the Bill and Melinda Gates foundation [http://www.gatesfoundation.org (OPP1151018)]. The funders

**Abstract**

The repeat region of the *Plasmodium falciparum* circumsporozoite protein (CSP) is a major vaccine antigen because it can be targeted by parasite neutralizing antibodies; however, little is known about this interaction. We used isothermal titration calorimetry, X-ray crystallography and mutagenesis-validated modeling to analyze the binding of a murine neutralizing antibody to *Plasmodium falciparum* CSP. Strikingly, we found that the repeat region of CSP is bound by multiple antibodies. This repeating pattern allows multiple weak interactions of single F_{\text{AB}} domains to accumulate and yield a complex with a dissociation constant in the low nM range. Because the CSP protein can potentially cross-link multiple B cell receptors (BCRs) we hypothesized that the B cell response might be T cell independent. However, while there was a modest response in mice deficient in T cell help, the bulk of the response was T cell dependent. By sequencing the BCRs of CSP-repeat specific B cells in inbred mice we found that these cells underwent somatic hypermutation and affinity maturation indicative of a T-dependent response. Last, we found that the BCR repertoire of responding B cells was limited suggesting that the structural simplicity of the repeat may limit the breadth of the immune response.

**Author summary**

Vaccines aim to protect by inducing the immune system to make molecules called antibodies that can recognize molecules on the surface of invading pathogens. In the case of malaria, our most advanced vaccine candidates aim to promote the production of antibodies that recognize the circumsporozoite protein (CSP) molecule on the surface of the invasive parasite stage called the sporozoite. In this report we use X-ray crystallography to
determine the structure of CSP-binding antibodies at the atomic level. We use other tech-
niques such as isothermal titration calorimetry and structural modeling to examine how
this antibody interacts with the CSP molecule. Strikingly, we found that each CSP mole-
cule could bind 6 antibodies. This finding has implications for the immune response and
may explain why high titers of antibody are needed for protection. Moreover, because the
structure of the CSP repeat is quite simple we determined that the number of different
kinds of antibodies that could bind this molecule are quite small. However a high avidity
interaction between those antibodies and CSP can result from a process called affinity
maturation that allows the body to learn how to make improved antibodies specific for
pathogen molecules. These data show that while it is challenging for the immune system
to recognize and neutralize CSP, it should be possible to generate viable vaccines targeting
this molecule.

Introduction
Malaria caused by *Plasmodium falciparum* causes the deaths of around 430,000 people each
year [1]. The most advanced vaccine candidate for malaria is the RTS,S/AS01 vaccine which
consists of a truncated version of the sporozoite-surface circumsporozoite protein (CSP),
packaged in a Hepatitis C core virus-like particle delivered in AS01—a proprietary liposome
based adjuvant [2]. Phase II and Phase III clinical trials have repeatedly demonstrated that the
g vaccine gives around 50% protection against clinical malaria in field settings for the first year
following vaccination [3]. The bulk of protection is attributed to antibodies targeting the CSP
repeat epitope included within the vaccine, with some contribution from CD4+ T cells [4]. It
is still unclear why the antibody response to CSP is only partially protective. We lack structural
information about how neutralizing antibodies bind to CSP and knowledge on the breadth
and nature of the B cell response elicited.

Antibodies to CSP were first identified as potential mediators of protection following semi-
nal studies that showed that immunization with irradiated sporozoites could induce sterile pro-
tection against live parasite challenge [5,6]. In the early 1980s, monoclonal antibodies (mAbs)
isolated from mice immunized with sporozoites were found to be capable of blocking invasion
of hepatocytes [7] and directly neutralizing parasites by precipitating the surface protein coat (a
process known as the circumsporozoite reaction) [8]. These antibodies were then used to clone
CSP, one of the first malaria antigens identified [8,9]. The N- and C-terminal domains of CSP
from all *Plasmodium* species are separated by a repeat region, which was the target of the origin-
al mAbs [9—11]. In the 3D7 reference strain of *P. falciparum*, the CSP repeat has 38 repeats: 34
asparagine-alanine-asparagine-proline (NANP)-repeats interspersed with 4 asparagine-valine-
aspartate-proline (NVDP) repeats that are concentrated towards the N-terminus [12] though
different isolates can contain slightly different numbers of repeats [13]. One of the most effec-
tive *P. falciparum* sporozoite neutralizing antibodies identified in these early studies was 2A10
which can block sporozoite infectivity *in vitro* [7] and *in vivo* mouse models utilizing rodent
*P. berghei* parasites expressing the *P. falciparum* CSP repeat region [14,15].

While CSP binding antibodies have been shown to be able to neutralize sporozoites and
block infection, it has also been proposed that CSP is an immunological “decoy” that induces a
suboptimal, but broad, T-independent immune response due to the CSP repeat cross-linking
multiple B cell receptors (BCRs) [16]. However, it remains unknown if the repetitive regions
of CSP can cross-link multiple BCRs as they are not as large as typical type- II T-independent
antigens [17]. Moreover, the ability to induce a T-independent response does not preclude a
T-dependent component to immunity as well: various oligomeric viral surface proteins can induce both short-lived T-independent responses and subsequent affinity matured IgG responses [18,19]. Furthermore, the very little published data on the sequences of CSP binding antibodies does not convincingly support activation of a broad B cell repertoire: a small study of five *P. falciparum* CSP mouse monoclonal antibodies (mAbs) identified some shared sequences [20]. In humans, a study that generated mAbs from three individuals who received RTS,S found that the three antibodies studied had distinct sequences though these all used similar heavy chains [21].

We therefore set out to test the hypothesis that the CSP repeat can bind multiple antibodies or BCRs and drive a T-independent immune response. To do this we undertook a comprehensive biophysical characterization of the 2A10 sporozoite-neutralizing antibody that binds to the CSP repeat. We found that this antibody binds with an avidity in the nano-molar range which was unexpected as previous studies using competition ELISAs with peptides predicted a micro-molar affinity [22,23]. Strikingly, isothermal titration calorimetry (ITC), structural analyses, and mutagenesis-validated modeling revealed that the CSP repeat can be bound by around six antibodies suggesting that the repeat may potentially crosslink multiple BCRs on the surface of a B cell. However, analysis of CSP-specific B cells revealed that CSP-specific B cells can enter Germinal Centers (GCs) and undergo affinity maturation contradicting the notion that the response to CSP is largely T-independent. Moreover, we found that the BCR repertoire of CSP-binding B cells is quite limited which may restrict the size and effectiveness of the immune response.

Results

Characterization of the thermodynamics of 2A10-antigen binding

We began our analysis by performing isothermal titration calorimetry (ITC) to understand the interaction between 2A10 and CSP. For ease of expression we used a recombinant CSP (rCSP) construct described previously which was slightly truncated with 27 repeats compared to 38 in the 3D7 reference strain [12,24]. ITC experiments were run on the purified 2A10 antibody and the purified 2A10 antigen-binding fragment (F_AB) fragment to test the thermodynamic basis of the affinity of 2A10 F_AB towards CSP. Experiments were also performed on the 2A10 F_AB fragment with the synthetic peptide antigen (NANP)6, which is a short segment of the antigenic NANP-repeat region of CSP (Table 1; Fig 1). The binding free energies (ΔG) and dissociation constants (K_D) were found to be -49.0 kJ/mol and 2.7 nM for the full 2A10 antibody with CSP, -40 kJ/mol and 94 nM for the 2A10F_AB with CSP, and -36.4 kJ/mol and 420 nM for the 2A10 F_AB with the (NANP)6 peptide.

Surprisingly, we did not observe a typical 1:1 antibody/F_AB domain:antigen binding stoichiometry (Table 1). We found that each (NANP)6 peptide was bound to by ~2 F_AB fragments (2.8 repeats per F_AB domain). With the rCSP protein we observed that ~11 F_AB fragments could bind to each rCSP molecule, (2.5 repeats per F_AB domain. Finally, when the single-domain F_AB fragment is replaced by the full 2A10 antibody (which has two F_AB domains), we observe binding of 5.8 antibodies per rCSP molecule (4.7 repeats per antibody). Therefore all complexes exhibit approximately the same binding stoichiometry of two F_AB fragments/domains per ~5 repeat units. These results suggest that the antigenic region of CSP constitutes a multivalent antigen and that repeating, essentially identical, epitopes must be available for the binding of multiple F_AB domains.

It is not possible to separate affinity from avidity in this system, although it is apparent that there is a substantial benefit to the overall strength of binding between the antibody and antigen through the binding of multiple F_AB domains. The F_AB:rCSP complex and the 2A10:rCSP
complex had similar enthalpy and entropy of binding (Table 1), but the 2A10:rCSP complex had a lower overall ΔG binding, corresponding to a lower dissociation constant (2.7 nM vs. 94 nM for FAB:rCSP). The observation that this antibody-antigen (Ab-Ag) interaction is primarily enthalpically driven is consistent with the general mechanism of Ab-Ag interactions [25]. It is clear that the dissociation constant ($K_d$) of a single FAB domain to the (NANP)$_6$ peptide is substantially higher (420 nM), and that the avidity, the accumulated strength of the multiple binding events between the FAB domains of the antibody and the CSP repeat, is the basis for the lower $K_d$ value observed in the 2A10:rCSP complex. Thus, the characteristic repeating pattern of the epitope on the CSP antigen allows multiple weak interactions with 2A10 FAB domains to accumulate, which yields a complex with a high avidity dissociation constant in the low nM range.

**Table 1. Thermodynamic parameters for interactions between 2A10 FAB, 2A10 and antigens.**

<table>
<thead>
<tr>
<th></th>
<th>(NANP)$_6$:FAB</th>
<th>rCSP:FAB</th>
<th>rCSP:2A10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_a$ (M$^{-1}$)</td>
<td>$(2.37 \pm 0.91) \times 10^6$</td>
<td>$(1.07 \pm 0.39) \times 10^7$</td>
<td>$(3.6 \pm 2.7) \times 10^8$</td>
</tr>
<tr>
<td>$K_d$ (nM)</td>
<td>$420 \pm 160$</td>
<td>$94 \pm 34$</td>
<td>$2.7 \pm 2.1$</td>
</tr>
<tr>
<td>ΔH (kJ/mol complex)</td>
<td>$-113 \pm 5$</td>
<td>$-1245 \pm 112$</td>
<td>$-1175 \pm 44$</td>
</tr>
<tr>
<td>TΔS (kJ/mol complex)</td>
<td>$-76.6 \pm 4.9$</td>
<td>$-1205 \pm 112$</td>
<td>$-1126 \pm 44$</td>
</tr>
<tr>
<td>ΔG (kJ/mol complex)</td>
<td>$-36.4 \pm 1.0$</td>
<td>$-40.0 \pm 0.9$</td>
<td>$-49.0 \pm 1.9$</td>
</tr>
<tr>
<td>$n$ (FAB/2A10: Ag)</td>
<td>$2.16 \pm 0.06$</td>
<td>$10.8 \pm 0.7$</td>
<td>$5.8 \pm 0.1$</td>
</tr>
</tbody>
</table>

Parameters were determined by ITC at 25°C. Errors for $n$ (Ag: FAB), $K_a$ and ΔH (complex) are 95% confidence intervals estimated from a single titration; errors for other parameters were propagated.

https://doi.org/10.1371/journal.ppat.1006469.t001

Structural analysis of the (NANP)-repeat region and the 2A10 FAB

To better understand the molecular basis of the multivalent interaction between 2A10 and rCSP, we performed structural analysis of the components. Previous work indicated that the

![Fig 1. ITC data for interactions between 2A10 FAB and antigens.](https://doi.org/10.1371/journal.ppat.1006469.g001)
NANP-repeat region of CSP adopts a flexible rod-like structure with a regular repeating helical motif that provides significant separation between the N-terminal and the C-terminal domains [26]. Here, we performed far-UV circular dichroism (CD) spectroscopy to investigate the structure of the (NANP)₆ peptide. These results were inconsistent with a disordered random coil structure (S1 Fig). Rather, the absorption maximum around 185 nm, minimum around 202 nm and shoulder between 215 and 240 nm, is characteristic of intrinsically disordered proteins that can adopt a spectrum of states [27].

The lowest energy structures of the (NANP)₆ repeat were predicted using the PEP-FOLD de novo peptide structure prediction algorithm [28]. The only extended state among the lowest energy structures that was consistent with the reported spacing of the N- and C-terminal domains of CSP [26], and which presented multiple structurally similar epitopes was a linear, quasi-helical structure, which formed a regularly repeating arrangement of proline turns (Fig 2A). The theoretical CD spectrum of this conformation was calculated (S1 Fig), qualitatively matching the experimental spectra: the maximum was at 188 nm, the minimum at 203 nm and there was a broad shoulder between 215 and 240 nm. To investigate the stability of this conformation, we performed a molecular dynamics (MD) simulation on this peptide, which showed that this helical structure could unfold, and refold, on timescales of tens of nanoseconds, supporting the idea that it is a low-energy, frequently sampled, configuration in solution (S1 Movie, S2 Fig). We also observed the same characteristic hydrogen bonds between a carbonyl following the proline and the amide nitrogen of the alanine, and the carbonyl group of an asparagine and a backbone amide of asparagine three residues earlier, that are observed in the crystal structure of the NPNA fragment [29]. Thus, this configuration, which is consistent with previously published experimental data, is a regular, repeating, extended conformation that would allow binding of multiple F_AB domains to several structurally similar epitopes.

To better understand the interaction between the 2A10 and the (NANP)-repeat region, we solved the crystal structure of the 2A10 F_AB fragment in two conditions (S1 Table), yielding structures that diffracted to 2.5 Å and 3.0 Å. All of the polypeptide chains were modeled in good quality electron density maps (Fig 2B), except for residues 134–137 of the light chain. This loop is located at the opposite end of the F_AB fragment to the variable region and not directly relevant to antigen binding. The 2.5 Å structure contained a single polypeptide in the asymmetric unit, whereas the 3.0 Å structure contained three essentially identical chains. Superposition of the four unique F_AB fragments from the two structures revealed that the variable antigen binding region is structurally homogeneous, suggesting that this region might be
relatively pre-organized in the 2A10 F<sub>AB</sub>. This is consistent with the observation that antibodies typically undergo relatively limited conformational change upon epitope binding [25]. Indeed, a recent survey of 49 Ab-Ag complexes revealed that within the binding site, the heavy chain Complementarity Determining Region (CDR)-3 was the only element that showed significant conformational change upon antigen binding and even this was only observed in one third of the antibodies [30].

Attempts to obtain a crystal structure of a complex between 2A10 F<sub>AB</sub> and the (NANP)<sub>6</sub> peptide were unsuccessful; unlike binary Ab-Ag interactions, in which the Ab will bind to a single epitope on an antigen and produce a population of structurally homogeneous complexes that can be crystallized, in this interaction we are dealing with an intrinsically-disordered peptide, the presence of multiple binding sites on the peptide, and the possibility that more than one 2A10 F<sub>AB</sub> domain can bind the peptide. Therefore it is difficult to obtain a homogeneous population of complexes, which is a prerequisite for crystallization. Attempts to soak the (NANP)<sub>6</sub> peptide into the high-solvent form of 2A10 F<sub>AB</sub>, in which there were no crystal packing interactions with the binding-loops, caused the crystals to dissolve, again suggesting that the heterogeneity of the peptide and the presence of multiple epitopes produces disorder that is incompatible with crystal formation.

Modeling the interaction of the 2A10 F<sub>AB</sub> with the NANP-repeat region and testing the model through site-directed mutagenesis

Although it was not possible to obtain a crystal structure of the 2A10-(NANP)<sub>6</sub> peptide complex, the accurate structures of the 2A10 F<sub>AB</sub> fragment, the (NANP)<sub>6</sub> peptide, and the knowledge that antibodies seldom undergo significant conformational changes upon antigen binding [30], allowed us to model the interaction, which we tested using site-directed mutagenesis. Computational modeling of Ab-Ag interactions has advanced considerably in recent years and several examples of complexes with close to atomic accuracy have been reported in the literature [31]. Using the SnugDock protein-protein docking algorithm [31], we obtained an initial model for binding of the peptide to the CDR region of the 2A10 F<sub>AB</sub> fragment (Fig 2C). We then performed, in triplicate, three 50 ns MD simulations on this complex to investigate whether the interaction was stable over such a time period (S2 Movie, S3 Fig). These simulations confirmed that the binding mode that was modeled is stable, suggesting that it is a reasonable approximation of the interaction between these molecules. To experimentally verify whether our model of the 2A10 F<sub>AB</sub>-(NANP)<sub>6</sub> peptide interaction was plausible, we performed site-directed mutagenesis of residues predicted to be important for binding. Our model predicted that the interaction with (NANP)<sub>6</sub> would be mainly between CDR2 and CDR3 of the light chain and CDR2 and CDR3 of the heavy chain (Fig 2C).

In the light chain (Fig 3A and 3B), Y38 is predicted to be one of the most important residues in the interaction; it contributes to the formation of a hydrophobic pocket that buries a proline residue and is within hydrogen bonding distance, via its hydroxyl group, to a number of backbone and side-chain groups of the peptide. Loss of this side-chain abolished binding. Y56 also forms part of the same proline-binding pocket as Y38, and loss of this side-chain also resulted in an almost complete loss of binding. R109 forms a hydrogen bond to an asparagine residue on the side of the helix; mutation of this residue to alanine results in a partial loss of binding. Y116 is located at the center of the second proline-binding pocket; since loss of the entire side-chain through an alanine mutation would lead to general structural disruption of the F<sub>AB</sub> fragment, we mutated this to a phenylalanine (removing the hydroxyl group), which led to a significant reduction in binding. Finally, S36A was selected as a control; the model indicated that it was outside the binding site, and the ELISA data indicated that had no effect on (NANP)<sub>6</sub> binding.
Within the heavy chain (Fig 3C and 3D), mutation of N57 to alanine led to complete loss of binding, which is consistent with it forming a hydrogen bond to a side-chain asparagine but also being part of a relatively well packed region of the binding site that is mostly buried upon binding. T66 is located on the edge of the binding site and appears to provide hydrophobic contacts through its methyl group with the methyl side-chain of an alanine of the peptide; mutation of this residue resulted in a partial loss of binding. Interestingly, mutation of E64, which is located in an appropriate position to form some hydrogen bonds to the peptide resulted in a slight increase in binding, although charged residues on the edge of protein:protein interfaces are known to contribute primarily to specificity rather than affinity [32]. Specifically, the cost of desolvating charged residues such as glutamate is not compensated for by the hydrogen bonds that may be formed with the binding partner. Y37 is located outside the direct binding site in the apo-crystal structure; the loss of affinity could arise from long-range effects, such as destabilization of the position of nearby loops. In general, the effects of the mutations are consistent with the model of the interaction.

The multivalency of the CSP repeat region

The binding mode of the F_AB fragment to the (NANP)_6 peptide is centered on two proline residues from two non-adjacent NANP-repeats (Fig 3A and 3C). These cyclic side-chains are hydrophobic in character and are buried deeply in the core of the F_AB antigen binding site,
into hydrophobic pockets formed by Y38 and Y56 of the light chain and the interface between the two chains. In contrast, the polar asparagine residues on the sides of the helix are involved in hydrogen bonding interactions with a number of polar residues on the edge of the binding site, such as N57 of the heavy chain. Due to the twisting of the (NANP)_6 repeat, the binding epitope of the peptide is 2.5–3 alternate NANP repeats, with a symmetrical epitope available for binding on the opposite face (Fig 4A). Thus, this binding mode is consistent with the stoichiometry of the binding observed in the ITC measurements, where we observed a stoichiometry of two 2A10 F\textsubscript{AB} fragments per (NANP)_6 peptide. To investigate whether this binding mode was also compatible with the indication from ITC that ~10.7 2A10 F\textsubscript{AB} fragments, or six antibodies (containing 12 F\textsubscript{AB} domains) could bind the CSP protein (Table 1), we extended the peptide to its full length. It is notable that the slight twist in the NANP helix results in the epitope being offset along the length of the repeat region, thereby allowing binding of ten 2A10 F\textsubscript{AB} fragments (Fig 4B). Six 2A10 antibodies can bind if two antibodies interact by a single F\textsubscript{AB} domain and the other four interact with both F\textsubscript{AB} domains. The observation that the F\textsubscript{AB} fragments bind sufficiently close to each other to form hydrogen bonds also explains the observation from the ITC that the complexes with rCSP, which allow adjacent F\textsubscript{AB} fragment binding, have more favorable binding enthalpy, i.e. the additional bonds formed between adjacent F\textsubscript{AB} fragments further stabilize the complex and lead to greater affinity (Table 1). Thus, the initially surprising stoichiometry that we observe through ITC appears to be quite feasible based on the structure of the NANP-repeat region of the rCSP protein and the nature of the rCSP-2A10 complex. It is also clear that the effect of antibody binding to this region would be

Fig 4. The multivalency of the NANP repeat region of the CSP protein. (A) An (NANP)_6 peptide results in the presentation of two symmetrical epitopes, formed by alternating repeats (cyan and magenta), allowing binding by two F\textsubscript{AB} domains, in keeping with the stoichiometry observed by ITC. (B) The full 27-mer repeat region results in the presentation of at least 10 separate epitopes and the twist of the helix results in displacement along the length of the repeat region, which allows binding of up to 10 separate F\textsubscript{AB} fragments, consistent with 4 antibodies bound by both F\textsubscript{AB} domains, and two bound by a single F\textsubscript{AB} domain.

https://doi.org/10.1371/journal.ppat.1006469.g004
to prevent the linker flexing between the N- and C-terminal domains and maintaining normal physiological function, explaining the neutralizing effect of the antibodies.

Identification of endogenous (NANP)$_n$ specific B cells to determine the BCR repertoire

We next set out to determine the implications of our structure for the B cell response to CSP. Because the CSP protein could conceivably cross-link multiple B cell receptors (BCRs) we hypothesized that the B cell response might be T-independent. As a tool to test this hypothesis we used (NANP)$_n$-based tetramers to identify and phenotype antigen specific B cells in mice immunized with P. berghei sporozoites expressing the repeat region of the P. falciparum CSP (P. berghei CS$^P$) [15]. The tetramers are formed by the binding of 4 biotinylated (NANP)$_9$ repeats with streptavidin conjugated phycoerythrin (PE) or allophycocyanin (APC). To validate our tetramer approach, mice were immunized with either P. berghei CS$^P$ or another line of P. berghei with a mutant CSP (P. berghei CS$^{SM}$) that contains the endogenous (P. berghei) repeat region, which has a distinct repeat sequence (PPPNPND)$_n$. (NANP)$_n$-specific cells were identified with two tetramer probes bound to different conjugates to exclude B cells that are specific for the PE or APC components of the tetramers which are numerous in mice [33]. We found that mice immunized with P. berghei CS$^P$ sporozoites developed large tetramer double positive populations, which had class switched (Fig 5A and 5B). In contrast, the number of tetramer double positive cells in mice receiving control parasites was the same as in unimmunized mice; moreover these cells were not class switched and appeared to be naive precursors indicating that our tetramers are identifying bona-fide (NANP)$_n$-specific cells (Fig 5B and 5C). Further analysis of the different populations of B cells showed that most B cells present at this time-point were GL7$^+$ CD38$^-$ indicating that they are GCB cells in agreement with results from a recent publication [34] (Fig 5B and 5D). Given that T cells are required to sustain GC formation beyond ~3 days these data indicate that a T-dependent response can develop to CSP following sporozoite immunization [35].

The B cell response to the (NANP)$_n$ repeat has both T-independent and T-dependent components

Our previous data showing GC formation among (NANP)$_n$ specific B cells was indicative of a T-dependent response. To determine whether there might also be a T-independent component to the B cell response we immunized CD28$^{-/-}$ mice as well as C57BL/6 controls with P. berghei CS$^P$ radiation attenuated sporozoites (RAS) and measured serum (NANP)$_n$ specific antibody by ELISA and the B cell response using our Tetramers. CD28$^{-/-}$ mice have CD4+ T cells but they are unable to provide help to B cell responses [36]. Interestingly4 days post immunization there were comparable IgM and IgG anti-(NANP)$_n$ responses in the CD28$^{-/-}$ mice and control animals (Fig 6A), indicative of a T-independent component to immunity. However by day 27 post immunization there was no detectable IgM or IgG antibody specific for (NANP)$_n$ in the CD28$^{-/-}$ mice suggesting the T-independent response is short-lived. We further analyzed (NANP)$_n$ specific B cell responses using our tetramers, in particular examining the number and phenotype (plasmablast vs GC B cell) of activated IgD$^-$ Tetramer$^+$ cells (Fig 6B). In agreement with our antibody data, similar numbers of antigen specific B cells were seen at 4 days post immunization in the CD28$^{-/-}$ and control mice and most of these cells were plasmablasts (Fig 6C). However by 7 days post immunization the number of antigen specific cells declines in the CD28$^{-/-}$ mice as the T dependent GC reaction begins to predominate. Thus CSP on the surface of sporozoites is able to induce short-lived T-independent B cell response, but subsequently T-dependent responses predominate.
We wanted to know if to induce a T-independent response it was necessary for CSP to be presented on the surface of the sporozoite or if free rCSP was sufficient. We found that indeed rCSP could induce a T-independent response as evidenced by similar IgM and IgG levels and IgD-Tetramer+ responses 4 days post immunization in control and CD28-/- mice (Fig 6D and 6E). Finally we were concerned that there may be some residual CD4+ T cell help in the CD28-/- mice so we performed experiments in which we used the antibody GK1.5 to deplete CD4+ T cells [37]. In agreement with our previous data we found that sporozoites (live or RAS) and rCSP induced IgM responses in CD4 depleted mice, though we were unable to detect a significant IgG response (S4 Fig). We also detected primed antigen specific B cells in GK1.5 treated mice following RAS or rCSP immunized mice 4 days post-immunization, albeit at lower levels than in mice treated with isotype control antibodies (S4 Fig). Overall our data with GK1.5 depleted mice support our results in the CD28-/- model.

Fig 5. CSP-specific B cells enter the germinal center following sporozoite immunization. BALB/C mice were immunized with either 5 x 10^5 P. berghei CS5M (expressing the endogenous P. berghei CSP repeat) or 5 x 10^5 P. berghei CSF (expressing the circumsporozoite protein from P. falciparum) live sporozoites under CQ cover. 12 days later the B cell response was analyzed by flow cytometry and putative (NANP)n-specific cells were identified using PE and APC conjugated tetramers. (A) Representative flow cytometry plots showing the identification of (NANP)n-specific (Tetramer+) cells. (B) Representative flow cytometry plots showing the proportion of Tetramer+ cells that have class switched and entered a GC. (C) Quantification of the number of class switched Tetramer+ cells under different immunization conditions. (D) Quantification of the number of GC Tetramer+ cells under different immunization conditions. Data from a single representative experiment of 2 repeats, analyzed by one-way ANOVA with Tukey’s post test.

https://doi.org/10.1371/journal.ppat.1006469.g005

We wanted to know if to induce a T-independent response it was necessary for CSP to be presented on the surface of the sporozoite or if free rCSP was sufficient. We found that indeed rCSP could induce a T-independent response as evidenced by similar IgM and IgG levels and IgD "Tetramer+" responses 4 days post immunization in control and CD28-/- mice (Fig 6D and 6E). Finally we were concerned that there may be some residual CD4+ T cell help in the CD28-/- mice so we performed experiments in which we used the antibody GK1.5 to deplete CD4+ T cells [37]. In agreement with our previous data we found that sporozoites (live or RAS) and rCSP induced IgM responses in CD4 depleted mice, though we were unable to detect a significant IgG response (S4 Fig). We also detected primed antigen specific B cells in GK1.5 treated mice following RAS or rCSP immunized mice 4 days post-immunization, albeit at lower levels than in mice treated with isotype control antibodies (S4 Fig). Overall our data with GK1.5 depleted mice support our results in the CD28-/- model.
Fig 6. The B cell response to CSP has a T-independent component. CD28−/− and control C57BL/6 mice were immunized with P. berghei CSP or rCSP in alum. Sera were taken and the spleens analyzed for antigen specific B cells using tetramers 4, 7 and 27 days post-immunization. (A) IgM and IgG (NANP)n ELISA responses following RAS immunization (B) Representative flow cytometry plots 7 days post RAS immunization showing the gating of different B cell populations among Tetramer+ cells. (C) Absolute numbers of (i) total Tetramer+ IgD-ve (ii) Tetramer+ Plasmablasts and (iii) Tetramer+ GC B cells post RAS immunization. (D) Antibody responses and (E) absolute numbers of Tetramer+ IgD- B cells 4 days post immunization with rCSP. Log-transformed data pooled from 2 independent experiments for each immunization (>3 mice/group/timepoint) were analyzed using linear mixed models with day and genotype/immunization as experimental factors and the individual experiment as a random factor; only significant differences are shown.
A restricted repertoire of BCRs can bind to the (NANP)_n repeat

Our ability to identify and sort (NANP)_n specific B cells with our tetramers also allows us to examine the repertoire of antibodies that can bind the (NANP)_n by sequencing the BCRs of the identified cells. While the repeat structure of CSP has been hypothesized to induce a broad polyclonal response based on data that the CSP repeat can absorb most of the sporozoite binding activity of human sera from immune individuals [23,38], an alternative hypothesis is that the antigenically simple structure of the repeat epitope might only be recognized by a small number of naïve B cells. We therefore sorted (NANP)_n-specific cells 35 days post immunization of BALB/C mice with sporozoites. We performed this analysis in BALB/C mice as this is the background of mice from which the 2A10 antibody was derived. We then prepared cDNA from the cells and amplified the heavy and kappa chain sequences using degenerate primers as described previously [39,40]. Heavy and kappa chain libraries were prepared from 4 immunized mice as well as from 3 naïve mice from which we bulk sorted B cells as controls. We obtained usable sequences from 3 of the 4 mice for both the heavy chain and kappa chain. Analysis of the heavy chain revealed that in each mouse 3 or 4 V regions dominated the immune response (Fig 7A). The V regions identified (IGHV1-20; IGHV1-26; IGHV1-34 and IGHV5-9) were generally shared among the mice. As a formal measure of the diversity of our V region usage in the (NANP)_n specific cells and the bulk B cells from naïve mice we calculated the Shannon entropy for these populations. This analysis formally demonstrated that the diversity of the antigen specific B cells was significantly lower than the diversity of the repertoire in naïve mice (Fig 7B). We further found that each V region was typically associated with the same D and J sequences even in different mice. For example, IGHV1-20 was typically associated with J4, IGHV5-9 with J4 while in different mice IGHV1-34 was variously paired with J1 or J4 (Fig 7C). Similar results were obtained for the kappa chain with the response dominated by IGKV1-135; IGKV5-43/45; IGKV1-110; IGKV1-117 and IGKV14-111 (Fig 7D and 7E). The V regions were typically paired with the same J regions even in different mice (Fig 7F), for example IGKV5.43/45 was typically paired with IGKJ5 or IGKJ2 and IGKV1-110 was typically paired with IGKJ5, although IGKV1-135 was typically more promiscuous. One limitation of our high throughput sequencing approach is that the degenerate primers only amplified ~70% of the known IGHV and IGKV sequences in naïve mice, suggesting that we may not capture the full diversity of the response. However, comparison with the 5 published antibody sequences (S2 and S3 Tables) that include IGHV-1-20, IGKV5-45 and IGKV1-110 reveals that we are likely capturing the bulk of the antibody diversity. Together these data suggest that the number of B cell clones responding to CSP may be limited, potentially reducing the ability of the immune system to generate effective neutralizing antibodies.

CSP-binding antibodies undergo somatic hypermutation to improve affinity

Finally we were interested in knowing if the GC reaction we could see following sporozoite immunization was inducing higher affinity antibodies. We therefore examined our deep sequencing data to determine if CSP-specific antibodies had undergone somatic hypermutation (SHM) that would be indicative of B cells specific for CSP entering the GC. Taking advantage of the fact that our kappa chain primers capture the entire V-J sequences of the antibodies we sequenced we asked: 1) if the kappa chains shared between immune animals differed from the germline (providing evidence of SHM) and 2) if the mutations were conserved between different mice indicative of directed selection. Analysis of the reads from the kappa chains of the three immune mice showed that these had a much higher degree of mutation than bulk B cells from naïve mice, demonstrating SHM in the CSP-specific antibodies (Fig 8A). We further
Fig 7. Limited diversity of (NANP)ₙ specific antibodies. BCR sequences were amplified from Tetramer⁺ cells sorted from BALB/C mice 35 days after immunization with live *P. berghei* CSPf sporozoites under CQ cover as well as bulk (B220⁺) B cells from naive BALB/C mice. (A) IGHV gene usage from among B cells from a representative naive mouse (grey bars) and Tetramer⁺ cells from immune mice (red, blue and yellow bars). (B) Shannon’s diversity calculated for the diversity of IGHV region usage among bulk B cells and Tetramer⁺ cells. (C) Circos plots showing the IGHV-IGHJ pairings in a representative naive mouse and 3 immune mice. (D) IGKV gene usage from among B cells from a representative naive mouse (grey bars) and Tetramer⁺ cells from
examined each specific common kappa chain in turn (IGVK1-110; IGKV1-135; IGVK5-43/45) comparing the sequences obtained from naïve B cells and (NANP)_n specific cells in immune mice. This analysis showed that while, as expected, sequences from naïve mice contained few mutations, the sequences from immune mice had much higher levels of SHM. Importantly mutations were found to be concentrated in the CDR loops, and were frequently shared by immunized mice providing strong circumstantial evidence for affinity maturation (Fig 8B; data for IGVK1-110 only shown).

To directly test if CSP-binding antibodies undergo affinity maturation we expressed the predicted germline precursor to the 2A10 antibody (2A10 gAb) in HEK293T cells. We identified the predicted germline precursors of the 2A10 heavy and light chains using the program V-quest [41] (S5 and S6 Figs). This analysis identified the heavy chain as IGHV9-3; IGHD1-3; IGHJ4 and the light chain as IGKV10-94;IGKJ2, with the monoclonal antibody carrying 6 mutations in the heavy chain and 7 in the light chain. The 2A10 gAb had considerably lower binding in ELISA assays compared to the 2A10 mAb itself (Fig 8C), indicative that affinity maturation had taken place in this antibody. To determine the relative contribution of mutations in the heavy and light chain to enhancing binding we also made hybrid antibodies consisting of the mAb heavy chain and the gAb light chain and vice versa. Interestingly mutations in the light chain were almost entirely sufficient to explain the enhanced binding by the mAb compared to the gAb (Fig 8C).

To identify the specific mutations that were important we introduced the mutations individually into the gAb light chain construct. We prioritized mutations that were shared with the 27E antibody which has previously been found to be clonally related to 2A10 having been isolated from the same mouse and which shares the same germline heavy and light chains as the 2A10 mAb [20]. We found that two mutations (L114F and T117V) in the CDR3 of the light chain appeared to account for most of the gain in binding (Fig 8C). The effect of these antibodies appeared to be additive rather than synergistic as revealed by experiments in which we introduced these mutations simultaneously (Fig 8D). A further mutation close to the light chain CDR2 (H68Y) also caused a modest increase in binding. As expected mutations in the heavy chains appeared generally less important for increasing binding though M39I, N59I and T67F all gave modest increases in binding (Fig 8E). Collectively our data suggest that CSP repeat antibodies can undergo SHM in GCs resulting in affinity maturation, however the antibody response may be limited by the number of naïve B cells that can recognize and respond to this antigen.

Discussion

Here we provide an analysis of the structure of a Plasmodium falciparum sporozoite-neutralizing antibody (2A10). Having obtained this structure we further modeled the binding 2A10 with its antigen target, the repeat region of CSP, and provide a thermodynamic characterization of this interaction. Finally, we used novel tetramer probes to identify and sort antigen specific B cells responding to sporozoite immunization in order to measure the diversity and maturation of the antibody response. We found that the avidity of 2A10 for the rCSP molecule was in the nanomolar range, which was much higher than the affinity previously predicted from competition ELISAs with small peptides [22,23]. This affinity is a consequence of the multivalent nature of the interaction, with up to 6 antibodies being able to bind to each rCSP...
molecule. Our model suggests that to spatially accommodate this binding the antibodies must surround CSP in an off-set manner, which is possible due to the slight twist in the helical structure that CSP can adopt. It is notable that the twisted, repeating arrangement of the CSP linker is the only structure that would allow binding in the stoichiometry observed through the ITC. We further found that the diversity of the antibody repertoire to the CSP repeat was limited, perhaps due to the relative simplicity of the target epitope. However, these antibodies have undergone affinity maturation to improve affinity, potentially allowing protective immune responses to develop.

Using ITC we determined the dissociation constant of 2A10 for rCSP to be 2.7 nM, which is not unusual for a mouse mAb. However it is a tighter interaction than that predicted from competition ELISAs, which predicted a micro-molar affinity [22,23]. However, these competition ELISAs were performed with short peptides rather than rCSP. Indeed, when we performed ITC with a short peptide and F_{AB} fragments we too obtained a dissociation constant in the micro-molar range (0.42 μM). The difference between the F_{AB} binding to the peptide and the tight interaction of the antibody binding to full length CSP appears to be driven by a high

Fig 8. CSP-binding antibodies undergo somatic hypermutation and affinity maturation. (A) Violin plots showing the number of mutations per kappa chain read from bulk B cells from 3 individual naïve mice and sorted (NANP)_n specific B cells from sporozoite immunized mice (B) Skyscraper plots showing the location of mutations away from germline in the IGKV1-110 gene in a naïve mouse and in sorted (NANP)_n specific cells in three sporozoite immunized mice. (C) ELISA binding to the (NANP)_n peptide of recombinant antibodies corresponding to the 2A10 mAb, the predicted germline precursor, and hybrid antibodies containing the 2A10 heavy chain (mHC) paired with the germline light chain (gLC) and the 2010 light chain (mLC) paired with germline heavy chain (gHC). (D) Predicted mutations in the gLC were introduced to the germline precursor and their effect on binding to (NANP)_n measured by ELISA (E) Predicted mutations in the gHC were introduced to hybrid antibodies consisting of the mLC and the gHC heavy chain and their effect on binding to (NANP)_n measured by ELISA.

https://doi.org/10.1371/journal.ppat.1006469.g008
avidity, multivalent interaction. There is also additional enthalpic stabilization (per F\textsubscript{AB}
domain) in the 2A10:CSP complex, although this is partially offset by the increased entropic
cost associated with combining a large number of separate molecules into a single complex.
One caveat of these data is that we used a slightly truncated repeat in our recombinant CSP,
however it is likely that longer repeats will have further stabilization of the interaction that
could result in even higher affinity interaction between CSP and binding antibodies.

The mechanism of sporozoite neutralization remains unclear, however our structural data
may provide some insights. Repeat specific antibodies can directly neutralize sporozoites
(without complement or other cell mediators) in the circumsporozoite reaction [8,42]. More-
over F\textsubscript{AB} fragments alone are sufficient to block invasion [42,43]. However, it is well estab-
lished that activation of complement and cell mediated immunity is important for the action
of blood stage-specific antibodies [44,45]. It has also been suggested that the CSP repeat might
act as a hinge allowing the N-terminal domain to mask the C-terminal domain which is
believed to be important for binding to and invading hepatocytes [10]. Cleavage of this N ter-
minal domain is therefore required to expose the C-terminal domain and facilitate invasion
[10]. Antibody binding as observed here may disrupt this process in several ways, either by
opening the hinge to induce the premature exposure of the C-terminal domain. Alternati-
vely since the repeat region is directly adjacent to the proteolytic cleavage site, anti-repeat antib-
odies might function by sterically hindering access of the protease to CSP, thus preventing
sporozoite invasion of the hepatocyte. One possible consequence of the requirement for muti-
valency to increase the avidity of the antibody, is that antibodies with different binding modes
may interfere with each other limiting their effectiveness.

Our results uncovering how neutralizing antibodies bind to CSP has several implications
for understanding the development of the immune response to CSP. Notably the finding that
the CSP molecule can be bound by multiple antibodies/B cell receptors raises the possibility
that this molecule can indeed crosslink multiple BCRs and potentially act as a type-II T
independent antigen [17]. We find that indeed there is a T-independent component to the
response to CSP, though T cells are required to sustain the immune response beyond day 7. As
such the response to CSP appears follow a similar process to that seen for several oligomeric
viral entry proteins, which induce a mix of T-independent and T-dependent responses [18,19].
It maybe that T-independent responses are driven by the density of CSP molecules on the spo-
rozoite surface; however, rCSP can also induce a small T-independent response. This suggests
that the CSP protein alone is sufficient to crosslink multiple BCRs on the B cell surface which
is consistent with our structural model. Interestingly, the RTS,S/AS01 vaccine based on that
contains 18 CSP repeats and does appear to induce high titers of anti-CSP antibodies which
initially decline rapidly and are then more stable [4,46]. This may be consistent with the induc-
tion of a short-lived a type-II T-independent plasmablast response (accounting for the initial
burst of antibodies), followed by a T-dependent response (which may be the basis of the more
sustained antibody titers). The relative contributions of short-lived antibody production and
long-term B cell memory to protection is an area for future investigation.

The finding of a limited repertoire in the BCR sequences specific for the (NANP)\textsubscript{n}
repeat contradicts previous suggestions that the response to CSP might be broad and polyclonal [38].
One explanation for this limited antibody diversity is that the antigenic simplicity of the CSP
repeat region limits the range of antibodies that are capable of responding. A prior example of
this is the antibodies to the Rhesus (Rh) D antigen. The RhD antigen differs from RhC by only
35–36 amino acids, resulting in the creation of a minimal B cell epitope [47]. The repertoire of
antibodies that can bind this epitope are accordingly limited and mainly based on the VH3-33
gene family [48]. Another potential explanation for a limited antibody repertoire could be that
the (NANP)\textsubscript{n} repeat shares structural similarity with a self-antigen as is speculated to happen
with meningococcus type B antigens [49], however it is not clear what this self-antigen might be. One potential outcome of this finding is that if each B cell clone has a finite burst size this may limit the magnitude of the overall B cell response.

One area for future investigation is to determine the binding modes and sporozoite neutralizing capacities of other antibodies in the response. It is clear that not all CSP-repeat binding antibodies have the same capacity for sporozoite neutralization [7]. As such the finding of a limited repertoire of responding B cells may lead to the possibility that some people have holes in their antibody repertoires limiting their ability to make neutralizing antibodies. This may explain why, while there is a broad correlation between ELISA titer of antibodies to the CSP repeat and protection following RTS,S vaccination, there is no clear threshold for protection [4].

While our work has been performed with mouse antibodies, there are major similarities between mouse and human antibody loop structure [50]. The main difference between the two species is the considerably more diverse heavy chain CDR3 regions that are found in human antibodies [51]. Consequently, this leads to a much larger number of unique clones found in humans compared to mice. However, the number of different V, D and J genes and the recombination that follows are relatively similar between humans and mice [52]. From our data it can be observed that while the BCR repertoire was restricted in the V gene usage, these different V gene populations were represented in multiple unique clones, suggesting that increasing the number of clones is unlikely to substantially increase V-region usage. Our analysis was performed on inbred mice which may also limit repertoire diversity, however studies on the human IGHV locus reveal that in any given individual ~80% V region genes are identical between the maternal and paternal allele i.e. heterozygosity is not a major driver of human V region diversity [53,54]. It is notable that all 4 human monoclonal antibodies described to date from different volunteers share the use of the IGHV3-30 gene family [21,22], suggesting that in humans as well as mice there may indeed be a constrained repertoire of responding B cells.

Overall our data provide important insights into how the antibody response to CSP develops. Our results also help explain why relatively large amounts of antibodies are required for sporozoite neutralization and suggest that the ability to generate an effective B cell response may be limited by the very simplicity of the repeat epitope. These data support previous suggestions that CSP may be a suboptimal target for vaccination. However, we also find that CSP binding antibodies can undergo somatic hypermutation and reach high affinities. This suggests if we can develop vaccination strategies to diversify the repertoire of responding B cells and favor the GC response it may be possible to generate long-term protective immunity targeting this major vaccine candidate antigen.

Methods
Ethics statement
All animal procedures were approved by the Animal Experimentation Ethics Committee of the Australian National University (Protocol numbers: A2013/12 and A2016/17). All research involving animals was conducted in accordance with the National Health and Medical Research Council’s (NHMRC) Australian Code for the Care and Use of Animals for Scientific Purposes and the Australian Capital Territory Animal Welfare Act 1992.

Mice, immunizations and cell depletions
BALB/C, C57BL/6 or CD28−/− [55] mice (bred in-house at the Australian National University) were immunized IV with 5 x 10^4 P. berghei CS^M sporozoites expressing mCherry [56] or 5 x
10⁴ *P. berghei* Cs⁹⁹ sporozoites dissected by hand from the salivary glands of *Anopheles stephensi* mosquitoes. Mice were either infected with live sporozoites and then treated with 0.6mg chloroquine IP daily for 10 days or immunized with irradiated sporozoites (15kRad). For immunization with rCSP, 30μg rCSP was emulsified in Injext Alum according to the manufacturer’s instructions (ThermoFisher Scientific) and delivered intra-peritoneally. All mice received only a single immunization in these experiments. To deplete CD4+ T cells mice were treated with two doses of 100μg GK1.5 antibody on the 2 days prior to immunization (BioXCell); control mice received an irrelevant isotype control antibody (LTF2; BioXCell).

**Flow cytometry and sorting**

Single cell preparations of lymphocytes were isolated from the spleen of immunized mice and were stained for flow cytometry or sorting by standard procedures. Cells were stained with lineage markers (anti-CD3, clone 17A2; anti-GR1, clone RB6-8C5 and anti-NKp46, clone 29A1.4) antibodies to B220 (clone RA3-6B2), IgM (clone II/41), IgD (clone 11-26c2a), GL7 (clone GL7), CD38 (clone 90), CD138 (clone 281-2) and (NANP)₉ tetramers conjugated to PE or APC. Antibodies were purchased from Biolegend while tetramers were prepared in house by mixing biotinylated (NANP)₉ peptide with streptavidin conjugated PE or APC (Invitrogen) in a 4:1 molar ratio. Flow-cytometric data was collected on a BD Fortessa flow cytometer (Becton Dickinson) and analyzed using FlowJo software (FlowJo). Where necessary cells were sorted on a BD FACS Aria I or II machine.

**Sequencing of (NANP)ₙ specific cells and BCR analysis**

Single cell suspensions from the spleens of immunized mice were stained with (NANP)ₙ tetramers and antibodies to B cell markers as described in the supplementary experimental procedures. Antigen specific cells were sorted on a FACS ARIA I or II instrument prior to RNA extraction with the Arturus Picopure RNA isolation kit (Invitrogen) and cDNA preparation using the iScript cDNA synthesis kit (BioRad). BCR sequences were amplified using previously described heavy and kappa chain primers including adaptor sequences allowing subsequent indexing using the Nexetera indexing kit (Illumina). Analysis was performed in house using R-scripts and the program MiXCR as described in supplementary experimental procedures.

**Binding of antibody variants**

Variants of the 2A10 antibody were expressed in HEK293 T cells (a kind gift of Carola Vinuesa, Australian National University) as described in the supplementary experimental procedures. Binding to the CSP repeat was tested by ELISA and ITC using standard techniques as described in the supplementary experimental procedures.

**Statistical analysis**

Statistical analysis was performed using Prism6 (GraphPad) for simple T tests and one-way ANOVAs from single experiments. Where data were pooled from multiple experiments, analysis was performed using linear mixed models in R version 3.3.3 (R foundation for Statistical Computing). Linear mixed models are a regression analysis model containing both fixed and random effects: fixed effects being the variable/treatment under examination, whilst random effects are unintended factors that may influence the variable being measured. If significance was found from running a linear mixed model, pair-wise comparisons of the least significant differences of means (LSD) was undertaken to determine at which level interactions were
occurring. Statistical significance was assumed if the p-value was < 0.05 for a tested difference. (ns = not significant, * = p < 0.5, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001).

Accession numbers

Sequence data generated in this paper is deposited at the NCBI sequence read archive (SRA) with accession number SRP092808 as part of BioProject database accession number PRJNA352758. Atomic coordinates and related experimental data for structural analyses are deposited in the Protein Data Bank (PDB) with PDB codes 5SZF and 5T0Y.

Supporting information

S1 Fig. Theoretical (A) and experimental (B) CD spectra of the (NANP)₆ peptide. The computational prediction of the spectra (A) was performed using DichroCalc [57], the experimental spectra was measured at 222 nm at 25°C. A peak at 185 nm, minimum at 205 nm and shoulder between 215 and 240 nm are consistent with an intrinsically disordered, but not random coil, structure.

S2 Fig. Cluster analysis for MD simulations of (NANP)₆ peptide. Conformations were clustered by concatenating the trajectory and performing a Jarvis-Patrick analysis. The clusters are sorted by their RMSD from the first cluster (starting geometry). As shown, Run 2 is stable in the starting geometry for several ns, while Run 3 diverged, then reconverged to the starting geometry, where it was stable for several ns. These data suggest the quasi-helical structure observed from the ab initio calculations is stable, and can be spontaneously sampled, on a timescale of several ns.

S3 Fig. Cluster analysis for MD simulations of (NANP)₆ peptide. Molecular dynamics simulation of the (NANP)$_c$:F$_{AB}$ complex. Root mean square deviation (RMSD) of the (NANP)$_c$:F$_{AB}$ complex as a function of time. Independent simulations are shown in green, black and red.

S4 Fig. The B cell response to CSP has a T-independent component. Mice either treated with an anti-CD4 depleting antibody or an isotype control were immunized with either P. berghei CSPf RAS, live P. berghei CSPf under CQ cover or rCSP. (A) 4 days later the IgM and IgG response to the (NANP)$_n$ repeat was analyzed by ELISA (B) At the same time the number of IgD$^-$/Tetramer$^+$ B cells was quantified in the spleen. Data are from a single experiment, analyzed using linear models with immunization/treatment as the experimental factor.

S5 Fig. Alignment of 2A10 heavy chain and the predicted germline sequence. Residues that are mutated away from the predicted germline sequence in more one or more other antibody heavy chain (2E7 or 3D6) are highlighted in red, mutations that are predicted to be involved in binding to CSP are highlighted in blue.

S6 Fig. Alignment of 2A10 heavy chain and the predicted germline sequence. Residues that are mutated away from the predicted germline sequence in both 2A10 and the related 2E7 antibody are highlighted in red, mutations that are predicted to be involved in binding to CSP are highlighted in blue.
S1 Movie. Molecular dynamics simulation of the solution structure of the (NANP)$_6$ peptide. Excerpt from (NANP)$_6$ run 3. The trajectory was fitted to minimize alpha-carbon RMSD and then passed through a low-pass filter with a filter length of 8 frames to reduce temporal aliasing. (MP4)

S2 Movie. Molecular dynamics simulation of the interaction of the (NANP)$_n$ repeat with the 2A10 F$_{AB}$. Excerpt from 2A10:(NANP)$_6$ run 3. The trajectory was fitted to minimize alpha-carbon RMSD and then passed through a low-pass filter with a filter length 8 frames to reduce temporal aliasing. (MP4)

S1 Table. Data collection and refinement statistics for the crystal structures of 2A10 F$_{AB}$ presented in this work. (DOCX)

S2 Table. Heavy chain CDR sequences of CSP binding antibodies. (DOCX)

S3 Table. Light chain CDR sequences of CSP binding immunoglobulins. (DOCX)

S1 Methods. Contains details of extended methods and associated references. (DOCX)

Acknowledgments

We thank the C3 Crystallisation Centre at CSIRO for help with crystal formation and the Australian Synchrotron and beamline scientists for help with data collection. We thank Michael Devoy, Harpreet Vohra and Catherine Gillespie of the Imaging and Cytometry Facility at the John Curtin School of Medical Research for assistance with flow cytometry and multi-photon microscopy.

Author Contributions

Conceptualization: Colin J. Jackson, Ian A. Cockburn.

Data curation: Henry J. Sutton, Mandeep Singh, Aaron Chuah.


Funding acquisition: Colin J. Jackson, Ian A. Cockburn.


Project administration: Colin J. Jackson, Ian A. Cockburn.

Resources: Thomas S. Peat, Colin J. Jackson, Ian A. Cockburn.

Software: Aaron Chuah.

Supervision: Colin J. Jackson, Ian A. Cockburn.


Writing – original draft: Colin J. Jackson, Ian A. Cockburn.

Writing – review & editing: Joe A. Kaczmarski, Ben Clifton, Joshua Mitchell, Johanna N. Dups.

References


4.3 Supplementary Information for “T-dependent B cell responses to Plasmodium induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein”
Interactions between neutralising antibodies and the circumsporozoite protein from Plasmodium falciparum

Figure S1. Theoretical (A) and experimental (B) CD spectra of the (NANP)6 peptide.

The computational prediction of the spectra (A) was performed using DichroCalc, the experimental spectra was measured at 222 nm at 25 °C. A peak at 185 nm, minimum at 205 nm and shoulder between 215 and 240 nm are consistent with an intrinsically disordered, but not random coil, structure.
Conformations were clustered by concatenating the trajectory and performing a Jarvis-Patrick analysis. The clusters are sorted by their RMSD from the first cluster (starting geometry). As shown, Run 2 is stable in the starting geometry for several ns, while Run 3 diverged, then reconverged to the starting geometry, where it was stable for several ns. These data suggest the quasi-helical structure observed from the ab initio calculations is stable, and can be spontaneously sampled, on a timescale of several ns.
Figure S3. Cluster analysis for MD simulations of (NANP)6 peptide.

Molecular dynamics simulation of the (NANP)6:FAB complex. Root mean square deviation (RMSD) of the (NANP)6:FAB complex as a function of time. Independent simulations are shown in green, black and red.
Mice either treated with an anti-CD4 depleting antibody or an isotype control were immunized with either *P. berghei* CSPf RAS, live *P. berghei* CSPf under CQ cover or rCSP. (A) 4 days later the IgM and IgG response to the (NANP)n repeat was analyzed by ELISA (B) At the same time the number of IgD- Tetramer+ B cells was quantified in the spleen. Data are from a single experiment, analyzed using linear models with immunization/treatment as the experimental factor.
### Figure S5. Alignment of 2A10 heavy chain and the predicted germline sequence.

Residues that are mutated away from the predicted germline sequence in more than one or more other antibody heavy chain (2E7 or 3D6) are highlighted in red, mutations that are predicted to be involved in binding to CSP are highlighted in blue.
Figure S6. Alignment of 2A10 heavy chain and the predicted germline sequence.

Residues that are mutated away from the predicted germline sequence in both 2A10 and the related 2E7 antibody are highlighted in red, mutations that are predicted to be involved in binding to CSP are highlighted in blue.
Interactions between neutralising antibodies and the circumsporozoite protein from Plasmodium falciparum

S1 Table. Data collection and refinement statistics for the crystal structures of 2A10 F\textsubscript{AB} presented in this work.

<table>
<thead>
<tr>
<th></th>
<th>2A10 F\textsubscript{AB} 2.52</th>
<th>2A10 F\textsubscript{AB} 3.01</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PDB ID</strong></td>
<td>5SZF</td>
<td>5T0Y</td>
</tr>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>I\textsubscript{4}1\textsubscript{3}2</td>
<td>P432;2</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a, b, c) (Å)</td>
<td>204.21</td>
<td>231.68, 231.68, 81.78</td>
</tr>
<tr>
<td>(\alpha, \beta, \gamma) (°)</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>37.28-2.52</td>
<td>39.73-3.01</td>
</tr>
<tr>
<td></td>
<td>(2.62-2.52)</td>
<td>(3.12-3.01)</td>
</tr>
<tr>
<td>(R\text{merge})</td>
<td>0.178 (1.836)</td>
<td>0.315 (1.913)</td>
</tr>
<tr>
<td>(R\text{pim})</td>
<td>0.028 (0.406)</td>
<td>0.085 (0.517)</td>
</tr>
<tr>
<td>(CC\text{1/2})</td>
<td>0.999 (0.620)</td>
<td>0.991 (0.718)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>100 (100)</td>
<td>99.9 (99.5)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>39.2 (22.0)</td>
<td>14.6 (14.6)</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>37.28-2.52</td>
<td>39.73-3.01</td>
</tr>
<tr>
<td></td>
<td>(2.61-2.52)</td>
<td>(3.12-3.01)</td>
</tr>
<tr>
<td>No. reflections</td>
<td>24796 (2440)</td>
<td>44596 (4352)</td>
</tr>
<tr>
<td>(R\text{work} / R\text{free})</td>
<td>0.2251/0.2483</td>
<td>0.2248/0.2467</td>
</tr>
<tr>
<td>No. atoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>3288</td>
<td>9856</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>20</td>
<td>90</td>
</tr>
<tr>
<td>Water</td>
<td>46</td>
<td>25</td>
</tr>
<tr>
<td>Wilson (B)-factor</td>
<td>46.80</td>
<td>60.66</td>
</tr>
<tr>
<td>R.m.s. deviations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>0.57</td>
<td>0.55</td>
</tr>
<tr>
<td>Ramachandran favored (%)</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>Ramachandran outliers (%)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Values in parentheses are for highest-resolution shell
## S2 Table: Heavy chain CDR sequences of CSP binding antibodies

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Species</th>
<th>Heavy chain</th>
<th>CDR1</th>
<th>CDR1 group</th>
<th>CDR2</th>
<th>CDR2 group</th>
<th>CDR3</th>
<th>CDR3 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A10</td>
<td>Mouse</td>
<td>HV9-3*02</td>
<td>KASGTF....TNYGIN*</td>
<td>H1-13-1</td>
<td>WINTI..TEEPT</td>
<td>H2-10-1</td>
<td>ARGSEFGRLVY</td>
<td>N/A</td>
</tr>
<tr>
<td>PfNPNAI</td>
<td>Human</td>
<td>HV3-30-3</td>
<td>AASGFTF....SSYAMH</td>
<td>H1-13-1</td>
<td>VISYD..GSNKY</td>
<td>H2-10-2</td>
<td>DRDSSSYFDS</td>
<td>N/A</td>
</tr>
<tr>
<td>3D6</td>
<td>Mouse</td>
<td>HV9-2-1*01</td>
<td>KASGSPF....PDSSMP</td>
<td>H1-13-1</td>
<td>WINTA..TGEPT</td>
<td>H2-10-1</td>
<td>GGGGGPWFAY</td>
<td>N/A</td>
</tr>
<tr>
<td>2C11</td>
<td>Mouse</td>
<td>HV1-20 OR 37</td>
<td>KASGYSF....TGSFMN</td>
<td>H1-13-1</td>
<td>RINPN..DGYTF</td>
<td>H2-10-1</td>
<td>GKNHGATDY</td>
<td>N/A</td>
</tr>
<tr>
<td>1E9</td>
<td>Mouse</td>
<td>HV1-20 OR 37</td>
<td>KASGYSF....TGSFMN</td>
<td>H1-13-1</td>
<td>RILPY..NGDTF</td>
<td>H2-10-1</td>
<td>GYVYDGYYATDY</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>HV5-9</td>
<td>AASGFTF....SSYTMS</td>
<td>H1-13-1</td>
<td>TISSG..GGNTY</td>
<td>H2-10-1</td>
<td>VARIABLE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>HV1-20</td>
<td>KASGYSF....TGYFMN</td>
<td>H1-13-1</td>
<td>RINPY..NGDTF</td>
<td>H2-10-1</td>
<td>VARIABLE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>HV1-26</td>
<td>KASGTYTF....TDYYMN</td>
<td>H1-13-1</td>
<td>DINPN..NGGTS</td>
<td>H2-10-1</td>
<td>VARIABLE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>HV1-34</td>
<td>KASGTYTF....TDYYMN</td>
<td>H1-13-1</td>
<td>YIYPN..NGGNG</td>
<td>H2-10-1</td>
<td>VARIABLE</td>
<td></td>
</tr>
</tbody>
</table>

*Letters in bold denote residues show to be required for binding.*
## S3 Table: Light chain CDR sequences of CSP binding immunoglobulins.

*Letters in bold denote residues show to be required for binding

<table>
<thead>
<tr>
<th>Species</th>
<th>Light chain</th>
<th>CDR1</th>
<th>CDR1 group</th>
<th>CDR2</th>
<th>CDR2 group</th>
<th>CDR3</th>
<th>CDR3 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A10 Mouse</td>
<td>KV10-94</td>
<td>SASQGIG......SNYLN*</td>
<td>L1-11-2</td>
<td>FYTSTLYS</td>
<td>L2-8-1</td>
<td>QQYSRFPYYV</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>PfNPNAI</td>
<td>KV1-5</td>
<td>RASQSI......SSWLA</td>
<td>L1-11-2</td>
<td>YDASSLES</td>
<td>L2-8-1</td>
<td>QQYNYSGLT</td>
<td>L3-10-1</td>
</tr>
<tr>
<td>3D6 Mouse</td>
<td>KV6-20</td>
<td>KASENV......VTYVS</td>
<td>L1-11-1</td>
<td>YRASNRYT</td>
<td>L2-8-1</td>
<td>GQQSSYPYT</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>2C11 Mouse</td>
<td>KV5-43</td>
<td>RASQNI......SNNLH</td>
<td>L1-11-1</td>
<td>TYASQSIS</td>
<td>L2-8-1</td>
<td>QQSNSWPLT</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>1E9 Mouse</td>
<td>KV1-110</td>
<td>RSSQSLGHS.HGNTYLH</td>
<td>L1-16-1</td>
<td>YKVSNRFS</td>
<td>L2-8-1</td>
<td>SQSTQLRT</td>
<td>L3-8-1</td>
</tr>
<tr>
<td>High Mouse</td>
<td>KV1-117</td>
<td>RSSQIVHS.NGNTYLE</td>
<td>L1-16-1</td>
<td>YKVSNRFS</td>
<td>L2-8-1</td>
<td>FQGSHVPPTF</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>Throughput</td>
<td>KV1-110</td>
<td>RSSQSLVHS.NGNTYLH</td>
<td>L1-16-1</td>
<td>YKVSNRFS</td>
<td>L2-8-1</td>
<td>SQSTHLPLTF</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>Sequencing</td>
<td>KV1-135</td>
<td>KSSQSSLDS.DGKTYLN</td>
<td>L1-16-1</td>
<td>YLVSKLDS</td>
<td>L2-8-1</td>
<td>WQGTHFPPTF</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>Mouse</td>
<td>KV5-43</td>
<td>RASQSI......SNNLH</td>
<td>L1-11-1</td>
<td>KYASQSIS</td>
<td>L2-8-1</td>
<td>QQSNSWPLTF</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>Mouse</td>
<td>KV5-45</td>
<td>RASQSI......SNYLH</td>
<td>L1-11-1</td>
<td>KYASQSIS</td>
<td>L2-8-1</td>
<td>QQSNSWPLTF</td>
<td>L3-9-cis7-1</td>
</tr>
<tr>
<td>Mouse</td>
<td>KV14-111</td>
<td>KASQDI......NSYLS</td>
<td>L1-11-1</td>
<td>YRANRLVD</td>
<td>L2-8-1</td>
<td>LQYGEPFPPTF</td>
<td>L3-9-cis7-1</td>
</tr>
</tbody>
</table>
S1 Methods: Supplementary Methods

Generation of antibody variants

Constructs containing minigenes for the monoclonal and germline heavy (isotype: IgG2A) and light chains of the 2A10 antibody in a pcDNA3.1+ backbone were ordered commercially (Biomatik). Mutations described in the figure legends were introduced using the QuickChange II site directed mutagenesis kit according to the manufacturers instructions (Agilent). To generate antibodies HEK293T cells grown in DMEM supplemented with Nutridoma-SP (Roche) were transfected with 15 μg of each of the heavy and light chain plasmids in 0.06mg/ml branched PEI in 120mM NaCl. 3 and 6 days following transfections supernatants were collected, concentrated over a 100kDa Ultra-15 centrifugal filter unit, Ultracell-100 membrane (Amicon). Antibody concentrations were determined by sandwich ELISA on coats plated with anti-mouse kappa (Southern Biotech) as capture antibodies and horseradish peroxidase conjugated anti-mouse IgG2A (KPL) as detection antibodies.

ELISA

Binding of 2A10 antibody variants was determined in solid phase ELISA. Briefly, Nunc Maxisorp Plates (Nunc-Nucleon) were coated overnight with 1μg/ml streptavidin followed by binding of biotinylated (NANP)₉ peptide for 1 hour. After blocking with 1% BSA, serial dilutions of the antibodies were incubated on the plates for 1 hour and after washing, incubated with HRP conjugated anti-IgG2A antibodies (KPL). For the analysis of sera from immunized mice data were expressed as the area under the curve (AUC) which was calculated in Genstat, using the log(dilution) on the x axis and the Absorbance at 405nm on the y axis. The mean AUC from a group of naïve control mice in each experiment was subtracted from the AUC of each immunized mouse to remove background.

High throughput sequencing of (NANP)₉ specific B cell receptors
RNA was extracted using the Arcturus Picopure RNA isolation kit and cDNA prepared using the iScript cDNA synthesis kit (Biorad) according to the manufacturer’s instructions. BCR sequences were amplified using previously described heavy and kappa chain primers including adaptor sequences allowing subsequent indexing using the Nextera indexing kit (Illumina). Amplification conditions were 1 cycle at 95°C for 5 minutes followed by 50 cycles at 95°C for 1 minute, 43°C for 1 minute and 72 for 1.5 minutes then finally 1 cycle at 72°C for 5 minutes before holding at 4°C to cool. Following initial amplification PCR products were cleaned up using AMPure XP beads (Beckman Coulter) according to the manufacturer's instructions. Subsequently the cleaned up libraries were used as templates for the indexing step using the Nextera indexing kit (Illumina). Indexing PCR was performed using the following setting 72°C for 30 seconds then 95°C for 30 seconds followed by 15 cycles of 95°C for 15 seconds, 63°C for 30 seconds then finally 73°C for 3 minutes before holding at 4°C. Samples are then cleaned up for a second time using the AMPure XP beads. The amount of each library was determined using a Caliper™ GX II and 5 μL of each library at 2 nM was pooled sequenced using the Illumina MiSeq sequencer performing 2x 300bp paired reads.

**Sequencing analysis**

Trimmomatic was used to clean up and remove unwanted paired end forward and reverse reads from the raw Fastq files generated by the MiSEQ. This involved cutting the Nextera sequencing adaptors from the read, cutting bases from the start or the end of the read if they had a quality score lower then 3, removing reads when the average quality within a window of 4 base pairs drops below a quality score of 20 and removing any reads below 150 bp (for kappa chain reads) or 50bp for heavy chain reads were dropped. The program MiXCR [1] was then used to analyze the cleaned paired end forward and reverse files. Forward and reverse reads were t aligned to known mouse V(D)J genes using the default align command. From these alignments clonotypes were built using MiXCR’s assemble command based on the CDR3. For kappa chains, additional clonotypes was built based of the entire VJ transcript. These clonotypes were exported into .txt files using the export command. Further analysis into VDJ usage and
diversity was done using the R package tcR [2]. SHM analysis was done using in-house scripts to analyze the data generated by the best V (and J) sequences.

**Protein purification**

2A10 and 2A10 F\textsubscript{AB} fragment were produced from hybridomas by Genscript (Piscataway, NJ) and purified using Protein A before being resuspended in PBS with 0.02% Sodium Azide and shipped as a lyophilized powder. His tagged rCSP was expressed in \textit{E. coli} by Genscript (Piscataway, NJ) and purified from the supernatant of the cell lysate prior to being shipped in PBS with 10% Glycerol. Prior to ITC, CD and X-ray crystallographic analysis, 2A10, 2A10 F\textsubscript{AB} fragment and rCSP were purified by size-exclusion chromatography using a HiLoad 26/600 Superdex 200 column (GE Healthcare). 2A10 and the 2A10 F\textsubscript{AB} fragment were eluted in 25 mM TRIS pH 7.2, 100 mM NaCl. rCSP was eluted in 50 mM TRIS pH 7.2, 200 mM NaCl, and then transferred into 25 mM TRIS pH 7.2, 100 mM NaCl using a PD 10 desalting column (GE Healthcare) immediately prior to ITC experiments. Protein purity was confirmed using SDS-PAGE.

**Isothermal Titration Calorimetry.**

ITC experiments were performed using a Nano-ITC low volume calorimeter (TA Instruments) at 25 °C, with stirring at 250 rpm. Protein solutions were prepared in TRIS buffer and degassed before use. For the F\textsubscript{AB}-(NANP)\textsubscript{6} titration, 50 µM 2A10 F\textsubscript{AB} was titrated with 1 × 1.2 µL, then 20 × 2.0 µL injections of 400 µM (NANP)\textsubscript{6}. For the F\textsubscript{AB}-CSP titration, 8.1 µM 2A10 F\textsubscript{AB} was titrated with 1 × 1.2 µL, then 20 × 2.0 µL injections of 5.9 µM CSP. For the 2A10-CSP titration, 8.8 µM 2A10 was titrated with 1 × 1.2 µL, then 28 × 1.5 µL injections of 9.0 µM CSP. Data were analyzed in NanoAnalyze software (TA Instruments); the baseline-subtracted power was integrated, and the integrated heats were fit to the independent binding sites model to obtain the stoichiometry of the interaction (n), the association constant (K\textsubscript{a}), and the enthalpy of binding (ΔH). The background heat was included as an adjustable parameter in the model. 95%
confidence intervals for $n$, $K_a$ and $\Delta H$ were estimated by simulating 500 replicate datasets and fitting them to the independent binding sites model, as implemented in NanoAnalyze software.

**Protein crystallography**

The 2A10 F$_{AB}$ fragment was concentrated to either 15 or 24 mg/mL using 100 kDa centrifugal filter units (Millipore). High throughput crystallisation screens were set up at the C3 crystallization facility, CSIRO (Melbourne). Crystals formed in conditions of 2 M ammonium sulfate, 0.1 M trisodium citrate (condition A), pH 5.5 and 2 M ammonium sulfate, 0.1 M bis-tris chloride, pH 6.5 (condition B). Crystals were added to cryo buffer (reservoir conditions with addition of 35% glycerol) and flash-cooled in liquid nitrogen. X-ray diffraction data were collected the MX1 beamline of The Australian Synchrotron. Crystals from condition A crystallized in the $I4_132$ space group and diffracted to 2.52 Å with one F$_{AB}$ monomer in the asymmetric unit. Crystals from condition B crystallized in the $P4_322$ spacegroup and diffracted to 3.01 Å with three F$_{AB}$ monomers in the asymmetric unit. The structures were solved by molecular replacement using PHASER with the PDB ID: 2BRR as the search model for the heavy chain and PDB ID: 1EMT [3-5] as the search model for the light chain. Iterative cycles of manual model building and refinement were performed using Coot 0.8.2 [6]and phenix.refine [7]. Coordinates and structure factors were deposited in the Protein Data Bank with accession codes 5ZSF (condition A) and 5T0Y (condition B).

**Circular dichorism**

To determine the solution structure of the (NANP)$_6$ peptide, far-UV CD was utilized. The peptide was diluted in 100 mM NaCl, 25 mM Tris, pH 7.2 to a concentration of 0.2 mg/mL and scanned from 180 - 260 nm in 0.5 nm steps at 20 °C on an Applied Photophysics ChiraScan circular dichroism spectrometer. The structure of the peptide was predicted using the PEP-FOLD *de novo* peptide structure prediction server using default settings [8]. Only one low energy structure exhibited repeating order; this structure was then used to calculate the predicted CD spectrum using DichroCalc considering 2 backbone charge transitions, side chain
Interactions between neutralising antibodies and the circumsporozoite protein from Plasmodium falciparum transitions, and with an ab initio parameter set [9].

Computational modelling of the 2A10:(NANP)$_6$ interaction

The 2.52 Å structure of the 2A10 F$_{AB}$ fragment and the ab initio predicted structure of the (NANP)$_6$ peptide were used to model of the complex. First, an initial approximate model was generated using the GRAMM-X protein:protein docking web server [10], using default settings. The best model from the GRAMM-X output was then used as input for Rosetta SnugDock [11], again using default parameters. To model the full complex, the (NANP)$_6$ peptide structure was duplicated and partially superimposed, to extend it to 27 repeats. The complex between the 2A10 F$_{AB}$ fragment and an epitope was then overlaid in a repeating fashion along the repeating unit. Superposition was carried out using Pymol 1.8.2.3 (Schrodinger, LLC, USA).

For molecular dynamics simulations, both the (NANP)$_6$ peptide structure and each peptide in the 2A10:(NANP)$_6$ complex models were capped with acetyl and amine groups. The peptide was solvated in SPC water in a truncated dodecahedral box with a distance of 5 nm between periodic images to allow the peptide some flexibility before encountering its periodic image. Meanwhile, the 2A10:(NANP)$_6$ complex was solvated in a truncated dodecahedral box with a distance of 3 nm between periodic images. Sodium and chloride ions were added to both systems to make up 200 mM salt solutions. All simulations were performed with GROMACS 5.1.2 [12] in the GROMOS 54A7 forcefield [13] on an in-house compute server with 2 Nvidia Tesla K20 GPUs and 32 CPU cores. Long-range electrostatics were treated with the Particle Mesh Ewald method and the Van der Waals cut-off was set to 1.4 nm. The temperature was coupled to a virtual water bath at 300 K with a velocity rescale thermostat. The Berendsen barostat was used during equilibrations with a time constant of 2 fs; production runs were pressure coupled with a Parrinello-Rahman barostat with a time constant of 10 fs. A 2 fs time step was used throughout. Simulations were initially equilibrated with a 1 ns (500 000 steps) simulation in which alpha carbons were position restrained with a force constant of 1000 kJ mol$^{-1}$ nm$^{-1}$. The position restraints were relaxed over a series of 5 further 1 ns equilibrations with restraints of 500, 100, 50, 10 and 0 kJ mol$^{-1}$ nm$^{-1}$. Finally, production runs were performed
for 100 ns (50 000 000 steps). Equilibration and production simulations were performed in triplicate for each system.
Supplementary References


4.4 Paper 2: Avid binding by B cells to the *Plasmodium* circumsporozoite protein repeat suppresses responses to protective subdominant epitopes

4.4.1 Publication status

This manuscript presented in this section has not been published. However, it is available at BioRxiv at https://doi.org/10.1101/2020.01.12.903682. Supporting information is provided in Section 4.5.

4.4.2 Author's contribution

I designed and performed the surface plasmon resonance saturation experiment which helped to show that truncated CSP9 could bind fewer mAb 2A10 than the longer CSP27. I prepared protein samples used in these experiments and analyzed and interpreted the data. I contributed to the editing of the manuscript and produced the SPR figure.

4.4.3 Contributions from others

The majority of the work presented in this paper was performed by my co-authors, primarily from the Cockburn (ANU John Curtin School of Medical Research) research group.
Avid binding by B cells to the *Plasmodium* circumsporozoite protein repeat suppresses responses to protective subdominant epitopes

Fiona J. Lewis*1, Deepyan Chatterjee*1, Joe Kaczmarski2, Xin Gao1, Yeping Cai1, Hayley A. McNamara1, Henry J. Sutton1, Colin J. Jackson2, Ian A. Cockburn1†

1. Department of Immunology and Infectious Disease, John Curtin School of Medical Research, The Australian National University, Canberra, ACT 2602, Australia.

2. Research School of Chemistry, The Australian National University, Canberra, ACT 2602, Australia.

* these authors contributed equally to this manuscript

†To whom correspondence should be addressed: ian.cockburn@anu.edu.au

Keywords: B cells; Immunodominance; *Plasmodium falciparum*; malaria; circumsporozoite protein; vaccines.
Abstract:

Antibodies targeting the NANP/NVDP repeat domain of the *Plasmodium falciparum* circumsporozoite protein (CSPRepeat) can confer protection against malaria. However, it has also been suggested that this repeat domain exists as a decoy that distracts the immune system from mounting protective responses targeting other domains of CSP. Here we show that B cell responses to the repeat domain are indeed ~10 fold higher than responses to the N- and C- terminal regions of CSP after sporozoite immunization. We investigated the role of the number of CSPRepeat-specific naïve precursor B cells and high avidity binding by B cells in driving the immunodominance of the CSPRepeat. Using adoptive transfer of germline precursors specific for the CSPRepeat, we found that increasing precursor number did indeed increase the responses to the repeat region, but not to the detriment of responses to other epitopes. To investigate the role of avid binding by B cells to the CSPRepeat in driving immunodominance we generated CSP9: a truncated CSP molecule with just 9 NANP repeats. Compared to near full length CSP molecules, CSP9 induced lower BCR signalling in CSPRepeat-specific cells and induced stronger responses to non-repeat epitopes. Finally, we found mice immunized with CSP9 molecules were strongly protected against mosquito bite challenge. Collectively these data demonstrate that the CSPRepeat does function as an immunodominant decoy and that truncated CSP molecules may be a promising avenue for future malaria vaccines.
Significance Statement

Malaria kills approximately 420,000 individuals each year (1). Our best vaccine, RTS,S/AS01 is based on the circumsporozoite protein that coats the surface of the parasite. However, this vaccine is only partially protective. Here we show that responses to a repeat region in the circumsporozoite dominate the immune response. However, immunizing with a circumsporozoite protein with a shortened repeat region induces a more diverse immune response, which could be an avenue to make better malaria vaccines.
Introduction

Our most advanced malaria vaccine RTS,S/AS01 aims to induce antibodies that target the repeat region of the circumsporozoite protein (CSP), which covers the surface of the \textit{Plasmodium} sporozoite (2-4). The rationale for this approach comes from the observation that immunization with radiation attenuated sporozoites confers sterile protection against malaria, and that the humoral response induced by irradiated sporozoites is dominated by anti-CSP antibodies (5-8). Early studies demonstrated that monoclonal antibodies (mAbs) targeting the repeat regions of the CSP molecule (CSP\textsubscript{Repeat}) protected mice against challenge with the rodent parasite \textit{P. berghei} (9-11). More recently, human monoclonal antibodies targeting the \textit{P. falciparum} CSP\textsubscript{Repeat} have also been shown to be protective in preclinical mouse models (12-14).

Despite the protective capacity of CSP\textsubscript{Repeat} specific Abs, it has also been argued that the CSP\textsubscript{Repeat} is an immunodominant “decoy” distracting the immune system from making protective responses against other epitopes within CSP or other proteins on the sporozoite surface (15, 16). Evidence for the immunodominance of the CSP\textsubscript{Repeat} initially came from early studies which showed that a short (NANP)$_3$ peptide based on this domain could absorb most sporozoite binding activity of sera from hyperimmune individuals (8). In support of the concept that the responses to CSP\textsubscript{Repeat} are sub-optimal, large amounts of anti-CSP\textsubscript{Repeat} mAbs are required for protection in preclinical challenge models, while in RTS,S clinical trials protection requires very high amounts (>50 \textmu g/ml) of anti-CSP\textsubscript{Repeat} antibody (12-14, 17). In contrast, antibody responses to other regions of CSP are less well understood. One small epidemiological study associated increased levels of antibodies targeting a truncated CSP\textsubscript{Nterm} peptide with protection from clinical disease (18). Subsequently, a mouse mAb, 5D5,
targeting an epitope within the N terminus of CSP was found to be protective against sporozoite challenge (19). More recently, human mAbs targeting the junction between the CSPNterm and CSPRepeat were found to be protective (12, 13). Antibodies targeting the C-terminal domain CSP (CSPCterm) have been associated with protection by the RTS,S vaccine in clinical trials (20, 21), though individual mAbs targeting this domain have not been found to confer protection (14).

There has been increased interest in the factors that drive B cell immunodominance and how these can be manipulated for improved vaccination outcomes. Recent findings in influenza and HIV immunology have revealed the existence of broadly neutralising antibodies (bnAbs), however these target rare subdominant epitopes. In HIV it was recently shown that transferred B cells carrying a germline version of the bnAb VRC01 could be induced to compete successfully in germinal centers (GCs) if the number of naïve precursors was artificially increased or if stimulated with a high polyvalent antigen that bound the B cells with greater avidity (22). For influenza it has been shown that broadly neutralizing responses to the stem regions of haemagglutinin (HA) can be favored over responses to the immunodominant - but highly variable - head region by immunization with stem-only constructs, even if delivered alongside full length HA (23).

Given the highlighted roles for antigen polyvalency and precursor numbers in driving B cell immunodominance, we investigated whether these factors drive the CSPRepeat to be immunodominant. The finding that CSPRepeat can be bound avidly by B cells from a range of immunoglobulin gene families suggested that there may be high numbers of precursors for this domain (12, 13, 24, 25). More suggestively, we and others have shown that the CSP repeat can be bound by 6 or more specific antibodies (12, 24, 26, 27), and it has been
demonstrated that the repeat can crosslink multiple BCRs to enhance B cell signalling (28).

Accordingly, we tested the roles of these factors in driving the dominance of the Ab response against the CSP\textsubscript{Repeat} and determined if we could manipulate the immunodominance hierarchy to develop better vaccination protocols.
Results

The circumsporozoite protein repeat domain is immunodominant

To formally test the immunodominance of responses to the CSP\textsubscript{Repeat} over the CSP\textsubscript{Nterm} and CSP\textsubscript{Cterm} we immunized mice with irradiated Pb-PfSPZ parasites which carry a full length \((4NVDP/38NANP)\) \textit{P. falciparum} CSP gene in place of the endogenous \textit{P. berghei} CSP (Fig. S1) (29). At days 4, 7, 14 and 28 post-immunization sera were taken for antibody analysis by ELISA with domain-specific peptides, and spleens were taken for cellular analysis by flow cytometry. IgG responses to the CSP\textsubscript{Repeat} were significantly higher than responses to either of the other domains, with a significant response developing to the CSP\textsubscript{Cterm} only after 28 days (Fig. 1A).

One concern is that ELISA measurements of antibody responses to the different domains are not directly comparable. Therefore, we used tetramer probes to track the total numbers of B cells responding to each domain of CSP over time, and their phenotype by flow cytometry (Fig. 1B, Fig. S2A). The response to sporozoites is characterized by an early plasmablast (PB) response that wanes and leaves a prolonged GC reaction (24). Four days after immunization of mice with Pb-PfSPZ, the number of CSP\textsubscript{Repeat}\textsuperscript{+} CD138\textsuperscript{+} PBs was \(\sim\)10 fold higher than the number of CSP\textsubscript{Nterm}\textsuperscript{+} or CSP\textsubscript{Cterm}\textsuperscript{+} PBs (Fig. 1Ci). By day 7, a pronounced GC reaction developed and the number of CSP\textsubscript{Repeat}\textsuperscript{+} GL7\textsuperscript{+} B cells was \(\sim\)10 fold higher than responses to the other domains, which was sustained until day 28 (Fig 1Cii). We further analyzed the number of IgD\textsuperscript{−} IgM\textsuperscript{−} (Switched Ig; SwIg) CD38\textsuperscript{+} memory B cells and found that the immunodominance of the response to the CSP\textsubscript{Repeat} response extended into the memory phase (Fig 1Ciii).
Increasing anti-CSP precursor B cell number does not suppress responses to other antigens

A previous study highlighted the fact that increasing the number of precursors for an antigen led to a concomitant increase in the number of cells entering GCs; the same study also highlighted roles for antigen valency in allowing responses to successfully compete in the GC (22). To investigate the roles of precursor number and antigen valency we developed a system in which we could modulate the number of precursors or the valency of two competing antigens in the context of CSP. We achieved this by conjugating the 4-hydroxy-3-nitrophenyl (NP) acetyl-hapten to recombinant CSP via crosslinking on lysine. Since lysine residues are not found in the CSP repeat region, the NP-hapten would bind exclusively to the N and C terminal domains (Fig. S3A). For these experiments we used a slightly truncated CSP molecule carrying 27 repeats (3NVDP and 24 NANP), designated CSP27 (Fig S1). As proof of principle we found that immunization with CSP27 conjugated to 2 NP moieties (CSP27-NP2) was able to induce strong anti-NP and anti- CSP_repeat IgG responses (Fig. S3).

We were then able to modulate the number of precursors for CSP_repeat-specific B cells using Ig knockin Igh^2A10 B cells which carry the germline heavy chain of the CSP_repeat specific 2A10 antibody (McNamara et al. submitted). Similarly, we could modulate the number of NP-specific B cells using the established B1-8hi mouse system (30, 31). Finally, we could alter the valency of the response to NP or the repeat by conjugating more NP molecules per CSP, or by reducing the length of the CSP_repeat domain.

To determine the role of precursor number in driving immunodominance we adoptively transferred defined numbers of CSP_repeat tetramer^+^, CD45.1 Igh^2A10^ cells into CD45.2 C57BL/6 mice (Fig. S2B). Mice were then immunized with CSP27-NP2, IgG responses were
measured 7, 14 and 21 days post immunization, and the number of NP and CSP_repeat specific cells were quantified by flow cytometry on day 21 (Fig. 2A). We hypothesized that increased anti-CSP_repeat precursor number would not only increase the response to the CSP_repeat but also suppress the NP immune response. As expected, increasing the number of CSP_repeat-specific precursors increased the total number of CSP_repeat binding B cells and CSP_repeat-specific GC B cells responding to this antigen (Fig. 2B and C). Perhaps surprisingly, the magnitude of the antibody response was unaltered (Fig. 2D). However, there was no concomitant decrease in the overall B cells and GC B cell response to NP (Fig. 2B-C), and the magnitude of the IgG antibody response to NP was also unaffected by the addition of Ighg2A10 cells (Fig. 2E).

Finally, because we could distinguish our transferred cells from the endogenous response by the expression of CD45.1, we were able to determine that the endogenous response to CSP was not suppressed by the addition of enhanced number of germline precursors specific for CSP (Fig. 2F).

We also performed the converse experiment and altered the number of anti-NP precursors relative to the number of CSP_repeat specific cells via the addition of B1-8hi cells (Fig. S2C). Notably, B1-8hi cells differ from Ighg2A10 cells not only in their specificity but also because they carry the high affinity mature Ighv1-72 heavy chain that confers strong binding to NP. Nonetheless, in agreement with the previous experiment, antibody titres to the CSP_repeat were generally unaffected by the transfer of additional naive B-18hi cells (Fig S4A), while increasing the number of NP cells had no significant effect on the number of CSP_repeat specific B cells responding and becoming GC B cells (Figure S5B-D). However, in contrast to the previous experiment, the additional of B1-8hi cells did not increase the overall magnitude of the antigen specific B cell and GC response to NP itself; rather, the transferred high affinity B1-8hi B cells displaced the endogenous cells from the response though this
effect rapidly saturated after the transfer of just $1 \times 10^4$ B1-8^{hi} cells (Fig S4E). Again, in contrast to the previous experiment, this displacement resulted in higher titres of antibodies targeting NP overall, perhaps due to the high affinity of the transferred cells (Fig. S4F).

Reducing the valency of immunodominant antigens allows subdominant responses to expand

Given the repeating nature of the CSP Repeat we next tested whether the ability of long CSP molecules to crosslink multiple B cell receptors (BCRs) might drive the immunodominance of this domain. Accordingly, we developed a construct that carried just 9 NANP repeats (CSP9; Fig. S1). Using a surface plasmon resonance saturation experiment, we found that CSP9 could only bind 2-3 2A10 antibodies, compared to the 5-6 bound by CSP27, which is in line with previous structural and biophysical data from our laboratory (Fig. 3A).

Importantly, this reduced binding corresponded to a reduction in BCR signalling as calcium fluxes were lower when Igh^{2A10} cells were pulsed with CSP9 compared to CSP27 (Fig. 3B-C).

To test whether the reduction in BCR signalling by CSP9 corresponded to a reduction in immunodominance of the CSP Repeat we compared responses to CSP27-NP2 and NP-haptenated CSP9 (CSP9-NP2). CSP9-NP2 had significantly elevated NP-specific IgG compared to NP2-CSP27, particularly on days 14 and 21 post-immunisation (Fig. 3D).

Analysis of the cellular NP and CSP Repeat responses via flow cytometry at day 21 revealed that truncation of the CSP Repeat not only reduced the number of cells responding to the CSP Repeat, but also allowed for a significant increase in the number of NP-specific total B cells and GC cells supporting the increase in antibody titres (Fig. 3E-G). Moreover, the CSP Repeat IgG titres were significantly decreased for NP2-CSP9 immunised mice compared to
NP2-CSP27 (Fig. 3H). Overall, this data supports the importance of valency in determining immunodominance as, decreasing the repeat length shifted the response away from the CSP\textsubscript{Repeat} and towards NP.

We next performed the converse experiment and compared responses to CSP27, CSP27-NP2, CSP27-NP6 and CSP27-NP10. In these conditions the length of the CSP\textsubscript{Repeat} antigen is fixed but the valency of NP varies. In agreement with the previous finding, increasing the NP:CSP ratio not only increased the response to NP (Fig S5A) but also decreased the level of antibodies to the CSP\textsubscript{Repeat} (Fig S5B). Increasing the NP:CSP-ratio resulted in a switch in the immunodominance hierarchy; in mice immunized with CSP27-NP2, the CSP\textsubscript{Repeat} response was immunodominant, while the NP response dominated the CSP\textsubscript{Repeat} response upon immunization with CSP27-NP10 (Fig S5C-E). Collectively these data indicate a powerful role for antigen valency in driving the immunodominance of repeating antigens.

Repeat-truncated P\textsubscript{f}CSP molecules induce diverse antibody responses that protect against parasite challenge

One prediction of our data is that immunization CSP9 will induce stronger responses to the CSP\textsubscript{Nterm} and CSP\textsubscript{Cterm} compared to CSP27. Moreover, if these non-CSP\textsubscript{Repeat} responses have anti-parasitic effect then immunization with CSP9 should be more protective than immunization with CSP27. Accordingly, we immunized mice 3 times at 5-week intervals with CSP9 and CSP27 formulated in alum before challenging mice with Pb-PfSPZ via mosquito bites 5 weeks after the final immunization (Fig. 4A). We also included an additional group which were immunized with CSP9\textsubscript{NVDP} - a recombinant protein which also had a 9-mer repeat but included 3 NVDP repeats (Fig. S1) - as it has been suggested that
(NANPNVDP)_n binding antibodies might confer superior protection to pure (NANP)_n binding antibodies (13).

Overall, IgG responses to PfCSP (as measured via ELISA on CSP27 coated plates) were similar in magnitude across all immunized groups (Fig. 4B). Because differences between immunized mice and control mice are large (as are differences between pre-immune and post-immune sera), but not of particular interest, this group was removed from subsequent analysis as was the pre-immune timepoint to avoid skewing the statistical models.

Examination of the specificity of the IgG responses at the domain level revealed significant differences in responses to the different immunogens: CSP9 and CSP9NVDP induced significantly stronger IgG responses to CSPNterm and CSPCterm compared to CSP27 while responses to the repeat were similar (Figure 4C and D). We also performed more specific ELISAs with peptides corresponding to the regions bound by 5D5 and the CSPNterm/CSPRepeat junction (Fig. S6A). We found that the truncated CSP9 and CSP9NVDP induced stronger responses to the 5D5 peptide which lies within the CSPNterm (Fig S6B) but that responses to the junction peptide were similar across all immunogens (Fig. S6C). Surprisingly, while CSPRepeat specific antibody responses were initially lower in CSP9 immunized mice compared to CSP27 immunized mice, the repeat specific responses were similar after the third dose (Fig. 4E). Nonetheless, the ratio of antibody titres to the CSPNterm and CSPCterm domains to the CSPRepeat domains were significantly higher in the CSP9 immunized mice compared to CSP27 immunized mice at the time of challenge (Fig 4F and G).

Forty-two hours after mosquito bite challenge, livers were excised from the mice and the parasite burden measured by RT-PCR. Both CSP9 and CSP27 induced significant reductions in the mean parasite burden in the liver, whilst CSP9NVDP did not (p = 0.069; Fig. 4H).
Overall 10/11 (91%) control mice had detectable parasite 18SRNA, compared to only 5/15 (33%) CSP9 immunized mice, which was statistically significant (p=0.005 by Fisher’s exact test; Fig. 4I). In both CSP27 and CSP9NVDP, 9/15 (60%) mice had detectable parasite RNA, which was not significantly different the control group (p=0.17 by Fisher’s exact test; Fig. 4I). The failure of CSP9NVDP to induce better protection than the CSP9 was surprising to us, but did not reflect a failure to induce NVDP binding antibodies as these antibodies were significantly higher in both CSP27 (which includes NVDP repeats) and CSP9NVDP immunized mice than in the CSP9 immunized mice (Fig. S6D and E). Overall, these data indicate that strategies which reduce the immunodominance of the CSP repeat can induce more diverse antibody responses and confer superior protection to vaccination strategies which utilise full length or near full length CSP.
Discussion

Here we show that the primary driver of the immunodominance of the CSP\textsubscript{Repeat} over other domains within CSP is the avidity of the binding between long repeats and BCRs on the surface of antigen specific B cells. Reducing the valency of the CSP\textsubscript{Repeat} allows the development of stronger B cell and antibody responses to other epitopes. To demonstrate the importance of this observation for vaccination we used truncated CSP molecules to immunize against malaria in a pre-clinical model. We found that mice immunized with truncated CSP developed stronger responses to the CSP\textsubscript{Nterm} and CSP\textsubscript{Cterm} domains, both of which have been associated with protective antibody responses. Moreover, mice immunized with truncated CSP molecules were protected against live parasite challenge, and the magnitude of this protection was greater than in mice immunized with nearly full length CSP, though this difference did not reach statistical significance.

Our findings shed light on the key drivers of B cell immunodominance. There have been comparatively few studies on B cell immunodominance, but T cell immunodominance studies have highlighted critical roles for TCR-peptide MHC affinity and the number of naïve precursors specific for a particular antigen (reviewed in (32)). A common strategy for enhancing antibody responses is to generate polyvalent antigens, such as virus-like particles, to enhance immunity (33). Accordingly, we investigated how the multivalent nature of the CSP\textsubscript{Repeat} affects the response to this epitope. Our finding that reducing the length of the CSP\textsubscript{Repeat} allows responses to other antigens to develop suggests that not only does the long repeat drive a large response to this antigen itself, but also allows it to supress other responses. The discovery of this “immunodomination” effect provides direct support for the decoy hypothesis, though the exact mechanism for this effect is unclear. Our observation that
CSP27 drives stronger BCR signalling than CSP9 suggests that this antigen may drive stronger expansion of the B cell response at the outset. An alternative – non-mutually exclusive hypothesis - is that CSP molecules carrying long repeats may be readily taken up by CSPRepeat specific B cells, allowing these B cells to outcompete CSPNterm and CSPCterm specific B cells for T cell help (34, 35). In agreement with this, it has been previously shown that competition for restricted T cell help can suppress responses to rare or subdominant epitopes (36).

We also investigated the roles of the number of antigen-specific precursors in determining the magnitude of the resulting immune response. The number of naïve precursors specific for a particular epitope was identified as a possible driver of CD4^+ and CD8^+ T cell immunodominance in mice in studies with peptide immunization or vesicular stomatitis virus infection (37, 38). However, other studies in humans (with HIV and hepatitis B virus) and mice (with influenza A virus) have not reported as strong a relationship as initially reported (39-41), perhaps due to compensation by other factors such as the affinity of the T cell receptor-MHC peptide interaction (32). In agreement with a previous study on the relationship between B cell precursor number and the subsequent GC response (22), we found that artificially increasing the number of precursors by transferring knock-in germline B specific for the CSPRepeat cells did increase the magnitude of the GC response to this epitope. However, surprisingly, this enhanced response did not occur at the expense of responses to other epitopes and did not result in a significantly enhanced antibody response. One reason for the lack of immunodomination after the transfer of large numbers of CSPRepeat specific B cells may be that precursor B cells for the competing antigen in this system (NP) are in excess. In agreement with this, NP specific B cells are reported to be present at high
frequency (~1/4000) in C57BL/6 mice (42). Thus, it will thus be important to determine if the
same effect is seen when the number of precursors for the subdominant epitope is limiting.

Interest in developing “universal” vaccines for viruses such as influenza and HIV has led to
increased focus on manipulating immunodominance hierarchies. This is because responses to
conserved influenza epitopes (such as the HA stem) or conserved targets of bnAbs are
subdominant. For HIV there is strong interest in developing multivalent immunogens
targeting rare bnAb germline precursors to enhance the frequency of these cells (43, 44). For
influenza, a possible vaccine strategy is to develop HA immunogens lacking the
immunodominant, but variable head domain (23). One approach that might be possible, based
upon studies in model systems would be to delete B cells specific for non-protective epitopes
by injecting pure epitopes not linked to a T cell epitope (45). For CSP based malaria vaccines
it is probably not desirable to remove the responses to the repeat altogether as antibodies
targeting this domain are clearly protective, however a rebalancing of the immune response
may be desirable. The current RTS,S vaccine may benefit from having a truncated repeat (of
only 18 NANP moieties), though – critically - it lacks the CSP_{Nterm} domain and does not have
any NVDP repeats (46). Perhaps surprisingly, we did not find that mice which mounted
stronger responses to the NANPNVDP repeats were better protected than those which had
responses more focussed on pure NANP repeats. This is despite the fact that mAbs that are
strongly cross-reactive between both repeat motifs have been shown to protect against
Plasmodium infection (13).

Collectively, our results provide insights into the factors that drive the immunodominance of
different B cell epitopes; notably, that repeat epitopes can induce larger responses at the
expense of subdominant epitopes. Our results suggest that truncated CSP molecules carrying
all domains of the protein, may be a promising approach to the generation of a next generation CSP based vaccine.
Materials and Methods

Full details of materials and methods used in this study are given in the supplementary information.

Ethics statement

All animal procedures were approved by the Animal Experimentation Ethics Committee of the Australian National University (Protocol numbers: A2016/17; 2019/36). All research involving animals was conducted in accordance with the National Health and Medical Research Council's Australian Code for the Care and Use of Animals for Scientific Purposes and the Australian Capital Territory Animal Welfare Act 1992.

Mice and Parasites

C57BL/6 mice, B1-8 mice (31) and Ighg2A10 (McNamara et al. submitted) were bred in-house at the Australian National University. Mice were immunized IV with 5 x 10^4 irradiated (15kRad) Pb-PfSPZ (29) dissected by hand from the salivary glands of Anopheles stephensi mosquitoes.

Proteins and Immunizations

For immunization with CSP27, CSP9 or CSP9NVDP, or CSP27-NP conjugates 30 µg protein (or as described in the relevant figure legend) was emulsified in Imject™ Alum according to the manufacturer’s instructions (ThermoFisher Scientific) and delivered intra-peritoneally.

Flow Cytometry
Single cell preparations of lymphocytes were isolated from the spleen of immunized mice and were stained for flow cytometry or sorting by standard procedures. Cells were stained with antibodies as outlined in Table S1 and CSP domain specific tetramers conjugated to PE or APC. Tetramers were prepared in house by mixing biotinylated (NANP)_9 peptide with streptavidin conjugated PE or APC (Invitrogen) in a 4:1 molar ratio. Flow-cytometric data was collected on a BD Fortessa flow cytometer (Becton Dickinson) and analyzed using FlowJo software (FlowJo).

**ELISA**

Binding of 2A10 antibody variants was determined in solid phase ELISA. Briefly, Nunc Maxisorp Plates (Nunc-Nucleon) were coated overnight with 1μg/ml streptavidin followed by binding of biotinylated peptide for 1 hour. After blocking with 1% BSA, serial dilutions of the antibodies were incubated on the plates for 1 hour and after washing, incubated with HRP conjugated anti-IgG antibodies (KPL). For the analysis of sera from immunized mice data were expressed as the area under the curve (AUC).

**Surface plasmon resonance**

Surface plasmon resonance saturation experiments were performed on a Biacore 8K instrument (GE Healthcare) at 25 °C using a Series S Sensor Chip NTA (GE Healthcare). CSP27 and CSP9 were immobilized on separate channels on the sensor chip surface as per the manufacturer’s recommendations. A saturating solution of mAb 2A10 was then passed over the chip for 400 s, using a flow rate of 30 μl/min, followed by a 400 s dissociation period. The binding stoichiometry (n, molar ratio of antibody to antigen in the complex under saturating concentrations of mAb 2A10) was estimated as described previously (47).
Quantitation of parasite RNA

*P. berghei* 18S rRNA was quantified from the livers of mice 42 hours after challenge via the bites of Pb-PfSPZ infected *Anopheles stephensi* mosquitoes as described previously (48).

Statistical analysis

Statistical analysis was performed in GraphPad Prism for simple analyses without blocking factors; all other analyses was performed in R (The R Foundation for Statistical Computing) with details of statistical tests in the relevant figure legends.
Acknowledgments:

We thank Michael Devoy, Harpreet Vohra and Catherine Gillespie of the Imaging and Cytometry Facility at the John Curtin School of Medical Research for assistance with flow cytometry and multi-photon microscopy. We also thank Theresa Neeman of the ANU statistical consulting unit for assistance with statistical analysis of the data. This work was funded by the Bill and Melinda Gates Foundation (OPP1151018) and the National Health and Medical Research Council (GNT1158404). Authors declare no competing interests; All data is available in the main text or the supplementary materials.
Figure Legends

Fig. 1. Responses to the CSP\textsubscript{Repeat} are immunodominant over responses to other domains. C57BL/6 mice were immunized with 5 x 10\textsuperscript{4} irradiated Pb-PfSPZ sporozoites, blood and spleens were taken at 4, 7, 14 and 28 days post-immunization for analysis by ELISA and flow cytometry using probes specific for each domain of CSP (CSP\textsubscript{Cterm}, CSP\textsubscript{Repeat} and CSP\textsubscript{Nterm}). (A) IgG responses to each domain measured by ELISA, data shown as area under the curve. (B) Representative flow cytometry plots from a single mouse at the day 7 timepoint showing the gating of PBs, GC B cells and SwIg Mem for antigen specific IgD- B cells identified using tetramers specific for the CSP\textsubscript{Cterm}, CSP\textsubscript{Repeat} and CSP\textsubscript{Nterm}; values are percentages (C) Quantifications of the absolute numbers of IgD- B cells using NANP, R1+ and N\textsubscript{R1-91} antibody panels. (C) Absolute numbers of (i) Plasmablasts, (ii) GC B cells and (iii) SwIg memory cells in each mouse for each antigen (domain). Data for all panels are represented as mean ± SD pooled from two independent experiments (n=3-5 mice/timepoint/experiment); all data were analyzed via 2-way ANOVA, with experiment and mouse included in the model as fixed factors, ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test and significant pairwise comparisons are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001.

Fig. 2. Increasing CSP\textsubscript{Repeat}-specific precursor number does not alter immunodominance. 0, 1x10\textsuperscript{4}, 3x10\textsuperscript{4}, or 9x10\textsuperscript{4} of CD45.1 Ig\textsubscript{h2A10} cells were adoptively transferred into C57BL/6 mice followed by immunization with 30 µg CSP27-NP2 in alum. Sera were taken on days 7, 14 and 21 and spleens analyzed 21 days post-immunization. (A) Representative flow cytometry plots showing gating of total IgD\textsuperscript{+} and GC B cells specific for
NP or the CSP\textsubscript{Repeat}; values are percentages (B) Absolute numbers of NP probe\textsuperscript{+} and CSP\textsubscript{Repeat} tetramer\textsuperscript{+} IgD\textsuperscript{−} B cells. (C) Absolute numbers of NP probe\textsuperscript{+} and CSP\textsubscript{Repeat} tetramer\textsuperscript{+} GC B cells. (D) Total IgG response to CSP\textsubscript{Repeat} measured via (NANP)\textsubscript{9} ELISA. (E) Total IgG response to NP measured via NP(14)BSA ELISA. (F) Absolute numbers of CSP\textsubscript{Repeat} tetramer\textsuperscript{+} CD45.1\textsuperscript{+} Igh\textsubscript{g2A10} and CD45.1\textsuperscript{−} endogenous cells. Data are represented as mean ± SD pooled from two independent experiments (n≥3 mice/group/experiment); all data were analyzed via 2-way ANOVA, with experiment and mouse included in the model as fixed factors. ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001.

**Fig. 3. Decreasing the valency of the CSP\textsubscript{Repeat} alters the immunodominance hierarchy**

Recombinant CSP9 was purified and conjugated to NP at a 1:2 ratio to generate CSP9-NP2, mice were immunized with either CSP9-NP2 (23 µg) or CSP27-NP2 (30 µg) in alum; sera were taken on days 7, 14 and 21 and spleens analyzed 21 days post-immunization. (A) Approximate binding stoichiometry of the 2A10:Ag complex formed when a saturating concentration (2 µM) of mAb 2A10 was passed over immobilized CSP27 or CSP9; data shows mean ± SD of 2 technical replicates (n=2). (B) Calcium flux of sorted Igh\textsubscript{g2A10} cells incubated with Indo-1 dye and stimulated with CSP27, CSP9 or OVA-HEL, and the Ca\textsuperscript{2+} flux measured; near the end of the acquisition Ionomycin was added as a positive control; data shows the mean ± SD of 3 experimental replicates with summary data. (C) Summary data from B analysed via pairwise t-test, mean ± SD shown. (D) Total IgG response to NP measured via NP(14)BSA ELISA. (E) Representative flow cytometry plots showing gating of total IgD\textsuperscript{−} and GC B cells specific for NP or the CSP\textsubscript{Repeat}; values are percentages. (F)
Absolute numbers of NP probe$^+$ and CSP$^{\text{Repeat}}$ tetramer$^+$ IgD$^-$ B cells. (G) Absolute numbers of NP probe$^+$ and CSP$^{\text{Repeat}}$ tetramer$^+$ GC B cells. (H) Total IgG response to CSP$^{\text{Repeat}}$ measured via (NANP)$_9$ ELISA. Data for panels D-H are represented as mean ± SD pooled from two independent experiments (n≥4 mice/group/experiment); these data were analyzed via 2-way ANOVA, with experiment and mouse included in the model as fixed factors. ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001.

Fig. 4. Immunization with CSP9 induces more diverse antibody responses and confers protection against sporozoite challenge. C57BL/6 mice were immunized three times at 5 weekly intervals with 30 µg CSP9, CSP9$^{\text{NVPD}}$, CSP27 or alum only control, mice were challenged via mosquito bite with Pb-PfSPZ sporozoites and parasite burden measured by RT-PCR; blood was drawn prior to each immunization and challenge for analysis of the antibody response. (A) Schematic of the experiment. (B) Overall IgG responses to CSP27. (C) Overall IgG responses to CSP$^{\text{Nterm}}$. (D) Overall IgG responses to CSP$^{\text{Repeat}}$. (E) Overall IgG responses to CSP$^{\text{Cterm}}$. Data in B-E was from experiments with 5 mice/experiment/group, analysed via 2-way ANOVA with experiment and mouse as blocking factors, ANOVA p values are listed below or adjacent to each graph; pairwise comparisons between groups (averaged over time) were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001. (F) Ratio of the IgG response to CSP$^{\text{Nterm}}$:CSP$^{\text{Repeat}}$ at week 15 for the different immunized groups. (G) Ratio of the IgG response to CSP$^{\text{Cterm}}$:CSP$^{\text{Repeat}}$ at week 15 for the different immunized groups. Data for panels F and G were analyzed via one-way ANOVA with with experiment and mouse as
blocking factors. (H) Parasite 18S RNA in the livers of mice 42 hours post challenge; data pooled from 3 experiments with 3-5 mice/experiment/group and analysed via Kruskal-Wallis test with Dunn’s multiple comparisons test. (I) Proportion of mice from (H) with no detectable parasite RNA in each group, pairwise comparisons were made via Fisher’s exact test.
References and notes:


Figure 1

a) **anti-CSP IgG**

- ELISA Response (AUC)
  - Day 4, Day 7, Day 14, Day 28
  - Anti-CSP Normal
  - Anti-CSP Repeat
  - Anti-CSP Nterm

- p(Antigen) < 0.001
  - p(Day) < 0.001
  - p(Ag:Day) < 0.001

b) **Day 7: Gated on IgD**

- CD19+ B cells and CD138+ Plasmablasts

- PBs
  - CSP Nterm Tet-PE
  - CD138-PECy7
  - FSC-A
  - IgM mem/naïve
  - SwIg Memory

- GC B cells
  - CSP Cterm Tet-APC
  - CD138-A700
  - CD38-A405
  - GL7-A405
  - SSC-A

- CSP Repeat Tet-APC
  - CSP Repeat Tet-PE
  - CSP Nterm Tet-APC

- CSP Nterm Tet-PE
  - CSP Repeat Tet-APC
  - CD38-A700
  - GL7-A405
  - SSC-A

- CSP Cterm Tet-APC
  - CSP Cterm Tet-PE
  - CD38-A700
  - GL7-A405
  - SSC-A

- CSP Repeat Tet-APC
  - CSP Repeat Tet-PE
  - CD38-A700
  - GL7-A405
  - SSC-A

- CSP Nterm Tet-APC
  - CSP Nterm Tet-PE
  - CD38-A700
  - GL7-A405
  - SSC-A

C i) **Plasmablasts**

- Ag-specific PBs/spleen
  - p(Antigen) < 0.001
  - p(Day) < 0.001
  - p(Ag:Day) < 0.001

C ii) **GC B cells**

- Ag-specific GC B cells/spleen
  - p(Antigen) < 0.001
  - p(Day) < 0.001
  - p(Ag:Day) < 0.001

C iii) **SwIg Mem B cells**

- Ag-specific SwIg Mem B cells/spleen
  - p(Antigen) < 0.001
  - p(Day) < 0.001
  - p(Ag:Day) < 0.001

Fig. 1. **Responses to the CSP Repeat are immunodominant over responses to other domains**

C57BL/6 mice were immunized with 5 x 10⁴ irradiated Pb-PfSPZ sporozoites, blood and spleens were taken at 4, 7, 14 and 28 days post-immunization for analysis by ELISA and flow cytometry using probes specific for each domain of CSP (CSP Cterm, CSP Repeat, and CSP Nterm). (A) IgG responses to each domain measured by ELISA, data shown as area under the curve. (B) Representative flow cytometry plots from a single mouse at the day 7 timepoint showing the gating of PBs, GC B cells and SwIg Mem for antigen specific IgD- B cells identified using tetramers specific for the CSP Cterm, CSP Repeat and CSP Nterm; values are percentages (C) Quantifications of the absolute numbers of IgD- B cells using NANP, R1+ and N81-91 antibody panels. (C) Absolute numbers of (i) Plasmablasts, (ii) GC B cells and (iii) SwIg memory cells in each mouse for each antigen (domain). Data for all panels are represented as mean ± SD pooled from two independent experiments (n=3-5 mice/timepoint/experiment); all data were analyzed via 2-way ANOVA, with experiment and mouse included in the model as fixed factors, ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test and significant pairwise comparisons are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001.
Fig. 2. Increasing CSP<sub>repeat</sub>-specific precursor number does not alter immunodominance. 0, 1x10⁴, 3x10⁴, or 9x10⁴ of CD45.1 Igh<sub>2A10</sub> cells were adoptively transferred into C57Bl/6 mice followed by immunization with 30 µg CSP27-NP2 in alum. Sera were taken on days 7, 14 and 21 and spleens analyzed 21 days post-immunization. (A) Representative flow cytometry plots showing gating of total IgD<sup>-</sup> and GC B cells specific for NP or the CSP<sub>repeat</sub>; values are percentages (B) Absolute numbers of NP probe<sup>+</sup> and CSP<sub>repeat</sub> tetramer<sup>+</sup> IgD<sup>-</sup> B cells. (C) Absolute numbers of NP probe<sup>+</sup> and CSP<sub>repeat</sub> tetramer<sup>+</sup> GC B cells. (D) Total IgG response to CSP<sub>repeat</sub> measured via (NANP)<sub>9</sub> ELISA. (E) Total IgG response to NP measured via NP(14)BSA ELISA. (F) Absolute numbers of CSP<sub>repeat</sub> tetramer<sup>+</sup> CD45.1<sup>+</sup> Igh<sub>2A10</sub> and CD45.1<sup>-</sup> endogenous cells. Data are represented as mean ± SD pooled from two independent experiments (n=3 mice/group/experiment); all data were analyzed via 2-way ANOVA, with experiment and mouse included in the model as fixed factors. ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001.
**Figure 3**

*Fig. 3. Decreasing the avidity of the CSP\textsubscript{Repeat} alters the immunodominance hierarchy* Recombinant CSP9 was purified and conjugated to NP at a 1:2 ratio to generate CSP9-NP2; mice were immunized with either CSP9-NP2 (23 µg) or CSP27-NP2 (30 µg) in alum; sera were taken on days 7, 14 and 21 and spleens analyzed 21 days post-immunization. (A) Approximate binding stoichiometry of the 2A10:Ag complex formed when a saturating concentration (2 µM) of mAb 2A10 was passed over immobilized CSP27 or CSP9; data shows mean ± SD of 2 technical replicates (n=2). (B) Calcium flux of sorted Igh\textsubscript{g2A10} cells incubated with Indo-1 dye and stimulated with CSP27, CSP9 or OVA-HEL, and the Ca\textsuperscript{2+} flux measured; near the end of the acquisition Ionomycin was added as a positive control; data shows the mean ± SD of 3 experimental replicates with summary data. (C) Summary data from B analysed via pairwise t-test, mean SD shown. (D) Total IgG response to NP measured via NP(14)BSA ELISA. (E) Representative flow cytometry plots showing gating of total IgD\textsuperscript{-} and GC B cells specific for NP or the CSP\textsubscript{Repeat}; values are percentages. (F) Absolute numbers of NP probe\textsuperscript{*} and CSP\textsubscript{Repeat} tetramer\textsuperscript{*} IgD\textsuperscript{-} B cells. (G) Absolute numbers of NP probe\textsuperscript{*} and CSP\textsubscript{Repeat} tetramer\textsuperscript{*} GC B cells. (H) Total IgG response to CSP\textsubscript{Repeat} measured via (NANP)\textsubscript{9} ELISA. Data for panels D-H are represented as mean ± SD pooled from two independent experiments (n=4 mice/group/experiment); these data were analyzed via 2-way ANOVA, with experiment and mouse included in the model as fixed factors. ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001.
**Figure 4**

C57BL/6 mice were immunized with three times at 5 weekly intervals with 30 µg CSP9, CSP9NVDP, CSP27 or Alum only control, mice were challenged via mosquito bite with PbPfSPZ sporozoites and parasite burden measured by RT-PCR; blood was drawn prior to each immunization and challenge for analysis of the antibody response. (A) Schematic of the experiment. (B) Overall IgG responses to CSP27. (C) Overall IgG responses to CSPNterm. (D) Overall IgG responses to CSPCterm. (E) Overall IgG responses to CSPRepeat. Data from B-E was from experiments with 5 mice/experiment/group, analysed via 2-way ANOVA with experiment and mouse as blocking factors, ANOVA p values are listed below or adjacent to each graph; pairwise comparisons between groups (averaged over time) were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001. (F) Ratio of the IgG response to CSPNterm:CSPRepeat at week 15 for the different immunized groups. (G) Ratio of the IgG response to CSPCterm:CSPRepeat at week 15 for the different immunized groups. Data for figure F and G was analyzed via one-way ANOVA with with experiment and mouse as blocking factors. (H) Parasite 18S RNA in the livers of mice 42 hours post challenge; data pooled from 3 experiments with 3-5 mice/experiment/group and analysed via Kruskal-Wallis test with Dunn’s multiple comparisons test. (I) Proportion of mice from (H) with no detectable parasite RNA in each group, pairwise comparisons were made via Fisher’s exact test.

**Fig. 4.** Immunization with CSP9 induces more diverse antibody responses and confers protection against sporozoite challenge. C57BL/6 mice were immunized with three times at 5 weekly intervals with 30 µg CSP9, CSP9NVDP, CSP27 or Alum only control, mice were challenged via mosquito bite with PbPfSPZ sporozoites and parasite burden measured by RT-PCR; blood was drawn prior to each immunization and challenge for analysis of the antibody response. (A) Schematic of the experiment. (B) Overall IgG responses to CSP27. (C) Overall IgG responses to CSPNterm. (D) Overall IgG responses to CSPCterm. (E) Overall IgG responses to CSPRepeat. Data from B-E was from experiments with 5 mice/experiment/group, analysed via 2-way ANOVA with experiment and mouse as blocking factors, ANOVA p values are listed below or adjacent to each graph; pairwise comparisons between groups (averaged over time) were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001. (F) Ratio of the IgG response to CSPNterm:CSPRepeat at week 15 for the different immunized groups. (G) Ratio of the IgG response to CSPCterm:CSPRepeat at week 15 for the different immunized groups. Data for figure F and G was analyzed via one-way ANOVA with with experiment and mouse as blocking factors. (H) Parasite 18S RNA in the livers of mice 42 hours post challenge; data pooled from 3 experiments with 3-5 mice/experiment/group and analysed via Kruskal-Wallis test with Dunn’s multiple comparisons test. (I) Proportion of mice from (H) with no detectable parasite RNA in each group, pairwise comparisons were made via Fisher’s exact test.
4.5 Supplementary Information for “Avid binding by B cells to the *Plasmodium* circumsporozoite protein repeat supresses responses to protective subdominant epitopes”
Supplementary Information

This supplement contains:

- Supplementary Materials and Methods
- Table S1: Antibodies used for flow cytometry in this study
- Supplementary References
- Figures S1-S6
Supplementary Materials and Methods

Lymphocyte isolation and flow cytometry

Spleens were collected from mice after cervical dislocation, and single cell suspension was prepared after passing through a 70 µM cell strainer into FACS buffer. Cells were briefly blocked for 30 minutes using 1µg/mL Streptavidin and 10 µg/mL TruStain fcX antibody diluted in FACS buffer. An antibody master mix was prepared in FACS buffer to stain the surface antigens of cells as per standard protocol using antibodies and B cell tetramers (prepared in house) listed in Table S1 for 30 minutes in dark. Cells were then lysed using ACK lysis buffer and washed twice before staining with 1% 7AAD as Live/dead dye. Samples were acquired with a BD Fortessa or X20 flow cytometer, and data was analysed using FlowJo software. The universal gating strategy is shown Figure S2, after which the further gating to study the B cells of interest is provided in the results.

Measuring antigen-specific sera IgG titres using ELISA

Blood was collected from mice via either tail vein or retro-orbital bleeds and left to clot overnight at 4 °C, and the following day sera was collected by spinning the blood at 2000 g for 15 minutes at 4 °C. Next, Maxisorp Nunc-Nucleon 96 flat bottom plates were coated with 1µg/ml streptavidin overnight at 4 °C in dark. The following day, plates were washed in wash buffer consisting 0.05 % tween20 in PBS, and were incubated for 1 hour with different biotinylated peptides listed in pertaining to diverse sections of CSP protein. The peptide sequences were as follows, CSP_{Nterm}: Biotin-Ahx-NKNNQGNGQGHNMPNDPNRNVDENANANSAVKNNNNEEPSDKHIKEYLNKIQNSS

31  TEWSPCSVTCGNGIQVRI KPGSANKPKDELDYANDIEKKICKMEKCS; CSP_{Repeat}: Biotin-Ahx-NANPNANPNANPNANPNANPNANPNANPNANPNANP, and CSP_{Cterm}:
Biotin-Ahx- QEYQCYGSSSNTRVLNELNYDNAGTNLYNELEMNYYG

KQENWYSLLKNSRLGENDDGNEDNEKLRKPKHKKL KQPADG, additional peptides also carrying a Biotin-Ahx linker are described in Figure S6. Wells were blocked for 1 hour before incubating with initial sera (dilution of 1/100 times, followed by 1 in 4 serial dilutions down the plate) for one hour. After washing, anti IgG detection antibody conjugated to HRP was diluted 1/2000 times and was incubated for one hour. The plates were washed, then developed with Peroxidase Substrate Kit for 15 minutes, and the reaction was stopped using 50 µL/ well stop solution, consisting of 10% SDS in PBS. Absorbance at 405 nm (A405) was measured using an Infinite PRO Tecan plate reader, and the data was expressed as area under the curve (AUC) calculated in Prism 7 from the log(dilution) on the x axis and the A405 on the y axis, fitting a sigmoidal curve.

Conjugation of NP to CSP

CSP27 or CSP9 was initially concentrated using an Amicon ultra-15 centrifugal filter unit (10 kD molecular weight cut-off, MWCO). Briefly, the filter was washed by spinning 10 mL of MQ at 4000 g for 20 minutes at 4 °C then discarding the flow through and any retained water. 2 mg of CSP27 or 1 mg CSP9 was added to the membrane and made up to 10 mL in 3% NaHCO₃, then spun at 4000 g for 60 minutes at 4 °C. The flow-through was discarded and the retained protein made up to 10 mL in 3% NaHCO₃, then respun at 4000 g for 60 minutes at 4 °C. The flow-through was discarded and the retained protein was made up to final volume of 1 mL per mg of original starting material in 3% NaHCO₃. This was then transferred into a pre-soaked 3.5 kD MWCO dialysis tubing and dialysed overnight, then for 4 hours in 1 L 3% NaHCO₃ at 4 °C.
Next, NP-ε-Aminocaproyl-Osu (NP-CAP-Osu) was dissolved in DMF to a final concentration of 10 mg/mL. We determined empirically that a 2:1 ratio of NP:CSP in the final conjugated product required a 20:1 ratio of NP:CSP at the conjugation step. For other NP:CSP ratios (6:1 and 10:1) this 10-fold increase during conjugation was also used; i.e. 60:1 and 100:1, respectively. The dialysed CSP and required NP-CAP-Osu were combined in sterile 2 mL tubes, covered with foil to protect from light, then rotated for 4 hours at RT to conjugate. The conjugated products were then dialysed in 1 L 3% NaHCO₃ at 4 °C for 5 hours, overnight then 4 hours before dialysis in 1 L PBS for 4 hours and overnight. Conjugated products were stored at 4 °C with 0.1% sodium azide to prevent bacterial growth.

Absorption spectroscopy for quantifying NP

Since NP-CAP-Osu has an absorption maximum of 430nm with extinction coefficient of 4230, the concentration of NP was determined via NanoDrop. Briefly, 1.5 µL aliquots of CSP27-NP/ CSP9-NP were analysed on a NanoDrop ND-1000 Spectrophotometer to measure protein at 280 nm and NP dye at 430 nm. An average of two reads was used to calculate the concentration of NP.

Sandwich ELISA for calculating the concentration of CSP, post NP-CSP conjugation

The absorbance of NP influenced the absorbance of proteins at 280 nm, therefore NanoDrop was not an accurate method for quantification of protein concentration (CSP27/CSP9) in the conjugated products. Instead, the concentration of CSP27 or CSP9 was determined by sandwich ELISA, taking advantage of histidine-tags (His-tag) attached to both CSP27 and CSP9 (Figure S1). Briefly, 96-well Maxisorp Nunc-Nucleon plates were coated overnight with 1 µg/mL of anti His-tag antibody. The following day, plates were washed and blocked with 1% BSA for 1 hour. The NP-CSP conjugated products were initially diluted 1/100 in 1%
BSA along with a 1/100 dilution of a CSP27 or CSP9 standard with known concentration, and serially diluted 1 in 5 times down the plate. Next, a 1/2000 dilution of 2A10 primary antibody (1) was incubated for one hour, washed and incubated with secondary antibody (Anti IgG detection antibody conjugated to HRP) for one hour. After washing, the plates were developed with Peroxidase Substrate Kit for 15 minutes and read at 405 nm using a Tecan Infinite 200Pro plate reader. The reaction was stopped using 50 µL/ well stop solution, consisting of 10% SDS in PBS. The concentration of CSP27 or CSP9 in the conjugated products was interpolated from a sigmoidal standard curve of CSP27 or CSP9.

Preparing B cell tetramers for detecting diverse CSP specific or NP specific B cell response

To detect B cells to specific epitopes in CSP we used tetramers based on peptide probes described above for the CSP\textsubscript{Repeat}, CSP\textsubscript{Nterm}, or CSP\textsubscript{Cterm} region conjugated to phycoerythrin (PE) or allophycocyanin (APC). The tetramers were generated by mixing biotin-conjugated peptides with streptavidin-conjugated PE or streptavidin-conjugated APC in a 4:1 molar ratio. Briefly, 2.17 nM of peptide was made up to 50 µL in PBS. Then 8.68 nM of PE or APC were added in 4 equal aliquots every 15 minutes, incubating at room temperature (RT) in darkness between aliquots. Tetramers were stored at 4 °C in dark until use.

NP-specific B cells were detected using 4-hydroxy-5-iodo-3-nitrophenol (NIP) conjugated to PE or APC. Briefly, 1 mg of Native R-Phycoerythrin protein or Natural Allophycocyanin protein were transferred into pre-soaked 3.5 kD MWCO dialysis tubing and dialysed for 5 hours, overnight, then for 4 hours in 1 L 3% NaHCO\textsubscript{3} at 4 °C. NIP-\(\varepsilon\)-Aminocaproyl-Osu (NIP-CAP-Osu) was dissolved in DMF to a concentration of 10mg/mL. The NIP-CAP-Osu was added to the dialysed PE or APC at a ratio of 20 µg:1 mg and rotated at RT for 4 hours protected from light with aluminium foil. The conjugated NP-PE and NP-APC were then
dialysed in 1 L 3% NaHCO₃ at 4 °C for 5 hours, overnight then 4 hours before dialysis in 1 L PBS for 4 hours and overnight. NP probes were stored at 4 °C in dark till use.

**Immunizations with NP-CSP conjugates to study immunodominance**

Immunisations were conducted with the following amounts of antigen: NP-CSP27 = 15 µg/mouse, NP-CSP9 = 11.93 µg/mouse. Negative control groups were immunized with vehicle (PBS in alum). Antigens were emulsified in alum (2:1 volumetric ratio antigen: alum) to a total volume of 150 µL per mouse. The resultant solution was vortexed slowly at RT for 30 minutes to ensure the immunisations were fully emulsified. Immunisations were delivered intraperitoneally (IP), 150 µL total delivered in 75 µL aliquots on each side of abdomen into C57BL/6 recipient mice.

**Immunization with truncated CSP constructs**

Three different constructs of CSP were used for immunization, they include CSP9NVDP, CSP9, and CSP27 (Figure S1). C57BL/6 recipient mice were randomly separated in four groups, three of them had 15 mice per group, and were immunized with one of the CSP constructs. The fourth group had 11 mice and was the negative control. Each mouse received 30 µg of a CSP construct. These were emulsified in alum to a 2:1 volumetric ratio of antigen: alum and a resultant solution of 200 µL, before vortexing at RT for 30 minutes to ensure complete emulsification. Mouse were then IP immunized with 200 µL, 100 µL on left and right side, respectively. The immunisation regimen consisted of one priming and two boaster doses, each separated by an interval of 5 weeks. One day before immunization, blood was collected from mice via tail vein or retro-orbital bleeds for assessing antibody response. 2 weeks after the mice received their final booster, they were challenged via mosquito bite.
**Challenge of mice via Anopheles stephensi bites**

Controlled malaria infection challenge was performed via bite of *Anopheles stephensi* mosquitoes infected with Pb-PfSPZ a *P. berghei* parasite strain that expresses *P. falciparum* CSP (2). Also, parasite intrinsically expresses either GFP or mCherry, thus the infected mosquitoes were visually identified under microscope, and at least 5/4 of these were sorted onto separate containers and topped up with another 5/6. These were fed with sucrose for the first 6 hours and then with water solution a day before the challenge. The following day, mice were anaesthetised and placed on top of the containers to allow the mosquitoes to blood feed for 30 minutes. Next, 42 hours post mosquito bite, mice were euthanised via cervical dislocation and liver was collected, washed twice in PBS, and homogenised in 4 mL Denatured Working Stock.

**Quantification of parasite 18S rRNA in the liver**

To extract parasite rRNA, 60 µL of 2 M Sodium Acetate was added to a 600 µL aliquot of homogenised liver and vortexed to mix. Then 750 µL Acid Phenol:Chloroform was added and vortexed before incubating on ice for 15 minutes. The samples were spun at 15000 g for 20 minutes at 4 °C and the upper aqueous phase transferred into a clean 1.5 mL tube. The RNA was precipitated via addition of 400 µL isopropanol, vortexing then incubating at -20 °C for 1 hour. The RNA was pelleted at 15000 g for 20 minutes at 4 °C then the pellet washed twice with 1 mL cold 70% ethanol (EtOH) then dried for 10 minutes before resuspension in ultra-pure water. RNA concentration was measured via Nanodrop, and was diluted to make 100 µL aliquots at 50 ng/µL and stored at 4 °C.

cDNA was synthesised from the RNA using iScript cDNA Synthesis Kit according to the manufacturer’s protocol. Briefly, each sample was run in a 20 µL reaction in individual
dome-capped PCR tubes containing 10 µL ultra-pure water, 4 µL 5 x iScript Reaction Mix, 1 µL iScript Reverse Transcriptase and 5 µL RNA (50 ng/µL). The samples were run on an Eppendorf ProS Mastercycler at 25 °C for 5 minutes, 42 °C for 30 minutes, 85 °C for 5 minutes then hold at 4 °C.

RT qPCR was run using Power SYBR Green PCR Master Mix. Briefly, a PCR mastermix was made for all samples plus 2 no template controls (NTC) and 5 standards (STD) with the following volume per one reaction: 4.4 µL ultra-pure water, 5 µL Power SYBR Green PCR Master Mix, 0.05 µL P. berghei forward primer and 0.05 µL P. berghei reverse primer. 38 µL aliquots of mastermix were transferred into 1.5 mL tubes for each condition; 2 NTC, 5 STD and x cDNA samples. For DNA templates for P. berghei 18S qPCR 2 µL of the following was added for each condition; NTC – ultra-pure water, STD – plasmid standards (10^7, 10^6, 10^5, 10^4 and 10^3), samples – cDNA. The PCR was plated in triplicates of 10 µL/well in a MicroAmp 384 well reaction plate. The plate was sealed with MicroAmp optical adhesive film and spun at 500 g for 15 seconds to ensure all samples were at the bottom of each well. The qPCR was run on a 7900HT Fast Real-Time PCR System using the following conditions; 50 °C for 2 minutes, 95 °C for 10 minutes then 40 cycles of 95 °C for 15 seconds and 60 °C for 1 minute followed by 95 °C for 15 seconds and 60 °C for 15 seconds. The above qPCR reaction was repeated for glyceraldehyde 3-phosphate dehydrogenase (GapDH) using GapDH primers. However, the standards consisted instead of a pool of 2 µL cDNA from each sample that was serially diluted for the following concentrations; 1.0, 0.5, 0.25, 0.125, 0.0625. qPCR data was read using sDS2.4 software, P. berghei 18S was normalised to the GapDH reference gene before further calculating means in Prism 7.

*Calcium flux measurement in repeat specific B cells*
CSPrepeat-specific B cells were FACS purified from Igh\textsuperscript{g2A10} knock-in mouse splenocytes (McNamara \textit{et al.} submitted), and were cultured in complete RPMI for 16 hours. Then the cells were labelled in RPMI media containing 2 µM/ml Indo-1 and 7AAD for 20 min at 37 °C. Following 2 washes, the signal at BUV395 channel (Indo-1 bound) and BUV496 channel (Indo-1 free) was collected for 60 s to define baseline Ca\textsuperscript{2+} levels as the ratio of Indo-1 (bound/free). B cells were then stimulated for additional 360 s with either 0.5 µM/ml CSP or 10 µg/ml OVA-HEL. 1 µg/ml ionomycin was used as positive control. Data were collected in a FACS Fortessa instrument (BD) and analyzed using the Kinetics tool in FlowJo software (Tree Star). To analyzed the result, the baseline Ca\textsuperscript{2+} level was defined by the mean value of Indo-1 (bound/free) within 0-60 s. The calcium influx was then calculated by the mean values of every 30 s divided by the baseline value, and the plot was made by the calcium influx versus the mean value of this time frame.

Surface Plasmon Resonance

Surface plasmon resonance saturation experiments were performed on a Biacore 8K instrument (GE Healthcare) at 25 °C using a Series S Sensor Chip NTA (GE Healthcare) and SPR running buffer (10 mM HEPES, 150 mM NaCl, 50 µM EDTA, 0.05 % v/v Tween 20, pH 7.4). Solutions of His\textsubscript{6}-tagged CSP27 and CSP9 were prepared in SPR running buffer at concentrations of 0.1 µg/ml and 0.4 µg/ml, respectively. His\textsubscript{6}-tagged CSP27 and CSP9 were immobilized on separate channels on the sensor chip surface as per the manufacturer’s recommendations: a pre-conditioned chip was first activated with 500 µM NiCl\textsubscript{2}, and a solution of the His\textsubscript{6}-tagged ligand was subsequently passed over the chip using a flow rate of 5 µl/min for 120 s. This yielded approximately 150 RU and 50 RU of immobilized CSP27 and CSP9, respectively. A saturating solution of mAb 2A10 (2 µM in SPR running buffer)
was then passed over the chip for 400 s using a flow rate of 30 µl/min, followed by a 400 s
dissociation period. The increase in response units (RU) corresponding to ligand
immobilization (RU_{lig}) and analyte binding (RU_{analyte}) in the reference-subtracted (reference =
blank surface) sensorgrams was measured, and the binding stoichiometry (n, molar ratio of
antibody to antigen in the complex under saturating concentrations of mAb 2A10) was
estimated using Eq. (1) as previously described (3), using molecular weights (MW)
calculated using ProtParam (4): CSP27 (35.4 kDa), CSP9 (28.2 kDa) and mAb 2A10 (145.9
kDa).

\[
n = \frac{RU_{analyte}}{RU_{ligand}} \times \frac{MW_{ligand}}{MW_{analyte}}
\]  

(1)

All buffers were filtered and degassed prior to use. Following each cycle, the chip was
completely regenerated using sequential washes of 500 mM imidazole, 350 mM EDTA (pH
8.5) and 100 mM NaOH. Each experiment was performed in duplicate (n=2), on separate
channels on the SPR chip.
### Table S1: Antibodies used for flow cytometry in this study

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Conjugate</th>
<th>Clone</th>
<th>Source</th>
<th>Catalogue</th>
<th>Conc</th>
<th>Dilution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TruStain fcX antibody:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rat anti-mouse CD16/32</td>
<td></td>
<td>93</td>
<td>Biolegend</td>
<td>101320</td>
<td>0.5 mg/mL</td>
<td>1/50</td>
</tr>
<tr>
<td>Anti-mouse CD38</td>
<td>APC A700</td>
<td>90</td>
<td>eBioscience</td>
<td>56-0381-82</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse IgM CD45.2</td>
<td>APC Cy7</td>
<td>II/41</td>
<td>Invitrogen</td>
<td>47-5790-82</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse CD45.2</td>
<td>APC Cy7</td>
<td>A20</td>
<td>Biolegend</td>
<td>110716</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse CD19</td>
<td>BUV395</td>
<td>1D3</td>
<td>BD Horizon</td>
<td>563557</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse IgD B220</td>
<td>BV605</td>
<td>11-26c.2a</td>
<td>BioLegend</td>
<td>405727</td>
<td>0.2 mg/mL</td>
<td>1/400</td>
</tr>
<tr>
<td>Anti-mouse CD45.2</td>
<td>BV605</td>
<td>RA3-6B2</td>
<td>BioLegend</td>
<td>103244</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse CD45.1</td>
<td>FITC</td>
<td>A20</td>
<td>Biolegend</td>
<td>110706</td>
<td>0.5 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse CD45.2</td>
<td>FITC</td>
<td>104</td>
<td>Invitrogen</td>
<td>11-0454-82</td>
<td>0.5 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse GL7</td>
<td>Pacific Blue</td>
<td>GL-7</td>
<td>Invitrogen</td>
<td>48-5902-82</td>
<td>0.2 mg/mL</td>
<td>1/100</td>
</tr>
<tr>
<td>Anti-mouse CD45.2</td>
<td>PE</td>
<td>104</td>
<td>Biolegend</td>
<td>109807</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse CD138</td>
<td>PE Cy7</td>
<td>281-2</td>
<td>Biolegend</td>
<td>142514</td>
<td>0.2 mg/mL</td>
<td>1/100</td>
</tr>
<tr>
<td>Anti-mouse CD3</td>
<td>PerCP Cy5.5</td>
<td>17A2</td>
<td>Biolegend</td>
<td>100218</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse CD11b</td>
<td>PerCP Cy5.5</td>
<td>M1/70</td>
<td>Biolegend</td>
<td>101228</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse CD11c</td>
<td>PerCP Cy5.5</td>
<td>N418</td>
<td>Biolegend</td>
<td>117328</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>Anti-mouse Ly-6G/Ly-6C (GR1)</td>
<td>PerCP Cy5.5</td>
<td>RB6-8C5</td>
<td>Biolegend</td>
<td>108428</td>
<td>0.2 mg/mL</td>
<td>1/200</td>
</tr>
<tr>
<td>7AAD Cell Viability Dye</td>
<td>PerCP Cy5.5</td>
<td>-</td>
<td>Biolegend</td>
<td>420404</td>
<td>50 g/mL</td>
<td>1%</td>
</tr>
</tbody>
</table>
Supplementary References


Supplementary Figure 1

(A) Sequence of the 3D7 circumsporozoite protein, different domains are labelled with different colours: leader sequence, orange; N-terminal domain, blue; repeat, NANP – light red, NVDP – dark red; C terminal domain, purple. (B) Schematics of the different constructs used in this study, numbers refer to the 3D7 sequence, insertions/substitutions are marked as text.

Supplementary Figure 1: Circumsporozoite protein sequences and constructs used in this study
Supplementary Figure 2: Gating strategies for flow cytometry

(A) General gating strategy applied to all flow cytometry data prior to gating antigen specific cells. Firstly, lymphocytes were gated based on their relative size and granularity (forward scatter area versus side scatter area). From the lymphocyte population, single cell events were gated on relative forward scatter area versus forward scatter height. From the single cell population, live cells were gated on PerCP Cy5.5 negative population (this is negative for 7AAD viability dye and also negative for lineage markers CD3, CD11b, CD11c and GR1). From the live cells, B cells and plasmablasts were gated as CD19⁺ and/or CD138⁺ cells. From the B cell population, activated cells were gated as IgD⁻ cells. The activated cells were then further gated for antigen specificity. Plots of further gating are provided in results figures.

(B) Gating strategy for quantifying Ighg2A10 NANP tetramer⁺ B Cells. Lymphocytes were gated based on their relative size and granularity (forward scatter area versus side scatter area). From the lymphocyte population, single cell events were gated on relative forward scatter area versus forward scatter height. From the single cell population, single cell events were gated on relative forward scatter area versus forward scatter height. From the single cell population, B cells were gated as B220⁺ cells. From the B cell population, CSP_repeat⁻tet⁻ cells were gated as cells double-positive for both the tetramer and congenic marker CD45.1.

(C) Gating strategy for quantifying B1-8⁺ NP⁺ B cells. The lymphocyte, single cell and B cell populations were gated as described in (A). From the B cell population, NP⁺ B cells were gated as cells double-positive for both the NP marker and congenic marker CD45.1.
Supplementary Figure 3: CSP27-NP2 elicits both CSPRepeat and NP specific IgG responses
C57BL/6 mice were immunised with CSP27-NP2 or CSP27 only control. Sera were taken 7, 14 and 21
days post-immunisation and the antibody response to each epitope measured via ELISA. (A) Schematic
of CSP27 showing the locations of lysine residues, indicated by asterisks. (B) Total IgG response to CSPRepeat measured via (NANP)_9 ELISA. (C) Total IgG response to NP measured via NP(14)BSA ELISA. Data are represented as mean ± SD pooled from three independent experiments (n=3 mice/group/experiment); ELISA data were analysed separately for each antigen in R Studio using a mixed linear model with Immunogen (CSP27-NP2/CSP27) and day as experimental factors, experiment as a fixed factor and mouse as a random factor. Two-way ANOVA p values are listed underneath each graph, pairwise comparisons were made via Tukey post test and are represented using symbols; * p<0.05, ** p<0.01, *** p<0.001.
Supplementary Figure 4: Increasing NP-specific precursor number does not alter immunodominance.

0, 1x10^4, 3x10^4, or 9x10^4 of CD45.1 B1-8hi cells were adoptively transferred into C57Bl/6 mice followed by immunization with CSP27-NP2 in alum. Sera were taken on days 7, 14 and 21 and spleens analyzed 21 days post-immunization. (A) Total IgG response to CSP_repeat measured via (NANP)_9 ELISA. (B) Representative flow cytometry plots showing gating of total IgD- and GC B cells specific for NP or the CSP_repeat values are percentages (C) Absolute numbers of NP probe+ and CSP_repeat tetramer+ IgD- B cells. (D) Absolute numbers of NP probe+ and CSP_repeat tetramer+ GC B cells. (E) Total IgG response to NP measured via NP(14)BSA ELISA. (F) Absolute numbers of CSP_repeat tetramer+ CD45.1+ Ighg2A10 and CD45.1- endogenous cells. Data are represented as mean ± SD from a single experiment (n=3 mice/group); all data were analyzed via 2-way ANOVA, with mouse included in the model as a fixed factor. ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test though no pairwise differences were significant.
Supplementary Figure 5

Increasing the level of NP conjugation to CSP alters the immunodominance hierarchy. C57BL/6 mice were immunized with either CSP27, CSP27-NP2, CSP27-NP6 or CSP27-NP10. Sera were taken on days 7, 14 and 21 and spleens analyzed 21 days post-immunization. (A) Total IgG response to NP measured via NP(14)BSA ELISA. (B) Total IgG response to CSP_repeat measured via (NANP)9 ELISA. (C) Representative flow cytometry plots showing gating of total IgD- and GC B cells specific for NP or the CSP_repeat* values are percentages (D) Absolute numbers of NP probe* and CSP_repeat tetramer* IgD B cells. (E) Absolute numbers of NP probe* and CSP_repeat tetramer* GC B cells. Data are represented as mean ± SD pooled from two independent experiments (n=4 mice/group/experiment); these data were analyzed via 2-way ANOVA, with experiment and mouse included in the model as fixed factors. ANOVA p values are listed below or adjacent to each graph. Pairwise comparisons were performed using a Tukey post-test with significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001.
Supplementary Figure 6: Additional analysis of mice immunized with truncated CSP molecules.

Sera from the mice described in Figure 4A were taken for additional ELISA analysis. (A) Sequence and location within the CSP molecule of additional peptides corresponding to potential targets of protective antibodies. (B) Overall IgG responses to the 5D5 epitope. (C) Overall IgG response to the junction between the CSPNterm and CSPRepeat. (D) Overall IgG response to a peptide corresponding to the first 32 amino acids of the 3D7 CSP repeat domain including multiple NVDP repeats. (E) Ratio of the week 15 response to the NVDP peptide and the standard CSPRepeat peptide between the different immunization groups. Data from panels B-D was analyzed from 2 experiments with 5 mice/experiment/group analyzed via 2-way ANOVA with experiment and mouse as blocking factors, ANOVA p values are listed below or adjacent to each graph; pairwise comparisons between groups (averaged over time) were performed using a Tukey post-test and significant values are represented as symbols; * p<0.05, ** p<0.01, *** p<0.001. Data from panel E was from 2 experiments with 5 mice/experiment/group analyzed via one-way ANOVA with experiment as a blocking factor.
Chapter 5. Using ITC to characterise molecular interactions involving novel components of the nonribosomal synthetase machinery.
5 Using ITC to characterise molecular interactions involving novel components of the nonribosomal synthetase machinery.

Key Ideas: Nonribosomal peptide synthesis, protein modularity, protein-substrate interactions.

Techniques: Isothermal titration calorimetry

5.1 Preface

Nonribosomal peptides (NRPs) are a class of peptide secondary metabolites that have a broad range of biological activities and pharmacological properties (Süssmuth and Mainz, 2017). NRPs are a very diverse and complex class of compounds; they can be linear, cyclic or branched, and can contain non-proteinogenic amino acids. Each NRP is synthesised in a mRNA-independent manner by a specialised nonribosomal peptide synthetase (NRPS). NRPSs are large, modular, multi-domain and/or multi-chained complexes where each core module in the NRPS functions to couple a single amino acid to the peptide chain. Each of these NRPS core modules consists of an adenylation (A) domain for the activation and loading of the amino acid, a thiolation or peptidyl carrier protein (T/PCP-domain) and a condensation (C) domain. Additional enzyme modules in the NRPS modify the resulting peptide sequence through epimerisation, formylation, methylation, heterocyclisation, reduction or oxidation reactions.

NRPs are important pharmaceutical compounds. Drugs centered around NRP core structures are used as last-resort antibiotics, immune suppressors and cytostatics, with sales worth billions of (US) dollars each year. Since these compounds are often difficult to synthesise using conventional chemistry approaches, understanding the biosynthesis of these compounds has great important for the pharmaceutical industry.
The two papers presented in this chapter explore the structure and function of two novel domains involved in NRP synthesis. Understanding how these domains interact with other modules within the NRPS complex and with substrates and products is critical to understanding the molecular mechanisms that underly their biological function and NRP biosynthesis. I contributed to these pieces of work by performing ITC experiments that were important for characterising such interactions. Coupled with crystal structures of these domains, this data providing important information about how these domains work to produce pharmacologically-relevant compounds.
5.2 Paper 1: *Drosophila melanogaster* nonribosomal peptide synthetase Ebony encodes an atypical condensation domain

5.2.1 Publication status

This manuscript presented in this section has been published by *PNAS*. Supporting information is provided in Section 5.3.

5.2.2 Author’s contribution

I designed and performed the ITC experiments for this publication. I analyzed the data and produced the figures and tables presenting the ITC data. I helped with the interpretation of the data and the editing of the final manuscript.

5.2.3 Contributions from others

The majority of the work presented in the following paper was done by members of Associate Professor Max Cryle’s group at Monash University. Professor Colin Jackson helped with the interpretation of the data and writing the manuscript.
**Drosophila melanogaster** nonribosomal peptide synthetase Ebony encodes an atypical condensation domain

Thierry Iзорe,1, Julien Tailhades,1, Mathias Henning Hansen,1, Joe A. Kaczmarski, Colin J. Jackson, and Max J. Cryle

1The Monash Biomedicine Discovery Institute, Monash University, Clayton, VIC 3800, Australia; 2Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia; EMBL Australia, Monash University, Clayton, VIC 3800, Australia; and 3Research School of Chemistry, The Australian National University, Acton, ACT 2601, Australia

Edited by Mohamed A. Marahiel, Philipps-Universität Marburg, Marburg, Germany, and accepted by Editorial Board Member Michael A. Marletta December 27, 2018 (received for review July 5, 2018)

The protein Ebony from *Drosophila melanogaster* plays a central role in the regulation of histamine and dopamine in various tissues through condensation of these amines with ⍺-alanine. Ebony is a rare example of a nonribosomal peptide synthetase (NRPS) from a higher eukaryote and contains a C-terminal sequence that does not correspond to any previously characterized NRPS domain. We have structurally characterized this C-terminal domain and have discovered that it adopts the aryl-alkylamine-N-acetyl transferase (AANAT) fold, which is unprecedented in NRPS biology. Through analysis of ligand-bound structures, activity assays, and binding measurements, we have determined how this atypical condensation domain is able to provide selectivity for both the carrier protein-bound amino acid and the amine substrates, a situation that remains unclear for standard condensation domains identified to date from NRPS assembly lines. These results demonstrate that the C terminus of Ebony encodes a euukaryotic example of an alternative type of NRPS condensation domain; they also illustrate how the catalytic components of such assembly lines are significantly more diverse than a minimal set of conserved functional domains.

Published under the PNAS license.

**Significance**

Nonribosomal peptide synthesis is responsible for the formation of many important peptide natural products in bacteria and fungi; it typically utilizes a modular architecture of repeating catalytic domains to produce these diverse peptide structures. The protein Ebony from *Drosophila melanogaster* is a rare example of such a nonribosomal peptide synthetase from a higher eukaryote, where it plays a central role in the regulation of amine neurotransmitters. Here, we reveal that the C-terminal portion of Ebony encodes an atypical peptide bond-forming nonribosomal peptide synthetase domain. Structural analysis shows that this domain adopts a fold not predicted by its primary sequence, and indicates how this domain maintains its high degree of substrate specificity.

Author contributions: T.I., J.T., and M.J.C. designed research; T.I., J.T., M.H.H., and J.A.K. performed research; T.I., J.T., J.A.K., C.J.J., and M.J.C. analyzed data; and T.I. and M.J.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. M.A.M. is a guest editor invited by the Editorial Board.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank, [www.pdb.org](http://www.pdb.org) (PDB ID codes 6DYM, 6DYN, 6DYO, and 6DYS).

1To whom correspondence may be addressed: Email: thierry.izore@monash.edu or max.cryle@monash.edu.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1811194116][1]/DCSupplemental.

Published online January 31, 2019.

www.pnas.org/cgi/doi/10.1073/pnas.1811194116
followed by sequences usually not found in archetypical NRPS assembly lines (4). Examples of such NRPS enzymes are mostly those involved in mammalian lysine metabolism (5): AASDH, a 2- aminoaoacidic 6-semialdehyde dehydrogenase harboring an unusual A-PCP-PQQ arrangement (where PQQ represents a sequence containing seven binding motifs for pyrroloquinoline quinone); and Lys2 (6), an α-aminoacid reductase with an A-PCP-NADPH-binding domain architecture. In addition to these well-characterized mammalian enzymes, Ebony from D. melanogaster contains an A-PCP di-domain followed by an uncharacterized sequence with no sequence homology to any known proteins based on standard search using domain prediction servers (Fig. 1).

Ebony is an 879-residue protein (98.5 kDa) expressed in both glial and cuticular cells (7, 8). In glial cells, Ebony is involved in histamine regulation (the main neurotransmitter in the optical nerve system) and plays an essential role in neurotransmitter inactivation through conversion to carmine [β-alanyl-histamine; Fig. 1 (9)]. Similarly, in cuticular cells Ebony catalyzes the condensation of β-alanine with dopamine to form β-alanyl-dopamine, a metabolite involved in the pigmentation and sclerotization of the insect cuticle. Mutants display strong phenotypes, with alteration of vision (10), circadian regulation of locomotor activity (11), and cuticle selerotization in affected flies.

Ebony is an unusually fast NRPS enzyme (12), which can achieve a condensation reaction up to 60,000 times faster than the archetypical NRPS tyrocidine synthetase. While the A-domain of Ebony is specific for β-alanine (13), the C-terminal domain appears to be versatile and can use a wide range of amines containing a planar ring. The C-terminal sequence of Ebony is thus of great interest among condensation-type domains given its versatility in assembling various novel sequences. While the C-terminal domain of Ebony appears to represent a previously unknown example of an NRPS condensation domain and displays intriguing catalytic properties, we sought to structurally characterize this domain. For this, we solved the crystal structure of the Ebony C-terminal domain both in its apo form and in complex with the amine substrates dopamine and histamine along with the resultant products β-alanyl-dopamine and caranine (β-alanyl-histamine).

Our results demonstrate that the Ebony C-domain, unlike standard NRPS C domains [e.g., VibH (14); Fig. 1], unexpectedly adopts the arylylalnine-N-acetyl transferase (AANAT) fold that was not directly apparent from standard sequence homology searches, and provides an understanding of the mechanism of selectivity of this condensation domain for both the PCP-bound amino acid and aromatic amine substrates that fits the biological functions of Ebony.

**Results**

As the condensation function of Ebony appeared to rely on the unusual C-terminal portion of this enzyme, we concentrated on the characterization of this atypical condensation-like domain. Previous work has shown that Ebony is highly prone to degradation during E. coli expression, and we used this fact to our advantage to identify an optimal C-terminal construct based on proteolysis of the full-length protein during overexpression. This region, encompassing the residues from Leu666 to the C-terminal residue Lys879 (referred to here as C_N for the amine-selecting C domain; Fig. 2), was well behaved and highly soluble when expressed with a C-terminal 6xHis tag. With the ability to access significant amounts of highly pure protein, we then turned to the structural characterization of this domain.

**Structural Characterization and Substrate Binding of the Ebony Condensation-Like Domain.** The optimized Ebony condensation-like domain (C_N) crystallized readily, forming numerous needle clusters in a wide range of conditions. To obtain diffraction-quality crystals, several rounds of condition optimization combined with crystal seeding were required, which yielded crystals in space group P2(1)(1) that diffracted to 2.0 Å. A molecular replacement model seeded from the *Paramecium bursaria* chlorella virus polyamine acetyltransferase (15) was successfully generated by the Robetta server (16), which makes use of both ab initio modeling and homology search routines. This proved necessary after crystallization of selenomethionine-labeled C_N protein as well as molecular replacement using potential homologs identified by Phyre2 (17) had failed. The density map thus obtained allowed us to build a model of C_N with high confidence and optimal geometry (SI Appendix, Table S1). The overall fold of the C_N domain is highly reminiscent of members of the AANAT family (18) despite a very low level of sequence identity (<20%, SI Appendix, Fig. S1); the closest structural homolog identified was the chlorella virus polyamine acetyltransferase (15), which aligned with an overall rmsd of 2.1 Å over 145 Ca atoms. The Ebony C_N domain is composed of seven α-helices arranged around a central β-sheet, with the 7-stranded β-sheet splayed between strands 4 and 5 (Fig. 2). While this structural feature in the AANAT superfamily serves to create a binding pocket for the pantetheine moiety of the acetyl-CoA (substrate) (18) (Fig. 2), in the case of Ebony it is instead required to accommodate the β-alanine-loaded phosphopantetheine arm from the upstream donor PCP domain. Although the structure of C_N is clearly highly divergent from the typical NRPS condensation domain—which is essentially a dimer of two chloramphenicol acetyltransferase (CAT) domains (Fig. 1)—the general feature of the binding site between two highly symmetrical half-domains is maintained in this C-type domain.

The two closest structural homologs of the Ebony C_N domain in *Drosophila* are the dopamine (19) (Fig. 2) and agmatine (20) AANATs (SI Appendix, Fig. S1), which display rmsd values of 2.4 and 2.5 Å despite the low sequence identities of these enzymes (16 and 12%, respectively). While the core fold of these enzymes is well conserved between C_N and the members of the AANAT family, there are substantial differences in the loop regions. The most striking difference is a mobile loop connecting β-strand 3 and α-helix 4 in Ebony, which is folded into a helix of different lengths in the structures of AANAT enzymes (SI Appendix, Fig. S2). This region is located on the opposite side of the protein to the donor PCP substrate-binding site and forms part of a highly acidic surface composed of residues Glu748, Glu750, Asp752, Glu762, Glu765, and Glu768 (Fig. 2). This acidic (and hence negatively charged) region is centered on an active site access channel anticipated to be involved in binding the positively charged dopamine and aromatic amine substrates.
charged amine substrates utilized by Ebony (Fig. 2). This hypothesis is supported by the related AANAT enzymes displaying a similar charged region and utilizing amine substrates comparable to Ebony. The alteration of a secondary structure within this region to a flexible loop presents one possible route for accelerating the access of amine substrates to the C\textsubscript{N} active site.

On the opposite side of the domain, the binding site for the aminoacyl-PCP domain is a relatively flat and hydrophobic surface, which stands in contrast to the highly positively charged “cradle” required to accommodate the phosphate groups of the CoA substrates in AANAT enzymes (Fig. 2 and SI Appendix, Fig. S1). The nature of this putative PCP interaction interface is in agreement with the vast majority of such PCP interaction interfaces identified within NRPS machineries to date [i.e., mainly hydrophobic (2)], likely because of the role played by the PCP-bound prosthetic linker in accessing the active sites of various NRPS domains. Replacement of a CoA substrate with a PCP-bound substrate in the case of Ebony also implies that binding of the thioester tethered β-alanyl substrate to C\textsubscript{N} is controlled by the activity of the upstream A domain through the hydrolysis of ATP during amino acid activation—a mechanism known as the A-domain alternation cycle (21, 22). The rapid rate of activity reported for Ebony appears to be governed by the activity of the upstream A domain, with the C\textsubscript{N} domain also then able to catalyze peptide bond formation at a rate significantly faster than typical NRPS C\textsubscript{N} domains (12). From isothermal calorimetry (ITC) measurements (SI Appendix, Table S2 and Fig. S7), Ebony C\textsubscript{N} has a dissociation constant (\(K_d\)) for dopamine of ~30 μM (\(K_d\) of ~60 μM for the β-alanyl-dopamine), which is consistent with processing under steady-state conditions given the concentration range of dopamine in cuticular cells (23). In contrast, Ebony C\textsubscript{N} has a substantially lower affinity for histamine (\(K_d\) ~600 μM), which again is consistent with the millimolar concentrations of histamine released during neurotransmission in the brain and optic lobes (670 mM in synaptic vesicles (24)). Interestingly, Ebony C\textsubscript{N} has a slightly higher affinity for carcinine (product; \(K_d\) ~220 μM) than for the substrate, histamine. Given the physiological role of histamine as a neurotransmitter that is released in “bursts,” this allows product concentration to regulate the activity of Ebony (24, 25), with such product inhibition also observed in other enzymes that modulate neurotransmitter levels [such as acetylcholinesterase (26)]. Thus, these affinity measurements are consistent with the physiological roles of the substrate molecules.

Substrate- and Product-Bound States of the Ebony C\textsubscript{N} Domain. To gain insight into how the Ebony C\textsubscript{N} domain functions to generate peptide bonds between PCP-bound β-alanine and dopamine/histamine, we determined four additional cosubstrate structures of this domain in either substrate-bound (histamine and dopamine) or product-bound [carcinine (β-alanyl-histamine) and β-alanyl-dopamine] states; the relatively weak binding of histamine required higher soaking concentrations, consistent with the ITC measurements. All complexes produced clearly defined electron density for the additional ligands, which were easily identified in the C\textsubscript{N} catalytic channel between β-strands 4 and 5 (see difference density maps and polder validation maps; SI Appendix, Fig. S3). The aromatic rings present in both the substrates and the β-alanine–conjugated products dock into a perfectly tailored hydrophobic cage as well as the E696L mutant of the central coordinating glutamate residue. In C\textsubscript{N} condensation assays (assessing the formation of β-alanyl-dopamine; see below), mutant E696L showed no enzymatic activity, indicating the importance of this residue for coordination of the aromatic moiety of the amine acceptor. Of the two phenylalanine mutants, F689A showed only ~5% of the activity of the wild-type enzyme, while F761A retained almost 50% activity (Fig. 4). These results directly correlate with the relative distance of the phenylalanine rings from the amine substrate, with F689 closer (~3.8 Å) than F761 (~4.2 Å) (Fig. 3). Such residues are also conserved in members of the AANAT superfamily (SI Appendix, Fig. S1), which implies a general role for them in binding the amine substrates for these enzymes.

Inside the hydrophobic cage, the catechol moiety of dopamine is coordinated by residues Glu696, which hydrogen-bonds to both hydroxyl groups (2.9 and 3.1 Å; Fig. 3). In the histamine-bound structure, the amine hydrogen on the histidine ring is coordinated via a single interaction with Glu696 (3.0 Å; Fig. 3), which, combined with the increased distance to F761 (5.5 vs. 4.2 Å), explains the lower affinity of C\textsubscript{N} for histamine. Although typical NRPS condensation domains utilize a conserved active site histidine residue [albeit one whose role is somewhat unclear (3)], no such direct interaction is present in C\textsubscript{N}. A superposition of the ligand-bound structures shows that the positioning of the...
aromatic ring of the amines is maintained—most likely due to the hydrophobic cage as discussed above—while there are much greater differences in the position of the aliphatic chain and the terminal amino group between the dopamine and the histamine molecules (Fig. 3). When ligand-bound CN structures are compared with the dopamine AANAT in complex with acetyl-CoA [PDB ID 4TE3 (19)], it becomes clear that the histamine-bound structure represents the most likely substrate position for the condensation reaction. Indeed, the proximity of the reactive amine group and the angle of attack that this orientation provides are suitable for the reaction to take place (SI Appendix, Fig. S4). Also, the position of the β-alanyl moieties of the products found bound to CN is consistent with attack of the amine (based on the histamine structure). Furthermore, the position of the peptide bond formed between the amine and β-alanine in both structures superimposes well on the carbonyl group of the acetyl-CoA substrate reported in the dopamine-AANAT structure (SI Appendix, Fig. S4). However, one important difference is that the β-alanyl moiety of the products bound to CN projects into a hydrophilic cavity that is not present in the dopamine-AANAT structure.

Given the interactions of amino acid side chains with the amine group of β-alanine in this region (Fig. 3), it appears that this cavity may be an important determinant of the specificity of CN for β-alanine (and a subset of related compounds) over α-amino acids (see below). Analysis of the structurally related dopamine-AANAT (3TE4) postulates an unusual Glu-Ser-Ser catalytic triad being involved in amide bond formation in AANAT catalysis. In CN, the serine residues are replaced by threonine residues (Thr828 and Thr832) but the glutamate residue is not conserved. The postulated role of these threonine residues in AANAT catalysis is not directly supported in the activity of CN due to the position of the amine groups of the substrate-bound CN structures solved here. Indeed, the amine moiety in histamine, which adopts the most likely position for interception of the thioester moiety of the PCP-bound substrate, is instead coordinated by the oxygen atoms of backbone carbonyl groups (residues Phe787 and Thr826) as well as several water molecules that themselves are further coordinated with the hydrophilic side chains of residues Glu768, His785, and Thr825 (Fig. 3). Mechanistic investigations of an enzyme different from, yet structurally related to, CN (serotonin NAT) have implicated the equivalent residue to His785 in the mechanism of this NAT (27), although the histamine-bound CN structure does not support direct interaction between this histidine residue and the amine group of histamines to allow deprotonation. Rather, the structures of CN indicate that it is likely that this water-mediated interaction network is also able to orient the amine group of the bound substrate in such a way as to promote thioester attack. This does not require rearrangement of the amine substrate, as is required for the postulated mechanism for dopamine AANAT, which appears
unlike given the effectiveness of the hydrophobic cage in binding the aromatic side chain of the amine substrates in this case.

To further investigate the role of these residues in the activity of the Cε domain, H785F and E768Q mutants were prepared. The H785F mutant could not be expressed in soluble form, which highlights a likely structural role of His785 in domain folding and stability (including interactions with Glu748 and Glu768). Mutation of Glu768 to Gln did not significantly reduce the activity of Cε (∼75% residual activity) (Fig. 4; numbers in bold throughout reference the structures in Fig. 4), supporting the indirect interaction of this residue with the amine group of β-alanine through a network of ordered water molecules as determined in the product-bound structures of the Cε domain. It has been suggested that the differences observed in catalytic mechanisms of AANAT-like enzymes result from the case of the thioester aminolysis reaction, which does not require a specific catalytic site (27). This argument also appears to hold for traditional NRPS C domains, where the role of the central active site histidine residue is debated across different systems (3). One clear difference, however, is that in standard NRPS C domains the central histidine residue is believed to directly coordinate the substrates during peptide bond formation. In the case of Cε, this is not the case, and the impact of the protein backbone appears to be mostly indirect, serving instead to assist in orienting the amine substrate via a coordinated water network.

**Amino Acid Specificity of the Ebony Cε Domain.** Within NRPS-catalyzed biosynthesis, amino acid selectivity is largely determined by the activity of specific adenylation domains within the NRPS machinery (1). C domains are believed to play a reduced role in the selection of peptide structure, but rather play a role as stereochemical gatekeepers together with the related epimerization (E) domains (1). Given the structure of Cε and the atypical structure of the β-alanine substrate, we were curious how selective this domain is for different amino acid acceptor substrates. We first confirmed the acceptance of β-alanyl-CoA as a substrate for the β-alanine-loaded PCP domain (28) by Ebony Cε in reactions along with dopamine as the amine acceptor; we confirmed the formation of β-alanyl dopamine through comparison with an authentic standard (SI Appendix, Fig. S5). Next, we utilized the same assay with the Cε domain and a set of 6 β-alanyl-CoA analogs, 7 α-amino acid CoAs, and glycine CoA (Fig. 4 and SI Appendix, Fig. S6). These experiments showed that, of the 14 substrates tested, only 3 out of the 6 β-alanine analogs were accepted, while α-amino acids/glycine were not accepted. The lack of acceptance of α-amino acids by Cε appears to be caused by steric hindrance around the α-position of these amino acids through the active site channel formed by parts of β-strands 4 and 5.

The product-bound Cε structures reveal that the side chains of Ser786, Thr825, and Asn827 all coordinate the amine group of β-alanine (Fig. 3) and aid the correct orientation of the PCP-bound (or CoA-bound) amino acid in the Cε active site. The side chain of Asn827, in particular, appears very important to this process, as it is held via hydrogen-bonding interactions from His873 [2.9 Å; His873 also interacts with Thr843 (3.4 Å)] in such a way that it fits into a β-sheet-type orientation that is then also adopted by the backbone of the bound β-alanine molecule and the β-strand 4 Met789 (Fig. 3). The lack of Cε activity with glycine can be rationalized by the loss of interaction between the terminal amino acid and these residues. This is also supported by the lack of acceptance of 2 by Cε, as the terminal amine is missing in this compound. Furthermore, the addition of a methyl substitution to the amine moiety of β-alanine in 4 prevents this compound from being a viable substrate for the Cε domain. This is likely reconciled by the additional steric bulk that this substitution introduces around the crucial amine group. The importance of coordination with the amine moiety in β-alanine is further supported, albeit indirectly, by the acceptance of 3; this indicates that the hydrogen-bonding interaction with the protein—although required—can accommodate alternate groups (such as a carboxylate) through alteration of the interacting water network. The methylene extended compound 5 is also accepted by Cε, which can also be rationalized by a rearrangement of the water network that maintains the essential interactions between the terminal amine and the Cε domain. The influence of the Cε binding site on the acceptance of branched substrates is clearly seen in the acceptance of (R)-β-homoalanine 7 as a substrate while the enantiomer 6 is not accepted. Inspection of the product-bound Cε structures shows that the S-methyl group sterically clashes with the carbonyl group of Asn827 but that the R-methyl group is easily accommodated in the cavity toward the more distant Met789 side chain (Fig. 3). These results indicate that the Cε domain requires a substrate with a terminal moiety able to hydrogen-bond effectively within the substrate-binding channel. The ability of the Cε domain to tolerate longer substrates as well as some branching within substrates depends entirely upon the ability of the narrow channel of the amino acyl-PCP binding site to accept them. However, these results indicate the potential for the Cε domain to accept substrates other than β-alanine, which is enforced in full-length Ebony by the selectivity of the A domain.

**Discussion**

The modular architecture of NRPS assembly lines leads to tremendous diversity within the products assembled by them, with the specificity of the assembly lines largely ascribed to the functions of adenylation domains (1). It has been recognized that C domains—which are essential catalytic NRPS domains—can play much wider roles in generating product diversity within NRPS pathways (3). Examples of atypical catalytic functions for traditional NRPS C domains include the generation of a β-lactam ring in nocardicin biosynthesis (29), multiple-step heterocyclization reactions, and elimination and rearrangement to produce methoxyvinyl-containing amino acids (30); the noncatalytic roles for these domains include the recruitment of multiple cytochrome P450 enzymes to perform oxidative cross-linking of aromatic side chains within glycopeptide antibiotic biosynthesis (31). It is then logical that within traditional NRPS-type architectures, there is significant potential to identify additional functions within “standard” catalytic domains.

Our results demonstrate that Ebony Cε is a condensation domain in NRPS machinery, having specific selectivity requirements for the PCP-bound amino acid. The Cε domain also adopts a totally different fold, belonging to the AANAT superfamily of enzymes despite very low levels of sequence identity. The relative chemical ease of the reaction performed by both Ebony Cε and standard NRPS C domains is reflected in the active sites found in both folds, although in the case of Ebony Cε it appears that the enzyme controls the orientation of the amine nucleophile through a water network rather than a central histidine residue. Use of a highly charged active site channel for the amine substrate in Ebony Cε somewhat diverges from the structure of traditional C domains with related amine acceptor substrates such as VibH. This likely stems from the rapid rate of catalysis required by Ebony compared with NRPS systems from secondary metabolism pathways. The structure of Ebony Cε also appears to be highly rigid, with little if any rearrangement of the protein upon substrate binding. Such rigidity also stands in contrast to traditional NRPS C domains, in which the relative orientation of the two CAT-like subdomains generates differences in the accessibility of the acceptor site. The ability of traditional C domains to adopt open and closed conformations (2, 3) with regard to the acceptor site is one way that an NRPS assembly line can generate directionality during synthesis across
multiple modules and hence multiple peptide bond formation steps. In the case of Ebony, the ability to prealign substrates in a rigid active site leads to a significant increase in the maximal rate of this reaction, which is crucial for its function in vivo. Although binding affinities for the amine substrates differ considerably [K<sub>d</sub> ∼30 μM (dopamine) vs. ∼600 μM (histamine)] in accordance with their different physiological roles (SI Appendix, Table S2), the use of the AANAT fold to perform this condensation reaction appears to be an effective way for this eukaryotic NRPS to control the rate of reaction by dispensing with a traditional, significantly slower C domain.

In utilizing a different fold to perform the role traditionally held by a C domain, Ebony C<sub>N</sub> is reminiscent of other recent examples of reactions found in NRPS catalysis within different enzyme folds [e.g., NRPS offloading via a penicillin-binding protein (32)]. Our results obtained with Ebony C<sub>N</sub> serve once again to demonstrate that significant diversity exists in the enzymatic machinery behind NRPS pathways and that these alternate enzymatic systems provide clear advantages for the specific function of these assembly lines. Given the importance of the products of NRPS pathways for human health, it is crucial that we gain an understanding of the different strategies adopted by naturally occurring assembly lines for peptide synthesis if we are to undertake the reengineering of NRPS assembly lines to produce new molecules with targeted structures and function.

**Methods**

**Protein Expression and Purification.** The Ebony C<sub>N</sub> construct was cloned via PCR into the pHIS-17 plasmid, expressed in *E. coli* and purified using NINTA affinity and gel filtration. Mutants were generated using standard PCR procedures, expressed, and purified as wild type (for details see SI Appendix).

**Protein Crystallization and Structure Determination.** Datasets were collected at the Australian Synchrotron (Victoria, Australia) on either beamline MX1 or MX2 equipped with an Eiger detector (Dectris) (SI Appendix, Table S1) (33–37). For details see SI Appendix.

**Compound Synthesis.** Compounds were synthesized using standard peptide synthesis procedures (for details see SI Appendix).

**Activity Assays.** Peptide bond formation using Ebony C<sub>N</sub> was assessed with different substrates (for details see SI Appendix). ITC experiments are described in the SI Appendix.

**ACKNOWLEDGMENTS.** We thank D. Maksel and K.W.G. Kong (Monash Macromolecular Crystallisation Facility) for crystal-screening experiments; Australian Synchrotron beamline scientists (MX1 and MX2) for support discussion; D. Steer for protein MS analysis; and S. Stamatatos for protein preparation. This research was undertaken in part using the MX2 beamline at the Australian Synchrotron, part of Australian Nuclear Science and Technology Organisation, and made use of the Australian Cancer Research Foundation detector. J.A.K. acknowledges support from an Australian Government Research Training Program Scholarship. The authors acknowledge the support of Monash University, EMBL Australia, and the National Health and Medical Research Council (APP1140619 (to M.J.C.).)

**Funding.** This study was supported by the Australian Research Council (to M.J.C.).

**Author contributions.** A. P. and P. J. S. performed the majority of the biochemical and structural experiments, discussed the results, and contributed to the writing of the manuscript, and P. J. S. prepared the figures. L. Z. assisted with the mass spectrometry and the activity assays. M. J. C. supervised the work and contributed to the writing of the manuscript. S. D. T. directed the project and contributed to the writing of the manuscript. S. D. T. and M. J. C. coordinated the structural studies and contributed to these in the manuscript. All authors discussed the results and contributed to the writing of the manuscript.
5.3 Supplementary Information for “Drosophila melanogaster nonribosomal peptide synthetase Ebony encodes an atypical condensation domain”
Drosophila melanogaster non-ribosomal peptide synthetase Ebony encodes an atypical condensation domain

Thierry Izoré\textsuperscript{a,b,c,1}, Julien Tailhades\textsuperscript{a,b,c}, Mathias Henning Hansen\textsuperscript{a,b,c}, Joe A. Kaczmarski\textsuperscript{d}, Colin J. Jackson\textsuperscript{d}, Max J. Cryle\textsuperscript{a,b,c,1}

\textsuperscript{a} The Monash Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia;
\textsuperscript{b} Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia;
\textsuperscript{c} EMBL Australia, Monash University, Clayton, Victoria 3800, Australia;
\textsuperscript{d} Research School of Chemistry, The Australian National University, Acton, ACT 2601, Australia.

Corresponding authors:
Thierry Izoré (thierry.izore@monash.edu), Max J. Cryle (max.cryle@monash.edu)

SUPPLEMENTARY INFORMATION:
Construct cloning:

A construct encoding residues L666 to K879 of Ebony (*Drosophila melanogaster*, EMBL-EBI European Nucleotide Archive (ENA) ID: AJ224446.1) was amplified by PCR from a synthetic gene (Eurofins) using the following primers:

Forward 5’GGTGGTCATATGCTGAAGATGGAAGCGGTC3’
Reverse 5’ATGATGATGGATCCCTTGCCCACTTCTTTCCAGTGTACA3’

The amplicon was analyzed on a 1% agarose gel in TBE buffer and the DNA subsequently gel-extracted and further purified on column using the PROMEGA Wizard® SV Gel and PCR Clean-Up System. After digestion with NdeI and BamHI restriction enzymes, the DNA fragment and the digested pHis17 vector were cleaned-up with the above-mentioned kit and used in a ligation experiment. The ligation reaction was then transformed according to standard procedures into *E. coli* NEB5α cells. After overnight growth on LB-agar plate supplemented with ampicillin, colonies were screened by sequencing for correct insertion of the Ebony fragment.
Generation of Ebony C<sub>N</sub> mutants:

All mutants were generated following standard Quick-Change site-directed mutagenesis procedures. Primers used are listed here:

**F761A_Reverse:**
5’-CTTCGCAAAAACCTCCAGAAACTGCTACAATCAGTAACTTCGATTTGATGT-3’

**F761A_Forward:**
5’-ACATCAAATCGAAGTTACTGATTGTAGCTGAGTTTTCTGGAGTGTGTTCGAAG-3’

**F689A_Reverse:**
5’-GTTCTAAATCCGCTTTGTGTAGGCCGAAGCACAATAATGTCGATGA-3’

**F689A_Forward:**
5’-TCATCGACATTATTGTGCTTCGCGCTACAACAAAGCGGATTAGAAC-3’

**E768Q_Reverse:**
5’-GTTATCGCGGATGCTGGGCAAACCTCCAGAAACTC-3’

**E768Q_Forward:**
5’-GAGTTTCTGGAGTTTTGCGTCGCGGATAAC-3’

**H785F_Reverse:**
5’-GGTACCCATCATGAAGAAAACAAGATTTGGTTCAGCCCT-3’

**H785F_Forward:**
5’-AGGGCTGAACCAAATCTTGTTTTCCTTCATGATGGTGACC-3’

**E696L_Forward:**
5’-CAACAAAGCGGATTTACTGCAGTGGCTCAAACCG-3’

**E696L_Reverse:**
5’-CGGTTTGAGCCACTGAGCTAAATCCGCTTTGGT-3’
Protein purification:

C$_N$ Ebony pH17 wildtype construct and mutants were used for protein expression in *E. coli* BL21 (DE3) (Invitrogen) with auto-induction media supplemented with ampicillin, in which cultures were grown at 37°C to an OD$_{600}$ nm of 0.3 before the temperature was lowered to 18°C and growth continued overnight. Cells were then harvested, resuspended in lysis buffer (50 mM Tris.HCl, 300 mM NaCl, 10 mM Imidazole (pH 7.4) with Protease Inhibitors (EDTA-free Protease Inhibitor Cocktail, Sigma), lysed by sonication and centrifuged to remove cell debris. The cleared supernatant was loaded onto a 5 mL HisTRAP HP column (GE-Healthcare). The column was subsequently washed with lysis buffer and the protein was eluted via an imidazole gradient (0.01 to 1 M) over 20 column volumes. Fractions containing the protein of interest were pooled and concentrated before being loaded on a S300-HR size exclusion chromatography column (GE-Healthcare) pre-equilibrated in a buffer composed of 50 mM Tris.HCl (pH 7.8), 200 mM NaCl; fractions were subsequently assessed by SDS-PAGE. Fractions containing pure protein were pooled, concentrated via centrifugal filtration and flash frozen in liquid nitrogen at a concentration of 10.6 mg/mL.
**Compound synthesis:**

**α- / β-aminoacyl-CoA.** Boc-protected α- / β-amino acid (1 eq.) and COMU (1 eq.) were dissolved in DMF containing TEA (2 eq.) for 15 min before dropwise addition of coenzyme A (1.1 eq.) in DMF containing TEA (2 eq.). After 15h, crude Boc-aminoacyl-CoA was precipitated by addition of ice cold Et₂O and washed using centrifugation (3x). Crude Boc-β-aminoacyl-CoA was purified by preparative RP-HPLC (ACN gradient 10% to 40% over 30 min). Fractions were analyzed by LCMS, combined and freeze dried to afford Boc-aminoacyl-CoA (15-35% yield). Cleavage of the Boc group was performed for 1h in TFA/ TIS/ H₂O (95/ 2.5/ 2.5, v/ v'/ v''; 1mL) and the solution concentrated before crude β-aminoacyl-CoA was precipitated by addition of ice cold Et₂O that was then washed by centrifugation (3x). Finally, the TFA salt was exchanged into HCl to prevent unwanted pH change during soaking experiments via three cycles of dissolution in ACN/ water/ HCl (49.95/ 49.95/ 0.01, v/ v'/ v'') and freeze drying. This afforded α- / β-aminoacyl-CoA.HCl (54% yield).

**β-alanyl-dopamine (BADA).** Boc-β-alanine (1 eq.) and COMU were dissolved in DMF containing TEA (1 eq.) for 15 min before dropwise addition of dopamine hydrochloride (1.2 eq.) dissolved in DMF containing TEA (1 eq.). After 15h the reaction was quenched with saturated NaHCO₃ and extracted with ethyl acetate (3x). The combined organic layers were washed with 1M HCl (3x) and brine, dried over Na₂SO₄, filtered and concentrated under reduced pressure. Crude BADA was purified by flash chromatography using a gradient of ethyl acetate in hexane from 0% to 50% ethyl acetate (+10% over 2.5 CV). Fractions were assessed by TLC, combined and concentrated to provide Boc-β-alanyl-dopamine (92.5% yield). Cleavage and salt exchange were performed as for β-alanyl-CoA to afford β-alanyl-dopamine.HCl (63% yield).
Protein crystallization and structure determination:

Initial screening was conducted at the Monash Molecular Crystallization Facility (MMCF) with subsequent optimization performed in 48-well sitting-drop plates. Crystallization trials of Ebony C$_N$ at a concentration of 10.6 mg/mL in a 1:1 ratio (v/v) with the crystallization solution (2 µl drops) led to an initial optimized condition composed of 0.1 M HEPES pH 7.5, 24% PEG 400, 0.15 M CaCl$_2$. Crystals thus obtained were used for micro-seeding experiments that afforded well diffracting crystals in a condition composed of 0.1 M HEPES pH 7.5, 14% PEG 400, 0.2 M CaCl$_2$. Crystals were cryo-protected by passaging through reservoir solution supplemented with 30% (v/v) glycerol and flash frozen in liquid nitrogen. Ligand bound structures were obtained using a soaking approach in which crystals were soaked 20 min in 0.1 M HEPES pH 7.5, 14% PEG 400, 0.2 M CaCl$_2$ supplemented with either 2 mM dopamine, 2 mM carcinine, 2 mM β-alanyl-dopamine or 10 mM histamine (2 mM was insufficient to detect histamine in soaked crystals for this ligand). After 20 min, crystals were cryo-protected and frozen as described above. All datasets were collected at the Australian Synchrotron (Clayton, Victoria, Australia) either on beamline MX1 (1) or beamline MX2 equipped with an Eiger detector (Dectris) at 100 K (SI Table 1). Data processing was performed using XDS (2) followed by AIMLESS as implemented in CCP4 (3). Phases for the apo C$_N$ dataset were obtained by the Phaser Molecular Replacement module in PHENIX (4) using a model generated by the ROBETTA Web-server from the PDB file 3QB8 (5). Phases for the datasets with bound substrates and products were obtained from the refined structure of the apo enzyme. Structures were built and refined using COOT (6) for model building and PHENIX-refine for refinement (using BUSTER refinement to generate ligand densities (7)). Polder maps were generated in PHENIX (8). All graphics were generated with Pymol.
**Supplementary Table 1:** Crystallographic data for collection and refinement of $C_N$ structures

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Apo-$C_N$</th>
<th>$C_N$-histamine</th>
<th>$C_N$-dopamine</th>
<th>$C_N$-carcinine</th>
<th>$C_N$-β-alanyl- dopamine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P 2(1)</td>
<td>P 2(1)</td>
<td>P 2(1)</td>
<td>P 2(1)</td>
<td>P 2(1)</td>
</tr>
<tr>
<td>Cell dimensions (Å)</td>
<td>a = 44.96, b = 56.49, c = 50.52 (*), α = 90.0, β = 113.23, γ = 90.0</td>
<td>a = 44.58, b = 57.47, c = 50.13 (*), α = 90.0, β = 113.23, γ = 90.0</td>
<td>a = 44.97, b = 57.19, c = 50.43 (*), α = 90.0, β = 114.0, γ = 90.0</td>
<td>a = 44.52, b = 56.05, c = 49.73 (*), α = 90.0, β = 112.9, γ = 90.0</td>
<td>a = 44.48, b = 56.49, c = 49.93 (*), α = 90.0, β = 113.5, γ = 90.0</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>2.00</td>
<td>2.10</td>
<td>2.80</td>
<td>2.45</td>
<td>2.30</td>
</tr>
<tr>
<td>Beamlines</td>
<td>MX-2</td>
<td>MX-1</td>
<td>MX-2</td>
<td>MX-1</td>
<td>MX-1</td>
</tr>
<tr>
<td>Rpim</td>
<td>0.06 (0.39)</td>
<td>0.05 (0.40)</td>
<td>0.1 (0.50)</td>
<td>0.1 (0.47)</td>
<td>0.05 (0.23)</td>
</tr>
<tr>
<td>I/σ(I)</td>
<td>8.2 (2.3)</td>
<td>11.9 (2.0)</td>
<td>6.1 (1.8)</td>
<td>8.0 (1.7)</td>
<td>14.2 (3.7)</td>
</tr>
<tr>
<td>CC1/2</td>
<td>0.7</td>
<td>0.63</td>
<td>0.59</td>
<td>0.65</td>
<td>0.856</td>
</tr>
<tr>
<td>R/Rfree</td>
<td>0.18/0.20</td>
<td>0.18/0.25</td>
<td>0.20/0.25</td>
<td>0.18/0.23</td>
<td>0.19/0.21</td>
</tr>
<tr>
<td>Ramachandran Favored (%)</td>
<td>99</td>
<td>99</td>
<td>95</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>Molprobity Score</td>
<td>1.27 (99th percentile)</td>
<td>1.18 (100th percentile)</td>
<td>1.47 (100th percentile)</td>
<td>1.35 (100th percentile)</td>
<td>1.21 (100th percentile)</td>
</tr>
</tbody>
</table>

PDB ID | 6DYM | 6DYN | 6DYO | 6DYM | 6DYS |

\[ R_{p,l,m.} = \frac{\sqrt{\sum_{l=1}^{L} \sum_{m=1}^{M} [I_{l,m} - \langle I_{l,m} \rangle]^2}}{\sum_{l=1}^{L} \sum_{m=1}^{M} I_{l,m}} \]

Numbers into bracket refer to the highest resolution shell
Isothermal Titration Calorimetry:

All isothermal titration calorimetry (ITC) experiments were performed on a Nano-ITC low-volume calorimeter (TA Instruments); details of instrument calibration have been described previously (9). ITC experiments were performed at 25 °C with stirring at 150 rpm. Protein and ligand solutions were prepared in matched ITC buffer (50 mM TRIS pH 7.8, 200 mM NaCl, 5 % v/v glycerol) and degassed before use. Solutions of dopamine, histamine, β-alanyl-dopamine and carcinine (β-alanyl histamine) were prepared volumetrically. Protein concentrations were calculated spectroscopically using the extinction coefficient 26930 M⁻¹ cm⁻¹ as calculated using ExPASy ProtParam server (10). Titrations were performed as follows: for dopamine, 34 µM Ebony, 1 x 1 µl injection, followed by 18 x 2.5 µl injections of 0.9 mM dopamine; for histamine, 174 µM Ebony, 1 x 1 µl injection, followed by 18 x 2.5 µl injections of 2.7 mM histamine; for β-alanyl-dopamine, 170 µM Ebony, 1 x 1 µl injection, followed by 18 x 2.5 µL injections of 1.8 µM β-alanyl-dopamine; for carcinine, 90 µM Ebony, 1 x 1 µl injection, followed by 14 x 3 µl injections of 0.9 mM carcinine. Raw data, fits and affinity constants are provided in Supplementary Figure 7. Data were analyzed using NITPIC (11), SEDPHAT (12) and GUSSI (13); the baseline-subtracted power was integrated, and the integrated heats were fit to the single binding site model \((A + B \leftrightarrow AB)\) hetero-association model to obtain the association constant \((K_a)\). Fitting was achieved by iteratively cycling between Marquardt-Levenberg and Simplex algorithms in SEDPHAT until modelling parameters converged. Data represents individual titrations. 68.3% confidence intervals were calculated using the automatic confidence interval search with projection method using F-statistics in SEDPHAT (12).
**Supplementary Table 2:** Isothermal titration calorimetry results. Affinities were determined by ITC at 25 °C with stirring at 150 rpm. 68.3% confidence intervals (1 Std. Dev.) are given for $K_a$ and $K_d$. Raw data are provided in Supplementary Figure 7.

<table>
<thead>
<tr>
<th></th>
<th>Dopamine $K_a$ (M$^{-1}$, x 10$^3$)</th>
<th>Histamine $K_a$ (M$^{-1}$, x 10$^3$)</th>
<th>β-alanyl-dopamine $K_a$ (M$^{-1}$, x 10$^3$)</th>
<th>Carcinine $K_a$ (M$^{-1}$, x 10$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_a$ (M$^{-1}$, x 10$^3$)</td>
<td>31.3 (26.2, 37.4)</td>
<td>1.6 (0.9, 2.2)</td>
<td>17.7 (14.2, 21.0)</td>
<td>4.6 (4.00, 5.2)</td>
</tr>
<tr>
<td>$K_d$ (μM)</td>
<td>32.0 (26.9, 38.1)</td>
<td>638.9 (446.7, 1090.7)</td>
<td>56.4 (47.7, 70.4)</td>
<td>218.6 (194.4, 250.8)</td>
</tr>
</tbody>
</table>

**Activity assays:**

Peptide bond formation using Ebony CN was assessed with different substrates. These include β-alanyl-CoA (1), alanyl-CoA, aspartyl-CoA, threonyl-CoA, leucyl-CoA, tryptophanyl-CoA, prolyl-CoA, phenylalanyl-CoA, glycyl-CoA, and 6 analogues of β-alanyl-CoA (2-7) 100 μL reactions consisted of the Ebony CN (400 μM), aminoacyl-CoA (10 mM) and dopamine (10mM) in 50 mM Tris-HCl (pH 7.4), 100 mM NaCl. After 3 h incubation at 25 °C, the reactions were diluted with ACN (200 μL), incubated for 10 min at 60 °C and the samples centrifuged to remove insoluble protein before being
dried, resuspended in 50 µL of 10% ACN in water and analyzed by LCMS. All assays have been made in triplicates.

**LCMS Analysis:**

Compounds were analyzed using a Shimadzu High Performance Liquid Chromatograph coupled to Mass Spectrometer LCMS-2020 (ESI, operating both in positive and negative mode) equipped with a SPD-20A Prominence Photo Diode Array Detector and a LC-20AD solvent delivery module. Analysis used a Waters XBridge BEH300 Prep C18 column (10 µm, 4.6 x 250 mm) at a flow rate of 1 mL/min. The solvents used were water + 0.1% formic acid (solvent A) and HPLC-grade ACN + 0.1% formic acid (solvent B).

**Preparative HPLC:**

Compound purification was performed using a Shimadzu High Performance Liquid Chromatograph equipped with a SPD-M20A Prominence Photo Diode Array Detector and two LC-20AP pumps. Purification used a Waters XBridge BEH300 Prep C18 column (5 µm, 19 x 150 mm) at a flow rate of 10 mL/min. The solvents used were water + 0.1% TFA (solvent A) and ACN + 0.1% TFA (solvent B).
Supplementary Figure 1: a) Structure-based alignment of Ebony C_N, Dopamine-NAT, Polyamine-NAT, Agmatine-NAT that indicates the very low sequence identity despite a very well conserved structure. Purple stars represent residues from the “hydrophobic cage”, green stars represent residues from the catalytic site. b) Cartoon representation of
Ebony Cₙ, Dopamine-NAT, Polyamine-NAT, Agmatine-NAT structures. c) Charge-colored surface representation of Ebony Cₙ, Dopamine-NAT, Polyamine-NAT, Agmatine-NAT structures. Note the differences in the CoA-binding site (NATs) and the corresponding hydrophobic flat surface on Ebony Cₙ.
**Supplementary Figure 2**: Cartoon representations of Ebony C_N along with Agmatine-NAT, Polyamine-NAT and Dopamine-NAT showing the differences in the region controlling access to the catalytic channel (colored in blue): note the differences in structure of this region across the four homologues.
Supplementary Figure 3: Representation of electron densities corresponding to the bound substrate/product molecules within the various Ebony C_N structures. Left: Unbiased
densities showing maps obtained after molecular replacement using the structure of apo-Ebony. Green/ red density maps represents Fo-Fc densities (sigma level of 3 for histamine, β-alanyldopamine and carcinine, 2.5 for dopamine); blue density map is 2Fo-Fc density (sigma level of 1.5). All generated using refinement in BUSTER (7). Right: Polder maps (8) for all ligands at the end of the refinement. Cross-Correlation coefficients (CC) are shown for (1): calculated Fobs with ligands, (2): calculated Fobs without ligand, (3): real Fobs data. A larger value for CC(1;3) compared to CC(1;2) or CC(2;3) is considered strong support for the presence of the ligand. To confirm these values were not due to bias occurring during refinement, Polder maps were also calculated from a hybrid model where each ligand has been positioned into the Ebony C\textsubscript{N} apo structure and refined against the apo dataset. The Polder maps produced for these hybrid models all indicated low CC(1;3) values, which indicates no ligand density (but rather bulk solvent) is present in the apo structure. This further supports the presence of all ligands in the soaked datasets.
**Supplementary Figure 4:** a) Comparison of histamine-bound Ebony C<sub>N</sub> (cyan) and dopamine NAT in complex with acetyl-CoA (yellow) showing that the amino-group of the histamine substrate bound would be ideally positioned to attack the incoming PCP-bound β-alanine thioester group based on the structure of the acetyl-CoA molecule bound in the dopamine NAT structure. b) Carcinine-bound Ebony C<sub>N</sub> (green) superimposed with dopamine NAT in complex with acetyl-CoA (yellow) showing that the carbonyl groups in both acetyl-coA and carcinine occupy very similar positions.
**Supplementary Figure 5:** Triplicate LCMS analyses indicating the reconstitution of Ebony C_N domain for dopamine together with β-alanyl-CoA (1) as substrates. Top panels show reconstitution of the wild type enzyme as well as the negative control, in which no C_N domain was present. Selected ion current displayed corresponds to the protonated product 1a corresponding to that of the substrate 1 (i.e. for C_N-catalyzed conversion of 1,
m/z = 225.1 (1a)), with the reaction outlined in the box. Lower panels show the same reactions performed for various mutants of the C_N domain, which have varying degrees of influence on the reaction outcome. The retention time for an authentic standard of 1a is indicated in the panels (t_R = 6.27 min).
Supplementary Figure 6: Triplicate LCMS analyses indicating the tolerance of Ebony C_N domain for various modified CoA substrates (2-7) in a condensation assay together with dopamine to produce products 2a-7a. Selected ion current displayed corresponds to the protonated product corresponding to that of the specific substrate (i.e. for C_N-catalyzed conversion of 2, m/z = 210.1 (2a); for C_N-catalyzed conversion of 3, m/z = 254.1 (3a); for C_N-catalyzed conversion of 4, m/z = 239.1 (4a); for C_N-catalyzed conversion of 5, m/z = 239.1 (5a); for C_N-catalyzed conversion of 6, m/z = 239.1 (6a); for C_N-catalyzed conversion of 7, m/z = 239.1 (7a)).
Supplementary Figure 7. Representative ITC isotherm data for interactions between Ebony $C_N$ and small molecules. a) Titration of Ebony $C_N$ with dopamine. b) Titration of Ebony $C_N$ with histamine. c) Titration of Ebony with $\beta$-alanyl-dopamine. d) Titration of Ebony $C_N$ with carcinine. The upper panels represent baseline-corrected power traces. By convention, negative power corresponds to exothermic binding. The middle panels represent the integrated heat data fitted to the single binding sites model in SEDPHAT.
(12). The bottom panels show the residuals of the fit. Error bars are standard error in the integration of the peaks as calculated by NITPIC (11). Figures were produced using GUSSI (13).
Supplementary References:


5.4 Paper 2: Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics

5.4.1 Publication status

This manuscript presented in this section has been published by *Nature Communications*. Supporting information is provided in Section 5.5.

5.4.2 Author’s contribution

I designed and performed the ITC experiments for this publication. I analyzed the data and produced the figures and tables presenting the ITC data. I helped with the interpretation of the ITC data with respect to the other data presented in this work. I helped with editing the final manuscript.

5.4.3 Contributions from others

The majority of the work presented in the following paper was done by members of Associate Professor Max Cryle’s group at Monash University. Professor Colin Jackson helped with the interpretation of the ITC data.
Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics


Kistamicin is a divergent member of the glycopeptide antibiotics, a structurally complex class of important, clinically relevant antibiotics often used as the last resort against resistant bacteria. The extensively crosslinked structure of these antibiotics that is essential for their activity makes their chemical synthesis highly challenging and limits their production to bacterial fermentation. Kistamicin contains three crosslinks, including an unusual 15-membered A-O-B ring, despite the presence of only two Cytochrome P450 Oxy enzymes thought to catalyse formation of such crosslinks within the biosynthetic gene cluster. In this study, we characterise the kistamicin cyclisation pathway, showing that the two Oxy enzymes are responsible for these crosslinks within kistamicin and that they function through interactions with the X-domain, unique to glycopeptide antibiotic biosynthesis. We also show that the kistamicin OxyC enzyme is a promiscuous biocatalyst, able to install multiple crosslinks into peptides containing phenolic amino acids.

1Department of Biochemistry and Molecular Biology, The Monash Biomedicine Discovery Institute, Monash University, Clayton, VIC 3800, Australia. 2EMBL Australia, Monash University, Clayton, VIC 3800, Australia. 3Interfaculty Institute of Microbiology and Infection Medicine Tübingen, Microbiology/Biotechnology, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany. 4Department of Biomolecular Mechanisms, Max Planck Institute for Medical Research, Jahnstrasse 29, 69120 Heidelberg, Germany. 5Department of Chemistry, The University of Queensland, St Lucia, QLD 4072, Australia. 6Monash Biomedical Proteomics Facility, Monash University, Clayton, VIC 3800, Australia. 7Research School of Chemistry, The Australian National University, Acton, ACT 2601, Australia. 8German Centre for Infection Research (DZIF), Partner Site Tübingen, 72076 Tübingen, Germany. 9These authors contributed equally: Thierry Izoré, Dumitrita Iftime, Julien Tailhades. Correspondence and requests for materials should be addressed to E.S. (email: evi.stegmann@biotech.uni-tuebingen.de) or to M.J.C. (email: max.cryle@monash.edu)
The glycopeptide antibiotics (GPAs) are a series of complex, peptide-based antibiotics that have been utilised in a clinical setting since the late 1950s and are exemplified by vancomycin. Structurally, GPAs consist of a heptapeptide backbone that is heavily crosslinked via the side chains of aromatic amino acids within the peptide sequence (Fig. 1a). The crosslinking of the peptide backbone rigidifies the GPA peptide aglycone, allowing this to form a non-covalent complex with the dipeptide terminus of the cell wall precursor lipid II. By binding lipid II, GPAs are able to prevent the formation of the bacterial cell wall, hence leading to their antimicrobial activity. The crosslinks within GPAs are therefore essential for the activity of these antibiotics and in addition are also the most challenging part of the total synthesis of these important molecules. Given the difficulties of their commercial synthesis through traditional synthetic means, all current production of GPAs remains reliant on the bacterial fermentation of the natural producer strain, which in turn makes understanding this machinery of great practical as well as scientific interest.

GPAs are produced by the action of two powerful biosynthetic systems. First, a non-ribosomal peptide synthetase (NRPS) machinery produces the linear heptapeptide GPA backbone, the core of which largely consists of non-proteinogenic phenylglycine residues. Second, an oxidative cyclisation cascade consisting typically of three to four cytochrome P450 (Oxy) enzymes (one enzyme per ring formed, see Fig. 1b) is then recruited to the NRPS-bound heptapeptide in order to perform the stepwise cyclisation of the peptide into a rigid, active antibiotic. The recruitment of the Oxy enzymes to the NRPS machinery is mediated by a unique P450 recruitment domain found within the NRPS, known as the X-domain. Different GPAs are classified based on the residues found in their peptide backbone, the crosslinking pattern of the final GPA and the modifications installed following construction of the crosslinked peptide aglycone. Most common GPAs with appreciable antibiotic activity are found in Types I-IV: Type I/II include GPAs with three crosslinks within the peptide backbone and aliphatic or aromatic residues at positions 1/3, respectively. Type III and IV GPAs contain an additional crosslink between aromatic residues at positions 1/3 of the peptide, with Type III/IV differentiated by the absence or presence of an aliphatic chain linked to a peripheral sugar moiety (Fig. 1a).

Due to the clinical utility of GPAs and the challenges associated with the synthesis of these complex molecules, significant research efforts have been made to understand the cyclisation process of GPAs from both an in vivo and in vitro standpoint. These studies have revealed that each Oxy enzyme is responsible for the installation of one crosslink at the heptapeptide stage (Fig. 1b) and that a specific order of activity exists (OxyB, OxyE, OxyA, OxyC), suggesting a potential for expensive catalysts for the reasons, there is great interest in an unusual class of GPAs—known as Type V GPAs—that include the representative members complestatin and kistamicin. Type V GPAs, whilst the most structurally divergent GPA class, share sufficient similarities in their structure and biosynthesis to be classed as GPAs in spite of their lack of glycosylation and reduced antibacterial activity (MIC values reported for complestatin and kistamicin against S. aureus of ~2 μg mL⁻¹ to 32 μg mL⁻¹ as opposed to 0.5–1 μg mL⁻¹ for vancomycin). Type V GPAs display divergent activity when compared to Type I-IV GPAs, including antiviral activity for both kistamicin and complestatin and as well as other types of potential antibacterial activity, clearly makes such atypical GPAs of great interest for diversification of the activity of GPAs.

One major structural difference in Type V GPAs is the DE aryl crosslink formed from the tryptophan residue at position 2 of the heptapeptide, whilst kistamicin in particular also contains an additional unusual ring, an A-O-B ring formed from a C-terminal 4-hydroxyphenylglycine (Hpg) residue. Furthermore, the kistamicin gene cluster only contains two Oxy enzymes despite the presence of three crosslinks in the final product, whilst the NRPS machinery also contains features that are not found in other GPAs. Given these differences in structure and biosynthetic machinery found in the oxidative cyclisation cascade of kistamicin biosynthesis, we have engaged in a detailed structural, biochemical and functional characterisation of this crosslinking cascade in this unique GPA.

Here, we show that the three crosslinks that occur during the kistamicin peptide construction are catalysed by only two Oxy enzymes, with the activity of the kistamicin Oxy enzymes being distinct from those found in the biosynthesis of typical Type I-IV GPAs. Despite these differences, the kistamicin Oxy enzymes still rely on recruitment to the peptide bound to the NRPS by interaction with the X-domain, a unique recruitment domain found in GPA biosynthesis. Through the biochemical studies and structural characterisation of an OxyAkis/X-domain complex we can now show that not only is the X-domain mediated Oxy recruitment mechanism conserved across all known types of GPA biosynthetic machineries but also that the interface between all Oxs and the X-domain is conserved, supporting a shuffling mechanism for Oxy activity. Within the kistamicin cyclisation cascade, we show that the kistamicin OxyC enzyme (OxyCkis) acts first, catalysing insertion of the C-O-D ring, which is followed by the actions of OxyAkis installing the DE aryl crosslink. Whilst it is not possible to unambiguously assign the insertion of the A-O-B ring to OxyCkis activity, supporting evidence from our in vivo and in vitro activity studies indicates that OxyCkis is a highly promiscuous enzyme that is capable of peptide bicyclisation and hence the most probable candidate for this biotransformation.

Results

The kis cluster reveals a divergent biosynthetic apparatus. Due to the unusual structure of kistamicin when compared with typical type I-IV GPAs, we initially sequenced the genome of Actinomadura parvosa subsp. kistnae (Nonomuraea sp. ATCC55076) to identify the kistamicin biosynthetic gene cluster. The kistamicin (kis) biosynthetic gene cluster spans ~60 kb with 33 open reading frames (Supplementary Fig. 1). During the course of our investigations, this cluster was also revealed by the Seyedsayamdeh and Gulder groups. Strikingly, the biosynthetic gene cluster of this GPA encodes only two cytochrome P450 enzymes (KisN and KisO) implicated in the crosslinking of the aromatic side chains of the peptide, despite there being three crosslinks in kistamicin. A phylogenetic analysis of the Oxy enzymes from kistamicin biosynthesis revealed that these fall into the OxyA (KisN) and OxyC (KisO) families and do not cluster with OxyB, the first enzyme involved in cyclisation cascade of the Type I-IV GPAs (Supplementary Fig. 2). This was unexpected, given that OxyB-catalysed insertion of the C-O-D ring is an essential prerequisite for the activity of subsequent Oxy enzymes in Type I-IV systems. The OxyAkis enzyme is expected to install the DE ring based on comparison to the complestatin system, which led us to hypothesise that the OxyCkis enzyme was responsible for installation of both the A-O-B as well as C-O-D rings in kistamicin.

Beyond the Oxy enzymes, the kistamicin NRPS machinery maintains the unique GPA X-domain for Oxy recruitment to the NRPS-bound peptide, but contains several differences to typical GPA NRPS assembly lines, including in modules 1
Fig. 1 GPA structures and NRPS/Oxy interplay. a Structure of different types of GPAs shown for vancomycin (Type I), A47934 (Type III), teicoplanin (Type IV), kistamicin (Type V) and complestatin (Type V). Major structural differences between kistamicin and Type I-IV GPAs include: the replacement of the AB ring required for the activity of standard GPAs with an enlarged A-O-B ring; the replacement of the D-O-E ring with a DE ring that incorporates a tryptophan residue; the opposite configuration of the 3rd peptide residue (D instead of L); and the lack of glycosylation. C-terminal residues of GPAs are coloured for 3,5-dihydroxyphenylglycine (3,5-Dpg, light blue) and 4-hydroxyphenylglycine (4-Hpg, violet). P450 catalysed intramolecular crosslinks are indicated in pink. b P450-catalysed crosslinking cascade performed by first OxyB, (then optional OxyE), OxyA and finally OxyC into type I-IV GPA heptapeptides. Numbering of amino acids is listed according to the timing of amino acid incorporation, i.e. from 1–7. c Part of the respective GPA biosynthetic gene clusters involving the NRPS (yellow) and the P450 enzymes (pink); other genes are also indicated (white). The NRPSs are divided into seven modules with a number of catalytically active domains (yellow circles). In the last module, the Oxy-recruiting domain X is present (light pink). Additional domains (E domains (blue), MT domain (green)) and missing domains (red cross) are indicated. Presence and type of Oxy within the GPA biosynthesis is shown (pink circles). Kistamicin has three crosslinks, but only two Oxy encoding genes in the cluster, an OxyA and an OxyC enzyme. Domain definitions: A, adenylation domain; C, condensation domain; E, epimerisation domain; PCP, peptidyl carrier protein; MT, methyltransferase; X, Oxy-recruiting domain; TE, thioesterase domain.

Expanded A–O–B ring

Lack of glycosylation

Kistamicin

Trp-containing DE ring

Teicoplanin

Vancomycin

A47934

Complestatin

Type V

Type IV

Type III

Type I

Two oxy genes
– three cross-links

Cross-linking cascade
in GPA I-IV

1. OxyB

2. OxyA

3. OxyC

(OxyE)
oxy gene deletions reveal the kistamicin cyclisation cascade.

Given the biosynthetic differences implicit in the sequence of the kistamicin gene cluster when compared to other GPA clusters, we first isolated and characterised kistamin from the A. parvosata producer strain. Our structural analysis indicated that there was no difference between the structure of the compound that we had isolated and that previously reported (Supplementary Figs. 4–11, Supplementary Table 1).18.

To ascertain the function of the Oxy enzymes within the kistamicin cluster, gene deletion mutants of OxyA<sub>kis</sub> (A. parvosata ΔkisN) and OxyC<sub>kis</sub> (A. parvosata ΔkisO) were prepared (Supplementary Figs. 12–16, Supplementary Table 2–4). After cultivation, kistamicin production was found to be abolished in both strains (Fig. 2, Supplementary Fig. 17). HRMS analysis of the isolates showed that the OxyC<sub>kis</sub> deletion strain only produces linear hexa- and heptapeptides, whilst the OxyA<sub>kis</sub> deletion strain produced monocylic hexa- and heptapeptides (Fig. 2). MS/MS analysis of these metabolites showed that the ring was inserted between Tyr-6 and Hpg-4, which corresponds to the C-O-D ring that is the first ring installed in all GPAs studied to date (Supplementary Figs. 18–23).10,25,26. This also indicates that the order of ring insertion activity matches that from other GPAs, with the DE ring required before insertion of the A-O-B ring can occur.10,25,26. The appearance of monocylic hexapeptide is in agreement with experiments from other GPA producer strains that have shown that the initial ring insertion is the most facile and can occur at the hexapeptide stage if the NRPS machinery has stalled due to deletion of Oxy enzymes.28. Production of kistamin by the OxyA<sub>kis</sub> (A. parvosata ΔkisN) deletion mutant strain was recovered through complementation of the gene on an integrative plasmid (Supplementary Figs. 17 and 19). These results support OxyC<sub>kis</sub> acting first during the cyclisation of kistamin, installing the C-O-D ring into the kistamin peptide.

We next investigated lower abundance species in the OxyA<sub>kis</sub> (A. parvosata ΔkisN) deletion mutant for evidence of bicyclic peptide formation. HRMS showed that low levels of bicyclic hexapeptide intermediates were indeed present in the OxyA<sub>kis</sub> deletion strain (Fig. 2). Given that the seventh residue (involved in the A-O-B ring) is missing from this peptide, we next used MS/MS analysis to determine that a crosslink was formed between Tyr-6 and Hpg-4 in addition to the Tyr-6/ Hpg-4 C-O-D crosslink present in the monocylic species (Supplementary Figs. 24–26). This suggests that the OxyC<sub>kis</sub> enzyme can be promiscuous and is capable of installing two crosslinks in a single peptide, with the second ring then formed after initial C-O-D ring insertion. However, OxyC<sub>kis</sub> is not able to form the DE ring, which in turn is required for A-O-B ring formation.

![Fig. 2 Production of kistamic intermediates in A. parvosata WT and Δoxy mutants.](image-url)

- **a** A. parvosata wildtype (WT) produces kistamin A, a tricyclic heptapeptide.
- **b** A. parvosata ΔoxyA (ΔkisN) produces several intermediates from kistamin biosynthesis, including monocylic hexa-/heptapeptides and different bicyclic hexapeptides (all structures contain the Tyr-6/ Hpg-4 C-D ring).
- **c** A. parvosata ΔoxyC (ΔkisO) produces only the linear hexa-/heptapeptide intermediates from kistamin biosynthesis.

(missing the core sub-domain of the adenylation domain), 3 (addition of an epimerisation (E)-domain) and 4 (missing an E-domain). The presence of an E-domain in module 3 is in agreement with the reported stereochemistry of the 3<sup>rd</sup> (Dpg) residue in kistamin, whilst the lack of an E-domain in module 4 is highly unusual given the presence of a (D)-configured amino acid in position four of all GPAs characterized to date. Given the importance of this residue, alteration of the stereochemistry of this position to the (L)-form would appear highly unlikely (although would be necessary to explore, vide infra), thus implying a dual C/E function for one of the C-domains between modules 3/4 or 4/5. A phylogenetic analysis of the kistamin C-domains did not reveal obvious clues to the identity of the dual C/E-domain, and confirmation of the identity of this domain awaits future characterisation (Supplementary Fig. 3).
Complementation of the previously published oxyA and oxyC deletion mutants from Type I (balhimycin)25 and Type IV (A47934)10 GPA producer strains by the kistamicin oxyA/kis/ oxyC/kis genes did not lead to recovery of antibiotic production in these strains as assessed by inhibition zone assays.

Computational analysis of kistamicin crosslinking. Based on these complementation studies, we next examined whether the inability of OxyCkis to complement the Type I/IV OxyC knockout strains was due to differences in ring strain in the different GPA structures or whether the type of crosslink installed (aryl vs phenolic) was an inherent property of different Oxy enzymes. To understand whether the outcome of crosslinking, namely AB vs A-O-B ring formation is related in any way to the strain energies of the rings, we performed molecular mechanics and density functional theory computations (Fig. 3, see Supplementary Figs. 27–31 and Supplementary Information for details). Upon comparison of models of native kistamicin, featuring a 15-membered A-O-B ring (1), with two unnatural analogues containing 13- or 12-membered AB rings (2/3), we found unexpectedly that the A-O-B ring was significantly less stable (by about 10 kcal/mol) than the smaller-ring AB isomers. Interestingly, a similar observation was made when comparing the native 16-membered C-O-D ring of kistamicin with a 14-membered CD analogue. In this case, the 14-membered biaryl-linked CD structure corresponds to the known crosslinking pattern in the arylomycins. A third example mirroring the preference for biaryl linkages was observed when comparing the native 12-membered AB ring of Type I/IV GPAs with a 14-membered A-O-B analogue.

These results suggest that a low stability of phenolic crosslinks relative to diaryl crosslinks is a common feature among GPAs possessing Dpg or Hpg residues at the C-terminus. Indeed, this trend parallels the intrinsic energy difference between diphenyl ethers and biaryl alcohols (the latter being more stable than the former). In the case of kistamicin, an additional contributing factor to the high energy of the A-O-B ring is found to arise from ring strain; in particular, formation of this ring requires a typical P450 fold, which is largely composed of α-helices7. The active site heme iron is coordinated by the thiolate of the conserved, proximal Cys332 residue, whilst one region of β-sheets5. The active site heme iron is coordinated by the thiolate of the conserved, proximal Cys332 residue, whilst the I-helix that runs across the top of the heme contains the

**Interaction of the Oxy enzymes with the kistamicin X-domain.** To study the potential interactions between the domains of the final NRPS module and the Oxy enzymes, we designed, expressed and purified OxyA/kis, OxyC/kis, the Xkis-domain alone as well as naturally fused to the preceding peptidyl carrier protein (PCP)-domain (PCP-Xkis)9,29. Due to the similar size of the X-domain and Oxy enzymes, we performed N-terminal labelling of the X-domain with fluorescein to allow a specific wavelength to be monitored for each protein. These experiments showed evidence of an interaction between the kistamicin Oxy enzymes and the X-domain independent of the presence of the adjacent PCP domain. This was evidenced by the formation of bands of higher molecular weight in the lanes where X-domain and Oxy enzymes were present (Fig. 4a–c). Whilst the common interaction partner is the X-domain itself, the native PAGE experiments also suggest that the presence of the PCP domain (albeit unloaded) may increase affinity for the Oxy enzymes.

With the limited resolution of these measurements, we next performed isothermal titration calorimetry (ITC) experiments to quantify the binding interactions occurring in the kistamicin system (Fig. 4d–f). These experiments showed that both Oxy enzymes associated with the X-domain, with the K_{d} obtained for both in the low micromolar range (OxyA/kis 6.1 μM, OxyC/kis 10.4 μM). These values match the affinities determined for the teicoplanin system using alternative biochemical methods9,30, and indicate that binding mostly driven by entropic effects, presumably due to the release of ordered water molecules from the interface. Given that the Oxy/NRPS interactions appear to be conserved despite the unusual structure of kistamicin and the non-standard crosslinking cascade, we next sought to understand this process at a molecular level.

**Characterisation of the kistamicin OxyA/ X-domain complex.** In order to gain further insight into the nature and sites of the Oxy/X-domain interactions from kistamicin biosynthesis, we were able to obtain crystals of a complex of the OxyA/kis protein and X-domain from kistamicin biosynthesis that diffracted to a resolution of 2.6 Å (Supplementary Table 5, Fig. 5). Analysis of the separate protein domains showed that both the OxyA/kis enzyme9,31–37 and the X-domain9 maintained their anticipated folds, with little difference to comparable reported structures (Fig. 5–6; Supplementary Tables 6–7). OxyA/kis shows the archetypal P450 fold, which is largely composed of α-helices aside from one region of β-sheets8. The active site heme iron is coordinated by the thiolate of the conserved, proximal Cys332 residue, whilst the I-helix that runs across the top of the heme contains the

---

**Fig. 3** Computational analysis of alternative A-B ring crosslinks. Comparison of the relative energies of kistamicin (containing the native 15-membered A-O-B ring and modelled as 1) relative to the energies of kistamicin analogues 2 and 3 which contain 13- and 12-membered AB rings, respectively. The ring derived from A-O-B crosslinking is significantly less stable than the smaller rings derived from AB crosslinking. Source data are provided as a Source Data file.
Glu/Gln pair of residues (Glu228, Gln229) that are required in P450 enzymes to maintain an effective protonation cascade during oxygen activation. The closest structural homologue to OxyA$_{X}$ was identified as the related GPA cyclisation enzyme OxyA$_{F}$ (RMSD 1.5 Å), which includes the same orientations of the capping F- and G-helices above the active site (Supplementary Fig. 32, Supplementary Table 6) [34]. The structure of OxyA$_{X}$ reveals gaps in the observable density of several of the mobile regions surrounding the active site, which is unsurprising given the lack of substrate peptide bound to this P450 (Fig. 6).

The X$_{X}$ domain within the OxyA$_{X}$ complex adopts the typical C-domain fold composed of a pseudo-dimer of chloramphenicol acetyl transferase-like domains (Fig. 5–6). It is similar to the teicoplanin X-domain (solved in complex with OxyB$_{tei}$) in sequence (53.4%) and in structure, showing an RMSD of only 1.6 Å (Supplementary Fig. 32, Supplementary Table 7) [9]. The high structural similarity between these two X-domains stands in contrast to the much higher RMSD values seen with other condensation or epimerisation domains (Supplementary Table 7) [38–43], which is typically due to differences in the relative orientations of the N- and C-terminal subdomains [6]. The only noticeable differences between the structures of X$_{X}$ and X$_{tei}$ stem from some slight motions at the end of the loop from region encompassing residues 1399–1411 (visible in the OxyA$_{X}$ complex, not in the OxyA$_{tei}$ complex) and differences in the turn between β-sheets (residues 1436–1441). The X$_{X}$ domain shows the same mutated active site residues as seen in the X$_{tei}$ structure, which includes the mutation of crucial histidine and glycine residues (HHxxxDG to HRxxxDE) that render the X-domain inactive for condensation or epimerisation activity as well as blocking the access channel to the active site itself (Fig. 6).

In terms of the interaction between the two proteins, the X$_{X}$ domain remains closely associated to the PxxD motif at the start of the OxyA$_{X}$ F-helix, but now the X-domain has rotated in towards the C-terminus of OxyA$_{X}$ increasing the number of favourable protein–protein contacts around the OxyA$_{X}$ D/E helices and connecting loop (Fig. 5). Burial of a substantial surface...
area is observed in both complexes, consistent with the positive binding entropy, with the OxyAkis/Xkis complex displaying a slightly larger buried surface area than the OxyBtei/Xtei complex (976 Å² vs 725 Å², respectively; Supplementary Table 8). There are a number of H-bond and salt bridge interactions (17 and 20, respectively) in both complexes, which are likely to be important in controlling binding specificity. The OxyAkis/Xkis complex is consistent with the ITC data and serves to confirm both the general binding mode of Oxy enzymes to X-domains during GPA biosynthesis and also the importance of the polar interactions in the interface, presumably to maintain specificity for Oxy/X binding.

**In vitro reconstitution of kistamicin Oxy enzyme activity.** To explore the cyclisation activity by the kistamicin Oxy enzymes in vitro, we first established that these enzymes maintained the essential heme thiolate ligation state following *E. coli* over-expression by obtaining reduced, carbon monoxide-bound UV/vis spectra showing the typical P450 absorbance of active enzymes, albeit with ~50% protonated, inactive P420 component (Supplementary Fig. 33). To explore the activity of OxyAkis and OxyCkis, we synthesised a series of nine peptides based on the kistamicin sequence that varied in length from 3–7 amino acids. Structurally, these explored the effects of altering the stereochemistry of Hpg-4, and also the role of the stereochemistry/structure of Dpg-3. All peptides were synthesised using conditions reported to avoid Hpg/Dpg racemisation. Synthesis was performed on hydrazine resin to enable generation of the corresponding peptidyl CoAs following resin activation and thioester formation with CoA (Supplementary Figs. 34–43, Supplementary Table 9). Following purification, the peptides were loaded onto the PCP-Xkis construct using the phosphopantetheinyl transferase Sfp from *B. subtilis* and
subsequently incubated with the enzymes OxyAkis, OxyCkis or OxyBtei (or a combination thereof) and appropriate redox system before cleavage of the thioesters with methylamine and analysis by LCMS (Fig. 7, Supplementary Dataset 1, Supplementary Figs. 44–52). OxyCkis proved to be highly active on tetrapeptides (K3–6, K4–7) that covered the C-O-D ring (positions 4–6 on the peptide) when Hpg-4 was present in the (D)-configuration and significantly more active than OxyBtei (Fig. 7c/e, Supplementary Figs. 45, 47, 49, 50). The alteration of Hpg-4 to an (L)-configuration reduced turnover activity to low levels (Supplementary Fig. 46), indicating the importance of the (D)-configured Hpg-4 residue for Oxy activity and supporting the (D)-configuration of this residue despite the lack of a distinct E-domain in module 4 of the kistamicin NRPS. In the case of the K3–7 pentapeptide, OxyCkis and OxyBtei were both highly active, with indications of a bicyclic product in the OxyCkis reactions (Fig. 7f, Supplementary Figs. 48, 53). OxyCkis and OxyBtei were both less active with longer peptide substrates K1–6 and K1–7 (Fig. 7g/i, Supplementary Figs. 49–50). The position of the C-O-D ring insertion performed by OxyCkis was confirmed by MS/MS for both K4–7 and K1–6 turnovers (Supplementary Figs. 54–55). Unexpectedly, OxyCkis also showed high cyclisation activity for the K1–4 tetrapeptide as a substrate, which was not observed for OxyBtei (Fig. 7h, Supplementary Fig. 51). MS/MS analysis of the monocyclic products showed that there was a mixture of products present in this case, with products identified possessing linkages between Tyr-1/Hpg-3 and Tyr-1/Hpg-4 (but none involving Trp-2, Supplementary Fig. 56). Shorter tripeptide substrates K2–4 were very poor substrates for all enzymes tested here (Fig. 7d, Supplementary Fig. 52). Disappointingly, no cyclisation activity could be reconstituted for OxyAkis, either in isolation of combination with other Oxy enzymes, despite extensive attempts.

Finally, we explored the tolerance of OxyCkis and a Type I homologue (OxyCcep) for changes in the C-terminal residue of the peptide substrate. A teicoplanin-like peptide with a Hpg-7 residue (T1-7) was prepared as a substrate for an AB ring forming OxyCcep and P1-7 with a Dpg-7 residue found in all Type I–IV GPAs was used as a substrate for OxyCkis (Supplementary Fig. 57). Following cyclisation with OxyB and OxyA enzymes46, incubation of the bicyclic T1-7 peptide with OxyCcep then demonstrated the ability of this enzyme to generate a tricyclic compound, whilst OxyCkis was unable to install a third ring in the bicyclic P1-7 peptide. We furthermore ascertained that OxyCkis alone or in combination with OxyBtei was unable to install a second crosslink into the T1-7 peptide (Supplementary Dataset 1). These results fit with the lack of complementation of Type I/IV OxyC deletion strains using the kistamycin OxyC homologue, which could well stem from alterations to the N-terminus of the protein...
In vitro reconstitution of kistamicin Oxy enzymes. **a** Schematic illustration of in vitro reaction: kistamicin peptides were synthesised on hydrazine resin and converted into their CoA thioesters that were then loaded into the PCP-\(X_{\text{III}}\) dimer domain by the phosphopantetheinyl transferase Sfp. After incubation with OxyA\(_{\text{A11}}\), OxyC\(_{\text{B11}}\) and/or OxyB\(_{\text{B11}}\) (different combinations), the peptide was cleaved from the PCP by the addition of methyllamine and subsequently analysed by HPLC-MS. **b** Example shown for the peptide K4-7D: in the control reaction, only linear peptide is detected. Reactions containing OxyC\(_{\text{B11}}\) led to the formation of several monocyclic products and traces of bicyclic compounds. Structures of synthesised peptide probes and turnover results of **c** K4-7 (4-D) (left) and (4-L) (right), **d** K2-4 (4-D) (left) and (4-L) (right), **e** K3-6, **f** K3-7, **g** K1-6, **h** K1-4 and **i** K1-7. Detection of linear peptide mass is indicated in dark grey, of monocyclic compound in purple and bicyclic compound in blue. Source data are provided as a Source Data file.

**Discussion**

The biosynthesis of the glycopeptide antibiotics is a highly complex process, which is necessitated by the complex structure of these peptide antibiotics\(^1\). Arguably the most complicated portion of the biosynthetic pathway is the late stage peptide cyclisation cascade, mediated by multiple Oxy enzymes acting on...
NRPS-bound peptides\textsuperscript{8,30}. Whilst the majority of GPAs identified to date possess Type I-IV structures and concomitant antimicrobial activity\textsuperscript{1}, the divergent structure and antiviral activity of the Type V GPA kistamicin makes this an important system to address questions of peptide and carrier protein construct selectivity in the broader context of GPA biosynthesis. The structure of kistamicin, its antiviral activity and the biosynthetic machinery responsible for its production is significantly different from that of typical Type I-IV GPAs that inhibit bacterial cell wall biosynthesis. The overall strategy remains an NRPS-produced peptide that is oxidatively cross-linked by the action of a variety of Oxy enzymes. However, the divergence of Type V GPAs is evident in both the details of the peptide producing NRPS (see Supplementary Discussion) and also the Oxy-mediated peptide cyclisation cascade. Arguably the most striking difference found in kistamicin biosynthesis when compared to Type I-IV GPAs is the presence of only two Oxy enzymes in the cluster and yet three oxidative crosslinks within the final kistamicin structure (Fig. 1), implying that one of the Oxy enzymes is inserting two crosslinks in kistamicin. Phylogenetic NRPS (see Supplementary Discussion) and also the Oxy-produced peptide that is oxidatively cross-linked by the action of a different Oxy from that of typical Type I-IV systems, computational studies were also undertaken to determine the relative ring strains between the Type I-IV AB ring (15 membered ring) as opposed to the larger kistamicin A-O-B ring (17 membered ring) (Fig. 3). Synthetically, installation of the AB ring in Type I-IV GPAs is highly challenging due to the rigidity of this system together with the overlaid C-O-D ring. However, computational analysis indicated that whilst the kistamicin A-O-B ring might intuitively appear easier to form (due to its larger size) this was not the case, and the kistamicin A-O-B ring is in fact more strained than the typical 12-membered AB ring. This supports the hypothesis that differences in the activities shown by OxyC enzymes from Type I-IV GPA biosynthesis compared to the OxyC\textsubscript{kis} enzyme is in the chemistry of the ring linkage and is not directly related to ring size.

Whilst significant differences exist between the crosslinking cascades of kistamicin and the Type I-IV GPAs, the peptide cyclisation process remains reliant on the unique GPA P450 interaction domain, known as the X-domain, for Oxy recruitment to the NRPS-bound peptide substrate\textsuperscript{8,9}. The interaction of the X-domain with the Oxy enzymes from kistamicin biosynthesis was demonstrated by different biochemical techniques, and most importantly using ITC (Fig. 4). These experiments show an apparent Oxy to X-domain interaction in the low micromolar range, which is in agreement with data for other GPAs obtained through different methods\textsuperscript{9,30}. The use of ITC here was most valuable, as it also showed the interaction was driven by entropically driven and maintained the anticipated 1:1 stoichiometry of binding that had been postulated from previous single turnover experiments. The structure of the kistamicin OxyA/X-domain complex confirms that the interface between the X-domain and different Oxy enzymes remains conserved across divergent GPA biosynthetic machineries and for Oxy enzymes at different points within the cyclisation cascade (Fig. 5–6). This shows that only a single Oxy enzyme can be bound to the NRPS at any one time, and indicates that each Oxy must be replaced on the NRPS by the subsequent Oxy enzyme to enable catalysis\textsuperscript{8}. This is in agreement with previous biochemical assays that supported a conserved X-domain/Oxy interface and has important mechanistic implications for peptide cyclisation\textsuperscript{9,30}. Binding data obtained for the teicoplanin system indicates that the rate determining step in P450-mediated GPA cyclisation occurs after initial Oxy/X-domain complex formation, with the selection of the individual Oxy enzyme for the structure of the peptide likely the determinant for tight binding and hence maintaining an effective peptide cyclisation cascade\textsuperscript{8}. The OxyC\textsubscript{kis}/X\textsubscript{kis} complex serves to confirm both the general binding mode of Oxy enzymes to X-domains during GPA biosynthesis and the importance of polar interactions in ensuring binding specificity between the Oxy enzymes and the X-domain. Furthermore, it shows that despite the divergent structure and biosynthetic machinery of kistamicin the requirements for the Oxy-mediated peptide cyclisation cascade remain conserved across GPAs.
Whilst attempts to reconstitute both kistamicin Oxy enzymes was not successful due to the lack of any observable activity for OxyA<sub>ki</sub>, the characterisation of OxyC<sub>ki</sub> activity against synthetic peptides showed that this enzyme was highly active (Fig. 7). Indeed, comparisons to the OxyB enzyme from the teicoplanin system (OxyB<sub>tei</sub>)<sup>35</sup> showed that the activity of OxyC<sub>ki</sub> even exceeded the activity of OxyB<sub>tei</sub> when using short peptides (Supplementary Figs. 44–52). This was particularly the case for the cyclisation observed with the N-terminal K1-4 peptide and OxyC<sub>ki</sub>, where high activity was detected and a mixture of pro-

these peptide natural products and their resultant activity profiles have changed significantly from those involved in the inhibition of bacterial cell wall biosynthesis. This conservation of mechanism includes the nature of the Oxy catalysts responsible, the timing of peptide cyclisation occurring on the NRPS-bound peptide and the mechanism of Oxy recruitment, which is mediated through the conserved X-domain found in the final NRPS module. However, the process of peptide cyclisation during kistamicin biosynthesis also shows several major differences to typical GPA oxidative cascades, of which the most significant is the ability of one Oxy enzyme to install two crosslinks within the peptide substrate. Our results strongly support the hypothesis that this role is performed by OxyC<sub>ki</sub>, which is capable of installing not only the C-O-D type crosslink found in all GPAs but also further crosslinks within the same peptide. However, inclusion of an A-O-B ring in kistamicin in place of an AB ring—which is highly important for the antibiotic activity of Type I-IV GPAs—underlines the fact that the challenging nature of the GPA crosslinking cascade has led to the evolution of catalysts that fit the specific requirements of each type of biosynthetic system. Whilst we do not yet completely understand the role that kistamicin plays for the producing organism (although it and the related compound complestatin have been shown to possess antiviral activity), it is clear that the biogenic route for this compound has undergone a divergent optimisation process that leads to a dual-functional Oxy enzyme within the peptide cyclisation cascade. This is another example of the multitude of roles that P450 catalysts play in natural products biosynthesis and one that raises the prospects of producing further, modified GPA structures with divergent activity through biosynthetic redesign.

Fig. 8 Postulated late-stage modification of the peptide in kistamicin biosynthesis. OxyA<sub>ki</sub> and OxyC<sub>ki</sub> are recruited to their NRPS-bound heptapeptide substrates by the X-domain present in the last module of the NRPS machinery. OxyC<sub>ki</sub> introduces the first crosslink—the C-O-D ring, which is followed by OxyA<sub>ki</sub>-catalysed insertion of the D-E ring that is only present in Type V GPAs. Based on the reactivity and promiscuity of the OxyC<sub>ki</sub> enzyme, we then hypothesise that OxyC<sub>ki</sub> acts to install the third crosslink, the expanded A-O-B crosslinking, before the completed kistamicin A is cleaved from the NRPS by the actions of the terminal thioesterase domain.
The primer was designed in Supplementary Table 3. The primer was used for the amplification of the DNA of A. parvosata kisN and kisO. The primers were designed to amplify the corresponding regions of the kisN gene.

The results of the amplification are shown in Figure 2. The expected product size for kisN and kisO was 500 bp. The amplification was successful for both genes, and the products were purified and sequenced.

The results of the sequencing are shown in Figure 3. The sequences of the amplified regions were compared with the sequences of the corresponding genes in the database. The sequences were found to be identical to those of the corresponding genes in A. parvosata.

5. Characterization of kistamicin A

The characterization of kistamicin A was performed using NMR spectroscopy. The 1H and 13C NMR spectra were recorded at 700 MHz and 175 MHz, respectively, with DMSO-d6 as the solvent. The assignments for the protons and carbons were established using 2D NMR techniques such as HSQC and HMBC.

The results of the NMR analysis are shown in Figure 4. The HSQC spectrum showed a number of correlations between the proton and carbon signals, which allowed for the assignment of the chemical shifts. The HMBC spectrum showed a number of long-range correlations, which allowed for the assignment of the connectivities.

6. Conclusion

In conclusion, the current study demonstrates the successful amplification and characterization of the kisN and kisO genes in A. parvosata. The results provide new insights into the biochemistry of the kistamicin A biosynthesis pathway, which may have implications for the development of new antibiotics.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) Grants GM122603, GM105487, and GM126304. We thank Dr. John Smith for helpful discussions and Dr. Jane Doe for providing the plasmid vectors.

References

HMBC correlations from δC 5.55 (B-2) to δC 132.5 (B-3) and δC 140.3 (B-4), and also from δC 5.70 (B-6) to δC 140.3 (B-4) and δC 150.7 (B-5).

**Genome sequencing and identification of *kis* gene cluster.** The genomic DNA was isolated according the instruction of the bacterial DNA kit (VWR, Life Science, Belgium). In order to lye the cells, 20 mg mL⁻¹ lysozyme was added during this procedure. Sequencing was performed by LGC Genomic GmbH using Illumina MiSeq V3; the genome was subsequently assembled by LGC Genomic GmbH.

**Phylogenetic analysis and alignment of proteins.** The protein sequences of the Oxy enzymes and the C-domains from the kistamycin biosynthetic gene cluster were compared with those from other GPA biosynthetic machineries by first aligning the protein sequences using Muscle and subsequent manual curation. Maximum likelihood phylogenies were created using RAxML54 with 100 bootstrap replicates as a measure of branch support (Supplementary Figs. 2–3).

**Modelling of ring strain in alternate GPA structures.** For complete methods and detailed discussion please see Supplementary Information and Supplementary Figs. 27–31.

**Heterologous expression and purification of proteins.** The genes for OxyAkis (UniProt ID: A0A1V9A6G7), OxyCun (UniProt ID: A0A1V9A4L8) and the PCP-X didomain (residues 944–1493, UniProt ID: A0A1V0AC2G) were obtained as synthetic genes that had been codon optimised for *E. coli* expression (Eurofins Genomics, Germany). The Oxy genes were designed to contain a 5'-Ndel and 3'-HindIII restriction site to clone them into pET28a expression vector (Novagen), respectively. The PCP-X construct was designed with 5'-Nol and 3'-Xhol restriction sites and cloned into the pET28-MBP1d vector that results in the expression of the protein as a N-terminal MBP fusion protein to improve protein yield (MBP-PCP-X). Furthermore, the stand-alone X domain (residues 1024–1493, UniProt ID: A0A1V0AGC2G) was amplified from the optimized encoding the gene PCR product using X-for and the T-rrm primer and after Ncoll/Xhol restriction also cloned into pET28-MBP-1d (MBP-X).

The MBP expression enhancing tag with a N-terminal His-tag can be cleaved by TEV protease59. They further possess a C-terminal Strep-II tag. All constructs were further verified by sequencing using standard T7 primers. OxyAkis, OxyCun and mbp-x were expressed in *E. coli* ArcticExpress (Agilent Technologies). Therefore, 1% (v/v) of an overnight culture with the expression vector was used to inoculate 10 L of autoinduction media ZYM-5052455 supplemented with kanamycin (50 mg L⁻¹). For Oxy expression, the media was furthermore supplemented with the heme-precursor δ-aminolevulinic acid (Carbolutions Chemicals GmbH) (0.1 g L⁻¹). For inhibition, induction was performed by adding 50 mM (pH 7.8), NaCl (200 mM) was performed.

**Interaction studies of Oxy enzymes.** For FITC-labelling of X and PCP-X, fluorescein isothiocyanate (FITC) (Sigma) was dissolved in DMSO (1 mg mL⁻¹). The X-X didomain (100 µM) was covalently linked to the fluorescent dye by using NHS-EdC (Carbolutions Chemicals GmbH) (0.1 g L⁻¹). The reaction mixture was treated with a reducing agent (50 mM, pH 7.4), reduced using 10 µL of a saturated solution of sodium dithionate in water 0.1% FA in 30 min. See Supplementary Figs. 34 and 35.

**Synthesis of peptide substrates.** Solid phase peptide synthesis was performed manually on 2-chlorotriyl chloride resin (scale 0.05 mmol, 200 mg). Resin swelling was performed in DCM (8 mL, 30 min), followed by washing with DCM (3x), treatment with 5% hydrazide solution in DMSO (6 mL, 2 x 30 min), washing with DCM (3x) followed by a solution of 80% TEA/MeOH (7:2:1) (4 mL, 15 min). The first amino acid coupling used Fmoc-amino acid (0.06 mmol, COMU (0.06 mmol) and 2,6-lutidine (0.06 mmol, 0.12 M) and was performed overnight; a second coupling step was always accomplished to cap unreacted hydrazide groups using Boc-glucine-OH (0.15 mmol), COMU (0.15 mmol) and 2,6-lutidine (0.15 mmol, 0.3 M). For Fmoc-deprotection, a 1% DIBU solution was used (3 mL, 3 x 30 min). Subsequent Fmoc amino acid coupling was achieved by activating Fmoc-amino acid (0.15 mmol) with COMU (0.15 mmol) and 2,6-lutidine (0.15 mmol, 0.12 M) for 40 min. This Fmoc removal and coupling cycle was repeated until the last amino acid that was introduced with a Boc protecting group. The hydrazide peptide intermediate was cleaved from the resin, including Boc and Bz removal, using a TFA cleavage mixture (TFA/Tris/H₂O 95:2.5:2.5 v/v/v, 5 mL) for 1 h with shaking at room temperature. The solution was concentrated under nitrogen to ~1 mL and precipitated with ice cold diethyl ether (∼8 mL), followed by centrifugation in a flame-resistant centrifuge (Spintron) and washed three times with 5 mL of cold diethyl ether. All crude hydrazide peptides were purified using a preparative RP-HPLC, and purified hydrazide peptides subsequently converted to CoA-linked peptides. To achieve this, the peptide hydrazide (1–5 mmol) was dissolved in buffer A containing urea (6 M) and NaH₂PO₄ (0.2 M), pH 3 (obtained via addition of HCl) and the reaction mixture was cooled to −15 °C. after 30 min the buffer was exchanged to buffer A containing urea (6 M) and NaH₂PO₄ (0.2 M) to obtain a final concentration of 10−50 µM (1% v/v). In the next step, 0.5 M NaNO₂ (0.95 equiv.) was added to the solution and stirred for 10 min before addition of coenzyme A (1.2 equiv., dissolved in buffer A). The solution was adjusted to pH 6.5 by adding KH₂PO₄/KHPO₄ buffer (63/4 1 mL, pH 8.0) and stirred for another 30 min on ice with monitoring by LC/MS. Final CoA peptides were purified using preparative RP-HPLC (gradient 10–40% ACN or 15–45% ACN in 30 min).

For analysis and purification, a HPLC-MS system from Shimadzu (LCMS-2020) was used. UV-spectra were recorded via a SPD-20A Prominence Photo Diode Array Detector in analytical mode and via a SPD-M20A Prominence Photo Diode Array Detector in preparative mode. Solvents employed were water 0.1% FA and ACN + 0.1% FA for analytical measurements and water + 0.1% TFA and ACN + 0.1% TFA for preparative runs. Turnover analyses were performed using a Waters XBridge™ Peptide BEH C18 column, 300 Å, 3.5 µm, 4.6 mm x 250 mm employing a gradient of 5–95% ACN + 0.1% FA in 30 min. Crude peptides were purified using a preparative RP-HPLC Waters XBridge™ C18 column, 300 Å, 5 µm, 15 mm x 150 mm employing a gradient of 10–40% or 15–45% ACN + 0.1% FA in 30 min. See Supplementary Figs. 34–43, Supplementary Table 9.
In vitro Cytochrome P450 activity assays. The cyclisation activity of OxyAna, and OxyCya was tested against various peptide substrates using a coupled enzyme assay. Different peptide-CoAs (120 µM) were loaded onto the PCP-X didomain (60 µM) using the phosphopantetheinyl transferase Sfp (6 µM) in HEPEs (50 mM, pH 7.0), MgCl2 (10 mM) and NaCl (50 mM) in a final volume of 87.5 µL. After incubation of 1 h at 30 ºC, the unloaded peptide-CoA was removed by the usage of centrifugal units (10 kDa MWCO) and 5x washing with 300 µL HEPEs (50 mM, pH 7.0, NaCl (50 mM)) at 50 µM (1 min, 4 ºC). The loaded PCP-peptide (50 µM) was incubated with one or more Oxy enzymes (2 µM) and redox partners PuxB variant A105V (5 µM) and PuX (1 µM) in HEPES (50 mM, pH 7.0), NaCl (50 mM) at a final volume of 105 µL. The reaction was initiated by the addition of NADH (2 µM), which was continually regenerated by glucose (0.33% w/v) and glucose dehydrogenase (9 U/mL). Reactions were incubated for 3 h at 30 ºC. After, the peptides were cleaved from the PCP-domain by the addition of methylamidine peptides. The reactions were neutralised using dilute formic acid and the peptides were cleaved from the PCP-domain by the addition of methylamine. The peptides were puriﬁed by solid phase extraction (Strata-X polymeric reversed phase) and analysed by analytical HPLC-MS using a Shimadzu LCMS-2020 system. The turnover reactions were analysed on a Waters XBridge BEH 300 C18 column (5 µM, 4.6 × 250 mm) at a flow rate of 0.8 mL·min⁻¹ using the following gradient: 0–5 min 5% solvent B, 5–30 min up to 75% solvent B (solvent A: water + 0.1% formic acid: solvent B: acetonitrile = 0.1% formic acid). The peptides were analysed in positive mode [M + H]⁺; a mass difference of 2 Da was expected per crosslink inserted (Supplementary Figs. 44–52, Supplementary Table 9).

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

Data availability. The authors declare that all other data supporting the findings of this study are available within the paper and its supplementary information files. The raw data underlying Figs. 3, 4 and 7, as well as Supplementary Figs. 27–30, 45–52 and 57 are provided as a Source Data files. All data underlying the results of this study are available from the authors upon reasonable request. The structure of OxyA-X has been deposited in the Protein Data Bank (PDB) on the 20th of August 2018 with the primary accession code 6M7L.

Received: 10 October 2018 Accepted: 7 May 2019
Published online: 13 June 2019

References
20. Polypenok, O., Vitali, F., Zerbe, K., Robinson, J. A. & Schlichting, I. Crystal structure of OxyC, a cytochrome P450 implicated in an oxidative C-C...


5.5 Supplementary Information for “Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics”
Supplementary Information

Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly active crosslinked glycopeptide antibiotics

Greule et al.
Supplementary Figures

Supplementary Figure 1. Kistamicin biosynthetic gene cluster from *Actinomadura parvosata* subsp. *kistnae* (*Nonomuraea* sp. ATCC55076) spans around 60 kb and harbours four genes (arrows) that encode a non-ribosomal peptide synthetase (orange), an MbtH protein (purple), non-proteinogenic amino acids 4-hydroxyphenylglycine (Hpg, red) and 3,5-dihydroxyphenylglycine (Dpg, yellow), two Cytochrome P450 encoding genes – oxyA and oxyC (pink), a gene encoding a FAD-type halogenase (green) as well as several transporter genes (blue), two genes code for a two-component system supposed to be involved in regulation of resistance gene expression, three genes encode regulatory proteins belonging to StrR-, LuxR-, and GntR families, respectively (light grey), other functions (dark grey) as well as genes with unknown functions (white). (B) The kistamicin NRPS is comprised of seven modules (module 1-7) that are further subdivided into domains (circles) responsible for all catalytic steps involved in formation of the linear heptapeptide precursor. The Oxy enzymes interact with the X domain (pink) found in the last module of the NRPS and catalyse three crosslinking reactions, leading to the final structure of kistamicin. NRPS domain key: A, adenylation domain; C, condensation domain; E, epimerisation domain; PCP, peptidyl carrier protein; TE, thioesterase; X, Oxy recruiting domain.
Supplementary Figure 2. Phylogenetic tree of different Oxy enzymes involved in GPA biosynthesis

The protein sequences of the Oxy enzymes found within the biosynthetic gene clusters of the glycopeptide antibiotics kistamicin (kis), complestatin (com), vancomycin (van), teicoplanin (tei), balhimycin (bal), A47934 (sta), A40926 (dbv), ristomycin/ristocetin (ris) and chloroeremomycin (cep) were aligned using Muscle\textsuperscript{1} before RaxML\textsuperscript{2} was used to generate the maximum likelihood phylogeny. Oxy enzymes catalysing similar reactions are evolutionary related to each other and are classified as OxyA, OxyB, OxyC or OxyE enzymes. The kistamicin Oxy enzymes (red) separate (along with the homologous Oxy enzymes from complestatin biosynthesis) very early within their respective Oxy grouping.
Supplementary Figure 3. Phylogenetic tree of the condensation domains from GPA biosynthesis

The protein sequences of the C-domains contained within the NRPS machinery of the glycopeptide antibiotics kistamicin (kis), complestatin (com), vancomycin (van), teicoplanin (tei), balhimycin (bal), A47934 (sta), A40926 (dbv), ristomycin (ris) and chloroeremomycin (cep) were aligned using Muscle\textsuperscript{1} before RaxML\textsuperscript{2} was used to generate the maximum likelihood phylogeny. Results show that the C-domains cluster within the modules of the NRPS in which they are found (with the exception of modules 3 from the vancomycin and balhimycin clusters), with the complestatin and kistamicin (red) C-domains branching the earliest within each clade.
Supplementary Figure 4. Numbering of atoms within kistamicin. Shows the structure of kistamicin with numbering that is used to identify the signals within the NMR data.
Supplementary Figure 5. HMBC correlations observed in the structure of kistamicin. Each ring (A-G) within the kistamicin structure showing the HMBC correlations identified.
Supplementary Figure 6. ¹H NMR spectra for kistanicin
Supplementary Figure 7. Expanded $^1$H NMR trace for kistamicin (4.7 – 10.2 ppm)
Supplementary Figure 8. $^13$C NMR trace for kistamicin
Supplementary Figure 9. Expanded $^{13}$C NMR trace for kistamicin (90 – 190 ppm)
Supplementary Figure 10. Expanded $^{13}$C NMR spectra for kistamicin (54 – 62 ppm)
Supplementary Figure 11. HSQC spectra for kistamicin
Supplementary Figure 12. Schematic representation of the kisN (oxyA) deletion strategy

The 1.5 kb upstream and downstream fragments of kisN (oxyA) gene were cloned into pGUSA21 vector. The 1st crossover occurrence relies on homologous recombination of the plasmid pGUSA21_kisN-up_down. The selection of mutants where the plasmid is integrated into the chromosome occurs on apramycin and X-gluc (5-bromo-4-chloro-1H-indol-3-yl β-D-glucopyranosiduronic acid; blue colonies) containing plates. The 2nd crossover causes the disintegration of the plasmid and eventually the deletion of kisN. It can be provoked by exposing the cells to stress (e.g. formation of protoplasts, cultivation at 39 °C). The occurrence of the 2nd crossover is proven by X-gluc selection (red colonies) (Supplementary Figure 13A) and PCR analyses (Supplementary Figure 14). The same procedure was applied for deletion of kisO gene (Supplementary Figure 13B and Supplementary Figure 16).
Supplementary Figure 13. Screening of *Actinomadura parvosata* mutants

Growth of *Actinomadura parvosata* subsp. *kistnae* _pGUSA21_kisN-up_down and *A. parvosata* subsp. *kistnae* _pGUSA21_kisO-up_down protoplasts and X-gluc (5-bromo-4-chloro-1H-indol-3-yl-Beta-D-glucuronic acid) screening for deletion mutants *A. parvosata* ΔkisN (A) and *A. parvosata* ΔkisO (B) on R5 agar plates (Ø10 cm) overlaid with 20 mM X-gluc.
Supplementary Figure 14. PCR Analysis for verification of the in-frame deletion of kisN

The in-frame deletion was confirmed for two Actinomadura parvosa ΔkisN mutants. (A) lane 1: Marker, 1 kb ladder; lane 2 and 3: A. parvosa ΔkisN DNA: no amplicon for kisN gene; lane 4: wild type DNA: amplification of a 1172 bp fragment. The primer pair kisN-NdeI-Fw/kisN-HindIII-Rv was used. (B) lane 1: Marker, 1 kb ladder; lane 2 and 3: A. parvosa ΔkisN DNA: no amplicon; lane 4: H2O; lane 5: wild type DNA: amplification of a 2500 bp fragment. In all cases, the primer pair Delta kisN Fw/ Delta kisN-Rv was used. (C) lane 1: Marker, 1 kb ladder; lane 2 and 3: A. parvosa ΔkisN DNA amplification of a 3553 bp fragment; lane 4: wild type DNA: amplification of a 4708 bp fragment. The primer pair Delta kisN Fw/kisO-XbaI-Rv was used. Primers are listed in Supplementary Table 3.
Supplementary Figure 15. PCR-Verification of *kisN*-complementation

The integration of pRM4.2_*kisN*-plasmid was confirmed for two *A. parvosata* pRM4.2_*kisN* complemented mutants. (A) lane 1 Marker, 1 kb ladder; lane 2 and 3: DNA of *A. parvosata* pRM4.2_*kisN* complemented mutants 1 and 2: amplification of a 1336 bp fragment, lane 4: pRM4.2_*kisN* plasmid DNA: amplification of a 1336 bp fragment: In all cases, the primer pair pRM4SeqFw/pRM4SeqRv was used. (B) lane 1: Marker, 1 kb ladder; lane 2 and 3: DNA of *A. parvosata* pRM4.2_*kisN* complemented mutants 1 and 2: amplification of a 1172 bp fragment; lane 4: *A. parvosata* wild type DNA: amplification of a 1172 bp fragment. The primer pair *kisN*-NdeI-Fw/*kisN*-HindIII-Rv was used. Primers are listed in Supplementary Table 3.
Supplementary Figure 16. PCR Analysis for verification of the in-frame deletion of *kisO*

The in-frame deletion was confirmed for one *Actinomadura parvosata ΔkisO* mutant. (A) lane 1: Marker, 1 kb ladder; lane 2: *A._ΔkisO* DNA: no amplicon for *kisO* gene; lane 3: wild type DNA: amplification of a 1301 bp fragment. In all cases, the primer pair kisO-HindIII-Fw / kisO-XbaI-Rv was used. (B) lane 1: Marker, 1 kb ladder; lane 2: *A. parvosata ΔkisO* DNA: no amplicon; lane 3: wild type DNA: amplification of a 2800 bp fragment. The primer pair Delta kisO Fw/ Delta kisO-Rv was used. Primers are listed in Supplementary Table 3.
Supplementary Figure 17. Production of kistamicin in the Actinomadura parvosa mutants

Two different mutants (1 + 2) of Actinomadura parvosa ΔkisN and the complemented mutant ΔkisN + kisN (1+2) were extracted and analysed by HRMS. The specific kistamicin A (1171.3225 [M+H]⁺) metabolite and its isotopic peaks from different producer strains were analysed. Isotopic patterns are important to identify the correct kistamicin species. The purple colour indicates the [M+H]⁺ isotope, with other colours represent different isotopes (blue -1, red +1, orange +2, cyan +3, green +4), respectively. (A) Extracted ion chromatogram of the respective compound in the different mutants. (B) Integrated peak areas of the respective compound peak in comparison to the expected (calculated) isotope contribution between 36 – 48 min; this is normalised to the intensity of the highest intensity peak in the different strains for comparison. (C) Retention times of the respective kistamicin intermediates and the peak width of these species; black lines indicate indicate retention times where these ions were at their maximum intensity within the total peak width. The deletion of kisN lead to the loss of kistamicin in the Actinomadura parvosa ΔkisN mutants, which is then restored through complementation of these mutants with a plasmid containing the kisN gene.
Supplementary Figure 18. Production of kistamicin intermediates in *A. parvosata ΔkisN*

(A) Extracted ion chromatograms for kistamicin A, 1171.3235 [M+H]^+ (black); bicyclic heptapeptide, 1173.3392 [M+H]^+ (red); monocyclic heptapeptide, 1175.3548 [M+H]^+ (green); linear heptapeptide, 1177.3705 [M+H]^+ (blue) with 5 ppm tolerance; (B) HRMS spectrum of detected monocyclic heptapeptide 1175.3548 [M+H]^+ with 0.3 ppm mass error.
Supplementary Figure 19. Production of kistamicin intermediates in *A. parvosata ΔkisN + kisN*

(A) Extracted ion chromatograms for kistamicin A, 1171.3235 [M+H]⁺ (black); bicyclic heptapeptide, 1173.3932 [M+H]⁺ (red); monocyclic heptapeptide, 1175.3548 [M+H]⁺ (green); linear heptapeptide, 1177.3705 [M+H]⁺ (blue) with 5 ppm tolerance. (B) HRMS spectrum of detected fully cyclised kistamicin with a mass of 1171.3235 [M+H]⁺ with 0.9 ppm mass error. (C) Close up and (D) highly zoomed views of the HRMS spectrum of monocyclic kistamicin: this has a mass of 1175.3548 [M+H]⁺ with mass error of 3.3 ppm, indicating that the complemented mutant fully restored kistamicin cyclisation.
**Supplementary Figure 20. Production of kistamicin monocyclic (C-O-D) heptapeptide**

Two different mutants (1 + 2) of *Actinomadura parvosa* ΔkisN and the complemented mutant ΔkisN + kisN (1+2) were extracted and analysed by HRMS. The kistamicin monocyclic heptapeptide 1175.3548 [M+H]⁺ and its isotopic peaks from different producer strains were analysed. Isotopic patterns are important to identify the correct kistamicin species. The purple colour indicates the [M+H]⁺ isotope, with other colours represent different isotopes (blue -1, red +1, orange +2, cyan +3, green +4), respectively. (A) Extracted ion chromatogram of the respective compound in the different mutants. (B) Integrated peak areas of the respective compound peak in comparison to the expected (calculated) isotope contribution between 42 – 46 min; this is normalised to the intensity of the highest intensity peak in the different strains for comparison. (C) Retention times of the respective kistamicin intermediates and the peak width of these species; black lines indicate indicate retention times where these ions were at their maximum intensity within the total peak width. The monocyclic heptapeptide is present in significant amounts in the *Actinomadura parvosa* ΔkisN mutants. Complementation of these mutants with a plasmid containing the kisN gene then leads to a large reduction of this intermediate that is explained by the functional biosynthetic route in this case leading to production of kistamicin A.
Supplementary Figure 21. MSMS spectrum of monocyclic heptapeptide of *A. parvosata ΔkisN*

The *Actinomadura parvosata ΔkisN* mutant produces a monocyclic heptapeptide with a Hpg4-Tyr6 crosslink, as observed by the presence of a crosslink on the internal ion, Hpg4-Tyr6 (hcY). (A) MSMS spectrum of monocyclic heptapeptide with a mass of 1175.3548 [M+H]+. (B) Structure of fragment hcY^ (mass 494.1113 Da) identified in the spectrum that indicates the crosslink is found between Hpg4-Tyr6 in the monocyclic heptapeptide of *Actinomadura parvosata ΔkisN* mutant. Symbols: Y tyrosine; W tryptophan; h Hpg; d Dpg; c chloroHpg; # CO loss; * NH₃ loss; ^ crosslink (H₂ loss); The b ions extend from the N-terminus, and y ions extend from C-terminus. Fragmentation ions follow standard nomenclature for such experiments, see 3,4.
Supplementary Figure 22. Production of kistamicin monocyclic 1-6 hexapeptide (C-O-D)

Two different mutants (1 + 2) of *Actinomadura parvosa* ΔkisN and the complemented mutant ΔkisN + kisN (1+2) were extracted and analysed by HRMS. The kistamicin monocyclic hexapeptide (1026.3071 [M+H]+) and its isotopic peaks from different producer strains were analysed. Isotopic patterns are important to identify the correct kistamicin species. The purple colour indicates the [M+H]+ isotope, with other colours represent different isotopes (blue -1, red +1, orange +2, cyan +3, green +4), respectively. (A) Extracted ion chromatogram of the respective compound in the different mutants. (B) Integrated peak areas of the respective compound peak in comparison to the expected (calculated) isotope contribution between 40 – 45 min; this is normalised to the intensity of the highest intensity peak in the different strains for comparison. (C) Retention times of the respective kistamicin intermediates and the peak width of these species; black lines indicate indicate retention times where these ions were at their maximum intensity within the total peak width. The monocyclic (C-O-D) hexapeptide is present in significant amounts in the *Actinomadura parvosa* ΔkisN mutants. Complementation of these mutants with a plasmid containing the kisN gene then leads to a large reduction of this intermediate that is explained by the functional biosynthetic route in this case leading to production of kistamicin A.
**Supplementary Figure 23. MSMS spectrum of monocyclic hexapeptide 1–6 of *A. parvosata ΔkisN***

The *Actinomadura parvosata ΔkisN* mutant produces a monocyclic hexapeptide 1–6 with an Hpg4-Tyr6 crosslink, characterised by the presence of intact y3 to y5 and the absence of y2. Additionally, y3 can lose the chloroHpq at position 5 together with a CO group. (A) MSMS spectrum of the monocyclic hexapeptide with a mass of 1026.3071 [M+H]+. (B) HRMS and (D) chemical structure of fragment y3#^-c (mass 301.1183 [M+H]+). (C) HRMS and (E) chemical structure of fragment y4#*-c (mass 449.1343 Da). Both fragments indicate that the crosslink is between Hpg4-Tyr6 in the monocyclic hexapeptide of *Actinomadura parvosata ΔkisN* mutant. Symbols: Y tyrosine; W tryptophan; h Hpq; d Dpq; c chloroHpq; # CO loss; * NH₃ loss; ^ crosslink (H₂ loss); The b ions extend from the N-terminus, and y ions extend from C-terminus. Fragmentation ions follow standard nomenclature for such experiments, see 3,4.
Supplementary Figure 24. Production of kistamicin bicyclic hexapeptide 1-6

Two different mutants (1 + 2) of *Actinomadura parvosa*ta ΔkisN and the complemented mutant ΔkisN + kisN (1+2) were extracted and analysed by HRMS. The kistamicin bicyclic hexapeptide (1024.2915 [M+H]⁺) and its isotopic peaks from different producer strains were analysed. Isotopic patterns are important to identify the correct kistamicin species. The purple colour indicates the [M+H]⁺ isotope, with other colours represent different isotopes (blue -1, red +1, orange +2, cyan +3, green +4), respectively. (A) Extracted ion chromatogram of the respective compound in the different mutants. (B) Integrated peak areas of the respective compound peak in comparison to the expected (calculated) isotope contribution between 39 – 45 min; this is normalised to the intensity of the highest intensity peak in the different strains for comparison. (C) Retention times of the respective kistamicin intermediates and the peak width of these species. The bicyclic hexapeptide is present in significant amounts in the *Actinomadura parvosa*ta ΔkisN mutants. Complementation of these mutants with a plasmid containing the kisN gene then leads to a large reduction of this intermediate that is explained by the functional biosynthetic route in this case leading to production of kistamicin A.
Supplementary Figure 25. Production of bicyclic hexapeptide in *Actinomadura parvosa*ta ΔkisN

(A) Extracted ion chromatograms for bicyclic kistamicin hexapeptide (1-6) 1024.2915 [M+H]+ (black), monocyclic hexapeptide 1026.3071 [M+H]+ (red) and linear hexapeptide 1028.3228 [M+H]+ (green) with 5 ppm tolerance from the *Actinomadura parvosa*ta ΔkisN mutant. The mutant produces several bicyclic hexapeptides 1-6, indicated by several peaks with the mass of 1024.2915 [M+H]+ shown in black. The HRMS spectrum of (B) monocyclic hexapeptide at 43.9 min and (C) bicyclic hexapeptide at 42.28 min.
Supplementary Figure 26. MSMS spectrum of one bicyclic hexapeptide in *A. parvosata ΔkisN*

The *Actinomadura parvosata ΔkisN* mutant produces several bicyclic hexapeptides 1-6 as shown in Supplementary Figure 25. (A) MSMS spectrum of the bicyclic hexapeptide at 40.92 min. It demonstrates a Tyr1-Trp2 and Hpg4-Tyr6 crosslinks. The Hpg4-Tyr6 crosslink is characterised by the presence of intact y3 and y4 with crosslinks and the absence of y2. Additionally, y3 and y4 can lose the chloroHpg at position 5 together with a CO (and NH$_3$) group. The fragmentation pattern shows again the m/z: y3#^-c = 301.1183 (B) and y4#^-c = 449.1343 (C) like in the monocyclic hexapeptide (see Supplementary Figure 23). The presence of a b2 ion with a crosslink (b2^) and related ions confirms the crosslink occurs within the N-terminal two residues (Tyr1-Trp2). Fragment b2^ indicates the presence of a crosslink between Tyr1-Trp2 in the one of the bicyclic forms of the hexapeptide produced by the *Actinomadura parvosata ΔkisN* mutant; the exact position of the crosslink is not clear. MSMS of other bicyclic species indicate the common presence of Tyr1 in these different links, including Tyr1-Dpg3 (data not shown). (D) Chemical structure of b2^.

Symbols: Y tyrosine; W tryptophan; h Hpg; d Dpg; c chloroHpg; # CO loss; * NH$_3$ loss; ^ crosslink (H$_2$ loss); The b ions extend from the N-terminus, and y ions extend from C-terminus. Fragmentation ions follow standard nomenclature for such experiments, see 3,4.
**Supplementary Figure 27. Kistamicin with different A-B ring crosslinking containing Hpg-7**

Comparison of the potential energies and free energies of the kistamicin 15-membered A-O-B ring relative to those of 13- and 12-membered AB ring analogues. Source data are provided as a Source Data file.

**Supplementary Figure 28. Kistamicin with different C-D ring crosslinking**

Comparison of the potential energies and free energies of the kistamicin 16-membered C-O-D ring relative to those of a 14-membered CD ring analogue. Source data are provided as a Source Data file.
Supplementary Figure 29. Kistamicin with different A-B ring crosslinking containing Dpg-7
Comparison of the potential energies of pekiskomycin (modelled as 9), containing the native 12-membered AB ring, relative to the pekiskomycin analogue 10 containing a 14-membered A-O-B ring. Source data are provided as a Source Data file.

11: Diphenyl ether 12: 2-Hydroxybiphenyl

$E_{rel} = 0 \text{ kcal/mol}$  $E_{rel} = -7.6 \text{ kcal/mol}$

$G_{rel} = 0 \text{ kcal/mol}$  $G_{rel} = -6.3 \text{ kcal/mol}$

Supplementary Figure 30. Diphenyl ether vs. 2-hydroxybiphenyl
Relative potential energies and free energies of diphenyl ether 11 and 2-hydroxybiphenyl 12. Source data are provided as a Source Data file.

Supplementary Figure 31. Geometry of the A-O-B ring in kistamicin
Comparison of the geometry of the A-O-B ring in kistamicin relative to the fully relaxed geometry. The substituents on the ether oxygen and $\alpha$-carbon have been modelled as hydrogens.
Supplementary Figure 32. Structural comparisons of the OxyA_kis/X_kis complex
(A) Overlay of the Oxy enzyme in the OxyA_kis/X_kis complex with the OxyB_teil/X_teil complex showing only the position of X_teil relative to the OxyA_kis/X_kis complex (X_teil shown in grey, OxyA_kis/X_kis complex coloured as in Figure 3, helices shown as cylinders). (B) Overlay of the X domains from the OxyA_kis/X_kis complex and OxyB_teil/X_teil complexes (coloured as in panel (A)). (C) Overlay of Oxy homologues on OxyA_kis structure (OxyA_kis shown in yellow; from left OxyA_teil (coloured green), OxyE_teil (coloured orange), OxyB_teil (coloured magenta), OxyC_van (coloured aquamarine)).
Supplementary Figure 33. CO spectra of OxyA and OxyC
UV/vis absorption spectra (black) and reduced, CO-complex (red) shown for purified OxyA<sub>KL</sub> (A) and OxyC<sub>KL</sub> (B). Both proteins are able to be reduced by sodium dithionite treatment that is then subsequently used to generate the reduced CO complex, as is shown by the shift of the absorption maximum from 418 nm to 450 nm.
Supplementary Figure 34. Synthesised peptide K1-7-CoA

(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide K1-7-CoA.
Supplementary Figure 35. Synthesised peptide K4-7D-CoA

(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide K4-7D-CoA. In red, the central Hpg residue, which is synthesised in D- and L-form.
RT: 17.880 min in a gradient of 5-35% in ACN (0.1% FA)

Supplementary Figure 36. Synthesised peptide K4-7L-CoA
(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide K4-7L-CoA. In red, the central Hpg residue, which is synthesised in D- and L-form.
Supplementary Figure 37. Synthesised peptide K2-4D-CoA

(A) Structure, (B) HPLC chromatogram and (C) MS spectrum (bottom) of synthesised peptide K2-4D-CoA. In red, the central Hpg residue, which is synthesised in D- and L-form.
Supplementary Figure 38. Synthesised peptide K2-4L-CoA

(A) Structure, (B) HPLC chromatogram and (C) MS spectrum (bottom) of synthesised peptide K2-4L-CoA. In red, the central Hpg residue, which is synthesised in D- and L-form.
Supplementary Figure 39. Synthesised peptide K3-6-CoA
(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide K3-6-CoA.
Supplementary Figure 40. Synthesised peptide K3-7-CoA

(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide K3-7-CoA.
Supplementary Figure 41. Synthesised peptide K1-4-CoA
(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide K1-4-CoA.
Supplementary Figure 42. Synthesised peptide P1-7-CoA
(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide P1-7-CoA.

RT: 19.136 min in a gradient of 10-40% in ACN (0.1% FA)

$[\text{M-2H}]^2/2 = 894.85$ Da; $\text{MW}_{\text{calculated}} = 1791.70$ Da
RT: 19.133 min in a gradient 15-15-45% in ACN (0.1% FA)

[M+H]^+ = 1839.42 Da; MW\text{calculated} = 1839.48 Da

[M+H]^2+/2 = 920.7 Da; MW\text{calculated} = 920.25 Da

Supplementary Figure 43. Synthesised peptide T1-7-CoA

(A) Structure, (B) HPLC chromatogram and (C) MS spectrum of synthesised peptide T1-7-CoA.
Supplementary Figure 44. Overview of extracted ion chromatograms of the Oxy-catalysed turnover
Extracted ions from analysis in positive mode [M+H]^+ of different kistamicin peptides shown with high magnification. Peaks with monocyclic mass are indicated by a pink star, peaks with bicyclic mass by a blue star. *peaks were subtracted from the calculation because they lie under the main peaks. MA, methylamine, linear mass (black), monocyclic mass (pink), bicyclic mass (blue) and tricyclic mass (green). Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure.
Supplementary Figure 45. XIC of the Oxy-catalysed turnover of kistamicin tetrapeptides K4-7D

Extracted ions from analysis in positive mode [M+H]+ of linear tetrapeptide K4-7D methylamide (MA) with a mass of 642.2 Da (black), monocyclic methylamide product with the mass of 640.2 Da (pink) and bicyclic methylamide product with the mass of 638.2 Da (blue) following methyamine cleavage. Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 42. XIC of the Oxy-catalysed turnover of kistamicin tetrapeptides K4-7L

Extracted ions from analysis in positive mode [M+H]^+ of linear tetrapeptide K4-7L methylamide (MA) with a mass of 642.2 Da (black), monocyclic methylamide product with the mass of 640.2 Da (pink) and bicyclic methylamide product with the mass of 638.2 Da (blue) following methylene cleavage. Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 47. XIC of the Oxy-catalysed turnover of kistamicin tetrapeptide K3-6
Extracted ions from analysis in positive mode [M+H]+ of linear tetrapeptide K3-6 methylamide with a mass of 642.2 Da (black), monocyclic methylamide (MA) product with the mass of 640.2 Da (pink) and bicyclic methylamide product with the mass of 638.2 Da (blue) following methyamine cleavage. Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 48. XIC of the Oxy-catalysed turnover of kistamicin pentapeptide K3-7
Extracted ions from analysis in positive mode [M+H]+ of linear pentapeptide K3-7 methylamide with a mass of 791.3 Da (black), monocyclic methylamide (MA) product with the mass of 789.3 Da (pink) and bicyclic methylamide product with the mass of 787.3 Da (blue) following methyamine cleavage. Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 49. XIC of the Oxy-catalysed turnover of kistamicin hexapeptide K1-6

Extracted ions from analysis in positive mode [M+H]⁺ of linear hexapeptide K1-6 methylamide with a mass of 1007.4 Da (blue), monocyclic methylamide (MA) product with the mass of 1005.4 Da (cyan) and bicyclic methylamide product with the mass of 1003.4 Da (green) following methylene cleavage. Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 50. XIC of the Oxy-catalysed turnover of kistamicin heptapeptide K1-7
Extracted ions from analysis in positive mode [M+H]^+ of the linear heptapeptide K1-7 methylamide (MA) with a mass of 1156.43 Da (black), monocyclic methylamide product with the mass of 1154.42 Da (pink) following methylamine cleavage, bicyclic product with the mass of 1152.40 Da (blue) and tricyclic product with the mass of 1150.39 Da (green). Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 51. XIC of the Oxy-catalysed turnover of kistamicin tetrapeptide K1-4
Extracted ions from analysis in positive mode [M+H]^+ of linear tetrapeptide K1-4 methylamide with a mass of 679.29 Da (black), monomeric methylamide (MA) product with the mass of 677.27 (pink) and bicyclic methylamide product with the mass of 675.26 Da (blue) following methyamine cleavage. Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 52. XIC of the Oxy-catalysed turnover of kistamicin tripeptides K2-4
Extracted ions from analysis in positive mode [M+H]+ of linear tripeptide methylamide with a mass of 532.21 Da (black) and monocyclic methylamide (MA) product with the mass of 530.2 Da (pink). (A) K2-4D, (B) K2-4L. Enzyme components present in each reaction are indicated adjacent to the Y-axis in each trace within the figure. Source data are provided as a Source Data file.
Supplementary Figure 53. HRMS analysis of OxyC turnover product K3-7

In the *in vitro* K3-7 assay with OxyC, there is a monocyclic pentapeptide and numerous bicyclic peptides present. (A) Base peak chromatogram of all ions (black), and extracted ion chromatograms of linear 791.3035 [M+H]+ (red), monocyclic 789.2879 [M+H]+ (green) and bicyclic 787.2722 [M+H]+ (blue) at 10 ppm. HRMS highlighting the linear (B), monocyclic (C) and bicyclic (D) pentapeptide. The abundant monocyclic co-elutes with both other versions, and forms low level gas-phase dimers.
Supplementary Figure 54. Analysis of OxyC turnover product K4-7

In the *in vitro* K4-7 assay with OxyC, there are the linear tetrapeptide 642.2558 [M+H]⁺ and one monocyclic tetrapeptide 640.2402 [M+H]⁺. (A) HPLC chromatograms, from top to bottom: TIC, Base peak full scan, Base peak MS2 scans, XIC linear tetrapeptide 642.2558 [M+H]⁺, XIC monocyclic tetrapeptide 640.2402 [M+H]⁺, XIC bicyclic tetrapeptide 638.22 [M+H]⁺. MSMS spectra of linear K4-7 (B) and monocyclic tetrapeptide (C). The monocyclic tetrapeptide appears to have a Hpg4-Tyr6 crosslink, as the Hpg7-MA is cleaved off. Fragmentation ions follow standard nomenclature for such experiments, see 3,4.
Supplementary Figure 55. Analysis of OxyC turnover product K1-6

In the in vitro K1-6 assay with OxyC, there is the linear hexapeptide 1007.3934 [M+H]+, several monocyclic hexapeptides 1005.3777 [M+H]+ and low levels of bicyclic hexapeptide 1003.3621 [M+H]+.

(A) Base peak chromatogram of linear (black), monocyclic (red) and bicyclic (green) hexapeptide targeted PRM, with selected peaks labelled. Representative MSMS spectra of labelled peaks of linear (B, peak 1 (37.7 min)) and monocyclic peptides (C, peak 2 (39.5 min)). The monocyclic spectrum shown supports a Hpg4-Tyr6 crosslink. The peak at 314.1492 m/z, labelled with @, represents a loss of Hpg and CO from the y3^ ion, which due to the sequence is somewhat ambiguous; however, it is more likely to result from a Hpg4-Tyr6 crosslink through loss of the Hpg5 residue and CO, rather than a loss of CO from a y2^- representing a Hpg5-Tyr6 crosslink as there is no loss of methylamine. Possible fragments are shown in (D), indicating different fragments with similar masses that makes analysis difficult. Several other linkages were present in the other monocyclic and bicyclic peaks. Fragmentation ions follow standard nomenclature for such experiments, see 3,4.
Supplementary Figure 56. Analysis of OxyC turnover product K1-4

In the in vitro K1-4 assay with OxyC, there are the linear tetrapeptide 679.2875 [M+H]⁺ and two different kind of monocyclic tetrapeptides 677.2718 [M+H]⁺. (A) Base peak chromatogram of linear (black) and monocyclic tetrapeptide (red). (B-D) MSMS of linear (B, peak 1 (34.2 min)) and monocyclic tetrapeptide (C, peak 2 (31.6 min), and D, peak 3 (34.3 min). Peak 2 appears to have a Tyr1-Hpg3 crosslink. Peak 3 represents a Tyr1-Hpg4 linkage, due to 255 ion abundance (linked Tyr and Hpg) and losses of Hpg and Trp from the parent with CO loss and virtual absence of both y1 and b3. Fragmentation ions follow standard nomenclature for such experiments, see 3,4.
Supplementary Figure 57. Comparison of the activity of OxyC homologues

A turnover assay was performed with a GPA type I OxyC (from chloroeremomycin) and type V OxyC (from kistamicin) to check AB/ A-O-B ring formation in peptides with different C-terminal residues. Peptides tested include a teicoplanin related peptide T1-7 bearing a Hpg7 residue (A) and a pekiskomycin (Type I) peptide P1-7 bearing a Dpg-7 residue (B). These assays show that the chloroeremomycin OxyC homologue is active for AB ring formation in the peptide with a Hpg-7 residue (13-membered ring based on comparison to reported activity for balhimycin OxyC homologue), whilst the kistamicin OxyC homologue is not active for A-O-B ring formation in a peptide with a Dpg-7 residue. Peptide preparation and cyclisation assays performed using reported protocols. Source data are provided as a Source Data file.
Supplementary Tables

Supplementary Table 1. $^1$H and $^{13}$C NMR spectroscopic data and HMBC correlations for Kistamicin A (700 MHz, methanol-$d_4$) Numbering of Kistamicin A according Supplementary Figure 4, HMBC correlation Supplementary Figure 5 and NMR spectra Supplementary Figures 6-11.

<table>
<thead>
<tr>
<th>Units</th>
<th>$^1$H</th>
<th>$^{13}$C</th>
<th>HMBC</th>
<th>Units</th>
<th>$^1$H</th>
<th>$^{13}$C</th>
<th>HMBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-1</td>
<td>-</td>
<td>135.9</td>
<td>-</td>
<td>D-1</td>
<td>-</td>
<td>131.1</td>
<td>-</td>
</tr>
<tr>
<td>A-2</td>
<td>7.73</td>
<td>134.7</td>
<td>A-6, A-4</td>
<td>D-2</td>
<td>7.26</td>
<td>121.3</td>
<td>D-5, D-6</td>
</tr>
<tr>
<td>A-3</td>
<td>7.57</td>
<td>124.4</td>
<td>A-1</td>
<td>D-3</td>
<td>6.96</td>
<td>118.1</td>
<td>D-1, D-4,</td>
</tr>
<tr>
<td>A-4</td>
<td>-</td>
<td>156.3</td>
<td>-</td>
<td>D-4</td>
<td>-</td>
<td>149.4</td>
<td>-</td>
</tr>
<tr>
<td>A-5</td>
<td>7.24</td>
<td>125.8</td>
<td>A-1, A-4</td>
<td>-</td>
<td>-</td>
<td>145.7</td>
<td>-</td>
</tr>
<tr>
<td>A-6</td>
<td>7.83</td>
<td>131.8</td>
<td>A-2, A-4</td>
<td>D-6</td>
<td>5.24</td>
<td>112.9</td>
<td>D-2, D-4,</td>
</tr>
<tr>
<td>A-1’</td>
<td>2.92, 3.61</td>
<td>41.7</td>
<td>A-2’, A-1,</td>
<td>D-1’</td>
<td>5.43</td>
<td>58.7</td>
<td>D-1, D-6</td>
</tr>
<tr>
<td>A-2’</td>
<td>4.33</td>
<td>59.3</td>
<td>-</td>
<td>D-2’</td>
<td>-</td>
<td>170.6</td>
<td>-</td>
</tr>
<tr>
<td>A-3’</td>
<td>-</td>
<td>170.6</td>
<td>-</td>
<td>D-NH*</td>
<td>8.55</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A-NH</td>
<td>8.17</td>
<td>-</td>
<td>E-1’</td>
<td>E-1</td>
<td>-</td>
<td>136.8</td>
<td>-</td>
</tr>
<tr>
<td>B-1</td>
<td>-</td>
<td>130.4</td>
<td>-</td>
<td>E-2</td>
<td>7.96</td>
<td>135.1</td>
<td>E-3, E-4,</td>
</tr>
<tr>
<td>B-2</td>
<td>5.55</td>
<td>132.5</td>
<td>B-3, B-4,</td>
<td>E-3</td>
<td>-</td>
<td>130.6</td>
<td>-</td>
</tr>
<tr>
<td>B-3</td>
<td>-</td>
<td>136.4</td>
<td>-</td>
<td>E-4</td>
<td>-</td>
<td>154.2</td>
<td>-</td>
</tr>
<tr>
<td>B-4</td>
<td>-</td>
<td>140.3</td>
<td>-</td>
<td>E-5</td>
<td>6.84</td>
<td>128.1</td>
<td>E-1, E-3,</td>
</tr>
<tr>
<td>B-5</td>
<td>-</td>
<td>150.7</td>
<td>-</td>
<td>E-6</td>
<td>7.38</td>
<td>130.4</td>
<td>E-2, E-4</td>
</tr>
<tr>
<td>B-6</td>
<td>5.70</td>
<td>109.8</td>
<td>B-2, B-4,</td>
<td>E-1’</td>
<td>4.99</td>
<td>55.5</td>
<td>E-1, E-6</td>
</tr>
<tr>
<td>B-1’</td>
<td>6.40</td>
<td>56.5</td>
<td>B-3</td>
<td>E-2’</td>
<td>-</td>
<td>171.3</td>
<td>-</td>
</tr>
<tr>
<td>B-2’</td>
<td>-</td>
<td>171.3</td>
<td>-</td>
<td>E-NH</td>
<td>9.53</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B-NH*</td>
<td>8.37</td>
<td>-</td>
<td>-</td>
<td>F-1</td>
<td>-</td>
<td>141.8</td>
<td>-</td>
</tr>
<tr>
<td>C-2</td>
<td>6.99</td>
<td>125.8</td>
<td>C-3, C-4,</td>
<td>F-2,6</td>
<td>6.15</td>
<td>105.2</td>
<td>F-3,5, F-4,</td>
</tr>
<tr>
<td>C-3</td>
<td>-</td>
<td>111.8</td>
<td>-</td>
<td>F-3,5</td>
<td>-</td>
<td>159.7</td>
<td>-</td>
</tr>
<tr>
<td>C-4</td>
<td>-</td>
<td>127.2</td>
<td>-</td>
<td>F-4</td>
<td>5.78</td>
<td>102.4</td>
<td>F-2,6</td>
</tr>
<tr>
<td>C-5</td>
<td>7.99</td>
<td>121.9</td>
<td>C-9</td>
<td>F-1’</td>
<td>4.94</td>
<td>58.0</td>
<td>F-1, F-2,6</td>
</tr>
<tr>
<td>C-6</td>
<td>6.84</td>
<td>123.4</td>
<td>C-4, C-7,</td>
<td>F-2’</td>
<td>-</td>
<td>172.1</td>
<td>-</td>
</tr>
<tr>
<td>C-7</td>
<td>-</td>
<td>133.1</td>
<td>-</td>
<td>F-NH</td>
<td>8.75</td>
<td>8.75</td>
<td>-</td>
</tr>
<tr>
<td>C-8</td>
<td>7.32</td>
<td>116.5</td>
<td>C-4, C-6,</td>
<td>G-1</td>
<td>-</td>
<td>126.9</td>
<td>-</td>
</tr>
<tr>
<td>C-9</td>
<td>-</td>
<td>139.1</td>
<td>-</td>
<td>G-2,6</td>
<td>6.70</td>
<td>131.8</td>
<td>G-4, G-1’</td>
</tr>
<tr>
<td>C-1’</td>
<td>3.01, 3.59</td>
<td>30.5</td>
<td>C-3, C-4</td>
<td>G-3,5</td>
<td>6.31</td>
<td>116.8</td>
<td>G-1, G-4</td>
</tr>
<tr>
<td>C-2’</td>
<td>6.29</td>
<td>54.9</td>
<td>-</td>
<td>G-4</td>
<td>-</td>
<td>157.8</td>
<td>-</td>
</tr>
<tr>
<td>C-3’</td>
<td>-</td>
<td>170.9</td>
<td>-</td>
<td>G-1’</td>
<td>2.90, 3.45</td>
<td>38.4</td>
<td>G-2,6, G-</td>
</tr>
<tr>
<td>C-NH*</td>
<td>9.29</td>
<td>-</td>
<td>-</td>
<td>G-2’</td>
<td>4.06</td>
<td>56.9</td>
<td>-</td>
</tr>
<tr>
<td>C-1-NH*</td>
<td>10.28</td>
<td>-</td>
<td>-</td>
<td>G-3’</td>
<td>-</td>
<td>175.6</td>
<td>-</td>
</tr>
</tbody>
</table>

* observed in methanol-$d_3$ (where all the amide protons were visible)
### Supplementary Table 2. Bacterial strains and plasmids used in this study

<table>
<thead>
<tr>
<th>Strains</th>
<th>Characteristics</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E. coli NovaBlue</strong></td>
<td>recA1, endA1, gyrA96, thi-1, hsdR17 (rK12, mK12&lt;sup&gt;+&lt;/sup&gt;) supE44, relA1, lac [F&lt;sup&gt;+&lt;/sup&gt;, proAB, lacI&lt;sup&gt;q&lt;/sup&gt;, lacZΔM15, Tn10] (Tet&lt;sup&gt;R&lt;/sup&gt;)</td>
<td>Novagen</td>
</tr>
<tr>
<td><strong>E. coli ET12567</strong></td>
<td>DNA methylation deficient donor strain for conjugation: F2dam13::Tn9, dcm-6, hsdM&lt;sup&gt;R&lt;/sup&gt;, recF143, zji-202::Tn10, gaiK2, galT22 ara-14, lacY1, xyl-5, leuB6, thi-1, tonA31, rpsL136, hisG4, tsx-78, mtl-1, glnV44</td>
<td>MacNeil, et al., 1992. &lt;sup&gt;9&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>E. coli ArcticExpress</strong></td>
<td>E. coli B F&lt;sup&gt;−&lt;/sup&gt;ompT hsdS&lt;sup&gt;(rB&lt;sup&gt;−&lt;/sup&gt;, mB&lt;sup&gt;−&lt;/sup&gt;)&lt;/sup&gt; dcm&lt;sup&gt;−&lt;/sup&gt; Tef, gal λ (DE3) endA Hte [cpn10 cpn60 Genta]</td>
<td>Agilent Technologies</td>
</tr>
</tbody>
</table>

#### Plasmids

<table>
<thead>
<tr>
<th>Plasmids</th>
<th>Characteristics</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>pJet1.2/blunt</td>
<td>rep(pMB1), bla(Amp&lt;sup&gt;R&lt;/sup&gt;), eco47IR, P&lt;sub&gt;lacUV5&lt;/sub&gt;, T7 RNA polymerase promoter. Part of the CloneJET&lt;sup&gt;™&lt;/sup&gt; PCR Cloning Kit used for high efficiency cloning of PCR products</td>
<td>Thermo Fisher Scientific Inc.</td>
</tr>
<tr>
<td>pRM4.2</td>
<td>pSET152ermEp*derivate (ΦC31 integration vector with ermE&lt;sup&gt;R&lt;/sup&gt; promoter, Apr&lt;sup&gt;R&lt;/sup&gt;) with artificial RBS, used for construction of pRM4_kisN and pRM4_KisO</td>
<td>Menges, et al., 2007. &lt;sup&gt;10&lt;/sup&gt;</td>
</tr>
<tr>
<td>pUB307</td>
<td>self-transmissible plasmid that mobilizes other plasmids in trans for DNA transfer into hosts: RP4, kan&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Bennett, et al., 1977. &lt;sup&gt;11&lt;/sup&gt;</td>
</tr>
<tr>
<td>pGUSA21</td>
<td>Promoter probe vector, pSETGUS with deleted KpnI fragment containing TipA promoter. gusA, Δint, ΔattB, MCS from pUC21. Used for gene inactivation</td>
<td>Myronovskyi et al., 2011. &lt;sup&gt;12&lt;/sup&gt;</td>
</tr>
<tr>
<td>pGUSA21_KisN-up_down</td>
<td>pGUSA21 deletion plasmid of kisN containing 1.5kb flanking regions each up-and downstream of the kisN gene</td>
<td>This study</td>
</tr>
<tr>
<td>pGUSA21_KisO-up_down</td>
<td>pGUSA21 deletion plasmid of kisO containing 1.5kb flanking regions each up-and downstream of the kisO gene</td>
<td>This study</td>
</tr>
<tr>
<td>pRM4_kisN</td>
<td>pRM4 with kisN cloned in Ndel + HindIII site</td>
<td>This study</td>
</tr>
<tr>
<td>pRM4_KisO</td>
<td>pRM4 with kisO cloned in HindIII + XbaI site</td>
<td>This study</td>
</tr>
</tbody>
</table>
### Supplementary Table 3. Primers used in this study

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence</th>
<th>Amplicon length bp</th>
<th>PCR Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>KisN-down Fw</td>
<td>CGCCTCTAGAGTCCCGTGAGTCATCAC</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>KisN-down Rv</td>
<td>GCGGCATATGGGGTGCCATCCTT</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>KisN-up Fw</td>
<td>CCGGACATATGACGGCCCGCGTCCCGC</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>KisN-up Rv</td>
<td>CGCAGGGCTTTCAGGGCCCGAGGGACT</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>KisO-down Fw</td>
<td>GCACGTCTAGAGGAGCAGACTCCCCTGTG</td>
<td>1519</td>
<td></td>
</tr>
<tr>
<td>KisO-down Rv</td>
<td>GACCATATAGCAGCTCTTCACAGCAGAGAGAGAG</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>KisO-up Fw</td>
<td>GCTGCATATGGGGCCCGGAGGCGCAGG</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>KisO-up Rv</td>
<td>CTGCAAGCTTGTGCGGGGCGGTCGCGGG</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>KGus-Fw</td>
<td>CGCCGTTGGATACACCAAGG</td>
<td>3271</td>
<td></td>
</tr>
<tr>
<td>KGus-Rv</td>
<td>GTGGACGCGGAGTCACTTGCG</td>
<td>1336</td>
<td></td>
</tr>
<tr>
<td>pRM4SeqFw</td>
<td>CACACGTGAGCGCCCGAGATGCTGTTG</td>
<td>1172</td>
<td></td>
</tr>
<tr>
<td>pRM4SeqRv</td>
<td>GTGTTAAAACGACGGCCAGTGGCAAGAGAGAGAG</td>
<td>1301</td>
<td></td>
</tr>
<tr>
<td>kisN-NdeI-Fw</td>
<td>GCTGCATATGGTGCGGCGGAGG</td>
<td>1172</td>
<td></td>
</tr>
<tr>
<td>kisN-HindIII-Rv</td>
<td>CTCAAGGCTTTCACACAGGTCACGGCCAG</td>
<td>1301</td>
<td></td>
</tr>
<tr>
<td>kisO-HindIII-Fw</td>
<td>CTGCAAGGCTTAGAGGGAATGTAGCGCCAG</td>
<td>1301</td>
<td></td>
</tr>
<tr>
<td>kisO-Xbal-Rv</td>
<td>GCGGTCTAGATGCATCGGCGGTCGCGGC</td>
<td>1172</td>
<td></td>
</tr>
<tr>
<td>Delta kisN Fw</td>
<td>GCGGCACCTGTCCGGCTCA</td>
<td>2500 (wt)</td>
<td></td>
</tr>
<tr>
<td>Delta kisN Rv</td>
<td>CGGAGCTGAGGCGGCGGCGG</td>
<td>2500 (wt)</td>
<td></td>
</tr>
<tr>
<td>Delta kisN Fw</td>
<td>GCGGCACCTGTCCCGCTCA</td>
<td>2500 (wt)</td>
<td></td>
</tr>
<tr>
<td>Delta kisN Rv</td>
<td>CGGAGCTGAGGCGGCGGCGG</td>
<td>2500 (wt)</td>
<td></td>
</tr>
<tr>
<td>Delta kisO f</td>
<td>CATCTGAGGCGCCACCCGGGTAC</td>
<td>2800 (wt)</td>
<td></td>
</tr>
<tr>
<td>Delta KisO r</td>
<td>CGCCCTCGGATTCGCGGCGCCG</td>
<td>2800 (wt)</td>
<td></td>
</tr>
<tr>
<td>X-for</td>
<td>ATTATACGATGCGGCGGCGGCCGTCCCGTCCG</td>
<td>1519</td>
<td></td>
</tr>
<tr>
<td>T7-rev</td>
<td>GCTAGTTATTGCCTACGGGG</td>
<td>1519</td>
<td></td>
</tr>
</tbody>
</table>
### Supplementary Table 4: Media used in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Mass (g/l)</th>
<th>Further information</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFM-agar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mannitol</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>soy meal full fat</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>agar</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>LB Broth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tryptone</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>yeast extract</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>NaCl</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>RS medium</td>
<td></td>
<td>pH 7.2; for plates 16g/L agar</td>
</tr>
<tr>
<td>Sucrose</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>K2SO4</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>MgCl2 (6H2O)</td>
<td>10.12</td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Difco Casaminoacids</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Trace elements</td>
<td>2ml</td>
<td></td>
</tr>
<tr>
<td>Difco Yeast extract</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>TES Puffer</td>
<td>5.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.893 L ion free H2O</td>
</tr>
<tr>
<td></td>
<td></td>
<td>autoclave separately</td>
</tr>
<tr>
<td>Supplementation R5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CaCl2</td>
<td>80 mL</td>
<td>3.68%</td>
</tr>
<tr>
<td>KH2PO4</td>
<td>10 mL</td>
<td>0.54%</td>
</tr>
<tr>
<td>L-Proline</td>
<td>15 mL</td>
<td>20%</td>
</tr>
<tr>
<td>Trace elements</td>
<td>1 L</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>autoclave separately</td>
</tr>
<tr>
<td>FeCl3 x 6H2O</td>
<td>200 mg</td>
<td></td>
</tr>
<tr>
<td>Na2B4O7 x 10H2O</td>
<td>10 mg</td>
<td></td>
</tr>
<tr>
<td>(NH4)6Mo7O24 x 4H2O</td>
<td>10 mg</td>
<td></td>
</tr>
<tr>
<td>CuCl2 x2H2O</td>
<td>10 mg</td>
<td></td>
</tr>
<tr>
<td>MnCl2 x4H2O</td>
<td>10 mg</td>
<td></td>
</tr>
<tr>
<td>ZnCl2</td>
<td>40 mg</td>
<td></td>
</tr>
<tr>
<td>ZYM 5052 Media</td>
<td></td>
<td>pH 7.4</td>
</tr>
<tr>
<td>Tryptone</td>
<td>10 g</td>
<td></td>
</tr>
<tr>
<td>Yeast extract</td>
<td>5 g</td>
<td></td>
</tr>
<tr>
<td>Na2HPO4</td>
<td>25 mM</td>
<td></td>
</tr>
<tr>
<td>KH2PO4</td>
<td>25 mM</td>
<td></td>
</tr>
<tr>
<td>NH4Cl</td>
<td>50 mM</td>
<td></td>
</tr>
<tr>
<td>NaSO4</td>
<td>5 mM</td>
<td></td>
</tr>
<tr>
<td>MgSO4</td>
<td>2 mM</td>
<td></td>
</tr>
<tr>
<td>Glycerol</td>
<td>5 g</td>
<td></td>
</tr>
<tr>
<td>D-Glucose</td>
<td>0.5 g</td>
<td></td>
</tr>
<tr>
<td>Lactose</td>
<td>2 g</td>
<td></td>
</tr>
</tbody>
</table>
**Supplementary Table 5. Crystallography data for the OxyA<sub>ksi</sub>/X<sub>ksi</sub> complex**

<table>
<thead>
<tr>
<th>Data Collection</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Beamline</td>
<td>MX1, AS</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9537</td>
</tr>
<tr>
<td>Space group</td>
<td>C2</td>
</tr>
</tbody>
</table>

**Cell parameters**

- a, b, c (Å): 118.8, 87.3, 96.0
- x, y, z (°): 90.00, 95.77, 90.00

<table>
<thead>
<tr>
<th>Data statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>50.0 – 2.65</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.7 (98.9)</td>
</tr>
<tr>
<td>Multiplicity&lt;sup&gt;1&lt;/sup&gt;</td>
<td>3.8 (3.8)</td>
</tr>
<tr>
<td>(I) / σ(I)&lt;sup&gt;1&lt;/sup&gt;</td>
<td>21.3 (2.1)</td>
</tr>
<tr>
<td>R&lt;sub&gt;merge&lt;/sub&gt;&lt;sup&gt;1&lt;/sup&gt;</td>
<td>0.04 (0.50)</td>
</tr>
<tr>
<td>R&lt;sub&gt;free&lt;/sub&gt;&lt;sup&gt;1&lt;/sup&gt;</td>
<td>0.026 (0.46)</td>
</tr>
<tr>
<td>CC1/2&lt;sup&gt;1&lt;/sup&gt;</td>
<td>0.99 (0.83)</td>
</tr>
</tbody>
</table>

**Refinement**

- R / R<sub>free</sub><sup>2</sup> | 0.212 / 0.264 |
- Bond length rmsd (Å) | 0.004 |
- Bond angle rmsd (°) | 0.634 |
- Favored (%)<sup>3</sup> | 94.7 |
- Disallowed (%)<sup>3</sup> | 0.7 |
- MOLPROBITY score | 2.27 (92<sup>rd</sup> percentile) |

**PDB ID**

6M7L

Stereo image of a portion of the electron density map for the X-domain (maps: 2Fo-Fc (blue, sigma level 1.5), positive Fo-Fc (green, sigma level 3.0) and negative Fo-Fc (red, sigma level 3.0)).

<sup>1</sup> Values in parentheses refer to the highest recorded resolution shell.

<sup>2</sup> 5% of reflections were randomly selected before refinement.

<sup>3</sup> Percentage of residues in the Ramachandran plot
Supplementary Table 6. Similar proteins to OxyA from kistamicin biosynthesis
Top ranking and carrier-protein interacting P450 structures with high levels of homology to the OxyA<sub>kis</sub> in complex with the X-domain as identified by Dali.\textsuperscript{13}

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Chain</th>
<th>Z-score</th>
<th>RMSD Ca [Å]</th>
<th>% Identity</th>
<th>Description (donor organism)</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>5HH3</td>
<td>C</td>
<td>44.2</td>
<td>1.5</td>
<td>48</td>
<td>OxyA&lt;sub&gt;tei&lt;/sub&gt; (Actinoplanes teichomyceticus)</td>
<td>14</td>
</tr>
<tr>
<td>3O1A</td>
<td>A</td>
<td>43.5</td>
<td>2.2</td>
<td>50</td>
<td>OxyE&lt;sub&gt;tei&lt;/sub&gt; (CYP165D3, Actinoplanes teichomyceticus)</td>
<td>15</td>
</tr>
<tr>
<td>1UED</td>
<td>A</td>
<td>43.2</td>
<td>2.1</td>
<td>44</td>
<td>OxyC&lt;sub&gt;van&lt;/sub&gt; (Amycolatopsis orientalis)</td>
<td>16</td>
</tr>
<tr>
<td>5EX8</td>
<td>A</td>
<td>43.0</td>
<td>2.1</td>
<td>39</td>
<td>StaF (Streptomyces toycacensis)</td>
<td>17</td>
</tr>
<tr>
<td>1LGF</td>
<td>A</td>
<td>42.9</td>
<td>2.0</td>
<td>43</td>
<td>OxyB&lt;sub&gt;van&lt;/sub&gt; (Amycolatopsis orientalis)</td>
<td>18</td>
</tr>
<tr>
<td>5EX6</td>
<td>C</td>
<td>42.7</td>
<td>1.9</td>
<td>44</td>
<td>StaH (Streptomyces toycacensis)</td>
<td>19</td>
</tr>
<tr>
<td>6FSH</td>
<td>D</td>
<td>41.4</td>
<td>2.0</td>
<td>44</td>
<td>Hybrid OxyB&lt;sub&gt;tei&lt;/sub&gt; BC/FG&lt;sub&gt;van&lt;/sub&gt; (Actinoplanes teichomyceticus)</td>
<td>20</td>
</tr>
<tr>
<td>4UBS</td>
<td>A</td>
<td>41.2</td>
<td>2.6</td>
<td>37</td>
<td>CYP105D7 (Streptomyces avermitilis)</td>
<td>21</td>
</tr>
<tr>
<td>5Y1I</td>
<td>B</td>
<td>41.1</td>
<td>2.2</td>
<td>35</td>
<td>GfsF (Streptomyces graminofaciens)</td>
<td>22</td>
</tr>
<tr>
<td>2ZBZ</td>
<td>A</td>
<td>40.9</td>
<td>2.8</td>
<td>33</td>
<td>CYP105A1 (Streptomyces griseolus)</td>
<td>23</td>
</tr>
<tr>
<td>4TVF</td>
<td>A</td>
<td>40.6</td>
<td>2.1</td>
<td>45</td>
<td>OxyB&lt;sub&gt;tei&lt;/sub&gt; (Actinoplanes teichomyceticus)</td>
<td>24</td>
</tr>
<tr>
<td>3BUJ</td>
<td>A</td>
<td>39.4</td>
<td>2.2</td>
<td>32</td>
<td>Cal02 (Micromonospora echinospora)</td>
<td>25</td>
</tr>
<tr>
<td>3EJB</td>
<td>H</td>
<td>38.5</td>
<td>2.5</td>
<td>34</td>
<td>P450&lt;sub&gt;ful&lt;/sub&gt; (CYP107H1, Bacillus subtilis)</td>
<td>26</td>
</tr>
<tr>
<td>4LOE</td>
<td>A</td>
<td>36.4</td>
<td>2.5</td>
<td>24</td>
<td>P450&lt;sub&gt;sky&lt;/sub&gt; (CYP163B3, Streptomyces sp. ACTA 2897)</td>
<td>27,28</td>
</tr>
<tr>
<td>3MGX</td>
<td>B</td>
<td>33.7</td>
<td>3.1</td>
<td>28</td>
<td>OxyD (Amycolatopsis balhimycina)</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 7. Similar proteins to the X domain from kistamicin biosynthesis

Top ranking structures with high levels of homology to the Xkis-domain in complex with OxyAkis as identified by Dali.13

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Chain</th>
<th>Z-score</th>
<th>RMSD Ca [Å]</th>
<th>% Identity</th>
<th>Description (donor organism)</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>4TX2</td>
<td>B</td>
<td>49.1</td>
<td>2.1</td>
<td>55</td>
<td>X-domain from teicoplanin Tcp12 synthetase (Actinoplanes teichomyceticus)</td>
<td>30</td>
</tr>
<tr>
<td>4TX3</td>
<td>B</td>
<td>48.9</td>
<td>1.6</td>
<td>55</td>
<td>X-domain from teicoplanin Tcp12 synthetase (Actinoplanes teichomyceticus)</td>
<td>30</td>
</tr>
<tr>
<td>2JGP</td>
<td>A</td>
<td>38.4</td>
<td>3.2</td>
<td>24</td>
<td>PCP and C-domain from tyrocidine synthetase (Brevibacillus brevis)</td>
<td>31</td>
</tr>
<tr>
<td>SU89</td>
<td>A</td>
<td>35.5</td>
<td>2.8</td>
<td>38</td>
<td>A, PCP and C domains from baccilibactin synthetase (Geobacillus sp. Y4.1MC1)</td>
<td>32</td>
</tr>
<tr>
<td>4ZXI</td>
<td>A</td>
<td>34.5</td>
<td>3.2</td>
<td>23</td>
<td>C-A-PCP-TE termination module from holo-AB3403 synthetase (Acinetobacter baumannii)</td>
<td>33</td>
</tr>
<tr>
<td>5DU9</td>
<td>A</td>
<td>33.9</td>
<td>4.9</td>
<td>19</td>
<td>C-domain from calcium-dependent antibiotic synthetase bound to acceptor mimic (Streptomyces coelicolor A3(2))</td>
<td>34</td>
</tr>
<tr>
<td>4JN3</td>
<td>B</td>
<td>32.7</td>
<td>4.9</td>
<td>18</td>
<td>C-domain from calcium-dependent antibiotic synthetase (Streptomyces coelicolor A3(2))</td>
<td>35</td>
</tr>
<tr>
<td>5T3D</td>
<td>A</td>
<td>31.7</td>
<td>3.5</td>
<td>19</td>
<td>C-A-PCP-TT termination module from enterobactin synthetase (Escherichia coli)</td>
<td>33</td>
</tr>
<tr>
<td>2VSQ</td>
<td>A</td>
<td>31.0</td>
<td>4.7</td>
<td>21</td>
<td>C-A-PCP-TE termination module from surfactin synthetase (Bacillus subtilis)</td>
<td>36</td>
</tr>
</tbody>
</table>

Supplementary Table 8. Comparison of the protein-protein interface within the two Oxy/X-domain complexes identified to date as performed by PISA.37

<table>
<thead>
<tr>
<th>Complex</th>
<th>Residues (X-domain)</th>
<th>Residues (Oxy)</th>
<th>Interface (Å²)</th>
<th>ΔG (kcal/mol)</th>
<th>ΔG P-value a</th>
<th>H-bonds</th>
<th>Salt bridges</th>
</tr>
</thead>
<tbody>
<tr>
<td>OxyB-Xtei</td>
<td>24</td>
<td>23</td>
<td>725</td>
<td>2.2</td>
<td>0.761</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>OxyA-Xkis</td>
<td>19</td>
<td>33</td>
<td>976</td>
<td>2.6</td>
<td>0.740</td>
<td>17</td>
<td>20</td>
</tr>
</tbody>
</table>

a Indication of hydrophobic nature of the interface, > 0.5 indicates the interface is less hydrophobic than usual for a protein-protein interface
**Supplementary Table 9. Theoretical molecular weight of synthesised kistamicin peptides**

Masses of starting material peptide-CoAs and their subsequent modifications during turnovers.

MA: peptide generated from methylamine cleavage to generate methylamide peptide

<table>
<thead>
<tr>
<th>Peptide</th>
<th>-CoA</th>
<th>MA linear</th>
<th>MA 1 ring</th>
<th>MA 2 rings</th>
<th>MA 3 rings</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1-7</td>
<td>1891.51</td>
<td>1155.43</td>
<td>1153.42</td>
<td>1151.40</td>
<td>1149.39</td>
</tr>
<tr>
<td>K4-7</td>
<td>1377.32</td>
<td>641.25</td>
<td>639.23</td>
<td>637.22</td>
<td>-</td>
</tr>
<tr>
<td>K2-4</td>
<td>1267.28</td>
<td>531.20</td>
<td>529.20</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>K3-6</td>
<td>1377.32</td>
<td>641.25</td>
<td>639.23</td>
<td>637.22</td>
<td>-</td>
</tr>
<tr>
<td>K2-7</td>
<td>1526.36</td>
<td>790.29</td>
<td>788.28</td>
<td>786.27</td>
<td>-</td>
</tr>
<tr>
<td>K1-6</td>
<td>1742.46</td>
<td>1006.39</td>
<td>1004.37</td>
<td>1002.36</td>
<td>1000.34</td>
</tr>
<tr>
<td>K1-4</td>
<td>1414.35</td>
<td>678.28</td>
<td>676.26</td>
<td>674.25</td>
<td>-</td>
</tr>
<tr>
<td>P1-7</td>
<td>1790.42</td>
<td>1054.35</td>
<td>1052.33</td>
<td>1050.32</td>
<td>1048.30</td>
</tr>
<tr>
<td>T1-7</td>
<td>1838.48</td>
<td>1102.41</td>
<td>1100.39</td>
<td>1098.38</td>
<td>1096.36</td>
</tr>
</tbody>
</table>
Supplementary Discussion

In terms of the peptide-producing NRPS, these differences are found in the 1st, 3rd and 4th modules, whilst the incorporation of Tyr-1, Trp-2 and Hpg-7 lead to a divergence in peptide structure (Figure 1, Supplementary Figure 1). In module 1, the lack of activation domain here implies that the tyrosine residue activated and loaded onto this position must be derived from an additional source. Previous work has shown that carrier protein (CP) domains can be aminoacylated by tRNA synthetases, and that A-domains can be utilised to load tRNAs, indicating a degree of crossover between these systems: this is one possible mechanism to account for the lack of the A_{core}-domain 1 in kistamicin biosynthesis. Furthermore, the D-configuration of this amino acid implies that the epimerisation of this residue would be occurring during initial peptide bond formation in the neighbouring condensation (C) domain, which is in agreement with early work on the chloroeremomycin (Type I) system. Further important differences found in the kistamicin NRPS when compared to Type I-IV GPAs are the placement of E-domains within modules 3 and 4: specifically, the presence of an E-domain in module 3 and a lack of such a domain in module 4. Whilst a module 3 E-domain has previously been reported in the NRPS producing A47934, this domain was shown to have been inactivated in this system, leading to the structure of A47934 matching other GPAs in possessing an (L)-configured residue at position 3 of the peptide. In the case of kistamicin, the presence of this E-domain agrees with the structural analysis performed for the initial identification of kistamicin that demonstrated a (D)-configured 3,5-dihydroxyphenylglycine (Dpg) residue at this position. The results of OxyC_{kis} reconstitution shows that peptides with residue 3 in a (D)-configuration are preferred over those with this residue in the (L)-configuration, which is consistent with the activity of this E-domain. The stereochemistry of the central Hpg residue installed by module 4 of the GPA NRPS is a key feature in all GPA structures, as this forms the central connection between both the C-O-D ring and D-O-E/DE ring of the GPA. All GPA structures to date have shown this central residue to be (D) configured. However, the kistamicin cluster does not possess an E-domain in this module (Figure 1). Given the central nature of this residue, alteration of the stereochemistry of this position to the (L)-form would appear highly unlikely, with OxyC_{kis} reconstitution again showing the importance of a (D)-configuration of Hpg-4. This implies a dual C/E function for one of the neighbouring C-domains that are involved with peptide bond formation at this residue, although the identity of this potential dual C/E-domain is not clear from phylogenetic analyses (Supplementary Figure 3). When comparing kistamicin with the only other characterised Type-V GPA complestatin, the relative similarity of these molecules is not repeated in their NRPS architectures, as modules 1, 3 and 4 in the complestatin NRPS conform to that of a standard GPA. Differences found within the complestatin NRPS – such as the extended linker between the X- and thioester (TE) domains in the final module, unique methyl transferase domain and
the silent additional E-domain within the 6th module – are not repeated within the kistamicin cluster, indicating significant diversity within the biosynthetic machineries of these two Type-V GPAs (Figure 1). 45
Supplementary Methods

OxyE complementation in *A. parvosata* WT

Due to the results from the *A. parvosata* ΔoxyA (ΔkisN) strain showing that OxyC_kis could install an addition crosslink between the Tyr-1 and Trp-2/Dpg-3 residues of the C-O-D crosslinked peptide, we also tested whether the addition of a Type IV OxyE enzyme to the kistamicin WT producer would be able to reproduce the insertion of an extra ring at the N-terminus of kistamicin. To this end, we ascertained that the heterologous expression of oxyE (staG) gene from the A47934 Type IV GPA into the *A. parvosata* WT kistamicin producer does not support insertion of an F-O-G ring between Tyr-1 and Dpg-3 in kistamicin. This lack of activity could stem from a number of reasons, concerning either the timing of OxyA ring formation restricting the activity of OxyE or changes in the peptide structure or the stereochemistry of the Dpg-3 residue.

Molecular Modelling of the rings in kistamicin

Molecular modelling calculations were undertaken to compare the stability of native kistamicin (containing a 15-membered A-O-B ring) relative to kistamicin analogues containing 13- or 12-membered AB rings. The peptides were modelled as structures 1–3, where the N-terminal tyrosine residue was replaced by an acetyl group (Figure 3). First, the conformations of each peptide were explored using the Macrocycle Conformational Sampling algorithm of MacroModel 11.7. The conformer sampling was conducted with the OPLS_2005 forcefield, using the GB/SA model to simulate solvation in water. The default settings of the Macrocycle Conformational Sampling protocol were used, in which 5000 cycles of large-scale low mode searches were performed on a set of seed structures obtained from 5000 cycles of MD simulated annealing, with eigenvectors determined for each new global minimum, and included the "enhanced" option for torsional sampling (which samples certain C–N and C–O single bonds with higher rotational barriers). For each molecule, the three most stable conformers identified by this protocol were then submitted to single-point energy calculation with density functional theory (DFT), performed with M06-2X/6-31G(d) in Gaussian 09. The conformer having the lowest DFT energy was used for analysis. This series of calculations provided the relative potential energies reported in Figure 3, which predict that the native kistamicin structure, containing the 15-membered A-O-B ring (1), is about 10 kcal/mol less stable than the analogues containing the 13- (2) and 12-membered (3) AB rings. In a similar series of calculations where the conformer sampling was performed with the OPLS3 forcefield rather than with OPLS_2005, the order of stabilities of 1–3 was the same, the spread between the most stable and least stable isomers was smaller (5 kcal/mol).

Each molecule was then truncated to give just the A-O-B or AB ring as shown in Supplementary Figure 27. Free valences were capped with hydrogen atoms. The 15-, 13-, and 12-membered macromonocycles 4–6 were subjected to full geometry optimization with M06-2X/6-31G(d). This series of calculations showed that the 15-membered A-O-B ring substructure of kistamicin (4) was again about 10 kcal/mol less stable than the 12- or 13-membered AB rings of the analogues (5, 6). This suggests that the origins of the difference in stability between native kistamicin and the smaller-ring AB analogues are localized largely within the A-O-B/AB ring.
A similar analysis was performed for the kistamicin C-O-D ring (Supplementary Figure 28). The native 16-membered C-O-D ring 7 was compared with the 14-membered CD ring analogue 8. In this case, the C-O-D ring was 3–4 kcal/mol less stable than the CD ring. The 14-membered biaryl structure in 8 is analogous to that found in arylomycins.

We also explored ring size effects in pekiskomycin, a GPA which natively contains an AB ring (Supplementary Figure 29). The native structure 9, containing a 12-membered AB ring, was compared with a ring-expanded analogue 10 containing a 14-membered A-O-B ring. In this case the native AB-containing structure was 2.7 kcal/mol more stable than the larger-ring A-O-B analogue.

Each of these sets of calculations reveals a common trend, namely that larger-ring diaryl ethers are less stable than the isomeric smaller-ring biaryl alcohols. This trend can, in part, be attributed to the intrinsic stabilities of diaryl ethers relative to biaryl alcohols. For example, calculations on diphenyl ether 11 and 2-hydroxybiphenyl 12 (Supplementary Figure 30) predict that ether 11 is about 7 kcal/mol less stable than biaryl alcohol 12, the same trend observed in the GPAs above.

Ring strain also contributes to the relative stabilities of diaryl ether versus biaryl alcohol-containing macrocyclic rings in GPAs. In kistamicin, the B ring is a locus of deformation: as shown in Supplementary Figure 31, the aryl ring is puckered and the ether oxygen and α-carbon are bent significantly out of plane.

The Cartesian coordinates for 1–12 are provided in the raw data folder. For species 1, 2, 3, 9 and 10, where a DFT single-point energy calculation was performed using the OPLS_2005 optimized geometry, the M06-2X/6-31G(d) single-point electronic potential energy (E) is reported. For species 4, 5, 6, 7, 8, 11 and 12, which were fully optimized with DFT, both the M06-2X/6-31G(d) electronic potential energy (E) and Gibbs free energy (G, 298.15 K, 1 atm) are reported.
Supplementary References


54 Gaussian 09, Revision E.01 (Gaussian, Inc., Wallingford, CT, 2013).
Chapter 6. Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors.
6 Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors

Key ideas: Protein engineering, protein design, applications

6.1 Preface

“What I can’t create I don’t understand”

Richard Feynman

events underpins our ability to engineer protein-based tools that have novel, useful functions – from new enzymes for use in the pharmaceutical industry to protein tools that allow researchers to detect the presence of small molecules in biological samples. In this review paper, we discuss several structural and evolutionary approaches that can be taken for the development of fluorescence-based small molecule biosensors.

The design of biosensors is not trivial. An effective small-molecule biosensor must bind to its target molecule with an affinity and specificity suitable for the desired application. While some sensors can be constructed using naturally occurring binding proteins that already selectively bind the target molecule, others (especially those targeting non-physiological ligands) must be engineered or designed so that they can detect the novel ligands. Successful design of the desired binding activity relies on a good understanding of how the architecture of protein binding sites influence the properties of molecular recognition events, but also on how protein dynamics affect ligand binding and resulting output signal. Further, biosensors must couple binding events with a measurable output, such as a change in a fluorescent signal, and the overall architecture of the biosensor can greatly influence what type of signal is produced.

As discussed in this review, current and emerging approaches for the design and optimisation of small molecule biosensors make use of evolutionary information (such as
Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors

sourcing suitable natural protein scaffolds, or using ancestral sequence reconstruction to generate thermostable and engineerable starting designs), improved fluorescent molecule designs, modular architectures, linkers and dynamics, and high-throughput screens.

Figure 6-1. Graphical Abstract from “Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors”.

<table>
<thead>
<tr>
<th>1. Select or engineer suitable components</th>
<th>2. Generate sensor architecture</th>
<th>3. Test and select</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="#">Diagram</a></td>
<td><a href="#">Diagram</a></td>
<td><a href="#">Diagram</a></td>
</tr>
</tbody>
</table>

Recognition domain | Linker Regions | Reporter Domain (Fluorescent Protein)

4. Optimise
6.2 Review paper: Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors

6.2.1 Author’s contribution

I was one of four co-first authors involved in researching and writing this review paper and played a leading role in the planning and preparation of the manuscript. I wrote the majority of the abstract, introduction, conclusions, and the sections on selecting and engineering suitable recognition domains and the optimization of the complete sensor. In addition, I contributed to the writing and editing of other sections of this paper. I produced the graphical abstract and Figure 1.

6.2.2 Contributions from others

The three other co-first authors, Joshua Mitchell, Matthew Spence and Vanessa Vongsouthi, contributed significantly to the planning, research, writing and editing of this review paper. Professor Colin Jackson helped to plan the overall structure of the review and edited the final manuscript.
Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors
Joe A Kaczmarski¹, Joshua A Mitchell¹, Matthew A Spence¹, Vanessa Vongsouthi¹ and Colin J Jackson

Biosensors that selectively report on the presence of specific small molecule analytes have applications in many fields of research, medicine and biotechnology. Here, we review recent advances and emerging approaches in the design and optimisation of genetically encoded fluorescence-based small molecule biosensors. We discuss how natural sensory proteins can be exploited to produce novel biosensors and the strategies for optimizing ligand specificity and fluorescence readout. Finally, we provide insight into high-throughput sensor optimisation and discuss the challenges that are faced when designing novel biosensors.

Address
Research School of Chemistry, The Australian National University, 137 Sullivans Creek Road, Acton, ACT 2601, Australia

Corresponding author: Jackson, Colin J (colin.jackson@anu.edu.au)
¹ These authors contributed equally to this work.

Introduction
The development of robust and sensitive genetically encoded biosensors, which can reliably report on the detection of small molecules in vivo and in situ with good spatiotemporal resolution, is of interest to research fields such as medical diagnostics, synthetic biology, and agriculture and environmental monitoring. To be effective, a biosensor needs to be specific for the target molecule, provide an output with a high signal-to-noise ratio (SNR) and good spatiotemporal resolution, be sensitive over biologically relevant concentrations and not significantly change the biological environment, in which it is applied. In this review, we discuss current and emerging structural and evolutionary approaches to the design and optimisation of novel genetically encoded protein-based biosensors that couple the binding of a target analyte at a recognition domain with changes in optical output from fluorescent proteins (FPs). The output from fluorescent biosensors is easily measured by fluorescence spectroscopy and can produce spatiotemporal resolution suitable for non-invasively probing complex biochemical events in real time.

The de novo design of biosensors remains challenging. Accordingly, current design efforts typically take a nature-inspired modular approach to biosensor design, mix-and-matching natural or engineered recognition domains and FP s to create new sensors. While some properties of biosensors can be introduced by careful selection or structure-guided design of individual sensor components, rational engineering efforts can be hindered by a lack of high-resolution structural information, or by gaps in our understanding regarding complex qualities such as protein allostery. In such cases, iterative rounds of high-throughput screening (HTS) or directed evolution using the complete biosensor can optimize selectivity, sensitivity, stability, kinetics, orthogonality and dynamic range (Figure 1).

Small molecule biosensors in nature
Understanding the diversity and evolution of naturally occurring sensory proteins can guide the design of novel biosensors. Nature employs a limited repertoire of protein folds to construct the complex sensory machinery that organisms rely on for responding to changing physio-chemical conditions. For example, four-helical bundle (4HB), cache and Per-Arnt-Sim (PAS) domains collectively account for approximately 80% of the recognition domains responsible for sensing small molecules among model bacteria, while other folds such as periplasmic solute-binding proteins (SBPs), GAF and calmodulin-like (CaM) domains are comparatively rare [1] (Figure 2a). Ligand-binding domains are typically coupled with response elements such as DNA-binding domains, kinases and ATP-binding cassette (ABC) transporters to create the modular biosensors that are central to cell regulation and signalling (Figure 2b).

Natural sensory proteins and their components can be exploited to construct artificial biosensors. For example, natural allosteric transcription factors (TFs) can be repurposed as small molecule biosensors by using them to
regulate the expression of reporter genes, such as GFP [2]. Novel TF-based biosensors can be generated by combining natural ligand-binding and DNA-binding domains [3*], engineering novel ligand specificity into the recognition domain or through promoter engineering [4**,5]. TF-based biosensors are, however, susceptible to promiscuous cross-reactivity with endogenous transcriptional machinery and often suffer from poor temporal resolution [6]. In contrast, fluorescence-based biosensors that directly link ligand binding at the recognition domain with changes in the optical output from flanking FPs can be used to report on analyte dynamics which occur over subsecond timescales.

**Fluorescence-based biosensor architectures and readout**

Fluorescence-based biosensor designs can be classified based on the number of integrated FPs, or by the type of optical output they produce. Single-FP sensors (Figure 3a) link the natural or engineered sensitivity of the fluorophores within the FPs to changes in environmental variables (‘intrinsic’), such as pH and halide ion concentration [7,8], or to changes in the conformation of attached recognition domains (‘extrinsic’). Extrinsic single-FP biosensors are typically constructed by fusing a suitable recognition domain with a circularly permuted FP (cpFP), or by inserting the ligand-binding domain between two halves of a split FP [9]. Single-FP biosensors are traditionally intensiometric, with ligand-induced changes in fluorescence intensity being measured at a single wavelength (Figure 3c). Intensiometric biosensors are highly sensitive, with high SNRs and dynamic ranges. However, intensiometric readouts do not provide absolute quantitative information regarding analyte concentrations, and are easily affected by imaging and instrumental artefacts, as well as changes in the concentration of the sensor.

One approach to overcoming these problems is to engineer FPs that exhibit dual excitation or emission behaviour. For example, dual emission, extrinsic single-FP sensors for monitoring ammonium transport were generated by making structure-guided mutations to the linkers between a FP and a membrane transporter in an established sensor [10]. Another approach involves fusing single-FP biosensors with a spectrally distinct FP that acts as an internal reference, as in the recently described GCaMP-Rs [11] and ‘Matryoshka’ biosensors (Figure 3b) [12*]. Matryoshka biosensors can be constructed from suitable recognition domains in a single cloning step, by insertion of a single cassette containing an internal reference FP that is nested within the peptide loop of the reporter cpFP. Another novel approach, which has been used to generate ratiometric Ca²⁺ biosensors, is based on the reversible exchange of heterodimeric binding partners of red and green dimerisation-dependent FPs [13].

The most common double-FP sensors are those based on Förster Resonance Energy Transfer (FRET). FRET is a mechanism of non-radiative energy transfer that occurs when the emission spectrum of an excited fluorophore (donor) overlaps with the absorption spectrum of another fluorophore (acceptor). FRET sensors can be easily constructed by fusing two FPs to a suitable recognition domain (Figure 3b). However, since FRET is highly
Figure 2

(a) Natural recognition domains

SBP (e.g. E. coli LacI, ATP-transporters)

CaM (Calmodulin)

4HB (e.g. E. coli Tsr, Tar chemosensing complexes)

(b) Natural biosensors

Chemoeffector molecules (input)

Recognition domain (e.g. 4HB, PAS, Cache, GAF, SBP)

Response domain (e.g. DNA-binding motif, ATP-transporter, protein kinase)

Genetic induction (e.g. E. coli LacI)

Nutrient transport (e.g. ABC transporter)

Phosphorylation cascade/chemotaxis (e.g. E. coli Tsr complex)

Adapted phenotype (Output)

Current Opinion in Structural Biology

Natural biosensors and small molecule recognition domains.

(a) Common recognition domain folds. Both SBPs and CaM undergo large conformational changes in the presence of their cognate ligands and are regularly used as starting scaffolds for novel fluorescent biosensors. 4HB domains are among the most abundant and important sensory proteins in natural biosensors. (b) Typical architectures of natural biosensors involved in cell signalling and regulation. The recognition domain (green) detects small molecule metabolites (red) and translates that signal to the response domain (magenta). Once activated, the response domain elicits a phenotypic response, for example: altered transcriptional profile, nutrient transport or initiation of phosphorylation cascades in chemotaxis and quorum sensing.

Since the discovery of GFP, engineering efforts have produced a continuously expanding palette of GFP-like FPs with different colours, fluorescent properties and physical characteristics. These FPs are suited to a range of applications and allow for simultaneous multi-parameter measurements. Circular permutation, site-directed mutagenesis, and structure-guided evolution have been used to produce FPs with improved brightness, photostability, quantum yield and maturation rates [18–20]. Although developments in the field of FP-engineering have slowed, there have been a number of notable studies published recently. Dou et al. [21] presented the de novo design of a fluorescent beta-barrel protein that is significantly smaller than GFP, using a process combining Rosetta-based design, molecular docking, yeast-surface display, next-generation sequencing and X-ray crystallography. In another study, semi-rational mutagenesis and colony screening were used to design a bright cyan-excitable orange fluorescent protein that can be simultaneously used with GFP in dual-emission microscopy in vivo [22]. While these new reporter
domains are likely to find various applications in the field in the near future, cyan and yellow FPs (and their derivatives) remain the most commonly used FP pair for FRET sensors, since they provide optimal spectral overlap [14].

Selecting and engineering suitable recognition domains
In their simplest form, recognition domains consist of a single sensing domain that binds the ligand of interest. While significant progress has been made towards the de novo design of protein scaffolds for the selective binding of small molecules [23], most recognition domains are based on natural ligand-binding proteins that undergo ligand-induced conformational changes. For example, the large Ca$^{2+}$-dependent conformational changes of CaM and troponin C continue to be exploited to produce a wide range of FRET and split-GFP calcium biosensors [24,25,12**]. The SBP fold, which undergoes a large hinge-bending conformational change upon ligand binding, is another popular scaffold for the design of FRET sensors for small molecules [26*,27]. Recognition domains have recently been created from other dynamic
binding protein scaffolds, such as hormone receptors [28]. Ligand-induced conformational changes can also be engineered from more rigid ligand-binding scaffolds through the addition of domains to create novel architectures. Such sensor designs include SNIFITS, in which the binding of the target molecule displaces an FP-associated intramolecular ligand [29] and designs that incorporate additional frames of protein folding to facilitate ligand-dependent FRET [30].

Starting from a thermostable protein scaffold can help to accommodate destabilizing function-switching mutations. Although biosensors have been engineered from thermophilic binding proteins from *Thermotoga maritima*, thermophilic binding proteins are seldom good starting scaffolds for biosensor design as they are often active only in the high temperatures that the parent organism has adapted to. Sequence-based engineering strategies that leverage phylogenetic information, such as consensus design and ancestral sequence reconstruction (ASR), can generate protein scaffolds that have greater thermostability, and are often more promiscuous, than their contemporary counterparts [31,32] and are viable design strategies in the absence of a suitable, naturally sourced starting scaffold. For example, Whitfield *et al.* [33] used an ancestrally reconstructed SBP as the starting point to engineer a robust and selective FRET biosensor that can accurately report t-arginine concentrations in rat brain slices under physiological conditions.

Perhaps the greatest challenge in biosensor engineering is introducing novel ligand specificity into existing scaffolds. Diverse strategies can be used to yield functional and selective biosensors with desired ligand affinities; examples of these range from binding-pocket grafting [34], structure-based rational design [26*], computational approaches and directed evolution [35]. There remains no best design approach for engineering ligand selectivity. Instead, engineering strategies are dictated by the requirements of the mature biosensor and properties of the starting scaffold. For example, Zhang *et al.* [26*] engineered a novel glycine FRET-based biosensor through iterative rounds of structure-based rational design, whereas Taylor *et al.* [4**] engineered novel biosensors using an *in vivo* high-throughput screening-selection system.

**Designing and modelling linkers**

For sensors that rely on reporting a conformational change, the relative positioning of the domains is important. This can be fine-tuned by circularly permuting the recognition domain [36] or fluorophores [37,38], by engineering contacts between reporter domains [39,40], or most commonly by engineering the linker sequences connecting the different domains. Linker sequences are generally compared on the basis of their length and flexibility [41*]. Flexible linkers, epitomized by glycine-serine repeats such as [GGS]$_n$, are largely unstructured and are thought to tether the domains together like a rope, only minimally constraining their movement beyond keeping the fused termini within some distance of each other. Rigid linkers, most commonly the alpha-helical [EAAAK]$_n$ repeat but also proline-rich sequences, are highly structured and thought to constrain the interdomain distance to a set value. Linkers with intermediate flexibility can be made by mixing glycine-serine and helical repeats [42] or by increasing serine content relative to glycine [41*].

Linker choice is highly system-dependent. SBP-based FRET sensors generally benefit from very short linkers that don’t let the FP move independently of the SBP’s conformational change [15,3*], but relatively long rigid linkers have been shown to dramatically improve the performance of other sensors [26*]. A number of sensors have been designed in which the linker is an essential part of the recognition element; these are still often optimized by tuning linker length and flexibility [43–45]. The complexity of engineering appropriate linkers for a sensor makes computational design an appealing prospect. Random coil models derived for synthetic polymers have been applied to model the behaviour of some linkers [46,45,41*], providing an efficient and intuitive way to quantify linker flexibility (Figure 4). A few groups have used biophysical force fields through molecular dynamics, but force field quality, the considerable size of a fusion protein and the long timescales associated with domain movement are formidable barriers [45]. As a result of the difficulty of rational and computational design, libraries of linkers are often recombined into sensor constructs [28,41*].

**Optimizing complete biosensors**

The selection and engineering of the recognition, reporter and linker modules can initially be considered separately. However, biosensor components often need to be further optimized in the context of the complete sensor. Indeed, in the absence of generally applicable strategies for the *de novo* design of novel fluorescence-based biosensors, several iterative rounds of optimisation are sometimes required to create effective small molecule biosensors that can be used *in vivo*. Many modern design approaches generate large libraries (10$^6$–10$^9$ variants) of sensors and then select or screen for desirable characteristics. Fortunately, fluorescence-based biosensors are particularly well suited for high-throughput screening methods for optimizing biosensor properties. For example, random mutagenesis followed by selection has also been used to create sensors with altered binding specificities [47**]. Nadler *et al.* [48] described a library-based approach for identifying allosteric hotspots for the insertion of cpGFP into recognition domains based on FACS screening and next-generation sequencing. Similarly, Younger *et al.* [44] recently
Gaussian chain models of flexible linkers GGS and SSS. Amino acid residues are modelled as links in a freely jointed chain. The angles between these links are chosen randomly from a uniform distribution. The lengths of these links are specified by the characteristic ratio $c_m$, which quantifies the flexibility of the chain ($c_m = 1.9$ for [GGS], $c_m = 3.4$ for [SSS]). The probability density distribution of the linker’s end-to-end distance can be calculated analytically (grey curve) providing a model of domain separation of fusion proteins by long, flexible linkers. Four example, chains are computed for each linker and plotted with one end at the origin (black circle) and the other on the positive x-axis (coloured crosses). The model shows that increasing linker length and decreasing glycine content does more to broaden the distribution than actually increasing the end-to-end distance, as most conformations of the linker ball up on themselves. Note that the Gaussian chain model is most appropriate in the limit of long, flexible linkers; shorter or more rigid linkers are better modelled by more detailed methods such as the worm-like chain.

Presented an approach to generate and select biosensors based on transposon-mediated protein fusion. Directed evolution approaches have also been used to generate biosensors from a range of ligand-binding domains [49], or to create biosensors that can report on molecular dynamics at the surface of cells [35].

**Concluding remarks**
The successful design of novel small molecule fluorescence-based biosensors requires the optimisation of many properties, including selectivity, sensitivity, stability, kinetics and dynamic range. Biosensors with a range of characteristics have been constructed by fusing one or more suitable FPs with naturally occurring or engineered ligand-binding domains. While SBP-based FRET sensors remain popular, several new biosensor architectures, such as Matryoshka biosensors and SNIFITs, have provided additional, generalizable platforms for the rapid development of fluorescence-based biosensors from a range of recognition domains. The design of novel ligand specificity, as well as the optimisation of the relative positioning of sensor domains, remains among the most challenging aspects of designing biosensors, but can be aided by high-resolution X-ray structures, new computer modelling algorithms, high-throughput screening of large libraries of variants, and insights from molecular evolution studies which seek to identify the molecular determinants that underlie ligand-specificity and protein dynamics. In coming years, we expect to see many new generalizable protocols for the construction of analyte
biosensors, including approaches based on de novo design, which will help to minimize the need for the costly empirical optimisation that traditional approaches have relied upon.

**Conflict of interest statement**
Nothing declared.

**Acknowledgements**
GJJ acknowledges support from the ARC Discovery Project and Future Fellowship schemes, and the NHMRC. JAK, JAM, MAS, and VV are supported by Australian Government Research Training Program (RTP) Scholarships.

**References and recommended reading**

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


The authors fuse ligand-binding domains with DNA-binding domains to create novel classes of transcription factors. They produced two novel benzene-sensing transcription-factor-based biosensors.


The authors engineered novel ligand specificity into a natural transcription factor using computational protein design, saturation mutagenesis and high throughput screening. This permitted them to change ligand specificity without disrupting the allosteric performance of the transcription factor.


The authors developed a generalizable protocol for the construction of novel ‘Matryoshka’ ratiometric biosensors in a single cloning step. The approach employs a single cassette containing a reference FP nested within a reporter FP and was used to convert existing intensiometric sensors into ratiometric sensors.


The authors used Rosetta for the first time to design a fluorescent beta-barrel protein which is smaller than GFP. The engineering approaches used in this paper show potential for use in designing novel families of both fluorescent and binding proteins.


The authors developed the ratiometric sensor specific for glycine from a promiscuous GABA binding domain and improved its performance with a rigid linker. They applied it to test theories about glycine’s role as a neurotransmitter in the hippocampus.


The authors develop a simple, intuitive model that describes the behaviour of long, flexible linkers. They parameterize the model with respect to composition and length of a commonly used family of linker repeat sequences.


The authors present a general method for developing small molecule biosensors in eukaryotes, which is based on the conditional stabilisation of a destabilized fusion protein upon ligand binding. They demonstrate that these biosensors can be used in yeast, mammalian, and plant cells, and then show that they can be used to improve biosynthesis of progesterone and regulate CRISPR activity in mammalian cells.


Chapter 7. Conclusions, implications and future directions
7 Conclusions, implications and future directions

The work presented in this thesis has focussed on using biophysics, biochemistry and bioinformatics to explore the evolutionary drivers of diverse molecular recognition events. Using this approach, we have provided insight into the structural and thermodynamic factors that underly the biological function of a diverse set of proteins – from P₄₅ signal transduction proteins to anti-PfCSP_repeat antibodies. The greatest impact of this work will arise from using this new information to guide future studies in which the structure and functions of these molecules are modified, enhanced and/or targeted to provide useful outcomes in areas such as improving crop yields, designing novel enzymes, improving malaria vaccines, or developing new antibiotics. Indeed, I hope that the work presented in this thesis has highlighted the powerful connection between understanding the molecular drivers of protein-ligand interactions and developing solutions to some of the most challenging global problems (such as feeding a growing global population (Chapter 2), finding ways to overcome antibiotic resistance (Chapter 5) and the eradication of infectious disease (Chapter 4)).

The findings presented in this thesis are already being used to guide on-going studies and have been supported by data from other groups. In the following sections I discuss how the work presented in each chapter of this thesis has contributed to the field, cite recent publications that have built on the ideas explored in this work, and provide ideas about future directions for these projects.
7.1 New insight into SbtB and the P\textsubscript{II} protein fold

7.1.1 On the function of SbtB

In Chapter 2 I presented work describing the structural basis for the binding of effector molecules, including ATP, to a cyanobacterial P\textsubscript{II}-like protein, SbtB. Unlike the previous study on an SbtB from a related cyanobacteria (SbtB6803), which did not show any differences between ligand bound states (Selim et al., 2018), our work revealed a clear ATP-induced structural change in the flexible T-loops, allowing us to develop a testable model for how SbtB regulates the SbtA bicarbonate transporter in response to changing conditions. These results are guiding on-going experiments in the Price Group (RSB, ANU), which are investigating the effect of nucleotides and other effector molecules (such as Ca\textsuperscript{2+}) on the association and dissociation of the SbtA-SbtB complex. Unpublished data from these experiments support the predictions made by us in Chapter 2; specifically, that ATP destabilises the SbtA-SbtB complex, while AMP/ADP binding stabilises the complex (Dean Price, personal communication, 2019).

Several questions about the function of SbtB remain unanswered. For example, what role does T-loop phosphorylation have in the SbtB-mediated regulation of SbtA (Spât et al., 2015)? Is SbtB required for a functional transgenic CCM, and will the conditions inside the chloroplasts of C3 plants be compatible with its function? Future studies could focus on exploring these questions. It is also unclear if SbtB has any non-SbtA targets. However, it has been noted that \(\beta\)-type SbtBs sometimes co-purify from cyanobacteria with another small, yet uncharacterised protein (Dean Price, personal communication, 2019). Characterising this unknown protein would be an obvious next step for uncovering other potential targets of SbtB.

7.1.2 On the function and evolvability of SbtB and the P\textsubscript{II} fold

Proteins with the P\textsubscript{II} fold make up one of the most widely distributed groups of signal transduction proteins in nature (Huergo et al., 2013); the work presented in Chapter 2 extends our understanding of non-canonical P\textsubscript{II} proteins and further demonstrates the evolvability of this fold. In particular, it shows how the function of P\textsubscript{II} proteins can be readily tweaked through the accumulation of mutations or insertions within the flexible regions of the protein without
compromising the structural integrity of the overall fold. This high level of “structural polarity” – that is, a rigid, thermostable core that affords mutational robustness (Bloom et al., 2006) combined with dynamic loops and termini that facilitate promiscuous functions (James and Tawfik, 2003b; Petrovic et al., 2018) and that accumulate mutations that modulate substrate specificities and reactivities (Dellus-Gur et al., 2013) – makes the P$_{II}$ protein fold especially evolvable (Figure 7-1). Indeed, conformational flexibility and the ability for the binding site to adopt more than one configuration also allows for binding of more than one ligand. Furthermore, this flexibility means that the binding site residues can be solvent exposed when the protein is in the open state (and thus be more tolerable to function-altering mutations) and can still form solvent-excluded sites in the closed conformation (which is normally required for function). In addition, because residues involved in binding come from parts of the protein that move independently of each other, the P$_{II}$ fold couples ligand binding with a ligand-induced conformational changes – a characteristic that is central to the function of signalling proteins.

Because it has these characteristics, the P$_{II}$ fold has attracted attention as an ideal scaffold for protein engineering. For example, the P$_{II}$ fold has been used to create fluorescent sensors for the common P$_{II}$-ligands ATP (Berg et al., 2009) and 2-oxoglutarate (Chen et al., 2018). Furthermore, only a small number of mutations are required to produce P$_{II}$ proteins that bind novel ligands or that have altered regulatory function. For example, a single mutation in *Synechococcus elongatus* P$_{II}$ produced a protein that recognised citrate (Zeth et al., 2012); this variant was also used to engineer a cyanobacteria that could produce large amounts of a biotechnologically interesting compound, cyanophycin (Watzer et al., 2015). Structure-function studies on diverse P$_{II}$ and P$_{II}$-like proteins (such as the work on SbtB7001 presented in Chapter 2) improve our understanding of how mutations effect function in this adaptable fold and may aid future engineering applications.
Conclusions, implications and future directions

Figure 7-1. Structural polarity and the evolvability of protein folds. The evolvability of a protein is often attributed to two characteristics: mutational robustness and conformational flexibility. Protein folds, such as the $\Pi_2$-like fold (e.g. SbtB, Chapter 2), achieve both of these properties by having high structural polarity.

7.2 The role of remote mutations and dynamics in the evolution of a new enzyme activity

In Chapter 3 I presented work that explored the evolution of an enzyme from a non-catalytic solute binding protein (SBP). In this work we showed that the evolution of cyclohexadienyl dehydratase (CDT) occurred via several distinct steps. First, several residues involved in the ancestral function were retained throughout the evolutionary trajectory and coadapted for the new function. Second, active site substitutions introduced catalytic residues and altered the shape of the active site cavity. Finally, outer shell mutations were needed to optimise catalytic activity by either controlling the conformations of the active site residues or by altering the overall conformational landscape of the protein.

7.2.1 Comparison with recent studies and de novo enzyme design approaches

The first paper (Chapter 3.2) was published back-to-back in Nature Chemical Biology with a study that used a very similar approach to explore the structural basis of catalysis from a different non-catalytic protein fold. In this work, Kaltenbach and colleagues used ancestral protein reconstruction, directed evolution, X-ray crystallography, NMR and MD simulations to
explore the evolutionary emergence of a chalcone isomerase from a non-catalytic fatty-acid binding protein (Kaltenbach et al., 2018). Similar to our work, they found that rearrangements of active-site residues and repositioning of the substrate were critical early in the evolutionary trajectory, but that changes in conformational sampling of sidechains were also required for the emergence and optimisation of the new function. The results from these two studies mirror findings from directed evolution experiments and computational enzyme design, which demonstrate that the incorporation of catalytic residues into the active site of a protein is required but often not sufficient for efficient catalysis, and that fine-tuning of dynamics through the accumulation of remote mutations plays an important role in the emergence and improvement of new function (Campbell et al., 2016; Kaltenbach and Tokuriki, 2014; Zanghellini, 2014). The emergence of computational protein design approaches that take conformational dynamics into account (e.g. (Davey et al., 2017)) will certainly go a long way in helping the de novo design of efficient protein catalysts. However, due to the complex, non-linear relationship between the sequence of a protein and its conformational landscape, most computational enzyme design approaches will probably still rely (for some time at least) on high-throughput methods or directed evolution to uncover combinations of remote mutations that are required for the optimisation of catalysis.

7.2.2 Future directions

There are several ways that future studies could extend the work presented in Chapter 3. In the second paper (Chapter 3.3), we highlighted three key areas that differ between AncCDT-5 and PaCDT and proposed that mutations in these regions were likely important for shifting the conformational dynamics away from the catalytically-incompatible wide-open state along the historical pathway the PaCDT. Future studies could use site directed mutagenesis or directed evolution to explore the effect of individual mutations on the open/closed equilibrium of these proteins, and the role that epistasis played in shaping the change in dynamics and function (since it is likely that the significant change in dynamics that was observed has resulted from the accumulation of multiple permissive or complementary mutations). One obvious thing to do would be to truncate the C-terminal extension found in PaCDT and measure the effect this has on both the prephenate dehydratase activity and open/closed dynamics (using DEER or a single molecule technique) of this protein. The C-terminal extension is only found in some of the extant CDTs, so it is not necessary for CDT activity, but
it does make several additional inter-domain contacts in PaCDT which likely influence the open/closed dynamics of this protein. Interestingly, there are many examples of SBPs with extensions, loops or additional domains that create similar bridging interactions between the two domains (Bucher et al., 2011; Scheepers et al., 2016); it seems this may be a common mechanism by which the function of these proteins can be modified, while still maintaining the core SBP fold.

One other area that could be investigated further is the function of AncCDT-2 and the related proteins in clade II (such as Pu1068). Genomic context analysis suggested that these proteins are SBPs that possibly bind a ligand related to porphobilinogen (Dan Tawfik, personal communication, 2017), but we are yet to determine what molecule these proteins are binding despite extensive screening by DSF. It may be that these proteins require experimental conditions (e.g. pH, cofactors) different to those tested here to bind their ligand, or they bind some obscure ligand that we did not test. Considering that Pu1068 originates from *Pelagibacter ubique*, which has the shortest genome of any known living organism, these proteins are likely to have an important role in the cell, and may have a function that is of biotechnological interest. Future projects could focus on characterising what these proteins bind.

Altogether, the work presented in Chapter 3 highlighted the power of combining evolutionary analysis, structural characterisation and functional studies to provide insight into the biophysical determinants that underly protein function.
7.3 Investigating the properties of antibodies that bind PfCSP\textsubscript{Repeat} and implications for vaccine design

7.3.1 Parallels between the affinity maturation process and the evolution of new protein function by natural selection

Analogous to the evolution of protein function by natural selection (as discussed in Chapter 2 and 3, for example), the adaptive immune system relies on an iterative process of genetic diversification (through the combination of different light and heavy chains, the imprecise joining of these fragments, and the accumulation of mutations during somatic hypermutation, SHM) followed by selection to produce B cells and antibodies (Abs) that bind antigens with increasing affinity (Mesin et al., 2016). However, unlike traditional evolutionary processes, which require multiple generations of an organism for noticeable differences in function to develop, the humoral affinity maturation process can improve the affinity of circulating Abs 10- to 5,000-fold in a matter of days (Mishra and Mariuzza, 2018). Because of the immune system’s remarkable ability to produce high-affinity complementary binding surfaces in such a short period of time, studying Ab-antigen binding, as well as the germline lineages and affinity maturation pathways that lead to high-affinity binders, is an excellent way to gain insight into the rules that govern macromolecular recognition. Of course, understanding the molecular basis for Ab-antigen interactions, and how this affects the size, dynamics, composition and protective qualities of the immune response, is also central to the rational design of immunogens that can be used in vaccines.

In Chapter 4, I presented work that explored the binding of Abs and Ab antigen-binding fragments (Fabs) to the central repeat region of \textit{Plasmodium falciparum} circumsporozoite protein (PfCSP\textsubscript{Repeat}). In the following sections, I describe recent developments in this field, discuss implications that these findings have for the rational design of improved immunogens based on PfCSP, and highlight several challenges that need to be addressed by future studies.
7.3.2 Polyvalent binding of Fabs to PfCSP<sub>Repeat</sub>

In Chapter 4.2, we proposed that multiple copies of anti-PfCSP<sub>Repeat</sub> Fabs from the gold standard monoclonal Ab (mAb), 2A10, would be able to bind along the length of PfCSP<sub>Repeat</sub> if organised in a pseudo-helical arrangement around the repeat peptide (Figure 7-2a) (Fisher et al., 2017). This model was based on results obtained from ITC, crystallography, computational modelling and site-directed mutagenesis studies. Our model placed neighbouring Fabs within hydrogen bonding distance of each other, which was consistent with our ITC results that suggested avidity effects strengthened the 2A10-CSP complex.

Soon after our paper was published, a Cryo-EM structure of a truncated CSP molecule (rsCSP, (NPDP)(NANPNVDP)<sub>n</sub>(NANP)<sub>16</sub>) in complex with a human anti-PfCSP<sub>Repeat</sub> Fab, Fab311, was published (Figure 7-2b) (Oyen et al., 2018). This structure showed a number of similarities to the 2A10-rCSP model proposed in our paper. In particular, the Cryo-EM structure confirmed that multiple copies of anti-PfCSP<sub>Repeat</sub> Fabs could seamlessly bind along the length of rsPfCSP<sub>Repeat</sub>, with the Fabs radiating out from a central helical conformation of rsPfCSP<sub>Repeat</sub>. The Cryo-EM structure showed that Fab311 bound to an epitope 8 amino acids long (Oyen et al., 2018); we had predicted, primarily from ITC data, that Fab2A10 recognised an epitope 8 – 12 amino acids long (Fisher et al., 2017). Furthermore, similar to what was predicted in our work, Fab-Fab interactions were shown to be important for stabilising the Fab311-rsCSP complex, with key mutations accumulating at these interfaces during SHM (Oyen et al., 2018).

Despite the similarities between these two models, there are several noticeable differences. For example, the 2A10 model presented in Chapter 4 was based on the assumption that Fabs of 2A10 would bind a narrow (~7 Å diameter) helix of (NANP)<sub>n</sub> repeats formed by consecutive type 1 β-like turns; this peptide conformation had been observed in computational modelling and MD simulations of the peptide. In contrast, in the Cryo-EM structure of rsPfCSP-Fab311 the PfCSP<sub>Repeat</sub> sequence forms a much wider helix (~27 Å) that is formed by consecutive type I β-turns and Asn pseudo 3<sub>10</sub> turns (Figure 7-2c). A crystal structure of Fab311 in complex with a PfCSP<sub>Repeat</sub>-derived peptide also showed it binding a type I β-turn followed by a Asn pseudo 3<sub>10</sub> turn (Oyen et al., 2017), consistent with the mode of binding observed by Cryo-
EM. The binding sites of Fab311 and Fab2A10 are quite different (esp. in the sequence, length and conformation of CDRH3 and CDRL3) (Figure 7-3a), and the proposed mode of binding of \((\text{NANP})_n\) repeats to Fab2A10 (modelled through computational docking and MD simulations) is very different to that observed in structure of Fab311. A crystal structure of Fab2A10 in complex with a PfCSP\textsubscript{Repeat} peptide would certainly be useful in confirming the way in which mAb2A10, which remains the gold standard of anti-PfCSP\textsubscript{Repeat} mAbs, binds PfCSP\textsubscript{Repeat}.

**Figure 7-2. Multi-valent binding of Fabs to PfCSP\textsubscript{Repeat} confirmed by Cryo-EM.** (a) The model proposed by us in Chapter 4, showing the binding of multiple copies of Fab2A10 along the length of a narrow PfCSP\textsubscript{Repeat} peptide helix. (b) The model of Fab311-rsPfCSP obtained by docking the variable regions of the crystal structure of Fab311 (colored surfaces) into a Cryo-EM map of the complex (PDB 6MHG). Copies of Fab311 bind seamlessly along the length of rsCSP and inter-Fab contacts are stabilised by mutations accumulated by SHM. Figure adapted from (Oyen et al., 2018). (c) Stripping back the bound Fab311 molecules reveals the pseudo helical conformation of rsCSP that is bound in this complex, which forms consecutive type I \(\beta\)- and Asn pseudo 3\textsubscript{10} turns along its length.
7.3.3 Diverse paratopes and modes of binding revealed by recent studies of anti-PfCSP\textsubscript{Repeat} Fabs

While the work on Fab311 provided experimental evidence of one way that high avidity complexes between anti-PfCSP\textsubscript{Repeat} Fabs and PfCSP can form (and shared several characteristics with our 2A10-CSP model), it has become clear that not all anti-PfCSP\textsubscript{Repeat} Fabs recognise PfCSP\textsubscript{Repeat} in the same way. Since our publication, over 20 crystal structures of human anti-PfCSP\textsubscript{Repeat} Fabs in complex with CSP-derived peptides have been reported (Imkeller et al., 2018; Kisalu et al., 2018; Murugan et al., 2019; Oyen et al., 2017, 2018; Pholcharee et al., 2020; Tan et al., 2018; Triller et al., 2017), highlighting differences in the binding sites of anti-PfCSP\textsubscript{Repeat} Fabs and the diverse ways that Fabs can bind PfCSP\textsubscript{Repeat} (Figure 7-3c). Although the majority of the human anti-PfCSP\textsubscript{Repeat} Fabs characterised to date recognise an NPNA motif in the form of a type I \(\beta\)-turn at the core of their paratopes, different combinations of light and heavy chains and/or the accumulation of mutations through SHM in the flanking regions of the paratope lead to anti-PfCSP\textsubscript{Repeat} Fabs that have varying levels of cross reactivity for distinct epitopes on PfCSP (including sequences that contain NANP-like motifs, such as NVDP or NPDP) (Murugan et al., 2019), have larger epitopes (e.g. 12 amino acids recognised by Fab317 rather than the 8 recognised by Fab311 (Oyen et al., 2017, 2018)) and recognise an array of PfCSP\textsubscript{Repeat} conformations (although predominantly combinations of type I \(\beta\)-turns and Asn pseudo 3\textsubscript{10} turns) (Pholcharee et al., 2020).

Importantly, variations in the size, conformation and sequence of the epitopes that a Fab binds will lead to differences in the overall stoichiometry and architecture of the Fab-PfCSP complex. For example, the seamless, tight packing of Fab311 along the length of rsCSP is dependent on the ability of Fab311 to (i) form Fab-Fab interactions (reversion of residues at Fab-Fab interfaces to their germline-encoded sequences resulted in the loss of the tightly packed, spiral-like complex), (ii) bind alternating type I \(\beta\)- and Asn pseudo 3\textsubscript{10} turns (this creates a rsPfCSP\textsubscript{Repeat} helix that is the right size to accommodate the surrounding Fabs) and (iii) bind NVDP-containing peptides (allowing complete coverage of the rsPfCSP\textsubscript{Repeat} region) (Oyen et al., 2018). In contrast, the heterogenous class averages and non-seamless binding that is observed in nsEM of rsCSP-Fab317 (Oyen et al., 2018) and rsCSP-Fab397 (Pholcharee et al., 2020) complexes can likely be explained by the inability of these Fabs to bind NVDP-
Conclusions, implications and future directions

containing sequences in a seamless manner or due to steric clashes between neighbouring Fabs.

Figure 7-3. Diversity amongst anti-PfCSP<sub>Repeat</sub> Fabs. The binding sites of Fabs that bind PfCSP<sub>Repeat</sub> can vary considerably and determine the way they recognise epitopes on PfCSP<sub>Repeat</sub>. For example, comparison of crystal structures of (a) Fab2A10 (PDB 5SZF) and (b) Fab311 (PDB 6AXK) (Oyen et al., 2017) highlight clear differences in the lengths and conformations of CDRH3 (orange) and CDRL3 (green). The (NPNA)₃ peptide is shown as cyan sticks. (c) A selection of some of the human anti-PfCSP<sub>Repeat</sub> Fabs that have been structurally characterised, showing different modes of binding an (NANP)<sub>n</sub> repeat peptide (cyan). Most paratopes include a germline-encoded Trp or Arg residue (highlighted in red) that is key to the recognition of a core NPNA motif. PDB accession codes, germline lineages, and references are noted for each structure.
7.3.4 Connecting binding modes with protection levels

Although recent studies have certainly highlighted clear differences in the way that distinct FabS recognize PfCSP\textsubscript{repeat}, it remains unclear how these different binding properties impact the sporozoite-neutralizing activity of the associated Abs; this is largely because the mechanisms by which anti-PfCSP\textsubscript{repeat} Abs confer protection are not well understood. For example, while some anti-PfCSP\textsubscript{repeat} FabS have been shown to impair sporozoite motility at the dermal site of inoculation, thereby decreasing the number of sporozoites that enter the blood stream and reach the liver, the primary mechanism by which they do so is debated (Aliprandini et al., 2018; Bejon et al., 2005; Flores-Garcia et al., 2018; Mishra et al., 2012; White et al., 2013). Similarly, while some anti-PfCSP\textsubscript{repeat} Abs can hinder the CSP-mediated invasion of hepatocytes (Rodríguez-Galán et al., 2017; Thai et al., 2020; Tweedell et al., 2019), it is unclear how differences in the way that Abs bind to PfCSP\textsubscript{repeat} correlate with differences in their ability to block the sporozoite invasion of hepatocytes. Nor is it known whether full coverage of PfCSP\textsubscript{repeat} by Abs is required for neutralisation; and if it is, can competition between Abs with overlapping epitopes disrupt the complete masking of PfCSP\textsubscript{repeat} and impair the protective properties of these Abs? Future studies that focus on elucidating the molecular mechanisms by which anti-PfCSP\textsubscript{repeat} Abs confer neutralising protection (and factors that may disrupt this, such as competition between Abs) will certainly help structure-based vaccine design.

7.3.5 The immunodominance of PfCSP\textsubscript{repeat}: an immune smokescreen?

In Chapter 4 we introduced the idea that PfCSP repeats may have evolved as an immune decoy and that the dominant anti-PfCSP\textsubscript{repeat} humoral immune response could be hindering the production of a long-lasting response against more protective epitopes on PfCSP. This idea is not new; the PfCSP\textsubscript{repeat} “smokescreen effect” was proposed in the 1980s (Anders, 1986; Burkot et al., 1991). Indeed, the presentation of immunodominant motifs (especially tandem repeats) on cell surfaces is known to be a common mechanism used by pathogens to evade the hosts’ immune response, and removing immunodominant regions from immunogens has been important to the development of effective anti-viral vaccines that elicit immune responses against more protective, cryptic epitopes (Valkenburg et al., 2016).
There are several properties of PfCSP\textsubscript{Repeat} that make it particularly immunodominant, supporting the idea that it evolved as an immune smokescreen. The most obvious property is the high copy number and accessibility of the repeat epitopes. Second is the fact that a wide range of naïve B cells are particularly predisposed for recognition of PfCSP's NANP repeats (Fisher et al., 2017; Murugan et al., 2018; Triller et al., 2017). Conversely, it could be said that NANP repeats have properties that make them predisposed to binding a range of precursor BCRs found in people with diverse genetic backgrounds. The immune response to PfCSP\textsubscript{Repeat} is dominated by Abs encoded by VH3-33/VH3-30 germline genes (Murugan et al., 2019), which encode a Trp residue at the core of the Fab's binding site that facilitates the recognition of NPNA repeats via interactions with the Pro and Asn in the CSP peptide (e.g. Figure 7-3b). These central Trp residues are often complemented by neighbouring germline-encoded aromatic residues or disulfide bonds that pack against the Pro of the NPNA repeat (Murugan et al., 2019). Consequently, many VH3-30/VH3-33-encoded germline Abs (gAbs) already have a moderate affinity for NPNA motifs and require only a few somatic mutations to improve their affinity to levels associated with protection (Murugan et al., 2019; Triller et al., 2017), which explains why they are so dominant in the immune repertoire against PfCSP. Other common combinations of heavy/light chains can also lead to precursor Abs with a central Trp, a high proportion of aromatic CDR residues and reasonable affinity for NPNA repeats. Furthermore, Fabs in which the central Trp is replaced with an Arg are also found in the repertoire of anti-PfCSP\textsubscript{Repeat} Abs (e.g. Fab4493, IGHV3-49 (Murugan et al., 2019)), although these are more rare. Indeed, the flexibility and high antigenic complexity of PfCSP\textsubscript{Repeat}, as well as its ability to transiently sample various secondary structure motifs, likely facilitates binding to a range of precursor B cells and negatively impacts the efficient affinity maturation of B cells towards any one specific epitope (Murugan et al., 2018; Wardemann and Murugan, 2018). Interestingly, the repeat regions found in CSP from other \textit{Plasmodium} species share similar properties with PfCSP\textsubscript{Repeat} in that they are predominantly made up of a combination of low side-chain entropy residues (glycine, proline, alanine) and residues that can hydrogen bond to the main chain (glutamine, asparagine, aspartate); for example, the central repeats of \textit{P. vivax} contain repeat GNQPGANGA or GQPAGDRA(A/D) motifs. CSP from some strains of \textit{Plasmodium} also contain additional repeats in its N-terminal domain that share these properties; these regions may also potentially act as immune smokescreens in addition to the central repeat region, but this remains to be tested.
Immunodominant immune responses, such as that against PfCSP\textsubscript{Repeat}, can hinder the production of immune responses to other protective (typically more cryptic) epitopes by sterically blocking the more protective epitopes (either direct competition between B cells or via antibody feedback mechanisms) (McNamara et al., 2019; Uthaipibull et al., 2001). Alternatively, immunodominant responses can hinder responses to subdominant epitopes (on either the same protein or other proteins) by competing for immunological space during the Ab maturation phases (Chaudhury et al., 2014). Unfortunately, it is difficult to say whether or not the immune response to PfCSP\textsubscript{Repeat} is blocking more protective responses without first identifying superior protective epitopes. Regions that are conserved across Plasmodium species on the N- and C-terminal that are known to have physiological roles (RI cleavage site, RII\textsuperscript{+} etc) seem obvious candidates, but Abs that directly target these regions have not been characterised. The dearth of Abs that target these ‘silent epitopes’ (Patarroyo et al., 2015) suggests that these epitopes have evolved to evade the immune response, and are likely protective. While studies often make the general claim that Abs that target the C- or N-terminal regions are not protective, this idea is based on a few studies that have characterised mAbs that target regions on these domains that are not associated with any known function (Scally et al., 2018; Thai et al., 2020). Future studies could focus on testing if targeting the “silent” epitopes on the terminal domains of PfCSP provides better protection than anti-PfCSP\textsubscript{Repeat} Abs.

One site that has gained recent attention is the ‘junctional’ region between the N-terminal domain and PfCSP\textsubscript{Repeats} (containing the conserved NANP-like NPDP motif). Rare (mostly non-VH3-30/33 encoded) anti-PfCSP\textsubscript{Repeat} Abs that are cross reactive with junctional peptides tend to provide better protection than pure NANP binders, possibly because they block the nearby RI cleavage site (Kisalu et al., 2018; Tan et al., 2018). One (yet to be explored) idea is that the more immunodominant anti-PfCSP\textsubscript{Repeat} responses (such as high avidity NVDP binders that require few SHMs to develop) get in the way of Abs that bind to this potentially protective site (and that typically require more extensive SHM to develop (Murugan et al., 2019)). Indeed, the conserved presence of alternating NANP/NVDP motifs across all sequenced strains of P. falciparum is likely to bias selection towards NVDP binders, which may block access to the nearby conserved junctional motif, NPDP (Figure 7-4). The lack of NVDP motifs (rather than shortening of the repeat region) may actually be the reason why immunisation
with CSP9 produces higher levels of protection in mice (Chapter 4.3) – we are awaiting results from this experiment. However, a recent study also concluded that reducing the number of NANP repeats in an immunogen can lead to a more protective immune response (Langowski et al., 2020).

**Figure 7-4. NVDP motifs act as anchors in PfCSP.** The alternating NVDP/NANP motifs toward the N-terminal end of PfCSP\textsubscript{Repeat} likely act as anchors to lock Fab binding into specific frameshift positions. This likely explains why some anti-PfCSP\textsubscript{Repeat} Fabs can bind seamlessly along the length of PfCSP\textsubscript{Repeat}, while others are observed to bind in a non-seamless manner. The presence of the NVDP motifs could be involved in an immune smokescreen mechanism, directing the immune response away from epitopes that are potentially more protective, such as the NANP-like junctional DGNP/NPDP motifs. The role of NVDP motifs that are present in the middle of PfCSP\textsubscript{Repeat} (not shown) in some strains is unclear, but may either help or hinder the formation of high avidity complexes.

### 7.3.6 Summary and implications for vaccine design

Together, our work and the studies reviewed above highlight how the ontological, structural and biophysical properties of a Fab shape the way that it recognises epitopes on PfCSP, controls the overall stoichiometry and architecture of the immune complex, contributes to the neutralising activity of Abs, and is key to whether or not it persists a part of the long-term immune repertoire. Of course, the improvement of PfCSP-based vaccines will benefit from a number of diverse approaches, including the optimisation of immunisation protocols (which can also direct the immune response towards more subdominant epitopes) (Cirelli et al., 2019; Regules et al., 2016), the development of new adjuvants (Cawfield et al., 2019; Genito et al., 2017), and the multivalent display of antigens on the surface of virus-like particles (Chan et al., 2019; Collins et al., 2017; Urakami et al., 2017) or tobacco mosaic virus disks (Langowski et al., 2020). However, studies that focus on resolving the physiological structure...
Conclusions, implications and future directions

of PfCSP, identifying of the most protective epitopes on PfCSP (both in sequence and structure), determining the mechanisms by which anti-PfCSP<sub>Repeat</sub> Abs confer neutralising protection, and investigating competition between Abs will certainly go a long way in guiding the design of improved immunogens for use in the next-generation of malaria vaccines. Studies that provide insight into how Abs bind PfCSP (such as those presented in Chapter 4), combined with a better understanding of the protective epitopes on PfCSP, will be key to the structure-guided design of next-generation malaria vaccines.

7.4 Nonribosomal peptide synthesis machinery: applications and evolution

In Chapter 5 I presented two papers that explored the structure and function of nonribosomal peptide synthesis (NRPS) enzymes. Natural NRPS pathways produce diverse peptide products (including cyclised, branched, glycosylated and unnatural amino acid-containing peptides) through the recruitment of different combinations of enzyme modules, each of which (normally) catalyse a single step in the pathway. Because of this, natural NRPS systems are a gold-mine of enzymes that can be used for the low-cost production or functionalisation of peptide-based bioactive compounds. For example, individual enzymes from NRPS complexes have been used in vitro to make specific modifications to synthetically produced peptides (Tailhades et al., 2018). Therefore, the functional characterisation of diverse NRPS enzymes (especially atypical enzymes, such as the unusual C-terminal domain of Ebony presented in Chapter 5.2) is essential for expanding the repertoire of modifications that can be made to peptides via chemoenzymatic approaches. Further, understanding the structural and biophysical basis for the activity of these enzymes (e.g. through crystallography and ITC, Chapter 5.2) will help with future efforts to optimise, modify or target these functions.

The modules of NRPS complexes and polyketide synthases can also be mixed-and-matched to create complete biosynthetic pathways for the production of a desired peptide (e.g. via in vivo strain manipulation) (Klaus et al., 2019). However, such efforts are often hampered by the loss of intimate domain interactions between domains that are required for the precise control of chain transfer and elongation reactions (Awakawa et al., 2018). In Chapter 5.3 we
demonstrated that combining ITC and crystallography is a powerful approach for understanding interactions between NRPS proteins. A similar approach could be used to investigate the interactions between other NRPS modules. The information gained from these kinds of experiments, especially when combined with evolution-based approaches (Chevrette et al., 2020), will certainly aid efforts to optimise the interactions between domains through protein engineering.
7.5 Concluding remarks

As much as anything else, the work presented in this thesis highlights the complexities and persisting gaps in our understanding of the relationship between the sequence, structure, dynamics, and function of proteins. The large, complex, multidimensional free energy landscapes of even the simplest protein-solvent and protein-ligand-solvent thermodynamic systems are impossible to completely characterise, making it difficult to accurately predict the detailed properties of molecular recognitions events. Further, epistasis makes the effect of mutations largely unpredictable and leads to a complex, rugged fitness landscape. Advances in computational power and high-throughput techniques will certainly improve our ability to explore these large and complex landscapes, and will assist efforts to computationally design protein dynamics and novel function.

However, as exemplified by the work presented here, there is still much to gain from focusing on characterising and comparing smaller, functionally-relevant areas of these landscapes. X-ray crystal structures can capture different ground states of proteins and protein-ligand complexes, highlighting residues important for function and structural differences between related proteins. Molecular dynamics simulations and experimental techniques (such as DEER) provide an idea of the overall dynamic properties of the systems. In addition, biophysical techniques such as ITC highlight the dominant thermodynamic drivers of protein binding and function. These techniques can be effectively complemented by using evolution-based approaches, such as ancestral sequence reconstruction or the analysis of the affinity maturation of antibodies, to identify important functional intermediates and function-shifting mutations. Such studies give us a better appreciation of the complex relationship between protein sequence, structure, dynamics and function, and are essential for guiding efforts to rationally design, or modify, protein function.
8 References

Citation convention: *Cell*


Chan, J.-A., Wetzel, D., Reiling, L., Miura, K., Drew, D.R., Gilson, P.R., Anderson, D.A.,


Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. Berichte Der


antibody binding involves specific hydrogen bonds rather than nonspecific hydrophobic stickiness. Protein Sci. 12, 2183–2193.


References


1–15.


References


