

Molecular Dynamics for Synthetic Biology

Joshua Ayden Mitchell
July 2020


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



© Copyright by Joshua Ayden Mitchell 2020
All Rights Reserved

Author's declaration

I declare that the research presented in this Thesis represents original work that I carried out during my candidature at the Australian National University, except for contributions to multi-author papers incorporated in the Thesis. These papers are enumerated in table 1, below, and detailed Statements of Contribution are given in the chapters in which they appear.

Signed: 
Joshua A. Mitchell
Candidate
July 2020


Signed: 
Professor Colin J. Jackson
On behalf of all collaborating authors.
July 2020

Table 1: Publications included in this thesis.

Page	Title	Authors	Publication venue and status	Contribution by JAM
37	Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecular biosensors	J. A. Kaczmarek,* J. A. Mitchell ,* M. A. Spence,* V. Vongsouthi,* C. J. Jackson	Current Opinion in Structural Biology (in press)	~25%
52	Rangefinder: A semisynthetic FRET sensor design algorithm	J. A. Mitchell ,* J. H. Whitfield,* W. H. Zhang,* C. Henneberger, H. Janovjak, M. L. O'Mara, C. J. Jackson	ACS Sensors (in press)	~33%
76	Method for developing optical sensors using a synthetic dye-fluorescent protein FRET pair and computational modeling and assessment	J. A. Mitchell , W. H. Zhang, M. K. Herde, C. Henneberger, H. Janovjak, M. L. O'Mara, C. J. Jackson	Methods in Molecular Biology (in press)	~70%
88	Monitoring hippocampal glycine with the computationally designed optical sensor GlyFS	W. H. Zhang,* M. K. Herde,* J. A. Mitchell , J. H. Whitfield, A. B. Wulff, V. Vongsouthi, I. Sanchez-Romero, P. E. Gulakova, D. Minge, B. Breithausen, S. Schoch, H. Janovjak, C. J. Jackson, C. Henneberger	Nature Chemical Biology (in press)	~10%
129	A computationally designed fluorescent biosensor for D-serine	V. Vongsouthi,* J. H. Whitfield,* P. Unichenko, J. A. Mitchell , B. Breithausen, O. Khersonsky, L. Kremers, H. Janovjak, H. Monai, H. Hirase, S. J. Fleishman, C. Henneberger, C. J. Jackson	BioRxiv (preprint)	~5%

Page	Title	Authors	Publication venue and status	Contribution by JAM
165	T-dependent B cell responses to <i>Plasmodium</i> induce antibodies that form a high-avidity complex with the circumsporozoite protein	C. R. Fisher,* H. J. Sutton,* J. A. Kaczmarek, H. A. McNamara, B. Clifton, J. A. Mitchell , Y. Cai, J. N. Dups, N. J. D'Arcy, M. Singh, A. Chuah, T. S. Peat, C. J. Jackson, I. A. Cockburn	PLOS Pathogens (in press)	~5%

* These authors contributed equally.

Acknowledgements

This thesis was written in Ngunnawal country. I would like to recognise their continuing connection to land, waters and culture and pay my respects to their Elders, past and present. I extend that special respect to Aboriginal and Torres Strait Islander peoples who read this thesis or who contributed to it in ways to which I am blind. Sovereignty was never ceded.

The course of my PhD has taken me to two research groups, three homes, and four countries. I have made life-long friends and learnt more about myself than I thought there was to know.

I am exceedingly grateful to my supervisor, Professor Colin Jackson. Colin has always supported me and allowed me to make my own choices in my research, even to his detriment. He has trusted me to spend outrageous amounts of lab money on computers and been there for me when everything was too much, and even had my back in conflict. I don't think I've cried in his office but I've come close. I have thoroughly enjoyed his guidance, sarcasm, humour, and baking stories. May you never run out of pu'er.

I want to thank my partner Madelaine Jade Blackmore-Warren. I met Maddie during the course of my PhD and now I can't imagine life without them. They believe in me when I can't, listen happily to my excited ranting about obscure computational details, and have taught me how amazing being best friends with your partner can be. I hope I make you enough tea.

Thank you also to my former (and hopefully future!) colleagues Karmen Čondić-Jurkić, Jason Whitfield, Eleanor Campbell and Michael 'Pickle' Thomas. Each of you has been there for me when I've needed you many times over. Thank you also to each of you for reading and commenting on drafts of this thesis. Your support has been invaluable and I am fortunate to call you all friends.

Thank you Tim Crundall for being a great house mate and a close friend. I'll never forgive you for moving to Germany but I'm glad you're happy.

Thank you to Vanessa, Isabel, Joe, Lynn, Spence, James, Margot, Brendon, Galen, Davis, Elena, Ben, Junming, Mahakaran, Cass, Jake, Suriya, Elaaf, Hafna, and the rest of the Jackson group, past and present. For some reason you all seem to like me and I'm grateful for the atmosphere and friendship I've been able to rely on at the Jackson group. I hope the group continues to be the supportive, friendly place I know.

Thanks to Alex, Hugo, Nick, Heather, Lily, Simon, Andrew, Amanda, Giuseppe, Lauren,

ACKNOWLEDGEMENTS

Teresa, Rika, Mitch, Ben, and Peter for being great colleagues in the computational world.

Thanks to the authors of GROMACS, for providing a great, beautifully documented, open source MD engine. I got the reference manual bound at Officeworks and it is the only book that I've unpacked since moving. Thanks to Officeworks for saving my butt at least three times. And thanks to the Rust language community for allowing me to believe I could code systems software and inspiring me to contribute to Open Source.

Thanks to my parents, Leigh and Wayne, for having my back and loving me when everything fell apart, and for supporting me financially since. Thanks Dad for your proof-reading — I incorporated most of your edits. Thanks to my brother Joel and his partner Liz, Maddie's sister Sage and their nan Nora, who has graciously hosted us in her home the last few months, and my extended family Nana, Granddad, Cool Aunt Jen, John, Tash, and Amber. Gran and Pop, I miss you.

Thanks to the people of Canberra for electing six Greens to the Assembly.

Finally, thank you to my eternal high-school friends Lauren, Jesse, Dzifa, Faiyaz, Ben, and Kerry. Thanks for being there for me the past five years. See you at the tree.

Abstract

Synthetic biology is the field concerned with the design, engineering, and construction of organisms and biomolecules. Biomolecules such as proteins are nature's nano-bots, and provide both a shortcut to the construction of nano-scale tools and insight into the design of abiotic nanotechnology. A fundamental technique in protein engineering is protein fusion, the concatenation of two proteins so that they form domains of a new protein. The resulting fusion protein generally retains both functions, especially when a linker sequence is introduced between the two domains to allow them to fold independently. Fusion proteins can have features absent from all of their components; for example, FRET biosensors are fusion proteins of two fluorescent proteins with a binding domain. When the binding domain forms a complex with a ligand, its dynamics translate the concentration of the ligand to the ratio of fluorescence intensities via FRET.

Despite these successes, protein engineering remains laborious and expensive. Computer modelling has the potential to improve the situation by enabling some design work to occur virtually. Synthetic biologists commonly use fast, heuristic structure prediction tools like ROSETTA, I-TASSER and FoldX, despite their inaccuracy. By contrast, molecular dynamics with modern force fields has proven itself accurate, but sampling sufficiently to solve problems accurately and quickly enough to be relevant to experimenters remains challenging.

In this thesis, I introduce molecular dynamics to a structural biology audience, and discuss the challenges and theory behind the technique. With this knowledge, I introduce synthetic biology through a review of fluorescent sensors. I then develop a simple computational tool, Rangefinder, for the design of one variety of these sensors, and demonstrate its ability to predict sensor performance experimentally. I demonstrate the importance of the choice of linker with yet another sensor whose performance depends critically thereon. In chapter 6, I investigate the structure of a conserved, repeating linker sequence connecting two domains of the malaria circumsporozoite protein. Finally, I develop a multi-scale enhanced sampling molecular dynamics approach to predicting the structure and dynamics of fusion proteins. It is my hope that this work contributes to the structural biology community's understanding of molecular dynamics and inspires new techniques developed for protein engineering.

Table of Contents

Author's declaration	i
Acknowledgements	iv
Abstract	vi
List of Figures	xi
Glossary of Abbreviations	xiii
1 Introduction to Molecular Dynamics	1
1.1 Why Molecular Dynamics?	1
1.1.1 Molecular dynamics can be reliable when used carefully	1
1.1.2 Molecular dynamics complements other structural biology methods	2
1.2 How does computational biophysics work	3
1.2.1 Computers can represent proteins as numbers	3
1.2.2 At equilibrium, we can think about proteins statistically	4
1.2.3 The probability depends on the energy	5
1.2.4 The energy can be estimated from the state	7
1.2.5 Algorithms map out the probability density surface as energy	8
1.2.6 Choice of algorithm depends on scoring function and goals	10
1.3 How does MD work	10
1.3.1 The force field gives potential energy and forces on a single state	11
1.3.2 Dynamics give rise to new states	13
1.3.3 Dynamics produce an approximation of the trajectory	14
1.3.4 Equilibrium ensembles can be inferred from trajectories	16
1.4 Common traps and their solutions	18
1.4.1 Use a high-quality, appropriate force field	18
1.4.2 Check for sufficient sampling	25
1.4.3 Check that the produced trajectories are reasonable	26
1.4.4 Use appropriate compute hardware and software	26

1.4.5	Consider your boundary conditions and box size	27
1.4.6	Use an appropriate thermostat and barostat	29
1.4.7	Enhance sampling	32
1.5	Conclusion	36
2	Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors	37
2.1	Preface	37
2.2	Statement of contribution	39
2.2.1	Publication status	39
2.2.2	Authorship and contribution	39
2.3	Manuscript	40
3	Semisynthetic fluorescent biosensors via Rangefinder	48
3.1	Preface	48
3.1.1	Rangefinder sensors as a modelling testbed	49
3.1.2	The FRET orientation factor	50
3.2	Rangefinder: a semisynthetic FRET sensor design algorithm	52
3.2.1	Statement of contribution	52
3.2.2	Manuscript	53
3.2.3	Supporting information	58
3.3	Method for developing optical sensors using a synthetic dye–fluorescent protein FRET pair and computational modeling and assessment	76
3.3.1	Statement of contribution	76
3.3.2	Manuscript	77
4	Monitoring hippocampal glycine with the computationally designed optical sensor GlyFS	88
4.1	Preface	88
4.1.1	Fluorescence in Synthetic Biology	88
4.1.2	Fluorophore dynamics and linkers	89
4.1.3	Simplified modelling as a guide to intuition	91
4.2	Statement of contribution	92
4.2.1	Publication status	92
4.2.2	Authorship and contribution	92
4.3	Manuscript	93
4.4	Supporting information	104

5	A computationally designed fluorescent biosensor for D-serine	129
5.1	Preface	129
5.2	Statement of contribution	131
5.2.1	Publication status	131
5.2.2	Authorship and contribution	131
5.3	Manuscript	132
5.4	Supporting information	161
6	T-dependent B cell responses to <i>Plasmodium</i> induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein	165
6.1	Preface	165
6.1.1	Antibodies	166
6.1.2	The Circumsporozoite NANP repeat region	167
6.2	Statement of contribution	169
6.2.1	Publication status	169
6.2.2	Authorship and contribution	169
6.3	Manuscript	170
6.4	Supporting information	198
7	Multiscale Molecular Dynamics simulations of fusion proteins	209
7.1	Introduction	209
7.2	Methods	211
7.2.1	Linker simulations	213
7.2.2	Fluorophore parametrisation	214
7.2.3	Mutated structure derivation	215
7.2.4	Protein–Protein simulations	216
7.2.5	Free energy calculation	217
7.3	Results	219
7.3.1	Linker simulations	219
7.3.2	Multiscale technique	222
7.4	Discussion and future work	223
7.4.1	Combining force fields	226
7.5	Conclusion	226
8	Conclusions and future work	228
8.1	Conclusions	228
8.1.1	Rangefinder is a rapid semisynthetic sensor design method	228

TABLE OF CONTENTS

8.1.2	The domains of a fusion protein are highly dynamic	229
8.1.3	Linker length and composition has a large effect on sensor performance	229
8.1.4	The design of sensors for a multitude of target molecules	229
8.1.5	The NANP repeat linker of the <i>Plasmodium</i> circumsporozoite protein produces a multivalent immune response	229
8.1.6	A REST2 regime efficiently samples disordered peptides	230
8.1.7	Linker peptides have characteristic free energy landscapes	230
8.1.8	AWH is well-suited to sampling protein–protein interactions	230
8.1.9	A reweighting method can combine data from different resolutions of simulation	231
8.2	Future work	231
8.2.1	Multi-state modelling of sensors	231
8.2.2	Improvements to fusion protein modelling technique	232
8.2.3	Assess size of error induced by ignoring non-bonded linker–domain interactions	232
8.2.4	Compare fusion protein modelling technique to experiment	232
References		233

List of Figures

1.1	The synthetic ten amino acid mini-protein Chignolin at different levels of coarse graining.	3
1.2	The probability and free energy surfaces of Chignolin along a simple one-dimensional folding coordinate.	6
1.3	The funnel-shaped folding landscape of Chignolin in two dimensions.	9
1.4	The CHARMM and Amber protein force field family trees.	20
1.5	Periodic boundary conditions with cubic and rhombic dodecahedral boxes. . .	28
2.1	Graphical abstract for <i>Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors</i>	38
4.1	The Green Fluorescent Protein.	88
4.2	FRET efficiency plotted as a function of inter-fluorophore distance.	90
6.1	An antibody.	167
6.2	Theoretical and experimental CD spectra of the 6 peptide.	193
6.3	Cluster analysis for MD simulations of 6 peptide.	193
6.4	Molecular dynamics simulation of the 6:Fab complex.	194
6.5	The B cell response to CSP has a T-independent component.	194
6.6	Alignment of 2A10 heavy chain and the predicted germline sequence.	195
6.7	Alignment of 2A10 light chain and the predicted germline sequence.	196
6.8	Molecular dynamics simulation of the solution structure of the (NANP) ₆ peptide. .	197
6.9	Molecular dynamics simulation of the interaction of the (NANP) _n repeat with the 2A10 Fab.	197
7.1	Mapping of atoms in the ECFP fluorophore to MARTINI beads.	215
7.2	Collective variables used for PMF.	219
7.3	Single repeat linkers in different contexts.	220
7.4	Internal co-ordinates of trimeric linkers	221
7.5	Inter-terminal distance of AlexaFluor 532-labelled ECFP-AYW fusion protein .	222

7.6	FRET efficiency of ECFP-AYW fusion	223
-----	--	-----

Glossary of Abbreviations

Abbreviations

ABF	Adaptive Biasing Force
ATB	Automated Topology Builder
AWH	Accelerated Weight Histogram
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CG	Coarse-Grained
CGenFF	CHARMM General Force Field
CHARMM	Chemistry at Harvard Molecular Mechanics
CPU	Central Processing Unit
Cryo-EM	Cryo Electron Microscopy
CSVR	Canonical Sampling through Velocity Rescaling thermostat
CV	Collective Variable
ECFP	Enhanced Cyan Fluorescent Protein
Fab	Fragment antigen-binding
Fc	Fragment crystallisable
FP	Fluorescent Protein
FRET	Förster Resonance Energy Transfer
GAFF	General Amber Force Field
GFP	Green Fluorescent Protein family
GPU	Graphical Processing Unit
GROMOS	Groningen Molecular Simulation
H-REMD	Hamiltonian Replica Exchange Molecular Dynamics
LJ	Lennard-Jones
MD	Molecular Dynamics
MM	Molecular Mechanics
MTTK	Martyna-Tuckerman-Tobias-Klein barostat
NMR	Nuclear Magnetic Resonance
NPT	Constant Number of particles, Pressure and Temperature (isobaric-isothermal en-

	semble)
NVT	Constant Number of particles, Volume and Temperature (canonical ensemble)
NVE	Constant Number of particles, Volume and Energy (microcanonical ensemble)
OpenMM	Open Molecular Mechanics
OPLS	Optimised Potential for Liquid Simulation
PBP	Periplasmic Binding Protein
PME	Particle Mesh Ewald
PR	Parrinello-Rahman barostat
QM	Quantum Mechanics
QM/MM	Quantum Mechanics/Molecular Mechanics
REMD	Replica Exchange Molecular Dynamics
SBP	Solute Binding Protein
SI	International System of Units
SPC	Simple Point Charge
TIP3P	Transferable Intermolecular Potential (3 point)
TIP4P	Transferable Intermolecular Potential (4 point)
TIP5P	Transferable Intermolecular Potential (5 point)
VES	Variationally Enhanced Sampling
VMD	Visual Molecular Dynamics

GROMACS, MARTINI, and NAMD are proper nouns that are stylised as acronyms throughout this thesis, despite either having abandoned their meaning or never having one in the first place. This is consistent with the wishes of their respective maintainers and widespread use in the literature.

Symbols

The following mathematical symbols are used consistently throughout this thesis. Other symbols are defined where they appear.

n	Number of atoms in a thermodynamic system
Z	The partition function of a thermodynamic system (see equation 1.2)
T	The temperature of a thermodynamic system
t	A time, or a step in a trajectory
m	A mass
P	A probability
R_g	Radius of gyration

S	Entropy
Ω	Density of states
$\log(x)$	The natural logarithm of x (base e) such that $e^{\log(x)} = x$
\mathbf{x}	A vector of generalised coordinates; $3n$ real numbers representing the positions $(x_0, x_1, \dots, x_{3(n-1)})$ in 3D space of n atoms, such that atom i has coordinates $(x_{3i}, x_{3i+1}, x_{3i+2})$
$\dot{\mathbf{x}}$	The first derivative of \mathbf{x} with respect to time; ie, the generalised velocity
$\ddot{\mathbf{x}}$	The second derivative of \mathbf{x} with respect to time; ie, the generalised acceleration
\mathbf{p}	A vector of generalised momenta (see generalised coordinates \mathbf{x} , above)
\mathcal{H}	The Hamiltonian, or total energy. In the microcanonical ensemble (constant NVE) for a non-relativistic system, the Hamiltonian is the sum of the potential and kinetic energies $\mathcal{H} = V(\mathbf{x}) + \sum_i \frac{1}{2} m_i \ \dot{\mathbf{x}}_i\ ^2$
$\mathbf{F}(\mathbf{x})$	The forces on a system of atoms with coordinates \mathbf{x} (see equation 1.6)
$V(\mathbf{x})$	The potential energy of a system of atoms with coordinates \mathbf{x} (see equation 1.6)
\sum_i^n	The sum over all n atoms in the system, indexed by i (similarly for j, k etc.)
$\frac{d}{d\mathbf{x}} f(\mathbf{x})$	The gradient $\nabla f(\mathbf{x})$, that is, the sum of partial derivatives $\frac{d}{d\mathbf{x}} f(\mathbf{x}) = \sum_i^n \hat{\mathbf{x}}_i \frac{\partial}{\partial x_i} f(\mathbf{x})$ where x_i is the i th element in \mathbf{x} and $\hat{\mathbf{x}}_i$ is the unit vector in the direction of x_i .

Units and Constants

Except where noted, SI units are used throughout this thesis.

K_B	The Boltzmann constant, $1.380\,649 \times 10^{-23} \text{ J K}^{-1}$
N_A	Avogadro's constant, $6.022\,140\,76 \times 10^{23} \text{ mol}^{-1}$
e	Euler's number, $\approx 2.718\,281\,828\dots$
π	The circle constant, $\approx 3.141\,592\,65\dots$

Chapter 1

Introduction to Molecular Dynamics

1.1 Why Molecular Dynamics?

The same year that Laplace pre-empted molecular simulation with his description of an intelligence that could predict the future of the universe from the current state of every particle within it (1814), Fraunhofer laid the groundwork for crystallography with his invention of the modern spectroscope. Since the invention of the computer, molecular simulation has developed in stops and starts to fill gaps left by experimental structural biology.

1.1.1 Molecular dynamics can be reliable when used carefully

Molecular dynamics (MD) simulates the dynamics of a molecular system by stepping it through time with Newtonian mechanics and a mathematical model called a force field. Its implementation on low-cost Graphical Processing Units (GPUs)^{1–3} and dedicated hardware⁴ in the last decade has spurred a rapid improvement in the quality of force fields and sampling available to researchers.⁵ Modern forcefields can fold many proteins from a fully extended peptide,^{6,7} occasionally even with implicit solvent.⁸ They have been shown to improve homology models when used with care^{9,10} and are now widely used in model refinement.¹¹ Force fields have produced superior results to semi-empirical QM methods for alanine in water¹² and helped explain how cytoskeletal motors get their sense of direction.¹³ They have been applied to systems as large as the influenza viral envelope¹⁴ and the lipids of the mitochondrion.¹⁵ The commercial OPLS 3 force field can even achieve experimental accuracy for protein-ligand binding energies at a fraction of the financial cost.¹⁶

Despite these successes, results from MD commonly fail to be reproduced, either in experiment^{17,18} or even in simulation.^{19,20} MD requires both that the result in question be correct in the underlying model — the force field and other parameters — and also that sufficient sampling has been done to ensure that the correct equilibrium value has been attained. When

these requirements are not met, MD often simply gives an incorrect result. Reliable force fields are not available for all systems, or even all proteins, and even with recent hardware advances it is still limited to short time-scales and small systems that do not undertake chemical reactions. It is therefore essential to understand the limitations of the technique and ensure that one's approach is appropriate.

1.1.2 Molecular dynamics complements other structural biology methods

Experimental and computational work are complementary in principle and often in practice. Experimental work is 'top-down'; it starts with all the complexity of a phenomenon, and layers are stripped back until its workings are clear. Computational work is 'bottom-up'; it starts with nothing, and model elements are added until it resembles some aspect of the real world. When they meet in the middle, they validate each other, the experiment providing the ground truth and the simulation providing a comprehensive model. Molecular simulation can also provide empirical evidence for a higher-order theory; in tandem with experiment, this allows working theories to be validated from both directions.

This partnership works itself out in different ways depending on the technique. Crystallography yields high-resolution structures of many proteins that make reliable starting points for simulation; any differences in conformation or dynamics between the crystal and dilute solution^{21–23} can be understood via simulation. Cryo-EM can produce moderate-to-low-resolution (2.5–3.5 Å) structures, even of large, dynamic proteins, but lacks temporal resolution.²⁴ By contrast, MD can provide atomic, picosecond-by-picosecond ensembles that describe all the dynamics of the protein, including loop motions that cannot be seen by other structural methods. The comparatively low financial cost of MD makes it remarkably effective at filling in gaps left by experimental methods.³

NMR may be applied to similar problems as MD, being most suitable to very small proteins and peptides. The two techniques are therefore useful as validation for each other. In addition, MD provides atomic descriptions of every atom in the ensemble all the way up to the longest time-scales it can reach, allowing it to fill in gaps left by NMR.²⁵

In practice, molecular simulation has become a true partner to experimental structural biology only in the last several years.^{26,27} Highlights in this vein include the complete description of the conformational landscape of methyltransferase SETD8 by combining kinetically trapped crystallography with molecular dynamics,²⁸ and an investigation of β -lactamase dynamics across time-scales from picoseconds to milliseconds.²⁵

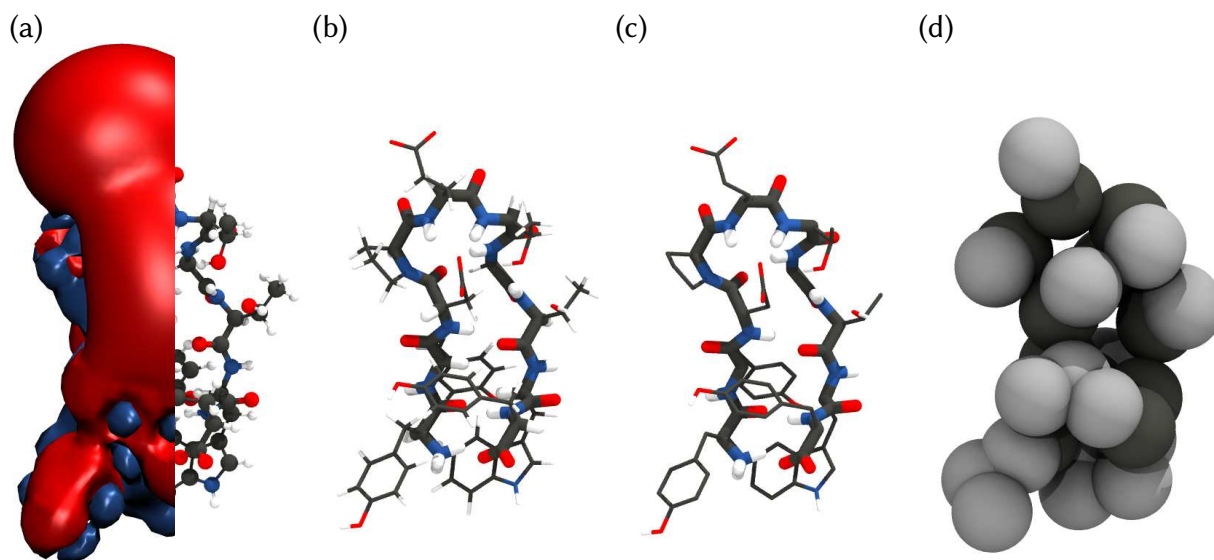


Figure 1.1: The synthetic ten amino acid mini-protein Chignolin²⁹ at different levels of coarse graining. (a) Quantum chemical approaches, which operate on the level of electrons, represented here as the APBS electrostatic surface. (b) All-atom representation. This representation has 166 atoms, each with three-dimensional position and velocity vectors; this implies a 498-dimensional configuration space and a 996-dimensional phase space, though in practice many of these dimensions are constrained or otherwise redundant. (c) United-atom representation. Non-polar hydrogens are represented implicitly within their parent carbon atoms. Polar hydrogens are represented explicitly. This reduces the size of configuration and phase spaces because the number of atoms is reduced. (d) MARTINI representation. Groups of approximately four heavy atoms along with their hydrogens are represented as a single ‘bead’, further reducing the size of configuration and phase space.

1.2 How does computational biophysics work

1.2.1 Computers can represent proteins as numbers

A computer requires a numerical representation of a protein to do any work on it. Most commonly, the protein’s state is recorded as a vector in so-called configuration space representing the positions of each atom, while the interactions between atoms are stored separately in a topology. The configuration space is an abstract vector space that spans all the possible combinations of positions of all the atoms in the system. Techniques that rely on velocities store them alongside the positions in a larger, but conceptually similar, ‘phase space’. Regions or points in these vector spaces correspond to states that the system could adopt.

It is also common to reduce the dimensionality of the system to make it easier to work with. This is called coarse-graining, in contrast to a detailed, fine-grained description (see figure 1.1). Common coarse-graining strategies include treating non-polar hydrogen atoms as part of their parent carbon atom (united-atom representation), treating water as a continuum rather than atoms (implicit solvent), or treating groups of atoms as single interaction sites³⁰ (MARTINI, CG-Rosetta). Indeed, treating the system as atoms can be considered a coarse-graining of

a quantum chemical picture. The ultimate coarse-graining is to describe the entire system with a single variable, and this is commonly done; reaction coordinates, folding coordinates, concentration, temperature, pressure, and fluorescence intensity are all examples of this. In this way, coarse-graining can be thought of as the goal of all computational biophysics.

1.2.2 At equilibrium, we can think about proteins statistically

While we think of proteins as having a defined fold, they are far from static. The path they trace out in configuration space is extremely difficult to track, given the sheer number of configurations open to even a folded protein, and experiments do not give access to this path. Fortunately, chemical systems move so fast, and are so numerous in even a tiny amount of solution, that we can apply the law of large numbers and consider them statistically.

The statistical picture assigns every point in configuration space some probability density that represents how likely one is to find the system in that state. This forms a probability density surface over all of configuration space and gives any finite region of the space a finite probability. The system spends most of its time in regions of high probability, but occasionally traverses a barrier of low probability to move to a new region. These traversals constitute the dynamics of the system, with time-scales determined by the height of the barriers, and when different regions have different properties of interest we consider them different functional states. Many proteins have a small region of such high probability that it spends more time in this confined region than everywhere else put together, and we call that region the folded state. The size and structure of the region describes the dynamics of the folded protein.

Computational biophysics therefore estimates the shape of this probability surface for a particular system under some particular coarse graining and assumptions. The questions it is equipped to answer can always be posed in terms of a probability surface; for example:

- Structure prediction: Assuming there exists a unique high-probability structure, what is it?
- Rigid docking: Which ligand pose is most probable, assuming the protein doesn't move?
- Mutation stability prediction: Which mutation will most improve the probability of this structure relative to some proxy of the unfolded state?
- Normal mode analysis: Assuming this structure is most probable, what other structures could this protein adopt?
- Reaction rate prediction: What energetic and entropic barriers exist to this path through phase space?

This statistical picture requires that the system be at equilibrium. Equilibrium can be broadly thought of as when the system has seen enough states that the law of large numbers applies

and a statistical treatment is valid. At equilibrium, the probability density is well defined in terms of the energy via the Boltzmann distribution, as we will see shortly. There are several equivalent ways to define when a set of points in phase space, in the parlance an ‘ensemble of states’, is at equilibrium:

1. The probability density distribution does not depend on time;
2. the free energy is at a minimum;
3. the entropy is at a maximum;
4. net forward and reverse rates for all processes are equal;
5. there is no net work being done on or by the system.

While MD can in principle be applied out-of-equilibrium, other biophysical techniques generally cannot, and today’s force fields are exclusively parametrised at equilibrium. Fortunately, most properties of interest to structural biology are well defined at equilibrium, and many are simple averages over the equilibrium distribution.

1.2.3 The probability depends on the energy

We require some way to compute the probability distribution of a complex molecular system. Ideally, we would compute the distribution analytically, but analytical solutions generally do not exist except for very simple systems such as the ideal gas. Therefore, we must sample actual states from the distribution, generate an ensemble that represents a proportionate sample of states, and compute properties from that. In other words, the probability distribution must be computed numerically.

The equilibrium probability distribution of a molecular system at constant temperature is given by the Boltzmann distribution, which relates the probability of a state P_i to the energy ϵ_i :

$$P_i = \frac{\exp(-\epsilon_i/K_B T)}{Z} \quad (1.1)$$

Where i indexes a state, some region (or point) in configuration space or phase space, coarse grained or otherwise; $\exp(x)$ is the exponential function e^x ; K_B is the Boltzmann constant; and T is the thermodynamic temperature. ϵ_i is the energy of state i , but its precise identity depends on both the thermodynamic ensemble and how the state is represented.

Z is the partition function, which is simply the normalisation constant which ensures that the probability for all states sums to unity. It is a constant for a given chemical system under some specified conditions, or otherwise constrained to a particular region S in phase or configuration space:

$$Z = \sum_{i \in S} \exp(-\epsilon_i/K_B T) \quad (1.2)$$

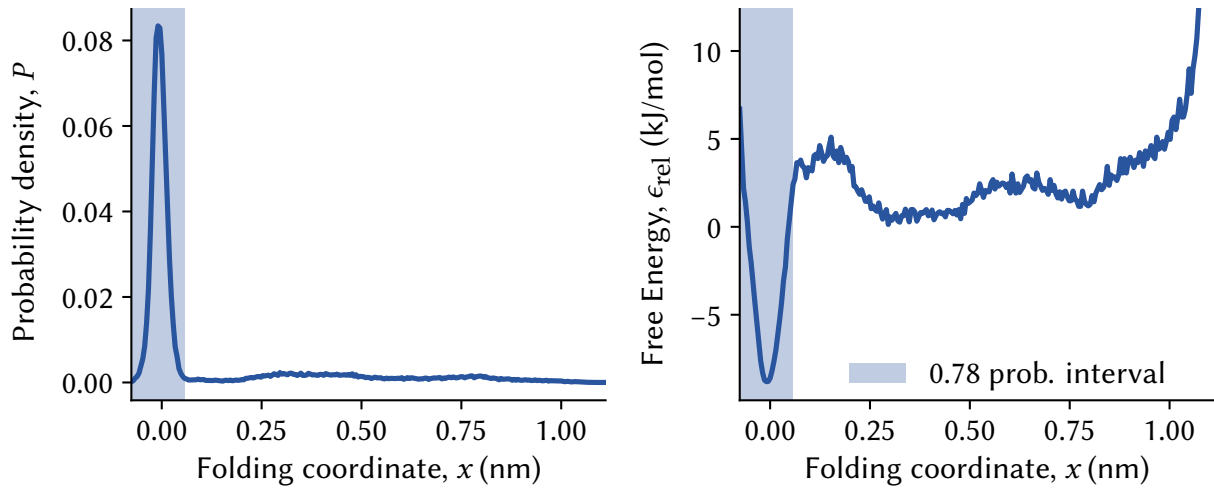


Figure 1.2: The probability and free energy surfaces of Chignolin along a simple one-dimensional folding coordinate. The other dimensions of configuration space are hidden in directions orthogonal to the page. The probability density (left) is a simple histogram of a converged REST2 simulation with the CHARMM22* force field; the free energy surface (right) is computed as described in equation 1.3 and then converted to units of kJ/mol and rezeroed. Note that relative to the probability density surface, the free energy surface is inverted and the local features are emphasised thanks to the negated logarithm function in the Boltzmann equation. Note also the deep free energy minimum/probability density maximum at $x = 0$ corresponding to the folded state

If we take the natural logarithm and solve for the energy:

$$\epsilon_i = -K_B T \log(P_i) - K_B T \log(Z)$$

$-K_B T \log(Z)$ is constant for a given system and temperature and is called the free energy (see figure 1.2). We can also consider dividing, or partitioning, the phase space of a system into multiple regions or ‘states’. Each state is then distributed according to the Boltzmann distribution by considering the partition function of that state as though it were the whole system. Similarly, the free energy of the states can be computed from their respective partition functions. The whole system then obeys the Boltzmann equation, with i indexing the sub-states and ϵ_i representing the free energy of those states.

If we are content with relative energies between states of the same system, the zero of energy may be set to the partition function. Experiments generally only give relative energies so this is convenient, and it drastically simplifies computation:

$$\epsilon_{i\text{rel}} = -K_B T \log(P_i) \quad (1.3)$$

In this view, energies are just a mathematically convenient representation of probabilities — the negative log likelihood. If several relative energies apply to a system, the total energy is

simply the sum, consistent with multiplication for combining independent probabilities:

$$\epsilon_i + \epsilon_j = -K_B T \log(P_i P_j) \quad (1.4)$$

In addition, relative energies depend only on the interesting states — contributions of other states are thrown away with the partition function.

The probability and energy are therefore intimately linked. The energy of a state is proportional to the negative logarithm of the probability, and the constant of proportionality can be computed as the partition function of all the states accessible to the system. This proportionality constant cancels out when considering only relative states of the same system, which makes them a pragmatic abstraction in biophysics. Moving up a probability density surface is exactly equivalent to moving down the equivalent energy surface, albeit notably steeper thanks to the logarithm. If we can compute the appropriate energy of a state, we can thus easily draw conclusions about that state's likelihood.

1.2.4 The energy can be estimated from the state

Many physical laws are nothing more than functions that go from configuration to force. The integral of force with respect to configuration gives the potential energy (equation 1.5), as in Coulomb's and Hooke's laws:

$$U(r) = \int F(r) dr \quad (1.5)$$

$$F(r) = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2}$$

$$U(r) = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r}$$

$$F(x) = kx$$

$$U(x) = \frac{1}{2} kx^2$$

When considering only configuration space properties under constant temperature, volume, and number of particles, the energy ϵ_i in equation 1.1 is simply the potential energy. For phase space properties, the kinetic energy must be included, easily computed from velocity:

$$U(\dot{x}) = \frac{1}{2} m \dot{x}^2$$

When the number of particles varies, the chemical potential must be included in ϵ_i , and if the volume v is allowed to vary in order to maintain a constant pressure p , the product pv is included to account for work done by the system. Likewise, the energy depends on the representation of the state. Any region in a coarse-grained representation encapsulates some generally larger, higher dimensional region of the fine-grained configuration space, and the entropy available in interactions between such regions must be accounted for as potential energy in the coarse representation. Coarse-grained energy functions may be unable to faithfully reproduce these interactions if they rely on details that have been lost.

Understanding energy as negative log likelihood can be used in reverse to compute energy functions from known statistics, and energies are also often derived in a coarse-graining process from quantum chemistry. Machine learning has recently found enormous success in producing energy estimates when the system is represented in an appropriate basis.³¹

Because the Boltzmann distribution cannot be analysed directly, all biophysical methods use some sort of function that goes from state to energy and is used to score the likelihood of structures. The scoring function is the model of the interactions within the system, and the accuracy of the method is always limited to the true answer given the scoring function, not that of reality. However, a scoring function alone can only give the relative likelihood of a state; to compute the absolute probability, we need some algorithm to provide context and find its place in the global probability distribution.

1.2.5 Algorithms map out the probability density surface as energy

A 50 kDa protein with 500 residues, each of which can either be in an α -helix or β -sheet, has 2^{500} , or about 10^{150} , available states. At this coarse level of detail, only one state at most could be considered folded, and almost all states have steric clashes or otherwise high energies, and therefore low probabilities. If we were to search naïvely, the universe's 10^{80} particles each computing a state every Planck time would take about 10^{19} years to find the folded state. This would be an inconvenient method for any biophysicist relying on modern farming techniques, as all the stars in the universe will run out of fuel in only 10^{14} years.

This ludicrous two-state-per-residue representation is obviously insufficient. Real proteins even of much smaller size have many more possible states, firstly to model coil residues and non-ideal secondary structures, but also side-chains and solvents. How can we locate and score the reasonable structures without having to test every one of many possible states? Given this interpretation, nature's ability to fold proteins in mere minutes is formidable. Computing the partition function might seem like a short cut, but in principle it requires converging a sum over all states and has the same difficulty.

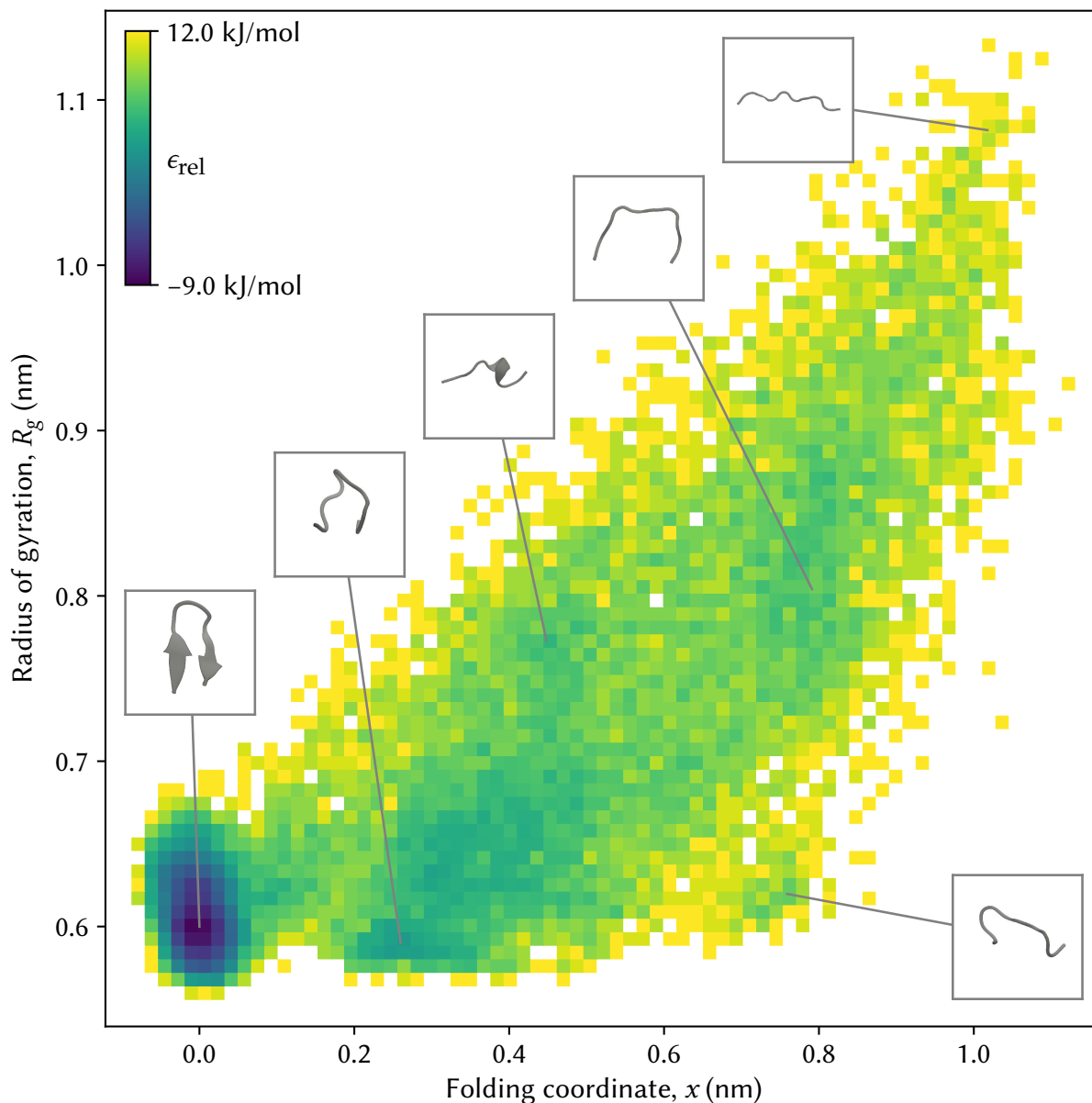


Figure 1.3: The funnel-shaped folding landscape of Chignolin in two dimensions. The funnel's mouth is flat beneath the page, and it falls from the regions of high free energy (yellow) to those of low free energy (blue). The x -axis uses the same folding coordinate as in figure 1.2, while the y -axis uses the radius of gyration. Frames from the simulation are shown alongside their positions in the free energy surface. Note that two states are clear in this image; a low free energy, highly concentrated folded state in the lower left ($x \approx 0.0$ nm, $R_g \approx 0.6$ nm), and a much more spread out unfolded state that covers a wide variety of conformations with various properties, including a prominent misfolded intermediate ($x \approx 0.25$ nm, $R_g \approx 0.6$ nm). The protein folds by falling into states of lower free energy, which here most obviously form a shallow funnel from the unfolded state into the misfolded intermediate. From the misfolded state, the barrier to the folded state is reduced, and once folded the reduced potential energy stabilises the state even in the face of the massive entropic contribution to the unfolded state from the sheer number of possible configurations. In addition, note the ruggedness of the surface; it is not a smooth funnel leading directly to the folded state, but meanders around and has many local minima.

Nature solves this problem by letting the shape of the probability density guide its search. Most states have very high energies, and this leads to a near-zero probability after exponentiation, so only a tiny fragment of phase space is truly available to the system. As long as the potential energy surface is smooth enough, the system can flow down the surface to the regions of lowest free energy. In other words, a protein folds by following a funnel-shaped³² energy surface from any one of a huge number of unfolded states to a much lower number of folded states (see figure 1.3).

1.2.6 Choice of algorithm depends on scoring function and goals

If the energy surface, or model thereof, is simple enough, then well-known optimisation algorithms can very efficiently fold proteins. By combining a careful choice of representation with machine learning, AlphaFold won CASP13 with a learned probability density surface that could be optimised by a simple gradient descent algorithm.³¹ Hand optimised potentials like Rosetta³³ commonly use variants of the Metropolis Monte Carlo³⁴ sampling algorithm. A biophysical method therefore cannot be reduced to either its scoring function or its algorithm.

Real proteins, and therefore highly accurate models of them, do not have potential energy surfaces smooth enough to optimise by gradient descent. In particular, dynamics cannot occur on such a surface, as the system cannot move between metastable functional states. In addition, such smooth descriptions cannot yet fold proteins reliably.³⁵ Monte Carlo methods can, in principle, model the multi-state dynamics found in real proteins, but in practice simulating the Newtonian dynamics of the system directly is more efficient. This simulation procedure is known as molecular dynamics (MD). This is not to say that nature has stumbled upon the best algorithm for the problem; many so called enhanced sampling algorithms have been developed that use MD sampling at their core, but dramatically improve the efficiency by controlling where that sampling takes place. Very recently, a machine learning approach that can sample directly from the Boltzmann distribution via a learned transformation from a Gaussian distribution has been developed; however even this method uses MD to generate training data.³⁶

1.3 How does MD work

The key insight behind molecular dynamics is that the same function can model both the structure and dynamics of a physical system: the energy. In addition to being the negative log likelihood in configuration space, the potential energy is the integral of the force with respect to position. This means that a scoring function which simply computes the potential energy

can be differentiated to yield the forces on every part of the system. These forces can then be used to step the system through time, sampling the local energy surface. Our knowledge of the world enters an MD simulation through the force field; simulating a natural process is a convenient way to sample phase space, not the goal of the technique.

MD is more efficient than other sampling methods for physical systems when there are many degrees of freedom. For example, Monte Carlo approaches to sampling a solvated protein must usually change only one atom per step and then compute the energy again, or else the step will most likely be rejected. MD allows every atom to be moved in a way that guarantees acceptance. While MD is routinely applied to small chemical systems using quantum mechanics to step through time, systems as large as even small peptides in explicit solvent necessitate a Newtonian treatment for efficiency. This treatment occurs at the level of atoms, which are parametrised as single interaction sites. While the corresponding gains in efficiency are enormous, Newtonian methods cannot describe processes that depend on the dynamics of individual electrons such as bond breaking or formation.

1.3.1 The force field gives potential energy and forces on a single state

A force field is a scalar function $V(\mathbf{x})$ that yields the potential energy at coordinates \mathbf{x} and is differentiable in \mathbf{x} . It encodes all the interactions and physical laws — except the laws of motion — of the system. Because it simply yields the potential energy, a force field is easier to parametrise than a scoring function; entropic free energy terms emerge spontaneously from the simulation, rather than being explicitly represented in the scoring function.

The negated derivative $\mathbf{F}(\mathbf{x})$ is called the force:

$$\mathbf{F}(\mathbf{x}) = -\frac{d}{d\mathbf{x}}V(\mathbf{x}) \quad (1.6)$$

The force field therefore not only encodes the potential energy, but also the forces. Force fields are usually constituted as the sum of individually differentiable terms, so the derivative can be computed analytically before a simulation. Force fields are extremely cheap to evaluate, both as energies and as forces, compared to both scoring functions and quantum chemical methods.

Today, force fields are simulated with periodic boundary conditions, which produce fewer modelling artefacts than walls or open boundaries. The system is constructed in a box, and atoms that leave one side of the box wrap around, re-entering the box from the other side. What is simulated is therefore a dilute liquid crystal of protein in water. Each pair of atoms is computed only once via the minimum image convention; only the interactions between an atom and its closest periodic image are calculated.

Modern force fields all use same basic mathematical functional form with different parameters, with some variance in how dihedral angles are treated. As an example, the CHARMM22* force field for a system of n atoms is given as a potential energy function below:

$$\begin{aligned}
 V = & \sum_{\text{bonds}} k_b(b - b_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{cmaps}} \text{cmap}(\phi, \psi) \\
 & + \sum_{\text{dihedrals}} k_\phi(1 + \cos[v\phi - \delta]) + \sum_{\text{impropers}} k_\omega(\omega - \omega_0)^2 \\
 & + \sum_i \sum_j^n \left(\frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \right)
 \end{aligned}$$

The first two lines give the bonded potentials. b , θ , ϕ , and ω are respectively the values of a particular bond length, bond angle, dihedral angle or improper dihedral angle in the current state. k_b , k_θ , k_ϕ , and k_ω are their respective force constants, and b_0 , θ_0 , ϕ_0 , and ω_0 are their equilibrium values. v and δ are respectively the dihedral angle's multiplicity and phase, while $\text{cmap}(\phi, \psi)$ is a function of the protein backbone dihedral angles ϕ and ψ that maps a particular pair of values to an correction energy.

The last line represents the non-bonded potentials. r_{ij} is the current distance between atoms i and j , which is used in the pairwise Lennard-Jones and Coulomb potentials whose computation dominates the time spent calculating a force field on a state. $4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$ is the Lennard-Jones potential between atoms i and j , parametrised by the energy well depth ϵ_{ij} and the distance σ_{ij} , which relates to the van der Waals radius $r_{vdW} = \sqrt[6]{2}\sigma_{ij}$. The Lennard-Jones potential models the total van der Waals interactions, including dispersion and the Pauli exclusion principle. $\frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}}$ is the Coulomb potential between atoms i and j with charges q_i and q_j .

The CHARMM22* force field specifies not just this functional form, but also all of the parameters that go into it — that is, the symbols above that do not depend on the current state. The CHARMM27, CHARMM36 and CHARMM36m force fields share the same form with modified parameters, and other biomolecular force fields have only minor differences. Parametrising a force field requires the fitting of thousands of parameters to accurately reproduce many properties of interest, and requires substantial computational investment just in testing.

The non-bonded potentials account for the vast majority of computation during any MD step because they must be calculated for every pair of atoms in the system, which would lead the number of calculations to grow with the square of the number of atoms. This contributes to the substantial efficiency improvement achieved by coarse-grained force fields. These potentials are therefore truncated at 1.2 nm, and a shift is applied that smoothly reduces the

forces and energies to zero at that cut-off, so that in practice the interactions can be computed in linear time.

Long-range Coulomb interactions can be important for biomolecular systems,³⁷ so rather than using a cut-off, electrostatics can be computed efficiently for the entire periodic system in reciprocal space with methods such as PME.³⁸ This allows long-range electrostatic interactions to act on proteins, and has been shown to be more realistic than other means of treating long-range electrostatics.¹⁹ Alternatives to PME include the similar Gaussian split Ewald method, used by the Desmond software;³⁹ and the venerable reaction field, which replaces long-range electrostatics with an isotropic field of constant dielectric constant.⁴⁰ Long-range interactions can usually be safely ignored for the Lennard-Jones potential as it decays much more quickly.

1.3.2 Dynamics give rise to new states

The unique formulation of a force field allows the system to be stepped through time. Given an initial state, forces and therefore accelerations are computed from the force field. Approximating the acceleration as varying linearly over a given time-step, the initial accelerations are the average acceleration over the time-step starting half a time-step before $t = 0$ and ending half a time-step after. This is used to compute the velocity at half a time-step, and that velocity is similarly used to update positions at the full time step. This process is repeated, stepping the system through time. Symbolically for a single atom:

$$\begin{aligned}\ddot{x}_t &= \frac{F(x_t)}{m} \\ \dot{x}_{t+\frac{1}{2}\Delta t} &= \dot{x}_{t-\frac{1}{2}\Delta t} + \ddot{x}_t \Delta t \\ x_{t+\Delta t} &= x_t + \dot{x}_{t+\frac{1}{2}\Delta t} \Delta t\end{aligned}$$

This algorithm is called the leapfrog integration algorithm and is a formulation of Newton's equations of motion. It is called an integrator because as Δt approaches 0, stepping through time like this becomes mathematically equivalent to integrating over time; it is therefore an example of a numerical integration algorithm. The leapfrog algorithm is an efficient, reversible, and symplectic algorithm that balances accurate integration with large time-steps.^{41–43}

For stepping through time to be equivalent to sampling, we must posit the ergodic hypothesis for our system. A system is ergodic if averaging over time is equivalent to averaging over the equilibrium distribution.⁴⁴ While ergodicity can be proved mathematically for simple systems, biomolecular systems are too complex, and ergodicity must be assumed from our intuition

about the world and the fact that real proteins do, in fact, equilibrate. If we take this step, and assume ergodicity, MD ceases to be mere simulation and instead becomes an efficient way to sample phase space. Compared to competing methods like Monte Carlo,⁴⁵ MD enables arbitrary movement around phase space and guarantees that progress is made with every calculation.

Understanding MD as an efficient sampling method, rather than a simulation technique, gives room to improve its efficiency by making it less like a simulation. This is called enhanced sampling (see section 1.4.7). Naïve MD is inefficient because most of its time is spent in regions of low energy and therefore high probability. Barriers between these regions are high in energy and transitions are rare, so the system can become trapped near the initial state. By spending less computer time there and somehow giving that time a greater statistical weight in the final ensemble, sampling can be dramatically accelerated.

Indeed, MD only achieves timescales relevant to chemical biology by abstracting quantum chemical effects into the force field's parameters so the system can be treated classically. This abstraction justifies both the choice of Newtonian equations of motion and the treatment of atoms as fundamental particles. Quantum MD is possible,⁴⁶ but it must use the more complex Schrödinger equation and must operate on the level of electrons and nuclei. In essence, MD results from a coarse graining of the electronic degrees of freedom, which results in a system whose dynamics are described well by Newtonian mechanics. Proteins are massive enough that they are well approximated by Newtonian mechanics, with a few exceptions. Critically, chemical reactions rely on the movement of electrons and cannot be treated classically, and will therefore not occur in classical MD. However, chemical reactions are easy to expect ahead of time and so their simulation can be avoided. Most other quantum effects are encoded in the force field as bond vibrations and the like, but parametrisation can be difficult when electron orbital shapes are important, such as in metal binding proteins or when the effect of polarisation is significant.

1.3.3 Dynamics produce an approximation of the trajectory

The output of a simple MD run is a series of points in phase space, each associated with a time, called a 'trajectory'. These points, or frames, are highly correlated with the neighbours in time, so almost all of them are discarded for efficiency. It would usually be conservative to preserve one frame per picosecond with a time-step of two femtoseconds. This digital trajectory is a discrete approximation of the true trajectory, which is the unique continuous path through phase space that satisfies the equations of motion and starts at the simulation's initial conditions.

Newton's equations follow the law of conservation of energy, so as the system moves to more probable states and lower potential energies the kinetic energy rises. This manifests as a rising temperature; in longer simulations, the temperature drifts as the system samples different regions of the potential energy landscape. As a consequence, the pressure can also vary depending on the equation of state $PV = nK_B T$. These drifts are only worsened by integration error, which adds a random drift to the total energy.

Drifting energies are realistic for an isolated system, but systems of biomolecular interest don't experience them because changes in kinetic energy in one part is balanced by heat flow from the rest of the system. The temperature of a protein in a simulation might be buffered by only a few molecular layers of water, as opposed to the rest of the solution in the pipette or water in the beaker or matter in the universe. The distribution of states sampled by a system differs depending on which thermodynamic variables are fixed and which are allowed to vary. In order to sample from more realistic distributions with a constant temperature or pressure, and to mitigate energy drift from integration error, the integration algorithm can be modified with a so-called thermostat or barostat (see section 1.4.6).

Many-body systems like proteins are chaotic, so tiny changes in starting conditions lead to trajectories that diverge exponentially with time.⁴⁴ Identical starting conditions with theoretically deterministic algorithms can produce uncorrelated trajectories within nanoseconds of simulation time on different hardware due to differences in numerical rounding. With a finite time-step, the integrator also introduces discretisation error, further deviating it from the true trajectory. This integration error comes from the approximation that the acceleration varies linearly over the time-step, and more detailed integrators can reduce it by considering further derivatives of x .⁴⁷ Integration error is ameliorated with a smaller time-step, and with a symplectic and reversible integrator such as leapfrog the error approaches zero in the limit as the time-step goes to zero.^{41,42} These integrators should therefore be favoured.

While a short time-step may improve integration error, it also reduces sampling efficiency and therefore increases statistical error for a given amount of computational resources. Thus, it is not the case that the most rigorous simulations have the shortest time steps. It is generally clear when a simulation's time step is too large, as the sudden, unrealistic changes in atom position from step to step violate assumptions made by the simulation software and lead to nearly immediate crashes. However, up to this point of instability, the sampling error almost always dominates at the scale of atoms and the largest possible stable time-step is therefore chosen. Coarse grained force fields, where interaction sites represent multiple atoms, can sometimes be stable at timesteps where the integration error is significant.⁴⁷ If extreme precision is required, a hybrid Monte Carlo scheme can eliminate integration error completely.^{48,49} There is also some hope that while the produced trajectory diverges from

the true trajectory of the given starting conditions, it shadows, or approximates, some other starting conditions.^{50,51}

1.3.4 Equilibrium ensembles can be inferred from trajectories

The goal of MD is to generate an ensemble of states in proportion to their equilibrium probabilities. Such an ensemble differs in one significant way from a trajectory: structures in an ensemble are not associated with time. Properties that are independent of time and can be computed from the ensemble are called equilibrium, structural, or thermodynamic properties, whereas those that depend on time are called dynamical or kinetic. Modern force fields are optimised to reproduce good equilibrium properties, and minor improvements can dramatically improve kinetics for particular systems.^{52,53} MD algorithms such as thermostats, barostats and enhanced sampling methods are usually targeted at equilibrium properties and have been found to disturb the kinetics of the system.^{52–54} Users should take extra care when investigating kinetic properties to ensure their method and force field are appropriate for the problem.

If our goal is to compute an ensemble, then it is important to remember that simulation is merely a means to that end. MD simulates the natural time-evolution of the system because it is an efficient means of sampling from the Boltzmann distribution, not because we want a video of a protein. While MD is reversible and deterministic in principle, in practice these are unimportant features of the method and most implementations discard them in favour of efficiency.

Simulations must be very long before they can produce more than a handful of statistically independent equilibrium samples. Microseconds of simulation of small soluble proteins is accessible on commodity hardware,^{3,55} and milliseconds on dedicated hardware.^{4,6} Most proteins take milliseconds to minutes to fold, and important structures like amyloids take decades to mature. The statistical error associated with under-sampling the true equilibrium distribution at such disparate time-scales often silently dominates the results of a simulation and has probably contributed to MD's reputation as being unreliable. In general, it is impossible to know in advance if something new would happen if we just simulated a little longer.

The statistical error can be further improved by running many replicas of the simulation, each with different starting conditions. This is commonly done by starting many different simulations with the crystal structure, but different initial velocities. Their initial correlation will rapidly decay, and after a short equilibration period one is left with an ensemble of independent trajectories. This ensemble weakens the reliance on the ergodic hypothesis⁴⁴ and has been found to accelerate sampling for the same amount of simulation time.⁵⁶ In addition,

modern parallel computer hardware is often more efficiently exploited by numerous replicas that can be run independently than by a single multi-node simulation.³

Starting structures are usually not drawn from a force field's equilibrium distribution. Modelling errors or artefacts from the starting structure, the differences between dilute solution and the environment of the crystal or the cell,⁵⁷ errors in the force field or approximations made in preparing the system for simulation can all contribute to this. Because we wish to draw samples from the Boltzmann distribution, it is important to discard data collected before the system reaches equilibrium. Thermodynamic variables such as pressure, temperature, and energy usually stabilise faster than properties of interest, so it is important to test for equilibration for the property being measured. Automated methods are available.⁵⁸

With an ensemble of equilibrium trajectories in hand, properties can be computed from the trajectory as though it were an ensemble of states. When considering uncertainty, however, it is important to remember that frames in a trajectory are not statistically independent. Sophisticated methods for calculating free energy landscapes from trajectories exist,⁵⁹ and for other properties autocorrelation, convergence analysis, and bootstrapping can all provide estimates of the effective number of independent samples present in a trajectory for a given property.⁶⁰

An ensemble of states also gives direct access to entropic properties of the system. For a state with Ω equiprobable or degenerate microstates per unit volume in phase space, the entropy has a simple form:

$$S = K_B \log(\Omega) \quad (1.7)$$

That is, the entropy simply counts the number of specific atomic configurations that contribute to a state in some coarse-grained picture of the system. Scoring function-based methods generally model proteins with single models, and so entropic effects must be directly modelled in the scoring function as though they were a form of potential energy. Thus, effects like solvation and the hydrophobic effect, loop flexibility and entropic stabilisation of ligand binding must be parametrised directly. Molecular dynamics force fields produce ensembles rather than single models, and so these effects emerge from the dynamics of the potential energy surface.

Note that when microstates are not equiprobable, such as when the temperature is constant and energy is constantly exchanged with the environment, the entropy takes on a slightly more complex but conceptually similar form:

$$S = -K_B \sum_i p_i \log(p_i) \quad (1.8)$$

Where microstate i has probability p_i . This so-called Gibbs entropy is equivalent to the Boltzmann entropy given in equation 1.7 when all microstates have equal probabilities and

unit volume in phase space.

1.4 Common traps and their solutions

1.4.1 Use a high-quality, appropriate force field

Even with perfect sampling, MD only ever describes the world of the force field. If a problem is not within the force field's narrow domain of applicability, the results may be irrelevant to the real world. Force field development has been going on for decades, but sufficient MD sampling to optimise them efficiently has only been available for less than a decade.

Proteins are the most important part of most biomolecular systems and are among the most difficult to parametrise. Thus, the choice of force field generally starts here. The history of force field development of the core CHARMM, Amber, GROMOS and OPLS families was reviewed recently by Riniker⁶¹ and Lemkul⁶² and is summarised in figure 1.4, and the trajectory of force field development has been discussed by Fröhling et al.⁵

CHARMM proteins

The CHARMM (CHemistry At Harvard Molecular Mechanics) force field family is distributed with the CHARMM MD software, and individual force fields are named for the version of the software they were first distributed with. They are parametrised with both structural information from quantum chemical calculations and experimental data.⁵ CHARMM22⁶³ was the first CHARMM all-atom force field for proteins. RNA parameters and dihedral correction maps were added in CHARMM27,^{64†} which improved backbone dihedral distributions. CHARMM27 has a notable bias for α -helix over β -sheet which was corrected in CHARMM36.⁶⁵ CHARMM36m⁶⁶ was then released to destabilise an unrealistic left-handed α -helix found in disordered proteins.¹⁸ CHARMM force fields are distributed by its authors in formats compatible with many different MD engines, making it a reliable parameter set that has steadily improved in quality and today approaches the state of the art.

There is one notable third-party optimisation of a CHARMM forcefield. CHARMM22*⁶⁷ was produced after optimisation using orders of magnitude more sampling than previously available on the purpose-built Anton supercomputer.⁴ CHARMM22* abandons CHARMM27's correction maps for backbone dihedral angles of residues other than glycine and proline with re-optimised dihedral parameters. It is a well-tested, accurate protein force field^{6,18,68–73} with excellent performance even on disordered proteins, despite being fitted exclusively to folded proteins.^{18,72}

[†]Also known as CHARMM22/CMAP when used without RNA parameters

Amber proteins

Amber force fields are distributed with the Amber software. Like CHARMM, they are parametrised with both QM and experimental data.⁵ New parameters are regularly released and rarely have time to be evaluated before they are updated.^{74–79} In addition, Amber force fields are a popular starting point for optimisations of force fields. ff99SB⁷⁶ in particular has produced a popular and accurate family of force fields, beginning with an optimisation of the backbone parameters in ff99SB*.⁸⁰ Later, several side chain dihedral potentials were improved and released as ff99SB-ILDN.⁸¹ These two improvements are commonly combined to form ff99SB*-ILDN,⁶⁷ and an optimisation of atomic charges targeting per-residue helical propensities led to ff99SB*-ILDN-q.⁸² ff99SB*-ILDN-q was then optimised alongside the TIP4P-D⁸³ water model aided by sampling from the Anton supercomputer to produce a99SB-*disp*,⁷² which includes both the protein and water parameters. Among other optimisations, a99SB-*disp* incorporates newly balanced water-protein dispersion forces to achieve a force field that accurately reproduces the properties of both folded and disordered proteins. a99SB-*disp* was also compared to several other modern force fields on benchmark systems that were not used in its parametrisation.⁷² Very recently, the protein–protein complex performance of a99SB-*disp* was further optimised with further Anton sampling to produce DES-Amber.⁸⁴ This involved changes to the force field's non-bonded parameters, and thus may be incompatible with other ff99SB-descended force fields.

Meanwhile, ff03⁷⁵ was developed on the basis of new quantum chemical calculations at a higher level of theory than the ff99SB family. Optimisations include ff03*,⁸⁰ ff03w,⁸⁵ and finally ff03ws.⁸⁶ Inspired by ff99SB-ILDN, the Amber team derived ff14SB⁷⁷ from ff99SB by re-parametrising backbone and side-chain dihedral parameters based on a large library of QM-optimised dipeptide conformations. ff14SB is the force field recommended by the Amber software team for all protein simulations. It was further optimised with a CHARMM-style dihedral correction map to improve performance with disordered peptides as ff14IDPs,⁸⁷ and later ff14IDPSFF.⁸⁸ The Amber team has also pursued a number of more systematic approaches to the production of force fields, resulting in ff14ipq,⁸⁹ ff15ipq⁷⁸ and FB15,⁷⁹ though none of these have supplanted the traditional optimisation approach.

Many Amber-descended force fields have been found to be very accurate for folded proteins, especially those based on ff99SB-ILDN.^{17,68,72} Specific force fields have been optimised to reproduce methyl rotation barriers for NMR relaxation data,⁵³ and NMR chemical shifts.⁹⁰ Amber force fields also have access to a wide variety of parameters for non-canonical amino acids⁹¹ and post-translational modifications,⁹² and the free AmberTools software package provides extensive tooling for the parametrisation of new compounds.

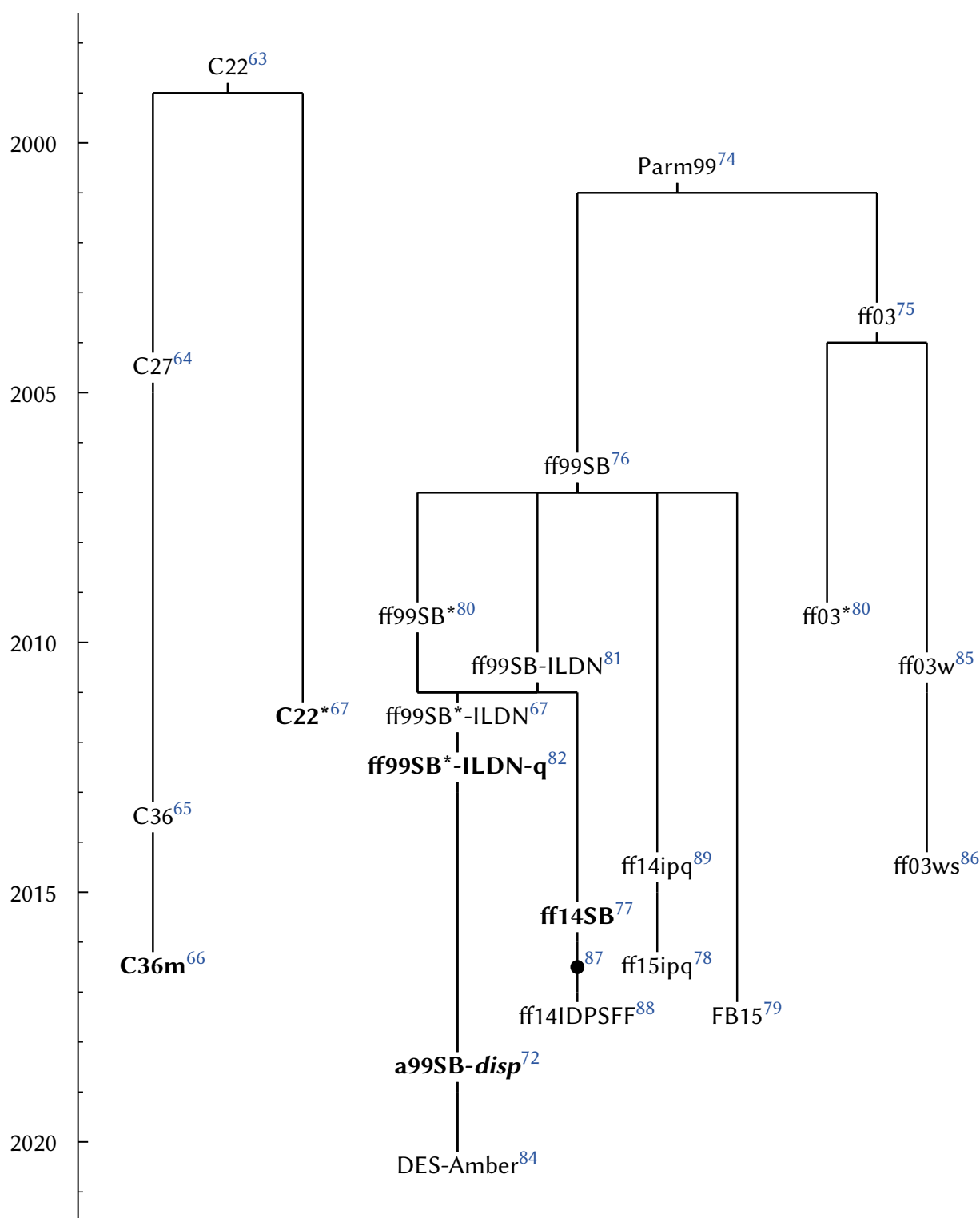


Figure 1.4: The CHARMM and Amber protein force field family trees. Citations are given as superscripts. Force fields in bold are well-tested, general-purpose force fields that have not been clearly superseded. The filled circle represents ff14IDPs.⁸⁷ Note that a99SB-disp and DES-Amber both use a more computationally expensive 4-point water model than the other force fields in this figure, which use 3-point water models. Note also that while ff14SB was motivated by the ff99SB-ILDN optimisation as shown, its actual parameters derive exclusively from ff99SB.

GROMOS proteins

GROMOS force fields are parametrised to reproduce experimental thermodynamic energies and largely eschew structural properties from quantum chemical calculations.⁵ They use a united atom format, with non-polar hydrogen atoms coarse-grained with their parent carbon atom into a single interaction site. This reduces the number of atoms in the system but has been discarded as a strategy by other force fields⁶¹ because of its marginal effect on performance in systems dominated by water and the increased flexibility provided by an all-atom model.⁹³ The most recent GROMOS protein force field is GROMOS 54a8,⁹⁴ which is a revision of the more popular GROMOS 54a7.⁹⁵ Comparative studies typically focus on CHARMM and Amber force fields, and those that include GROMOS suggest mixed protein performance..^{69,71,96–98}

OPLS proteins

OPLS is an early force field with parameters for many classes of biomolecules.⁹³ It was parametrised to reproduce experimental energies and QM structures.⁵ Protein parameters in the freely available version, OPLS-AA, has not been subject to the same level of optimisation as CHARMM and Amber.^{17,68,99} However, a commercially available release, OPLS3¹⁶ and its unpublished update OPLS3e combines OPLS's broad molecule base with lessons learned from CHARMM and Amber regarding modelling proteins. It can produce results for protein-ligand binding energies with experimental accuracy for many drug-like molecules¹⁶ and is distributed with the commercial Schrödinger software.

MARTINI

MARTINI¹⁰⁰ is notable for being a coarse-grained force field. Instead of treating each atom explicitly, a single 'bead' represents approximately four heavy atoms along with their associated hydrogen atoms. This coarse-graining provides an approximately thousand-fold acceleration in sampling compared to atomic-resolution simulation. This enhanced performance derives not only from the reduced number of interactions, but also from an increased stable time-step Δt and increased sampling owing to the smoother potential energy surface. While coarse-graining provides a substantial efficiency improvement, it comes at the cost of some loss in accuracy and precision.^{101,102} Several other coarse grained force fields for biomolecular systems exist,³⁰ but MARTINI is exceptional for its wide use and broad coverage of biomolecules and biopolymers.

MARTINI uses standard MD force field interactions and has a similar functional form to the above force fields, allowing it to run on existing MD software. Beads are organised according to their chemical properties, and multiple similar chemical groups can be represented by the

same bead. Only integer charges are represented by electrostatics, and all other non-bonded interactions are realised with the Lennard-Jones potential. Rather than each bead being parametrised individually and pair potentials being derived from mixing rules as in other force fields, the matrix of Lennard-Jones interactions between all bead pairs is parametrised directly to reproduce experimental thermodynamic properties for simple molecules. Bonded parameters are then optimised against an atomic-resolution force field.

MARTINI was originally intended for simulations of lipids^{103–106} but has since been extended to other biomolecules including proteins,^{107–110} carbohydrates,¹¹¹ nucleic acids^{112,113} and others. The treatment of water has also been the subject of much optimisation,^{114–117} and tweaks have been made for GPUs,¹¹⁸ implicit solvent,¹¹⁹ or in combination with atomic-resolution models.^{120,121} The simplicity of the representation allows bead selection to be automated, making parametrisation of new molecules simple. MARTINI has been used to discover a new ligand binding pathway in the well-studied light harvesting complex II¹²² and simulate entire viruses.^{123,124}

Although MARTINI shines brightest when applied to massive systems on time-scales inaccessible to atomic-resolution MD, it can also be applied to smaller systems. The dynamics of individual proteins can be improved with an elastic network¹⁰⁸ or by combining them with a structural Gō model.¹²⁵ Dynamic protein-protein interactions have been found to require some scaling down of protein bead interactions.^{126,127} Protein parameters have been the subject of some optimisation already,^{109,110} and improved protein performance is a major goal of the upcoming MARTINI 3.0 update.^{102,128,129}

Polarisable force fields

The force fields discussed above, with the exception of the polarisable MARTINI water model, are fixed-charge force fields. Fixed-charge force fields have a single, set charge value for each atom in the system that does not vary over the course of the simulation. Polarisation effects are parametrised implicitly in this charge, as well as in the bonded and Lennard-Jones parameters and in the dynamics of the water model. As most modern codes are tuned towards fixed-charge force fields, and since fixed charge force fields are much older and therefore more refined and better optimised, they generally yield better performance at present; for instance, polarisable force fields are not yet able to reliably produce folded protein structures from sequence alone. However, polarisable models are in development and may be necessary to obtain dramatic improvements over the state of the art in protein force field design.^{62,84,130}

Water models

The vast majority of any dilute biomolecular system is water, and solvent-protein interactions are very important to protein structure and dynamics. It is therefore essential that the water model used in protein simulations is both extremely fast and accurate in its interactions with the solute. Fortunately, the dynamics and structure of bulk water is less important and grants some flexibility for parametrisation. Ideally, a continuum model could be used in place of explicit water molecules, and work is being done in this direction especially by the Amber software.^{8,131,132} For the time being, an explicit solvent is much more accurate¹³³ and models solvent entropic effects directly, and the best protein force fields are optimised around explicit solvent.

Solute force fields are optimised around particular water models, and care should be taken when using a different model.¹³⁴ Therefore, it is safest to simply use the water model for which one's protein force field is parametrised. Most protein force fields are parametrised for the 3-point water model TIP3P,¹³⁵ although CHARMM uses a variant that includes Lennard-Jones interactions on all three atoms¹³⁶ and GROMOS uses an alternative 3-point water model called SPC.¹³⁷

Several 4-point water models, with the oxygen's charge moved to a fourth interaction site to simulate the lone pairs, have been parametrised for improved accuracy in different situations. These 4-point models are invariably based on TIP4P.¹³⁵ TIP4P-D⁸³ optimises the performance of disordered regions, which have been found to depend strongly on the precise strength of protein-water dispersion interactions,^{72,83,86} but does not reliably improve the force field it was parametrised for, CHARMM22*.¹⁸ TIP4P-Ew¹³⁸ was re-optimised for use with long-range electrostatics via PME rather than a Coulomb cut-off, similarly to the 3-point TIP3P-Ewald.¹³⁹ However, this appears to be an insignificant improvement, even where the effects of long-range electrostatics should be important.¹⁴⁰ TIP4P/2005¹⁴¹ improves on the original TIP4P parameters with the benefit of twenty years of computer development, but requires re-parametrisation of the solute to improve protein performance.^{85,140} 5-point water models, with interaction sites representing each of the oxygen atom's lone pairs, improve the accuracy of the bulk water, but not the performance of solutes compared to 4-point models.^{142,143}

Metal cations

Organo-metallic binding is neither purely ionic nor spherically symmetric, and therefore the usual force field approximations of spherically symmetric potentials acting on atomic point charges without electronic structure or bond formation are naturally less appropriate.¹⁴⁴ Nonetheless, fixed point charge models with optimised Lennard-Jones parameters are available

with all force fields. These are generally appropriate for monovalent cations,¹⁴⁵ but their performance degrades with divalent cations.¹⁴⁶ Multi-site models^{144,146,147} and modified potentials^{145,148} can improve the accuracy of divalent metal cations, but care should be taken when studying metal binding or when a metal is structurally important to a protein.

Drugs and other small biomolecules

Proteins and nucleic acids have the advantage of being built up from relatively few building blocks. If the parameters of the building blocks are sufficiently accurate and transferable, then the whole macromolecule is accurate. Small, drug-like biomolecules exhibit a much wider variety of properties and structures, and therefore require enormously more parameters to describe accurately. Despite this, structural biology often relies on the interaction between small molecules and a protein. OPLS 3e claims similar accuracy to experiment with near-arbitrary drug molecules,¹⁶ but is a proprietary, commercial force field with under-tested proteins. CHARMM is compatible with CGenFF,¹⁴⁹ and Amber with GAFF,¹⁵⁰ both of which are general force fields designed to be transferable across different drug-like molecules. They do not require any QM calculation of the target in day-to-day use, making them convenient, but their accuracy leaves much to be desired.^{151–153}

In addition to the GROMOS 2016H66 small molecule force field,¹⁵⁴ GROMOS force fields are uniquely compatible with the world-class Automated Topology Builder (ATB) for parametrisation of arbitrary molecules from automated QM calculations.^{155,156} The ATB produces much greater accuracy in a wider chemical space compared to CGenFF or GAFF, but is slower to parametrise in the best case. This cost is mitigated by the complexity of manually producing or refining parameters for the former when the automatically assigned parameters are a poor fit, which is not needed for the ATB as the process is thorough and automatic. The ATB is available as a web server.

While still in its infancy, the Open Force Field Consortium¹⁵⁷ has developed a methodology of parametrising general force fields without atom types, dramatically simplifying the specification of the force field. SMIRNOFF99Frosst¹⁵⁷ is compatible with Amber force fields and has accuracy comparable to GAFF2 with only one tenth the number of parameters.^{157,158} It is hoped that this approach will lead to substantially more accurate general force fields in the near future.

Force fields: Conclusion

The choice of force field is the most important step in an MD investigation, and this choice determines the reliability of all results no matter how rigorously other steps are carried out.

Fortunately, force field parameters for proteins have improved rapidly in recent years.

CHARMM22* is a reliable, well-tested force field for apo-proteins in salt water, or if accurate CHARMM parameters exist for any ligand. Amber ff99SB-*disp* or other descendants of Amber ff99SB*-ILDN also perform well, and are especially well suited to metallo-proteins and proteins with unusual amino acid residues when used with the appropriate additional parameters. OPLS 3e has excellent performance for drug-protein binding where it is affordable. Combining parameters between families of force fields should never be attempted, as different families have different strategies for parametrizing non-bonded interactions and are not mutually compatible. Finally, the most reliable results are those that can be replicated or averaged across different high-quality force fields.

1.4.2 Check for sufficient sampling

Even if the force field perfectly models a phenomenon, one must have sufficient sampling to reach all the important states contributing to that phenomenon for the resulting ensemble to describe it. Furthermore, transitions between the important states must happen enough that the relative populations of the states have a reasonably small uncertainty. There are many possible paths through configuration space, and an appealing narrative deduced from a single simulation may have very little to do with the real world or even the force field's model.²⁰ It may be helpful to think of MD as a sampling technique, rather than a simulation technique (see section 1.3.2).

Insufficient sampling can often, but not always, be revealed by assessing the uncertainty of quantities derived from simulation. This should be routine, as quantities should never be reported without uncertainty. However, computing the uncertainty from an MD simulation can be complicated by the fact that one cannot consider frames to be drawn independently from the underlying distribution. Grossfield et al.⁶⁰ have recently reviewed best practices in assessing uncertainty from trajectory data.

It is also important to consider expectations for a simulation; statistical approaches to uncertainty can only describe the uncertainty for regions of phase space that have already been reached. If a property's true value is intermediate to contributions from two states, and a simulation is started in one of these states, the uncertainty will steadily decrease with time as the state is fully explored. When the system makes a jump to the other state, the uncertainty will suddenly rise, despite the estimate of the property improving relative to the true value. The uncertainty may not fall to an acceptable level until many transitions between the two states have been observed. This is not an error, it merely reflects that there is no way to predict from a trajectory in one state alone that there exists a second important state. Thus, human

intuition and insight are essential.

One potentially under-appreciated way to improve the robustness of sampling is to perform multiple replicate simulations of the system.^{44,56,159} Not only does this improve utilisation of parallel hardware,³ it also decreases the reliance on the ergodic hypothesis⁴⁴ and makes it easier to discard narratives that apply to only one simulation.⁵⁶ Empirically, 5-10 replicas is a good starting point, and one is generally better off splitting the same amount of computer time over many replicas rather than one long simulation.⁵⁶

If one needs dramatically more sampling than is available with atomic-resolution simulation, a coarse-grained force field such as MARTINI^{100,129} should be considered. While the systemic errors introduced by this simplification may be substantial, they may be preferable to unknown statistical errors produced by under-sampling, especially for large systems.

1.4.3 Check that the produced trajectories are reasonable

Many errors resulting from unrealistic parameters and algorithms can be ignored by simulation software, but are quite apparent to a human observer.¹⁹ These can include (i) unphysical phase changes, such as frozen or evaporated water or gel-phase lipids; (ii) unexpected conformational changes, such as the rapid denaturation of proteins or loss of planarity in aromatic rings; (iii) vacuums spontaneously forming in condensed phase systems; (iv) violations of the second law of thermodynamics such as particle movement against concentration gradient. These errors can easily be carried forward throughout analysis and only noticed if they eventually cause an unbelievable result. It is therefore valuable to visually inspect simulations for such errors, at least until one has gained some level of confidence in their implementation of a method. One should also consider running automated validation tests on one's system¹⁶⁰ that can detect non-Boltzmann sampling and other unphysical behaviour.

1.4.4 Use appropriate compute hardware and software

Much of the improvement in force fields of recent years may be attributed to a massive increase in the availability of compute hardware that is more appropriate to MD than the standard CPU. Three stand-out force fields (CHARMM22*,⁶⁷ OPLS3e,¹⁶ and Amber ff99sb-*disp*⁷²) were all optimised with the Anton supercomputer,^{4,161} which was purpose-built to accelerate MD simulations by two orders of magnitude relative to the state of the art. At the same time, MD algorithms became available on GPUs, which had reached a point at which a single GPU could perform an MD simulation faster than an entire cluster of CPUs at a fraction of the price.³ Purpose-built computer hardware is not within the reach of most researchers, but the great gains in efficiency accomplished by GPUs make their use nearly essential for MD today.

Statistical error usually dominates in MD simulations, so choosing an efficient platform is essential. Some codes have recently emphasised correctness,¹⁶⁰ but most packages in use today are reasonably reliable,^{162,163} so it is generally best to use the fastest software and hardware that supports the desired method and force field. Changing software settings can make a substantial difference, so it is always useful to read the manual and understand how the simulation is executed by the software.

Popular codes today include:

1. GROMACS,^{164–170} a free and open source package focusing on speed, flexibility, and recently reliability;
2. Amber,^{171,172} a commercial package with great GPU and implicit solvent performance;
3. OpenMM,^{173–175} a free and open source software library with extensive support for nearly arbitrary potentials at great speed;
4. NAMD,¹⁷⁶ a fast and simple code that interoperates with popular visualisation software VMD; and
5. Desmond,¹⁷⁷ the free MD component of the proprietary Schrödinger software package whose commercial license includes the OPLS3e force field.

For extremely large-scale simulations of over 100 million atoms, the GENESIS software package should also be considered.¹⁷⁸

1.4.5 Consider your boundary conditions and box size

As it is impractical to perform a simulation of the entire universe, careful consideration must be applied to where in space a simulation ends and how these potentially discontinuous boundaries are treated. In biomolecular simulation with explicit solvent, periodic boundaries are used almost universally as they are simple to implement and produce minimal artefacts compared to alternatives like walls or open boundaries.

With periodic boundary conditions, the system is constructed in a triclinic box described by three box vectors. During simulation, the box is treated as though surrounded by copies of itself. Each atom has its copies separated from itself by integer multiples of the box vectors, and when one atom leaves the box in one direction it re-enters the box from the opposite side. Thus, the system is simulated in an infinite crystal lattice. This imposes some structure on the system as it is unable to move in ways that break this symmetry. These lattice errors can be substantial, especially when the net charge of the system changes over the course of the simulation;^{179–181} however, the significance of the artefact scales with box size and, for more typical simulations, is extremely mild for even moderately large boxes.

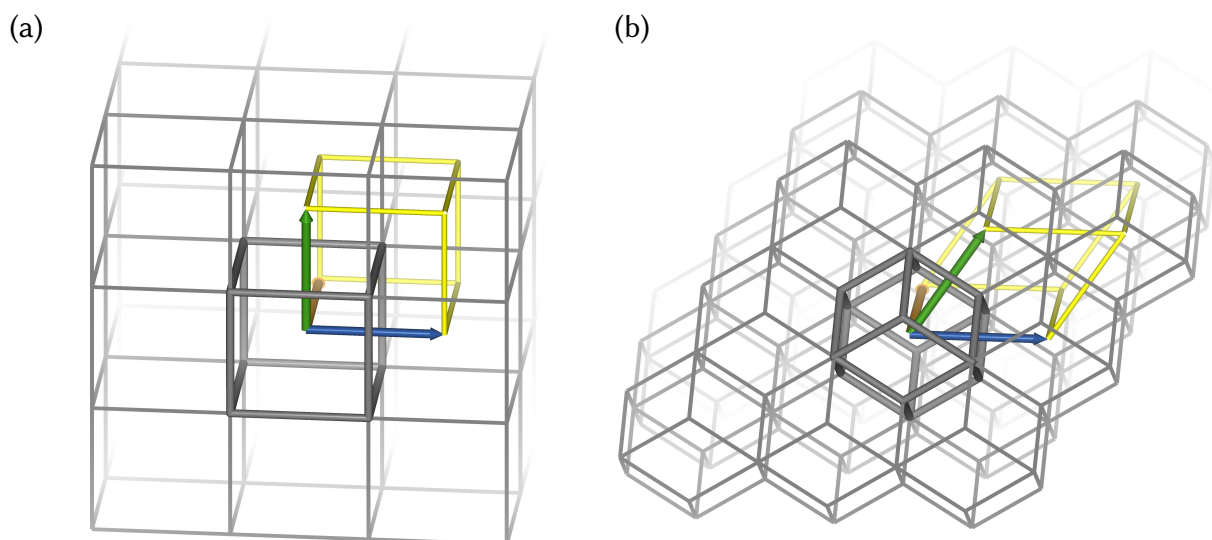


Figure 1.5: Periodic boundary conditions with (a) cubic and (b) rhombic dodecahedral boxes. Though the distance between a point and its image in the neighbouring cell is identical for both boxes, the rhombic dodecahedron has only about 71% of the cube's volume. Blue, orange and green arrows indicate box vectors \mathbf{a} , \mathbf{b} and \mathbf{c} . Grey lines indicate the edges of the box in its polyhedral representation. A single box is emphasised for clarity. Integer linear combinations of the box vectors translate coordinates from one box to another, and when centred on the origin the vectors outline one corner of the box's triclinic representation (yellow lines) so that the centres of the polyhedra correspond to the vertices of the triclinic representation. Note that for the cube, the polyhedral and triclinic representations are identical, while for the rhombic dodecahedron they appear very different.

Periodic boundary conditions touch on almost every aspect of the simulation. The distance between two atoms is not uniquely defined, as both atoms exist in every box; by convention, the shortest such distance is used, which may not be the one in which both atoms share a box. Long-range electrostatic interactions are commonly calculated with Ewald methods such as PME³⁸ that take advantage of periodic boundary conditions to efficiently compute every pairwise interaction in the box — and throughout the virtual crystal. If a constant pressure is required, the box size can even change over the course of the simulation, and this information must be propagated throughout the software.

There are also an infinite number of equivalent ways that the system of atoms can be represented (see the grey and yellow boxes in figure 1.5). A common concern for new biophysicists is that bonds can appear to be broken across boundaries in the output trajectory, when in fact the output is simply an inconvenient representation of a physically correct system. Codes provide tools to change between these representations, but producing comprehensible representations of systems of multiple solutes can be difficult.

Because of this wide impact, the initial box size must be chosen carefully. Too small, and lattice artefacts become significant; too large, and simulation time is wasted on uninteresting solvent molecules. A common rule-of-thumb is that the solute should be separated from

its periodic images by twice the non-bonded cut-off distance, or about 2 nm. It should be noted that this condition being satisfied at the start of the simulation does not guarantee that it will remain satisfied. Most importantly, proteins tumble on MD time-scales; this can be worked around by aligning the solute's longest axis with the box's shortest when computing the solvent buffer. Additionally, proteins can dramatically change shape through folding or conformational change, and constant-pressure simulations can experience changes in box size.

The shape of the box is also significant. While a cubic box is simple and traditional, a tumbling protein sweeps out a sphere, and so the corners of the cube are wasted. A rhombic dodecahedron is a triclinic box that can accommodate a spherical solute of the same size and with the same buffer as a cube with only 70.7% of the volume (see figure 1.5). This saves computer time by reducing the number of solvent molecules that must be simulated to fill the box. A rhombic dodecahedron with distance d between an atom and its periodic image has box vectors:

$$\begin{aligned}\mathbf{a} &= \langle d, 0, 0 \rangle \\ \mathbf{b} &= \langle 0, d, 0 \rangle \\ \mathbf{c} &= \langle \frac{1}{2}d, \frac{1}{2}d, \frac{1}{2}\sqrt{2}d \rangle\end{aligned}$$

If the greatest distance between two atoms in the solute is l , then d can be estimated for a cut-off distance r_c :

$$d = l + 2r_c$$

Note that water models often have different densities, and a box containing poorly equilibrated solvent may contract or expand when pressure coupling is applied. In addition, this procedure will produce unnecessarily large boxes for systems that are expected to collapse over the course of the simulation, such as peptides in the extended conformation; for these, $d = l + r_c$ may be more appropriate.

1.4.6 Use an appropriate thermostat and barostat

Thermostats and barostats are the algorithm used to keep the temperature and pressure of a simulation constant. They modify the integrator to change the evolution of some other properties of the system, generally velocity or box size, in order to have deviations in the appropriate thermodynamic variables decay over time.

The choice of thermostat and barostat can have a substantial effect on the accuracy of a simulation.^{19,163,182,183} Because they modify the integrator, they will always have an effect on the dynamics of the system. Some thermostats and barostats do not produce the correct distribution of temperatures and pressures according to the appropriate ensemble, or have a larger impact on dynamics than an alternative.

Simulations and thermodynamic ensembles are characterised by the thermodynamic variables they keep fixed. A simulation with an unmodified leapfrog algorithm has a constant number of particles, volume, and energy and so might be described as NVE. The distribution of states with constant NVE is called the micro-canonical ensemble. Adding a thermostat fixes the temperature by allowing the energy to vary and can be said to be NVT, or in the canonical ensemble. Additionally using a barostat fixes the pressure by freeing the volume, and so a simulation with both modifications is called the NPT ensemble or the isobaric-isothermal ensemble.

Thermostats

Langevin dynamics:^{184,185} Langevin dynamics adds a randomly generated ‘friction’ term to the velocity of each atom during each integration step. This can be used to either model friction with the solvent in implicit solvent simulations, or heat exchange with the surroundings. In the latter case, the friction term becomes a thermostat. Langevin dynamics is well understood and easy to model, and molecular dynamics without a thermostat is often thought of as Langevin dynamics in the zero-friction limit.⁴¹

Canonical Sampling through Velocity Rescaling thermostat:^{186–188} The CSVr thermostat, also known as the Bussi or Bussi-Donadio-Parrinello thermostat, stochastically scales all the velocities of the simulation such that the total kinetic energy matches a value drawn from the appropriate ensemble. It can be thought of as Langevin dynamics, tuned to be minimally invasive, with a single random variable per step rather than one for each atom.¹⁸⁷ It has been shown to be extremely efficient, reproducing canonical distribution temperatures with minimal disruption of dynamics and only one tunable parameter,^{183,189} and is a great choice for all simulations.

Nosé-Hoover chains:^{190–192} The Nosé-Hoover thermostat treats temperature as a virtual degree of freedom with its own equation of motion, and couples it to the velocities of the system.^{190,191} This was found to inhibit ergodicity for some systems and so was extended by Martyna and co-workers into a chained approach,¹⁹² which may still be problematic in replica

exchange.¹⁹³ It has the unique but dubious benefit of being deterministic, but is much more complex in its implementation than other temperature control algorithms.

Berendsen thermostat:¹⁹⁴ The Berendsen thermostat naïvely rescales velocities so that deviations in temperature decay at a user-configurable rate. It does not produce temperatures drawn from the canonical distribution, but is historically important as a predecessor to the CSVR thermostat that shares its strengths while being thermodynamically correct. Despite its violation of the equipartition principle,¹⁸³ the many erroneous artefacts it has been found to introduce,^{19,163,183,195} and its inability to produce the correct ensemble, the Berendsen thermostat is still in wide use. This may occasionally be appropriate for initially equilibrating systems when the available correct thermostats exhibit pathological convergence behaviour. In modern simulation packages, however, the Berendsen thermostat has been superseded by CSVR and Nosé-Hoover chains and should not be used.

Andersen thermostat:¹⁹⁶ The Andersen thermostat randomly reassigns the velocity of a single particle each integration step to a velocity drawn from the appropriate ensemble. In effect, this simulates collisions with the heat bath. While this is a simple method that produces the correct temperature distribution, it dramatically slows down kinetics — and therefore sampling — as particles all ‘forget’ their velocities over some short time period.

Alternate temperature control schemes: It has been proposed to use the solvent as a heat bath that doesn’t disturb dynamics at all, either by placing a thermostat only on the solvent^{182,197} or perhaps neglecting the thermostat altogether.¹⁹⁸ These approaches have not proved popular and may be under-tested.

Barostats

Monte Carlo barostat:^{199,200} The Monte Carlo barostat is a family of barostats in which the box vectors are resized randomly, and then this change is either accepted or rejected according the Metropolis criterion.³⁴ It is a simple barostat with rapid convergence properties that produces the correct ensemble. However, it abandons any notion of realistic pressure dynamics.

Berendsen barostat:¹⁹⁴ The Berendsen barostat scales the positions, velocities, and box vectors so that deviations from the target pressure vanish over some configurable time period. Like the Berendsen thermostat, this is simple and efficient but does not produce the correct

ensemble. Its efficient convergence and stability makes it a popular choice in system equilibration, where rapid changes in box size are required, but it should not be used when collecting data.²⁰¹

Parrinello-Rahman²⁰² and Martyna-Tuckerman-Tobias-Klein^{203,204} barostats: Similarly to the Nosé-Hoover thermostat, these barostats treat the pressure as an additional virtual degree of freedom coupled to box size and the positions of atoms. The MTTK barostat is an incremental improvement over PR;^{205,206} however, both are in wide use during data collection, especially when pressure dynamics are important. Both exhibit poor convergence properties, as large changes in box size can lead to oscillations and instability, and so should not be used when equilibrating the box size.

Stochastic Cell Rescaling barostat:²⁰⁷ Similarly to the CSVR thermostat, stochastic cell rescaling incorporates random noise in rescaling the box to adopt the advantages and simplicity of the Berendsen barostat while being thermodynamically correct. While it has only recently been developed and is not yet widely available in efficient codes, it is likely to be a robust barostat for the same reasons that CSVR is. It is an anticipated feature of GROMACS version 2021.

1.4.7 Enhance sampling

With the understanding that MD is first and foremost a convenient sampling technique for condensed-phase biomolecular systems, it becomes clear that simple integration algorithms may not be the most efficient. Enhanced sampling methods modify the basic dynamics of MD in various ways to accelerate sampling, either without altering the produced ensemble or by altering it in controlled ways that can be corrected in analysis. A wide variety of these methods were recently reviewed by Yang et al.²⁰⁸

Simulated annealing^{209,210}

Simulated annealing is a widely applicable optimisation algorithm inspired by physics and later applied to molecular dynamics. At low temperatures such as the temperature of interest, sampling is in the correct ensemble but is very slow as high-energy transition barriers are only sampled rarely. At higher temperatures, sampling is faster but in the incorrect ensemble. Indeed, increasing the temperature is equivalent to decreasing the scale of every potential term in the force field, making it flatter and accelerating movement. Simulated annealing exploits this by performing sampling at the temperature of interest for some time, and then heating

the system up to accelerate movement around the free energy surface. When the simulation is at a high temperature, data is not recorded, and sampling resumes when the system cools. This simple approach enhances sampling where it is slowed by having to negotiate energetic barriers in configuration space, as in biophysics.

Parallel tempering/temperature replica exchange^{211,212}

Parallel tempering, also known as temperature replica exchange (REMD), can be thought of as a parallel version of simulated annealing. Simulations of the same system are run in parallel at a range of temperatures like rungs on a ladder. With some frequency, adjacent replicas are allowed to swap coordinates according to the Metropolis criterion; that is, according to the Boltzmann probability that those coordinates would be found in each other's ensemble. The replica at the temperature of interest is then used in analysis. It forms a discontinuous trajectory, with jumps around configuration space accelerated by trips to higher temperatures, but is entirely drawn from the appropriate ensemble and has dramatically less correlation between frames.

Efficiency of sampling can be assessed by the efficiency of movement around the temperature ladder,²¹³ so temperatures must be chosen that are close enough together that fluctuations in energy overlap. However, higher maximum temperatures also lead to greater efficiency, so many replicas may be needed.²¹⁴ Replica exchange is relatively simple to implement, and does not require specific knowledge about the system beyond the fact that sampling is slow because of enthalpic barriers.

When setting up a replica exchange simulation, a number of parameters must be chosen; the exchange frequency, the number of replicas, and the temperature ladder. There is some debate about how to choose the exchange frequency. In theory, exchange should be attempted very frequently for efficiency and is thermodynamically correct when done properly;^{215,216} however, it has been suggested that in practice artifacts are introduced with exchange times below 1 ps,^{217,218} though this has not been shown with modern thermostats. The temperature ladder and number of replicas is simpler, however; the temperature ladder should be a geometric progression (though see also Gront et al.²¹⁹ and Gront²²⁰) starting at the temperature of interest and ending before the entropy dominates the free energy of interest, and the number of replicas should be chosen to maintain a respectable exchange acceptance ratio of about 20 – 30 %.^{214,221,222}

Hamiltonian replica exchange

Parallel tempering may be generalised by considering that temperature is not the only property that can affect sampling. Hamiltonian replica exchange uses a ladder with a modified energy function, or Hamiltonian, such that the bottom rung experiences normal sampling and the top rung's sampling is substantially accelerated. The accelerated Hamiltonian is gradually turned on over the other rungs to permit exchange.

Since sampling only occurs on the unmodified bottom rung, any Hamiltonian that accelerates sampling and has an overlapping Boltzmann distribution can be used, even if it is physically nonsense. This allows the use of knowledge about the system to accelerate sampling while requiring many fewer replicas than parallel tempering. For example, heating up the water solvating a protein does not improve sampling, so parallel tempering can be improved by turning it into a H-REMD scheme that simply heats the protein.^{223,224} As many fewer atoms are heated in this scheme, the same sampling increase is achieved with many fewer replicas. Similarly, a model of the free energy surface along a slow coordinate can be used to sample other faster coordinates.^{225,226}

Bias methods

Bias methods are a large family of methods^{227–234} that vary greatly in detail but all work according to the same biophysical principles. Bias methods target a model of the free energy surface in terms of some coarse-grained low-dimensional coordinate system, or collective variable. Sampling along this variable is slow in simple MD because the trajectory dwells in regions of low free energy and cannot traverse large free energy barriers. A model of the probability distribution along the collective variable is built up on the fly from sampling done over the course of the simulation, and this is transformed into an energetic bias that is applied to the system. This boosts the system out of areas it has already explored and encourages it into new regions of phase space. Over time, as the model of the free energy surface becomes more accurate, sampling converges to uniformity across the collective variable.

Bias methods accelerate sampling along the collective variables they are applied to. If those collective variables describe the slowest degrees of freedom, then bias methods accelerate overall sampling. It should be noted that bias methods do not produce unbiased trajectories; instead, they produce a trajectory with a known time-dependent bias, as well as an unbiased model of the free energy surface in collective variable space. This model can be used to re-weight the trajectory, or as a bias for H-REMD.²²⁵ Bias methods are also used alongside replica exchange methods to accelerate degrees of freedom for which collective variables are unavailable.²³⁵

These methods rely heavily on the choice of collective variable. If the wrong collective variable is chosen, stable convergence may be slow or the converged bias may not accelerate sampling overall. While this allows a researcher to leverage knowledge about the system to accelerate sampling, it is also prone to failure. Work has been done in the last several years to simplify and automate the selection of collective variables and even build them on the fly.^{236–239}

The most well known bias method is well-tempered metadynamics,²³² in which Gaussians of slowly reduced size are added to the bias at the system's current position with some frequency. Before it, umbrella sampling²²⁷ involved placing harmonic biases along a collective variable and simulating the system independently in each in a sort of metadynamics-by-hand. Adaptive biasing force²⁴⁰ is a bias method inspired by metadynamics and integrated into the Amber software package that uses a force-based spline as a bias. Supported similarly by GROMACS, the Accelerated Weight Histogram (AWH) method²³³ uses a histogram to construct the bias. AWH differs notably by starting with a sweep across the entire collective variable space to rapidly converge a rough estimate of the bias that is then further optimised in additional sweeps. Both methods have a reduced overhead compared to metadynamics in their respective packages. Variationally Enhanced Sampling²³⁴ formulates the construction of the bias as a minimisation problem, allowing great flexibility in the choice of biasing function. AWH and VES both support arbitrary target distributions defined in terms of their collective variables, allowing them to provide fine-grained corrections to an existing model.

Weighted Ensemble method

The Weighted Ensemble method is an extremely flexible, unbiased method that can be applied either to explore an unknown energy landscape²⁴¹ or to sample along a collective variable. The method exploits the fact that a rigorous re-sampling with re-weighting of a statistical ensemble preserves the distribution of the ensemble.²⁴² An ensemble of n trajectories of the same system, each with an initial weight of $1/n$, are begun.²⁴³ The simulations are run for a short time, then stopped. Trajectories that have reached interesting regions of configuration space are cloned, and their weights are divided among the children, while trajectories sampling regions of low free energy are culled and their weights concentrated on the survivors. This procedure is repeated and produces a swarm of weighted trajectory segments, which leads to an ensemble that converges rapidly in the ways defined as interesting. There is enormous flexibility in how simulations are re-sampled and which features are considered interesting,²⁴² and the method is easily applied to non-equilibrium systems.²⁴⁴

The weighted ensemble method is very similar to the popular Markov state model adaptive sampling approaches,²⁴⁵ but does not rely on the assumption of Markovian dynamics. This

makes it simpler, more flexible, and less error prone, and allows information coming from within a state to be retained in the final model. The method has been applied to an 11 minute process in 6 μ s of simulation⁵⁴ and was reviewed recently by Zuckerman *et al.*²⁴⁶

1.5 Conclusion

MD has rapidly developed in the last decade since the application of GPUs to the problem of molecular simulation. Force fields have reached the point of reliability for soluble folded proteins, and are rapidly improving. Enhanced sampling methods are easy to use and continue to be worked on. As such, MD is reaching a level of maturity that permits its wide use by non-experts. Given the computational expense of performing an MD simulation, the complexity of the technique, and the plethora of features available from today's software, users should be familiar with the goals and procedures of MD simulation. MD is sampling, not simulation; it is helpful to think of MD as a stochastic sampling method that produces states from an equilibrium ensemble, not a precise simulation of everything a system does from a given start point. This mindset reduces expectations that MD cannot yet meet, assists in choosing optimal parameters and methods, and emphasises the importance of a statistically rigorous interpretation of results.

Chapter 2

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

2.1 Preface

Biological organisms are nothing more than complex chemical systems that maintain themselves in an out-of-equilibrium steady state. A complete understanding of biology therefore necessarily involves a chemical picture of the cell. To that end, it is essential that biologists develop minimally invasive tools to identify and locate chemicals with great specificity and spatio-temporal resolution. This is especially true in fields that involve chemical messaging, such as neuroscience, where the precise location, timing, and nature of a chemical response to stimuli holds a lot of essential information.

Fluorescence is close to a holy grail of chemical sensing. Fluorescence can be used to surpass the diffraction limit of optical light²⁴⁷ allowing for excellent spatial resolution. Fluorescence can happen in picoseconds, which promises great theoretical temporal resolution, though effects like photobleaching can limit this in practice. Fluorescent probes themselves are non-invasive, as they are simply biologically benign chemicals, and fluorescent microscopes can be used on live specimens even in complex animals.²⁴⁸ In addition, fluorescent probes of different colours can be combined in a study to visualise multiple analytes simultaneously. Because of these properties, fluorescence is an essential tool in the biologist's kit that allows real-time imaging of subcellular chemical structures in a living organism.

Proteins provide a nanoscale platform with extraordinary chemical flexibility. Natural proteins possess exquisite sensitivity and specificity for their substrates,^{249,250} and directed

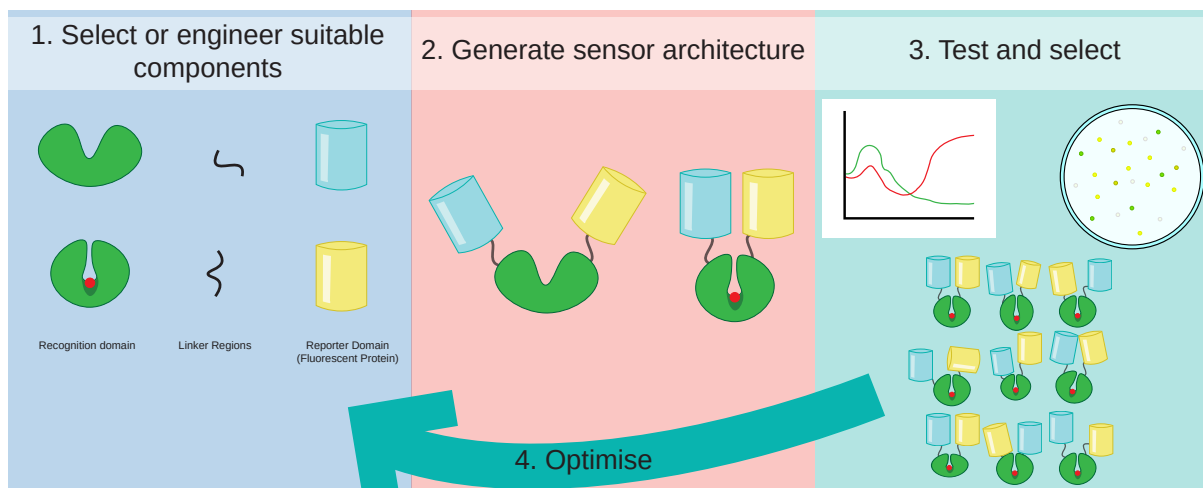


Figure 2.1: Graphical abstract for *Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors*

evolution grants the ability to alter and improve these properties in a timely manner.²⁵¹ Fluorescent proteins are available with a wide range of colours and other properties, and do not require chemical modification.^{252,253} While fluorescent sensors are not available from nature, the combination of fluorescent and sensing proteins into fusion proteins allows creation of biosensors that offer all of these benefits. In addition, like natural proteins, fluorescent fusion proteins are genetically encodable, and offer the possibility of expression by the system under study.

Fluorescent biosensors are a great testbed for new technologies in protein engineering. Rather than requiring complicated chemical assays, the quality of the sensor can be evaluated rapidly and easily through fluorescence. Advancements in engineering techniques at the whole-protein level such as circular permutation²⁵⁴ and PROSS²⁵⁵ can be applied to individual domains, while improvements in optimisation of binding sites are applicable to sensing domains. In general, a sensor can be put together quickly by considering a combination of existing domains, and then optimised over time.

In this review, we reviewed the state of the art and future directions in fluorescent biosensor engineering.

2.2 Statement of contribution

I declare that the research presented in this chapter represents original work that I carried out during my candidature at the Australian National University, except for contributions to multi-author papers incorporated in the chapter where my contributions are specified in this Statement of Contribution.

2.2.1 Publication status

This manuscript has been published with the title *Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors* in a themed issue of the journal **Current Opinion in Structural Biology** (2019, 57:31-38). The formatted article is reproduced in this chapter.

2.2.2 Authorship and contribution

The manuscript was authored by Joe A. Kaczmariski, Joshua A. Mitchell (the author), Matthew A. Spence, Vanessa Vongsouthi and Colin J. Jackson. JAK, JAM, MAS and VV contributed equally to the work. I contributed the section entitled ‘Designing and modelling linkers’, figure 4, proofing and editing of the entire work, and preparation of the manuscript for submission.



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

Joe A Kaczmariski¹, 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.

Current Opinion in Structural Biology 2019, 57:31–38

This review comes from a themed issue on **Engineering and design: synthetic signaling**

Edited by **Andreas Möglich** and **Harald Janovjak**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online 27th February 2019

<https://doi.org/10.1016/j.sbi.2019.01.013>

0959-440X/© 2019 Elsevier Ltd. All rights reserved.

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 FPs 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 physiochemical 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

Figure 1

				
Component to be engineered	Reporter domain (Fluorescent protein)	Recognition domain	Linkers	Complete Sensor
Desired property	Excitation/emission wavelengths Quantum Yield Photostability Maturation times	Specificity Affinity Thermostability	Length Rigidity	Specificity Intensiometric/Ratiometric? Response time Orthogonality Dynamic range
Engineering method	Circular permutation Site-directed mutagenesis Unnatural amino acid incorporation <i>De novo</i> design	<i>De novo</i> design Binding pocket grafting Computational design Statistical coupling analysis Consensus sequence design Ancestral protein reconstruction	Linker composition: [GGS] _n , [EAAAK] _n Computational design: Random coil models, molecular dynamics	Rational design Laboratory evolution

Current Opinion in Structural Biology

Biosensor properties and engineering approaches.

Engineering occurs at all stages of the biosensor design process. Individual components (recognition domain, linkers and reporter domains) can be engineered separately to achieve desired properties, but they normally need to be further optimized in context of the complete sensor. Engineering strategies range from structure-guided rational design of binding sites, to the high-throughput screening of large libraries of biosensor variants.

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

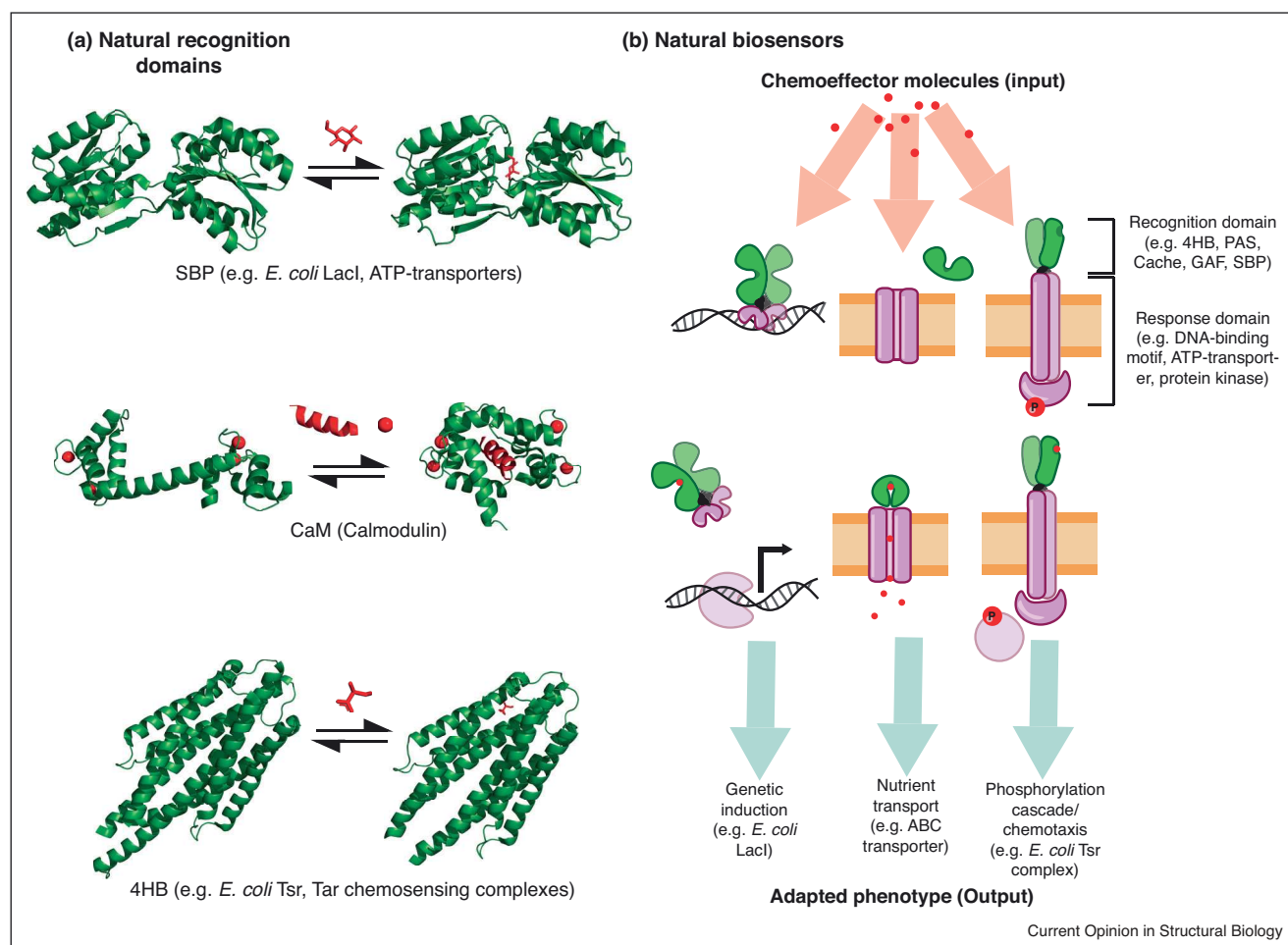
Fluorescence-based biosensor designs can be classified based on the number of integrated FPs, or by the type of optical output they produce. Single-FP sensors (Figure 3a) link the natural or engineered sensitivity of the fluorophores within the FPs to changes in environmental variables ('intrinsic'), 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 intensimetric, 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



Natural biosensors and small molecule recognition domains.

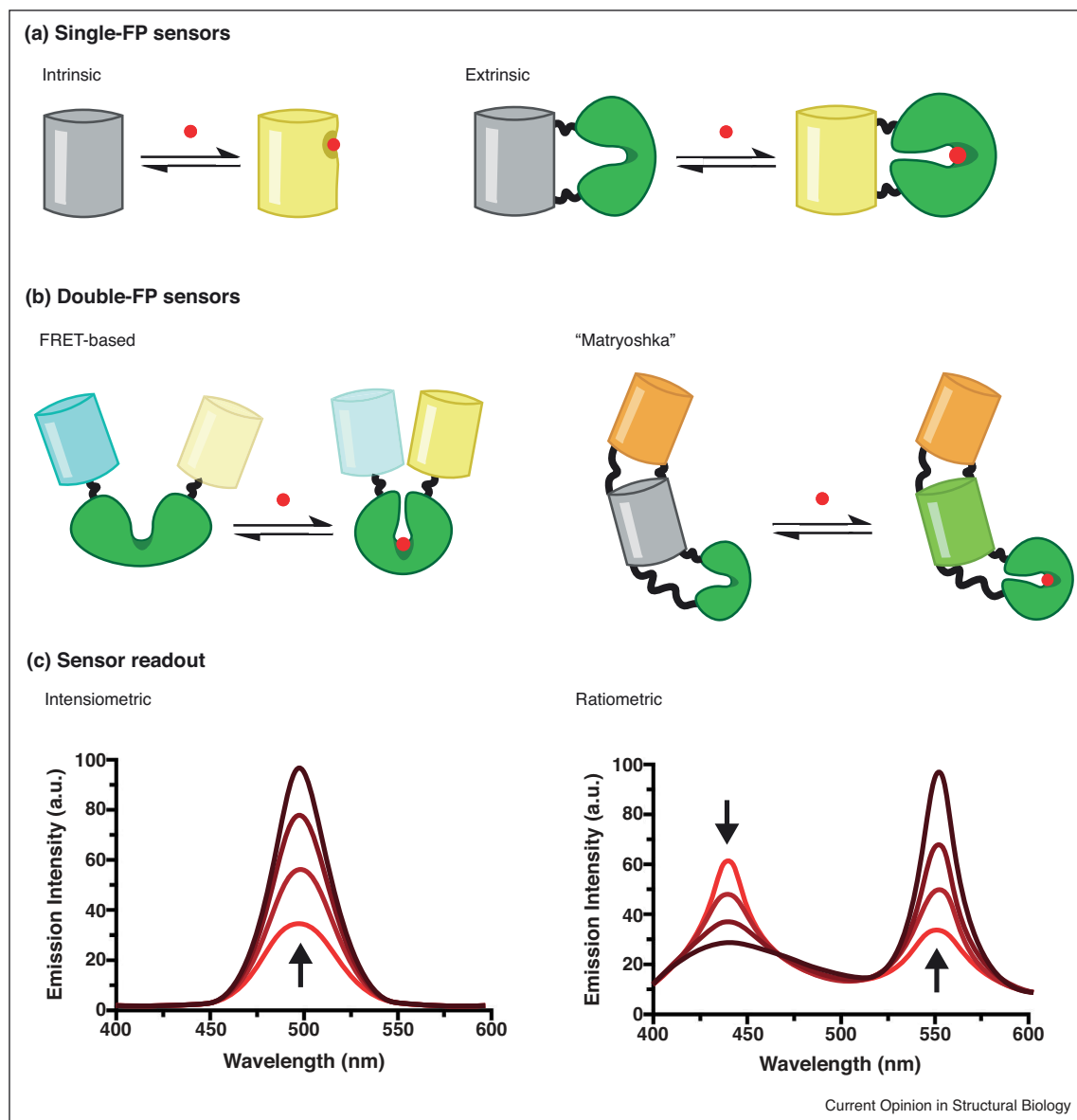
(a) Common recognition domain folds. Both SBPs and CaM undergo large conformational changes in the presence of their cognate ligands and are regularly used as starting scaffolds for novel fluorescent biosensors. 4HB domains are among the most abundant and important sensory proteins in natural biosensors. **(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.

sensitive to spectral overlap, distance and orientation between the donor and acceptor FPs, designing FRET sensors with high dynamic range and SNRs can be difficult [14]. Strategies to improve FRET-efficiency include engineering linkers (see **Linkers section**), the insertion of one FP into the recognition domain (tighter allosteric linkage and decrease in rotational averaging) [15], changing the sequential order of the donor and acceptor FPs [16] or engineering the spectral properties of the FPs [17].

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

Figure 3



Fluorescence-based biosensor architectures and readouts.

Common sensor architectures of (a) single-FP and (b) double-FP sensors. (c) Example readouts from intensiometric (left) and ratiometric (right) biosensors, showing the change in emission intensity as a function of increasing analyte concentration (light to dark red).

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 L-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[•]].

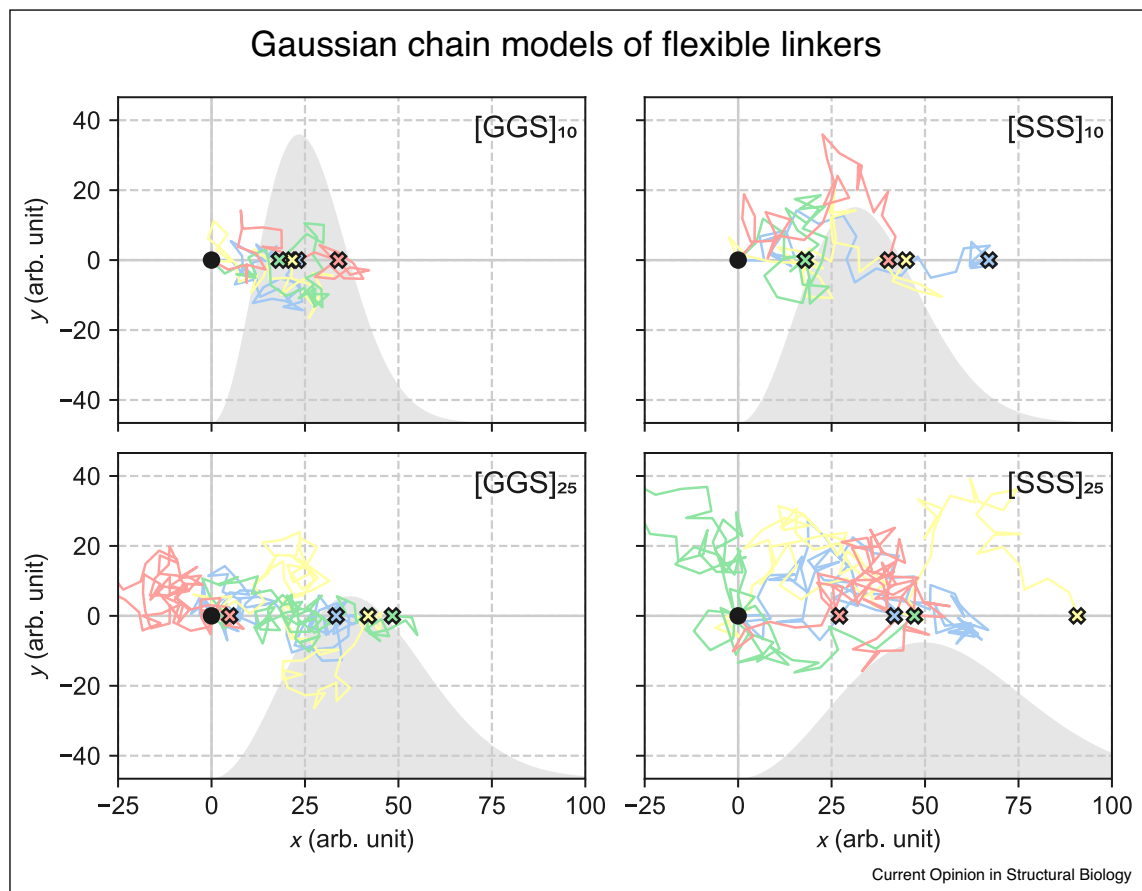
Flexible linkers, epitomized by glycine-serine repeats such as [GGS]_n, are largely unstructured and are thought to tether the domains together like a rope, only minimally constraining their movement beyond keeping the fused termini within some distance of each other. Rigid linkers, most commonly the 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 (10⁶–10⁹ 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

Figure 4



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_∞ which quantifies the flexibility of the chain ($c_\infty = 1.9$ for [GGS] $_n$, $c_\infty = 3.4$ for [SSS] $_n$). 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 wide 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

CJJ 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

1. Ortega Á, Zhulin IB, Krell T: **Sensory repertoire of bacterial chemoreceptors**. *Microbiol Mol Biol Rev* 2017, **81**.
2. Rogers JK, Guzman CD, Taylor ND, Raman S, Anderson K, Church GM: **Synthetic biosensors for precise gene control and real-time monitoring of metabolites**. *Nucleic Acids Res* 2015, **43**:7648-7660.
3. Juárez JF, Lecube-Azpeitia B, Brown SL, Johnston CD, Church GM: **Biosensor libraries harness large classes of binding domains for construction of allosteric transcriptional regulators**. *Nat Commun* 2018, **9**.
- The authors fuse ligand-binding domains with DNA-binding domains to create novel classes of transcription factors. They produced two novel benzoate-sensing transcription-factor based biosensors.
4. Taylor ND, Garruss AS, Moretti R, Chan S, Arbing MA, Cascio D, Rogers JK, Isaacs FJ, Kosuri S, Baker D *et al.*: **Engineering an allosteric transcription factor to respond to new ligands**. *Nat Methods* 2016, **13**:177-183.
- 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.
5. Mannan AA, Liu D, Zhang F, Oyarzún DA: **Fundamental design principles for transcription-factor-based metabolite biosensors**. *ACS Synth Biol* 2017, **6**:1851-1859.
6. Cheng F, Tang X-L, Kardashliev T: **Transcription factor-based biosensors in high-throughput screening: advances and applications**. *Biotechnol J* 2018, **13**:1700648.
7. Llopis J, McCaffery JM, Miyawaki A, Farquhar MG, Tsien RY: **Measurement of cytosolic mitochondrial, and Golgi pH in single living cells with green fluorescent proteins**. *Proc Natl Acad Sci* 1998, **95**:6803-6808.
8. Jayaraman S, Haggie P, Wachter RM, Remington SJ, Verkman AS: **Mechanism and cellular applications of a green fluorescent protein-based halide sensor**. *J Biol Chem* 2000, **275**:6047-6050.
9. Kost LA, Nikitin ES, Ivanova VO, Sung U, Putintseva EV, Chudakov DM, Balaban PM, Lukyanov KA, Bogdanov AM: **Insertion of the voltage-sensitive domain into circularly permuted red fluorescent protein as a design for genetically encoded voltage sensor**. *PLoS One* 2017, **12**:e0184225.
10. Ast C, De Michele R, Kumke MU, Frommer WB: **Single-fluorophore membrane transport activity sensors with dual-emission read-out**. *eLife* 2015, **4**.
11. Cho JH, Swanson CJ, Chen J, Li A, Lippert LG, Boye SE, Rose K, Sivaramakrishnan S, Chuong CM, Chow RH: **The GCaMP-R family of genetically encoded ratiometric calcium indicators**. *ACS Chem Biol* 2017, **12**:1066-1074.
12. Ast C, Foret J, Oltrogge LM, Michele RD, Kleist TJ, Ho C-H, Frommer WB: **Ratiometric Matryoshka biosensors from a nested cassette of green- and orange-emitting fluorescent proteins**. *Nat Commun* 2017, **8**.
- 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 intensimetric sensors into ratiometric sensors.
13. Ding Y, Li J, Enterina JR, Shen Y, Zhang I, Tewson PH, Mo GCH, Zhang J, Quinn AM, Hughes TE *et al.*: **Ratiometric biosensors based on dimerization-dependent fluorescent protein exchange**. *Nat Methods* 2015, **12**:195-198.
14. Bajar B, Wang E, Zhang S, Lin M, Chu J: **A guide to fluorescent protein FRET pairs**. *Sensors* 2016, **16**:1488.
15. Deuschle K, Okumoto S, Fehr M, Looger LL, Kozhukh L, Frommer WB: **Construction and optimization of a family of genetically encoded metabolite sensors by semirational protein engineering**. *Protein Sci* 2005, **14**:2304-2314.
16. Ohta Y, Kamagata T, Mukai A, Takada S, Nagai T, Horikawa K: **Nontrivial effect of the color-exchange of a donor/acceptor pair in the engineering of Förster Resonance Energy Transfer (FRET)-based indicators**. *ACS Chem Biol* 2016, **11**:1816-1822.
17. Klarenbeek J, Goedhart J, van Batenburg A, Groenewald D, Jalink K: **Fourth-generation epac-based FRET sensors for cAMP feature exceptional brightness photostability and dynamic range: characterization of dedicated sensors for FLIM for ratiometry and with high affinity**. *PLoS One* 2015, **10**:e0122513.
18. Goedhart J, Stetten D, von, Noirclerc-Savoye M, Lelimosin M, Joosen L, Hink MA, Weeren L, van, Gadella TWJ, Royant A: **Structure-guided evolution of cyan fluorescent proteins towards a quantum yield of 93%**. *Nat Commun* 2012, **3**.
19. Bajar BT, Wang ES, Lam AJ, Kim BB, Jacobs CL, Howe ES, Davidson MW, Lin MZ, Chu J: **Improving brightness and photostability of green and red fluorescent proteins for live cell imaging and FRET reporting**. *Sci Rep* 2016, **6**.
20. Shui B, Wang Q, Lee F, Byrnes LJ, Chudakov DM, Lukyanov SA, Sondermann H, Kotlikoff ML: **Circular permutation of red fluorescent proteins**. *PLoS One* 2011, **6**:e20505.
21. Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, Mao B, Fought GW, Lee MY, Gagnon LA *et al.*: **De novo design of a fluorescence-activating β -barrel**. *Nature* 2018, **561**:485-491.
- The authors use Rosetta for the first *de novo* design of 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.
22. Chu J, Oh Y, Sens A, Ataie N, Dana H, Macklin JJ, Laviv T, Welf ES, Dean KM, Zhang F *et al.*: **A bright cyan-excitable orange fluorescent protein facilitates dual-emission microscopy and enhances bioluminescence imaging *in vivo***. *Nat Biotechnol* 2016, **34**:760-767.
23. Huang P-S, Boyken SE, Baker D: **The coming of age of *de novo* protein design**. *Nature* 2016, **537**:320-327.
24. Cho J-H, Swanson CJ, Chen J, Li A, Lippert LG, Boye SE, Rose K, Sivaramakrishnan S, Chuong C-M, Chow RH: **The GCaMP-R family of genetically encoded ratiometric calcium indicators**. *ACS Chem Biol* 2017, **12**:1066-1074.
25. Thestrup T, Litzlbauer J, Bartholomäus I, Mues M, Russo L, Dana H, Kovalchuk Y, Liang Y, Kalamakis G, Laukat Y *et al.*: **Optimized ratiometric calcium sensors for functional *in vivo* imaging of neurons and T lymphocytes**. *Nat Methods* 2014, **11**:175-182.
26. Zhang WH, Herde MK, Mitchell JA, Whitfield JH, Wulff AB, Vongsouthi V, Sanchez-Romero I, Gulakova PE, Minge D, Breithausen B *et al.*: **Monitoring hippocampal glycine with the computationally designed optical sensor GlyFS**. *Nat Chem Biol* 2018, **14**:861-869.
- 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.

27. Marvin JS, Borghuis BG, Tian L, Cichon J, Harnett MT, Akerboom J, Gordus A, Renninger SL, Chen T-W, Bargmann CI *et al.*: **An optimized fluorescent probe for visualizing glutamate neurotransmission.** *Nat Methods* 2013, **10**:162-170.
 28. Rizza A, Walia A, Lanquar V, Frommer WB, Jones AM: ***In vivo* gibberellin gradients visualized in rapidly elongating tissues.** *Nat Plants* 2017, **3**:803-813.
 29. Brun MA, Tan K-T, Nakata E, Hinner MJ, Johnsson K: **Semisynthetic fluorescent sensor proteins based on self-labeling protein tags.** *J Am Chem Soc* 2009, **131**:5873-5884.
 30. DeGrave AJ, Ha J-H, Loh SN, Chong LT: **Large enhancement of response times of a protein conformational switch by computational design.** *Nat Commun* 2018, **9**.
 31. Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon APM, Wyss M: **The consensus concept for thermostability engineering of proteins: further proof of concept.** *Protein Eng* 2002, **15**:403-411.
 32. Clifton BE, Jackson CJ: **Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins.** *Cell Chem Biol* 2016, **23**:236-245.
 33. Whitfield JH, Zhang WH, Herde MK, Clifton BE, Radziejewski J, Janovjak H, Henneberger C, Jackson CJ: **Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction.** *Protein Sci* 2015, **24**:1412-1422.
 34. Scheib U, Shanmugaratnam S, Farias-Rico JA, Höcker B: **Change in protein-ligand specificity through binding pocket grafting.** *J Struct Biol* 2014, **185**:186-192.
 35. Limsakul P, Peng Q, Wu Y, Allen ME, Liang J, Remacle AG, Lopez T, Ge X, Kay BK, Zhao H *et al.*: **Directed evolution to engineer monobody for FRET biosensor assembly and imaging at live-cell surface.** *Cell Chem Biol* 2018, **25**:370-379.e4.
 36. Okada S, Ota K, Ito T: **Circular permutation of ligand-binding module improves dynamic range of genetically encoded FRET-based nanosensor.** *Protein Sci* 2009, **18**:2518-2527.
 37. Nagai T, Yamada S, Tominaga T, Ichikawa M, Miyawaki A: **Expanded dynamic range of fluorescent indicators for Ca^{2+} by circularly permuted yellow fluorescent proteins.** *Proc Natl Acad Sci U S A* 2004, **101**:10554-10559.
 38. Fritz RD, Letzelter M, Reimann A, Martin K, Fusco L, Ritsma L, Ponsioen B, Fluri E, Schulte-Merker S, van Rheenen J *et al.*: **A versatile toolkit to produce sensitive FRET biosensors to visualize signaling in time and space.** *Sci Signal* 2013, **6**:rs12.
 39. Merckx M, Golynskiy MV, Lindenburg LH, Vinkenburg JL: **Rational design of FRET sensor proteins based on mutually exclusive domain interactions.** *Biochem Soc Trans* 2013, **41**:1201-1205.
 40. Lindenburg LH, Malisauskas M, Sips T, van Oppen L, Wijnands SPW, van de Graaf SFJ, Merckx M: **Quantifying stickiness: thermodynamic characterization of intramolecular domain interactions to guide the design of Förster resonance energy transfer sensors.** *Biochemistry* 2014, **53**:6370-6381.
 41. van Rosmalen M, Krom M, Merckx M: **Tuning the flexibility of glycine-serine linkers to allow rational design of multidomain proteins.** *Biochemistry* 2017, **56**:6565-6574.
- 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.
42. Li G, Huang Z, Zhang C, Dong B-J, Guo R-H, Yue H-W, Yan L-T, Xing X-H: **Construction of a linker library with widely controllable flexibility for fusion protein design.** *Appl Microbiol Biotechnol* 2015, **100**:215-225.
 43. Stein V, Alexandrov K: **Protease-based synthetic sensing and signal amplification.** *Proc Natl Acad Sci* 2014, **111**:15934-15939.
 44. Younger AKD, Su PY, Shepard AJ, Udani SV, Cybulski TR, Tyo KEJ, Leonard JN: **Development of novel metabolite-responsive transcription factors via transposon-mediated protein fusion.** *Protein Eng Des Sel* 2018, **31**:55-63.
 45. Liu B, Åberg C, van Eerden FJ, Marrink SJ, Poolman B, Boersma AJ: **Design and properties of genetically encoded probes for sensing macromolecular crowding.** *Biophys J* 2017, **112**:1929-1939.
 46. van Dongen EMWM, Evers TH, Dekkers LM, Meijer EW, Klomp LWJ, Merckx M: **Variation of linker length in ratiometric fluorescent sensor proteins allows rational tuning of Zn(II) affinity in the picomolar to femtomolar range.** *J Am Chem Soc* 2007, **129**:3494-3495.
 47. Feng J, Jester BW, Tinberg CE, Mandell DJ, Antunes MS, Chari R, Morey KJ, Rios X, Medford JI, Church GM *et al.*: **A general strategy to construct small molecule biosensors in eukaryotes.** *eLife* 2015, **4**.
- 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.
48. Nadler DC, Morgan S-A, Flamholz A, Kortright KE, Savage DF: **Rapid construction of metabolite biosensors using domain-insertion profiling.** *Nat Commun* 2016, **7**:12266.
 49. Brandsen BM, Mattheisen JM, Noel T, Fields S: **A biosensor strategy for *E. coli* based on ligand-dependent stabilization.** *ACS Synth Biol* 2018, **7**:1990-1999.

Chapter 3

Semisynthetic fluorescent biosensors via Rangefinder

3.1 Preface

Solute Binding Proteins (SBPs) are a family of binding proteins described in detail in chapter 2. This large family has a broad range of substrates and forms a two-lobed ‘Venus flytrap’ fold. Unbound, the two lobes form an ‘open’ configuration providing access to the binding site; upon binding, the lobes close around the ligand. This conformational change makes them ideal for sensor development, and their conserved fold allows strategies developed on one SBP to be applied to a wide variety of ligands, but in principle these strategies can be used in any protein where a target event is coupled to a conformational change that changes the distance between two parts of the protein.

While the GFP family has diversified into a wide array of fluorescent proteins with many properties, synthetic fluorophores have access to many more chemistries and are capable of superior performance. Synthetic dyes exist with extraordinary brightness and with almost any desirable spectroscopic properties. Synthetic dyes can be easily attached to solvent-exposed cysteine residues via a thiol-maleimide Michael addition.^{256,257} This allows rapid, efficient and site-specific labelling of proteins, amplifying their fluorescent capabilities.

While labelling of a protein with a single synthetic dye is simple, site-specific labelling of multiple different fluorophores is much more challenging and may require introduction of entire domains^{258–260} or non-canonical amino acids.^{261–263} The ability to precisely locate multiple fluorophores is essential for the production of ratiometric sensors, which use the ratio of two fluorescence peaks as a readout rather than total intensity in order to control for sensor concentration. The Snifit sensor design²⁶⁴ in particular uses two synthetic dyes and has been applied to many analytes.^{265–268}

In my Honours thesis,²⁶⁹ I explored combining a genetically encoded ECFP²⁷⁰ with a solute

binding protein (SBP) labelled with a maleimide dye to produce a semisynthetic biosensor with extraordinary dynamic range and a relatively simple construction. I found that the precise labelling site on the SBP had an enormous impact on dynamic range, and considered coarse-grained MD simulations of the ECFP-SBP construct for justification of these dynamic ranges. Early in my PhD candidature, I formalised and repeated this MD-based methodology as a predictive tool for constructing sensors. This method is published in the book chapter *Method for developing optical sensors using a synthetic dye–fluorescent protein FRET pair and computational modeling and assessment*, and is reproduced in section 3.3. The `simulations.sh` script can be downloaded at <https://bit.ly/jam-simulations>, while the functionality of the `process-data.py` script is included in RangeFinder.

I then considered whether the extensive MD methodology used for these initial results was necessary for prediction. On the basis of the ensemble produced by MD, I designed a geometric method to construct a ‘typical’ location for the ECFP fluorophore based only on the SBP structure. I found that computing dynamic ranges considering only this point produced dynamic range predictions that were just as good as the ensemble, probably because the ensemble average FRET efficiency is similar to the FRET efficiency of a point near the ensemble average location. It is not clear whether this is a general property of FRET efficiencies or something peculiar to this system. With my colleagues Dr. Zhang and Dr. Whitfield, I assessed the performance of this method at predicting dynamic ranges of novel sensors. This work was published as *RangeFinder: a semisynthetic FRET sensor design algorithm* and is reproduced in section 3.2. In addition to the URL provided in the manuscript, the RangeFinder program is available on GitHub at <https://github.com/Yoshanuikabundi/rangefinder>. The methodology is explained in more detail in the supplementary information.

3.1.1 Rangefinder sensors as a modelling testbed

Not only do these dye-SBP-FP semisynthetic sensors combine great performance with easy manufacture, but they also offer a simplified model of genetically encoded FP-SBP-FP sensors. They work in essentially the same way; a conformational change of the SBP associated with binding of the analyte changes the efficiency of FRET excitation transfer between two fluorophores, resulting in a fluorescence intensity ratio that indicates the presence of the analyte. When the number of sensor molecules is large and the analyte concentration is close to the dissociation constant, this ratio becomes indicative of the concentration of analyte. However, dye-SBP-FP sensors replace a FP domain with a much smaller and less dynamic synthetic fluorophore. This allows the dynamics and location of the remaining FP domain to be studied independently of interactions from the former.

Indeed, this simplification was the inspiration behind the MD methodology used in these publications. The volume of a box large enough to fit two domains with an extended linker was substantial, and the complexity of the solvated system made a coarse-grained approach with the MARTINI force field preferable. Unfortunately, once the linker completed its hydrophobic collapse, the strong protein-protein interactions in this force field over-stabilised whatever conformation the two domains formed on first contact. In the paper, this is justified as kinetically controlled dimerisation, but in hindsight it seems more likely to be a reflection of known limitations of the force field.^{126,127} This is supported by the facts that the ensemble does not improve the predictions of the single point method and that interactions between the two proteins should be transient and non-specific given that they are derived not just from different species but from different domains of life.

Despite this failing, these simulations were essential in revising our understanding of protein biosensors. The literature before this paper was published modelled fluorescent protein domains as occupying space in a cone with its point separated from the SBP's terminal residue by a fully extended linker.²⁷¹ These simulations led our group to appreciate the full range of conformations open to a flexible linker and the preference for collapse of fusion proteins in general, as well as the wide variety of orientations that even a short linker permits in its attached domains. Rather than form a cone some distance away from the fused terminus, fluorescent proteins form a cloud around the terminus. This understanding was essential to the development of GlyFS (see section 4.1.2).

3.1.2 The FRET orientation factor

The simulations also highlight that the FRET orientation factor κ^2 is probably an under-appreciated element of sensor design. SBP-based biosensors rely on Förster Resonance Energy Transfer (FRET), a distance-dependent radiationless through-space transfer of excitation energy from a singlet donor fluorophore to a singlet acceptor fluorophore.²⁷² The binding-associated conformational change of the SBP changes the distance between fused fluorophores, which is reflected in a change in fluorescence intensity ratio of the two fluorophores when the donor is excited by the researcher. The Förster distance R_0 , or distance at which half of the donor's excitation events are transferred to the acceptor, depends on the orientation factor κ^2 as follows:²⁷³

$$R_0^6 = K_F \kappa^2 Q_D n^{-4} \int_0^\infty \epsilon_A(\lambda) F_D(\lambda) \lambda^4 d\lambda \quad (3.1)$$

Where N_A is Avogadro's number, n is the index of refraction of the medium and varies from ~1.3 in pure water to ~1.6 in more crowded environments,²⁷⁴ λ is a wavelength, $\epsilon_A(\lambda)$ is the molar absorption coefficient of the acceptor at wavelength λ , and $F_D(\lambda)$ is the fluorescence

intensity of the donor, normalised to integrate to unity over the given limits, at that wavelength. The integral describes the overlap between the donor emission and acceptor absorption spectra and is often presented in the literature as a constant J for a given pair of fluorophores. K_F is a constant given by the equation:

$$K_F = \frac{9 \log(10)}{128 \pi^5 N_A} \approx 8.7851 \times 10^{-28} \text{ mol}$$

The FRET efficiency, or proportion of donor excitation events that are transferred to the acceptor, is then given in terms of the distance r between fluorophores and the Förster distance:

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6} \quad (3.2)$$

While κ^2 does not appear in this last equation, it does affect the efficiency via equation 3.1. It is usually taken to be $2/3$,²⁷⁵ which is the value it takes on when the fluorophores are able to randomly re-orient faster than the picosecond time-scale of the transfer.²⁷⁶ This is a reasonable assumption for small molecules in solution, but for fusion proteins it fails on both counts; steric interactions between domains mean that fluorophore orientations are not sampled randomly, and tumbling occurs on a nanosecond time-scale. In principle, it can take on values from 0 – 4.

The orientation factor could be calculated from the ensemble method, but must be assumed for the single point Rangefinder method. However, we did not pursue this in this publication. Let **D** and **A** be the donor emission and acceptor absorption dipole moments, respectively, and let **R** be the vector connecting the donor to the acceptor. Then θ_T is the angle between **D** and **A**, and θ_D and θ_A are the angles between the respective dipole moments and **R**. κ^2 for a single pair of molecules then has a simple dependence on these angles:²⁷⁶

$$\kappa^2 = (\cos \theta_T - 3 \cos \theta_D \cos \theta_A)^2 \quad (3.3)$$

Because of the complex dependence of the efficiency on κ^2 (equations 3.1 and 3.2), the ensemble average $\langle \kappa^2 \rangle$ can only be substituted into equation 3.1 when the fluorophores re-orient much faster than the timescale of the energy transfer.^{276,277} As a result, for ensemble approaches, the orientation factor, interfluorophore distance, Förster distance, and FRET efficiency should be computed individually for each structure, and only ensemble averaged as the FRET efficiency E .

3.2 Rangefinder: a semisynthetic FRET sensor design algorithm

3.2.1 Statement of contribution

I declare that the research presented in this chapter represents original work that I carried out during my candidature at the Australian National University, except for contributions to multi-author papers incorporated in the chapter where my contributions are specified in this Statement of Contribution.

Publication status

This manuscript has been published with the title *Rangefinder: a semisynthetic FRET sensor design algorithm* in the journal **ACS Sensors** (2016, [1:1286–1290](#)). The formatted article with supporting information is reproduced in this chapter.

Authorship and contribution

The manuscript was authored by Joshua A. Mitchell (the author), Jason H. Whitfield, William H. Zhang, Christian Henneberger, Harald Janovjak, Megan L. O'Mara, and Colin J. Jackson. JAM, JHW and WHZ contributed equally to the work. This method was an extension of the book chapter described below. I performed all computational work, including molecular dynamics simulations and writing and documenting the titular algorithm. In addition, I devised the combined computational-experimental method, performed the labelling and fluorescence titrations, and assisted with protein purification. Finally, I constructed figures 1 and 3, the graphical abstract, supporting figures S1, S5 and S7 and supporting tables S2 and S3, and I assisted in drafting, writing and editing the work.

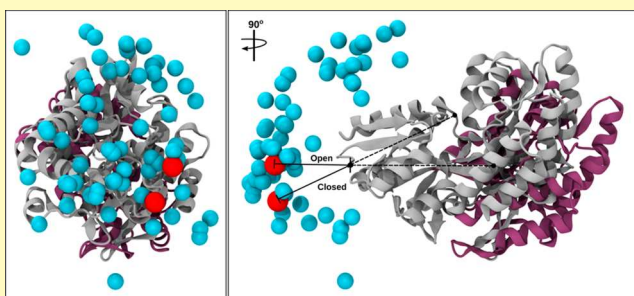
Rangefinder: A Semisynthetic FRET Sensor Design Algorithm

Joshua A. Mitchell,^{#,†} Jason H. Whitfield,^{#,†} William H. Zhang,^{#,†} Christian Henneberger,^{‡,§,||} Harald Janovjak,[⊥] Megan L. O'Mara,[†] and Colin J. Jackson^{*,†,||}[†]Research School of Chemistry, Australian National University, Canberra, 2601, Australia[‡]Institute of Neurology, University College London, London, WC1E 6BT, United Kingdom[§]German Center for Neurodegenerative Diseases (DZNE), 53175 Bonn, Germany^{||}Institute of Cellular Neurosciences, University of Bonn, 53113 Bonn, Germany[⊥]Institute of Science and Technology, 3400 Klosterneuburg, Austria

S Supporting Information

ABSTRACT: Optical sensors based on the phenomenon of Förster resonance energy transfer (FRET) are powerful tools that have advanced the study of small molecules in biological systems. However, sensor construction is not trivial and often requires multiple rounds of engineering or an ability to screen large numbers of variants. A method that would allow the accurate rational design of FRET sensors would expedite the production of biologically useful sensors. Here, we present Rangefinder, a computational algorithm that allows rapid in silico screening of dye attachment sites in a ligand-binding protein for the conjugation of a dye molecule to act as a Förster acceptor for a fused fluorescent protein. We present three ratiometric fluorescent sensors designed with Rangefinder, including a maltose sensor with a dynamic range of >300% and the first sensors for the most abundant sialic acid in human cells, *N*-acetylneuraminic acid. Provided a ligand-binding protein exists, it is our expectation that this model will facilitate the design of an optical sensor for any small molecule of interest.

KEYWORDS: arginine, biosensors, fluorescent dyes, FRET, maltose, periplasmic binding proteins, protein engineering, solute binding protein, Neu5Ac



Ratiometric FRET-based biosensors allow detection and quantitation of target analytes with excellent spatiotemporal resolution in physiological environments.^{1–5} The design of new sensors relies on the existence of a suitable binding protein for the analyte(s) of interest. Solute binding proteins (SBPs; SCOPe classification c.94.1) are among the largest known protein families,^{6,7} and members have been shown to bind ligands as diverse as amino acids,^{8,9} sugars,^{10–12} oligopeptides,¹³ and metal ions,¹⁴ with high specificity.

SBPs undergo conformational change upon ligand binding, which can be coupled to a change in FRET efficiency if the proteins are labeled or fused to fluorophores with overlapping fluorescence excitation and emission spectra.^{1–5,15} Despite broad structural conservation within the SBP superfamily, there is diversity in the magnitude and nature of the conformational changes that take place on ligand binding.^{16–20} Because of this, not all SBPs can be converted to sensors by fusing the N- and C-termini to fluorescent proteins.^{1,2,21} Several strategies have been developed to improve SBP-based FRET sensors. For example, relocation of the fluorophores relative to the SBP, via insertion of fluorescent proteins into loops on the binding core or circular permutation of the binding core itself, can allow improvement in the dynamic range (DR).^{3,5,22} Here, we define DR as the donor/acceptor fluorescence ratio (which depends

upon several factors, including FRET efficiency) of the sensor in the saturated “on” state, divided by the fluorescence ratio of the sensor in the unbound “off” state. A sensitive sensor, with a large dynamic range, will therefore undergo a large change in the fluorescence ratio upon ligand binding. We targeted a minimum dynamic range of 15% to ensure an acceptable signal-to-noise ratio for in vivo experiments.

Synthetic dye attachment, in contrast to fluorescent protein (FP) fusion, permits greater control over the fluorophore position and allows construction of sensors without protein remodeling.^{23,24} Unfortunately, identification of sites for fluorophore attachment (either dyes or FPs) is not trivial and often requires combinatorial testing and screening of variants.^{2,23,25,26} Sensors that incorporate dyes are often nonratiometric, preventing accurate quantification of analyte concentrations, but the use of two dyes to produce a ratiometric sensor complicates their construction, often necessitating the use of orthogonal protein–dye conjugation reactions.^{24,27,28} A semisynthetic FRET pair (dye/FP) allows

Received: September 14, 2016

Accepted: November 10, 2016

Published: November 10, 2016

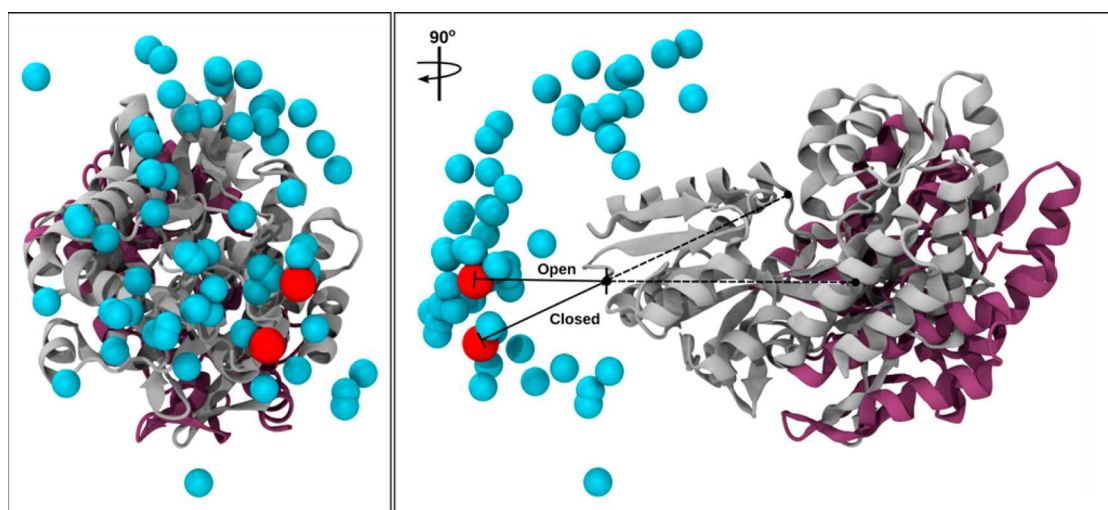


Figure 1. Front and side view of the fluorophore locations for the MBP construct predicted by MD simulation and Rangefinder. Coordinates of the MBP-CFP fusion protein from the end point of each of the 30 independent MARTINI simulations were fitted to the crystal structures of the open (purple, N-terminal lobe not shown) and closed (gray) SBP conformations, based on backbone RMSD of the SBP N-terminal lobe. For each of the 30 simulations, the location of the Trp63 fluorophore backbone, at the end point of the simulation, is shown as a blue sphere. The approximated ECFP positions used by Rangefinder (red spheres) are located 2 nm from the α -carbon of the SBP's N-terminus (black dot/vertical dash) positioned directly from the conformation's center of geometry for each conformation.

the production of ratiometric sensors with the superior optical properties of synthetic dyes in a single synthetic step.^{28,29}

Here, we describe a general method for the design of semisynthetic sensors (Figure S1). To demonstrate its utility, we have produced sensors for maltose, arginine, and sialic acid. Construction of these sensors involved a single round of computational screening with no experimental optimization. Maltose binding protein (MBP)^{11,30} has become a model system in sensor design,^{21,22,31,32} sialic acid (*N*-acetylneuraminic acid; Neu5Ac) binding protein (SAB)¹⁰ was selected because there are no Neu5Ac-specific SBP-based biosensors, and an ancestrally reconstructed arginine binding protein (AncQR)³³ was chosen to assess the accuracy of the algorithm when empirical structures of the SBPs are not available.

We sought a more accurate understanding of the positions adopted by the fluorescent protein, relative to the SBP in apo- and holo- states, to facilitate sensor design. We therefore performed coarse-grained molecular dynamics (MD) simulations with the MARTINI force-field on structures of enhanced cyan fluorescent protein³⁴ (ECFP)-solute binding protein (SBP) fusions, constructed by extending the disordered termini of the SBP structures in both the apo- and holo-states and fusing them to ECFP. Crystal structures were available for MBP (apo: 1JW4,³⁰ holo: 1ANF¹¹) and SAB (apo: 2CEX chain A,¹⁰ holo: 2CEY¹⁰). Unlike MBP and SAB, only the holo-structure was available for AncQR (4VZ1³³). A model of the apo- structure for AncQR was produced with i-Tasser³⁵ using the Gln-binding protein (1GGG³⁶) as a template.

Thirty 200 ns simulations were performed on each of the six starting structures. We observed the collapse of the disordered regions in all 180 simulations, resulting in the ECFP and SBP domains coming into contact. Notably, each replicate produced a distinct collapsed state (Figure 1); extending a sample of these simulations to a microsecond established that these states were stable. Although the individual simulations were not ergodic, the ensemble of final states represents an improved model for the ensemble of ECFP fluorophore positions for a given population of fusion proteins. This hemispherical

distribution indicates that while some individual sensor molecules might undergo large changes in FRET efficiency (high dynamic range), those with the ECFP collapsed near the hinge region or on the opposite face of the lobe will exhibit little to no change, effectively reducing the FRET signal for the population of molecules. This is consistent with the finding that constraining a fluorophore via linker truncation can improve dynamic range.²⁵

Because running a large number of MD simulations is time-consuming and requires substantial computational resources, we sought to develop an alternative approach that was less computationally intensive. The various collapsed states were structurally diverse, but were centered near the point of fusion. We reasoned that we could qualitatively approximate the ensemble by modeling an ECFP fluorophore 2 nm away from the N-terminus of the SBP, along a line drawn from the SBP's center of geometry and through its N-terminus. The approximated locations of the ECFP, for both apo- and holo-conformations, were typical of the ensembles of states that were generated (Figure 1). We then calculated the theoretical FRET efficiencies for sensors if the SBP was labeled with Alexa Fluor 532 C5 maleimide at every residue in the SBP domain in both the apo- and holo-conformations. For consistency, both approximated ECFP fluorophore positions were used to calculate efficiencies in each conformation; these two efficiencies were then averaged to give the efficiency for that conformation. With the averaged efficiencies, we calculated the predicted dynamic ranges for sensors and selected candidate positions for sensor construction (Figure S2). The Rangefinder algorithm is explained in greater detail in the Supporting Information. To investigate the effect of approximating the position of the ECFP, we also predicted DRs from ECFP fluorophore locations taken from six frames covering the last 50 ns of each MD simulation (30 each for apo- and holo-states). Efficiencies were calculated and averaged in the same way as for Rangefinder, but with 360 positions rather than just two (Figure S3).

To benchmark Rangefinder, we selected five dye attachment positions for both MBP and SAB, and two for AncQR (Figures S2–3). To investigate the accuracy of the model, we selected residues with a variety of predicted dynamic ranges. If we observed that residues were buried or were part of key structural motifs, they were excluded from selection.

We expressed each SBP, with the candidate site for dye conjugation mutated to a cysteine residue, as a fusion construct with an N-terminal cysteine-free ECFP (C48S, C70V) and used thiol-maleimide conjugation to attach the dye (Figure 2).³⁷ We

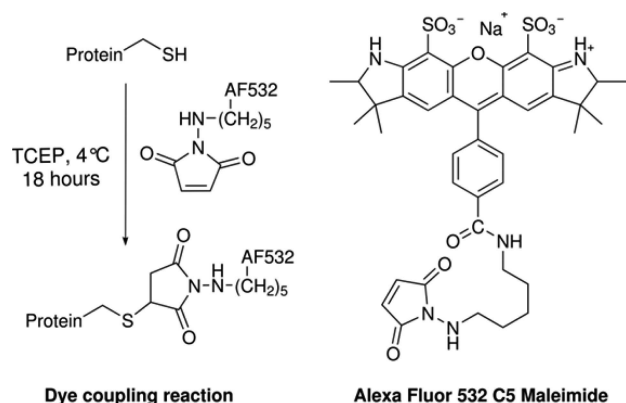


Figure 2. Scheme of the dye coupling reaction. Surface cysteine residues were reduced using TCEP (tris(2-carboxyethyl)phosphine) and Alexa Fluor AF532 C5 maleimide was added at 10-fold molar excess to the protein. The reactions were incubated for 18 h at 4 °C with gentle agitation.

did not add any linker residues to the fusion construct since long linkers have been suggested to reduce dynamic range.²⁵ Twelve variant proteins expressed in soluble form and maintained solubility after the labeling step. We tested the dynamic range and binding affinity of each variant by titrating them against their cognate ligand (Figures S4–5). Finally, we subjected a variant of the MBP construct that lacked any introduced cysteine residue to the same protocol, as a negative control. This variant did not display any acceptor fluorescence above background when the donor was excited, with or without the addition of ligand. However, absorption measurements of the labeled constructs indicated that the labeling efficiency was approximately 120%, which suggests that a small amount of unlabeled dye was present in the samples.

We evaluated our model as a screen for potential sensors by comparing the dynamic ranges predicted by both Rangefinder and the simulated ensemble to those that were determined experimentally (Figure S6, Table 1). The predicted DRs from both Rangefinder and the simulated ensembles correlate well with experimental values for both MBP and SAB (Pearson's correlation test, each with $p < 0.05$, Figure S7, Table S1).

For MBP ($R^2_{\text{RF}} = 0.88$, $R^2_{\text{Ens}} = 0.86$), the predictions were qualitatively accurate: the two positive predictions (Mal 381, Mal 437) yielded efficient sensors, while the three negative predictions did not exhibit significant change in fluorescence intensity upon addition of maltose. For SAB, the ensemble model produced highly correlated predictions ($R^2_{\text{RF}} = 0.96$, $R^2_{\text{Ens}} = 0.99$) although the dynamic range was systematically lower ($\sim 30\%$) for all designs. In the case of the AncQR sensors, designed with a homology model of the apo-structure, both sensors were responsive, albeit significantly less than predicted by Rangefinder (Table 1). Thus, X-ray crystallographic or

Table 1. Dynamic Range and Affinity of Sensor Constructs^a

variant	Ens.	RF	Exp.	K_d (mM)
Mal 381	0.48	0.48	0.51	5.8 ± 0.46
Mal 393	0.12	0.07	0.00	-
Mal 437	0.85	0.90	3.12	390 ± 72
Mal 482	0.08	0.05	-0.02	-
Mal 524	0.03	0.02	0.04	-
Sia 362	0.13	0.02	0.02	-
Sia 371	0.37	0.14	0.11	-
Sia 397	1.07	0.77	0.32	0.85 ± 0.02
Sia 404	0.51	0.29	0.17	0.44 ± 0.03
Sia 425	0.39	0.27	0.11	-
Arg 345	0.17	0.82	0.19	38 ± 5.7
Arg 365	0.74	2.04	0.14	25 ± 2.5

^a“Ens.” denotes simulation ensemble method, “RF” denotes Rangefinder. K_d data is shown in (μM) where the construct gave a substantial ratio change ($\sim 15\%$).

NMR structures should be used when possible. Even given the inaccuracies introduced through the homology modeling, Rangefinder was able to design functional sensors in the case of AncQR. Overall, the comparison between the predictions generated by Rangefinder and those generated from the ensembles produced by the computationally intensive coarse-grained MD simulations revealed Rangefinder to be comparably accurate, although Rangefinder was substantially less accurate at predicting the DRs of the AncQR models, relative to the ensemble method.

In addition to correctly predicting successful designs in a test-set of twelve proteins, in this work Rangefinder has produced two additional results of note. First, it has resulted in the construction of a ratiometric sensor for the common model system, MBP, with a dynamic range of $>300\%$. To the best of our knowledge, this is approximately 5-fold greater than the next largest dynamic range for a ratiometric MBP sensor in the literature (Figure 3A,C).²² This result highlights the potential of semisynthetic ratiometric sensors, which can exhibit DRs an order of magnitude greater than typical fluorescent protein based ratiometric sensors. Although the K_d of this sensor increased to $390 \mu\text{M}$, the large DR meant that significant

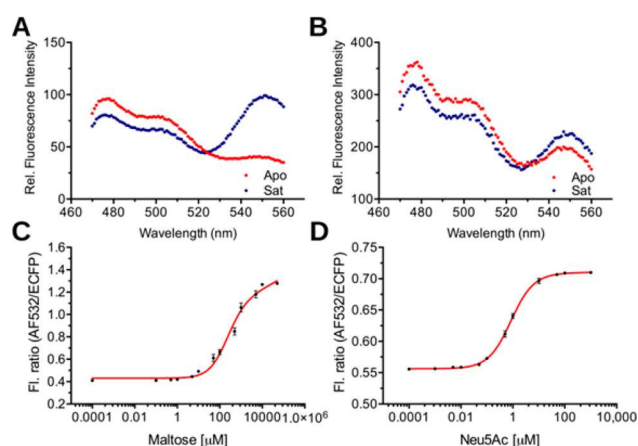


Figure 3. Characterization of the best performing constructs for MBP and SAB. (A) Mal 437 shows a 311% dynamic range with saturating maltose. (B) Sia 397 shows a 32% dynamic range with saturating Neu5Ac. (C) Fluorescence titration of Mal 437, indicating a K_d of $390 \mu\text{M}$. (D) Fluorescence titration of Sia 397 showing a K_d of $0.85 \mu\text{M}$.

changes in the fluorescence ratio of this sensor are observable from addition of 1 μM to 100 μM maltose (Figure 3). Second, the development of a sialic acid sensor will now allow for the detection and study of Neu5Ac (Figure 3B,D), which has been implicated in the regulation of neural networks³⁸ and may have significant roles in early human neurodevelopment.³⁹

RangeFinder is a straightforward algorithm that runs in seconds on a modern personal computer. It models the highly dynamic ECFP domain as a single point in a typical position. Compared to the ECFP, whose position can vary by up to 60 Å (Figure 2), the location of the much smaller covalently attached dye is relatively restricted and its dynamics are therefore not considered (it is modeled at the α -carbon of the residue of interest). Despite these approximations, the method is able to accurately predict sites for dye attachment that yield efficient sensors. The observation that the DR for Mal437 was extraordinarily high was somewhat surprising, as it significantly exceeded (312% vs 85/90%) the predictions from RangeFinder and the ensemble method. It is possible that there is an additional effect, not incorporated in our models, contributing to this large DR, such as constraint of the fluorescent dye at this particular position.

RangeFinder is designed for use with proteins of the SBP superfamily, which encompass thousands of diverse ligand binding proteins.^{6,7} However, the method is sufficiently generalizable that it could be adapted for virtually any structural fold that undergoes a conformational change on ligand binding. Additionally, RangeFinder has been used in this work to produce hybrid biosensors via the use of thiol chemistry to site selectively label introduced cysteine residues, which precludes their use in *in vivo* applications. However, the use of site-specific incorporation of unnatural amino acids that can undergo bio-orthogonal “click” chemistry reactions has been shown to allow *in vivo* dye attachment of dyes to biosensors and would be equally effective with RangeFinder.^{40–42} In summary, RangeFinder is a rapid and simple-to-use computational design tool to facilitate FRET biosensor construction and can reliably produce sensors for a diverse range of biological ligands.⁴³

■ ASSOCIATED CONTENT

■ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssensors.6b00576.

Materials and methods including the prediction of dynamic ranges, RangeFinder outputs, statistical analyses, and variant characterizations (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

ORCID

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

Author Contributions

[#]These authors contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.A.M., J.H.W., and W.H.Z. were supported by Australian Postgraduate Awards (APA), AS Sargeson Supplementary scholarships, and RSC supplementary scholarships. C.J.J. acknowledges support from a Human Frontiers in Science Young Investigator Award and a Discovery Project and Future Fellowship from the Australian Research Council. M.L.O. is supported by an Australian Research Council Discovery Project (DP130102153) and the Merit Allocation Scheme of the National Computational Infrastructure.

■ REFERENCES

- (1) Okumoto, S.; Looger, L. L.; Micheva, K. D.; Reimer, R. J.; Smith, S. J.; Frommer, W. B. Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 8740–8745.
- (2) Hires, S. A.; Zhu, Y.; Tsien, R. Y. Optical measurement of synaptic glutamate spillover and reuptake by linker optimized glutamate-sensitive fluorescent reporters. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 4411–4416.
- (3) Bogner, M.; Ludewig, U. Visualization of arginine influx into plant cells using a specific FRET-sensor. *J. Fluoresc.* **2007**, *17*, 350–360.
- (4) Gruenewald, K.; Holland, J. T.; Stromberg, V.; Ahmad, A.; Watcharakichkorn, D.; Okumoto, S. Visualization of Glutamine Transporter Activities in Living Cells Using Genetically Encoded Glutamine Sensors. *PLoS One* **2012**, *7*, e38591.
- (5) Whitfield, J. H.; Zhang, W. H.; Herde, M. K.; Clifton, B. E.; Radziejewski, J.; Janovjak, H.; Henneberger, C.; Jackson, C. J. Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* **2015**, *24*, 1412–1422.
- (6) Berntsson, R. P. A.; Smits, S. H. J.; Schmitt, L.; Slotboom, D.-J.; Poolman, B. A structural classification of substrate-binding proteins. *FEBS Lett.* **2010**, *584*, 2606–2617.
- (7) Fox, N. K.; Brenner, S. E.; Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **2014**, *42*, D304–D309.
- (8) Sun, Y.; Rose, J.; Wang, B.; Hsiao, C. The structure of glutamine-binding protein complexed with glutamine at 1.94 Å resolution: comparisons with other amino acid binding proteins. *J. Mol. Biol.* **1998**, *278*, 219–229.
- (9) Yao, N.; Trakhanov, S.; Quirocho, F. A. Refined 1.89-Å structure of the histidine-binding protein complexed with histidine and its relationship with many other active transport/chemosensory proteins. *Biochemistry* **1994**, *33*, 4769–4779.
- (10) Müller, A.; Severi, E.; Mulligan, C.; Watts, A. G.; Kelly, D. J.; Wilson, K. S.; Wilkinson, A. J.; Thomas, G. H. Conservation of structure and mechanism in primary and secondary transporters exemplified by SiaP, a sialic acid binding virulence factor from *Haemophilus influenzae*. *J. Biol. Chem.* **2006**, *281*, 22212–22222.
- (11) Quirocho, F. A.; Spurlino, J. C.; Rodseth, L. E. Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure* **1997**, *5*, 997–1015.
- (12) Vyas, N.; Vyas, M.; Quirocho, F. Comparison of the periplasmic receptors for L-arabinose, D-glucose/D-galactose, and D-ribose. Structural and Functional Similarity. *J. Biol. Chem.* **1991**, *266*, 5226–5237.
- (13) Levnikov, V. M.; Blagova, E. V.; Brannigan, J. A.; Wright, L.; Vagin, A. A.; Wilkinson, A. J. The structure of the oligopeptide-binding protein, AppA, from *Bacillus subtilis* in complex with a nonapeptide. *J. Mol. Biol.* **2005**, *345*, 879–892.
- (14) Pina, K.; Navarro, C.; Mcwaller, L.; Boxer, D. H.; Price, N. C.; Kelly, S. M.; Mandrand-Berthelot, M. A.; Wu, L. F. Purification and Characterization of the Periplasmic Nickel-Binding Protein NikA of *Escherichia coli* K12. *Eur. J. Biochem.* **1995**, *227*, 857–865.

- (15) Abraham, B. G.; Sarkisyan, K. S.; Mishin, A. S.; Santala, V.; Tkachenko, N. V.; Karp, M. Fluorescent Protein Based FRET Pairs with Improved Dynamic Range for Fluorescence Lifetime Measurements. *PLoS One* **2015**, *10*, e0134436.
- (16) Quijcho, F. A.; Ledvina, P. S. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. *Mol. Microbiol.* **1996**, *20*, 17–25.
- (17) Ger, M.-F.; Rendon, G.; Tilson, J. L.; Jakobsson, E. Domain-Based Identification and Analysis of Glutamate Receptor Ion Channels and Their Relatives in Prokaryotes. *PLoS One* **2010**, *5*, e12827.
- (18) Tang, C.; Schwieters, C. D.; Clore, G. M. Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature* **2007**, *449*, 1078–1082.
- (19) Bermejo, G. A.; Strub, M.-P.; Ho, C.; Tjandra, N. Ligand-free open-closed transitions of periplasmic binding proteins: the case of glutamine-binding protein. *Biochemistry* **2010**, *49*, 1893–1902.
- (20) Trakhanov, S.; Vyas, N. K.; Luecke, H.; Kristensen, D. M.; Ma, J.; Quijcho, F. A. Ligand-free and-bound structures of the binding protein (LivJ) of the Escherichia coli ABC leucine/isoleucine/valine transport system: trajectory and dynamics of the interdomain rotation and ligand specificity. *Biochemistry* **2005**, *44*, 6597–6608.
- (21) Fehr, M.; Frommer, W. B.; Lalonde, S. Visualization of maltose uptake in living yeast cells by fluorescent nanosensors. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 9846–9851.
- (22) Okada, S.; Ota, K.; Ito, T. Circular permutation of ligand-binding module improves dynamic range of genetically encoded FRET-based nanosensor. *Protein Sci.* **2009**, *18*, 2518–2527.
- (23) Namiki, S.; Sakamoto, H.; Iinuma, S.; Iino, M.; Hirose, K. Optical glutamate sensor for spatiotemporal analysis of synaptic transmission. *Eur. J. Neurosci.* **2007**, *25*, 2249–2259.
- (24) Medintz, I. L.; Goldman, E. R.; Lassman, M. E.; Mauro, J. M. A fluorescence resonance energy transfer sensor based on maltose binding protein. *Bioconjugate Chem.* **2003**, *14*, 909–918.
- (25) Deuschle, K.; Okumoto, S.; Fehr, M.; Looger, L. L.; Kozhukh, L.; Frommer, W. B. Construction and optimization of a family of genetically encoded metabolite sensors by semirational protein engineering. *Protein Sci.* **2005**, *14*, 2304–2314.
- (26) Fritz, R. D.; Letzelter, M.; Reimann, A.; Martin, K.; Fusco, L.; Ritsma, L.; Ponsioen, B.; Fluri, E.; Schulte-Merker, S.; van Rheenen, J.; Pertz, O. A versatile toolkit to produce sensitive FRET biosensors to visualize signaling in time and space. *Sci. Signaling* **2013**, *6*, rs12–rs12.
- (27) Hsieh, H. V.; Sherman, D. B.; Andaluz, S. A.; Amiss, T. J.; Pitner, J. B. Fluorescence resonance energy transfer glucose sensor from site-specific dual labeling of glucose/galactose binding protein using ligand protection. *J. Diabetes Sci. Technol.* **2012**, *6*, 1286–1295.
- (28) Brun, M. A.; Tan, K.-T.; Nakata, E.; Hinner, M. J.; Johnsson, K. Semisynthetic fluorescent sensor proteins based on self-labeling protein tags. *J. Am. Chem. Soc.* **2009**, *131*, 5873–5884.
- (29) Adams, S. R.; Harootunian, A. T.; Buechler, Y. J.; Taylor, S. S.; Tsien, R. Y. Fluorescence ratio imaging of cyclic AMP in single cells. *Nature* **1991**, *349*, 694–697.
- (30) Duan, X.; Quijcho, F. A. Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry* **2002**, *41*, 706–712.
- (31) Marvin, J.; Corcoran, E.; Hattangadi, N.; Zhang, J.; Gere, S.; Hellinga, H. The rational design of allosteric interactions in a monomeric protein and its applications to the construction of biosensors. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 4366–4371.
- (32) Zhou, L. Q.; Cass, A. E. Periplasmic binding protein based biosensors 1. Preliminary study of maltose binding protein as sensing element for maltose biosensor. *Biosens. Bioelectron.* **1991**, *6*, 445–450.
- (33) Clifton, B. E.; Jackson, C. J. Ancestral Protein Reconstruction Yields Insights into Adaptive Evolution of Binding Specificity in Solute-Binding Proteins. *Cell. Chem. Biol.* **2016**, *23*, 236–245.
- (34) Heim, R.; Tsien, R. Y. Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Curr. Biol.* **1996**, *6*, 178–182.
- (35) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **2014**, *12*, 7–8.
- (36) Hsiao, C.; Sun, Y.; Rose, J.; Wang, B. The Crystal Structure of Glutamine-binding Protein from Escherichia coli. *J. Mol. Biol.* **1996**, *262*, 225–242.
- (37) Suzuki, M.; Ito, Y.; Savage, H. E.; Husimi, Y.; Douglas, K. T. Protease-sensitive signalling by chemically engineered intramolecular fluorescent resonance energy transfer mutants of green fluorescent protein. *Biochim. Biophys. Acta, Gene Struct. Expression* **2004**, *1679*, 222–229.
- (38) Isaev, D.; Isaeva, E.; Shatskih, T.; Zhao, Q.; Smits, N. C.; Shworak, N. W.; Khazipov, R.; Holmes, G. L. Role of extracellular sialic acid in regulation of neuronal and network excitability in the rat hippocampus. *J. Neurosci.* **2007**, *27*, 11587–11594.
- (39) Wang, B.; Brand-Miller, J. The role and potential of sialic acid in human nutrition. *Eur. J. Clin. Nutr.* **2003**, *57*, 1351–1369.
- (40) Plass, T.; Milles, S.; Koehler, C.; Schultz, C.; Lemke, E. A. Genetically Encoded Copper-Free Click Chemistry. *Angew. Chem., Int. Ed.* **2011**, *50*, 3878–3881.
- (41) Lang, K.; Davis, L.; Torres-Kolbus, J.; Chou, C.; Deiters, A.; Chin, J. W. Genetically encoded norbornene directs site-specific cellular protein labelling via a rapid bioorthogonal reaction. *Nat. Chem.* **2012**, *4*, 298–304.
- (42) Seitchik, J. L.; Peeler, J. C.; Taylor, M. T.; Blackman, M. L.; Rhoads, T. W.; Cooley, R. B.; Refakis, C.; Fox, J. M.; Mehl, R. A. Genetically encoded tetrazine amino acid directs rapid site-specific in vivo bioorthogonal ligation with trans-cyclooctenes. *J. Am. Chem. Soc.* **2012**, *134*, 2898–2901.
- (43) Rangefinder is available for free download as a python script from <http://chemistry.anu.edu.au/research/groups/chemical-structural-biology/rangefinder>

Supporting information:

Rangefinder: A semi-synthetic FRET sensor design algorithm

Joshua A. Mitchell[‡], Jason H. Whitfield[‡], William H. Zhang[‡], Megan L. O'Mara, Colin J. Jackson*

Methods	2
<i>Generation of initial extended linker models.</i>	2
<i>Simulations.</i>	2
<i>Rangefinder algorithm.</i>	2
<i>Predictions of dynamic ranges.</i>	4
<i>DNA cloning and mutagenesis.</i>	5
<i>Expression and purification of protein.</i>	5
<i>Dye labelling reaction.</i>	6
<i>Fluorescence assays.</i>	6
<i>Labeling efficiency determination.</i>	6
Supporting Information Figures and Tables	7
<i>Figure S1.</i>	7
<i>Figure S2.</i>	8
<i>Figure S3.</i>	9
<i>Figure S4.</i>	10
<i>Figure S5.</i>	11
<i>Figure S6.</i>	12
<i>Figure S7.</i>	13
<i>Table S1. Primers used for site directed mutagenesis and cloning.</i>	14
<i>Table S2. Amino acid sequences of the ECFP-SBP fusion constructs.</i>	15
<i>Table S3. Amino acid sequences of the ECFP-SBP fusion constructs.</i>	16
<i>Table S4. Pearson's test for the Rangefinder and ensemble predictions of for the dynamic ranges of the MBP and SAB series variants.</i>	17
References:	18

Methods

Generation of initial extended linker models. Fluorescent proteins in the green fluorescent protein (GFP) family incorporate a three amino acid motif that is autocatalytically converted to a fluorophore.¹ The modified fluorophore residue of the crystal structure of ECFP (PDB 2WSN²) was converted to the unreacted motif Thr-Trp-Gly for compatibility with the MARTINI force field.³ The disordered C-terminal region was then extended in PyMOL⁴ by rotating around ϕ and φ dihedrals to form a linear linker region. Additional residues from the N-terminus that were absent from each SBP model (e.g. Ala1 from PDB ID 2CEX⁵) were restored. The N-terminus of each SBP model was extended in the same way as the C-terminus of the ECFP domain. The residues of the SBP models were renumbered as necessary and the models fused in PyMOL to generate initial models of the full structure with an extended linker.

Simulations. Simulations were conducted with the MARTINI 2.2³ force field in GROMACS 5.1.2.⁶ Systems were solvated with a non-polarizable water model including 10% antifreeze water and 200 mM NaCl in truncated dodecahedral boxes with 0.5 nm to the edge of the box, leading to a minimum distance between nearest solute atoms in adjacent cells of approximately three times the Coulomb and Van der Waals cut-offs. An elastic network was applied to all pairs of atoms within 0.5-0.9 nm of each other, excluding the disordered linker region. This network was applied to constrain each SBP in its open or closed conformation. Each fusion model was simulated 30 times for 200 ns with a 20 fs time step. Simulations otherwise followed the *rf-new* parameter set of de Jong et al.⁷ Linker collapse was observed within 100 ns, after which there was minimal movement of one domain relative to the other. These simulations took two to four hours per replica on an in-house machine equipped with a 32 core Intel Xeon 2.40 GHz processor, two NVIDIA Tesla K40m GPUs and 64 GB RAM. The cumulative simulation time per sensor was estimated to be 240 hours.

Rangefinder algorithm. Coordinates are read and processed with the Biopython library.⁸ First, the center of geometry of each SBP is calculated, and a model ECFP fluorophore bead is placed 2 nm away from the N-terminal alpha carbon along the

vector from the center of geometry to the N-terminal alpha carbon (Figure S1B). This gives a position 2 nm away from the terminus to which the ECFP is fused; using the center of geometry ensures that the fluorophore is not placed within the protein itself. This location was chosen because it is easily calculated by extrapolating from the center of geometry of the SBP, is within the distribution of fluorophore positions observed during the coarse-grained molecular dynamics simulations (Figure 1), and can be readily applied to all Venus flytrap-fold SBPs. Using this approximate location allows Rangefinder to give results immediately, without requiring extensive MD simulations or high performance computing resources.

The N-terminal lobes of both structures are superimposed, and the distances from each residue in the binding core to both hypothetical fluorophore beads was measured in order to account for movement of the N-terminus (Figure S1C). FRET efficiency is calculated for each distance according to equation (1):⁹

$$1) E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6}$$

Where E is the FRET efficiency, r is the interfluorophore distance from the model, and R_0 is the Förster distance (here taken as 4.8 nm). The two efficiencies measured from the closed state are averaged, as are those for the open state, yielding predicted FRET efficiencies for each conformation. These are then converted into expected dynamic ranges via equation (2):

$$2) \text{ Dynamic range} = \frac{E_{holo} - E_{apo}}{(1 - E_{holo})(1 + S(E_{apo} - 1))}$$

Where E_{apo} and E_{holo} are the average FRET efficiencies in each conformation, and S is a derived factor that represents the degree of spectral overlap between fluorophores, approximately 0.82 for ECFP-AF532:

$$3) S = \frac{1 - \frac{Q_D F_D}{Q_A F_A}}{\frac{\epsilon_A}{\epsilon_D} + 1} = \frac{\epsilon_D (Q_A F_A - Q_D F_D)}{Q_A F_A (\epsilon_A + \epsilon_D)}$$

Here Q_A and Q_D are the quantum yields of acceptor and donor, ϵ_A and ϵ_D are their molar attenuation coefficients at the excitation wavelength for an experiment, and F_A

and F_D are the emission intensities of the donor and acceptor in isolation at the acceptor emission peak after the spectra have been normalized to unity. FRET fluorophore design is complicated by the possibility of “cross-talk” between fluorophores. The acceptor may be directly excited by the incident light and donor emission may overlap with the acceptor peak. In principle, it is also possible for the acceptor to donate FRET to the donor and for acceptor emission to overlap the donor peak. As this does not occur for the donor-acceptor pairs under study here owing to the width of the donor spectra and narrowness of the acceptor spectra, these forms of cross-talk were assumed to be negligible for the derivation of these formulae.

Rangefinder output includes predicted dynamic ranges for each residue as a comma-separated text file. Residues used to superimpose the two structures, which do not change position, give predicted dynamic ranges of 0. It can take an arbitrary number of structures as inputs for each conformation, all of which will generate their own fluorophore position and set of predicted dynamic ranges. Thus, the algorithm, can be applied to MD simulations or individual structures of sensors.

Rangefinder will be widely applicable to binding proteins with a Venus flytrap fold that exhibit substantial conformational changes associated with binding. Its potential applications should extend to the structurally homologous Venus fly trap fold binding domains of some membrane-bound receptors. In addition, Rangefinder may be suitable for proteins of other folds, provided that they can be divided into two structural regions that move relative to each other upon binding, one of which must include the N-terminus.

Rangefinder was developed with the ECFP-AF532 FRET pair in mind, but supports other fluorophores as long as the Förster distance is provided. The constant S (Eq. 3) may also be specified, but is assumed to be 0.75 if omitted.

Predictions of dynamic ranges. The open and closed models used to generate extended linker models for each SBP above were used as input for the Rangefinder algorithm to generate predicted dynamic ranges.

To calculate dynamic ranges from simulation, six frames were taken from the last 50 ns of each simulation. For each frame, the coordinates of the backbone bead of

Trp63 (the modified amino acid that constitutes the fluorophore of ECFP) was recorded. All other frames were fitted to that frame by the C-terminal SBP domain, and distances from the selected bead to the backbone bead of each residue in the binding core were measured. This fitting-measurement process was repeated for each frame to generate a set of inter-fluorophore distances from each residue in each frame, relative to the ECFP fluorophore in all frames. Efficiencies and dynamic ranges were then calculated as above.

DNA cloning and mutagenesis. The arginine binding protein, AncQR, previously generated through ancestral protein reconstruction was cloned into a pETMCSIII vector¹⁰⁻¹¹. Maltose binding protein (MBP; P0AEX9) and Sialic acid binding protein (SAB; P44542) wild type genes with flanking sequences complementary to ECFP and the T7 terminator region of the vector (5' and 3' respectively) were synthesized by GeneArt (ThermoFisher Scientific) and cloned into the pETMCSIII vector¹¹ with the cysteine deficient ECFP already present. Site directed mutagenesis was performed by creating gene fragments using long mutagenic primers based on the QuikChange™ method¹² to introduce the relevant cysteine mutations. T7 terminator/ECFP-SBP-specific primers were used to create gene fragments with 40 bp overlap with each other and the vector backbone. These were assembled by Gibson assembly.¹³ A list of primers used in this work is provided in SI Table S1.

Expression and purification of protein. The MBP and AncQR variants were expressed in BL21 (DE3) *Escherichia coli* cells (New England Biolabs), grown for 48 hours at 20 °C in Lysogeny Broth (LB). The sialic acid binding protein variants were expressed using auto inducing M9 minimal media (auto-induction reagents: 0.5% glycerol, 0.05% glucose and 0.2% lactose) that was further supplemented with sodium sulfate (5 mM), L-amino acids (500 µg of each amino acid) and trace metals (500 µL of a 1000-fold stock).¹⁴ Cells were pelleted through centrifugation (4730 g in a VWR VX22G centrifuge using a Hitachi R9A rotor at 4 °C for 15 minutes) and resuspended in 50 mL buffer A (50 mM NaH₂PO₄, 200 mM NaCl, 20 mM imidazole pH 7.3) and lysed by sonication (Omni Sonic ruptor 400 ultrasonic homogenizer with OR-T-375 processing tip) for 5 minutes per sample at 50% pulse and 50% power. Samples were kept on ice during sonication. Purification was achieved *via* nickel affinity chromatography, using 5 mL Ni-NTA columns (GE lifesciences), equilibrated in buffer A for protein binding and 100% buffer B (50 mM NaH₂PO₄, 200 mM NaCl,

250 mM imidazole pH 7.3) for elution. All samples were dialyzed overnight at 4 °C (~18 hours) in 2 L dye reaction buffer C (50 mM NaH₂PO₄, 200 mM NaCl, pH 7.3) per sample. SAB variant buffers were exchanged and dialyzed for a further 7 hours.

Dye labelling reaction. Alexa Fluor® 532 C5 Maleimide (1 mg) (ThermoFisher Scientific) was dissolved in dye reaction buffer C (500 µL to give 2 mg/mL concentration). The dye labelling reaction (in 500 µL) was 3.5 µL (final concentration: 700 µM) Tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl) (100 mM stock solution) with a 10:1 excess of the AF532 dye to protein. 80 µM protein was used per reaction (therefore 800 µM dye – 167 µL of a 2 mg/mL (2.46 mM) stock solution in a 500 µL reaction). The volume was made up to 500 µL with buffer C. Reactions were incubated at 4 °C overnight (~18 hours) with gentle agitation. To remove excess dye, samples were buffer exchanged using PD-10 gravity columns (GE life sciences), equilibrated in buffer C, with samples eluted in 3.5 mL of buffer C. These samples were concentrated to 1 mL using Amicon spin concentrators (30 kDa cut-off) and buffer exchanged further with PD-10 columns, as done previously.

Fluorescence assays. These were performed using a Cary Varian spectrophotometer. All sensor constructs were excited at 433 ± 5 nm with spectra taken between 470 nm and 560 nm. Samples were allowed to equilibrate to room temperature (25 °C) for approximately 10 minutes before measurements were taken. To calculate the fluorescence ratio, peaks at 476 nm (ECFP) and 550 nm (AF532) were used. All ligands were dissolved in buffer C. Saturating spectra were obtained by the addition of saturating concentrations of ligand (minimum 10 mM ligand).

Labeling efficiency determination. Protein concentration was calculated using the UV 280 nm absorbance. The concentration of the dye was calculated using the Beer-Lambert law and AF532 absorbance at 528 nm with a Nanodrop ND1000 spectrophotometer. The molar ratio of these concentrations was used to calculate the labeling efficiency.

Supporting Information Figures and Tables

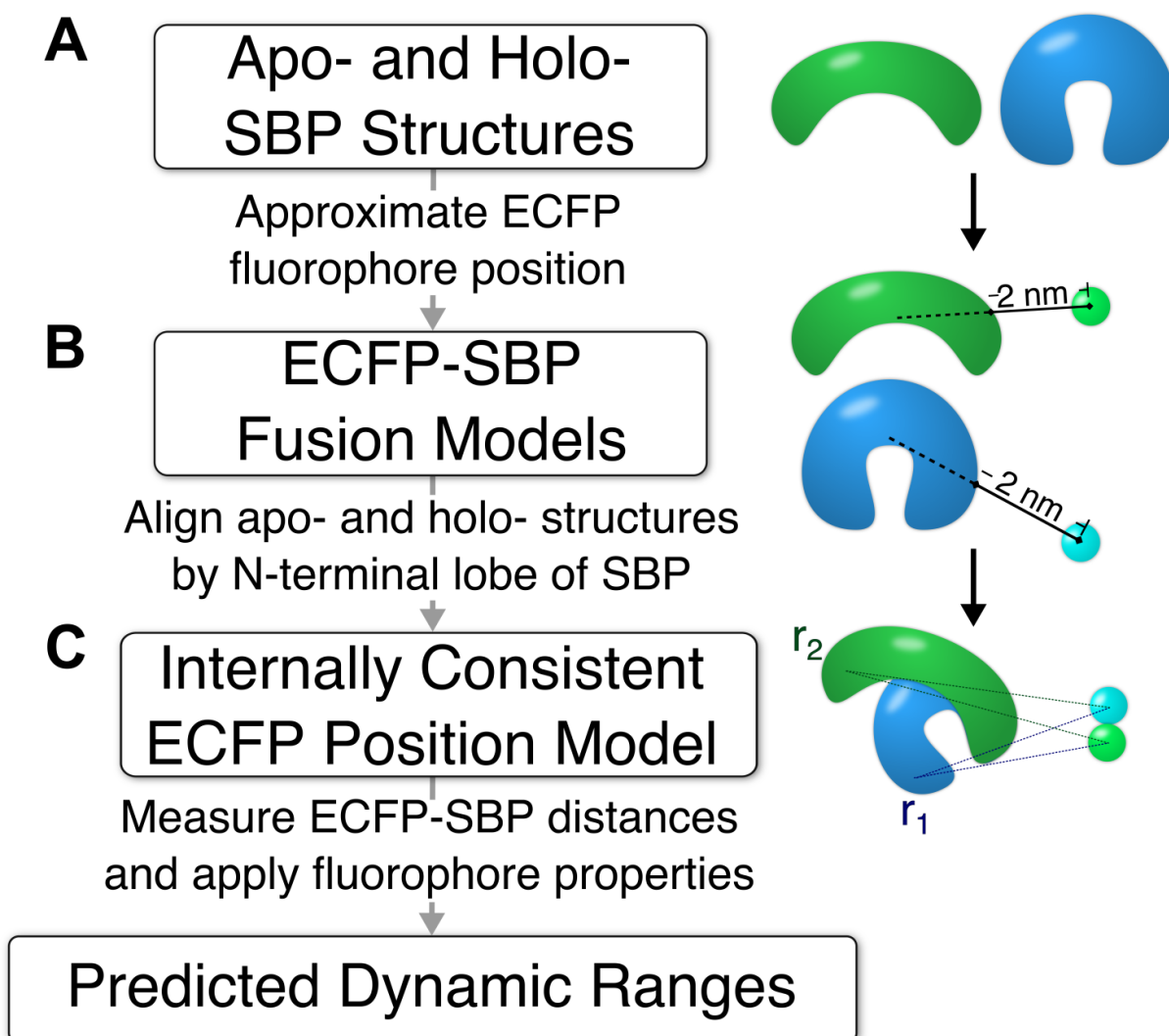


Figure S1. The methodology used in Rangefinder. A) Structures (experimental or modeled) of the SBP in both the open (apo-) and closed (holo-) conformations are collected and the position of the fused fluorescent protein is modeled independently for both SBP conformations. The ECFP position is taken as the point 2 nm away from the SBP's N-terminus along the axis from the SBP center of geometry through the N-terminus. B) These fusion models are then aligned by the N terminal lobe of the SBP and inter-fluorophore (ECFP – Alexa Fluor 532) distances are measured as the distance between every residue and the fluorescent protein location in both holo- (r_1) and apo- (r_2) protein conformations. C) Efficiencies are calculated from the distances r_1 and r_2 and the input Förster distance (4.80 nm by default) using equation 1. These efficiencies are then used in equation 2 to calculate the predicted dynamic ranges, which are outputted.

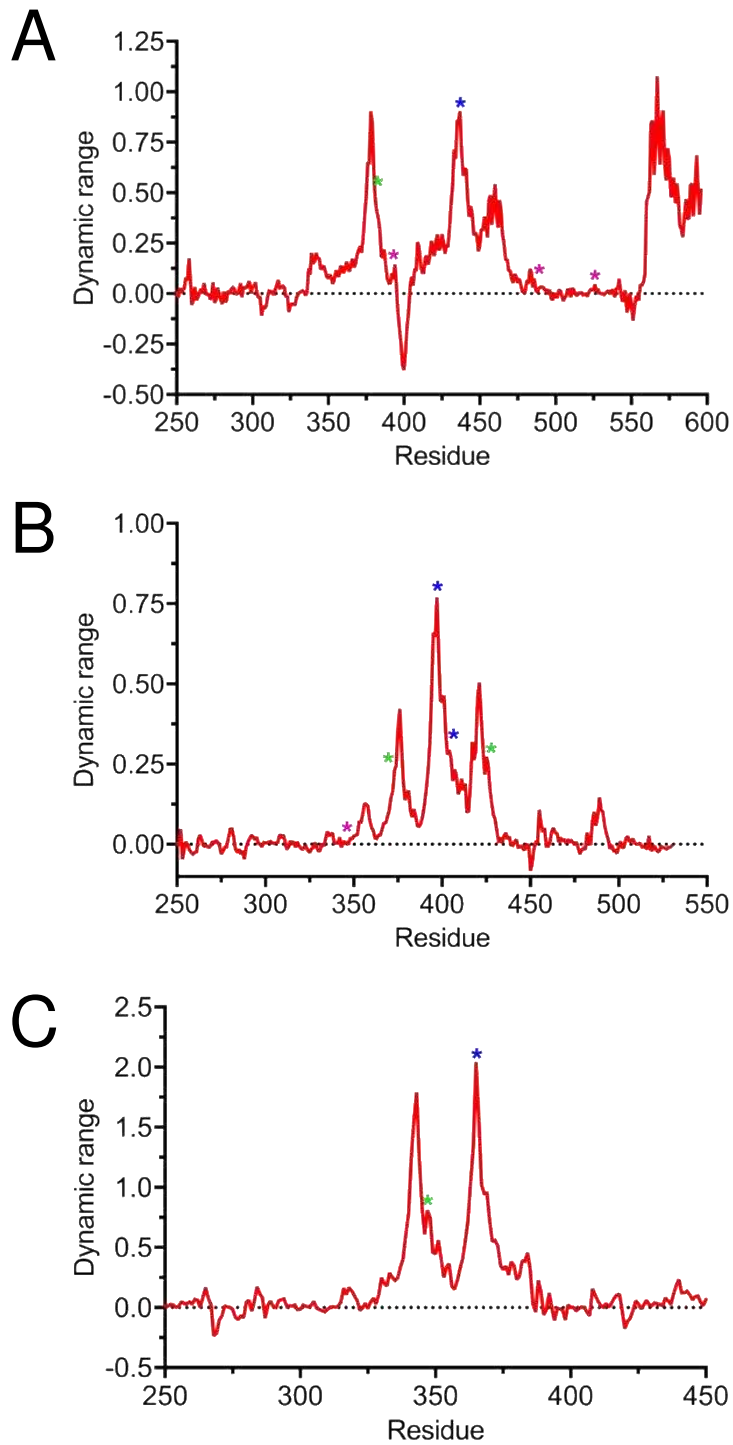


Figure S2. Predicted dynamic ranges for (A) MBP, (B) SAB, (C) AncQR from Rangefinder. Residues are numbered as part of the full ECFP fusion construct. *Denotes the sites selected for benchmarking, covering a range of predicted values to test Rangefinder's ability to be predictive. The blue sites are predicted to yield sensors with "high" dynamic ranges, the green sites "moderate" and the purple sites "low".

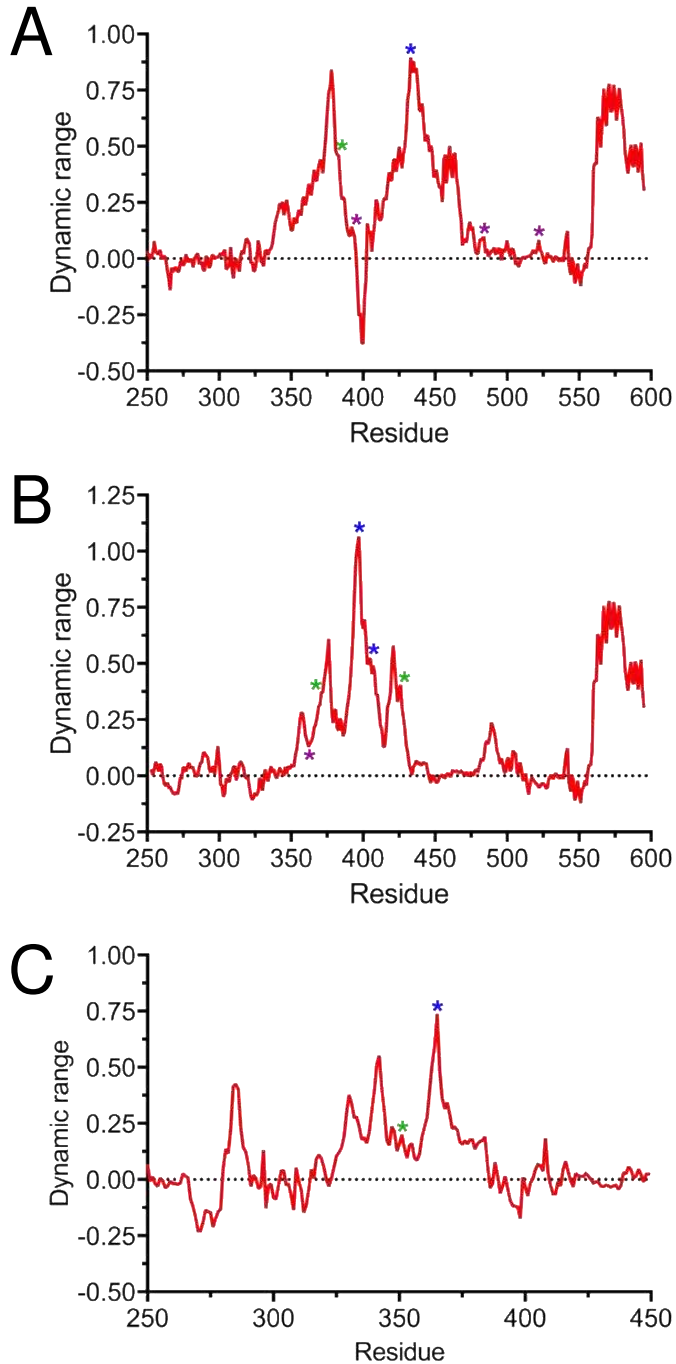


Figure S3. Predicted dynamic ranges for (A) MBP, (B) SAB, (C) AncQR from the computationally intensive ensemble method. Residues are numbered as part of the full ECFP-SBP fusion construct. *Denotes the sites selected for benchmarking, covering a range of predicted values to test Rangefinder’s ability to be predictive. The blue sites are predicted to yield sensors with “high” dynamic ranges, the green sites “moderate” and the purple sites “low”.

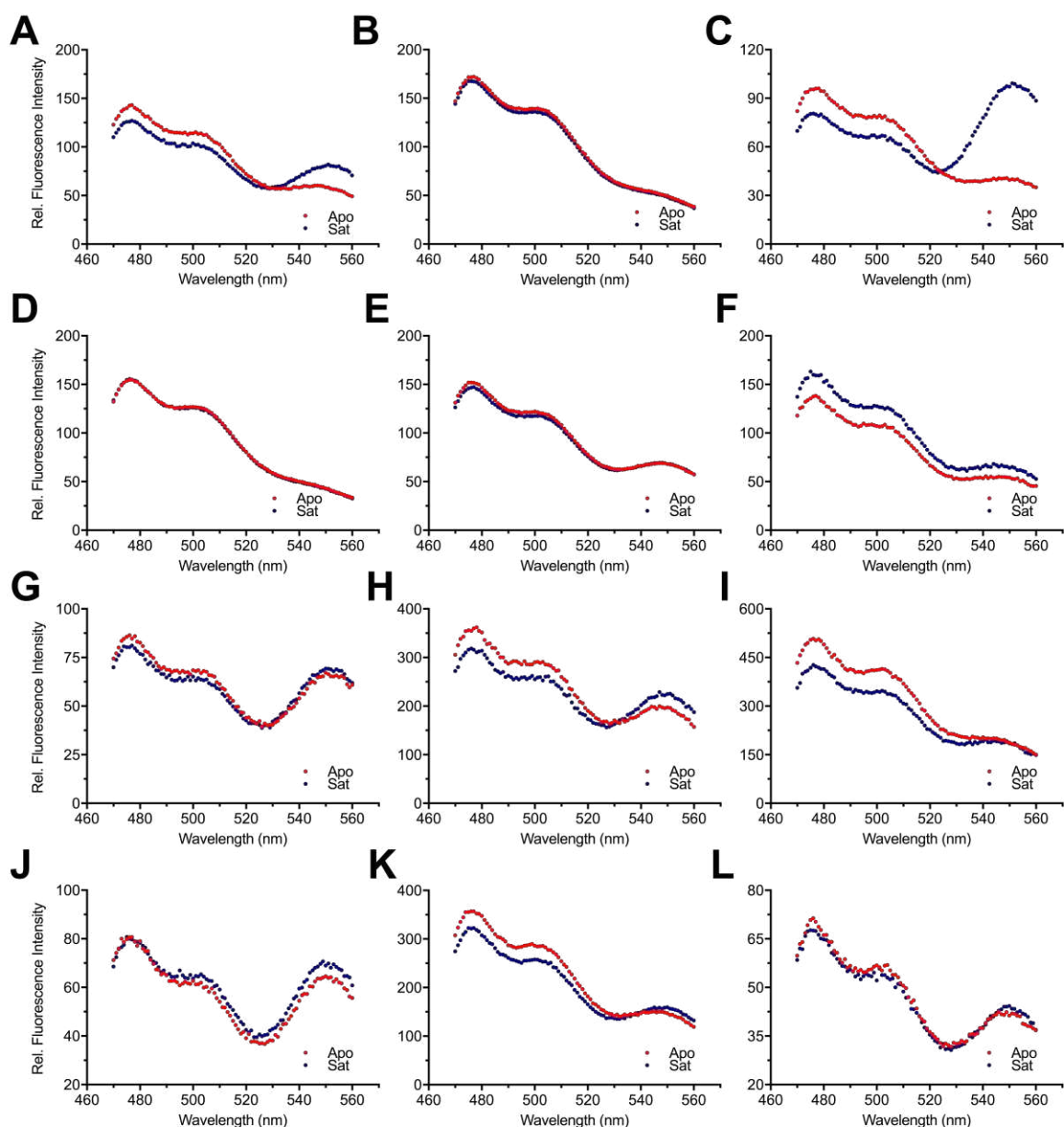


Figure S4. Fluorescence emission spectra for all variants. Spectra are shown as variants before (apo-state, red) and after the addition of saturating ligand (10-100 fold K_d ; holo-state, blue) with excitation at 433 nm. (A) Mal 381 spectrum showing 51% dynamic range (DR) upon addition of saturating maltose (predicted 48%). (B) Mal 393 spectrum 0% DR (predicted 7%). (C) Mal 437 spectrum, 311% DR (predicted 90%) (D) Mal 482 spectrum with -2% DR (predicted 5%). (E) Mal 524 spectrum with 4% DR (predicted 2%). (F) Sia 362 spectrum showing 2% DR (predicted 2%). (G) Sia 371 spectrum showing 11% DR (predicted 14%). (H) Sia 397 spectrum showing 32% DR (predicted 77%) (I) Sia 404 showing 17% DR (predicted 29%). (J) Sia 425 spectrum showing 11 % DR (predicted 27%). (K) Arg 345 spectrum showing 19 % DR (predicted 56%). (L) Arg 365 spectrum showing 14% DR (predicted 54 %).

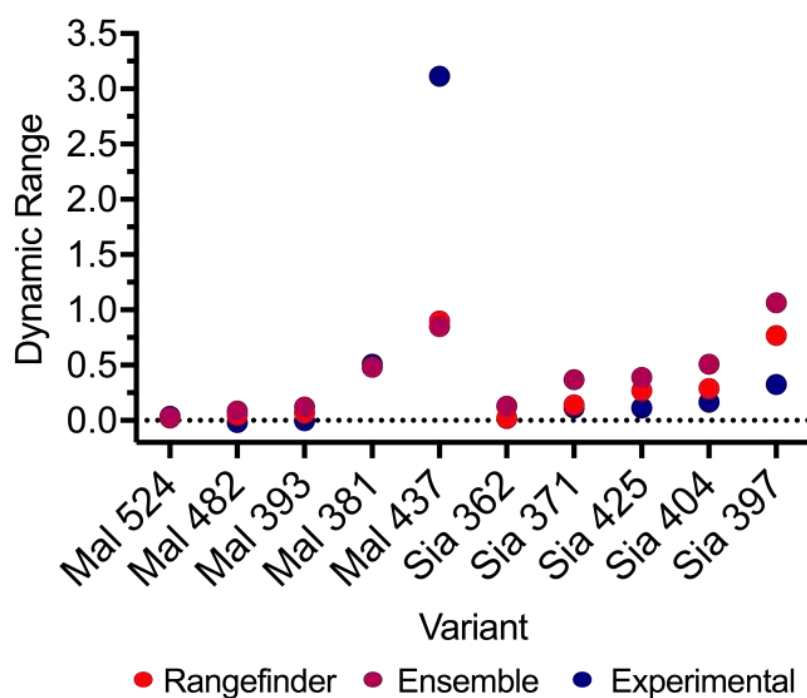


Figure S5. Overview of prediction accuracy for Rangefinder and the ensemble method for producing MBP- and SAB-based sensors. Experimental and predicted DRs are sorted by their predicted magnitude.

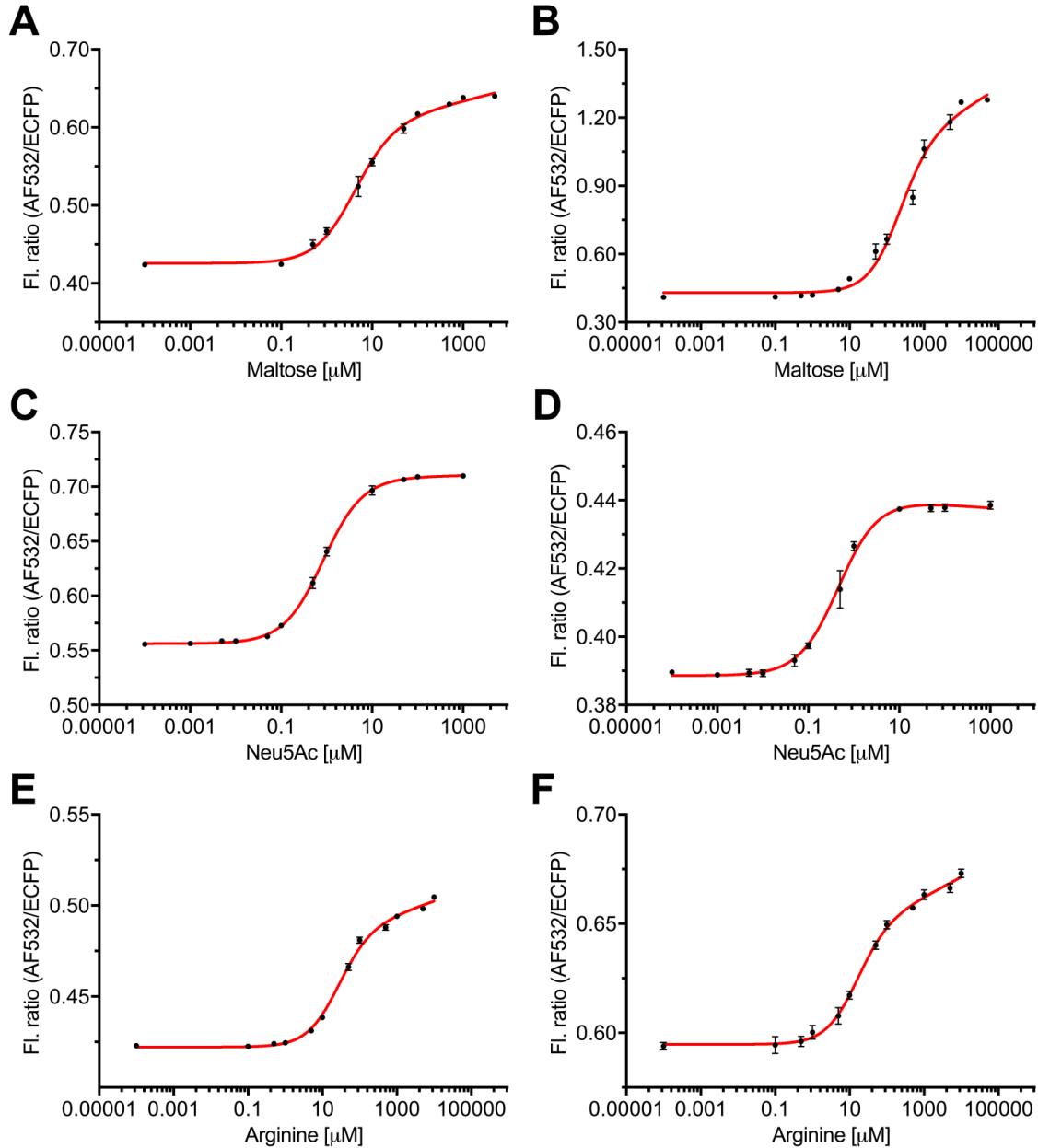


Figure S6. Fluorescence ratio (AF532/ECFP) titrations with increasing ligand concentrations. (A) Mal 381 with increasing concentrations of maltose. $K_d = 5.8 \pm 0.46 \mu\text{M}$. (B) Mal 437 with increasing concentrations of maltose. K_d of $390 \pm 72 \mu\text{M}$ (C) Sia 397 with increasing concentrations of Neu5Ac. K_d of $0.85 \pm 0.02 \mu\text{M}$ (D) Sia 404 with increasing concentrations of Neu5Ac. $K_d = .85 \pm 0.03 \mu\text{M}$. (E) Arg 345 with increasing concentrations of L-arginine. $K_d = 38 \pm 5.7 \mu\text{M}$. (F) Arg 365, with increasing concentrations of L-arginine. $K_d = 25 \pm 2.5 \mu\text{M}$.

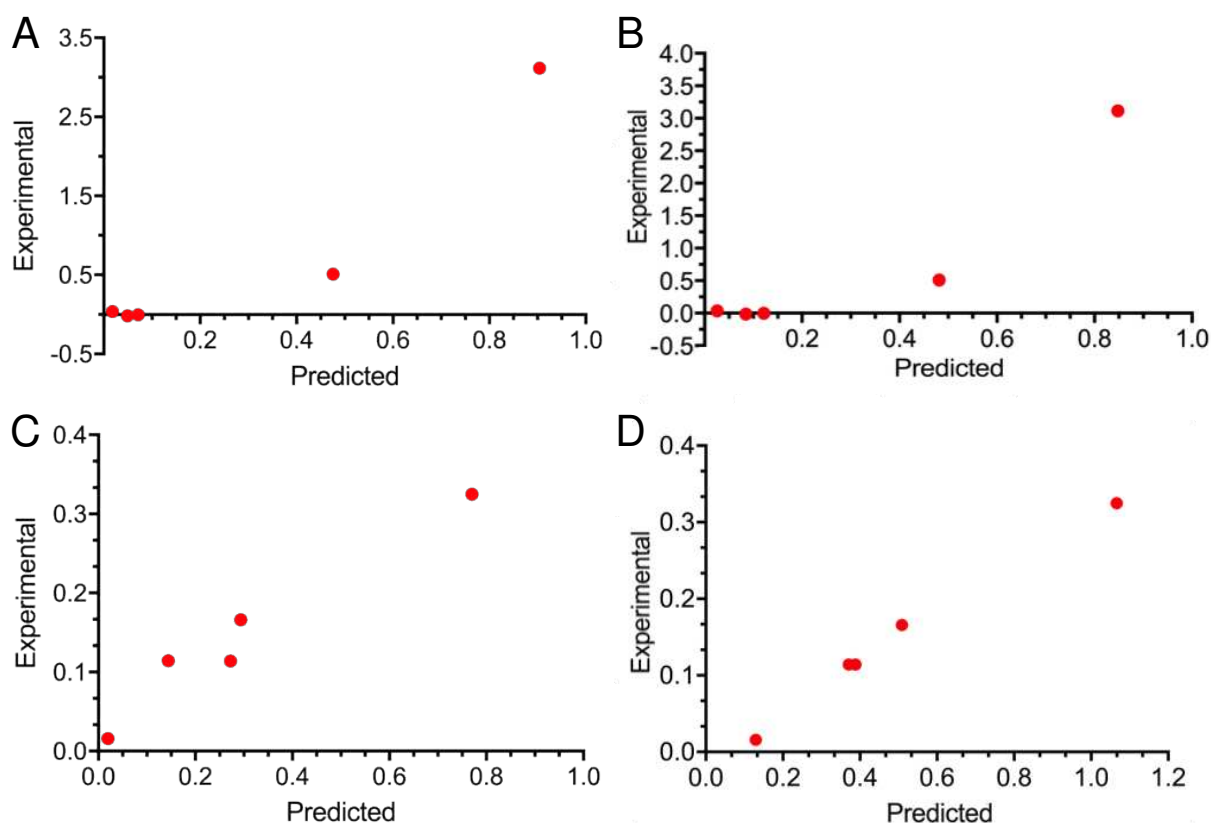


Figure S7. Correlation of Rangefinder Predicted vs. Experimental dynamic ranges for MBP and SAB. The MBP-Rangefinder (A) and MBP-ensemble (B) predictions correlated with $R_2 = 0.88$ and 0.86 , respectively, even when skewed by the unexpectedly large experimental dynamic range of Mal 437 (312%). Due to the experimental range of Mal 437 measuring approx. 3-fold higher, the relationship appears non-linear and masks the strong correlation of the other data points. The SAB-Rangefinder (C) and SAB-ensemble (D) predictions correlated with $R_2 = 0.96$ and 0.99 , respectively, showing linear relationships.

Table S1. Primers used for site directed mutagenesis and cloning.

NAME ^a	SEQUENCE
Fra-FOW^b	CAACGAGAAGCGCGATCACATGG
Fra-REV^b	GTTAGCAGCCGGATCTATCGATGCATGCCATGGTACCCGGGAGCTCGAATTC
Vec-FOW^c	GAATTCGAGCTCCCGGGTACC
Vec-REV^c	GATCCCGGCGGCGGTACGAAC
Sia 397-FOW	GCGTCTCCAACCCCTTGCGCATTTTCTGAGGTTTACTTGGCGTTG
Sia 397-REV	CAACGCCAAGTAAACCTCAGAAAATGCGCAAGGGGTTGGAGACGC
Sia 425-FOW	GCTGCCGTGCAGGCACAGTGCTTTTATGAGGTCCAAAAATTTCTG
Sia 425-REV	CAGAAATTTTGGACCTCATAAAAGCACTGTGCCTGCACGGCAGC
Sia 371-FOW	CGATAAATTCCATCGCAGATATGTGCGGCTTAAAGTTAAGAGTGC
Sia 371-REV	GCACTCTTAACCTTAAAGCCGCACATATCTGCGATGGAATTTATCG
Sia 362-FOW	CGGCAGACGACATCCAATTGCGCGATAAATTCATCGCAGATATG
Sia 362-REV	CATATCTGCGATGGAATTTATCGCGCAATTGGATGTGCTGTGCCG
Sia 404-FOW	GCATTTTCTGAGGTTTACTGCGCGTTGCAGACCAATGCTGTGGAC
Sia 404-REV	GTCCACAGCATTGGTCTGCAACGCGCAGTAAACCTCAGAAAATGC
Sia 476-FOW	CCGCGAAGTACCACACTTGCCTTTTTGTTGATGGCGAAAAGGATC
Sia 476-REV	GATCCTTTTCGCCATCAACAAAAAGGCAAGTGTGGTACTTCGCGG
Mal 381-FOW	GCACTTATGTTTAATTTATGCGAGCCGTATTTTACCTGGCCGTTG
Mal 381-REV	CAACGGCCAGGTAAATACGGCTCGCATAAATTAACATAAGTGC
Mal 393-FOW	CCTGGCCGTTGATAGCCGCATGCGGTGGATATGCGTTTAAGTAC
Mal 393-REV	GTACTTAAACGCATATCCACCGCATGCGGCTATCAACGGCCAGG
Mal 437-FOW	CAAGCATATGAATGCGGACACTTGCTACTCGATCGCCGAGGCTGC
Mal 437-REV	GCAGCCTCGGCGATCGAGTAGCAAGTGTCCGCATTTCATATGCTTG
Mal 482-FOW	GTATTGCCACATTTAAAGGTTGCCCATCAAACCATTCGTCGGC
Mal 482-REV	GCCGACGAATGGTTTTGATGGGCAACCTTTAAATGTGGGCAATAC
Mal 524-FOW	GACGAAGGTCTGGAGGCGGTAAATTGCGATAAACCCCTGGGTGC
Mal 524-REV	GCACCCAGGGGTTTATCGCAATTTACCGCCTCCAGACCTTCGTC
Arg-FOW^b	ATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCGGGATCCGTGGCAGACTGCGTGTTG
Arg 345-FOW	GTAAAAAAGTTGGTGTTCAAGTGCAGTACCGAGCGAACAGCATG
Arg 345-REV	CATGCTGTTGCTGCTGCTACCGCACTGAACACCAACTTTTTTAC
Arg 365-FOW	CAAAAGATGCCGGTGTTAAAGTGTGCAAATTCGACAACTTTAGCG
Arg 365-REV	CGCTAAAGTTGTGCAATTTGCACACTTTAACACCGGCATCTTTTG

^a Primers are labelled as Protein Variant-Mutation site-forward/reverse primer. The three letter codes are as follows: Fra- Generic primer complementary to the ECFP overlap region (forward primer) and T7 terminator region (reverse primer). Vec – vector (pEtMCSIII), Sia – SAB, Mal – MBP, Arg – AncQR. The number (###) denotes the mutated residue in the ECFP-SBP fusion construct. FOW denotes forward primers and REV denotes reverse primer sequences.

^b Primers used to generate the ends of each fragment complementary to the ECFP and pETMCSIII vector.

^c Primers used to create the vector backbone to yield 40 base pair overlaps complementary to the gene fragments (created using ^b primers). The Fra-FOW primer was paired with the protein variant-REV primer to create a fragment with the mutation at the 3' end and with the ECFP overlap at the 5' end. The Fra-REV primer was paired with the protein variant-FOW primers to create a DNA fragment with the mutation at the 5' end in a 40 bp region complementary to the first fragment and a region complementary to the plasmid T7 terminator sequence at the 3' end. These overlaps were sufficient to allow a one-pot Gibson assembly.

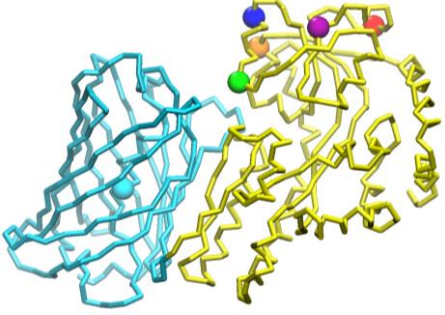
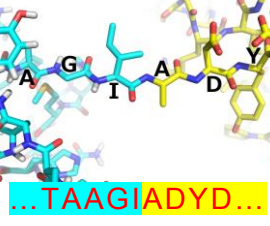
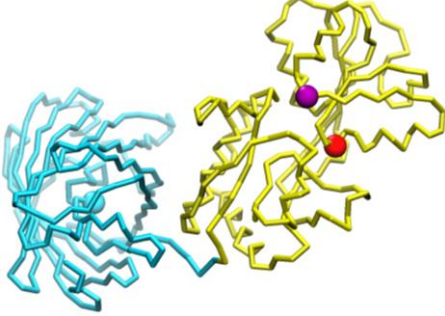
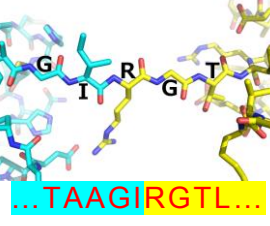
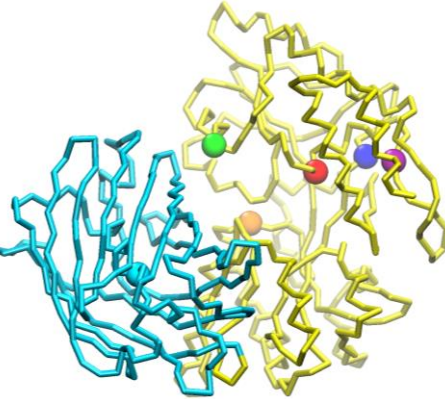

Table S2. Amino acid sequences of the ECFP-SBP fusion constructs.

FUSION CONSTRUCT	AA SEQUENCE ^{d,e}
ECFP-SAB	MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG KLTTLKFISTTGKLPVPWPTLVTTLTWGVQVFSRYPDHMKQH DFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNR IELKGIDFKEDGNILGHKLEYNYISHNVYITADKQKNGIKANF KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSAL SKDPNEKRDHMLLEFVTAAGIADYDLKFGMNAGTSSNEYK AAEMFAKEVKEKSQGGKIEISLYPSSQLGDDRAMLKQLKDGS LDFTFAESARFQLFYPEAAVFALPYVISNYNVAQKALFDTEF GKDLIKKMDKDLGVTLLSQAYNGTRQTTSN R AINSIADM K GL KLRVPNAATNLAYAKYVGASPTP M AFSEVY L ALQTNADVGGQ ENPLAAVQAQ K FYEVQKFLAMTNHILNDQLYLVSNETYKEL PEDLQKVVKDAAENAAKYHTKLFVDGEKDLVTFFEKQGVKI THPDLVPFKESMKPYYAEFVKQTGQKGESALKQIEAINP
ECFP-AncQR	MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG KLTTLKFISTTGKLPVPWPTLVTTLTWGVQVFSRYPDHMKQH DFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNR IELKGIDFKEDGNILGHKLEYNYISHNVYITADKQKNGIKANF KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSAL SKDPNEKRDHMLLEFVTAAGIRGTLRVGTEATFPFPFGFKD ENGKLVGFDIDLAKAIAKKLGVKVEFKPMDFDGIIPALQSGKI DVVIAGMTITEERKKQVDFSDPYFEAGQAIVVKKGNDSIKSL EDLKGGKKVGVL L GSTSEQHVKKVAKDAGVKV K KFDNFSEA FQELKSGRVDVVTDNAVALAYVKQNPAGVKIVGETFSGE PYGIAVRKGNSELLEKINKALEEMKKDGTYDKIYEKWFGE
ECFP-MBP	MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG KLTTLKFISTTGKLPVPWPTLVTTLTWGVQVFSRYPDHMKQH DFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNR IELKGIDFKEDGNILGHKLEYNYISHNVYITADKQKNGIKANF KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSAL SKDPNEKRDHMLLEFVTAAGIKIEEGKLVIWINGDKGYNGL AEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFW AHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDVAVRYNGKL IAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSA LMFNL Q EPYFTWPLIAAD D GGYAFKYENGKYDIKDVGVNDAG AKAGLTFLVDLIKNNKHMNADT D YSIAEAAFNKGETAMTINGP WAWSNIDTSKVNYGVTVLPTFKG Q PSKPFVGVLSAGINAAS PNKELAKEFLENYLLTDEGLEAVN K DKPLGAVALKSYYYEELA KDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASG RQTVDEALKDAQTRITK

^d Sequences are colour coded, amino acids highlighted in blue are the ECFP and those in yellow are the SBP. The termini were joined directly with no extra linker residues.

^e Mutated sites are illustrated as bold and underlined. Each of these sites was mutated to a cysteine and is numbered as the residue in the full ECFP-SBP fusion. In ECFP-SAB the mutations were: R362C, K371C, M397C, L404C, K425C. In ECFP-AncQR: L345C, K365C and in ECFP-MBP: Q381C, D393C, D437C, Q482C, K524C.

Table S3. Amino acid sequences of the ECFP-SBP fusion constructs.

FUSION CONSTRUCT	LABELLED RESIDUES ^f	FUSION SITE ^g	LEGEND
ECFP-SAB			R362C K371C M397C L404C K425C
ECFP-AncQR			L345C K365C
ECFP-MBP			Q381C D393C D437C Q482C K524C

^f Typical structures of SBP-FP fusion proteins taken from coarse-grain MD simulations and displayed as backbone beads only, with bonds between adjacent beads. The vertex of each bond is located at the centre of mass of the backbone atoms of the respective residue. ECFP domains are in cyan; SBP domains are in yellow. Spheres represent fluorophore locations: the ECFP fluorophore in cyan, and the labelled SBP residues in red, purple, green, blue and orange.

^g The s ECFP's C-terminus and SBP's N-terminus in an extended conformation. These structures were used as starting points for the coarse-grain MD simulations and demonstrate that both domains were directly fused without the use of additional linker residues in order to minimise the conformational freedom of the two domains and thereby maximise dynamic range. The relevant region of the construct's primary sequence is provided. (see also Table S2)

Table S4. Pearson's test for the Rangefinder and ensemble predictions of for the dynamic ranges of the MBP and SAB series variants.

	MBP		SAB	
	Rangefinder	Ensemble	Rangefinder	Ensemble
r	0.940	0.928	0.978	0.996
R squared	0.883	0.861	0.956	0.991
P value				
P (two tailed)	0.0176	0.0229	0.0039	0.0004
Significant? ($\alpha = 0.05$)	Yes	Yes	Yes	Yes
Number of XY pairs	5	5	5	5

References:

- (1) Tsien, R. Y. The green fluorescent protein. *Annu. Rev. Biochem* **1998**, 67, 509-544.
- (2) Lelimousin, M.; Noirclerc-Savoye, M.; Lazareno-Saez, C.; Paetzold, B.; Le Vot, S.; Chazal, R.; Macheboeuf, P.; Field, M. J.; Bourgeois, D.; Royant, A. Intrinsic dynamics in ECFP and Cerulean control fluorescence quantum yield. *Biochemistry* **2009**, 48, 10038-10046.
- (3) de Jong, D. H.; Singh, G.; Bennett, W. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* **2012**, 9, 687-697.
- (4) Schrödinger, L. The PyMOL molecular graphics system, version 1.8. **2015**,
- (5) Müller, A.; Severi, E.; Mulligan, C.; Watts, A. G.; Kelly, D. J.; Wilson, K. S.; Wilkinson, A. J.; Thomas, G. H. Conservation of structure and mechanism in primary and secondary transporters exemplified by SiaP, a sialic acid binding virulence factor from *Haemophilus influenzae*. *J. Biol. Chem.* **2006**, 281, 22212-22222.
- (6) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, 1, 19-25.
- (7) de Jong, D. H.; Baoukina, S.; Ingólfsson, H. I.; Marrink, S. J. Martini straight: Boosting performance using a shorter cutoff and GPUs. *Comput. Phys. Commun.* **2016**, 199, 1-7.
- (8) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, 25, 1422-1423.
- (9) Piston, D. W.; Kremers, G.-J. Fluorescent protein FRET: the good, the bad and the ugly. *Trends Biochem Sci* **2007**, 32, 407-414.
- (10) Whitfield, J. H.; Zhang, W. H.; Herde, M. K.; Clifton, B. E.; Radziejewski, J.; Janovjak, H.; Henneberger, C.; Jackson, C. J. Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* **2015**, 24, 1412-1422.
- (11) Neylon, C.; Brown, S. E.; Kralicek, A. V.; Miles, C. S.; Love, C. A.; Dixon, N. E. Interaction of the *Escherichia coli* replication terminator protein (Tus) with DNA: A model derived from DNA-binding studies of mutant proteins by surface plasmon resonance. *Biochemistry* **2000**, 39, 11989-11999.
- (12) Liu, H.; Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **2008**, 8, 1.
- (13) Gibson, D. G.; Young, L.; Chuang, R.-Y.; Venter, J. C.; Hutchison, C. A.; Smith, H. O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Meth.* **2009**, 6, 343-345.
- (14) Sivashanmugam, A.; Murray, V.; Cui, C.; Zhang, Y.; Wang, J.; Li, Q. Practical protocols for production of very high yields of recombinant proteins using *Escherichia coli*. *Protein Sci.* **2009**, 18, 936-948.

3.3 Method for developing optical sensors using a synthetic dye–fluorescent protein FRET pair and computational modeling and assessment

3.3.1 Statement of contribution

I declare that the research presented in this chapter represents original work that I carried out during my candidature at the Australian National University, except for contributions to multi-author papers incorporated in the chapter where my contributions are specified in this Statement of Contribution.

Publication status

This manuscript has been published with the title *Method for developing optical sensors using a synthetic dye–fluorescent protein FRET pair and computational modeling and assessment* as [chapter 6](#) of the book *Synthetic Protein Switches: Methods and Protocols*, edited by Viktor Stein. This book is volume 1596 of the Methods in Molecular Biology series published by Springer Science. The formatted article is reproduced in this chapter.

Authorship and contribution

The manuscript was authored by Joshua A. Mitchell (the author), William H. Zhang, Michel K. Herde, Christian Henneberger, Harald Janovjak, Megan L. O'Mara, and Colin J. Jackson. I developed the methodology, wrote all of the code, and performed experiments evaluating the method.

Chapter 6

Method for Developing Optical Sensors Using a Synthetic Dye-Fluorescent Protein FRET Pair and Computational Modeling and Assessment

Joshua A. Mitchell, William H. Zhang, Michel K. Herde,
Christian Henneberger, Harald Janovjak, Megan L. O'Mara,
and Colin J. Jackson

Abstract

Biosensors that exploit Förster resonance energy transfer (FRET) can be used to visualize biological and physiological processes and are capable of providing detailed information in both spatial and temporal dimensions. In a FRET-based biosensor, substrate binding is associated with a change in the relative positions of two fluorophores, leading to a change in FRET efficiency that may be observed in the fluorescence spectrum. As a result, their design requires a ligand-binding protein that exhibits a conformational change upon binding. However, not all ligand-binding proteins produce responsive sensors upon conjugation to fluorescent proteins or dyes, and identifying the optimum locations for the fluorophores often involves labor-intensive iterative design or high-throughput screening. Combining the genetic fusion of a fluorescent protein to the ligand-binding protein with site-specific covalent attachment of a fluorescent dye can allow fine control over the positions of the two fluorophores, allowing the construction of very sensitive sensors. This relies upon the accurate prediction of the locations of the two fluorophores in bound and unbound states. In this chapter, we describe a method for computational identification of dye-attachment sites that allows the use of cysteine modification to attach synthetic dyes that can be paired with a fluorescent protein for the purposes of creating FRET sensors.

Key words Synthetic dye, Optical sensor, Computational modeling, Förster resonance energy transfer

1 Introduction

Optical sensors have allowed for the investigation of physiological processes such as neurotransmission with both spatial and temporal resolution. FRET-based optical sensors are particularly useful, as they are capable of giving quantitative recordings independent of sensor concentration due to the ratiometric signal output of the sensor and the concentration independence of FRET donor lifetimes [1, 2]. Contemporary sensors typically use fluorescent proteins (FPs). However, some have used synthetic fluorescent

dyes rather than FPs as the signaling component of the sensor. As there is much greater control over the precise location of synthetic fluorophores, synthetic dyes can theoretically allow for even small conformational changes to produce a measurable FRET signal. While a large conformational change for a given protein is always desirable for sensor construction, it is often a necessity when using FPs, meaning that synthetic dyes are potentially applicable across a much larger range of proteins, rather than being restricted to those with large distance-based conformational changes.

Sensors that use synthetic dyes are either intensity-based sensors (non-FRET), which require multiple synthetic components, or must be developed through extensive high-throughput screening [3, 4]. For example, in addition to the use of two synthetic dyes, the Snifit-type sensor design requires the development and synthesis of a tethered competitive ligand that can occupy the binding active site, which necessarily introduces an extra design phase of engineering and screening [3]. On the other hand, the EOS-type sensor developed by Namiki et al. only requires a single synthetic component (a dye) [4], but still requires exhaustive screening of different residues to find a location that gives a strong signal. In addition, EOS sensors do not produce ratiometric output and are therefore not quantitative.

It is possible to improve on one or more of these design components when creating a synthetic dye-based sensor. Specifically, it has been shown that it is possible to create FRET sensors that are a combination of one FP and one synthetic dye [5], which is an improvement over both single fluorophore and two dye sensors as it is ratiometric and requires one fewer site-specific modification. Additionally, rather than using brute force high-throughput screening of all residue locations to identify a dye labeling site, we have expedited sensor development through the use of computational screening, which can reduce the number of possible dye labeling sites to a subset with a higher likelihood of yielding a functional sensor. We have used these computational techniques in tandem with synthetic dyes (via thiol-maleimide labeling of residues) to develop optical sensors with large dynamic ranges.

2 Materials

Whenever possible, prepare all stock solutions and buffers in ultra-pure water (MilliQ). For reagents that have poor solubility in water, dissolve them with the smallest possible proportion of organic solvent (i.e., 5% DMSO would be preferable to 10% DMSO) as some proteins may have poor stability in organic solvent. All solutions and reagents should be prepared as fresh as possible as some reagents will have short lifetimes when in solution, even at -20°C . All buffer solutions should be filtered through a membrane filter (pore size $0.45\text{ }\mu\text{m}$ or smaller) after preparation.

2.1 Software

1. Python 2.7.11 (www.python.org).
2. Bash 4.2.46 (www.gnu.org/software/bash/).
3. GROMACS 5.1.2 (www.gromacs.org) [6].
4. MARTINI 2.1 (<http://md.chem.rug.nl/>) [7].
5. PyMol 1.7.6.0 (<https://sourceforge.net/projects/pymol>).
6. GAWK 4.0.2 (www.gnu.org/software/gawk/).
7. DSSP 2.2.1 (<http://swift.cmbi.ru.nl/gv/dssp/>).
8. Grep 2.5.1 (<https://www.gnu.org/software/grep/>).
9. Curl 7.29.0 (<https://curl.haxx.se/>).

2.2 Cloning and Protein Purification Components

1. The gene of interest in an expression vector (*see* Subheading 3.3).
2. Primers containing the mutation of interest (*see* Subheading 3.3).
3. High-fidelity DNA polymerase kit.
4. Gibson assembly kit.
5. Thin-walled PCR tubes.
6. PCR thermocycler.
7. PCR purification kit.
8. Competent cells for cloning (Top10).
9. Competent cells for protein expression (BL21DE3).
10. Materials for colony PCR and analysis by gel electrophoresis:
 - (a) PCR master mix.
 - (b) T7 primers.
 - (c) 1% agarose gel made with SB buffer (46 g/L boric acid, 8 g/L sodium hydroxide), with a visualizing stain.
 - (d) DNA ladder mix.
 - (e) Agarose gel electrophoresis apparatus.

2.3 Dye Labeling Components

1. Buffer solution, Phosphate buffer (*see* **Note 1**): 0.05 M Sodium phosphate, 0.2 M NaCl. Adjust the pH to what is appropriate for both the protein of interest and the chemistry that is needed to label the protein with the synthetic dye (using either hydrochloric acid or sodium hydroxide).
2. TCEP stock solution: dissolve 0.1437 g of Tris(2-carboxyethyl) phosphine hydrochloride (TCEP-HCl) in 1 mL of buffer solution (501 mM stock solution) (*see* **Note 2**).
3. Dye stock solution: Add a suitable solvent to the synthetic dye to achieve a stock solution with a final concentration of 10 mM. In the case of the Alexa Fluor 532 C5 Maleimide, 1 mg was dissolved in 123 μ L of phosphate buffer (10 mM stock solution) (*see* **Note 3**).

4. Protein solutions: proteins should be concentrated as much as practical (ideally at least 500 μM) and exchanged or dialyzed into the same buffer solution as used to prepare the reagent stock solutions (*see* **Note 4**).

3 Methods

All experimental (noncomputational) work should be performed at 4 °C unless otherwise specified.

3.1 Computational Screening and Residue Selection

First, prepare models of the target fusion protein in both its bound and unbound conformations. This involves modeling both the solute-binding and fluorescent domains with the appropriate linker. Start with crystal structures or high-quality homology models of the desired binding core in both conformations and the fluorescent protein. Ensure no nonstandard residues exist in the models, as these are not parameterized in the MARTINI forcefield (*see* **Note 5**). Reconstruct any residues missing from the crystal structures at the termini by which they are fused. Then, construct the linker sequence as expressed experimentally (*see* **Note 6**). Complete the model by fusing the two domains (*see* **Note 7**). Save both SBP-linker-FP models as .PDB files.

Create a directory for each conformation, copy the appropriate .PDB file into each, and also copy simulations.sh into each (*see* **Note 8**). The script simulations.sh automatically prepares and runs a number of MARTINI simulations. It depends on all of the software in Subheading 2.1 except MARTINI, which is downloaded automatically by the script, provided all software dependencies are available in your \$PATH (*see* **Note 9**). The script can be configured by making a copy of it in each conformation directory and editing the leading “CONFIG” portion. Most defaults should be appropriate; however, the input_file, system_name and linker_residues variables should be set for your system (*see* **Note 10**). Configure and run the script, read and follow the directions given, then repeat for the other conformation.

Run the second script “process-data.py,” giving as the first two arguments the locations of each set of output files. The residue number of the central residue of the FP fluorophore and the range of residues that should be checked for sensor generation should also be given as arguments, respectively, with the -f and -r switches (*see* **Note 11**). process-data.py reads the given .PDB files, and predicts dynamic ranges for sensors formed by labeling each residue in the given range with a dye with configurable Forster distance. This output is stored by default as comma-separated values with appropriate headers in sensor_predict.csv and can be visualized graphically using a program such as Excel.

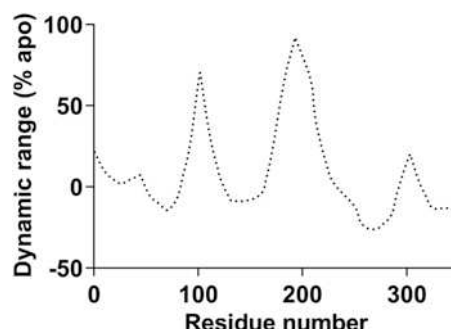


Fig. 1 A graphical representation of the script output. The data shown in this instance is hypothetical and is not based on a true set of calculations. The script determines the dynamic range for a hypothetical sensor that has been labeled with a fluorescent dye at a given residue. For example, for a construct with an ECFP fused at the N-terminus of the binding protein, a sensor construct that has been labeled at residue 50 is predicted to have a very small or nonexistent dynamic range. Conversely, for a sensor that has been labeled at residue 200, the dynamic range should be large, as it is approaching the theoretical maximum dynamic range possible for the construct

Select residues that are appropriate for cysteine mutagenesis (or mutagenesis to an appropriate residue). Note that this prediction does not account for any disruption of binding core function associated with chemical labeling. Therefore, a residue that yields the largest predicted dynamic range may not necessarily yield the best sensor, as the residue may have some structural or functional importance, which may be disrupted with mutagenesis. Residues with side chains oriented toward the solvent, or that are not a part of a structural motif should be selected preferentially. In the hypothetical data set example (Fig. 1), residue 200 is predicted to yield a large dynamic range upon labeling with a dye. Suppose, however, that for this hypothetical protein residue 200 is both not exposed to solvent and has its sidechain oriented toward the binding site of the protein (Fig. 2). Mutating and labeling this residue may abolish protein function, as the sterically bulky dye excludes the substrate from the active site. In contrast, although residue 100 has a smaller predicted dynamic range than residue 200, it is located on a flexible loop that does not have significant contact with other structural features of the binding protein. This makes this location preferable to residue 200, as labeling the residue is less likely to impact the correct function and dynamics of the binding domain, while still providing an excellent dynamic range.

3.2 Cloning and Purification of the Mutant Proteins

1. The sensor construct (SBP fused with the fluorescent protein) should be first cloned into an expression vector (with a T7 promoter). The sequence to be cloned should match the sequence used to model the sensor exactly, with the exception

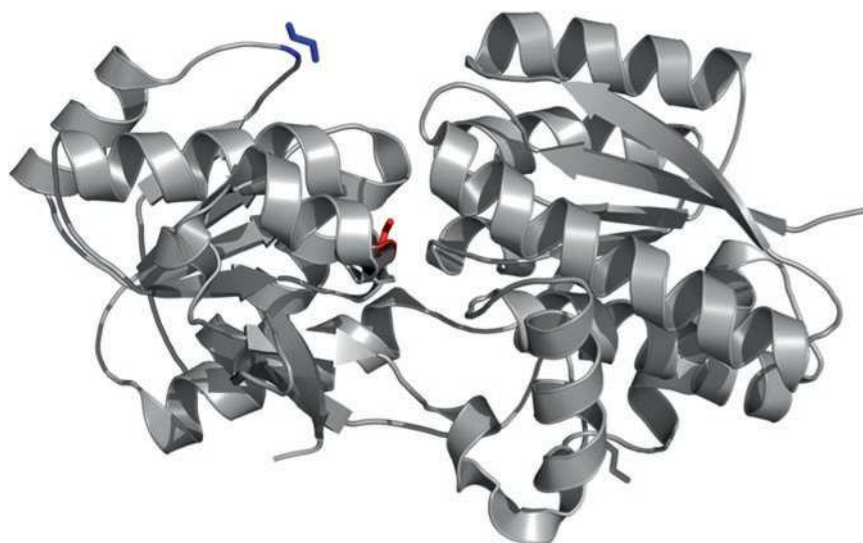


Fig. 2 A generic structure of an amino acid-binding protein (PDB 3IP9) used to illustrate that residue location and function needs to be considered along with the computational prediction. Residues 100 (*blue*) and 200 (*red*) are shown. Although labeling residue 200 would produce a sensor with the theoretically largest dynamic range (Fig. 1), in reality, as this would require the dye to occupy the ligand-binding site between the two domains, it would not result in a functional sensor. Alternatively, residue 100 is also predicted to yield a sensor with a significant dynamic range and is located on a flexible loop, where chemical modification and mutagenesis is less likely to negatively impact correct function of the binding protein

that there can be a histidine tag at the C-terminus of the fluorescent protein to facilitate protein purification.

2. Next, cysteine mutants of this sensor should be created at the residue locations identified by the computational screening (*see Note 12*). Any unwanted surface cysteines should be mutated to alternative residues to avoid nonspecific labeling (*see Note 13*). Although many cloning methods are suitable to introduce mutations, our preferred method is Gibson assembly.
3. In order to create the cysteine mutants through Gibson assembly, first synthesize or order a set of complementary primers (forward and reverse primers encoding the same sequence) that encompasses the residue of interest, with the total length of the primer between 30 and 50 nucleotides. The nucleotides coding the residue of interest should be changed to encode a cysteine residue, all other residues should match the template DNA exactly.
4. For the PCR, these primers will then be paired with the T7 promoter/terminator primers for PCR amplification. The T7 promoter forward primer is paired with the reverse mutagenic primer, while the T7 terminator reverse primer is paired with the mutagenic forward primer.

5. The PCRs using these primer sets should result in two fragments, with one fragment overlapping with the T7 promoter region and the mutagenic region, and the other fragment overlapping with the mutagenic region and the T7 terminator region. The mutagenic regions of these fragments will overlap and anneal during Gibson assembly. Gel purification is highly recommended to improve cloning efficiency, even if the agarose gel electrophoresis of the PCR product shows a single clear band.
6. Linear vector for use in the Gibson assembly reaction can be made through PCR using complementary primers of the T7 regions. Specifically, a T7 terminator forward primer and a T7 promoter reverse primer should be used. The linearized vector produced from this PCR reaction will have T7 regions that can overlap with the T7 regions of the gene fragments for the Gibson assembly reaction.
7. The PCR fragments (for both the sensor and vector) can then be combined into the Gibson master mix (in equimolar ratios). The typical incubation time for the master mix is 50 °C for 1 h.
8. The Gibson assembly mixture should then be transformed into competent cells that are designed for DNA cloning (e.g., TOP10 cells). If electrocompetent cells are used, the Gibson mix can be diluted with water to prevent arcing.
9. After confirming that the cloning is successful through sequencing, the gene should be transformed into an expression cell type (e.g., BL21DE3).
10. If the sensor construct contains a histidine tag, it can then be purified through nickel affinity chromatography.

3.3 Dye Labeling and Purification

1. To a volume of 849 μL buffer add 100 μL of the concentrated protein solution (assuming the concentration of the protein solution is 500 μM) and 1 μL of the TCEP solution. Wait approximately for 5 min to allow this solution to equilibrate to room temperature and for any disulfides to be reduced. This reaction can be performed in an Eppendorf tube or an equivalent (*see Note 14*).
2. To the reaction mixture add 50 μL of the dye stock. The final composition of the reaction mixture should contain approximately 50 μM of protein, 500 μM TCEP, and 500 μM dye. The reaction should be allowed to proceed overnight (16 h) at 4 °C in the dark with constant agitation (*see Note 15*).
3. After the reaction period, centrifuge the reaction mixture at high speed (18,000 $\times g$, 5 min) to separate out any precipitated dye or protein (*see Note 16*).
4. The mixture should then be purified through gel filtration, with a desalting column usually being sufficient (*see Note 17*).

5. After one round of purification with the desalting columns, the protein mixture should be re-concentrated and buffer exchanged using centrifugal protein concentrators, which will remove trace TCEP and further remove unreacted and free dye.
6. The protein should then undergo a final desalting step to ensure that no free dye or TCEP remains and labeling efficiency can be evaluated using the following equation.
7. Moles of *dye per mole* protein = $\frac{A}{\epsilon \times C}$

where for a given sample “*A*” is the absorbance at the peak excitation wavelength of the dye in use, “*ε*” is the extinction coefficient of the dye, and “*C*” is the concentration of protein.

4 Notes

1. Phosphate buffer might not be compatible with all proteins; substitute with an appropriate buffer if needed, but ensure that the pH range is compatible with the dye system in use. In the case of thiol-maleimide conjugations, the desired pH is between 7.0 and 7.5.
2. TCEP, or reducing agents in general, are typically only necessary for thiol based conjugation. TCEP is far preferable to other reducing agents such as DTT, as under normal conditions TCEP will not interfere with the maleimide-thiol reaction and also does not have issues with odor.
3. Sometimes there may be trace precipitate or undissolved dye in the stock solution. This is not a cause for concern so long as the precipitate is suspended homogenously prior to use. This trace precipitate will dissolve slowly over time as the reagent is consumed.
4. The protein should be purified to the highest degree possible, as other proteins could be labeled by the dye, which can affect the observed dynamic range of the protein sample. If additional purification is needed, size exclusion chromatography can be performed either before or after the labeling reaction.
5. For fluorescent proteins, this generally means mutating the fluorophore back to the three autocatalytic residues that form it; other proteins may incorporate nonstandard residues as a result of chemical modification for crystallization, etc. Unrecognized residues can be mutated to glycine by opening the .PDB file in a text editor, deleting the side chain atoms, and changing the residue names to GLY. They can then be mutated to the appropriate residue with PyMOL’s mutagenesis wizard.

6. We found this easiest to do by working out from both domains and allowing the constructed extended linker model to meet in the middle. Residues can be added to an existing model in PyMOL in Editing mode, accessed by clicking the mouse key shortcuts box in the lower right corner of the viewer window while in the default Viewing mode. Once in Editing mode, select the N-terminal nitrogen or C-terminal carbonyl carbon by clicking on it, then add residues by holding Alt and typing the single-letter code associated with the desired amino acid.
7. In PyMOL's Editing mode, select both atoms that should be bonded in the product, then run PyMOL's fuse command. For example if a model with an ECFP on the N terminus of the protein is desired, start by selecting both the amine of the N terminus of the protein and the carbonyl carbon of the C terminus of the ECFP. Then, while both atoms remain selected, enter "fuse" as a command input, which will generate an approximate model of the fusion protein. The fuse command may sometimes orient the proteins poorly; make sure to rotate the proteins as such that they do not overlap in physical space and the linker is fully extended. This can be done in editing mode by holding shift and right-clicking on a bond to rotate the associated torsion angle.
8. For example, you may be working in a subfolder of your home directory called ~/dyes. You should create new directories for each conformation, perhaps ~/dyes/open and ~/dyes/closed. You add the script and starting structure to each directory, and are left with the following files:


```
~/dyes/open/open-start.pdb
~/dyes/open/simulations.sh
~/dyes/closed/closed-start.pdb
~/dyes/closed/simulations.sh
```
9. The script will then generate folders for setup, each run, and the trimmed, fitted trajectories as pdb files:


```
~/dyes/open/setup/
~/dyes/open/run1/, ~/dyes/open/run2/,
~/dyes/open/run3/ etc.
~/dyes/open/results/
```
10. The version numbers given in the materials are known to work; other versions will probably work as well but have not been tested. Note that MARTINIZE.py, which is downloaded and run by simulations.sh, is not compatible with Python 3, and thus Python 2 must be available on your system for the setup steps. The typical name for the DSSP executable varies from system to system; simulations.sh can be told the correct name for your system either by editing the dssp_name variable or by passing the correct name as an argument

with the `-dssp_name` switch. Finally, GROMACS versions prior to 5.0 are not supported.

11. By default, `simulations.sh` prepares, equilibrates, and runs all simulations it is asked to without taking a break. However, if it is terminated, it can restart from the beginning of the last step it finished successfully, so it may be used (with some modification) as a resubmit script for systems like PBS. If preferred, the `-o` switch can be supplied to the script, which will perform the setup steps and generate individual run folders, but will not perform the full-length production runs. The full-length simulations can then be run on whatever hardware is appropriate by simply running `gmx grompp` on the provided `.MDP` files and starting structures.

The script first prepares the MARTINI coarse-grained force field topology, and converts the provided `.PDB` file to a coarse-grained model with the script `Martinize`. This includes construction of an elastic network around both protein domains, which keep their conformations constant and allow sampling of linker collapse. It then energy minimizes and solvates the model, including addition of neutralizing ions, anti-freeze MARTINI water, and experimental salt concentration if desired. It performs a second energy minimization, and then equilibrates the system with thermostat and barostat in several rounds of progressively weaker backbone-restrained MD. Force constants of 1000, 500, 100, 50, and 10 are used by the script. The production run involves 30 replicate 200 ns simulations. New velocities are generated for each of these runs in a final unrestrained equilibration step. These simulations take approximately 90 min each for a 600 residue fusion model running on dual Tesla K40s with 32 cores. Finally, it outputs the last 500 ns of each simulation, sampled in 1 ns intervals, as a `PDB` file to the results directory.

For instance, if **step 2** was performed in directories called `~/dyes/open` and `~/dyes/closed`, and the script is being run in `~/dyes`, per the example in **Note 8**, a typical call for a fusion model with the binding protein occupying residues 242–586 and the fluorophore at residue 63 might be:

```
python process-data.py open/results/*.  
pdb closed/results/*.pdb -f 63 -r 242-586
```

12. While cysteine mutagenesis is a functional method of chemically labeling a protein with a synthetic dye, alternative chemistries are possible through the use of unnatural amino acid incorporation. This can allow for biorthogonal labeling of proteins *in vivo*, and in principle, it can allow for the sensor to be genetically encoded in an organism capable of utilizing the required unnatural amino acid.

13. There are two conserved cysteine residues in the GFP family; we have had success with the mutations C49S and C71V with minimal fluorescence loss, as described by Suzuki et al. [8].
14. The reaction volume and reagents should be scaled appropriately, relative to the concentration of the protein stock solution.
15. When agitating the solution, care should be made that the solution does not begin to form froth or foam as this can lead to precipitation of protein.
16. Filtration can be used instead of centrifugation; however, there is typically some volume/yield loss when filtering small volumes.
17. PD-10 columns (GE healthcare) are usually sufficient to separate free dye from the protein. If the purity of the protein is a concern, the protein should be first buffer exchanged to remove excess TCEP and then purified with SEC (GE healthcare, Hiload 26/600 superdex 200pg, adequate for most proteins). This should separate labeled protein from free dye and any contaminant proteins.

References

1. Scanziani M, Hausser M (2009) Electrophysiology in the age of light. *Nature* 461(7266):930–939
2. Hou B-H, Takanaga H, Grossmann G, Chen L-Q, Qu X-Q, Jones AM, Lalonde S, Schweissgut O, Wiechert W, Frommer WB (2011) Optical sensors for monitoring dynamic changes of intracellular metabolite levels in mammalian cells. *Nat Protoc* 6(11):1818–1833. <http://www.nature.com/nprot/journal/v6/n11/abs/nprot.2011.392.html#supplementary-information>
3. Masharina A, Reymond L, Maurel D, Umezawa K, Johnsson K (2012) A Fluorescent sensor for GABA and synthetic GABAB receptor ligands. *J Am Chem Soc* 134(46):19026–19034. doi:10.1021/ja306320s
4. Namiki S, Sakamoto H, Iinuma S, Iino M, Hirose K (2007) Optical glutamate sensor for spatiotemporal analysis of synaptic transmission. *Eur J Neurosci* 25(8):2249–2259. doi:10.1111/j.1460-9568.2007.05511.x
5. Suzuki M, Tanaka S, Ito Y, Inoue M, Sakai T, Nishigaki K (2012) Simple and tunable Förster resonance energy transfer-based bioprobes for high-throughput monitoring of caspase-3 activation in living cells by using flow cytometry. *Biochim Biophys Acta* 1823(2):215–226. doi:10.1016/j.bbamcr.2011.07.006
6. Hess B, Kutzner C, Van Der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4(3):435–447
7. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink S-J (2008) The MARTINI coarse-grained force field: extension to proteins. *J Chem Theory Comput* 4(5):819–834
8. Suzuki T, Arai S, Takeuchi M, Sakurai C, Ebana H, Higashi T, Hashimoto H, Hatsuzawa K, Wada I (2012) Development of cysteine-free fluorescent proteins for the oxidative environment. *PLoS One* 7(5):e37551. doi:10.1371/journal.pone.0037551

Chapter 4

Monitoring hippocampal glycine with the computationally designed optical sensor GlyFS

4.1 Preface

4.1.1 Fluorescence in Synthetic Biology

Fluorescence came into its own as a method for studying biological systems with the discovery of the Green Fluorescent Protein (figure 4.1). GFP is a simple beta-barrel protein of about 27 kDa (238 residues) that catalyses the reaction of three of its own residues into a fluorophore. These three residues are located in the middle of a partially helical strand that passes through the centre of the barrel, which then protects the mature fluorophore from solvent quenching effects.

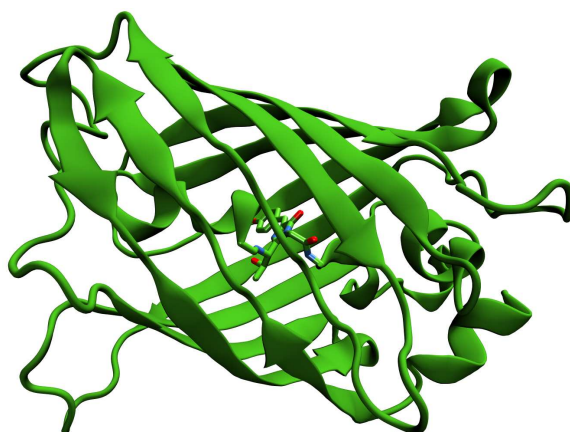


Figure 4.1: The Green Fluorescent Protein (PDB ID 4KW4). The 11-strand β -barrel surrounds a hydrophobic α -helix, at the centre of which the lies the SYG triplet that matures into the fluorophore. Here, the matured fluorophore is shown as sticks.

In this way, GFP exhibits protein fluorescence without requiring any cofactor or substrate apart from oxygen.^{278,279} Many fluorescent proteins of varying colours and properties have been developed in the lab by mutating GFP, and especially by modifying the fluorophore motif.^{252,253}

GFP-family fluorescent proteins are so revolutionary because they are entirely genetically encodable. As the amino acid sequence gives rise to fluorescence directly without needing any extrinsic influence after translation, any translation machinery capable of working with the 20 canonical amino acids is in principle able to produce a fluorescent protein. This makes them easy to use in organisms where fluorescent cofactors may be toxic or expensive, or where eukaryotic post-translational modification mechanisms are absent. In research applications, the fluorophore can even be expressed *in situ* by the model organism itself. When the protein is desired independently of the organism, this flexibility also opens the door to simplified expression systems.

GFP has been used as a cheap, specific and bio-compatible fluorescent label. By simply fusing a target protein to GFP in a genetically modified model organism, the dynamics of the target protein can be visualised under a microscope. This eliminates toxicity and specificity concerns that traditional stains exhibit and in some cases allows single molecules to be tracked in the cell in real time. While GFP-family proteins cannot attain the brightness and other fluorescent properties available to synthetic fluorophores, the opportunities made available by the promise of genetic encoding make them an essential and proven part of the fluorescent toolkit. In the paper presented below, two GFP variants are combined with a rationally engineered binding core to produce a ratiometric glycine sensor called GlyFS.

4.1.2 Fluorophore dynamics and linkers

Our previous development of semi-synthetic bio-sensors, discussed in chapter 3, highlighted the importance of the dynamics of the fluorophores within the fusion construct. Here, we extended these findings to optimise a genetically encoded sensor by varying the linker connecting it to the sensing domain. Linkers are used nearly universally in fusion proteins.²⁸⁰ At their simplest, long, unstructured linkers of dozens of flexible, hydrophilic residues are used to allow the two domains to fold independently.^{281,282} These long, flexible linkers are thought to act as a tether, allowing each domain to retain its intrinsic structure and dynamics while being relatively co-localised in space. Alternatively, linkers with a propensity for forming α -helices are used to keep the two domains at a fixed distance from each other.^{283,284}

The composition and length of the linker very often influences or controls the quality of a fusion protein. This is well known in fluorescent sensors, where the location of the domains

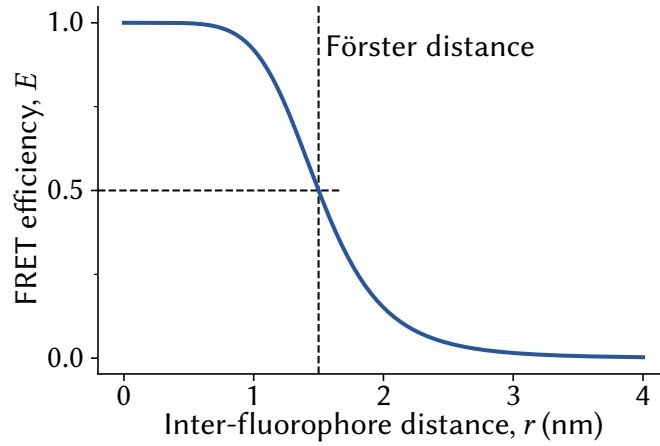


Figure 4.2: FRET efficiency plotted as a function of inter-fluorophore distance (equation 4.1). The efficiency is sigmoidal in distance, with a midpoint at the characteristic Förster distance. See also section 3.1.2.

plays an important role in performance,^{271,285,286} but is also found in other synthetic fusion proteins^{280,287–291} and even natural multidomain proteins.^{292–295} Rational design of linkers has been pursued, invariably by designing libraries and selecting the best construct in the lab.^{281,282,296}

For the FRET-based ratiometric sensors investigated here, the source of this dependence is relatively clear. Recall the FRET equations, discussed in section 3.1.2 and plotted in figure 4.2:

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6} \quad (4.1)$$

$$R_0^6 = K_F \kappa^2 Q_D n^{-4} \int_0^\infty \epsilon_A(\lambda) F_D(\lambda) \lambda^4 d\lambda \quad (4.2)$$

Optimising a sensor means maximising the difference between FRET efficiencies E in the bound and unbound states. The efficiency depends on the locations of the two fluorophores in at least four ways:

1. Straightforwardly, the distance between fluorophores r
2. The orientation factor of one fluorophore to the other, κ^2
3. The index of refraction n of the path the virtual photon takes, which is different in water and in protein
4. Any effects that the locations have on their fluorescence properties ϵ , Q or F

The first two of these are direct effects built into the equations, have clear directionality in their effects, are much easier to model, and are expected to have a larger influence on efficiency than the latter two, and are therefore the focus of our optimisation efforts.

Any effect the linker has on location can therefore translate into a change in performance. If the location effect is different in the two states this is obvious, but even if the effect is the same in both states this can move the average efficiency closer to 50%, where its gradient with respect to distance is greatest (see figure 4.2). The same change in inter-fluorophore distance associated with binding can then lead to a greater change in efficiency and observed fluorescence ratio. This unfortunately means that the directionality of the effects of location on FRET efficiency do not always translate into directionality in sensor performance. A linker that improves one sensor can therefore worsen another.

4.1.3 Simplified modelling as a guide to intuition

The design of the GlyFS began with Dr. William H. Zhang's re-engineering of a promiscuous GABA-binding protein Atu2422 into a glycine-specific binding protein. This was done via three binding site mutations and is described at length in the text. However, fusing this core to the fluorescent proteins ECFP and Venus produced a very poor sensor. It was only through linker engineering that GlyFS could be developed.

The same MD simulations that inspired Rangefinder (section 3.1.1) also inspired the linker choice that made GlyFS a satisfactory sensor. While these simulations were simplistic and not terribly accurate, they worked effectively as an intuition pump highlighting the wide variety of states available to the sensor as well as a proteins preference for associating with itself rather than being maximally solvated. MD simulations were able to achieve this because they are developed bottom-up from the underlying physics, and so did not rely on our preconceptions and top-down theories about how fusion proteins behave.

4.2 Statement of contribution

I declare that the research presented in this chapter represents original work that I carried out during my candidature at the Australian National University, except for contributions to multi-author papers incorporated in the chapter where my contributions are specified in this Statement of Contribution.

4.2.1 Publication status

This manuscript has been published with the title *Monitoring hippocampal glycine with the computationally designed optical sensor GlyFS* in the journal **Nature chemical biology** (2016, [14:861–869](#)). The formatted article with supporting information is reproduced in this chapter.

4.2.2 Authorship and contribution

The manuscript was authored by William H. Zhang, Michel K. Herde, Joshua A. Mitchell (the author), Jason H. Whitfield, Andreas B. Wulff, Vanessa Vongsouthi, Inmaculada Sanchez-Romero, Polina E. Gulakova, Daniel Minge, Björn Breithausen, Susanne Schoch, Harald Janovjak, Colin J. Jackson and Christian Henneberger. WHZ and MKH contributed equally to the work. I contributed an important insight that led to one of the major innovations in the GlyFS sensor (described in section [4.1.3](#)), performed MD simulations that did not appear in the final manuscript, and otherwise contributed to the design and analysis of the sensor and editing figures and writing. In addition, I contributed reviewer-requested clarifications and explanations and computer models of the sensor's unusual negative dynamic range, which appear in supplementary figures 3, 5, and 7.

Monitoring hippocampal glycine with the computationally designed optical sensor GlyFS

William H. Zhang^{1,9}, Michel K. Herde^{2,9}, Joshua A. Mitchell¹, Jason H. Whitfield¹, Andreas B. Wulff², Vanessa Vongsouthi¹, Inmaculada Sanchez-Romero³, Polina E. Gulakova⁴, Daniel Minge², Björn Breithausen², Susanne Schoch⁴, Harald Janovjak^{3,5,6}, Colin J. Jackson^{1*} and Christian Henneberger^{2,7,8*}

Fluorescent sensors are an essential part of the experimental toolbox of the life sciences, where they are used ubiquitously to visualize intra- and extracellular signaling. In the brain, optical neurotransmitter sensors can shed light on temporal and spatial aspects of signal transmission by directly observing, for instance, neurotransmitter release and spread. Here we report the development and application of the first optical sensor for the amino acid glycine, which is both an inhibitory neurotransmitter and a co-agonist of the N-methyl-D-aspartate receptors (NMDARs) involved in synaptic plasticity. Computational design of a glycine-specific binding protein allowed us to produce the optical glycine FRET sensor (GlyFS), which can be used with single and two-photon excitation fluorescence microscopy. We took advantage of this newly developed sensor to test predictions about the uneven spatial distribution of glycine in extracellular space and to demonstrate that extracellular glycine levels are controlled by plasticity-inducing stimuli.

Optical sensors for biologically active ions and small molecules have revolutionized many research areas, including neuroscience. They can visualize changes in ion concentration (for example, Ca^{2+}) and other intra- and extracellular signals and resolve them in time and space and in locations otherwise unavailable for direct experimental investigation. In particular, optical sensors have been used with great success to uncover the dynamics of neurotransmitter signaling in the central nervous system. A number of neurotransmitter probes have been developed, mainly for the excitatory neurotransmitter glutamate (for example, FLIPE¹, EOS² and iGluSnFR³), as well as one for the inhibitory neurotransmitter γ -aminobutyric acid (GABA)⁴. In contrast, to our knowledge, no such optical sensor is available for the abundant inhibitory neurotransmitter and NMDAR co-agonist glycine.

As a neurotransmitter, glycine plays two distinct roles in the central nervous system. First, it acts as an inhibitory neurotransmitter via ionotropic glycine receptors, which are abundant in the spinal cord and brainstem but also present throughout other brain areas like the hippocampus^{5,6}. Second, glycine, together with D-serine, is a co-agonist of excitatory glutamate receptors of the NMDAR subtype. These NMDAR co-agonists need to be present at sufficient concentrations for NMDARs to open in response to presynaptic release of glutamate^{7,8}, which then can trigger synaptic long-term plasticity, a cellular correlate of learning processes^{9,10}. In addition, glycine binding to NMDARs can prime the receptor for internalization¹¹, and its abundance can regulate the relative contribution of GluN2A and GluN2B containing NMDARs to NMDAR-dependent synaptic transmission¹². Therefore, dynamic changes of NMDAR co-agonist supply can profoundly modulate NMDAR-dependent

synaptic plasticity in the hippocampus^{13–15}. Co-agonist signaling has also been suggested to be spatially segregated. Whereas glycine was implicated in extrasynaptic NMDAR function and synaptic long-term depression (LTD), the co-agonist D-serine was demonstrated to primarily act on synaptic NMDARs and to support long-term potentiation (LTP)¹⁴. In addition, glycine can regulate NMDAR-dependent plasticity through its action on pre- and post-synaptic glycine receptors^{6,16,17}. Together, these observations imply that extracellular glycine levels are controlled in time and space in the synaptic microenvironment and that they dynamically control NMDAR-dependent plasticity.

Studies of co-agonist signaling have largely relied on electrophysiological recordings of NMDAR activity, often using somatic whole-cell patch clamp recordings. This approach has been instrumental in discovering fundamental mechanisms but has *per se* no spatial resolution. Additionally, NMDAR-mediated currents and potentials are indirect readouts of extracellular co-agonist concentration that cannot by themselves identify the NMDAR co-agonist and can be affected by changes of NMDAR function and number. An alternative method is microdialysis, which can distinguish between co-agonists¹⁸ but has little spatial resolution. In contrast, fluorescent sensors permit direct spatial and temporal study of neurotransmitter release and spread in intact tissue. A glycine fluorescent sensor should permit dissection of the mechanisms that govern spatial and temporal glycine signaling in the synaptic environment.

FRET-based biosensors allow measurement of ligand concentrations by combining a pair of donor–acceptor fluorophores with a ligand-binding domain, in which binding can induce a conformational change that changes optical properties by displacing the

¹Research School of Chemistry, Australian National University, Canberra, Australia. ²Institute of Cellular Neurosciences, University of Bonn Medical School, Bonn, Germany. ³Institute of Science and Technology Austria (IST Austria), Klosterneuburg, Austria. ⁴Institute of Neuropathology, University of Bonn Medical School, Bonn, Germany. ⁵Australian Regenerative Medicine Institute (ARMI), Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia. ⁶European Molecular Biology Laboratory Australia (EMBL Australia), Monash University, Melbourne, Australia. ⁷German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany. ⁸UCL Institute of Neurology, London, UK. ⁹These authors contributed equally: William H. Zhang, Michel K. Herde. *e-mail: colin.jackson@anu.edu.au; christian.henneberger@uni-bonn.de

fluorophores. Solute-binding proteins (SBPs) undergo a venus-fly-trap-like conformational change when binding a range of ligands in both prokaryotes and eukaryotes¹⁹. The FLIPE biosensor is an excellent example of an SBP being exploited to produce a sensor, whereby a naturally occurring L-glutamate-specific SBP was sandwiched between fluorescent proteins¹. Unfortunately, the same design strategy cannot be used to create a glycine-specific sensor owing to the absence of any characterized glycine-specific SBPs in nature.

In this study we have used computational protein design to engineer the glycine FRET sensor (GlyFS). Iterative rational design of a binding core provided us with a glycine-selective binding protein, which was then linked to FRET pairs. Optimization of rigid linkers allowed us to improve the responsiveness of the optical glycine sensor GlyFS. By combining optical glycine measurements and electrophysiology in acute hippocampal slices, we then revealed developmental changes of extracellular glycine concentrations, the enrichment of glycine outside synaptic regions and the increase of extracellular glycine after plasticity-inducing synaptic stimuli.

Results

Design of the optical glycine FRET sensor (GlyFS). A solute-binding protein (SBP) from *Agrobacterium tumefaciens*, Atu2422, which displays promiscuous binding activity for glycine, L-serine and GABA²⁰ (Fig. 1a,b), was used as the template for computational design, involving cycles of computational design with FoldX²¹ and ligand docking with Autodock²² (Methods; Supplementary Fig. 1). The binding site of Atu2422 was redesigned, focusing on introducing steric obstruction to prevent L-serine and GABA binding, thereby making it specific for glycine (Fig. 1a). Initial analysis suggested that reducing the size of Phe77 could allow Ala100 to be mutated to an aromatic residue that could block the binding of amino acids larger than glycine. An F77A/A100Y (AY) mutant was produced, which was confirmed by isothermal titration calorimetry (ITC) to no longer bind GABA with significant affinity (Fig. 1b; Supplementary Table 1; Supplementary Fig. 2). However, L-serine still bound with an affinity of $20.1 \pm 6.3 \mu\text{M}$, possibly by adopting an inverted conformation (Fig. 1a). Further simulations suggested that mutation of Leu202, located across the binding pocket from Ala100, to a larger amino acid could impede L-serine binding. ITC of an F77A/L202W (AW) variant confirmed that L-serine and GABA binding was significantly reduced, whereas glycine could still bind with a K_D of $2.3 \mu\text{M}$. However, glutamate bound to AW with significant affinity (Fig. 1b; Supplementary Table 1). It is likely that rotation of Trp202 allows glutamate to bind (Fig. 1a). To prevent this, the A100Y mutation was restored (mutant AYW), which yielded a specific glycine-binding protein with an affinity of $20.3 \pm 3.7 \mu\text{M}$. Competition ITC experiments (glycine binding in the presence of $500 \mu\text{M}$ L-serine, GABA or glutamate) confirmed that the protein was specific for glycine (Supplementary Table 1). Thus, rational design produced a binding core selective for glycine from a promiscuous template.

To construct the fluorescent sensor, AYW was cloned between enhanced cyan fluorescent protein (ECFP) and Venus-fluorescent protein (Venus), a classical and still-popular FRET pair²³, as described previously²⁴. A hexahistidine-tagged N-terminal biotin domain was also included for immobilization (Supplementary Fig. 3). The full-length fusion protein could be obtained in high purity through a combination of Ni^{2+} affinity chromatography and size-exclusion chromatography (Supplementary Fig. 4). The initial sensor design had flexible regions between the fluorescent proteins and the AYW domain and produced a maximum ratio-metric response to saturating concentrations of glycine (dynamic range) of $\sim 4\%$ (ECFP/Venus; Fig. 1c; Supplementary Fig. 3). The relative increase of donor over acceptor fluorescence intensity indicates that glycine binding to the sensor core reduces FRET between the fluorophores (Supplementary Fig. 5). We performed a series of

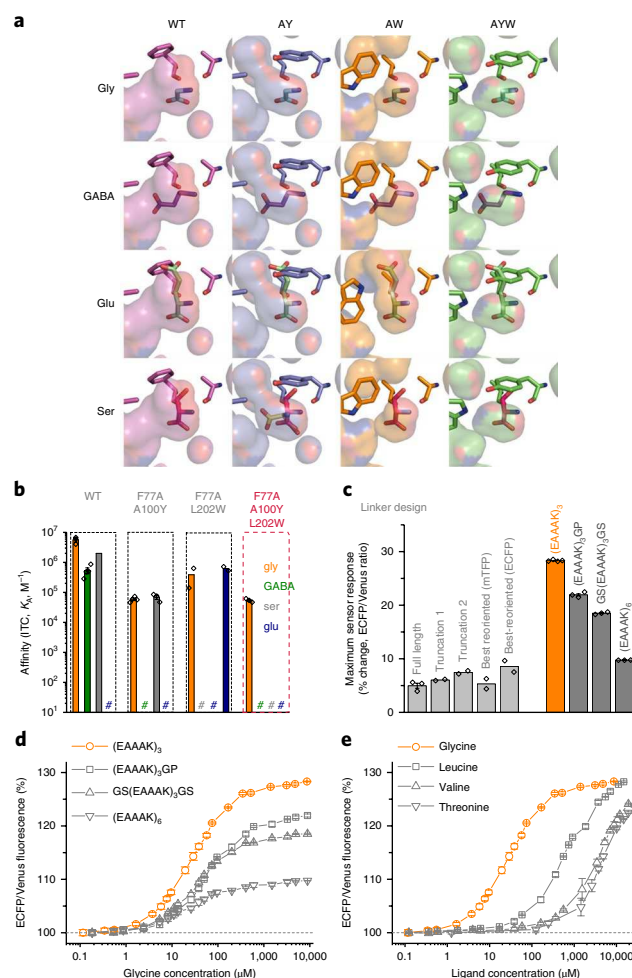


Fig. 1 | Design of the optical glycine FRET sensor (GlyFS). **a**, Design of the binding site. Modeled binding sites leading up to the glycine-specific mutant Ala/Tyr/Trp (AYW; WT, wild-type). Unwanted ligands are increasingly unable to bind as the binding pocket becomes more restricted. Shown in the glutamate-bound AW mutant is the flipping of residue 202 to accommodate glutamate. An alternative ligand conformation for the serine-bound AY mutant is also shown. **b**, Affinity of mutants for different ligands obtained from isothermal titration calorimetry (ITC; color coded: glycine, orange; GABA, green; L-serine, gray; glutamate, blue; $n = 3, 4, *, 1, 4, 2, 3, 1, 2, 2, 2, 3, 1, 3$ and 3 independent experiments from left to right; *, the estimated K_D for L-serine extrapolated from ref. ²⁰ is $\leq 500 \text{ nM}$; # represents no detection of binding in ITC in the presence of $500 \mu\text{M}$ ligand). F77A, A100Y and L202W mutations (red box) confer glycine selectivity to the sensor-binding site and protein (Supplementary Fig. 2). **c**, A set of fluorescent indicator proteins based on the FRET sensor pair ECFP and Venus using various linker regions were tested (ECFP-binding site-linker-Venus). The maximum change of the ECFP/Venus fluorescence intensity ratio in response to saturating glycine concentrations was quantified for different linkers. Linker region optimization increased the glycine sensor dynamic range ($n = 3, 2, 2, 2, 2, 4, 3, 3$ and 3 independent experiments, from left to right). The rigid linker (EAAAK)₃ was selected. See also Supplementary Figs. 3, 7 and 8 and text for details of linker-region design variants. **d**, Dose-response curves of glycine sensors with different linker regions to glycine (subset of data from **c**, fluorescence intensity ratios, ECFP/Venus). Sensors with the rigid linker (EAAAK)₃ display the highest dynamic range ($28.3 \pm 0.08\%$; $n = 4$ independent experiments, orange; all others in gray and $n = 3$ independent experiments). **e**, Dose-response curves of relevant GlyFS ligands ($n = 4, 3, 3$ and 3 independent experiments, ECFP/Venus fluorescence intensity ratio). All data are presented as mean or mean \pm s.e.m. as appropriate.

rational design steps to improve the sensor. Truncating the flexible regions²⁵ (Supplementary Fig. 3), repositioning the fluorophores by inserting them into different loops of circularly permuted AYW³ (Supplementary Figs. 3 and 6), and varying the relative angle of Venus and ECFP through the use of differently circularly permuted variants of the donor fluorophore²⁶ (Supplementary Table 2) did not improve the dynamic range beyond ~10% (Fig. 1c). We next optimized the linkers between the sensor region and the fluorescent proteins. Truncation of the linker to ECFP and replacement of the linker to Venus with a rigid triple repeat of an α -helical Glu/Ala/Ala/Ala/Lys linker²⁷ increased the dynamic range to $28.3 \pm 0.08\%$ ($n=4$, after size-exclusion chromatography), producing the sensor GlyFS (Fig. 1d; Supplementary Fig. 4). We note that a further increase in linker length resulted in a decrease in the FRET efficiency (Supplementary Figs. 7 and 8), which explains the concomitant decrease in dynamic range. This is consistent with previous work suggesting that changes to interfluorophore distance rather than orientation result in larger changes in FRET efficiency²⁸. Thus, optimizing interfluorophore distance via rigid linkers appears to be an efficient method to improve FRET sensors.

Although we had confirmed that AYW specifically bound glycine without detectable binding of glutamate, GABA or L-serine, we also tested their binding to the full sensor construct using fluorescence measurements, as well as the remaining 17 proteinogenic amino acids, GABA and D-serine (500 μ M throughout). In addition to glycine, the amino acids leucine, valine and threonine elicited small changes in the fluorescence ratio. No other potential ligands had an effect on the GlyFS fluorescence ratio (<1% ratio change, $n=2$ each). Dose-response curves of GlyFS for glycine, leucine, valine and threonine were recorded, revealing insignificant binding of leucine, valine and threonine in the concentration range of interest (0–50 μ M) and K_D values >10-fold higher than that for glycine (Fig. 1e) and well above the concentrations encountered in the extracellular space^{18,29}. The promiscuous binding of leucine and valine is not unexpected given that Atu2422 shares a common ancestor with leucine/valine-binding proteins²⁰ and because related binding proteins often display weak promiscuous binding of amino acids that were bound by their ancestral states³⁰.

Performance of GlyFS in hippocampal tissue. Next, GlyFS was immobilized in acute hippocampal slices through a biotinylation-based technique that anchors proteins exclusively in extracellular space (Methods and refs. ^{24,31}; Supplementary Figs. 9 and 10 for endogenous expression). To achieve this, a biotin tag was introduced into GlyFS (Supplementary Figs. 11 and 12). The performance of this modified sensor was then tested with two-photon excitation (2PE) fluorescence microscopy (800 nm) in a saline solution containing a range of glycine concentrations (Fig. 2a). This calibration yielded GlyFS affinities (K_D) to glycine and dynamic ranges similar to those observed before (Fig. 1). Tests with several independently produced batches of GlyFS showed that both K_D and dynamic ranges were stable between batches and on average around ~20 μ M and ~20%, respectively (Fig. 2b,c; Supplementary Table 3). Furthermore, GlyFS was not responsive to the NMDAR co-agonist D-serine at concentrations of up to 5 mM (Fig. 2d). Thus, this final GlyFS variant can be used with 2PE imaging, which allows its use deep within organized tissue like acute brain slices.

We then monitored extracellular glycine levels *in situ* by anchoring GlyFS in the extracellular space of the CA1 region of biotinylated acute hippocampal slices via a biotin-streptavidin linker (Fig. 3a; Methods; refs. ^{24,31}). This method of GlyFS delivery into extracellular space provided a stable readout of glycine levels, because the ECFP/Venus fluorescence ratio remains stable overall throughout the experiment (Fig. 3b) despite some GlyFS unbinding and washout over the time course of 45 min. We noticed a small but significant increase of the GlyFS ratio over 45 min (Fig. 3b), which

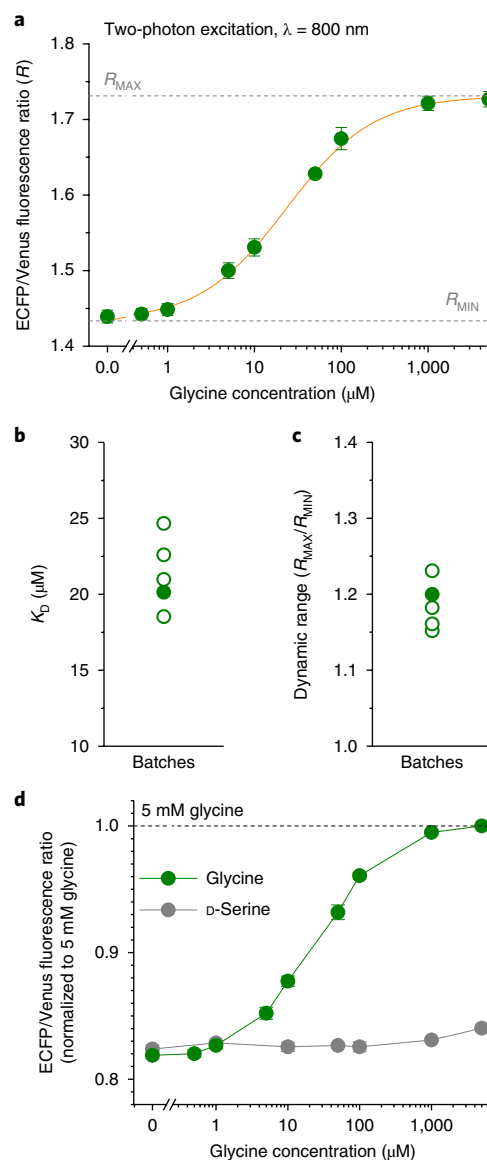


Fig. 2 | Characterization of GlyFS using two-photon excitation (2PE) fluorescence microscopy ($\lambda_{2PE} = 800$ nm). **a**, Calibration of a single batch of sensor in a cuvette ($n=6$ independent experiments). The ratio of ECFP and Venus fluorescence intensities (R) was obtained for a range of glycine concentrations. R_{MAX} and R_{MIN} denote the maximum and minimum ratios for a saturating glycine concentration of 5 mM and nominally zero glycine, respectively. **b**, The affinity of GlyFS to glycine was determined by fitting a Hill equation to calibration data for each sensor batch. The dissociation constant for glycine (K_D) was, on average, $21.4 \pm 1.1 \mu$ M ($n=5$ independently produced batches and experiments; filled circle corresponds to the calibration shown in **a**). **c**, The dynamic range of the sensor was estimated in five different sensor batches (average $18.5 \pm 1.4\%$; $n=5$ independently produced batches and experiments; filled circle corresponds to calibration shown in **a**). **d**, Comparison of binding of the two NMDAR co-agonists glycine and D-serine. GlyFS was exposed to increasing concentrations of glycine (green, $n=4$ independent experiments) and D-serine (gray, $n=3$ independent experiments) before saturation of the sensor by bath-applied glycine (5 mM) to test intactness of GlyFS. GlyFS displayed no significant changes in fluorescence in response to application of D-serine (maximum change of $+2.0 \pm 0.7\%$ at 5 mM D-serine; two-sided paired Student's t -test, $t(2) = -2.95$, $P = 0.10$; below $+1\%$ otherwise, $P \geq 0.15$). All data are presented as mean \pm s.e.m. where applicable.

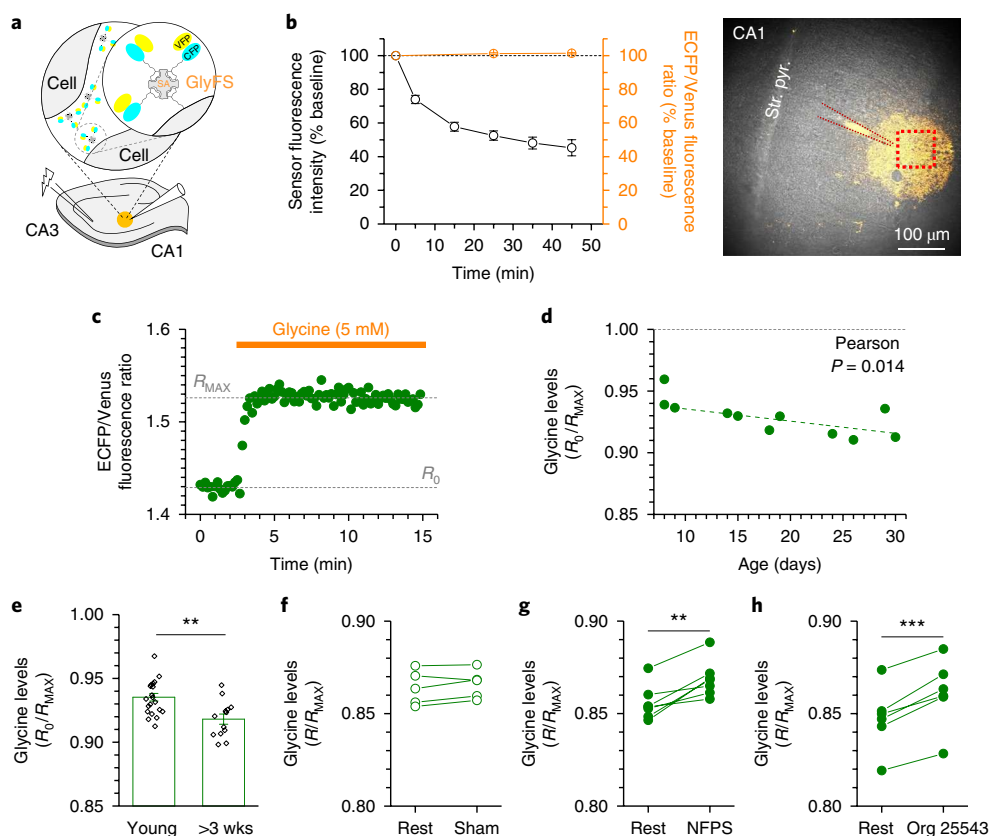


Fig. 3 | Measuring extracellular glycine levels using GlyFS in hippocampal tissue. **a**, Upper panel, GlyFS contains a biotin tag that is used to anchor the sensor in extracellular space via streptavidin (SA) to biotinylated membranes as described previously^{24,31} (see Methods). Lower panel, GlyFS was pressure-loaded into hippocampal brain slices (CA1, greyscale differential interference contrast (DIC) image, pyramidal cell layer str. pyr.) via a pipette (red dotted line, only partially in focus). This typically labels a circular region around the loading pipette with a diameter of 200 to 300 μm (Venus fluorescence overlay in yellow, single representative example for all experiments shown in **d–h**). Red box represents a typical region of interest (ROI). **b**, Recordings were started 15 min after GlyFS labeling. In a first set of experiments, the amount of intact sensor remaining in the extracellular space was quantified by normalizing GlyFS Venus fluorescence intensity to its initial value ($t = 0$ min, GlyFS loading at $t = -15$ min, black dots and axis). The decrease of GlyFS Venus fluorescence over time indicates that about half of GlyFS is washed out over 50 min ($n = 9$ independent experiments). In a second set of experiments, the stability of the sensor at a recording temperature of 34 $^{\circ}\text{C}$ was assessed by imaging at the beginning and after 25 and 45 min. The overall stable ECFP/Venus ratio indicates that GlyFS' ability to report extracellular glycine levels is not compromised by partial GlyFS unbinding from the tissue (orange dots and axis). We observed a small but significant increase of the GlyFS ratio by $1.4 \pm 0.40\%$ over 45 min compared to the initial GlyFS ratio (two-sided paired Student's t -tests vs. 0 min, after 25 min $t(4) = -2.65$ and $P = 0.057$, after 45 min $t(4) = -3.48$ and $P = 0.025$; $n = 5$ independent experiments). Also see corresponding results section. **c**, To estimate the resting concentration of glycine, GlyFS was imaged before, during and after saturation of the sensor with 5 mM glycine and the ECFP/Venus fluorescence intensity ratios R_0 and R_{MAX} were determined. R_0 approaches R_{MAX} as the resting glycine concentration increases. Therefore, R_0/R_{MAX} was used as a measure of the extracellular glycine resting levels. Shown is a single example (acute brain slice from 3 week old rat). This type of experiment was used, whenever R_{MAX} was established (example for all experiments shown in **d–h**). **d**, The developmental profile of the extracellular glycine concentration in the stratum radiatum was determined. Experiments were performed using a single batch of GlyFS (#3) to avoid variability introduced by changes of the dynamic range between sensor batches. A significant negative correlation between R_0/R_{MAX} and age was observed (Pearson $R = -0.71$, $P = 0.014$, $n = 11$ animals, 2–4 brain slices per animal). No correlation between R_{MAX} and age (Pearson $R = 0.077$, $P = 0.82$). **e**, To directly compare glycine levels in young and old animals, data were grouped according to age (younger and older than three weeks). Glycine levels were significantly lower in hippocampal slices obtained from older animals (two-sided Welch's t -test, $t(27) = 3.23$, $P = 0.0035$; $n = 20$ independent experiments from young and 13 from older animals; small circles represent individual data points, only sensor batch #3 as in **d**). No statistically significant difference between R_{MAX} (two-sided Welch's t -test, $t(22) = -0.16$, $P = 0.87$). **f–h**, The low extracellular resting levels of glycine in CA1 stratum radiatum are maintained by glycine transporters. **f**, The stability of glycine levels in the absence of any pharmacological manipulation was tested first. Glycine levels were determined at rest and 5 min later (sham). No significant change was detected (paired two-sided Student's t -test, $t(4) = -1.50$, $P = 0.21$; $n = 5$ independent experiments). **g**, The application of the specific GlyT1 inhibitor NFPS (5 μM , dissolved in DMSO, final DMSO concentration 0.05%) increased the resting glycine level significantly (paired two-sided Student's t -test, $t(6) = -4.69$, $P = 0.0034$; $n = 7$ independent experiments). **h**, Similarly, inhibition of GlyT2 by acute application of Org 25543 (1 μM , dissolved in water) increased extracellular glycine levels significantly (paired two-sided Student's t -test, $t(5) = -7.83$, $P = 0.00055$; $n = 6$ independent experiments). Together, these results show that the baseline activity of GlyT1 and 2 decreases extracellular glycine levels. All data are presented as mean \pm s.e.m. where applicable. ** $P < 0.01$; *** $P < 0.001$.

may be caused by low extracellular glycine accumulation during long recordings. This emphasizes the need for adequate controls for prolonged physiological experiments (see below).

The GlyFS fluorescence intensity ratio at rest (R_0) depends on the resting glycine concentration in extracellular space, as well as on, among other parameters, how ECFP and Venus fluorescence

is detected (for example, photomultiplier voltage), which can vary between experiments. Adopting a formalism used for estimating resting Ca^{2+} concentrations^{32,33} could provide glycine concentration estimates but requires determining R_{MIN} , the fluorescence ratio in the absence of glycine, and R_{MAX} , the fluorescence ratio of GlyFS saturated with glycine, in each experiment. Because measuring R_{MIN} is not straightforward (see Discussion), we used the ratio of R_0 and R_{MAX} determined by bath-application of 5 mM glycine (example in Fig. 3c) to obtain an estimate of resting glycine levels. This measure is independent of, for instance, imaging settings and is thus more suitable for comparing resting glycine levels and the effects of their pharmacological manipulation between experiments. GlyFS-reported resting glycine levels were first compared between the CA1 subregions stratum oriens (SO), stratum pyramidale (SP), stratum radiatum (SR) and stratum lacunosum moleculare (SLM) by measuring R_0/R_{MAX} as illustrated in Fig. 3c. No significant differences were detected (one-way ANOVA, $F(3,93) = 1.19$; $P = 0.32$; $n = 9, 9, 50$ and 29 individual brain slices for SO, SP, SR and SLM, respectively). Next, we investigated the developmental profile of extracellular glycine levels, because the effect of glycine degradation on NMDAR signaling decreases with age^{12,15}. Indeed, GlyFS-reported resting glycine levels (R_0/R_{MAX}) observed in the CA1 stratum radiatum were negatively correlated with the age of the animal (Fig. 3d) and were significantly lower in slices obtained from older animals compared to younger animals (Fig. 3e). Importantly, the intensity ratio of GlyFS saturated with glycine (R_{MAX}) did not display an age dependence (Fig. 3d,e, legend), indicating that GlyFS fluorescence properties do not change with the developmental stage of the tested tissue. We then monitored the GlyFS ratio during blockade of the glycine transporters 1 and 2 (GlyT1 and GlyT2, respectively) because these maintain extracellular glycine levels³⁴. Pharmacological blockade of GlyT1 and GlyT2 using NFPS and Org 25543, respectively, induced a significant increase of GlyFS-reported glycine levels, which was not observed in control experiments (Fig. 3f–h). GlyT1 and GlyT2 therefore actively lower extracellular glycine concentrations. These experiments establish GlyFS as a useful tool to investigate how extracellular glycine levels are controlled.

Mechanisms governing glycine levels. The suitability of GlyFS for studying spatial variations of glycine levels was explored next. We compared resting glycine levels around spines and dendritic shafts, because glycine was shown to be the primary co-agonist of extrasynaptic, but not synaptic, NMDARs¹⁴. We therefore labeled the neuropil with GlyFS and then filled a single CA1 pyramidal cell with the fluorescent dye Alexa Fluor 594 via the whole-cell patch pipette to visualize its dendritic tree and synaptic spines (Fig. 4a). We then zoomed in on a pseudo-randomly chosen dendritic segment and measured GlyFS fluorescence before and after bath application of 5 mM glycine to saturate GlyFS. This enabled us to determine the resting levels of glycine (R_0/R_{MAX}) in regions of interest (ROIs) around synaptic spines and at dendritic shafts (Fig. 4b). GlyFS-reported resting levels displayed considerable variability among both spines and dendritic shafts in individual experiments (Fig. 4c). Averages from a total of eight experiments revealed overall lower glycine levels at spines compared to dendritic shafts (Fig. 4d). Again, altered fluorescence properties of GlyFS did not underlie this finding, because R_{MAX} was not different between the two ROI types (Fig. 4d, legend). These differences between ROIs at dendritic shafts and spines are likely to underestimate the difference between synaptic and extrasynaptic glycine concentrations (see Discussion). Our results are in line with predictions about the extracellular landscape of glycine concentrations¹⁴ and provide direct evidence for a subregion-specific regulation of extracellular glycine levels in intact tissue.

Finally, we further characterized the activity-dependent mechanisms that control extracellular glycine levels. Because glycine is involved in NMDAR-dependent synaptic plasticity^{13–15}, we focused

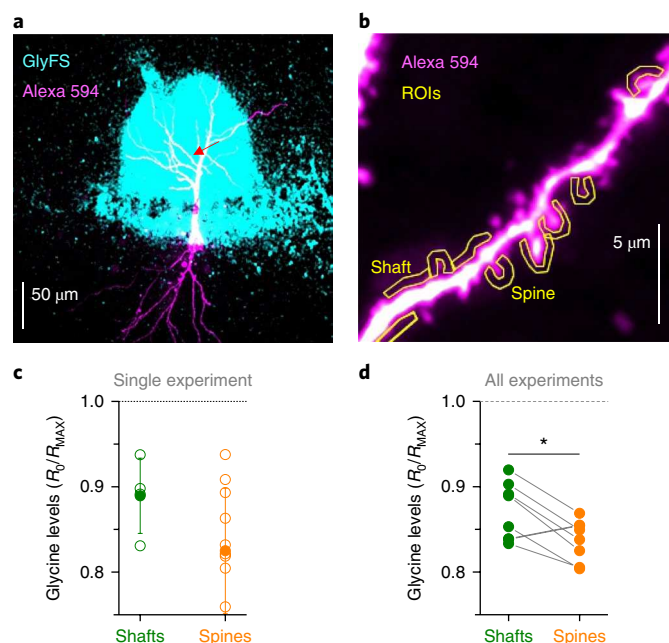


Fig. 4 | Extracellular glycine levels reported by GlyFS are lower at dendritic spines. **a**, A CA1 pyramidal cell was filled with a fluorescent marker Alexa Fluor 594 (magenta) via a whole-cell patch clamp pipette. Its dendrites were traced into a GlyFS-labeled area of CA1 stratum radiatum (cyan) after resealing and removal of the patch-clamp pipette. Red arrow indicates dendritic segment selected for analysis (shown in **b**). Representative example for experiments presented in this figure. **b,c**, Magnified dendrites with spines (Alexa Fluor 594, magenta). Same experiment as that in **a**. Regions of interest (ROIs, yellow) were defined around well-isolated spines and at dendritic shafts without visible spines. Resting glycine levels were estimated as described in Fig. 3c using bath application of saturating glycine concentrations (5 mM). R_0/R_{MAX} was calculated for each ROI. Results from this representative example (full data set in **d**) are displayed in **c** (single experiment; $n = 4$ dendritic shaft ROIs and 11 spine ROIs, empty circles). Filled circles represent mean \pm s.d. of this individual experiment (two-sided Welch's t -test of this example, $t(9.3) = 2.1$, $P = 0.068$). **d**, Results from all eight independent experiments ($n = 8$) with 3–14 spine and dendritic shaft ROIs per experiment. Despite the considerable variability in individual experiments, higher resting glycine levels (R_0/R_{MAX}) were detected on average at dendritic shafts compared to ROIs around spines (paired two-sided Student's t -test, $t(7) = 2.94$, $P = 0.022$). Similar results were obtained when R_0 values instead of R_0/R_{MAX} were compared (paired two-sided Student's t -test, $t(7) = 3.37$, $P = 0.012$). The latter was not the case for R_{MAX} (GlyFS saturated with exogenous glycine, paired two-sided Student's t -test, $t(7) = -0.26$, $P = 0.80$), indicating that the above result is not caused by altered GlyFS fluorescence properties in the different types of ROIs. $*P < 0.05$.

on plasticity-inducing stimuli. We stimulated Schaffer collateral CA3–CA1 synapses at high (HFS) and low frequency (LFS), suitable for induction of long-term potentiation (LTP) and long-term depression (LTD), respectively, while monitoring GlyFS fluorescence (Fig. 5a). We have previously demonstrated that biotinylation of acute slices does not affect synaptic transmission at this synaptic pathway²⁴. Furthermore, immobilizing GlyFS in the acute slices did not cause significant changes of CA3–CA1 synaptic transmission (Supplementary Fig. 13). HFS resulted in a significant increase of GlyFS-reported extracellular glycine levels (Fig. 5b,c), which was not observed when the HFS was omitted (Fig. 5d). The HFS-induced glycine increase was also not observed in the presence of the glycine transporter inhibitors NFPS and Org 25543 (Fig. 5e). In addition,

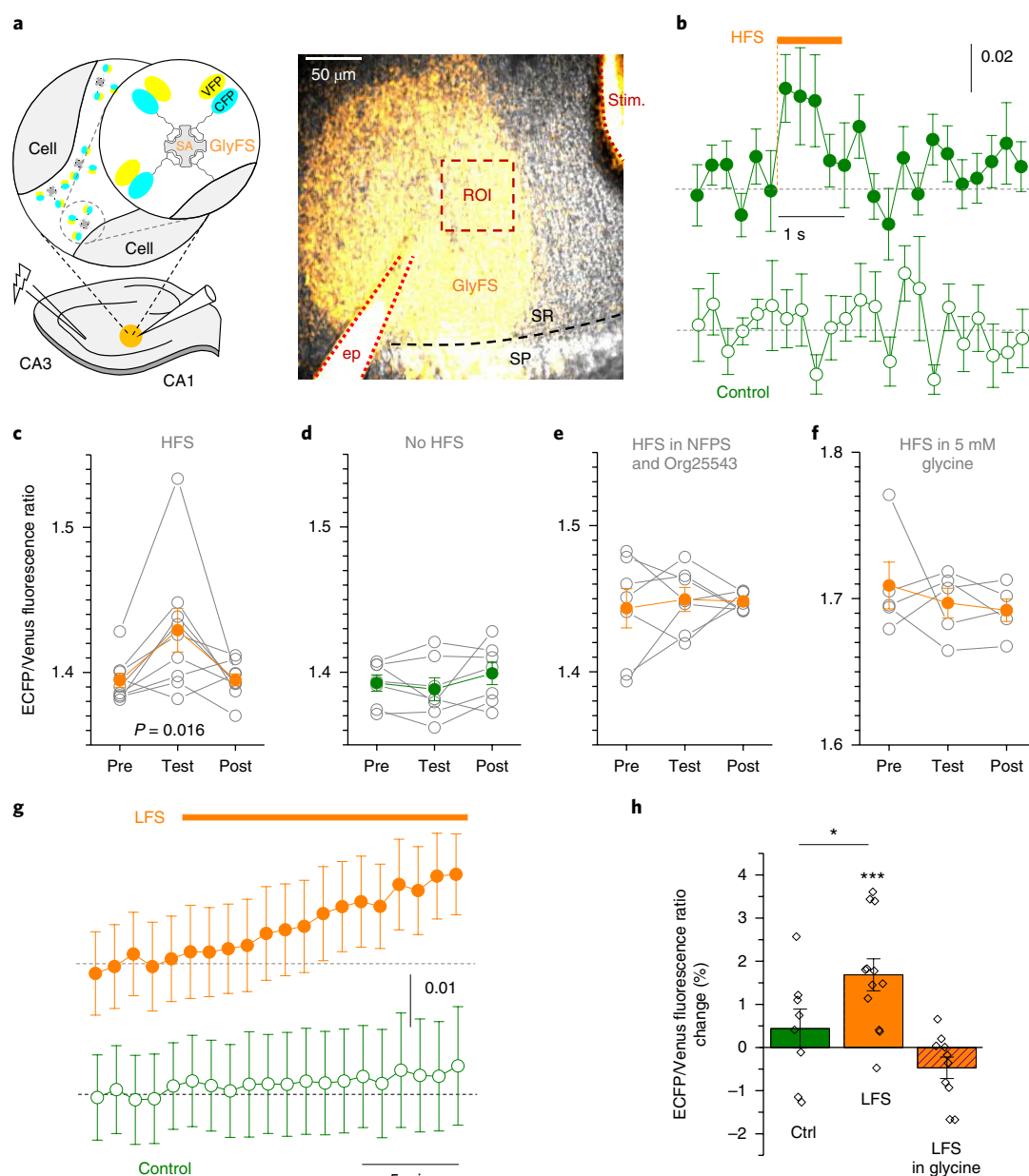


Fig. 5 | GlyFS identifies activity patterns that control extracellular glycine levels. **a**, Left panel, schematic illustration of GlyFS anchoring in extracellular space (SA, streptavidin). Right panel, a single representative example of GlyFS injection into hippocampal slices and monitoring of extracellular glycine levels in combination with electrophysiology (full data set in **b–h**). DIC image of CA1 stratum radiatum (SR) with Venus fluorescence overlay (ep, extracellular recording pipette; SP, stratum pyramidale; stim., stimulation pipette; ROI, region of interest). **b**, High-frequency stimulation (HFS, orange bar; 1 s, 100 Hz) induced a transient increase of glycine levels reported by GlyFS (upper trace, green filled circles; $n=9$ independent experiments). After acquisition of a baseline period (1 s), a HFS (1 s, 100 Hz) was delivered to the CA3–CA1 Schaffer collaterals. No such increase was observed in control experiments without HFS (lower trace, empty circles). **c**, Summary of experiments with HFS. The pre-test baseline, test response and a post-test baseline were determined. A significant increase of extracellular glycine levels was detected (paired two-sided Student's t -test, $t(8) = -3.04$, $P = 0.016$; $n = 9$ independent experiments). Individual experiments shown in gray; mean \pm s.e.m. in orange. **d**, No increase of GlyFS-reported glycine levels was detected when HFS was omitted (paired two-sided Student's t -test, $t(6) = 0.87$, $P = 0.42$; $n = 7$ independent experiments). Individual experiments shown in gray; mean \pm s.e.m. in green. **e**, No increase of GlyFS-reported glycine levels was observed in the presence of GlyT1/2 inhibitors NFPS and Org 25543 (paired two-sided Student's t -test, $t(6) = -0.51$, $P = 0.63$; $n = 7$ independent experiments). Individual experiments in gray; mean \pm s.e.m. in orange. **f**, Saturating the glycine binding site of GlyFS by bath application of exogenous glycine occluded GlyFS fluorescence changes (5 mM, paired two-sided Student's t -test, $t(4) = 0.56$, $P = 0.61$; $n = 5$ independent experiments). Individual experiments in gray; mean \pm s.e.m. in orange. **g**, The time course of the effect of low-frequency stimulation (LFS) of CA3–CA1 synapses (15 min, 2 Hz; $n = 12$ independent experiments) on GlyFS-reported glycine levels is shown in the upper panel (ECFP/Venus fluorescence intensity ratio). The control time course without LFS is shown in the bottom panel ($n = 8$ independent experiments). **h**, A significant increase of GlyFS-reported glycine levels was induced by LFS (orange, one-population two-sided Student's t -test, $t(11) = 4.54$, $P = 0.00084$; $n = 12$ independent experiments) but not in control recordings (ctrl, green, one-population two-sided Student's t -test, $t(7) = 0.97$, $P = 0.36$; $n = 8$ independent experiments) or in the presence of exogenous glycine (10 μ M, orange, hatched, one-population two-sided Student's t -test, $t(9) = -1.9$, $P = 0.090$; $n = 10$ independent experiments; control vs. LFS two-populations two-sided Student's t -test, $t(18) = -2.13$, $P = 0.047$). Small circles represent individual data points. Data are presented as mean \pm s.e.m. where applicable. * $P < 0.05$; *** $P < 0.001$.

this increase was also not detected when the experiment was performed in the presence of 5 mM extracellular glycine to saturate the GlyFS' glycine-binding site (Fig. 5f), showing that HFS increased extracellular glycine levels and did not affect GlyFS fluorescence by another mechanism. The latter observation is further supported by two additional experiments. First, a previously designed arginine sensor²⁴ using the same FRET pair did not respond to synaptic high-frequency stimulation (ECFP/Venus ratio change of $-0.19 \pm 0.27\%$, paired two-sided Student's *t*-test, $t(4) = 1.27$, $P = 0.27$, $n = 5$ independent experiments). Second, changes of extracellular $[K^+]$, $[Ca^{2+}]$ and pH observed previously during HFS did not affect the GlyFS fluorescence ratio (Supplementary Fig. 14). Together, these experiments reveal that, surprisingly, HFS results in an increase of extracellular glycine levels. Qualitatively similar results were obtained using LFS (Fig. 5g,h). Overall, monitoring of extracellular glycine levels using the newly developed GlyFS enabled us to uncover the dynamic modulation of its concentration and to directly reveal its spatial inhomogeneities in organized tissue.

Discussion

In this work, we have constructed the optical sensor GlyFS and demonstrated its effectiveness for visualizing the dynamic signaling and spatial distribution of the inhibitory neurotransmitter and NMDAR co-agonist glycine. This was accomplished through the use of computational aides such as FoldX²¹ and Autodock²², which facilitated the engineering of a novel glycine-specific binding domain. This reflects a growing capacity for computational design to yield binding proteins that are improved and more specific³⁵, and it was the only realistic approach to solve this problem because the subtle changes in binding affinity with a wide range of ligands cannot be easily screened via high-throughput approaches. Conversion of this binding core into a usable optical sensor was enabled through the use of a rigid (Gly/Ala/Ala/Ala/Lys)₃ linker, which converts angular changes into an increased interfluorophore displacement more efficiently than the more commonly used flexible Gly/Gly/Val/Ser/Lys/Gly/Glu linker by behaving as a lever, thereby increasing the dynamic range of the sensor. This is distinct from other approaches that have sought to increase dynamic range on the basis of fluorophore orientation, which did not yield positive results for this sensor. By focusing on maximizing changes in interfluorophore distance rather than orientation, sensors could be improved more reliably and quickly, which is facilitated through the ability of rigid linkers to convert angular changes to distance changes. In this study, we have produced a ratiometric sensor, because changes in the ratio of the ECFP and Venus fluorescence intensity are not affected by differences in the amount of sensor present in each sample/region of interest, allowing more accurate measurements. The production of the glycine-specific binding core should facilitate the production of other types of glycine sensors, as exemplified by the family of glutamate sensors produced from the bacterial glutamate-specific solute-binding protein (FLIPE¹, EOS² and iGluSnFR³).

For functional tests, a biotin tag was introduced into GlyFS to fix the sensor in the extracellular space by injecting a GlyFS-streptavidin mixture into the surface-biotinylated hippocampal tissue as described previously^{24,31}. As a consequence, all sensor is located in extracellular space and is fully exposed to only extracellular glycine. The method is applicable in preparations that can be biotinylated and injected with GlyFS such as cultures and acute brains slices and is potentially compatible with short-term *in vivo* imaging³⁶. The dynamic range of GlyFS and its K_D for glycine were readily obtained in cell-free solutions (Fig. 2). Under the assumption that anchoring GlyFS in the hippocampal tissue does not affect either property and by adopting the formalism for steady state Ca^{2+} concentration³³, the extracellular glycine concentration in the hippocampus can be estimated. It is given by $[Gly] = K_D \times (R - R_{MIN}) / (R_{MAX} - R)$, where R , R_{MIN} and R_{MAX} denote the ECFP/Venus

fluorescence intensity ratios in a ROI at rest, in the absence of glycine and in the presence of saturating glycine concentrations, respectively. Directly determining R_{MIN} *in situ* requires the complete removal of extracellular glycine from the slice preparation. However, a fast, easily diffusible, high-affinity glycine buffer suitable for bath application is not available, to our knowledge. Additionally, enzymatic degradation of glycine using glycine oxidase^{12,14,37} may not outcompete glycine efflux from cells in the slice because NMDAR-mediated potentials and currents are not completely blocked in experiments using enzymes to simultaneously degrade the NMDAR co-agonists D-serine and glycine¹⁴. We therefore recast the above equation in terms of the dynamic range $f = R_{MAX} / R_{MIN}$, yielding $[Gly] = K_D \times (R / R_{MAX} - 1 / f) / (1 - R / R_{MAX})$. R / R_{MAX} was directly obtained from experiments with an overall average of 0.87 ($n = 96$ individual measurements using GlyFS batches #2 and #3; data from Fig. 3, > 3 weeks, K_D from Fig. 2, Supplementary Fig. 15) giving an estimate for $[Gly]$ of 4.7 μM . This is close to values previously measured using microdialysis of about 8–12 μM in the frontal cortex and hippocampus *in vivo*^{18,38,39}, implying that the mechanisms governing extracellular glycine levels are largely preserved in acute slices. Our estimate primarily reflects the extrasynaptic glycine concentration, because the resting GlyFS fluorescence was averaged in these experiments over relatively large ROIs (Fig. 3a) in which > 95% of GlyFS is expected to be extrasynaptic (Methods; ref. ³¹).

GlyFS also enabled us to directly test predictions about compartmentalization of glycine levels. It was previously demonstrated that degradation of glycine affects primarily extrasynaptic and not synaptic NMDARs¹⁴, indicating that synaptic glycine levels should be lower than extrasynaptic levels. We confirmed this by directly measuring glycine levels in the vicinity of synaptic spines and at dendritic shafts. The observed difference is likely to underestimate spatial glycine gradients between synaptic, perisynaptic and extrasynaptic space for two reasons. First, the nanometer-scale structure of extracellular space in and around synapses cannot be fully resolved by diffraction-limited two-photon excitation⁴⁰. Therefore, a considerable amount of extrasynaptic GlyFS will contribute to any selected 'synaptic region of interest'. Second, a GlyFS signal quantified in a 'dendritic region of interest' could be contaminated by GlyFS fluorescence from nearby invisible synapses and their synaptic clefts and perisynaptic regions given the synapse density of $2 \mu m^{-3}$ in the densely packed neuropil of the CA1 stratum radiatum⁴¹. Full resolution of the glycine landscape could be achieved by increasing the imaging resolution. Because multiple fluorescent proteins can be excited and visualized using stimulated emission depletion (STED) microscopy⁴², 3D-STED microscopy⁴⁰ of GlyFS could allow super-resolution imaging of submicrometer glycine gradients in the synaptic environment.

GlyFS also enabled us to investigate the mechanisms controlling glycine levels. We found that pharmacological inhibition of either glycine transporter (GlyT1 or GlyT2) increased GlyFS-reported glycine levels. Indeed, GlyT1 is expressed in the hippocampus, where it is localized mainly in glial cells and to lesser extent in neurons^{6,43}. Our experiments demonstrate that extracellular glycine levels are actively lowered by GlyT1, which is in line with previous reports in which GlyT1 inhibition increased NMDAR-mediated postsynaptic currents and LTP^{44,45}. In contrast, GlyT2 is less strongly expressed in the hippocampus, although it has been detected using immunohistochemistry^{6,46}. We also found that extracellular glycine was increased by stimulation of CA3–CA1 Schaffer collateral synapses at low and high frequencies. The latter was not observed in the presence of GlyT1 and GlyT2 inhibitors, indicating that the glycine increase is most likely linked to changes of transporter function during stimulation. Indeed, GlyTs are believed to be the primary regulator of extracellular glycine in the hippocampus⁶. It is therefore likely that the glycine increase during neuronal stimulation is due to a transient reduction of glycine uptake or glycine transporter

reversal. This probably involves mainly GlyT1 for two reasons. First, GlyT1 is more abundant than GlyT2 in the hippocampus^{6,43}. Second, GlyT1 co-transporters glycine along with 1 Cl⁻ and 2 Na⁺ as opposed to the 1 Cl⁻ and 3 Na⁺ for GlyT2, allowing GlyT1 to reverse/export glycine more easily³⁴. The astroglial sodium rise resulting from astroglial glutamate uptake after neuronal activity⁴⁷ is a likely link between CA3–CA1 synaptic activity and changes of astroglial GlyT1 function.

What is the functional significance of the observed extracellular glycine transients? Our estimate of the mainly extrasynaptic resting glycine concentration of ~5 μM suggests that the co-agonist binding site of extrasynaptic NMDARs was close to saturation in these experiments (EC₅₀ of glycine at NMDARs: ~0.1 to 2.0 μM; refs^{7,48}). Indeed, tonic NMDAR-dependent currents in CA1 pyramidal cells, which are likely to be mediated by extrasynaptic NMDARs, did not change when extracellular NMDAR co-agonist levels were increased⁴⁹. The increase of extracellular glycine may therefore have little effect on extrasynaptic NMDAR function. Another abundant target for glycine in the hippocampus is glycine receptors, which can modify neuronal excitability, presynaptic release and synaptic plasticity^{6,16,17,50}. Therefore, the observed changes in glycine levels could alter network function via activation of glycine receptors.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at <https://doi.org/10.1038/s41589-018-0108-2>.

Received: 30 December 2017; Accepted: 21 June 2018;

Published online: 30 July 2018

References

- Okumoto, S. et al. Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. *Proc. Natl. Acad. Sci. USA* **102**, 8740–8745 (2005).
- Namiki, S., Sakamoto, H., Iinuma, S., Iino, M. & Hirose, K. Optical glutamate sensor for spatiotemporal analysis of synaptic transmission. *Eur. J. Neurosci.* **25**, 2249–2259 (2007).
- Marvin, J. S. et al. An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nat. Methods* **10**, 162–170 (2013).
- Masharina, A., Reymond, L., Maurel, D., Umezawa, K. & Johnsson, K. A fluorescent sensor for GABA and synthetic GABA(B) receptor ligands. *J. Am. Chem. Soc.* **134**, 19026–19034 (2012).
- Betz, H. Glycine receptors: heterogeneous and widespread in the mammalian brain. *Trends Neurosci.* **14**, 458–461 (1991).
- Xu, T.-L. & Gong, N. Glycine and glycine receptor signaling in hippocampal neurons: diversity, function and regulation. *Prog. Neurobiol.* **91**, 349–361 (2010).
- Johnson, J. W. & Ascher, P. Glycine potentiates the NMDA response in cultured mouse brain neurons. *Nature* **325**, 529–531 (1987).
- Schell, M. J. The N-methyl-D-aspartate receptor glycine site and D-serine metabolism: an evolutionary perspective. *Philos. Trans. R. Soc. Lond. B* **359**, 943–964 (2004).
- Citri, A. & Malenka, R. C. Synaptic plasticity: multiple forms, functions, and mechanisms. *Neuropsychopharmacology* **33**, 18–41 (2008).
- Nabavi, S. et al. Engineering a memory with LTD and LTP. *Nature* **511**, 348–352 (2014).
- Nong, Y. et al. Glycine binding primes NMDA receptor internalization. *Nature* **422**, 302–307 (2003).
- Ferreira, J. S. et al. Co-agonists differentially tune GluN2B-NMDA receptor trafficking at hippocampal synapses. *Elife* **6**, e25492 (2017).
- Henneberger, C., Papouin, T., Oliet, S. H. R. & Rusakov, D. A. Long-term potentiation depends on release of D-serine from astrocytes. *Nature* **463**, 232–236 (2010).
- Papouin, T. et al. Synaptic and extrasynaptic NMDA receptors are gated by different endogenous coagonists. *Cell* **150**, 633–646 (2012).
- Le Bail, M. et al. Identity of the NMDA receptor coagonist is synapse specific and developmentally regulated in the hippocampus. *Proc. Natl. Acad. Sci. USA* **112**, E204–E213 (2015).
- Chen, R.-Q. et al. Role of glycine receptors in glycine-induced LTD in hippocampal CA1 pyramidal neurons. *Neuropsychopharmacology* **36**, 1948–1958 (2011).
- Winkelmann, A. et al. Changes in neural network homeostasis trigger neuropsychiatric symptoms. *J. Clin. Invest.* **124**, 696–711 (2014).
- Hashimoto, A., Oka, T. & Nishikawa, T. Extracellular concentration of endogenous free D-serine in the rat brain as revealed by in vivo microdialysis. *Neuroscience* **66**, 635–643 (1995).
- Berntsson, R. P.-A., Smits, S. H. J., Schmitt, L., Slotboom, D.-J. & Poolman, B. A structural classification of substrate-binding proteins. *FEBS Lett.* **584**, 2606–2617 (2010).
- Planamente, S. et al. A conserved mechanism of GABA binding and antagonism is revealed by structure-function analysis of the periplasmic binding protein Atu2422 in *Agrobacterium tumefaciens*. *J. Biol. Chem.* **285**, 30294–30303 (2010).
- Van Durme, J. et al. A graphical interface for the FoldX forcefield. *Bioinformatics* **27**, 1711–1712 (2011).
- Morris, G. M. et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
- Bajar, B. T., Wang, E. S., Zhang, S., Lin, M. Z. & Chu, J. A guide to fluorescent protein FRET pairs. *Sensors (Basel)* **16**, 1488 (2016).
- Whitfield, J. H. et al. Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* **24**, 1412–1422 (2015).
- Deuschle, K. et al. Construction and optimization of a family of genetically encoded metabolite sensors by semirational protein engineering. *Protein Sci.* **14**, 2304–2314 (2005).
- Fritz, R. D. et al. A versatile toolkit to produce sensitive FRET biosensors to visualize signaling in time and space. *Sci. Signal.* **6**, rs12 (2013).
- Chen, X., Zaro, J. L. & Shen, W.-C. Fusion protein linkers: property, design and functionality. *Adv. Drug Deliv. Rev.* **65**, 1357–1369 (2013).
- Piston, D. W. & Kremers, G.-J. Fluorescent protein FRET: the good, the bad and the ugly. *Trends Biochem. Sci.* **32**, 407–414 (2007).
- Hamberger, A. & Nyström, B. Extra- and intracellular amino acids in the hippocampus during development of hepatic encephalopathy. *Neurochem. Res.* **9**, 1181–1192 (1984).
- Clifton, B. E. & Jackson, C. J. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem. Biol.* **23**, 236–245 (2016).
- Okubo, Y. et al. Imaging extrasynaptic glutamate dynamics in the brain. *Proc. Natl. Acad. Sci. USA* **107**, 6526–6531 (2010).
- Ermolyuk, Y. S. et al. Independent regulation of basal neurotransmitter release efficacy by variable Ca²⁺ influx and bouton size at small central synapses. *PLoS Biol.* **10**, e1001396 (2012).
- Maravall, M., Mainen, Z. F., Sabatini, B. L. & Svoboda, K. Estimating intracellular calcium concentrations and buffering without wavelength ratioing. *Biophys. J.* **78**, 2655–2667 (2000).
- Roux, M. J. & Supplisson, S. Neuronal and glial glycine transporters have different stoichiometries. *Neuron* **25**, 373–383 (2000).
- Tinberg, C. E. & Khare, S. D. *Computational Protein Design*. 363–373 (Humana Press, New York, NY, 2017).
- Tannous, B. A. et al. Metabolic biotinylation of cell surface receptors for in vivo imaging. *Nat. Methods* **3**, 391–396 (2006).
- Panatier, A. et al. Glia-derived D-serine controls NMDA receptor activity and synaptic memory. *Cell* **125**, 775–784 (2006).
- Horio, M. et al. Levels of D-serine in the brain and peripheral organs of serine racemase (Srr) knock-out mice. *Neurochem. Int.* **59**, 853–859 (2011).
- Matsui, T. et al. Functional comparison of D-serine and glycine in rodents: the effect on cloned NMDA receptors and the extracellular concentration. *J. Neurochem.* **65**, 454–458 (1995).
- Tønnesen, J., Inavalli, V. V. G. K. & Nägerl, U. V. Super-resolution imaging of the extracellular space in living brain tissue. *Cell* **172**, 1108–1121.e15 (2018).
- Rusakov, D. A. & Kullmann, D. M. Extrasynaptic glutamate diffusion in the hippocampus: ultrastructural constraints, uptake, and receptor activation. *J. Neurosci.* **18**, 3158–3170 (1998).
- Bethge, P., Chéreau, R., Avignone, E., Marsicano, G. & Nägerl, U. V. Two-photon excitation STED microscopy in two colors in acute brain slices. *Biophys. J.* **104**, 778–785 (2013).
- Cubelos, B., Giménez, C. & Zafra, F. Localization of the GLYT1 glycine transporter at glutamatergic synapses in the rat brain. *Cereb. Cortex* **15**, 448–459 (2005).
- Bergeron, R., Meyer, T. M., Coyle, J. T. & Greene, R. W. Modulation of N-methyl-D-aspartate receptor function by glycine transport. *Proc. Natl. Acad. Sci. USA* **95**, 15730–15734 (1998).
- Martina, M. et al. Glycine transporter type 1 blockade changes NMDA receptor-mediated responses and LTP in hippocampal CA1 pyramidal cells by altering extracellular glycine levels. *J. Physiol. (Lond.)* **557**, 489–500 (2004).
- Danglot, L., Rostaing, P., Triller, A. & Bessis, A. Morphologically identified glycinergic synapses in the hippocampus. *Mol. Cell. Neurosci.* **27**, 394–403 (2004).
- Langer, J. & Rose, C. R. Synaptically induced sodium signals in hippocampal astrocytes in situ. *J. Physiol. (Lond.)* **587**, 5859–5877 (2009).
- Chen, P. E. et al. Modulation of glycine potency in rat recombinant NMDA receptors containing chimeric NR2A/2D subunits expressed in *Xenopus laevis* oocytes. *J. Physiol. (Lond.)* **586**, 227–245 (2008).
- Le Meur, K., Galante, M., Angulo, M. C. & Audinat, E. Tonic activation of NMDA receptors by ambient glutamate of non-synaptic origin in the rat hippocampus. *J. Physiol. (Lond.)* **580**, 373–383 (2007).

50. Zhang, L.-H., Gong, N., Fei, D., Xu, L. & Xu, T.-L. Glycine uptake regulates hippocampal network activity via glycine receptor-mediated tonic inhibition. *Neuropsychopharmacology* **33**, 701–711 (2008).

Acknowledgements

We thank Dr. O'Mara (Australian National University) for helpful discussions. Research was funded by the Human Frontiers Science Program Young Investigator Award (HFSP to H.J., C.H., and C.J.J.; grant number: RGY0084/2012), German Academic Exchange Service (DAAD-Go8) Travel Fellowship (to C.H. and C.J.J.), NRW-Rückkehrerprogramm (to C.H.), the European Union (ITN EU-GliaPhD) and German Research Foundation (DFG, SFB1089 B03, SPP1757 HE6949/1 and HE6949/3, to C.H.).

Author contributions

W.H.Z., J.A.M., V.V., J.H.W. and C.J.J. designed, produced and analyzed the sensor. M.K.H., J.H.W., A.B.W., W.H.Z., D.M., B.B. and C.H. performed and analyzed all

experiments using two-photon excitation and electrophysiology in acute brain slices. M.K.H., J.H.W., I.S.-R., P.E.G., H.J., S.S. and A.B.W. performed studies on GlyFS expressed by cells. C.H., C.J.J. and H.J. designed the study. C.H., C.J.J. and W.H.Z. wrote the initial manuscript, to which all authors subsequently contributed.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41589-018-0108-2>.

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

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

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

Methods

DNA cloning and mutagenesis. Genes were ordered either from the Thermo Fisher GeneArt service with the desired sequence supplied in a nonexpression plasmid or from Integrated DNA Technologies as a gBlocks fragment. The desired DNA sequence was PCR amplified, gel purified and cloned through restriction digest and ligation (NdeI, EcoRI) into the vector PETMCS3, which enabled the expression of the binding protein by itself for the purposes of ITC analysis. The genes and relevant mutants were also cloned into the vector pDOTS10 through restriction digestion (SapI), which arranged the protein such that there was a biotin tag and ECFP linked at the N terminus and a VenusFP linked to the C terminus (also see ref. ²⁴). The restriction digestion method was also used for cloning into the Pertz kit array (Addgene, BspEI, NotI) and for replacement of the circularly permuted mTFP variants with ECFP in the same array (A1–A5). GlyFS was PCR amplified and cloned into pDisplay FLIPE-600n¹ (Addgene #13545, courtesy of W.B. Frommer) using SalI and XmaI restriction sites. To display GlyFS at the extracellular side of the plasma membrane, it was placed between the immunoglobulin κ-chain leader sequence and the PDGFR transmembrane domain.

Mutants/linker optimizations were generated through a Gibson assembly protocol in which fragments containing the desired mutation were created via PCR using large primers with at least 40 base-pair overlap. SnFR constructs were also cloned using Gibson assembly, with the circularly permuted FPs ordered from Integrated DNA Technologies as gBlocks.

Expression and purification of proteins. All proteins were expressed through transformation into BL21(DE3) *E. coli* cells and grown for 48–72 h at room temperature (20–25 °C) in 1 L autoinducing medium (yeast extract, 5 g; tryptone, 20 g; NaCl, 5 g; KH₂PO₄, 3 g; Na₂HPO₄, 6 g in 1,000 ml of water to which 10 ml autoclaved 60% glycerol, 5 ml autoclaved 10% glucose and 25 ml autoclaved 10% lactose were added) supplemented with 100 mg of ampicillin. In some cases, the full expression of the fluorescent protein constructs needed to be monitored by observing the ECFP/Venus spectra over time, which typically peaked at approximately 50 h of expression at 20 °C.

Cells were harvested through centrifugation, and the pellet was stored at –20 °C if it was not purified immediately after centrifugation. For purification, the pellet (frozen or otherwise) was suspended in buffer A (50 mM phosphate, 200 mM NaCl, 20 mM imidazole, pH 7.5), lysed through sonication, re-centrifuged at high speed (13,500 r.p.m. for 60 min at 4 °C) and the clarified supernatant collected. This was loaded onto a Ni-NTA/His-trap column, washed with 10-column volumes of buffer A (5 ml/min) and eluted with 4-column volumes of 100% buffer B (50 mM phosphate, 200 mM NaCl, 300 mM imidazole, pH 7.5), and the eluted protein was dialyzed against two exchanges of 4 L of buffer C (50 mM phosphate, 200 mM NaCl, pH 7.5). In the case of proteins with high affinity (determined by ITC) on-column refolding was performed, this followed the previous protocol with the exception that before elution with buffer B, the column was washed with 10 column volumes of buffer D (50 mM phosphate, 200 mM NaCl, 6 M guanidine, pH 7.5) and then returned from 100% buffer D slowly to 100% buffer A over a gradient for 3 h at room temperature at a lower flow rate (1 ml/min). After returning to 100% buffer A, the column was washed with an additional 10-column volumes of buffer A at a regular flow rate (5 ml/min) before elution with buffer B.

Further purification (when necessary) was performed on a HiLoad 26/600 Superdex 200 pg SEC column using buffer C. For FRET protein constructs, size-exclusion chromatography was performed in all cases and fractions tested for their maximum FRET range. Fractions that had a poor range compared to the maximal range were discarded, and the remaining viable fractions were pooled. In practice, this resulted in a batch of sensor that had a good dynamic range but not at the peak possible dynamic range.

Isothermal titration calorimetry. Binding studies were performed on a Nano-ITC at 25 °C with a stir rate of 250 r.p.m., and samples were degassed using a TA instruments degassing station (350 mm Hg). Protein concentrations used were between 50–100 μM, and ligand concentrations were between 0.5 to 5 mM depending on the affinity of the binding, with low ligand concentrations for higher affinity (<2 μM) and higher ligand concentrations for lower affinity (>2 μM). 3 μl injections of the ligand solution were injected every 200 s until a three-fold excess of ligand to protein was reached. The obtained data were processed with the NanoAnalyze 3.7.5 software provided.

Competition assays were performed by pre-incubating a known concentration of ligand B (the competitor) and then titrating the same protein sample with ligand A of a higher and known affinity. The observed affinity of ligand A (K_{Aobs}) was then used to determine the affinity of ligand B (K_B) using the equation below.

$$K_B = \left(\frac{K_A}{K_{Aobs}} - 1 \right) \times \frac{1}{[B]}$$

Fluorescence assays. Fluorescence titrations were performed on a Varian Cary Eclipse using a quartz narrow volume fluorescence cuvette. Samples underwent excitation at 433 nm and were scanned over a range of 460 nm to 560 nm for full

spectra analysis. ECFP/Venus ratios were determined using peak wavelength values of 525 nm (Venus) and 476 nm (ECFP).

Computational assessment of mutations. Mutations were assessed by first creating the mutation in YASARA with the FoldX v4.4.23 plugin enabled²¹, using the crystal structure of the protein of interest in the closed/bound state that had undergone the FoldX repair process and with bound ligand (if any) removed. The conformation with the lowest energy was selected from the set(s) generated through this method and was then analyzed with Autodock v4.2.6 (ref. ²²), within the Autodock tools suite, to assess the ability of desired and undesired ligands to fit/bind into the binding pocket of the protein. In scenarios in which more than one residue could provide the desired conformation or pocket shape, the one that was the least destabilizing or most stabilizing was selected. If this was not possible because the conformations/stabilizations were approximate, then the different residues were tested experimentally and one that offered the better binding profile/properties was selected.

Expression of GlyFS in cultured cells. HEK293T cells were cultured on glass coverslips in Dulbecco's modified Eagle medium (Thermo Fischer Scientific, USA), supplemented with FCS (10%) and penicillin/streptomycin (5%), at 37 °C in a 5% CO₂ atmosphere. The pDisplay-GlyFS plasmid was transfected into HEK293T cells using Lipofectamine 2000 (Invitrogen, USA) following the manufacturer's instructions. Primary rat cortical neurons, prepared as described previously from embryonic day 17–19 Wistar rat embryos²¹ and cultured on glass coverslips in Neurobasal medium (Thermo Fischer Scientific, USA) at 37 °C in a 5% CO₂ atmosphere, were transfected on DIV3–5 using the calcium phosphate method. Imaging experiments were performed 24–48 h after transfection (see Supplementary Figs. 9 and 10).

Brain slice preparation. Acute hippocampal slices were prepared from one- to five-week-old male Wistar rats as previously described²². All animals used in this study were housed under 12 h light/dark conditions and were allowed ad libitum access to food and water. Briefly, 300 μm thick acute hippocampal slices were obtained in full compliance with national and institutional regulations (Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen and University of Bonn Medical School) and guidelines of the European Union on animal experimentation. Slices were prepared in an ice-cold slicing solution containing (in mM): NaCl 60, sucrose 105, KCl 2.5, MgCl₂ 7, NaH₂PO₄ 1.25, ascorbic acid 1.3, sodium pyruvate 3, NaHCO₃ 26, CaCl₂ 0.5, and glucose 10 (osmolality 305–310 mOsm) and kept in the slicing solution at 34 °C for 15 min before being stored at room temperature in an extracellular solution containing (in mM) NaCl 131, KCl 2.5, MgSO₄ 1.3, NaH₂PO₄ 1.25, NaHCO₃ 21, CaCl₂ 2, and glucose 10 (osmolality 297–303 mOsm, pH adjusted to 7.4). This solution was also used for recordings. Slices were allowed to rest for at least 50 min. All solutions were continuously bubbled with 95% O₂/5% CO₂. Slice viability was tested electrophysiologically (see below). Slices not displaying prominent fEPSPs at low-to-moderate stimulus intensities were discarded.

Electrophysiology and sensor loading. GlyFS was lyophilized for long-term storage and transport for up to 2 months at ambient temperature (for shipping) and 4 °C (storage until use). Before experiments, the sensor was reconstituted in water and agitated for 30 min at room temperature before the buffer was changed to PBS (pH 7.4) with a PD-10 desalting column (GE Healthcare). GlyFS was then concentrated to 60–100 μM using centricons (Vivaspin 500, 10 kDa cutoff, Sartorius Stedim Biotech). No deterioration of sensor properties was detected over at least two months when reconstituted GlyFS was stored at 4 °C. For anchoring of optical sensor in brain tissue, cell surfaces within acute slices were biotinylated using a previously published procedure²⁴. Briefly, slice storage solution was supplemented with 50 μM Sulfo-NHS EZ Link Biotin (Thermo Fisher) for 45 min before washing and storage. Slices were transferred to a submersion-type recording chamber and superfused with extracellular solution at 34 °C. For injections of GlyFS into the tissue, patch-clamp pipettes (2–4 MΩ) were backfilled with PBS (pH 7.4) to which 50–85 μM GlyFS and 6–10 μM streptavidin (Life Technologies) had been added. The pipette was inserted ~70 μm deep into the tissue under visual control and GlyFS was pressure injected.

GlyFS-injected acute slices were allowed to recover for 10–15 min before recordings. The method of sensor delivery was chosen to restrict GlyFS to the extracellular space. This avoids, for instance, a contamination of optical recordings by genetically expressed GlyFS exposed to high intracellular glycine, which cannot be easily distinguished from extracellular GlyFS in intact neuropil using diffraction-limited imaging. The biotinylation approach is also unlikely to acutely induce astrogliosis, which, however, can be associated with viral transduction methods and protein overexpression⁵³ and may thus perturb glycine homeostasis. No impairment of synaptic transmission by the labeling method was detected previously²⁴. In addition, we have tested whether GlyFS specifically affects CA3–CA1 synaptic transmission, the model system of this study, but found no indication for that (Supplementary Fig. 13). Equivalent tests would be required for other experimental designs/models.

Linking a sensor via streptavidin to biotinylated tissue was previously shown by electron microscopy to lead to a primarily extra- and perisynaptic sensor localization³¹. This is expected if the sensor is linked to all surface membranes homogeneously, because the density of membrane surface per volume in the CA1 stratum radiatum neuropil is $\sim 14 \mu\text{m}^2/\mu\text{m}^3$ (ref. ⁴¹) and the total membrane surface facing the synaptic cleft of the two synapses found on average per μm^3 is, assuming a circular synaptic interface with a diameter of $0.3 \mu\text{m}$, only $4 \times \pi \times (0.15 \mu\text{m})^2/\mu\text{m}^3 = 0.3 \mu\text{m}^2/\mu\text{m}^3$. Therefore, it is reasonable to assume that the majority of GlyFS ($>95\%$) is located outside of synaptic clefts.

For extracellular recordings, the injection pipette or another patch pipette filled with extracellular solution were inserted into the CA1 stratum radiatum. Whole-cell recordings from CA1 pyramidal cells were obtained using standard patch pipettes ($2\text{--}4 \text{ M}\Omega$) filled with an intracellular solution containing (in mM) $\text{KCH}_3\text{O}_2\text{S}$ 135, HEPES 10, di-Tris-Phosphocreatine 10, MgCl_2 4, $\text{Na}_2\text{-ATP}$ 4, Na-GTP 0.4 (pH adjusted to 7.2 using KOH, osmolarity $290\text{--}295 \text{ mOsm}$). The membrane-impermeable dye Alexa Fluor 594 hydrazide ($200 \mu\text{M}$, Invitrogen) was added to the intracellular solution to visualize CA1 pyramidal cells, dendrites and spines. After $5\text{--}10 \text{ min}$ in whole-cell mode, the pipette was retracted gently to allow the cell's membrane to reseal. Data were recorded using MultiClamp 700B amplifiers, digitized (40 kHz) and stored for offline analysis. For stimulation experiments, a bipolar concentric stimulation electrode was placed in the stratum radiatum at the border between CA2/3 and CA1. The stimulus intensity was adjusted to evoke half-maximal field responses (fEPSPs). Channel or receptor blockers were added to the extracellular solution as indicated: strychnine ($1 \mu\text{M}$, Sigma Aldrich), D-APV ($50 \mu\text{M}$, Abcam), NFPS ($5 \mu\text{M}$, Tocris) Org25543 ($1 \mu\text{M}$, Tocris).

Two-photon excitation sensor imaging. Two-photon excitation sensor imaging was performed as previously described^{24,52,54}. GlyFS, CA1 pyramidal cells and their dendrites and spines were visualized by two-photon excitation (2PE) fluorescence microscopy. We used a FV10MP imaging system (Olympus) optically linked to a femtosecond pulse laser (Vision S, Coherent, $\lambda = 800 \text{ nm}$) integrated with patch-clamp electrophysiology (Multiclamp 700B, Molecular Devices) and equipped with a $25\times$ (NA 1.05) objective (Olympus). For titrations in solution, GlyFS was imaged in a meniscus of PBS at a laser power of 3 mW , and increasing amounts of glycine (in PBS) were added. For slice experiments, the laser power was adjusted for depth in the tissue to obtain, on average, $2\text{--}3 \text{ mW}$ in the focal plane. The glycine sensor GlyFS and Alexa Fluor 594 were both excited at 800 nm . GlyFS ECFP and Venus fluorescent protein and, in a subset of experiments, Alexa 594 fluorescence were separated using appropriate band-pass filters and dichroic mirrors. ECFP and Venus fluorescence signals were collected with photomultiplier tubes connected to a single photon counting board (PicoHarp, Picoquant). Their arrival times were recorded using Symphotime 1.5 software (Picoquant). Offline analysis was performed using OriginPro 2017 (OriginLab) and custom written scripts in Matlab R2017a (Mathworks). The ratio of ECFP and Venus fluorescence (R) was

calculated from the number of photons detected by the respective detectors in time bins of $\sim 220 \text{ ms}$. The photon count rate 11.5 ns to 12.5 ns after the laser pulse ($81\text{--}82 \text{ MHz}$ repetition rate) was used to reduce the contribution of emission-independent photons to analysis. Glycine was applied together with strychnine, NFPS, Org25543 and D-APV at a saturating concentration of 5 mM at the end of all experiments to determine R_{MAX} , the ECFP/Venus fluorescence emission ratio of the fully glycine-bound sensor. R_{MAX} was used for normalization between experiments to account for variable emission, excitation and emission detection between different brain slices and imaging depths. To reduce the potential effect of scattering of ECFP and Venus fluorescence we have performed all imaging experiments at a depth of $50\text{--}70 \mu\text{m}$ below the slice surface. For experiments investigating resting glycine concentrations at spines and dendritic shafts an image stack ($xy\ 34 \times 34 \mu\text{m}$, $z\ 0.5 \mu\text{m}$) was acquired of segments of the neuron's apical dendritic arbor 10 min after GlyFS injection. After sensor saturation with glycine for 15 min another image stack was acquired. During offline analysis, regions of interest of $1 \mu\text{m}^2$ area were analyzed around spines (arced ROIs around the spine circumference) and along dendritic shafts (box-shaped ROIs) in the imaging plane, in which they appeared optimally in focus.

Statistics. Data are reported as mean \pm s.e.m., unless stated otherwise, with n representing the number of independently performed experiments as explained along with individual data. For experiments in acute brain slices, one experiment was performed per acute slice. Experimental results were typically obtained from one to two acute slices per animal. Statistical tests were always two-sided and used as indicated. The significance level (P) is stated in figure legends and illustrated in figures by asterisks as described. Please see figure legends for further details. Due to the experimental design, the analysis was not performed in a blinded manner.

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

Data availability. All relevant data and materials are available from the authors upon reasonable request.

References

- Woitecki, A. M. H. et al. Identification of synaptotagmin 10 as effector of NPAS4-mediated protection from excitotoxic neurodegeneration. *J. Neurosci.* **36**, 2561–2570 (2016).
- Anders, S. et al. Spatial properties of astrocyte gap junction coupling in the rat hippocampus. *Philos. Trans. R. Soc. Lond. B* **369**, 20130600 (2014).
- Ortinski, P. I. et al. Selective induction of astrocytic gliosis generates deficits in neuronal inhibition. *Nat. Neurosci.* **13**, 584–591 (2010).
- Minge, D. et al. Heparan sulfates support pyramidal cell excitability, synaptic plasticity, and context discrimination. *Cereb. Cortex* **27**, 903–918 (2017).

Supplementary table 1

Ligand affinity between mutants (K_D, μM)					
	Wild Type	Leu202Trp	Phe77Ala, Ala100Tyr	Phe77Ala, Leu202Trp	Phe77Ala, Ala100Tyr, Leu202Trp
glycine	0.11 ± 0.06	32.7 ± 8.2	18.2 ± 4.0	2.25 ± 0.76	20.0 ± 3.7
GABA	2.1 ± 0.69	nd	nd	nd	nd
L-serine	< 2, *	34.5 ± 11.4	20 ± 6.3	nd	nd
glutamate	nd	-	nd	2.3 ± 1.02	nd
Competition ITC (glycine + presence of 500 μM GABA/L-ser/glu)					
GABA					20.3 ± 3.6
L-serine					20.4 ± 2.5
glutamate					20.5 ± 3.3

Binding affinities of mutants.

Summary of the binding properties of the various mutants at 25 °C. Affinities were determined either directly or through inhibition/competition of glycine. A ligand was considered to show no detectable (nd) binding if there was both no binding observed through a direct titration ITC experiment and if it showed no inhibition of glycine binding in a competition titration (* for L-serine affinity please see ¹). Data presented as mean \pm SEM (see Fig. 1B and legend for further experimental details). The affinity for glycine was also measured in the presence of 500 μM GABA, L-serine and glutamate. For these competition ITCs, means \pm 90% confidence interval obtained from analyzing ITC data (see methods section) are shown (n = 1, 2, 2 independent experiments for GABA, L-serine and glutamate, respectively).

Supplementary table 2

Sensor construct number (as per Fritz <i>et al.</i> ²)	Dynamic range (mTFP/Venus or ECFP/Venus, % of apo)	Sensor design structure
A1	-1.8%	mTFP-WT Venus-WT
A2	#	mTFP-WT Venus-157
A3	#	mTFP-WT Venus-173
A4	#	mTFP-WT Venus-195
A5	#	mTFP-WT Venus-229
A6	2.8%	mTFP-105 Venus-WT
A7	2.1%	mTFP-105 Venus-157
A8	#	mTFP-105 Venus-173
A9	1.1%	mTFP-105 Venus-195
A10	2.4%	mTFP-105 Venus-229
A11	4.4%	mTFP-159 Venus-WT
A12	3.2%	mTFP-159 Venus-157
B1	#	mTFP-159 Venus-173
B2	#	mTFP-159 Venus-195
B3	#	mTFP-159 Venus-229
B4	2.1%	mTFP-175 Venus-WT
B5	1.3%	mTFP-175 Venus-157
B6	#	mTFP-175 Venus-173
B7	#	mTFP-175 Venus-195
B8	#	mTFP-175 Venus-229
B9	#	mTFP-227 Venus-WT
B10	-2.2%	mTFP-227 Venus-157
B11	#	mTFP-227 Venus-173
B12	-1.4%	mTFP-227 Venus-195
C1	-2.3%	mTFP-227 Venus-229
A1-ECFP	4.2%	ECFP-WT Venus-WT
A2-ECFP	7.5%	ECFP-WT Venus-157
A3-ECFP	#	ECFP-WT Venus-173
A4-ECFP	8.7%	ECFP-WT Venus-195
A5-ECFP	7.5%	ECFP-WT Venus-229

Dynamic range of sensors created using the commercially available Pertz kit array (design type 1 from ²).

The table above shows the fluorescence ratio (mTFP/Venus or ECFP/Venus) in the presence of 1 mM glycine normalized to control (apo, 0 mM glycine) for each sensor construct. # indicates that the fluorescence intensity ratio changed by less than one percent. The sensor design structure refers to the circular permutation residue location of the fluorescent proteins used. For example, A3

corresponds to a wild type mTFP and a Venus fluorescent protein that has been circularly permuted at residue 173. The highest response of the original Pertz kit was ~ 4% and thus the biggest dynamic range was measured for A-11. When mTFP was replaced with ECFP, a dynamic range of 8.7 % could be obtained (A4-ECFP). The naming system (A1-C1) is based on the original sensor design array described by Fritz *et al.* ² as per the commercially available Pertz Kit.

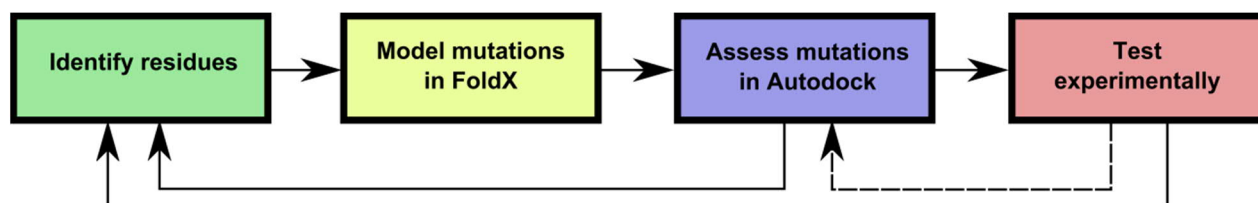
Supplementary table 3

	Single-photon excitation	Two-photon excitation
Affinity for glycine (K_D)	27.7 μ M	21.4 \pm 1.1 μ M
Dynamic range (R_{MAX}/R_{MIN})	28.3 %	18.5 \pm 1.4 %

Comparison of GlyFS properties between single and two-photon excitation.

Dynamic range (maximum change of GlyFS fluorescence intensity ratio [ECFP/Venus] from zero glycine to a saturating glycine concentration) and affinity (K_D) for single-photon excitation (single batch, experiments from Fig. 1) and two-photon excitation (mean \pm SEM, five independent experiments and batches as shown in Fig. 2). The K_D is relatively close to the extracellular resting glycine concentration (see Discussion) but far below the concentrations found in the cytosol and the peak glycine concentration in the synaptic cleft during glycinergic synaptic transmission, which are both in the millimolar range ^{3,4}.

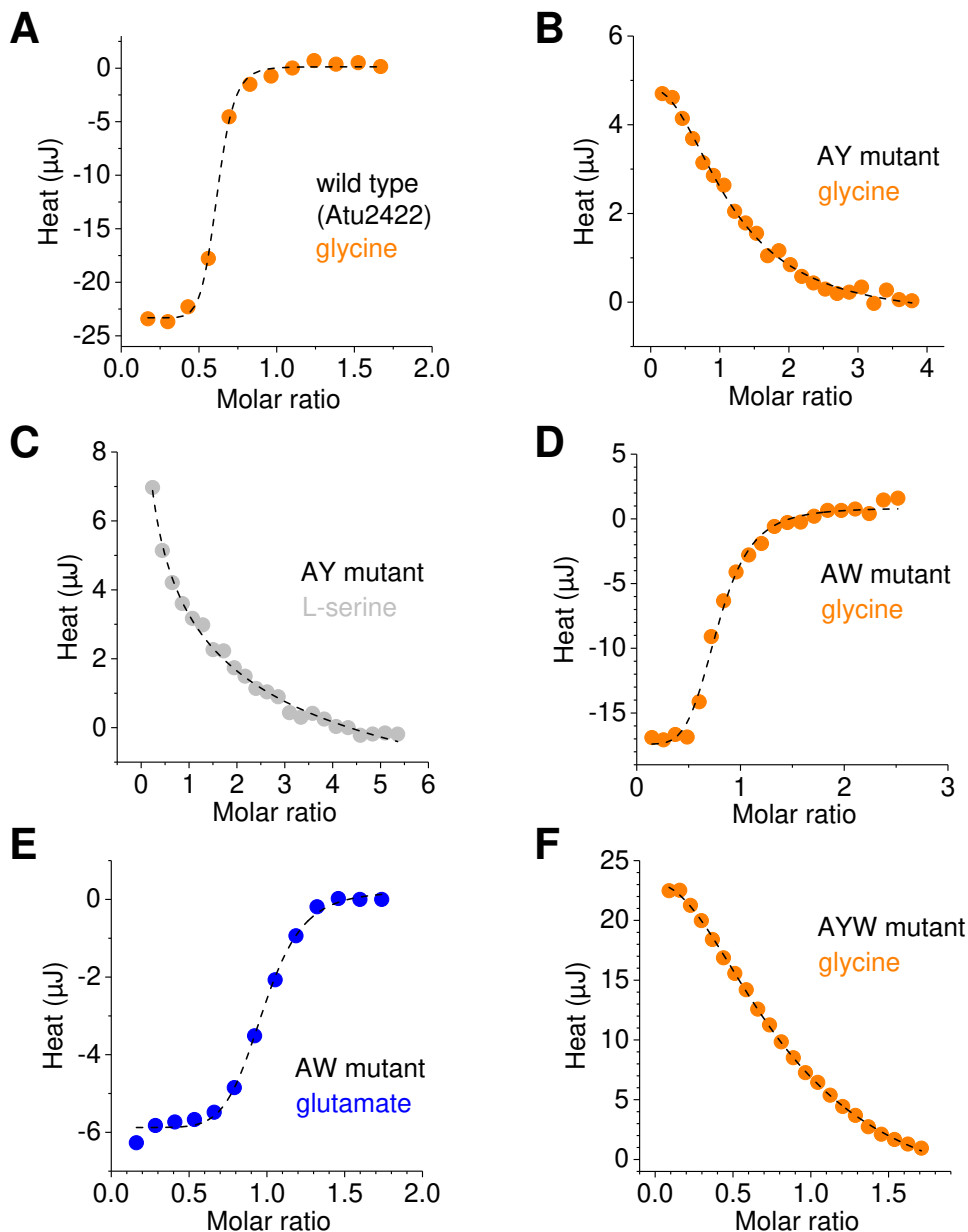
Supplementary figure 1



Workflow for engineering a glycine-specific binding protein

Engineering of the binding protein for glycine binding specificity was done as an iterative process, transitioning between computational modelling and experimental validation. The basic workflow is illustrated above. We first identified residues in the binding pocket that could potentially prevent unwanted ligands from binding (green) if modified. Next, mutations of these residues were modeled using FoldX to predict residue conformations ⁵ (yellow). In the third step (blue), Autodock was used to evaluate the capacity of ligands to bind into the modified active site ⁶ using the mutant models generated from FoldX. Depending on results, new residues were identified (green) or the mutant binding protein was then tested experimentally (orange). Experimental data was used to revise docking parameters, to identify important residues and residue behaviors, which all then guided further mutations.

Supplementary figure 2



Isothermal titration calorimetry (ITC) of the wild type binding protein and its mutants. The binding isotherms are displayed as heat per ligand injection vs. molar ratio. Mutant and ligand as indicated in each panel. Please see methods section for details. A single representative example is shown for each experiment. See Fig. 1B for summary data of all experiments.

A) ITC binding profile of the wild type binding protein (Atu2422) to glycine. Ligand binding is exothermic (enthalpy driven).

B) ITC binding profile of the Phe77Ala/Ala100Tyr mutant to glycine. Ligand binding is endothermic (entropy driven).

C) ITC binding profile of the Phe77Ala/Ala100Tyr to L-serine, with the same approximate affinity as glycine. Ligand binding is endothermic (entropy driven).

D) ITC binding profile of the Phe77Ala/Leu202Trp to glycine. Ligand binding is exothermic (enthalpy driven).

E) ITC binding profile of the Phe77Ala/Leu202Trp to glutamate, with the same approximate affinity as glycine. Ligand binding is exothermic (enthalpy driven).

F) ITC binding profile of the Phe77Ala/Ala100Tyr/Leu202Trp to glycine. Ligand binding is endothermic (entropy driven).

Also note that the selectivity of AYW was determined through direct and competitive ITC titrations and also through fluorescence titrations after the conversion of AYW into the sensor GlyFS. GlyFS was titrated (in duplicate) with a target ligand (the 20 canonical amino acids, D-serine and GABA) to a final concentration of 500 μ M in order to determine if there were any observable ratiometric changes. Ligands that resulted in a fluorescence ratio change $> 1\%$ (glycine, leucine, valine, threonine) underwent a dose-response titration in order to determine the affinity of the sensor for that ligand. Ligands, which did not produce an observable fluorescence ratio change of GlyFS, would have either no affinity, or would necessarily have a binding affinity substantially weaker than that of threonine (K_D of 6 mM).

Supplementary figure 3



B

	DR	mTFP/ ECFP	Flex.	SBP	Flex.	VFP
Full length	4%	...AGI	TLGMDELYKGGTGIM	DVV...IQQ	GGVSKGE	ELF...
Truncation 1	6%	...AGI	TLGMDELYKGGTGIM	DVV...IQQ	GG-----	ELF...
Truncation 2	8%	...AGI	-----GGTGIM	DVV...IQQ	GG-----	ELF...
Truncation 3	8%	...AGI	-----	DVV...IQQ	GG-----	ELF...
Truncation 4	0%	...AGI	-----	DVV...IQQ	-----	ELF...

C

	DR	mTFP/ ECFP	Flex.	SBP	Flex.	VFP
Best reoriented (mTFP)	5%	...AGI	SGM	DVV...IQQ	AAAM	ELF...
Best reoriented (ECFP)	9%	...AGI	SGM	DVV...IQQ	AAAM	ELF...

D

	DR	mTFP/ ECFP	SBP	Linker 2	VFP
(EAAAK) ₃	28%	...AGI	DVV...IQQ	-- (EAAAK) 3 --	VSKGELF...
(EAAAK) ₃ GP	22%	...AGI	DVV...IQQ	-- (EAAAK) 3GP	VSKGELF...
GS(EAAAK) ₃ GS	18%	...AGI	DVV...IQQ	GS (EAAAK) 3GS	VSKGELF...
(EAAAK) ₆	9%	...AGI	DVV...IQQ	-- (EAAAK) 6 --	VSKGELF...

The modular design of GlyFS variants and linker regions.

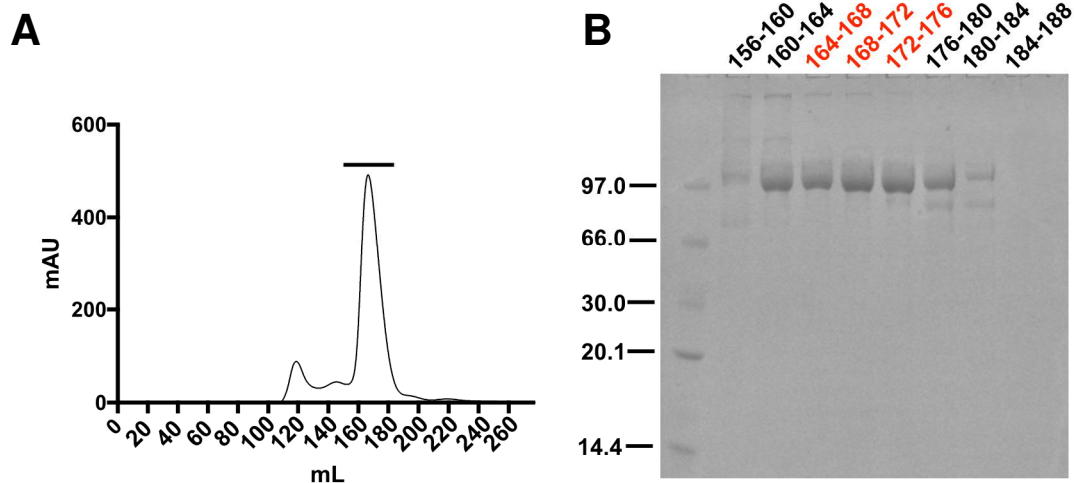
A) The N-terminal region of GlyFS comprises a biotin tag (white) for immobilization. The solute binding protein (SBP, black), in this case the glycine binding protein, is then sandwiched between two fluorescent proteins, ECFP (cyan) at the N-terminus and Venus (yellow) at the C-terminus. Linking the glycine binding protein to the fluorescent proteins are two linker regions (red), which are variable.

B) Five iterations with flexible linkers were constructed. Initial truncations were based on ⁷. Each variant has different flexible regions comprised on the N- or C-termini of the fluorescent proteins (cyan and yellow) and additional amino acids introduced as linkers (red). Relatively little change in the dynamic range (DR, % increase of fluorescence intensity ratio) was achieved in any of these variants (max gain of ~ 4%).

C) Using the method of Fritz and colleagues ² (design type 1), the positions at which the binding protein was fused to the fluorescent protein were randomized by circular permutation of the fluorescent proteins to test different orientations of the fluorescent proteins. This did not yield large changes in the dynamic range (DR) when used with monomeric Teal Fluorescent Protein (mTFP), nor when mTFP was replaced with ECFP.

D) The removal of any flexible linker to the ECFP and the introduction of rigid (EAAAK) linkers of different length led to substantial increases in dynamic range. The best variant included three EAAAK repeats with no additional linker residues.

Supplementary figure 4

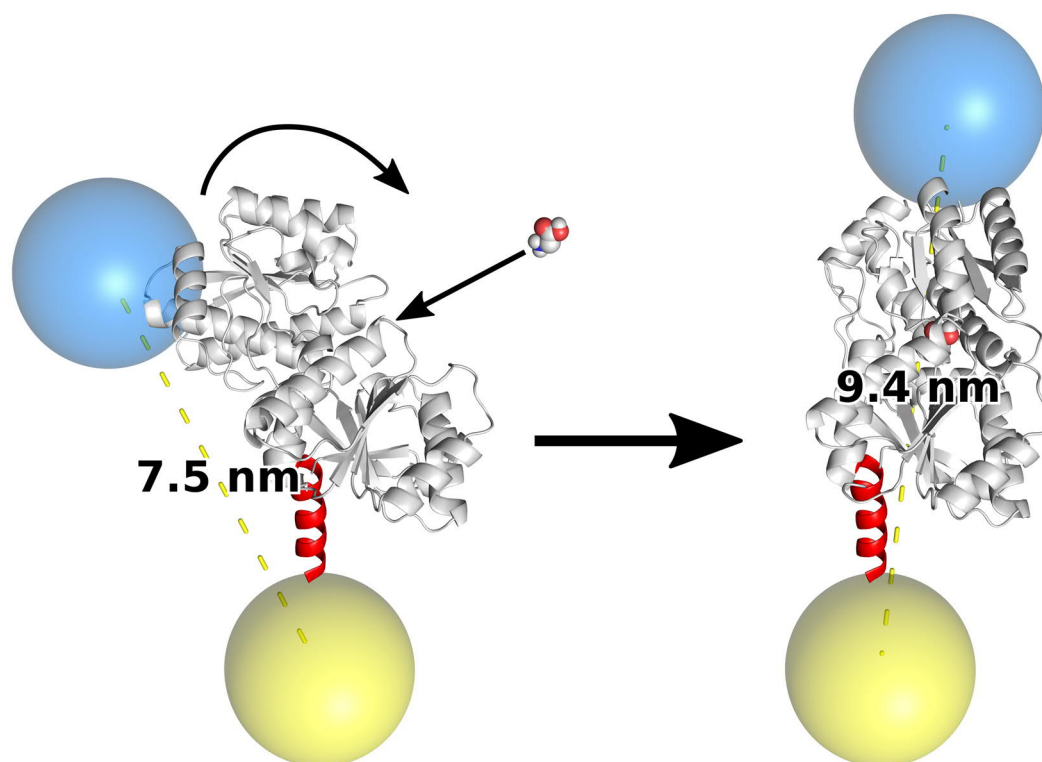


The purity of GlyFS used in experiments. Single example, routinely performed for individual sensor batches.

A) Size exclusion chromatography (hiload 26/600 superdex 200 pg) of the protein after Ni²⁺-affinity chromatography reveals the protein is largely pure after affinity chromatography. A single major peak was obtained between 156-188 mL.

B) Eight 4 mL fractions between 156-188 mL were analyzed by SDS-PAGE, revealing the most pure fractions to be between 164-176 mL (colored red). There was a single major band in these fractions, corresponding to the correct size for GlyFS (107 kDa). No breakdown products were observed. These fractions were pooled and used for subsequent imaging experiments.

Supplementary Figure 5

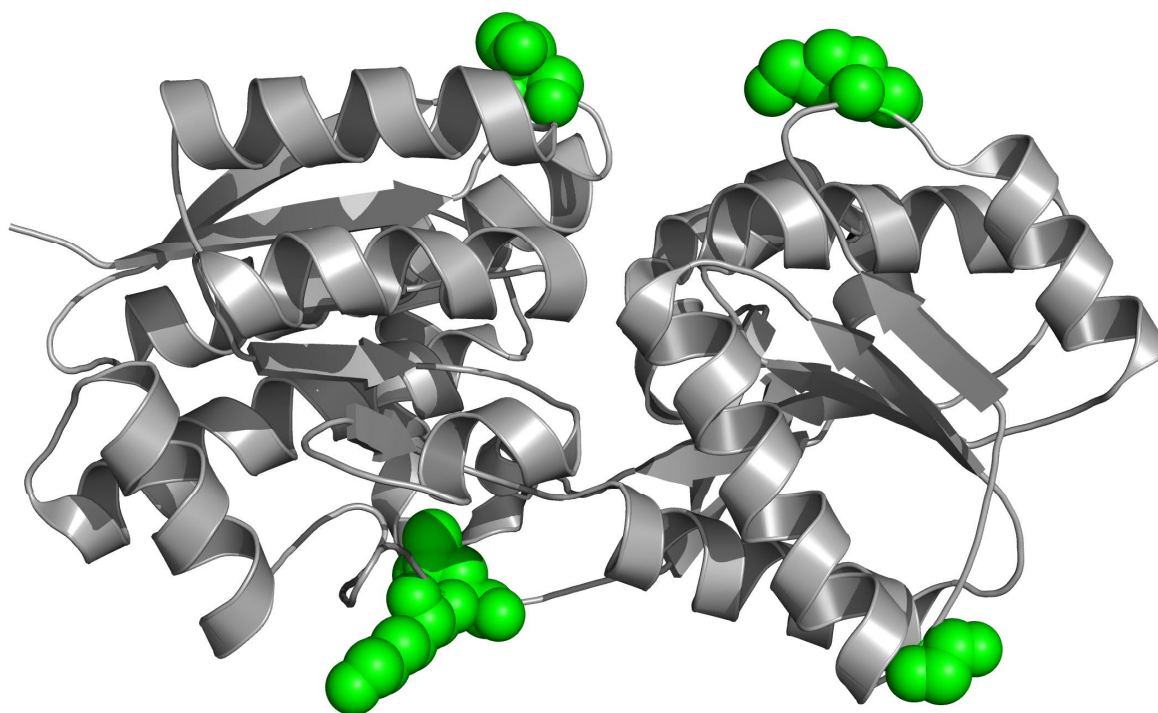


GlyFS' reduction of FRET efficiency after glycine binding.

Unlike many SBP-based ratiometric biosensors, GlyFS' FRET efficiency decreases in the presence of glycine (the ECFP/Venus fluorescence intensity ratio increases). In other words, the fluorophores move away from each other upon substrate binding. This is consistent with this structural model, which shows the protein's termini to be near in space to the hinge region. When the two lobes bend away from each other in the open state (left), the three domains form a triangle. This triangle straightens into a line upon glycine complexation (right), increasing the interfluorophore distance.

Models of the binding cores were constructed by homology. A crystal structure of the wild-type binding core (Atu2422, PDB ID 3IP5) ¹ bound to alanine was used as the template for the closed state. No such structure is available for the open structure, so the apo structure of the homologous leucine binding protein was used (PDB ID 1USG) ⁸. This template was validated by its sequence similarity (58% amino acid similarity; 40% amino acid identity) and the close structural similarity between ligand-bound states (PDB ID 1USK) ⁸ (1.0 Å C_α RMSD). The homology models were created by one-to-one threading on the Phyre2 web portal ⁹. The helical linker was added to the model using the Rosetta energy function and FoldIt ¹⁰. Spheres merely highlight the approximate size of the fluorescent proteins and the locations of the binding core's termini to demonstrate the putative change in interfluorophore distance.

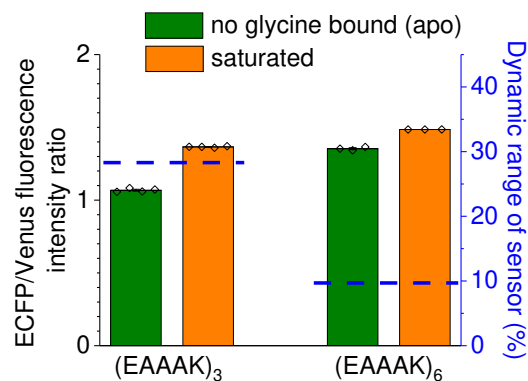
Supplementary figure 6



Circularly permuted ECFP (cpECFP) insertion locations tested.

The promiscuous solute-binding protein Atu2422 from *Agrobacterium tumefaciens*, with the cpECFP insertion locations shown as green spheres (residues Thr12, Gly166, Lys180, Lys326 and Leu327). The sensor design used here is based on the sensor developed by Marvin *et al.*¹¹ (intensity-based glutamate-sensing fluorescent reporter, iGluSnFR). These insertions did not create sensors with observable changes in fluorescent spectra upon saturation with glycine.

Supplementary figure 7

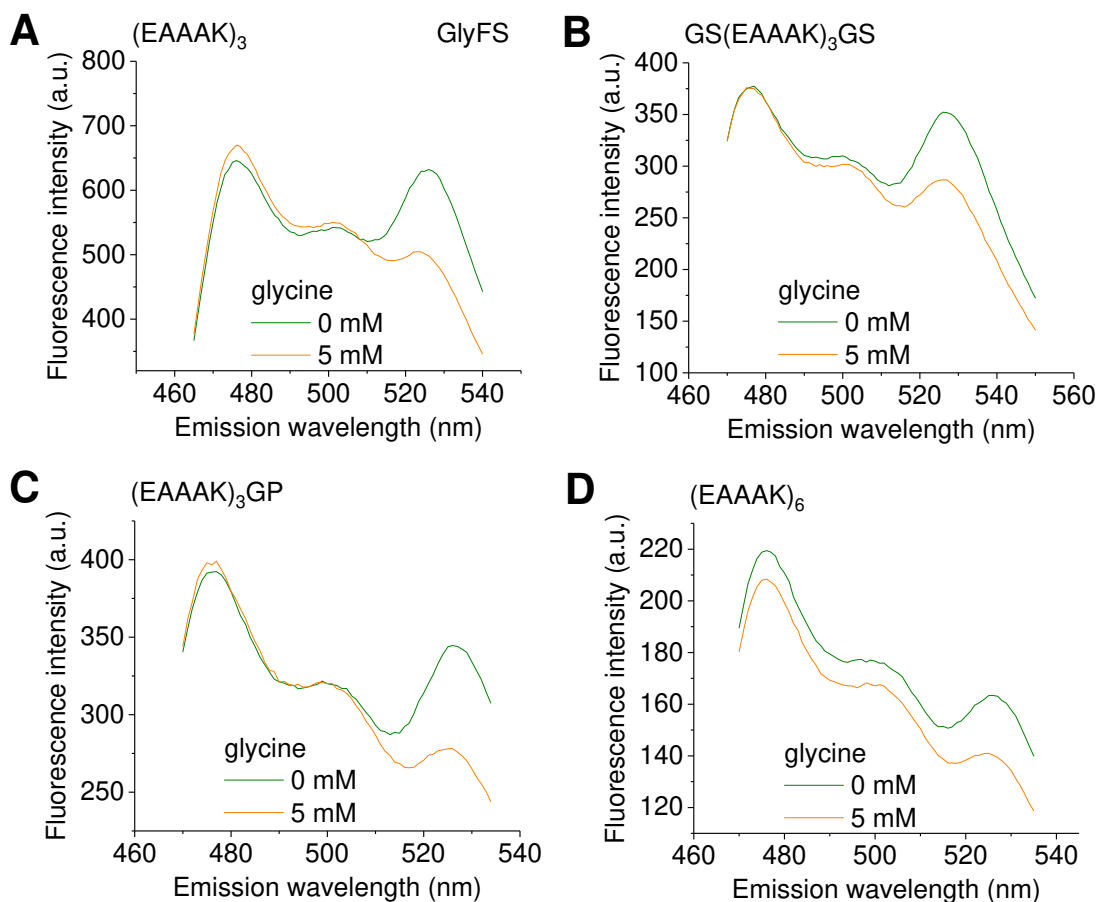


Dependence of FRET efficiency on linker length.

Binding of glycine decreases the FRET efficiency of GlyFS (increase of the ECFP/Venus fluorescence intensity ratio, also see supplementary figure 5). The (EAAAK)₃ (~ 2 nm) helical linker used in GlyFS works in part by tuning the interfluorophore distance to be close to the optimal Förster distance of the system. In this region of the FRET efficiency function ($E(r) = \frac{1}{1+(r/R_0)^6}$; $r \approx R_0$; $E(R_0) = 0.5$), a small change in the distance produces a large concomitant change in efficiency. Extending the linker to (EAAAK)₆ (~ 4 nm) overall increases the ECFP/Venus ratio (decreases FRET efficiency, right bars) and reduces the ECFP/Venus ratio changes from zero to saturating glycine concentrations thereby decreasing the dynamic range of the sensor, most likely by moving further away from the optimal Förster distance. The similarity between the saturated ratio of the ~ 2 nm linker and the apo ratio of the ~ 4 nm linker is consistent with the ~2 nm change in interfluorophore distance in the structural model (supplementary figure 5). However, no single Förster distance is compatible with any other combination of these distances, suggesting that orientation effects also play a crucial role in explaining changes in sensor FRET efficiencies.

Dynamic ranges of the sensor fluorescence ratios are indicated by dashed blue lines (right y-axis). All data taken from experiments displayed in Fig. 1C-D. Data are represented as mean \pm SEM (n = 4 independent experiments for (EAAAK)₃ and 3 for (EAAAK)₆, individual data points as overlay).

Supplementary figure 8



The emission spectra of the glycine sensor candidates with different linkers.

The emission spectra (arbitrary fluorescence intensity units, a.u.) are shown for the various sensor constructs in the absence (green) of glycine and in the presence of a saturating glycine concentration of 5 mM (orange). Excitation of ECFP at 433 nm. Linker displayed at the top of each panel. Representative single examples for corresponding data sets in Fig. 1C.

A) Spectrum of the sensor with the (EAAAK)₃ linker, the linker eventually used for the glycine sensor GlyFS.

B) Spectrum of the sensor with a linker modified for increased flexibility, GS(EAAAK)₃GS.

C) Spectrum of the sensor with a linker that alters fluorophore orientation through the addition of a proline residue, (EAAAK)₃GP.

D) Spectrum of the sensor with a longer linker (additional repeats), (EAAAK)₆.

Supplementary figure 9

GlyFS was genetically expressed in cultured human cells (HEK293) and neurons. Membrane targeting was achieved by introducing the sensor into the pDisplay vector (pDisplay-GlyFS)¹². See below for full DNA and protein sequences of pDisplay-GlyFS (color-code: **lg k-chain leader sequence**, **hemagglutinin A epitope**, **ECFP**, **binding core**, **rigid linker**, **VenusFP**, **myc epitope**, **PDGFR transmembrane domain**) and supplementary figure 10 for results.

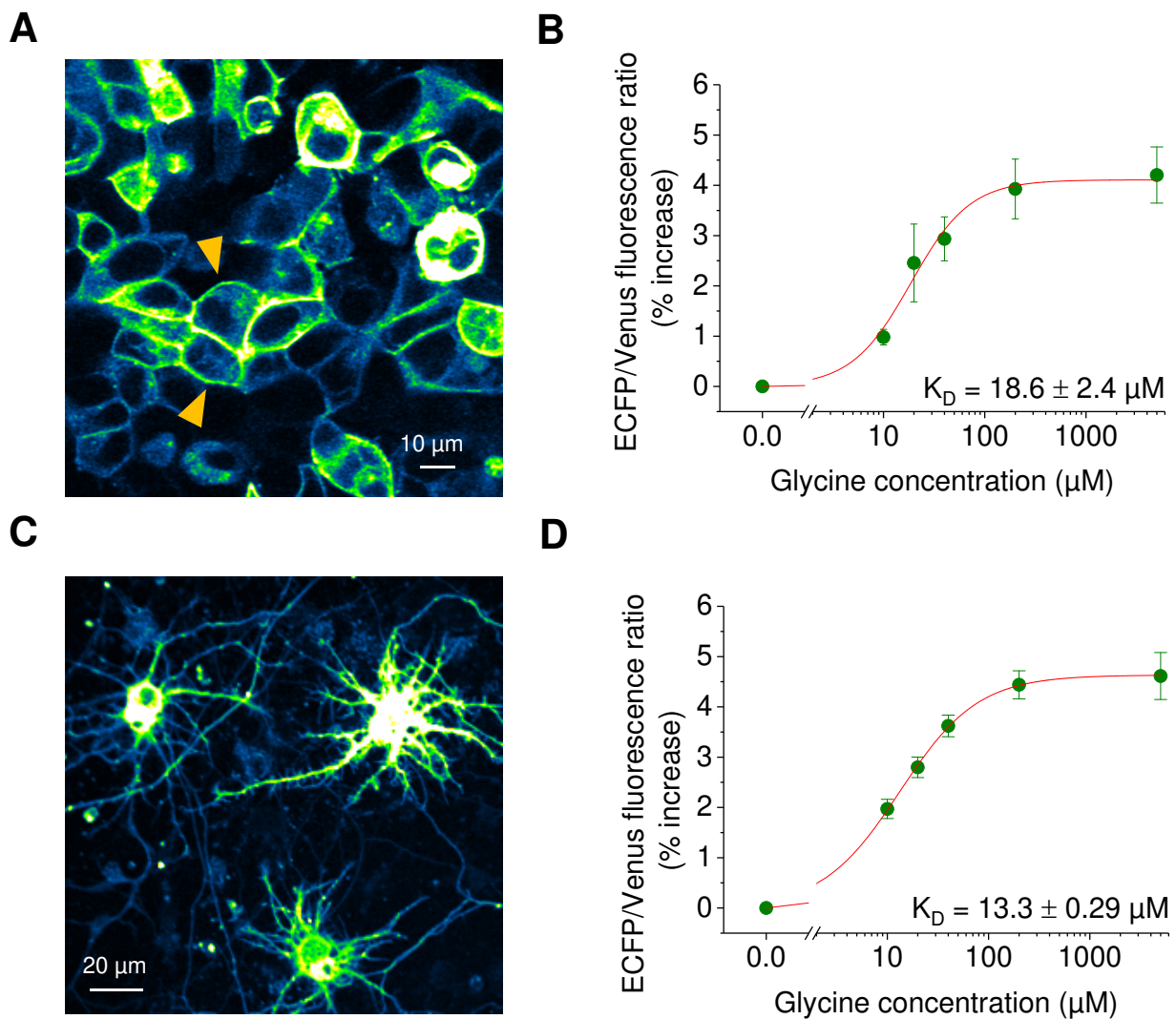
DNA

```
ATGGAGACAGACACACTCCTGCTATGGGTACTGCTGCTCTGGGTTCAGGTTCCACTGGTGACTATCCATATGATGTTCCAGATTATGC
TGGGGCCAGCCGGCCAGATCTCCCGGGGATCCGGGCCGCATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCATCTGG
TCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCCTGAAG
TTCACTGTCACCACCGGCAAGCTGCCCGTGCCTGGCCACCCTCGTGACCACCTGACCTGGGGCGTGAGTGCTTCAGCCGCTACCC
CGACCACATGAAGCAGCAGACTTCTTCAAGTCCGCCATGCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCA
ACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGC
AACATCCTGGGGCACAAGCTGGAGTACAACATACATCAGCCACAACGTCTATATCACCGCCGACAGCAGAAGAACGGCATCAAGGCCAA
CTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCCTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGC
TGCTGCCCCGACAACCACTACCTGAGCACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCTCTGCTGGAGTTC
GTGACCGCCCGCGGGATCGATGTTGTTATTGCAGTTGGTGCACCGCTGACCGGTCCGAATGCAGCATTTGGTGCACAGATTAGAAAAGG
TGCAGAACAGGCAGCAAAAGATATTAATGCAGCCGGTGGTATTAAATGGCGAGCAGATTAAATCGTCTGGGTGATGATGTTAGCGATC
CGAAACAGGGTATTAGCGTTGCCAATAAATTCGTTCAGATGGCGTTAAATTTGGTGGGTTCATGCGAACACGGGTGTTAGCATTCCG
GCAAGCGAAGTTTATGCAGAAAATGGTATTCTCGAGATTACACCGTATGCAACCAATCCGGTTTTTACCGAACGTGGTCTGTGGAATAC
CTTTCGTACCTGCGGCCGCGACGATCAGCAGGGTGGTATTGCAGGTAAATATCTGGCAGATCATTCAAAGATGCCAAAGTGGCCATCA
TCCATGATAAAACCCCGTATGGTCAGGGTCTGGCCGATGAAACCAAAAAGCAGCAATGCAGCGGGTGTACCGAAGTTATGTATGAA
GGTGTAAATGTGGCGATAAAGATTTTAGCGCACTGATCAGCAAAATGAAAGAAGCAGGCGTTAGCATTATCTATTGGGGTGGTTGGCA
TACCGAAGCAGGTCTGATTATTGTCAGGCAGCAGATCAGGGCCTGAAAGCAAACTGGTTAGCGGTGATGGTATTGTTAGCAATGAAC
TGGCAAGCATTGCCGGTGATGCAGTTGAAGGCACCCTGAATACATTTGGTCCTGATCCGACCCTGCGTCCGAAAAATAAAGAACTGGTT
GAAAAATTCAAAGCCGAGGCTTTAATCCGGAAGCATATACCTGTATAGCTATGCAGCAATGCAGGCAATTGCGGGTGCAGCCAAAGC
AGCAGGTAGCGTTGAACCGGAAAAAGTTGCAGAAGCACTGAAAAAAGGTAGCTTTCCGACCGCACTGGGTGAAATCAGCTTTGATGAAA
AAGGTGATCCTAAACTGCCTGGCTATGTGATGTATGAATGGAAAAAGGACCGGATGGCAAATTCACCTATATTCAGCAGGAAGCAGCA
GCAAAAGAAGCCGCTGCCAAAGAAGCGGCAGCGAAAGTGAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCATCCTGGTCGAGCT
GGACGGCGCAGCTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCTGAAGCTGATCT
GCACCACCGGCAAGCTGCCCGTGGCCCTGGCCACCCCTCGTGACCAACCCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCAC
ATGAAGCAGCAGCACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAA
GACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCC
TGGGGCACAAGCTGGAGTACAACATAACAGCCACAACGTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACTTCAAG
ATCCGCCACAACATCGAGGACGGCGGCGTGCAGCTCGCCGACCCTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCC
CGACAACCACTACCTGAGCTACCACTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCTCTGCTGGAGTTTCGTGACCG
CCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGTCGACGAACAAAACTCATCTCAGAAGAGGATCTCAATGCTGTGGGCCAG
GACACGCAGGAGGTATCGTGGTGCCACACTCCTTGCCCTTTAAGGTGGTGGTGATCTCAGCCATCCTGGCCCTGGTGGTGCTACCCAT
CATCTCCCTTATCATCCTCATCATGCTTTGGCAGAAGAAGCCACGTTAG
```

Protein

```
METDTLLLLWVLLWVPGSTGDXPYDVPDYAGAQPARSPGDPRMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEEDATYGLKTLK
FICTTGKLPVPWPTLVTTLTWGVQCFSRYPDHMKQHDFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDG
NILGHKLEYNYISHNVYITADKQKNGIKANFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLEF
VTAAGIDVVIAVGAPLTGPNAAFGAQIQKGAEQAAKDINAAGGINGEQIKIVLGDDVSDPKQGISVANKFVADGVKFVVGHANSVGSIP
ASEVYAENGILEITPYATNPVFTERGLWNTFRTCGRDDQGGIAGKYLADHFKDAKVAIIHDKTPYQGLADETKKAANAAGVTEVME
GVNVGDKDFSALISKMEAGVSIIYWGWHTEAGLIIRQAADQGLKAKLVSGDGIVSNELASIAGDAVEGTLNTFGPDPTLRPENKELV
EKFKAAGFNPEAYTLYSYAAMQAIAGAAKAAGSVEPEKVAEALKKGSFPTALGEISFDEKGDPKLPGYVMYEWKKGPDGKFTYIQQEAA
AKEAAAKEAAAKVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEEDATYGLKTLKLICTTGKLPVPWPTLVTTLTGYGLQCFARYPDH
MKQHDFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYSNVYITADKQKNGIKANFK
IRHNIEDGGVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLEFVTAAGITLGMDELYKVDEOKLISEEDINAVGQ
DTQEVIVVPHSLPFKVVISAILALVVLTIISLIILIMLWQKKPR
```

Supplementary Figure 10



Detection of extracellular glycine changes by membrane-targeted GlyFS expressed by HEK293 cells and neurons. Data is presented as mean \pm SEM in **B** and **D**.

A) HEK293 cells were transfected with pDisplay-GlyFS (see Methods section). Representative sample fluorescence image (ECFP) obtained using two-photon excitation at 800 nm two days after transfection in the nominal absence of extracellular glycine (from data set analysed in **B**). Note the prominent membrane labeling of HEK293 cells by pDisplay-GlyFS but also presence of some sensor in the cytosol (orange arrow heads).

B) Calibration of pDisplay-GlyFS expressed by HEK293 cells ($n = 5$ independent experiments from individual coverslips, from 2 separate cultures and transfections). The ratio of ECFP and Venus fluorescence intensities was calculated and expressed as % increase over 0 glycine. The K_D of 18.6 μM was obtained by fitting a Hill-type equation to the average of the five titrations. The maximum response was 4.2 %.

C) Dissociated neuronal cultures were transfected with pDisplay-GlyFS at day *in vitro* 3-5. Representative example fluorescence image of neurons expressing pDisplay-GlyFS three days after transfection in the nominal absence of extracellular glycine (from data set analysed in **D**).

D) Glycine dependence of pDisplay-GlyFS fluorescence. Expression in cultured neurons (n = 11 independent experiments from individual coverslips, from 2 separate cultures and transfections). The K_D of 13.3 μ M and maximum response (dynamic range) of 4.6 % were obtained as described in **B**. Experiments were performed in 50 μ M D-APV, 1 μ M strychnine, 1 μ M Org25543 and 5 μ M NFPS. In addition, the sodium channel inhibitor TTX (1 μ M) was added to the extracellular solution in some experiments, which had no detectable effect on the results.

The maximum response of expressed pDisplay-GlyFS is reduced compared to the maximum sensor response of purified GlyFS. The most parsimonious explanation is that diffraction-limited fluorescence microscopy of cell surfaces cannot distinguish between GlyFS facing extracellular space and intracellular GlyFS exposed to high intracellular glycine. (GlyFS can be safely assumed to be saturated with glycine in the presence of the millimolar glycine concentrations present intracellularly ³). The latter fraction of GlyFS does not respond to extracellular glycine concentration changes but contributes to the fluorescence readout thus reducing the apparent dynamic range (also see main text and biotin-mediated anchoring of GlyFS in the extracellular space).

Supplementary figure 11

Full DNA sequence of the glycine sensor GlyFS

Domains are color-coded: biotin tag (for further details please ¹³), ECFP, binding core, rigid linker, VenusFP.

DNA

```
ATGCGGGGTTCTCATCATCATCATCATGGTATGGCTAGCATGACTGGTGGACAGCAAATGGGTGCGGATCTGTACGACGATGACGA
TAAGGATCCGAAACTGAAGGTAACAGTCAACGGCACTGCGTATGACGTTGACGTTGACGTCGACAAGTCACACGAAACCCGATGGGCA
CCATCCTGTTTCGGCGGAGGCACCGGCGGCGCGCCGGCACCGGCAGCAGGTGGCGCAGGCGCCGGTAAGGCCGGAGAGGGCGAGATTCCC
GCTCCGCTGGCCGGCACCGTCTCCAAGATCCTCGTGAAGGAGGTGACACGGTCAAGGCTGGTCAGACCGTGCTCGTTCTCGAGGCCAT
GAAGATGGAGACCGAGATCAACGCTCCACCGACGGCAAGGTCGAGAAGGTCCTGGTCAAGGAGCGTGACGCGGTGCAGGGCGGTTCAGG
GTCTCATCAAGATCGGGGATCTCGAGCTCATCGAAGGCTCGAGCGGTTTCGGATCCGGGCCGCATGGTGAGCAAGGGCGAGGAGCTGTTT
ACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCAC
CTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGGCCTGGCCACCCTCGTGACCACCCTGACCTGGGGCG
TGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCAGACTTCTTCAAGTCCGCCATGCCGAAGGCTACGTCCAGGAGCGCACC
ATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGG
CATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACATACATCAGCCACAACGTCTATATACCGCCGACAGC
AGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGACGCTCGCCGACCCTACCAGCAGAACACC
CCCATCGGCGACGGCCCCGTGCTGCTGCCCCACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGA
TCACATGGTCTGCTGGAGTTCGTGACCGCCGCCGGGATCGATGTTGTTATTGACGTTGGTGCACCGCTGACCGGTCCGAATGCAGCAT
TTGGTGCACAGATTGAGAAAGGTGCAGAACAGGCAGCAAAAGATATTAATGCAGCCGGTGGTATTAATGGCGAGCAGATTAAATCGTT
CTGGGTGATGATGTTAGCGATCCGAAACAGGGTATTAGCGTTGCCAATAAATTCGTTGCAGATGGCGTTAAATTTGTGGTGGGTGATGC
GAACAGCGGTGTTAGCATTCCGGCAAGCGAAGTTTATGCAGAAATGGTATTCTCGAGATTACACCGTATGCAACCAATCCGGTTTTTA
CCGAACGTGGTCTGTGGAATACCTTTCTGACCTGCGGCCGCGACGATCAGCAGGGTGGTATTGCAGGTAAATATCTGGCAGATCATTTC
AAAGATGCCAAAGTGGCCATCATCCATGATAAAACCCCGTATGGTCAGGGTCTGGCCGATGAAACCAAAAAAGCAGCAAAATGCAGCGGG
TGTTACCGAAGTTATGTATGAAGGTGTTAATGTGGGCGATAAAGATTTTAGCGCACTGATCAGCAAAATGAAAGAAGCAGGCGTTAGCA
TTATCTATTGGGGTGGTTGGCATAACGAAGCAGGTCTGATTATTCTGTCAGGCAGCAGATCAGGGCCTGAAAGCAAAACTGGTTAGCGGT
GATGGTATTGTTAGCAATGAACTGGCAAGCATTGCCGGTGATGCAGTTGAAGGCACCCTGAATACATTTGGTCTGATCCGACCCTGCG
TCCGGAATAAAGAACTGGTTGAAAAATTCAAAGCCGAGGCTTAAATCCGGAAGCATATACCTGTATAGCTATGCAGCAATGCAGG
CAATTGCGGGTGACGCCAAAGCAGCAGGTAGCGTTGAACCGGAAAAAGTTGCAGAAGCACTGAAAAAAGGTAGCTTCCGACCGCACTG
GGTGAAATCAGCTTTGATGAAAAAGGTGATCCTAAACTGCCTGGCTATGTGATGTATGAATGGAAAAAAGGACCGGATGGCAAATTCAC
CTATATTGAGCAGCAAGCAGCAGCAAAAGAAGCCGCTGCCAAAGAAGCGGCAGCGAAAATGAGCAAGGGCGAGGAGCTGTTACCGGGG
TGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGC
AAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGGCCTGGCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTG
CTTCGCCCCTACCCCGACCACATGAAGCAGCAGACTTCTTCAAGTCCGCCATGCCGAAGGCTACGTCCAGGAGCGCACCATCTTCT
TCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGAC
TTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACATACAACAGCCACAACGTCTATATACCGCCGACAAGCAGAAGAA
CGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGCGGTGCAGCTCGCCGACCCTACCAGCAGAACACCCCATCG
GCGACGGCCCCGTGCTGCTGCCCCGACAACCACTACCTGAGCTACCAAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATG
GTCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAA
```

Supplementary figure 12

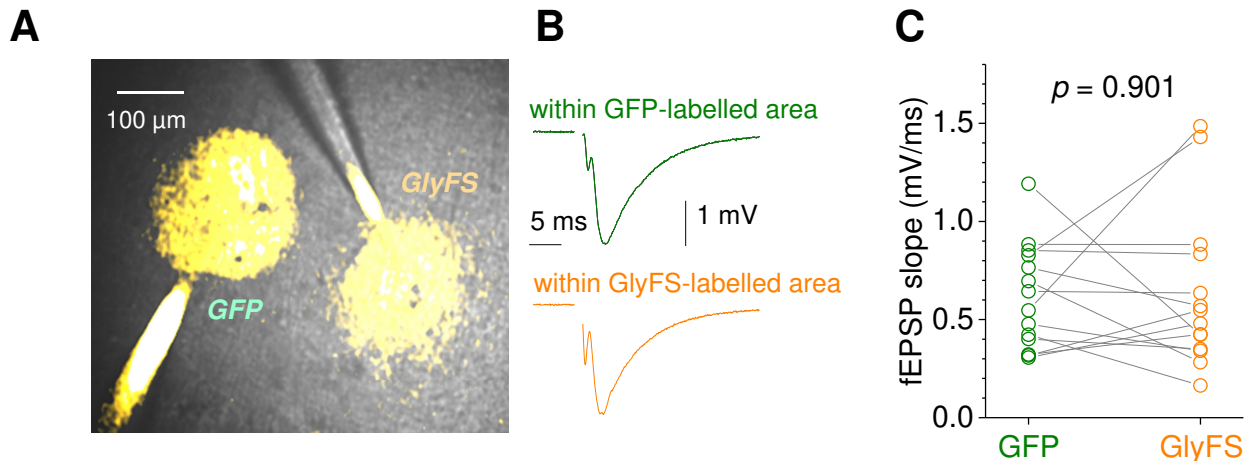
Full protein sequence of the glycine sensor GlyFS

Domains are color-coded: biotin tag (for further details please see ¹³), ECFP, binding core, rigid linker, VenusFP.

Protein

```
MRGSHHHHHHGMASMTGGQQMGRDLYDDDDKDPKLVTVNGTAYDVDVDVKSHENPMGTILFGGGTGGAPAPAAGGAGAGKAGEGEIP
APLAGTVSKIILVKEGDTVKAGQTVLVLEAMKMETEINAPTDGKVEKVLVKERDAVQGGQGLIKIGDLELIEGSSGSDPGRMVSKGEELF
TGVVPIILVELDGDVNGHKFSVSGEGEGDATYGKLTCLKFICTTGKLPVPWPTLVTTLTWGVQCFSRYPDHMKQHDFFKSAMPEGYVQERT
IFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLENYIISHNVYITADKQKNGIKANFKIRHNIEDGSQLADHYQQNT
PIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMLLEFVTAAGIDVVIIVGAPLTGPNAAFQAQIQKGAEQAAKDINAAGGINGEQIKIV
LGDDVSDPKQGISVANKFVADGVKFVVGHANSGVSIPASEVYAENGILEITPYATNPVFTERGLWNTFRTTCGRDDQQGGIAGKYLADHF
KDAKVAIIHDKTPYQGLADETKKAANAAGVTEVMYEGVNVGDKDFSALISKMKKEAGVSIIVWGGWHTAGLIIRQAADQGLKAKLVSG
DGIVSNELASIAGDAVEGTNLTFGPDPTLRPENKELVEKFKAAGFNPEAYTLYSYAAMQAIAGAAKAAGSVEPEKVAEALKKGSFPTAL
GEISFDEKGDPKLPGYVMYEWKKGPDGKFTYIQQEAHAKAAAKEAAAKVSKGEELFTGVVPIILVELDGDVNGHKFSVSGEGEGDATYG
KLTCLKICTTGKLPVPWPTLVTTLGYGQLCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGID
FKEDGNILGHKLEYNYNHNVYITADKQKNGIKANFKIRHNIEDGGVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHM
VLEFVTAAGITLGMDELYK
```

Supplementary figure 13



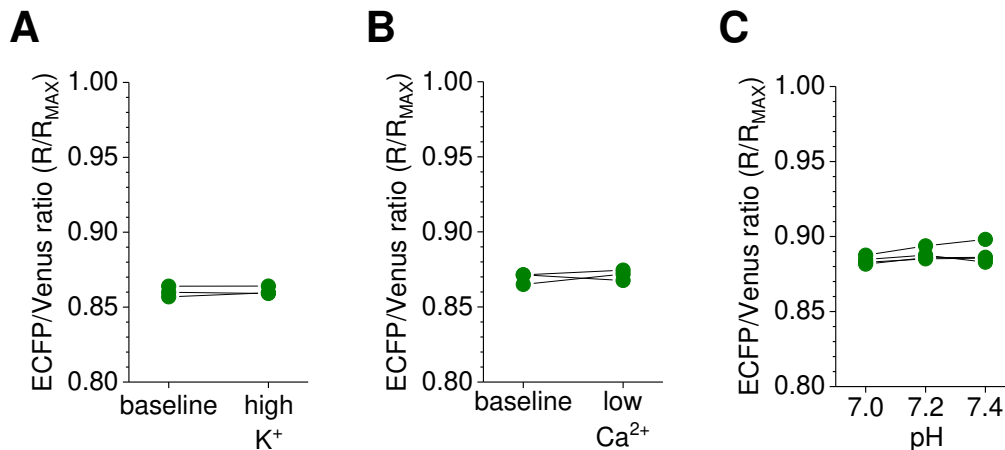
GlyFS has no effect on fEPSPs evoked by stimulation of CA3-CA1 Schaffer collateral connections.

A) We have previously established that tissue biotinylation does not affect synaptic transmission at CA3-CA1 synapses¹³, the model system used in this study. To test if glycine binding by GlyFS affects excitatory fEPSPs we compared synaptic transmission in GlyFS-labelled and GFP-labelled areas in the same slice. To this end, GlyFS-labelling (right pipette, as described, Fig. 3) was combined with anchoring GFP-labeled streptavidin nearby (biotinylated GFP, left pipette). Note that both GlyFS and GFP were excited by 2PE ($\lambda = 800$ nm) and fluorescence was collected in a single focal plane (giving differently-sized fluorescence profiles of dye inside the pipettes positioned at slightly different depths). Fluorescence of both is superimposed in yellow on the DIC image of the CA1 *stratum radiatum*. Stimulation pipette not visualized. Positions of GlyFS and GFP labelled areas relative to the stimulation electrode (near/far) were alternated between experiments. Representative example for experiments shown in **C**.

B) Representative sample fEPSP traces recorded through the pipettes used for injecting the GFP and GlyFS in **A** (see **C** for summary of full data set). Stimulus artifacts omitted for clarity. Stimulation pipette (not visible in **A**) was to the 'right' of the GlyFS recording electrode in this example.

C) No differences between fEPSP slopes in GFP and GlyFS labeled areas were detected (paired two-sided Student's t-test, $t(13) = -0.13$, $p = 0.90$, $n = 14$ independent experiments). Similarly, no differences between fiber volley amplitudes, a parameter related to the number of activated axons, were observed in these experiments (paired two-sided Student's t-test, $t(13) = 1.23$, $p = 0.24$, $n = 14$ independent experiments).

Supplementary figure 14



Changes of extracellular K⁺, Ca²⁺ and pH typical for high-frequency stimulation of CA3-CA1 synapses do not underlie observed GlyFS ratio changes (Fig. 5). Data are presented as mean \pm SEM below.

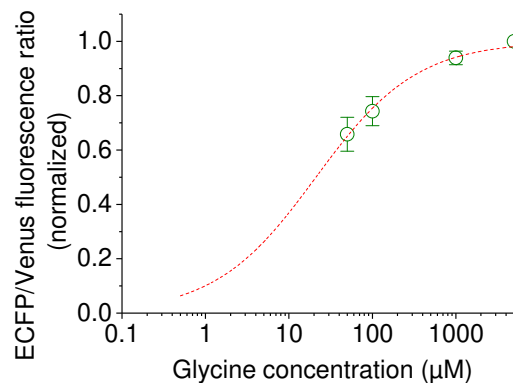
A) Extracellular potassium changes are closely associated with many types of neuronal activity. Therefore, we tested if GlyFS-reported glycine levels are affected by a stimulation-induced increase of extracellular [K⁺]_o (Fig. 5). The GlyFS fluorescence ratio was determined in baseline ambient K⁺ (PBS, 2.7 mM), after increasing [K⁺]_o by 10 mM to 12.7 mM, which is similar to the potassium ‘ceiling’ concentration during neuronal activity ¹⁴, and finally in the presence of 5 mM glycine to saturate GlyFS and to obtain R_{MAX} (see Results and Fig. 3C for experimental paradigm). The GlyFS ratio (R/R_{MAX}) was not significantly affected by increasing [K⁺]_o ($+0.077 \pm 0.22$ %, paired two-sided Student’s t-test, $t(2) = -0.60$, $p = 0.61$, $n = 3$ independent experiments).

B) High-frequency stimulation (HFS, Fig. 5) of CA3-CA1 synapses can reduce the extracellular [Ca²⁺]_o. The reduction of [Ca²⁺]_o during HFS (Fig. 5, 100 stimuli at 100 Hz) has been estimated based on the decrease of [Ca²⁺]_o induced by five stimuli at 100 Hz of ~ 40 -50 μ M, which was mainly mediated by NMDARs ¹⁵. During HFS synaptic transmission is becoming strongly depressed ¹⁶ and neurons cease to fire action potentials after ~ 16 stimuli ¹⁷, which will severely decrease any further NMDAR-dependent reduction of [Ca²⁺]_o. To test if a HFS-induced reduction of [Ca²⁺]_o affects GlyFS measurements, we reduced ambient Ca²⁺ from the 2.0 mM used in our extracellular solution by 0.2 mM (50 μ M / 5 stimuli \times 16 stimuli = 0.16 mM \approx 0.2 mM) in the presence of 1.3 mM Mg²⁺ (as in the acute slice experiments). No significant effect of lowering [Ca²⁺]_o was found ($+0.24 \pm 0.63$ %, $n = 3$ independent experiments, paired two-sided Student’s t-test, $t(2) = -0.67$, $p = 0.57$).

C) Repeated stimulation of CA3-CA1 Schaffer collaterals can lead to an acidification and a subsequent longer lasting alkalization of up to Δ pH ± 0.1 and the resting pH inside the slice can be lower than set by the extracellular perfusion solution ¹⁸. For these reasons, we tested if GlyFS is sensitive to pH changes from 7.0 to 7.2 to 7.4, a range twice as large as documented changes. No

statistically significant change was observed (one-way repeated-measures ANOVA, $F(2,6) = 2.97$, $p = 0.13$, $n = 4$ independent experiments, average change of R/R_{MAX} from pH 7.0 to 7.4 $+0.0040 \pm 0.0025$ or $+0.45 \pm 0.28 \%$).

Supplementary Figure 15



Calibration of GlyFS in acute slices.

In acute slices, the exact extracellular glycine concentration is not easily controlled experimentally and GlyFS measurements in the presence of zero glycine cannot be obtained because endogenous mechanisms such as glycine transporters maintain extracellular glycine levels (Fig. 3, Discussion). Therefore, the ECFP/Venus fluorescence intensity ratio of GlyFS immobilized in acute slices was recorded in the presence of the glycine transporter blockers NFPS (5 μM) and Org25543 (1 μM), the glycine receptor inhibitor strychnine (1 μM) and the NMDAR inhibitor D-APV (50 μM) to inhibit actions of applied glycine by activation/modulation of these receptors. Increasing concentrations of glycine starting at 50 μM were then applied. The rationale was that at these concentrations exogenously applied glycine will overwhelm any remaining endogenous mechanisms. The GlyFS fluorescence intensity ratio (R) was normalized by assuming that R_{SAT} (ECFP/Venus intensity ratio with GlyFS fully saturated with glycine) is reached at 5 mM exogenous glycine and R_0 (ECFP/Venus intensity ratio with no glycine bound to GlyFS) is given by $R_{\text{SAT}} / (1 + f)$, where f is dynamic range of GlyFS ($f = 0.185$, 18.5%, Fig. 2). R is then normalized by calculating $(R - R_0) / (R_{\text{SAT}} - R_0)$. Using this approach ($n = 4$ independent experiments), we obtained a K_D of GlyFS for glycine of ~ 21.1 μM ($R^2 = 0.98$, red dotted line represents fit), which is close to the K_D determined in free medium (Fig. 2). Data presented as mean \pm SEM.

References

1. Planamente, S. *et al.* A conserved mechanism of GABA binding and antagonism is revealed by structure-function analysis of the periplasmic binding protein Atu2422 in *Agrobacterium tumefaciens*. *J. Biol. Chem.* **285**, 30294–30303 (2010).
2. Fritz, R. D. *et al.* A Versatile Toolkit to Produce Sensitive FRET Biosensors to Visualize Signaling in Time and Space. *Sci. Signal.* **6**, rs12–rs12 (2013).
3. Supplisson, S. & Roux, M. J. Why glycine transporters have different stoichiometries. *FEBS Lett.* **529**, 93–101 (2002).
4. Beato, M. The Time Course of Transmitter at Glycinergic Synapses onto Motoneurons. *J. Neurosci.* **28**, 7412–7425 (2008).
5. Van Durme, J. *et al.* A graphical interface for the FoldX forcefield. *Bioinformatics* **27**, 1711–1712 (2011).
6. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
7. Deuschle, K. *et al.* Construction and optimization of a family of genetically encoded metabolite sensors by semirational protein engineering. *Protein Sci.* **14**, 2304–2314 (2005).
8. Magnusson, U., Salopek-Sondi, B., Luck, L. A. & Mowbray, S. L. X-ray Structures of the Leucine-binding Protein Illustrate Conformational Changes and the Basis of Ligand Specificity. *J. Biol. Chem.* **279**, 8747–8752 (2004).
9. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10**, 845–858 (2015).
10. Kleffner, R. *et al.* Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. *Bioinformatics* **33**, 2765–2767 (2017).
11. Marvin, J. S. *et al.* An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nat. Methods* **10**, 162–170 (2013).
12. Okumoto, S. *et al.* Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8740–8745 (2005).
13. Whitfield, J. H. *et al.* Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* **24**, 1412–1422 (2015).
14. Heinemann, U. & Dieter Lux, H. Ceiling of stimulus induced rises in extracellular potassium concentration in the cerebral cortex of cat. *Brain Res.* **120**, 231–249 (1977).
15. Rusakov, D. A. & Fine, A. Extracellular Ca²⁺ depletion contributes to fast activity-dependent modulation of synaptic transmission in the brain. *Neuron* **37**, 287–297 (2003).
16. Kim, E., Owen, B., Holmes, W. R. & Grover, L. M. Decreased afferent excitability contributes to synaptic depression during high-frequency stimulation in hippocampal area CA1. *J. Neurophysiol.* **108**, 1965–1976 (2012).

17. Grover, L. M., Kim, E., Cooke, J. D. & Holmes, W. R. LTP in hippocampal area CA1 is induced by burst stimulation over a broad frequency range centered around delta. *Learn. Mem.* **16**, 69–81 (2009).
18. Chen, J. C. & Chesler, M. Modulation of extracellular pH by glutamate and GABA in rat hippocampal slices. *J. Neurophysiol.* **67**, 29–36 (1992).

Chapter 5

A computationally designed fluorescent biosensor for D-serine

5.1 Preface

The design of a FRET-based biosensor is a complex task. Not only is a satisfactory dynamic range required, but also an appropriate binding affinity and specificity. For sensors based on a periplasmic binding protein scaffold, this means simultaneously optimising over multiple states of the binding protein. The closed state must tightly bind the analyte, but not other ligands, while the open state must be sufficiently different from the closed state to translate binding to a significant change in binding efficiency. However, both states are affected by changes made to the peptide sequence; improve the stability of the bound state, and one may inadvertently over-stabilise the closed state.

At least four states are relevant to the engineering of a PBP-based sensor:

1. The bound closed state
2. The bound open state
3. The unbound closed state
4. The unbound open state

When engineering for specificity, states bound to other ligands are also relevant. In general, for affinity, the bound states must be stabilised over the unbound states, and for dynamic range, the unbound open and bound closed states must both be stabilised relative to the other two. Therefore, the goal is to stabilise the bound closed state and the unbound open state, and destabilise the bound open state and the unbound closed state.

In order to produce high quality predictions of mutations that would improve sensor performance, all four states should be considered. Traditionally, sensors are designed based on crystal structures, which tend to be in closed states. When absent, the ligand is often added in

by docking. The tendency is therefore to privilege crystallised states over other states. This work demonstrates that it is possible to computationally derive the open state from a closed crystal structure simply and with sufficient accuracy to inform rational design using molecular dynamics with a modern force field. It therefore paves the way for future expansions on the theme of multi-state protein design.

5.2 Statement of contribution

I declare that the research presented in this chapter represents original work that I carried out during my candidature at the Australian National University, except for contributions to multi-author papers incorporated in the chapter where my contributions are specified in this Statement of Contribution.

5.2.1 Publication status

This manuscript has been released on the pre-print server BioRxiv with the title *A computationally designed fluorescent biosensor for D-serine*. The formatted pre-print with supporting information is reproduced in this chapter.

5.2.2 Authorship and contribution

The manuscript was authored by Vanessa Vongsouthi, Jason H. Whitfield, Petr Unichenko, Joshua A. Mitchell (the author), Björn Breithausen, Olga Khersonsky, Leon Kremers, Harald Janovjak, Hiromu Monai, Hajime Hirase, Sarel J. Fleishman, Christian Henneberger, and Colin J. Jackson. VV and JHW contributed equally to the work. I contributed guidance and advice on molecular dynamics simulations performed in the work.

A computationally designed fluorescent biosensor for D-serine

Vanessa Vongsouthi^{1†}, Jason H. Whitfield^{1†}, Petr Unichenko², Joshua A. Mitchell¹, Björn Breithausen², Olga Khersonsky³, Leon Kremers², Harald Janovjak⁴, Hiromu Monai⁵, Hajime Hirase⁵, Sarel J. Fleishman³, Christian Henneberger^{2,6,7}, Colin J. Jackson^{1*}

¹Research School of Chemistry, Australian National University, Canberra, Australia

²Institute of Cellular Neurosciences, Medical Faculty, University of Bonn, Bonn, Germany

³Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel

⁴Australian Regenerative Medicine Institute (ARMI), Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton/Melbourne, Australia

⁵Laboratory for Neuron-Glia Circuitry, RIKEN Center for Brain Science, Wako, Japan

⁶Institute of Neurology, University College London, London, United Kingdom

⁷German Center for Degenerative Diseases (DZNE), Bonn, Germany

[†]These authors contributed equally.

^{*}Corresponding authors.

Abstract

Periplasmic solute-binding proteins (SBPs) have evolved to balance the demands of ligand affinity, thermostability and conformational change to carry out diverse functions in small molecule transport, sensing and chemotaxis. Although the ligand-induced conformational changes that occur in SBPs make them useful components in biosensors, their complexity can be difficult to emulate and they are challenging targets for protein engineering and design. Here we have engineered a fluorescent biosensor with specificity for the signalling molecule D-serine (D-SerFS) from a D-alanine-specific SBP. Through a combination of binding site and remote mutations, the affinity, specificity and thermostability were optimized to allow the detection of changes in D-serine levels using two-photon excitation fluorescence microscopy *in situ* and *in vivo*. This work illustrates the multidimensional constraints that are imposed by the trade-offs between structural dynamics, ligand affinity and thermostability, and how these must be balanced to achieve desirable activities in the engineering of complex, dynamic proteins.

Introduction

D-amino acids have been discovered in a range of biological organisms (Corrigan, 1969). Free D-serine, in particular, was first found in mammalian brain tissue in the early 1990s using gas chromatography-mass spectrometry (GC-MS) (Hashimoto et al., 1992). Since then, there has been considerable interest in the physiological and pathophysiological role of this molecule. It is now well known that D-serine acts as a co-agonist of excitatory glutamate receptors of the N-methyl-D-

aspartate receptor (NMDAR) subtype. Together with the primary agonist, L-glutamate, D-serine binds to the 'glycine-binding site' on the NR1 binding domain of the NMDAR to activate it (Mothet et al., 2000; Panatier et al., 2006; Schell, Molliver, & Snyder, 1995). By the same mechanism, glycine also acts as a co-agonist of the NMDARs (Mothet et al., 2000; Panatier et al., 2006; Schell et al., 1995). Previous work has demonstrated the preferential gating of synaptic NMDARs by D-serine (Henneberger, Papouin, Oliet, & Rusakov, 2010; T. Papouin et al., 2012), and of extrasynaptic NMDARs by glycine (T. Papouin et al., 2012). Notably, it is synaptic NMDARs that are primarily responsible for inducing long-term potentiation (LTP) of excitatory synaptic transmission, a key mechanism of learning and memory (T. V. Bliss & Collingridge, 1993; T. V. P. Bliss & Cooke, 2011; Hardingham & Bading, 2010).

D-serine has been associated with a number of conditions centred on cognitive impairment and disturbances to NMDAR activity. D-serine levels in the brain are largely regulated by the activity of serine racemase (SR), which racemizes L-serine to produce D-serine, and D-amino acid oxidase (DAAO), which catabolises it (Pollegioni & Sacchi, 2010; H. Wolosker, Blackshaw, & Snyder, 1999). Abnormal levels of both SR and DAAO have been associated with schizophrenia and Alzheimer's disease (Basu et al., 2009; Labrie et al., 2009; Madeira et al., 2015). Recent studies in flies and mammals have also implicated D-serine in sleep regulation (Dai et al., 2019; Liu, Liu, Tabuchi, & Wu, 2016; Thomas Papouin, Dunphy, Tolman, Dineley, & Haydon, 2017; Tomita, Ueno, Mitsuyoshi, Kume, & Kume, 2015) and kidney disease (Wiriyaermkul et al., 2020). Despite the significance of D-serine in critical physiological and pathophysiological processes, aspects of the D-serine signalling pathway remain elusive. While it is thought that D-serine is a gliotransmitter released from astrocytes (Henneberger et al., 2010; Thomas Papouin, Henneberger, Rusakov, & Oliet, 2017; Yang et al., 2003), its cellular origin continues to be an intensely debated topic (Thomas Papouin, Henneberger, et al., 2017; Herman Wolosker, Balu, & Coyle, 2016). Progress towards understanding the D-serine pathway has been partly hindered by the lack of a method to study the transmitter dynamically with high spatial and temporal resolution. While methods such as microdialysis can provide precise quantification of D-serine from tissues (Shippenberg & Thompson, 2001), they tend to suffer from poor temporal resolution and can cause mechanical damage to the tissue under study (Beyene, Yang, & Landry, 2019; Ganesana, Lee, Wang, & Venton, 2017). Amperometric probes for D-serine based on DAAO have also been developed by several groups (Mohd Zain, Ab Ghani, & O'Neill, 2012; Pernot et al., 2008); however, these probes are subject to losses in sensitivity due to fouling of the outer biolayer and loss of enzymatic activity during operation (Dale, Hatz, Tian, & Llaudet, 2005), and also have a low spatial resolution.

Optical biosensors are an alternative to existing D-serine detection methods that have the potential to provide unparalleled spatial and temporal resolution when used with high resolution microscopy. Those based on Förster resonance energy transfer (FRET) are commonly a fusion of a ligand-binding domain to a pair of donor-acceptor fluorophores. Here, ligand-induced conformational changes in the binding domain lead to the displacement of the fluorophores and a measurable

change in FRET efficiency (Kaczmarek, Mitchell, Spence, Vongsouthi, & Jackson, 2019). Several FRET-based biosensors of this architecture have been engineered for the dynamic study of key amino acids in brain tissue, including L-glutamate (Okumoto et al., 2005), glycine (Zhang et al., 2018), and L-arginine (Whitfield et al., 2015). The development of such a sensor for D-serine has previously been limited, because no appropriate D-serine-specific solute-binding protein (SBP) is known.

The limitations of naturally occurring SBPs on biosensor design can be overcome by using protein engineering to modify existing SBPs towards desirable binding affinity and specificity. In the case of promiscuous SBPs, however, engineering for both strength and preference of a chosen on-target interaction can be challenging. While improvements in affinity alone can be driven by non-specific hydrophobic interactions, modifying specificity generally requires finer tuning of the charge and shape complementarity between the binding interface and the target ligand. This has been demonstrated in previous work where a SBP (Atu2422) that promiscuously bound glycine, L-serine and GABA, was engineered towards glycine-specificity through multiple rounds of computational design and experimental testing (Zhang et al., 2018). Also highlighted in this work was the challenge of achieving specificity in SBPs that promiscuously bind multiple, structurally similar ligands, requiring a combination of both positive and negative selection.

In this study, we used rational and computational protein design to engineer a D-serine FRET sensor (D-SerFS), using a D-alanine binding protein from *Salmonella enterica* (DalS) (Osborne et al., 2012) as a starting point. Iterative rounds of design and experimental testing produced a robust, D-serine-specific fluorescent sensor with high affinity for D-serine ($K_D = 7 - 9 \mu\text{M}$). *In-situ* testing of D-SerFS in acute hippocampal rat brain slices demonstrated a response (DR = 6.5%) to exogenously applied D-serine. Furthermore, we demonstrate that D-SerFS responds to extracellular changes of D-serine in the living animal.

Results

Homology-guided design of a D-serine-specific binding protein. A SBP from *Salmonella enterica*, DalS, that binds D-alanine and glycine (Osborne et al., 2012) was chosen as a starting point for the design of a D-serine-specific binding protein. The initial design was guided by a comparison to the NR1 binding domain of the NMDAR, a structural homologue of SBPs that contains the receptor's D-serine/glycine-binding site (Furukawa & Gouaux, 2003). A structural alignment of DalS (PDB 4DZ1) and NR1 (PDB 1PB8) (Fig. 1A) revealed key differences in the polarity and cavity-size of the two binding sites, particularly at the residues comprising the D-alanine methyl side chain pocket, F117, A147 and Y148 (vs. L146, S180 and V181 in NR1). Thus, these residues were targeted for mutagenesis towards D-serine specificity in DalS.

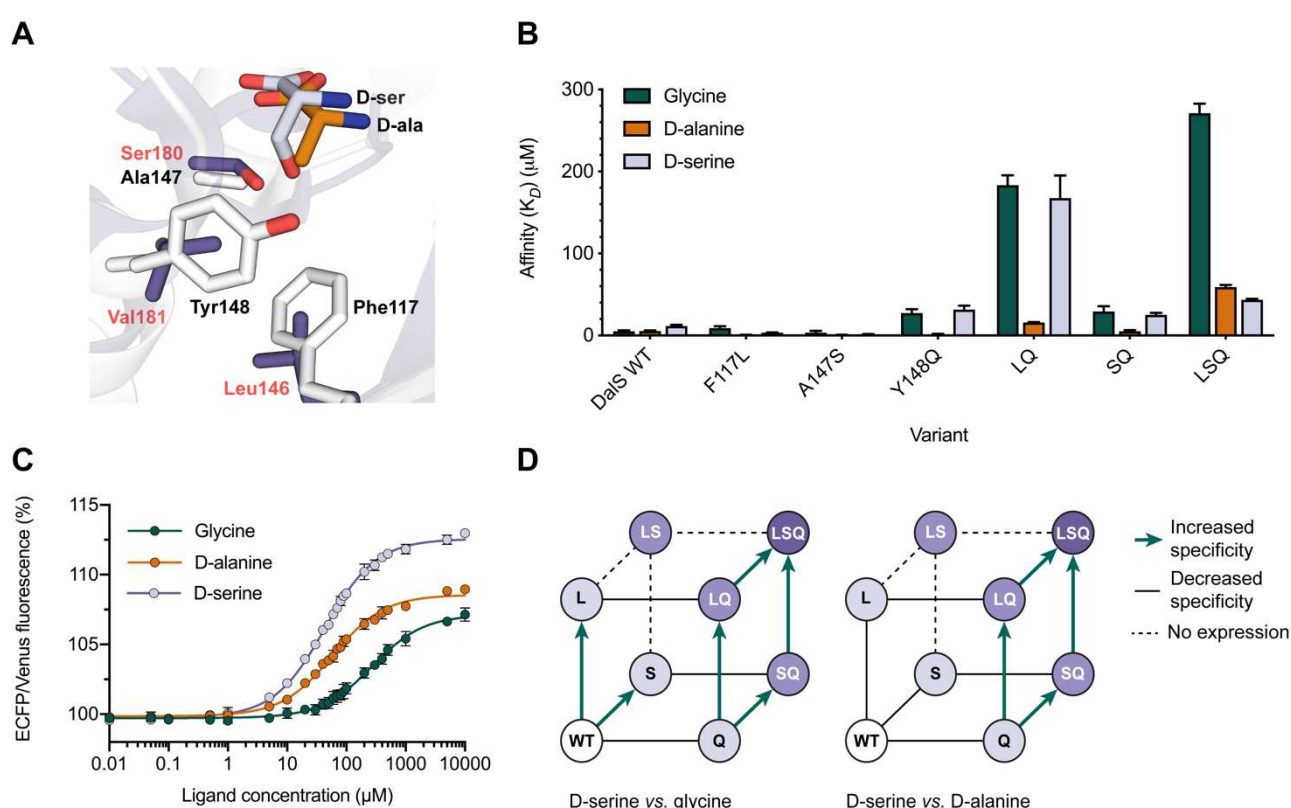


Figure 1. Homology-guided design of DalS. A) Structural alignment of binding site residues targeted for mutagenesis in DalS (white, black labels; PDB 4DZ1) to homologous residues in the NR1 binding domain (purple, red labels; PDB 1PB8). B) Binding affinities (μM) determined by fluorescence titration of wild-type DalS, and single, double and triple mutants, for glycine, D-alanine and D-serine ($n = 3$). C) Sigmoidal dose response curves for F117L/A147S/Y148Q (LSQ) with glycine, D-alanine and D-serine ($n = 3$). Values are the (475 nm/530 nm) fluorescence ratio as a percentage of the same ratio for the apo sensor. D) Schematic demonstrating the epistatic interactions between individual mutations contributing to D-serine specificity relative to glycine (left) and D-alanine (right). Increased D-serine specificity is represented by green arrows, whereas loss in specificity is represented by solid black lines. Dashed lines associated with the variant F117L/A147S (LS) represent the absence of experimental data for this variant due to no maturation of the Venus chromophore during expression.

Three mutations were initially considered: F117L, A147S and Y148Q. F117L was hypothesised to increase the size of the cavity and reduce van der Waals contacts with the methyl side chain of D-alanine, whilst A147S would increase the polarity of the binding site to form polar contacts with the hydroxyl side chain of D-serine. Y148 was mutated to glutamine rather than valine (as in the NR1 binding site) to increase both the polarity and size of the side chain pocket, reducing potentially favourable interactions with D-alanine and glycine.

To experimentally test the proposed design, DalS was first cloned between enhanced cyan fluorescent protein (ECFP) and Venus fluorescent protein (Venus); a commonly used FRET pair (Bajar, Wang, Zhang, Lin, & Chu, 2016). For immobilization, an N-terminal hexahistidine-tagged biotin domain was included in the construct. F117L, A147S and Y148Q were introduced to the wild-type FRET sensor in a combinatorial fashion. Soluble, full-length protein was obtained in high purity for all variants excluding the double mutant, F117L/A147S (LS). The LS mutant did not exhibit maturation of the Venus chromophore during expression, suggesting that the combination of the two mutations was sufficiently destabilising to the binding protein such that folding of the protein and the downstream Venus fluorescent protein were negatively affected.

Fluorescence titrations with D-alanine, glycine and D-serine were performed on the wild-type protein and the successfully purified variants to determine the binding affinities (K_D) (Fig. **1B**, Table **1**). The wild-type exhibited similar affinity to D-alanine and glycine ($\sim 5 \mu\text{M}$) and significant, but lower, affinity to D-serine ($12 \mu\text{M}$). Among the single mutants, F117L and A147S increased affinity to D-alanine and D-serine by similar amounts, with minimal effect on glycine affinity. The only mutation that had a large effect in isolation was Y148Q, which had the unintended effect of increasing affinity to D-alanine and decreasing affinity ~ 8 -fold to D-serine. In combination with Y148Q, F117L (LQ) decreased D-alanine affinity ~ 4.5 -fold from Y148Q, while only decreasing D-serine affinity 3-fold, however, A147S (SQ), decreased D-alanine affinity ~ 3 -fold, while increasing D-serine affinity ~ 3.5 -fold. The effects of the mutations in the double mutants on glycine vs. D-serine were largely identical. Finally, the combination of the three mutations in F117L/A147S/Y148Q (LSQ), generated a variant that was specific for D-serine, exhibiting a K_D for D-serine of $43 \pm 1 \mu\text{M}$, compared to $59 \pm 2 \mu\text{M}$ and $271 \pm 16 \mu\text{M}$ for D-alanine and glycine, respectively (Fig. **1C**). The 6.3-fold higher affinity for D-serine vs. glycine is the more important comparison because of the spatial and temporal overlap between these in the brain; D-alanine, in contrast, is present in concentrations near or below detection levels in brain tissue (Hashimoto et al., 1992; Popielek, Tierney, Steyn, & DeVivo, 2018). LSQ displayed a dynamic range of 13% in response to a saturating concentration of D-serine. The combinatorial analysis of the effects of these mutations is indicative of strong epistasis, i.e. the effects of the mutations are not additive and there is no smooth pathway of stepwise increases in D-serine specificity available by which we could arrive at the LSQ variant (Fig. **1D**). This strong epistasis highlighted the advantage of adopting a rational design approach to engineering the binding site, allowing for the introduction of all three mutations simultaneously, even though there were few encouraging signs in the affinities of the single and double mutants.

Table 1. Summary of the binding affinities for D-alanine, glycine and D-serine (μM), specificities and thermostabilities ($^{\circ}\text{C}$) for all variants of the wild-type sensor. D-serine specificity of the variants relative to glycine (α_{Gly}) and D-alanine ($\alpha_{\text{D-ala}}$) is defined as the ratio between the K_D for D-serine and that of the corresponding off-target ligand: $\frac{K_D \text{ for Glycine (or D-alanine)}}{K_D \text{ for D-serine}}$. Values are mean \pm s.e.m. throughout ($n = 3$).

Variant	Binding Affinity, K_D (μM)			Specificity		T_m ($^{\circ}\text{C}$)
	D-ala	Gly	D-ser	α_{Gly}	$\alpha_{\text{D-ala}}$	
WT	5.4 ± 0.4	5 ± 1	12 ± 1	0.42	0.45	68.0 ± 0.3
F117L	0.65 ± 0.03	9 ± 1	3.7 ± 0.2	2.40	0.18	63.1 ± 0.3
A147S	0.62 ± 0.03	4.0 ± 0.2	1.8 ± 0.4	2.22	0.34	66.9 ± 0.4
Y148Q	1.6 ± 0.3	75 ± 8	101 ± 3	0.74	0.02	62.0 ± 0.2
(LS) F117L/A147S [†]	-	-	-	-	-	-
(LQ) F117L/Y148Q	32 ± 2	340 ± 14	309 ± 19	1.10	0.10	60.3 ± 0.2
(SQ) A147S/Y148Q	5.4 ± 0.5	29 ± 2	28 ± 3	1.04	0.19	60.8 ± 0.2
(LSQ) F117L/A147S/Y148Q	59 ± 2	271 ± 16	43 ± 1	6.30	1.44	60.5 ± 0.2
(LSQ+E) LSQ + D216E	118 ± 3	429 ± 21	41 ± 1	10.46	2.88	62.2 ± 0.2
(LSQE+D) LSQE + A76D	130 ± 4	410 ± 100	48 ± 4	8.54	2.71	65.0 ± 0.1
(LSQED+K) LSQED + S60K	180 ± 1	540 ± 13	61.0 ± 0.2	8.85	2.95	66.3 ± 0.1
(LSQED+E) LSQED + S208E	120 ± 3	420 ± 93	40 ± 1	10.50	3.00	66.2 ± 0.1
(LSQED+K) LSQED + N200K	230 ± 63	1400 ± 1000	470 ± 270	2.98	0.49	69.4 ± 0.1
(LSQED+D) LSQED + T172D	-	-	-	-	-	71.0 ± 0.7
(LSQED+R) LSQED + S60R	202 ± 42	1200 ± 380	70 ± 17	17.14	2.89	79.0 ± 0.3
(LSQED+Y) LSQED + T197Y	13 ± 1	41 ± 17	7.0 ± 0.4	5.86	1.86	79.0 ± 0.3
(LSQED+RY) LSQED + S60R/T197Y	14 ± 1	49 ± 10	6.1 ± 0.5	8.03	2.30	*

[†] This variant did not yield mature fluorescent protein.

* Indicates that a sigmoidal transition in fluorescence ratio as a function of increasing temperature was not observed within the temperature range of the experiment ($20 - 85^{\circ}\text{C}$), suggesting a T_m higher than $\sim 85^{\circ}\text{C}$.

Increasing D-serine specificity of LSQ using ligand docking. In order to identify mutations to the binding site that would further increase the specificity of LSQ for D-serine, modelling of the LSQ binding protein with FoldX (Schymkowitz et al., 2005) and ligand docking with Glide (Friesner et al., 2004), were performed. As LSQ maintained similar affinities for D-alanine ($K_D = 59 \pm 2 \mu\text{M}$) and D-serine ($K_D = 43 \pm 1 \mu\text{M}$), both ligands were docked into the binding site of a model of LSQ and the top poses were analysed (Fig. 2A). In agreement with Osborne et al. (Osborne et al., 2012), the poses suggested that R102 and S97 stabilise the carboxylate group of all docked ligands. It appears that D216, N115, and Y173F, likely play roles in ligand binding and specificity given their proximity to the side chains. Thus, these residues were rationally mutated *in silico* for subsequent rounds of ligand docking.

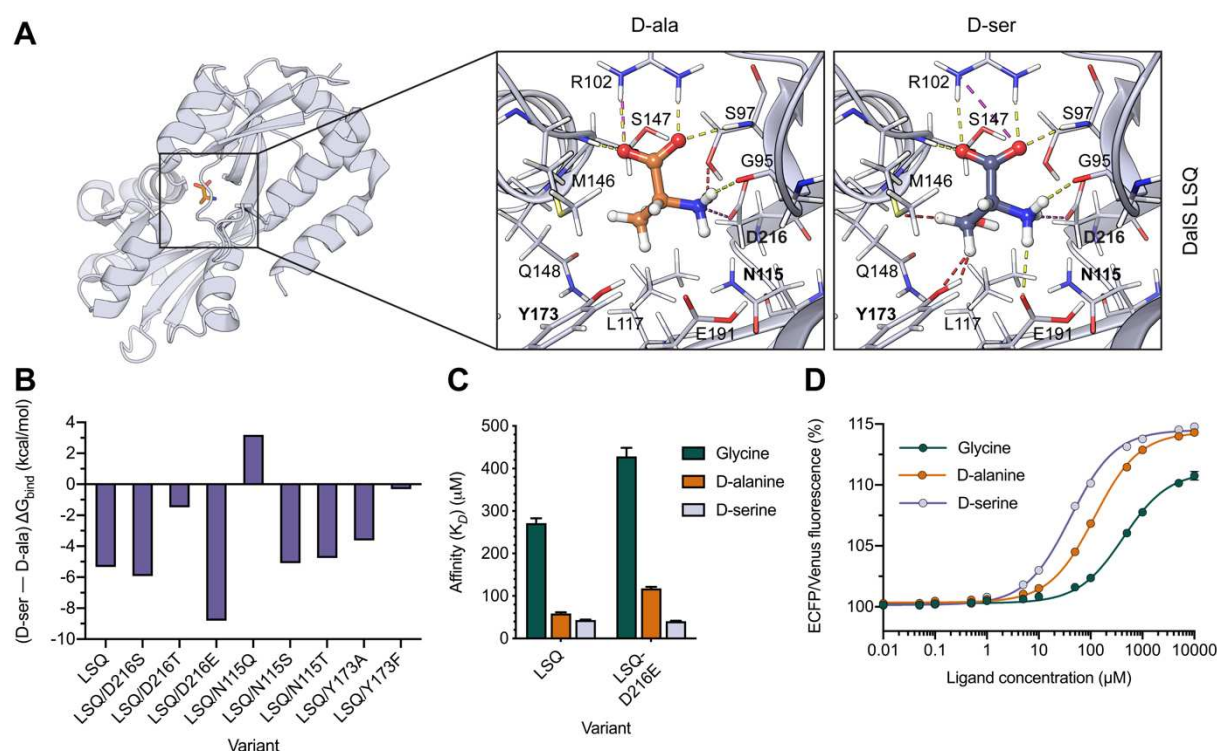


Figure 2. Ligand docking of the LSQ binding protein using Glide (Friesner et al., 2004). **A)** Top binding poses of D-alanine (orange, left) and D-serine (purple, right) in a FoldX-generated (Schymkowitz et al., 2005) model of DaLS with mutations F117L, A147S, and Y148Q (DaLS LSQ). Dashed lines represent non-covalent interactions between binding site residues and the ligands, including hydrogen bonds (yellow), salt-bridges (pink) and unfavourable clashes (red). Residues targeted for further rational design, N115, Y173 and D216, are bolded. **B)** Difference between ΔG_{bind} (kcal/mol) for D-serine and D-alanine, predicted using Prime MM-GBSA (Rapp, Kalyanaraman, Schiffriller, Schoenbrun, & Jacobson, 2011) for FoldX-generated models of DaLS LSQ and single mutants in the background of DaLS LSQ. **C)** Binding affinities (μM) determined by fluorescence titration of LSQ and LSQ/D216E (LSQE), for glycine, D-alanine and D-serine ($n = 3$). **D)** Sigmoidal dose response curves for LSQE with glycine, D-alanine and D-serine ($n = 3$). Values are the (475 nm/530 nm) fluorescence ratio as a percentage of the same ratio for the apo sensor.

We first mutated, *in silico*, each residue to the corresponding residue in the NR1 domain, giving rise to LSQ-D216S, -N115Q, and -Y173A. As the carboxylic acid functional group on D216 was predicted to form stabilising interactions with the amine group of both D-alanine and D-serine, the mutation of this residue to glutamate, D216E, was performed. The introduction of other polar residues at positions 115 and 216 was also explored, leading to the mutants LSQ-D216T, -N115T, and -N115S. Lastly, as the side chain of Y173 was predicted to clash with the D-serine ligand in LSQ, the mutation of this residue to phenylalanine, Y173F, was investigated. D-serine and D-alanine were docked into FoldX (Schymkowitz et al., 2005) models of the proposed mutants of LSQ and the generated poses were analysed by visual inspection. In order to quantify and compare the effects of the mutations on binding specificity, Prime MM-GBSA was used to predict the free energy of binding

(ΔG_{bind}) in kcal/mol for each ligand. Prime MM-GBSA can effectively predict the relative, rather than absolute, binding free energies for congeneric ligands (Lyne, Lamb, & Saeh, 2006; Rapp et al., 2011). Due to this, focus was placed on the difference between the ΔG_{bind} for D-serine and D-alanine for each mutant, where positive and negative values corresponded to preferential binding of D-alanine and D-serine, respectively (Fig. 2B). Using this approach, D216E appeared to be the most beneficial mutation for increasing D-serine specificity (ΔG_{bind} difference = -8.8 kcal/mol) relative to LSQ (ΔG_{bind} difference = -5.3 kcal/mol) and was selected for experimental characterisation. The remaining single mutations did not exhibit ΔG_{bind} differences that were indicative of increased D-serine specificity.

Experimental testing of LSQ/D216E (LSQE) revealed improved specificity towards D-serine. Fluorescence titrations showed that it decreased the affinity of the sensor for both D-alanine ($K_D = 118 \pm 3$ vs. 59 ± 2 μM) and glycine ($K_D = 429 \pm 21$ vs. 271 ± 16 μM), whilst the affinity for D-serine was slightly increased ($K_D = 41 \pm 1$ vs. 43 ± 1 μM) compared to LSQ (Fig. 2C – D). Furthermore, this improvement in specificity did not affect the dynamic range of the sensor. As a result, LSQE was taken forward for an additional round of engineering.

Table 2. Scores computed by Glide (Friesner et al., 2004) and Prime MM-GBSA (Rapp et al., 2011). Glide scores (kcal/mol) and relative binding free energies (ΔG_{bind}) predicted by Prime MM-GBSA for the top poses of D-alanine (D-ala) and D-serine (D-ser) are shown. The difference between the ΔG_{bind} of D-ser and D-ala represents the predicted specificity of each variant.

Variant	Glide Score (kcal/mol)		ΔG_{bind} (kcal/mol)		(D-ser – D-ala) ΔG_{bind} (kcal/mol)
	D-ala	D-ser	D-ala	D-ser	
WT	-6.6	-7.4	-27.9	-29.1	-1.2
LSQ	-6.0	-7.9	-22.5	-27.8	-5.3
D216S	-5.5	-5.0	-15.1	-21.0	-5.9
D216T	-6.0	-5.9	-13.3	-14.8	-1.5
D216E	-6.0	-6.3	-12.0	-20.8	-8.8
N115Q	-6.4	-8.1	-28.3	-25.1	3.2
N115S	-7.5	-8.7	-25.0	-30.1	-5.1
N115T	-7.3	-8.7	-23.9	-28.6	-4.8
Y173A	-6.1	-7.5	-20.1	-23.7	-3.6
Y173F	-6.9	-6.6	-12.8	-13.1	-0.3

Stabilising mutations in LSQE improve the affinity and specificity for D-serine. Preliminary testing of the LSQ variant under two-photon excitation (2PE) fluorescence microscopy showed a decrease in the dynamic range following the reconstitution of lyophilised sensor (SI. Fig. 1). As ECFP/Venus is a common FRET pair successfully used in previous sensors that had not suffered a loss in dynamic range following this process (Whitfield et al., 2015; Zhang et al., 2018), this effect was attributed to the inadequate stability of the binding protein. The thermostability of the sensor

variants was determined by measuring the fluorescence ratio (ECFP/Venus) as a function of increasing temperature (Fig. 3A). This analysis revealed that all of the variants were less thermostable than the wild-type by up to 8 °C, which we assume is very close to the point at which the protein can no longer fold, as was observed for the F117L/A147S variant. The destabilizing effects of these mutations are consistent with the observation that mutations to active/binding/core sites are more often than not destabilizing (Tokuriki, Stricher, Schymkowitz, Serrano, & Tawfik, 2007).

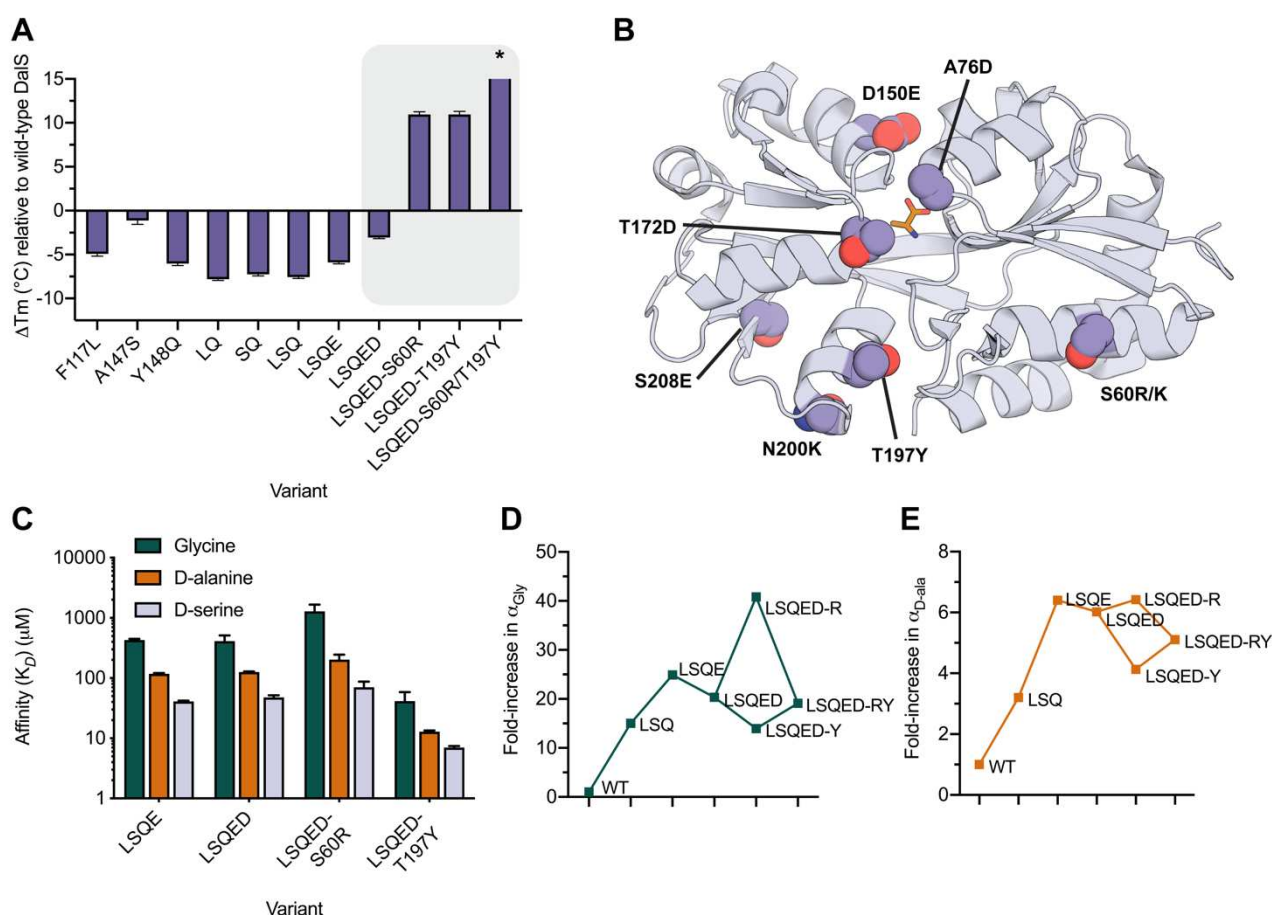


Figure 3. A) Change in melting temperature of DaIS variants relative to the wild-type (ΔT_m) (°C). Mutations to the binding site are destabilising. Stabilising mutations to LSQE predicted by FoldX (Schymkowitz et al., 2005) and PROSS (Goldenzweig et al., 2016) are highlighted (grey box). *The ΔT_m for LSQED-S60R/T197Y exists outside the axis limit as the unfolding of this variant could not be detected within the temperature range of the experiment (20 – 85 °C). B) Crystal structure of wild-type DaIS (PDB 4DZ1) with the positions of predicted stabilising mutations highlighted (spheres) and labelled. C) Binding affinities (μM) determined by fluorescence titration of LSQE, LSQE-A76D (LSQED), LSQED-S60R (LSQED-R), and LSQED-T197Y (LSQED-Y), for glycine, D-alanine and D-serine. D – E) Fold-increase in D-serine specificity compared to the wild-type for key variants in the design of D-SerFS, relative to glycine (D) and D-alanine (E). Specificity (α) is defined as: $\frac{K_d \text{ for Glycine (or D-alanine)}}{K_d \text{ for D-serine}}$.

In order to identify potentially stabilising mutations in the binding protein, two computational methods were used: the structure-based FoldX (Schymkowitz et al., 2005) and the sequence and structure-based PROSS (Goldenzweig et al., 2016). Residues on the surface (S60, A76, D150, T172, N200,

S208; Fig. **3B**) were initially targeted for mutation by FoldX as it is known that mutations to the core of a protein are, on average, more destabilizing (Tokuriki et al., 2007). The mutations to each of the other amino acids at each position were ranked by the predicted change in Gibbs free energy of folding ($\Delta\Delta G$), and the most stabilising mutation at each position that introduced a polar or charged residue to the surface was identified. This produced the following set of mutations: A76D, S208E, S60K, T172D, S60R, N200K, and D150E (Fig. **3B**, **Table 1**). In parallel, the PROSS webserver developed by Goldenzweig et al. (Goldenzweig et al., 2016) (available at: <https://pross.weizmann.ac.il>) was also used to identify stabilising mutations. The PROSS algorithm combines phylogenetic analysis and Rosetta-based design to provide users with combinatorial designs predicted to have improved stability relative to the input sequence and structure. A PROSS calculation was performed on the DalS crystal structure and residues with known importance to ligand binding were not allowed to design. This produced the following set of mutations: N32R, S60K, S61A, Q64K, S65E, H67G, L70C, T73V, A76S, G82A, Q88K, E110D, I114T, A123K, H125D, N131D, N133S, N136K, T176A, T176V, T197Y, K199L, L234D, Q242E, S245K, S245Q, G246A, G246E. The individual mutations identified by PROSS were also evaluated with FoldX. All the identified stabilising mutations were then ranked by the $\Delta\Delta G$ values calculated by FoldX (**SI. Table 1**).

Initially, the A76D mutation, which was predicted by FoldX to be the most stabilizing, was selected for experimental testing in the background of the LSQE variant. LSQE/A76D (LSQED) displayed a +3 °C improvement in thermostability ($T_m = 65.0 \pm 0.1$ °C) compared to LSQE and resulted in little change to the binding affinities (Fig. **3C**, **Table 1**). Given this improvement, all mutations with a FoldX $\Delta\Delta G$ that was more negative than 2.5 FoldX standard deviations (< -1.15 kcal/mol) were individually introduced to the background of LSQED. Two variants, LSQED/S208E and LSQED/S60K, displayed no significant improvements (< 1 °C) in thermostability (**Table 1**). The variants, LSQED/T172D and LSQED/N200K exhibited moderate improvements in thermostability ($T_m = 71.0 \pm 0.7$ °C and 69.4 ± 0.1 °C, respectively), while the variants LSQED/S60R (LSQED-R) ($T_m = 79.0 \pm 0.3$ °C) and LSQED/T197Y (LSQED-Y) (79 ± 0.3 °C) resulted in the greatest improvements in thermostability (Fig. **3A**, **Table 1**).

Notably, several of the mutations that were identified to increase thermostability were found to have pronounced effects on substrate affinity and specificity. For example, N200K resulted in a ~10-fold reduction in affinity for D-serine (**Table 1**). In contrast, S60R significantly decreased the affinity for glycine ($K_D = 1200 \pm 380$ μ M; Fig. **3C**). In a comparison of the fold-change in specificity of key variants relative to the wild-type (Fig. **3D – E**), S60R was found to be the most specific variant, exhibiting a 40-fold and 6-fold greater specificity for D-serine, relative to glycine and D-alanine, respectively, albeit with slightly reduced affinity for D-serine ($K_D = 70 \pm 17$ μ M; Fig. **3C**). The T197Y mutation was also unexpectedly beneficial for binding and resulted in a marked increase in affinity for D-serine ($K_D = 7.0 \pm 0.4$ μ M; Fig. **3C**), as well as for the native ligands D-alanine ($K_D = 13 \pm 1$ μ M) and glycine ($K_D = 41 \pm 17$ μ M). The T197Y mutant, LSQED-Y, maintained D-serine specificity,

displaying a 14-fold and 4-fold greater specificity for D-serine, relative to glycine and D-alanine, respectively, compared to the wild-type (Fig. **3D – E**). Interestingly, when the mutations S60R and T197Y were included together in the same variant (LSQED-RY), the affinities for D-serine ($K_D = 6.1 \pm 0.5 \mu\text{M}$), D-alanine ($K_D = 14 \pm 1 \mu\text{M}$) and glycine ($K_D = 49 \pm 10 \mu\text{M}$) closely resembled that observed for the T197Y mutation alone (**Table 1**). As no sigmoidal transition in fluorescence ratio as a function of increasing temperature was observed for this variant, this suggested that the thermostability is similar to, or exceeds, the upper limit of the temperature range of the experiment ($\sim 85^\circ\text{C}$; Fig **3A**).

The molecular basis for the effects of remote mutations on ligand affinity. The large effect of the T197Y mutation on improving the binding affinities to the three ligands despite being $\sim 13 \text{ \AA}$ from the binding site was unexpected, prompting further analysis of the molecular basis for this effect. There is no apparent structural explanation for why this remote mutation would change affinity based on the crystal structure of DalS, i.e. there is no mechanism by which it could change the shape of the ligand binding site. Because the crystal structure of DalS is in the ligand bound (closed) conformation, and it is known this protein family fluctuates between open and closed states, we investigated the open-closed conformational transition of the protein scaffold to examine whether the mutation was altering affinity by differentially stabilizing the ligand-bound (closed) state over the unbound (open) state. To do this, we first needed a model of the unbound, open state. Previous work on SBPs closely related to DalS have shown that molecular dynamics (MD) simulations of closed SBP crystal structures in the apo state can reasonably sample open conformational substates (Clifton et al., 2018; Kaczmariski et al., 2020). To simulate the closed to open conformational transition of DalS (PDB 4DZ1), MD simulations (100 ns x 10 replicates) were run with the ligand removed, and clustering analysis was performed to obtain a representative open conformation from the largest cluster. The $C\alpha$ – $C\alpha$ distance between residues A85 and K153 (either side of the binding site cleft), and radius of gyration, were calculated for all frames of the simulation (Fig. **4A**). The 85-153 $C\alpha$ – $C\alpha$ distance in the open conformation was 0.9 nm greater than that observed in the closed crystal structure (3.0 vs. 2.1 nm; Fig **4B**). In the X-ray crystal structures of a closely related SBP, AncCDT-1, the difference in $C\alpha$ – $C\alpha$ distance (at equivalent positions) between the open (PDB 5TUJ) and closed (PDB 5T0W) crystallographic states is similar (0.8 nm) (Clifton et al., 2018; Kaczmariski et al., 2020). The difference in the radius of gyration ($\sim 1 \text{ \AA}$) between the open and closed crystallographic states of AncCDT-1 was also comparable to that observed between the representative open conformation and the closed crystal structure of DalS.

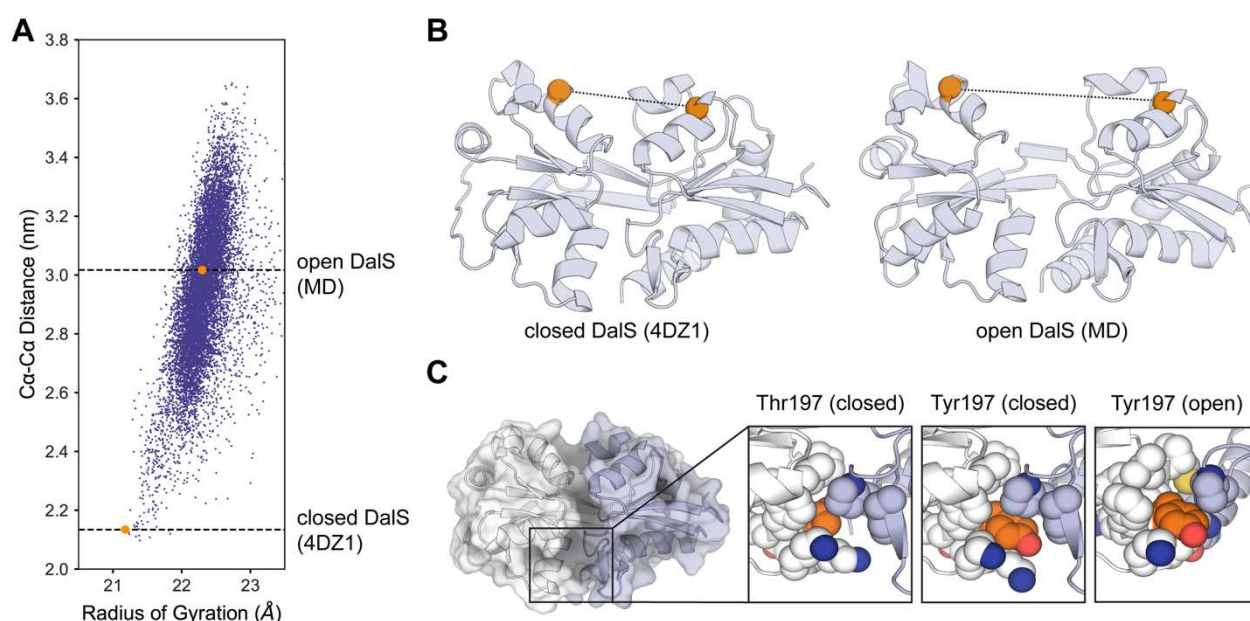


Figure 4. A) Conformational substates of DalS sampled by MD simulations. The Cα-Cα distance between A85 and K153 is plotted against the radius of gyration where each data point represents a single frame of the MD simulation (sampled every 0.1 ns). Points corresponding to the closed crystal structure of DalS (PDB 4DZ1) and the representative open conformation are highlighted in orange. B) The structure of closed DalS (left) is compared to the structure of the representative open conformation. Residues A85 and K153 are represented by orange spheres, highlighting the difference in Cα-Cα distance. C) Thr197 in the closed crystal structure is compared to FoldX-generated (Schymkowitz et al., 2005) models of the T197Y mutation in the open and closed conformations. Residue 197 (orange spheres) and residues within 4 Å (white spheres) are displayed, showing tighter packing of Tyr197 by surrounding residues in the closed conformation compared to the open conformation. The positioning of Tyr197 at the interface between Lobe 1 (purple) and Lobe 2 (white) is also highlighted.

T197Y was modelled into the representative open conformation and the closed crystal structure of DalS using FoldX (Schymkowitz et al., 2005). A comparison of Thr and Tyr at position 197 in the closed conformation (Fig. 4C) showed that the mutation to the larger residue, Tyr, fills a void in the protein between the two lobes and is likely to increase stability by improving hydrophobic packing. Given that residue 197 is positioned at the interface between the two lobes of the binding protein, the improved hydrophobic packing upon mutation to Tyr will stabilise the interaction between them and is likely to stabilize the closed state (relative to Thr). Contrastingly, in the open conformation (Fig. 4C), Tyr197 is less tightly packed by surrounding residues and is more solvent exposed. This is consistent with the predicted effects on stability as calculated by FoldX, with the calculated $\Delta\Delta G$ for introducing the T197Y mutation to the open and closed conformations being -0.7 and -1.6 kcal/mol, respectively. This was also investigated experimentally: the fluorescence spectra of the LSQED-Y mutant, which displayed a reduced dynamic range *in-vitro* (7%), is consistent with a reduced population of the open state (SI. Fig. 2). These results suggest that the remote mutation affects affinity by stabilizing the ligand bound state, which is also consistent with the observation that the affinity increases for all ligands rather than specifically for D-serine (Table 1).

Performance of D-SerFS in hippocampal tissue. The LSQED-Y variant, which is hereafter referred to as D-SerFS (D-serine FRET Sensor), was selected for *in situ* and *in vivo* testing as a biosensor due to its high D-serine affinity, specificity for D-serine over glycine, and thermostability. Furthermore, additional fluorescence titrations of D-SerFS *in-vitro* confirmed that the sensor did not bind other small molecules prominent in brain tissue (L-serine, GABA, L-glutamate, L-aspartate) with any significance (**SI. Fig. 3**). To test the compatibility of D-SerFS with 2PE fluorescence microscopy, the sensor was imaged at an excitation wavelength of 800 nm in free solution under the objective and titrated with increasing concentrations of D-serine. In these conditions, the sensor exhibited similar affinity and dynamic range to that observed *in vitro* previously (Fig. **5A**), confirming that D-SerFS would be suitable for use with 2PE and thus, within intact tissue. Imaging the sensor over time in a meniscus after applying a saturating concentration of D-serine (10 mM) demonstrated that the ECFP/Venus ratio remained stable over a time course of at least 30 minutes (Fig. **5B**).

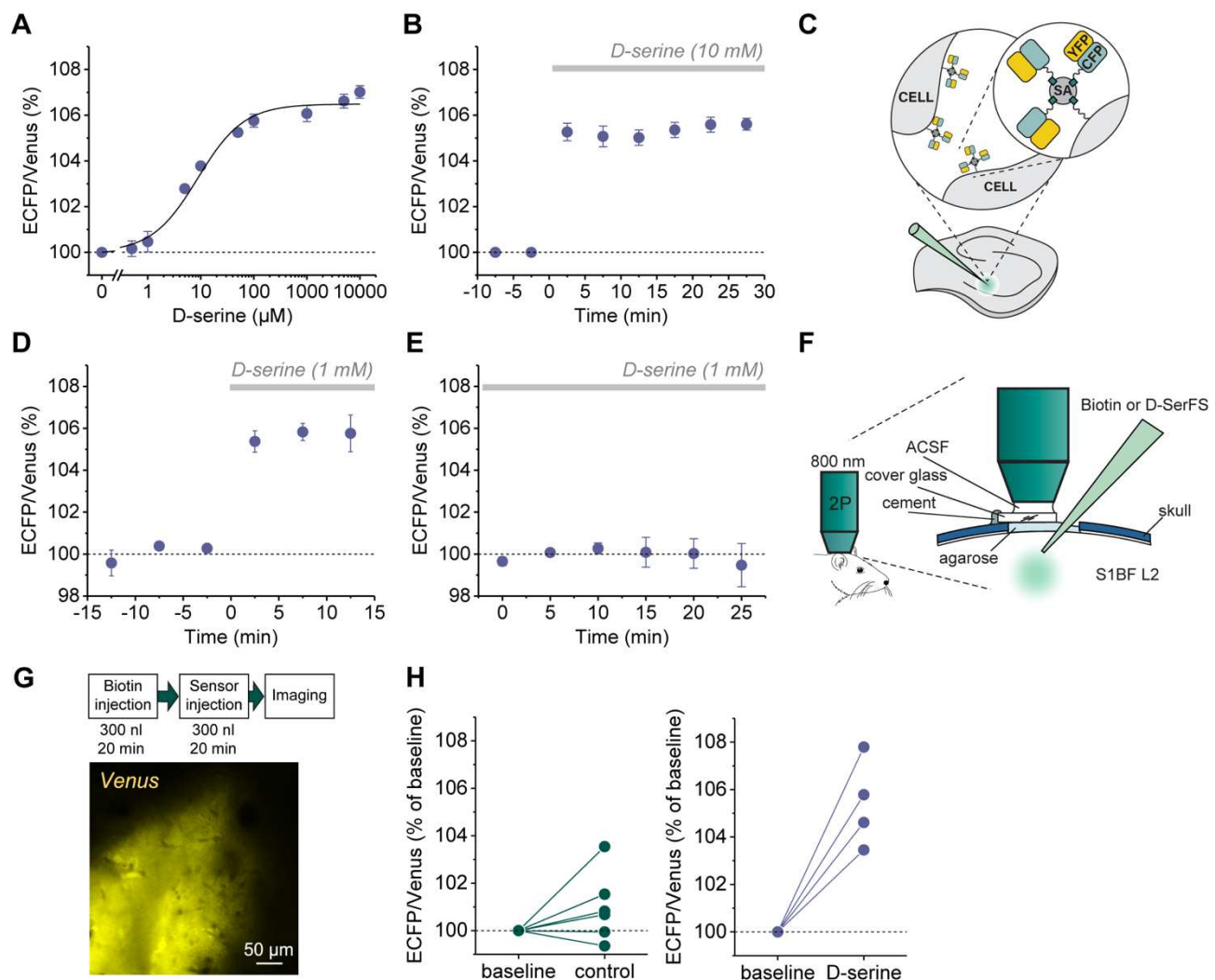


Figure 5. Characterisation, and in situ and in vivo testing of D-SerFS using 2PE fluorescence microscopy. A) Calibration curve of D-SerFS using 2PE ($\lambda_{exc} = 800$ nm) shows a dose-dependent change in ECFP/Venus (475 nm/530 nm) ratio. Fitting the relationship between fluorescence ratio and D-serine concentration with a Hill equation obtains a K_D of 8.8 ± 1.6 μ M (dynamic range 6.5 ± 0.16 %, $n = 3$ independent experiments). B) Monitoring of the ECFP/Venus (475 nm/530 nm) ratio in a meniscus after applying 10 mM D-serine demonstrates a stable fluorescence ratio over a time course of 30 minutes ($n = 5$). C) Illustration of D-SerFS immobilisation in brain tissue. Biotinylated D-SerFS is conjugated to streptavidin (SA) and anchored in the extracellular space of biotinylated slices, as described previously (Whitfield et al., 2015; Zhang et al., 2018). D) Imaging of D-SerFS in response to the exogenous application of 1 mM D-serine in rat brain hippocampal slices ($n = 3$). E) Monitoring of the ECFP/Venus ratio in rat brain hippocampal slices in the presence of 1 mM D-serine demonstrates a stable fluorescence ratio over a time course of 25 minutes ($n = 4$). F) Illustration of imaging of D-SerFS in layer two of somatosensory cortex (barrel field, S1BF) in a urethane-anesthetized mouse using 2PE fluorescence microscopy ($\lambda_{exc} = 800$ nm). G) Schematic of the in vivo testing procedure and image of Venus fluorescence following pressure-loading of D-SerFS into the brain. H) ECFP/Venus ratio recorded in vivo in control recordings (left panel, baseline recording first, control recording after 10 minutes, $n = 6$) and during D-serine application (right panel, baseline recording first, second recording after D-serine injection, 1 mM, $n = 4$). Values are mean \pm s.e.m. throughout.

Next, D-SerFS was tested in 350 μ m thick acute slices of rat hippocampus. The sensor was anchored in the extracellular space of the CA1 region of biotinylated acute hippocampal slices using

a biotin-streptavidin linker, as described previously (Zhang et al., 2018) (Fig. 5C). A rapid change in the ECFP/Venus fluorescence ratio was observed following the bath application of 1 mM D-serine to the hippocampal slice (Fig. 5D). Furthermore, the ratio did not change in the presence of saturating D-serine concentrations for up to 25 minutes (Fig. 5E), indicating that the indicator is stable in this environment.

The *in vivo* performance of D-SerFS was investigated next. A craniotomy was performed above the barrel field (BF) of the primary somatosensory cortex (S1) of urethane-anaesthetized mice. Twenty minutes after the initial application of biotin, D-SerFS was applied to the same area using a glass micropipette (Fig. 5F,G,H, D-serine). The sensor was imaged through the cranial window using 2PE fluorescence microscopy. D-serine (1 mM, 300 nl) was pressure applied to the layer two of somatosensory cortex (130-150 μ m depth) using a glass micropipette and the results were compared to control experiments where no D-serine was applied (Fig 5H, control). These experiments demonstrate that D-SerFS can be stably immobilized and visualized in the brain of living mice while maintaining its ability to report changes of the ambient D-serine concentrations. They thus pave the way for future experiments visualizing endogenous D-serine signalling.

Discussion. In this study, we have constructed the optical sensor D-SerFS and demonstrated its utility in detecting D-serine, an NMDAR co-agonist. In the absence of a suitable D-serine binding protein to use as a recognition domain for the sensor, we used rational and computational design to alter the specificity of an existing D-alanine/glycine binding protein from *Salmonella enterica* (DaIS) (Osborne et al., 2012). An initial round of rational design guided by homology to the NR1 binding domain of the NMDAR (Furukawa & Gouaux, 2003), followed by FoldX (Schymkowitz et al., 2005) modelling and Glide (Friesner et al., 2004) ligand docking, led to a variant of DaIS (LSQE) that displayed D-serine binding specificity (vs. other amino acids). The stepwise introduction of individual binding site mutations towards the proposed rational design demonstrated that there was no single mutational pathway to the improved LSQ variant that displayed improvements at each step owing to the epistatic interactions between these residues. Thus, it was only by introducing all three mutations simultaneously (F117L, A147S, Y148Q) that the improvements to D-serine specificity could be achieved, highlighting the advantage of rational computational design over evolutionary approaches in some instances. Moreover, this allowed us to achieve the improvement in specificity through testing a small number of variants rather than large libraries. However, the affinity of this first round variant (LSQ) for D-serine ($K_D = 41 \pm 1 \mu$ M) was too low to be useful in some biological contexts, such as in neuroscience. Moreover, measurement of the thermostabilities of the mutated variants revealed an overall decrease in thermostability with increased D-serine specificity. As the sensor was required to be robust against the effects of lyophilisation and reconstitution, as well as stable in the crowded environment of brain tissue, we employed the computational methods FoldX (Schymkowitz et al., 2005) and PROSS (Goldenzweig et al., 2016) to identify stabilising mutations to the binding protein. As the mutations predicted by FoldX and the Rosetta-ddG application have

been shown to overlap in only 12 – 25% of all predictions (Buß, Rudat, & Ochsenreither, 2018; Wijma et al., 2014), combining both algorithms had the potential to provide greater coverage of beneficial mutations. In the background of LSQE, the combination of a stabilising surface mutation identified using FoldX (A76D) and a mutation identified using PROSS (T197Y), produced D-SerFS (LSQED-Y), a variant of the sensor that exhibited high thermostability ($T_m = 79.0 \pm 0.3$ °C) and an improved binding affinity for D-serine ($K_D = 7.0 \pm 0.4$ µM), while maintaining specificity for D-serine.

Several stabilising mutations remote from the binding site provided unexpected improvements in binding affinity. In particular, T197Y improved the binding affinity for all ligands while maintaining D-serine specificity. Such effects may be attributed to shifts in the conformational equilibrium of the binding protein. As the binding affinity of a protein for a given ligand may be considered in terms of the equilibria between the open or closed, and apo or bound states, manipulating these equilibria at positions distant from the binding site can lead to changes in affinity, as demonstrated previously (Marvin & Hellenga, 2001). Indeed, investigation of the T197Y mutation in the closed crystal structure of DaIS (PDB 4DZ1) and a representative open conformation obtained from MD simulations, suggested that the mutation would stabilize the ligand bound (closed) conformation. The effect of the T197Y mutation on the relative populations of the open and closed conformations is also consistent with the observed decrease in dynamic range (13% in LSQED vs. 7% in LSQED-Y). This suggests a possible trade-off between improved binding affinity (*via* stabilisation of the closed conformation) and dynamic range in FRET sensor design. In this instance, the slightly reduced dynamic range caused by the T197Y mutation, was significantly outweighed by the improved thermostability and affinity for D-serine of D-SerFS, which allowed the use of the sensor *in-situ* and *in-vivo*.

Testing D-SerFS in acute hippocampal rat brain slices and in the somatosensory cortex *in vivo* demonstrated that the sensor is responsive to D-serine in the environment of both acute slices and *in vivo* and is compatible with 2PE fluorescence microscopy. D-SerFS displayed similar K_D and dynamic range to that observed *in-vitro*, indicating that stabilisation of the binding protein prevented the loss in dynamic range due to lyophilisation and reconstitution observed in preliminary experiments. Using our calibration data for 2PE in free solution and the maximum fluorescence change to saturating concentration of D-serine in brain tissue (see Zhang et al. 2018 and Whitfield et al. 2015), we can estimate the resting concentration of D-serine to be 1-2 µM. This is in line with previous reports using microdialysis, which also measured extracellular D-serine concentrations in the low micromolar range *in vivo* (rat striatum ~ 8 µM (Ciriacks & Bowser, 2004), rat frontal cortex ~ 6 µM (Matsui et al., 1995), mouse barrel cortex 4.2 µM (Takata et al., 2011)). Since astrocytic Ca^{2+} signalling is a trigger for D-serine release (Henneberger et al., 2010; Thomas Papouin, Dunphy, et al., 2017; Thomas Papouin, Henneberger, et al., 2017; Takata et al., 2011), our slightly lower estimates could be the result of deeper anaesthesia, which can disrupt astrocytic Ca^{2+} signalling (Thrane et al., 2012). Compared to the reported D-serine resting levels, the apparent K_D of D-SerFS

(7-9 μ M depending on method) is ideal to detect both increases and decreases of the ambient D-serine concentrations in physiological settings and disease models.

Conclusions

In this work, we have used computational protein design to engineer an existing glycine/D-alanine binding protein (DalS) towards D-serine specificity and greater stability for use in a FRET-based biosensor for D-serine (D-SerFS). We demonstrated that D-SerFS can be used to detect changes of extracellular D-serine levels in rat brain tissue, providing a new tool for the *in-situ* and *in-vivo* study of the transmitter. Further work on D-SerFS could involve improvement of the dynamic range of the sensor for further *in-situ* and *in-vivo* experiments. This work highlights the utility of computational design tools in engineering naturally occurring binding proteins towards novel specificities, particularly where low-throughput experimental tests of affinity are required to discern the subtle effects of binding site mutations on specificity.

Materials and Methods

DNA cloning and mutagenesis. The DalS (Osborne et al., 2012) wild-type gene was synthesized (GeneArt) for sensor cloning, codon-optimized for expression in *Escherichia coli*. Sensor constructs were cloned into a vector backbone denoted as 'pDOTS10'. This utilizes a vector system described previously (Okumoto et al., 2005) (Addgene Plasmid #13537), which contains a pRSET backbone with an N-terminal 6xHis tag and the insertion of a biotin tag from the PinPoint™ Xa-1 Vector (Promega, USA) in between the His tag and the first fluorescent protein. Endogenous *SapI* sites were removed from the ECFP-Venus cassette, and the binding protein gene YbeJ was replaced with a *SapI* linker (ATCAgaagagcactgcatggtGCGGCCGCcaccactctcgctcttcCCTC) designed around the method of Golden Gate cloning (Engler, Kandzia, & Marillonnet, 2008), whereby *SapI* restriction sites (GCTCTTC/GAAGAGC) are used to clone in a gene of interest. Reciprocal *SapI* sites were added to the 5' and 3' ends of the DalS gene by PCR for subsequent cloning into pDOTS10. Mutations were introduced using a combination of long mutagenesis primers and T7 promoter/terminator primers to create gene fragments with >40 base pair overlaps assembled *via* Gibson Assembly (Gibson et al., 2009). These variants were cloned into a new vector backbone (pETMCSIII), retaining the fluorescent proteins and biotin tag.

Protein expression and purification. All proteins were expressed through transformation into BL21(DE3) *E. coli* cells and grown for 72 – 96 h at 18 °C in 1 L autoinducing medium (yeast extract, 5g; tryptone, 20g; NaCl, 5g; KH₂PO₄, 3g; Na₂HPO₄, 6g in 1 L of water to which 10 ml autoclaved 60% (v/v) glycerol, 5 ml autoclaved 10% (w/v) glucose, 25 ml autoclaved 8% (w/v) lactose were added) supplemented with 100 mg ampicillin. The full expression of the fluorescent protein constructs needed to be monitored by observing the ECFP/Venus spectra over time, with Venus emission typically peaking at 72 h of expression at 18 °C.

Cells were harvested through centrifugation and the pellet was stored at -20 °C prior to purification. For purification, the pellet (frozen or otherwise) was suspended in buffer A (50 mM phosphate, 300 mM NaCl, 20 mM imidazole, pH 7.5), lysed by sonication, re-centrifuged at high speed (13,500 r.p.m. for 45 min at 4 °C) and the clarified supernatant was collected. The supernatant was loaded onto a 5 mL Ni-NTA/His-trap column pre-equilibrated in buffer A, washed with 10 column volumes of buffer A, 5 column volumes of 10% buffer B, and eluted with 100% buffer B (50 mM phosphate, 300 mM NaCl, 250 mM imidazole, pH 7.5). The eluted protein was dialyzed against 3 exchanges of 4 L of buffer C (20 mM phosphate, 200 mM NaCl, pH 7.5) at 4 °C. The dialyzed protein was further purified using a HiLoad 26/600 Superdex 200 pg SEC column using buffer C.

Fluorescence assays. Fluorescence titrations were performed on a Varian Cary Eclipse using a Quartz narrow volume fluorescence cuvette. Samples underwent excitation at 433 nm and were scanned over a range of 450 nm – 570 nm for full spectra analysis. ECFP/Venus ratios were

determined using peak wavelength values of 475 nm (ECFP) and 530 nm (Venus). Temperature dependent measurements were obtained using an Applied Photophysics Chirascan™ fluorescence photomultiplier with 433/3 nm excitation and peak fluorescence measured at 475 nm and 530 nm. K_D values were determined by fitting curves with the following equation:

$$y = (y_{min} + \frac{[ligand] * (y_{max} - y_{min})}{(EC_{50} + [ligand])})$$

Computational assessment of mutations. Mutations were assessed by first creating the mutation using the BuildModel command in FoldX (Schymkowitz et al., 2005), using the DalS crystal structure (Osborne et al., 2012) (PDB 4DZ1) that had undergone the FoldX Repair process and had the bound ligand removed. Prior to Schrödinger Glide (Friesner et al., 2004) docking, the Protein Preparation Wizard (Madhavi Sastry, Adzhigirey, Day, Annabhimoju, & Sherman, 2013) was used to assign bond orders, generate protonation states and optimize the hydrogen bonding network at pH 7.4. Minimization of the complex using the OPLS3e (Roos et al., 2019) force field was completed with heavy atoms restrained to 0.30 Å of the input structure. Ligands were downloaded in SDF format and LigPrep was used to generate possible ionization states of the ligand at pH 7.4 ± 2.0. In Receptor Grid Generation, the default scaling factor of the Van der Waals radii of receptor atoms (1.0) and partial charge cut-off (0.25) were used. Rotatable receptor hydroxyl and thiol groups were allowed to rotate. The default scaling factor of the Van der Waals radii for the ligand (0.80) and partial charge cut-off (0.15) were used. In Glide, the standard precision (SP) docking method was used with ligand sampling set to the flexible setting, the option of adding Epik state penalties to the docking score included, and post-docking minimization performed. For Prime MM-GBSA (Rapp et al., 2011) calculations, the VSGB solvent model and OPLS3e force field were used in the computation of binding free energy. No residues were allowed to flex during minimization. For the identification of stabilising mutations using FoldX (Schymkowitz et al., 2005), the PositionScan command was used to mutate the targeted positions to each of the other amino acids, using the DalS crystal structure that had undergone the FoldX Repair process and had the bound ligand removed as the input.

Molecular dynamics simulations and analysis. The crystal structure of DalS (Osborne et al., 2012) (PDB 4DZ1) was prepared for MD simulation by first building in missing residues using the mutagenesis wizard in PyMOL ("The PyMOL Molecular Graphics System, Version 2.3 Schrödinger, LLC,") and removing the D-alanine ligand. The repaired structure was then capped with an N-terminal acetyl group and C-terminal N-methyl amide group. The protein was solvated in a rhombic dodecahedral box with 13 634 TIP3P water molecules, 37 sodium ions and 41 chloride ions (200 mM salt solution). Simulations were performed using GROMACS version 2018.3 (Abraham et al., 2015) using the CHARMM36m forcefield (Best et al., 2012). Long-range electrostatics were treated with the Particle Mesh Ewald method and the Van der Waals cut-off was set to 1.2 nm. The

temperature was coupled to a virtual water bath at 300 K with a Bussi-Donadio-Parrinello 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 2 ns (1 000 000 steps) simulation and production runs were performed for 100 ns each (50 000 000 steps). 10 replicates of the equilibration and production simulations were performed. C α -C α distances and radius of gyration for each frame (sampled every 0.1 ns) was calculated using the ProDy package (Bakan, Meireles, & Bahar, 2011). All replicates were concatenated prior to clustering analysis. Clustering was performed using the gromos method for clustering with an RMSD cut-off of 0.2 nm. The middle structure for the largest cluster was taken as the representative open conformation.

Two-photon excitation (2PE) sensor imaging. Experiments were performed as previously described (Whitfield et al., 2015; Zhang et al., 2018) with minor changes. We used a FV10MP imaging system (Olympus) optically linked to a femtosecond pulse laser (Vision S, Coherent) equipped with a 25x (NA 1.05) objective (Olympus). See below for in vivo imaging. For titrations in solution, D-SerFS was imaged in a meniscus of PBS at a laser power of 3 mW and increasing amounts of D-SerFS (in PBS) were added. For slice experiments (see below for further details), the laser power was adjusted for depth in the tissue to obtain, on average, a fluorescence intensity corresponding to that of 2-3 mW laser power at the slice surface. The excitation wavelength was 800 nm throughout. Fluorescence of ECFP and Venus fluorescent protein was separated using appropriate band pass filters and dichroic mirrors and detected with photomultiplier tubes connected to a single photon counting board (PicoHarp, Picoquant). Their arrival times were recorded using Symphotime 1.5 software (Picoquant). Offline analysis was performed using OriginPro 2017 (OriginLab) and custom written scripts in Matlab R2017a (Mathworks). The ratio of ECFP and Venus fluorescence (R) was calculated from the number of photons detected by the respective detectors in time bins of ~ 220 ms. The photon count rate 11.5 ns to 12.5 ns after the laser pulse (81-82 MHz repetition rate) was subtracted to reduce the contribution of photons not originating from D-SerFS to analysis.

Brain slice preparation. Acute hippocampal slices were prepared from three- to five-week-old male Wistar rats as previously described (Anders et al., 2014; Zhang et al., 2018). All animals used in this study were housed under 12 h light/dark conditions and were allowed ad libitum access to food and water. Briefly, 350 μ m thick horizontal slices containing hippocampal formation were obtained in full compliance with national and institutional regulations (Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen and University of Bonn Medical School) and guidelines of the European Union on animal experimentation. Slices were prepared in an ice-cold slicing solution containing (in mM): NaCl 60, sucrose 105, KCl 2.5, MgCl₂ 7, NaH₂PO₄ 1.25, ascorbic acid 1.3, sodium pyruvate 3, NaHCO₃ 26, CaCl₂ 0.5, and glucose 10 (osmolarity 305–310 mOsm) and kept in the slicing solution at 34 °C for 15 min before being stored at room temperature in an extracellular

solution containing (in mM) NaCl 131, KCl 2.5, MgSO₄ 1.3, NaH₂PO₄ 1.25, NaHCO₃ 21, CaCl₂ 2, and glucose 10 (osmolarity 297–303 mOsm, pH adjusted to 7.4). This solution was also used for recordings. Slices were allowed to rest for at least 50 min. All solutions were continuously bubbled with 95% O₂/ 5% CO₂.

D-SerFS imaging in acute slices. D-SerFS was handled and immobilized in acute hippocampal slices as described before (Whitfield et al., 2015; Zhang et al., 2018). For long-term storage and transport D-SerFS was lyophilized and stored for up to 2 months at ambient temperature (for shipping) and 4° C (storage until use). Before experiments, the sensor was reconstituted in water, agitated for 30 min at room temperature before the buffer was changed to PBS (pH 7.4) with a PD-10 desalting column (GE Healthcare). D-SerFS was then concentrated to 60-100 μM using centricons (Vivaspin 500, 10 kDa cutoff, Sartorius Stedim Biotech). For anchoring of optical sensors in brain tissue, cell surfaces within acute slices were biotinylated using a previously published procedure (Whitfield et al., 2015; Zhang et al., 2018). Briefly, the slice storage solution was supplemented with 50 μM Sulfo-NHS EZ Link Biotin (Thermo Fisher) for 45 min before washing and further storage. Slices were transferred to a submersion-type recording chamber and superfused with extracellular solution at 34 °C. For injections of D-SerFS into the tissue, patch clamp pipettes (2-4 MΩ when filled with PBS saline) were backfilled with PBS (pH 7.4) to which 50-85 μM D-SerFS and 6-10 μM streptavidin (Life Technologies) had been added. The pipette was inserted ~70 μm deep into the tissue under visual control and D-SerFS was pressure-injected. D-SerFS-injected acute slices were allowed to recover for 10-15 minutes before recordings. To reduce the potential effect of scattering of ECFP and Venus fluorescence, we have performed all imaging experiments at a depth of 50-70 μm below the slice surface. Exogenous D-serine was applied via the bath perfusion (1 mM). The following inhibitors were present throughout these experiments to block excitatory synaptic transmission and action potential firing: TTX (1 μM, Sigma Aldrich), D-APV (50 μM, Abcam) and NBQX (10 μM, Tocris).

In vivo imaging and analysis. Imaging was performed in the barrel cortex of 8-12 week old C57BL/6 male and female mice as described previously (Monai et al., 2016; Monai et al., 2019). All experimental protocols were approved by the RIKEN Institutional Animal Care and Use Committee. All efforts were made to minimize the number of animals used and their suffering. Under initial isoflurane inhalation anesthesia (3% for induction, 1.5 % for maintenance), the metal frame was attached to the exposed skull using a dental acrylic (Fuji LUTE BC, GC Corporation, Super Bond C&B, Sunmedical) and the mouse head plate was fixed in the stereotactic apparatus and a small craniotomy (about 2.7 mm diameter, 1.5 mm posterior to bregma and 3.5 mm from the midline) was performed using a dental drill. The skull, but not the dura mater, above the right barrel cortex was carefully removed. The cortex was covered with a 1.5% low-melting agarose and a glass cover slip (2.7 mm in diameter) was placed on top and secured by dental cement leaving a small non-covered window for application of biotin, D-SerFS, and D-serine (**Fig. 5F**).

For recordings, mice were anesthetized with an intraperitoneal urethane injection (1.5 g/kg), head-fixed to a stereotactic stage and placed under a 2PE resonant scanner-based microscope (B-Scope, Thorlabs) equipped with a femtosecond Chameleon Vision 2 laser (Coherent) and an Olympus objective lens (XLPlan N 25X). Animals were kept at a constant temperature of 37° C throughout the surgical procedures and the whole experimental session by placing them on a heating blanket (BWT-100 A, Bio Research Center or TR-200, Fine Science Tools). Then the sensor was immobilized in the barrel field of the primary somatosensory cortex as described elsewhere (Okubo et al., 2010) with minor modifications. First, sulpho-NHS-SS-biotin (about 10 mM in PBS, 300 nl) was injected via a glass micropipette at the depth of about 130-150 µm from dura mater at a rate of 50-100 nl per minute. Second, a mixture of D-SerFS and streptavidin was injected (100 µM in PBS, 300 nl) into the same area 20 min after the biotin injection (**Fig. 5D**). Imaging was then performed 20 min after the sensor injection (800 nm excitation wavelength). In a subset of experiments, D-serine (300 nl, 1 mM) was applied to the same area using a glass micropipette. D-SerFS fluorescence was acquired using frame scanning (ThorImage, 512x512 pixels, frame rate 30 Hz). Data were analyzed offline using FIJI (Schindelin et al., 2012).

Statistics. Data are reported as mean \pm s.e.m. where n is the number of independently performed experiments, unless stated otherwise.

Acknowledgements

V.V., and J.A.M. acknowledge financial support from an Australian Government Research Training Program Scholarship. Research was funded by an ARC Discovery Project awarded to C.J.J. Research in the Fleishman lab was funded by the European Research Council (815379), the Israel Science Foundation (1844), the Milner Foundation and a charitable donation from Sam Switzer and family. The work was further supported by the Human Frontiers Science Program (HFSP; RGY0084/2012 to C.H., H.J. and C.J.J.), the German Academic Exchange Service (DAAD-Go8) Travel Fellowship (to C.H. and C.J.J.), the NRW-Rückkehrerprogramm (C.H.) and German Research Foundation (DFG; SFB1089 B03, SPP1757 HE6949/1, FOR2795 and HE6949/3 to C.H.; SPP1757 young investigator grant to P.U.).

Contributions

J.H.W., V.V., J.A.M., O.K., S.J.F., and C.J.J. designed, produced and analysed the sensor. P.U., B.B., L.K., and C.H. performed and analysed all experiments using two-photon excitation in acute brain slices. P.U., H.M., and H.H. performed and analysed all *in-vivo* experiments. H.J., C.J.J. and C.H. designed the study. C.J.J. and V.V. wrote the initial manuscript, to which all authors subsequently contributed.

Author information

These authors contributed equally: Vanessa Vongsouthi, Jason H. Whitfield

Conflict of interest

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

References

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2, 19-25. doi: <https://doi.org/10.1016/j.softx.2015.06.001>
- Anders, S., Minge, D., Griemsmann, S., Herde, M. K., Steinhäuser, C., & Henneberger, C. (2014). Spatial properties of astrocyte gap junction coupling in the rat hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1654), 20130600. doi: <https://doi.org/10.1098/rstb.2013.0600>
- Bajar, B. T., Wang, E. S., Zhang, S., Lin, M. Z., & Chu, J. (2016). A Guide to Fluorescent Protein FRET Pairs. *Sensors (Basel)*, 16(9). doi: <https://doi.org/10.3390/s16091488>
- Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*, 27(11), 1575-1577. doi: <https://doi.org/10.1093/bioinformatics/btr168>
- Basu, A. C., Tsai, G. E., Ma, C. L., Ehmsen, J. T., Mustafa, A. K., Han, L., . . . Coyle, J. T. (2009). Targeted disruption of serine racemase affects glutamatergic neurotransmission and behavior. *Molecular psychiatry*, 14(7), 719-727. doi: <https://doi.org/10.1038/mp.2008.130>
- Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., & MacKerell, A. D. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ , ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *Journal of Chemical Theory and Computation*, 8(9), 3257-3273. doi: <https://doi.org/10.1021/ct300400x>
- Beyene, A. G., Yang, S. J., & Landry, M. P. (2019). Review Article: Tools and trends for probing brain neurochemistry. *Journal of vacuum science & technology. A, Vacuum, surfaces, and films : an official journal of the American Vacuum Society*, 37(4), 040802-040802. doi: <https://doi.org/10.1116/1.5051047>
- Bliss, T. V., & Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407), 31-39. doi: <https://doi.org/10.1038/361031a0>
- Bliss, T. V. P., & Cooke, S. F. (2011). Long-term potentiation and long-term depression: a clinical perspective. *Clinics (Sao Paulo, Brazil)*, 66 Suppl 1(Suppl 1), 3-17. doi: <https://doi.org/10.1590/s1807-59322011001300002>
- Buß, O., Rudat, J., & Ochsenreither, K. (2018). FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Computational and Structural Biotechnology Journal*, 16, 25-33. doi: <https://doi.org/10.1016/j.csbj.2018.01.002>
- Ciriacks, C. M., & Bowser, M. T. (2004). Monitoring d-Serine Dynamics in the Rat Brain Using Online Microdialysis-Capillary Electrophoresis. *Analytical Chemistry*, 76(22), 6582-6587. doi: <https://doi.org/10.1021/ac0490651>
- Clifton, B. E., Kaczmariski, J. A., Carr, P. D., Gerth, M. L., Tokuriki, N., & Jackson, C. J. (2018). Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nature Chemical Biology*, 14(6), 542-547. doi: <https://doi.org/10.1038/s41589-018-0043-2>

- Corrigan, J. J. (1969). D-Amino Acids in Animals. *Science*, 164(3876), 142. doi: <https://doi.org/10.1126/science.164.3876.142>
- Dai, X., Zhou, E., Yang, W., Zhang, X., Zhang, W., & Rao, Y. (2019). D-Serine made by serine racemase in *Drosophila* intestine plays a physiological role in sleep. *Nature Communications*, 10(1), 1986. doi: <https://doi.org/10.1038/s41467-019-09544-9>
- Dale, N., Hatz, S., Tian, F., & Llaudet, E. (2005). Listening to the brain: microelectrode biosensors for neurochemicals. *Trends in Biotechnology*, 23(8), 420-428. doi: <https://doi.org/10.1016/j.tibtech.2005.05.010>
- Engler, C., Kandzia, R., & Marillonnet, S. (2008). A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE*, 3(11), e3647. doi: <https://doi.org/10.1371/journal.pone.0003647>
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., . . . Shenkin, P. S. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739-1749. doi: <https://doi.org/10.1021/jm0306430>
- Furukawa, H., & Gouaux, E. (2003). Mechanisms of activation, inhibition and specificity: crystal structures of the NMDA receptor NR1 ligand-binding core. *Embo j*, 22(12), 2873-2885. doi: <https://doi.org/10.1093/emboj/cdg303>
- Ganesana, M., Lee, S. T., Wang, Y., & Venton, B. J. (2017). Analytical Techniques in Neuroscience: Recent Advances in Imaging, Separation, and Electrochemical Methods. *Analytical Chemistry*, 89(1), 314-341. doi: <https://doi.org/10.1021/acs.analchem.6b04278>
- Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods*, 6(5), 343-345. doi: <https://doi.org/10.1038/nmeth.1318>
- Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., . . . Fleishman, S. J. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol Cell*, 63(2), 337-346. doi: <https://doi.org/10.1016/j.molcel.2016.06.012>
- Hardingham, G. E., & Bading, H. (2010). Synaptic versus extrasynaptic NMDA receptor signalling: implications for neurodegenerative disorders. *Nature Reviews Neuroscience*, 11(10), 682-696. doi: <https://doi.org/10.1038/nrn2911>
- Hashimoto, A., Nishikawa, T., Hayashi, T., Fujii, N., Harada, K., Oka, T., & Takahashi, K. (1992). The presence of free D-serine in rat brain. *FEBS Lett*, 296(1), 33-36. doi: [https://doi.org/10.1016/0014-5793\(92\)80397-y](https://doi.org/10.1016/0014-5793(92)80397-y)
- Henneberger, C., Papouin, T., Oliet, S. H. R., & Rusakov, D. A. (2010). Long-term potentiation depends on release of d-serine from astrocytes. *Nature*, 463(7278), 232-236. doi: <https://doi.org/10.1038/nature08673>

Kaczmariski, J. A., Mahawaththa, M. C., Feintuch, A., Clifton, B. E., Adams, L. A., Goldfarb, D., . . . Jackson, C. J. (2020). Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme. *bioRxiv*, 2020.2002.2003.932491. doi:

<https://doi.org/10.1101/2020.02.03.932491>

Kaczmariski, J. A., Mitchell, J. A., Spence, M. A., Vongsouthi, V., & Jackson, C. J. (2019). Structural and evolutionary approaches to the design and optimization of fluorescence-based small molecule biosensors. *Current Opinion in Structural Biology*, 57, 31-38. doi:

<https://doi.org/10.1016/j.sbi.2019.01.013>

Labrie, V., Fukumura, R., Rastogi, A., Fick, L. J., Wang, W., Boutros, P. C., . . . Roder, J. C. (2009). Serine racemase is associated with schizophrenia susceptibility in humans and in a mouse model. *Hum Mol Genet*, 18(17), 3227-3243. doi: <https://doi.org/10.1093/hmg/ddp261>

Liu, S., Liu, Q., Tabuchi, M., & Wu, M. N. (2016). Sleep Drive Is Encoded by Neural Plastic Changes in a Dedicated Circuit. *Cell*, 165(6), 1347-1360. doi:

<https://doi.org/10.1016/j.cell.2016.04.013>

Lyne, P. D., Lamb, M. L., & Saeh, J. C. (2006). Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring. *Journal of Medicinal Chemistry*, 49(16), 4805-4808. doi: <https://doi.org/10.1021/jm060522a>

Madeira, C., Lourenco, M. V., Vargas-Lopes, C., Suemoto, C. K., Brandão, C. O., Reis, T., . . . Panizzutti, R. (2015). d-serine levels in Alzheimer's disease: implications for novel biomarker development. *Translational Psychiatry*, 5(5), e561-e561. doi: <https://doi.org/10.1038/tp.2015.52>

Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des*, 27(3), 221-234. doi: <https://doi.org/10.1007/s10822-013-9644-8>

Marvin, J. S., & Hellinga, H. W. (2001). Manipulation of ligand binding affinity by exploitation of conformational coupling. *Nature Structural Biology*, 8(9), 795-798. doi:

<https://doi.org/10.1038/nsb0901-795>

Matsui, T.-a., Sekiguchi, M., Hashimoto, A., Tomita, U., Nishikawa, T., & Wada, K. (1995). Functional Comparison of d-Serine and Glycine in Rodents: The Effect on Cloned NMDA Receptors and the Extracellular Concentration. *Journal of Neurochemistry*, 65(1), 454-458. doi:

<https://doi.org/10.1046/j.1471-4159.1995.65010454.x>

Mohd Zain, Z., Ab Ghani, S., & O'Neill, R. D. (2012). Amperometric microbiosensor as an alternative tool for investigation of D-serine in brain. *Amino Acids*, 43(5), 1887-1894. doi:

<https://doi.org/10.1007/s00726-012-1365-0>

Monai, H., Ohkura, M., Tanaka, M., Oe, Y., Konno, A., Hirai, H., . . . Hirase, H. (2016). Calcium imaging reveals glial involvement in transcranial direct current stimulation-induced plasticity in mouse brain. *Nature Communications*, 7(1), 11100. doi: <https://doi.org/10.1038/ncomms11100>

Monai, H., Wang, X., Yahagi, K., Lou, N., Mestre, H., Xu, Q., . . . Hirase, H. (2019). Adrenergic receptor antagonism induces neuroprotection and facilitates recovery from acute ischemic stroke.

Proceedings of the National Academy of Sciences, 116(22), 11010. doi:

<https://doi.org/10.1073/pnas.1817347116>

Mothet, J.-P., Parent, A. T., Wolosker, H., Brady, R. O., Linden, D. J., Ferris, C. D., . . . Snyder, S. H. (2000). D-Serine is an endogenous ligand for the glycine site of the *N*-methyl-D-aspartate receptor. *Proceedings of the National Academy of Sciences*, 97(9), 4926-4931. doi:

<https://doi.org/10.1073/pnas.97.9.4926>

Okubo, Y., Sekiya, H., Namiki, S., Sakamoto, H., Iinuma, S., Yamasaki, M., . . . Iino, M. (2010). Imaging extrasynaptic glutamate dynamics in the brain. *Proceedings of the National Academy of Sciences*, 107(14), 6526. doi: <https://doi.org/10.1073/pnas.0913154107>

Okumoto, S., Looger, L. L., Micheva, K. D., Reimer, R. J., Smith, S. J., & Frommer, W. B. (2005). Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. *Proc Natl Acad Sci U S A*, 102(24), 8740. doi:

<https://doi.org/10.1073/pnas.0503274102>

Osborne, S. E., Tuinema, B. R., Mok, M. C., Lau, P. S., Bui, N. K., Tomljenovic-Berube, A. M., . . . Coombes, B. K. (2012). Characterization of DalS, an ATP-binding cassette transporter for D-alanine, and its role in pathogenesis in *Salmonella enterica*. *J Biol Chem*, 287(19), 15242-15250. doi: <https://doi.org/10.1074/jbc.M112.348227>

Panatier, A., Theodosis, D. T., Mothet, J.-P., Touquet, B., Pollegioni, L., Poulain, D. A., & Oliet, S. H. R. (2006). Glia-Derived d-Serine Controls NMDA Receptor Activity and Synaptic Memory. *Cell*, 125(4), 775-784. doi: <https://doi.org/10.1016/j.cell.2006.02.051>

Papouin, T., Dunphy, J. M., Tolman, M., Dineley, K. T., & Haydon, P. G. (2017). Septal Cholinergic Neuromodulation Tunes the Astrocyte-Dependent Gating of Hippocampal NMDA Receptors to Wakefulness. *Neuron*, 94(4), 840-854.e847. doi: <https://doi.org/10.1016/j.neuron.2017.04.021>

Papouin, T., Henneberger, C., Rusakov, D. A., & Oliet, S. H. R. (2017). Astroglial versus Neuronal D-Serine: Fact Checking. *Trends in Neurosciences*, 40(9), 517-520. doi: <https://doi.org/10.1016/j.tins.2017.05.007>

Papouin, T., Ladepeche, L., Ruel, J., Sacchi, S., Labasque, M., Hanini, M., . . . Oliet, S. H. (2012). Synaptic and extrasynaptic NMDA receptors are gated by different endogenous coagonists. *Cell*, 150(3), 633-646. doi: <https://doi.org/10.1016/j.cell.2012.06.029>

Pernot, P., Mothet, J.-P., Schuvailo, O., Soldatkin, A., Pollegioni, L., Pilone, M., . . . Marinesco, S. (2008). Characterization of a Yeast d-Amino Acid Oxidase Microbiosensor for d-Serine Detection in the Central Nervous System. *Analytical Chemistry*, 80(5), 1589-1597. doi: <https://doi.org/10.1021/ac702230w>

Pollegioni, L., & Sacchi, S. (2010). Metabolism of the neuromodulator D-serine. *Cell Mol Life Sci*, 67(14), 2387-2404. doi: <https://doi.org/10.1007/s00018-010-0307-9>

Popiolek, M., Tierney, B., Steyn, S. J., & DeVivo, M. (2018). Lack of Effect of Sodium Benzoate at Reported Clinical Therapeutic Concentration on d-Alanine Metabolism in Dogs. *ACS Chemical Neuroscience*, 9(11), 2832-2837. doi: <https://doi.org/10.1021/acschemneuro.8b00229>

The PyMOL Molecular Graphics System, Version 2.3 Schrödinger, LLC.

Rapp, C., Kalyanaraman, C., Schiffmiller, A., Schoenbrun, E. L., & Jacobson, M. P. (2011). A molecular mechanics approach to modeling protein-ligand interactions: relative binding affinities in congeneric series. *Journal of chemical information and modeling*, 51(9), 2082-2089. doi:

<https://doi.org/10.1021/ci200033n>

Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J. M., Lu, C., . . . Harder, E. D. (2019). OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *Journal of Chemical Theory and Computation*, 15(3), 1863-1874. doi: <https://doi.org/10.1021/acs.jctc.8b01026>

Schell, M. J., Molliver, M. E., & Snyder, S. H. (1995). D-serine, an endogenous synaptic modulator: localization to astrocytes and glutamate-stimulated release. *Proceedings of the National Academy of Sciences*, 92(9), 3948-3952. doi: <https://doi.org/10.1073/pnas.92.9.3948>

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., . . . Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7), 676-682. doi: <https://doi.org/10.1038/nmeth.2019>

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server issue), W382-W388. doi: <https://doi.org/10.1093/nar/gki387>

Shippenberg, T. S., & Thompson, A. C. (2001). Overview of microdialysis. *Current protocols in neuroscience*, Chapter 7, Unit7.1-Unit7.1. doi: <https://doi.org/10.1002/0471142301.ns0701s00>

Takata, N., Mishima, T., Hisatsune, C., Nagai, T., Ebisui, E., Mikoshiba, K., & Hirase, H. (2011). Astrocyte Calcium Signaling Transforms Cholinergic Modulation to Cortical Plasticity In Vivo. *The Journal of Neuroscience*, 31(49), 18155-18165. doi: <https://doi.org/10.1523/jneurosci.5289-11.2011>

Thrane, A. S., Rangroo Thrane, V., Zeppenfeld, D., Lou, N., Xu, Q., Nagelhus, E. A., & Nedergaard, M. (2012). General anesthesia selectively disrupts astrocyte calcium signaling in the awake mouse cortex. *Proceedings of the National Academy of Sciences*, 109(46), 18974-18979. doi: <https://doi.org/10.1073/pnas.1209448109>

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., & Tawfik, D. S. (2007). The Stability Effects of Protein Mutations Appear to be Universally Distributed. *J Mol Biol*, 369(5), 1318-1332. doi: <https://doi.org/10.1016/j.jmb.2007.03.069>

Tomita, J., Ueno, T., Mitsuyoshi, M., Kume, S., & Kume, K. (2015). The NMDA Receptor Promotes Sleep in the Fruit Fly, *Drosophila melanogaster*. *PLoS ONE*, 10(5), e0128101. doi: <https://doi.org/10.1371/journal.pone.0128101>

Whitfield, J. H., Zhang, W. H., Herde, M. K., Clifton, B. E., Radziejewski, J., Janovjak, H., . . . Jackson, C. J. (2015). Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Science*, 24(9), 1412-1422. doi: <https://doi.org/10.1002/pro.2721>

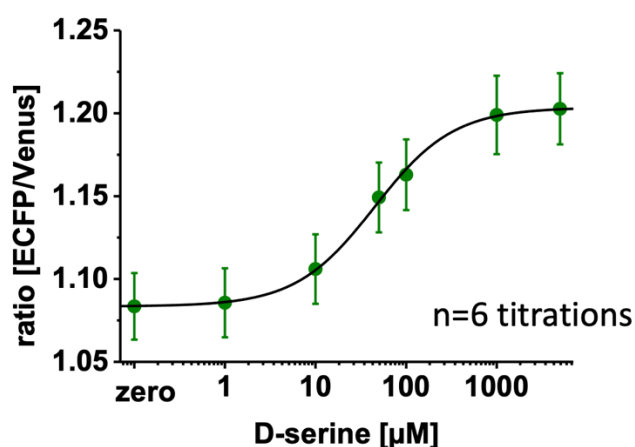
- Wijma, H. J., Floor, R. J., Jekel, P. A., Baker, D., Marrink, S. J., & Janssen, D. B. (2014). Computationally designed libraries for rapid enzyme stabilization. *Protein engineering, design & selection : PEDS*, 27(2), 49-58. doi: <https://doi.org/10.1093/protein/gzt061>
- Wiriyasermkul, P., Moriyama, S., Tanaka, Y., Kongpracha, P., Nakamae, N., Suzuki, M., . . . Nagamori, S. (2020). D-Serine, an emerging biomarker of kidney diseases, is a hidden substrate of sodium-coupled monocarboxylate transporters. *bioRxiv*, 2020.2008.2010.244822. doi: <https://doi.org/10.1101/2020.08.10.244822>
- Wolosker, H., Balu, D. T., & Coyle, J. T. (2016). The Rise and Fall of the d-Serine-Mediated Gliotransmission Hypothesis. *Trends in Neurosciences*, 39(11), 712-721. doi: <https://doi.org/10.1016/j.tins.2016.09.007>
- Wolosker, H., Blackshaw, S., & Snyder, S. H. (1999). Serine racemase: a glial enzyme synthesizing D-serine to regulate glutamate-N-methyl-D-aspartate neurotransmission. *Proc Natl Acad Sci U S A*, 96(23), 13409-13414. doi: <https://doi.org/10.1073/pnas.96.23.13409>
- Yang, Y., Ge, W., Chen, Y., Zhang, Z., Shen, W., Wu, C., . . . Duan, S. (2003). Contribution of astrocytes to hippocampal long-term potentiation through release of D-serine. *Proceedings of the National Academy of Sciences*, 100(25), 15194. doi: <https://doi.org/10.1073/pnas.2431073100>
- Zhang, W. H., Herde, M. K., Mitchell, J. A., Whitfield, J. H., Wulff, A. B., Vongsouthi, V., . . . Henneberger, C. (2018). Monitoring hippocampal glycine with the computationally designed optical sensor GlyFS. *Nature Chemical Biology*, 14(9), 861-869. doi: <https://doi.org/10.1038/s41589-018-0108-2>

Supplementary Table 1. Change in free energy of folding ($\Delta\Delta G$) calculated by both FoldX and Rosetta computational mutation scanning for all stabilising mutations predicted by FoldX PositionScan (blue) and PROSS (green). Mutations are ranked by increasing $\Delta\Delta G$ (kcal/mol) predicted by FoldX. Variants above the dashed line had a $\Delta\Delta G$ more negative than 2.5 FoldX standard deviations (< 1.15 kcal/mol) and were selected for experimental testing.

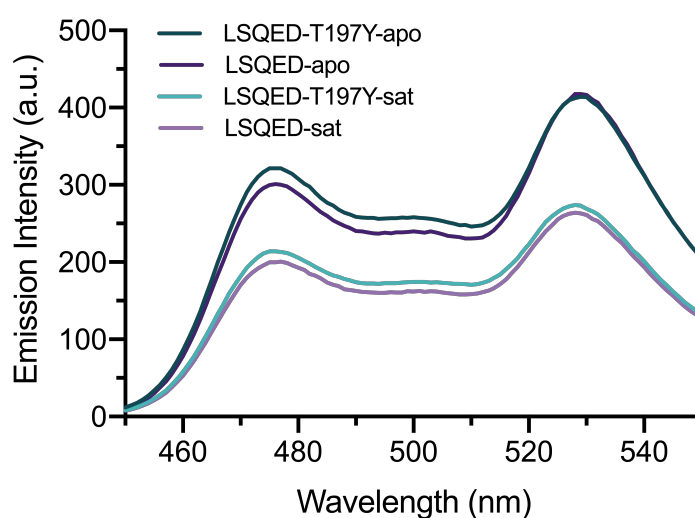
Variant	Predicted $\Delta\Delta G$		ΔT_m relative to LSQED ($^{\circ}\text{C}$)
	FoldX (kcal/mol)	Rosetta (R.e.u.)	
A76D	-2.30	-1.82	2.8 ^a
S208E	-1.94	8.22	1.2
S60K ^b	-1.60	-0.26	1.3
T172D	-1.41	4.78	6.0
S60R	-1.31	-0.33	14.0
T197Y	-1.25	-0.82	14.0
N200K	-1.17	-0.34	4.4
H67G	-1.04	-0.29	-
A76S	-0.87	-1.87	-
G82A	-0.85	-0.69	-
Q242E	-0.85	-0.40	-
K199L	-0.80	-0.60	-
G246A	-0.78	-0.62	-
S61A	-0.67	-1.12	-
D150E	-0.65	2.51	-
G246E	-0.65	-0.50	-
N32R	-0.63	1.07	-
S245K	-0.50	-1.62	-
S245Q	-0.43	-0.84	-
A123K	-0.33	-0.16	-
Q88K	-0.20	-0.49	-
T73V	-0.09	-0.78	-
N136K	-0.09	-0.18	-
H125D	-0.04	-0.59	-
T176A	0.04	-0.41	-
T176V	0.04	2.52	-
N133S	0.09	-1.16	-
Q64K	0.09	-0.36	-
E110D	0.24	-0.46	-
S65E	0.33	-0.36	-
N131D	0.44	0.16	-
I114T	0.46	-0.17	-
L234D	1.19	-0.24	-
L70C	2.39	1.02	-

^a ΔT_m of A76D is relative to LSQE. All other experimentally determined ΔT_m values are relative to LSQE/A76D (LSQED).

^bS60K was predicted by both FoldX PositionScan and PROSS.

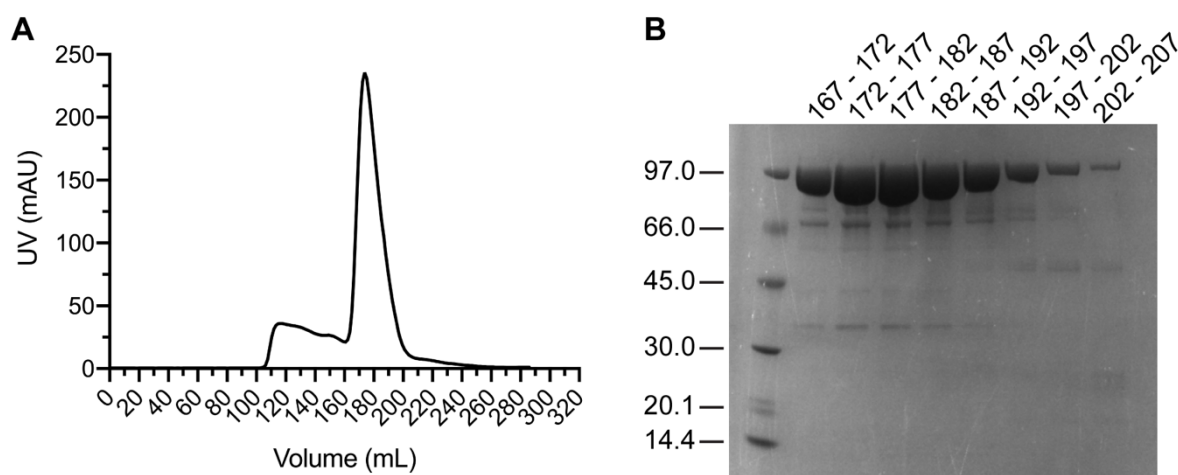


Supplementary Figure 1. Preliminary titration of the LSQ variant under two-photon excitation (2PE) fluorescence microscopy, demonstrating a reduced dynamic range ($\sim 10\%$) following lyophilisation and reconstitution compared to prior ($\sim 13\%$).



Supplementary Figure 2. Emission spectra (450 – 550 nm) of LSQED and LSQED-T197Y (D-SerFS) upon excitation of ECFP ($\lambda_{\text{exc}} = 433 \text{ nm}$). In the apo states, LSQED-T197Y (dark teal) exhibits a higher ECFP emission (475 nm) compared to LSQED (dark purple), while the Venus emission (530 nm) remains similar. The T197Y mutation results in an increase in the ECFP/Venus ratio of the apo state, shifting the ratio closer to that of the saturated sensor as the ECFP/Venus ratio increases with increasing D-serine concentration. This explains the decreased dynamic range of LSQED-T197Y and suggests a shift towards the closed state in the apo population.

Supplementary Figure 3. Fluorescence titration of LSQED-T197Y (D-SerFS) against L-serine, L-glutamate, L-aspartate, GABA, and D-serine and glycine for control. No binding of L-serine, L-glutamate, L-aspartate or GABA was detected. The purity of L-serine used was ~ 99%, thus, concentrations of 500 and 1000 μ M of L-serine possibly contain D-serine at concentrations of ~ 5 and ~ 10 μ M, respectively.



Supplementary Figure 4. A) Chromatogram from size-exclusion chromatography (HiLoad 26/600 Superdex 200 pg) of D-SerFS following Ni^{2+} -affinity chromatography. The main elution peak occurs between 167 – 207 mL. B) Eight 5 mL fractions from the elution peak (167 – 207 mL) were analysed for purity by SDS-PAGE. The most concentrated fractions occurred between 172 – 187 mL and contained minor impurities compared to the band corresponding to D-SerFS (97 kDa). These fractions were pooled for subsequent experiments.

Supplementary Figure 5. Full amino-acid sequence of D-Ser-FS

6xHis Tag, Biotin Purification Tag, ECFP, Linker 1, DalS LSQED-Y binding core, Linker 2, Venus

HHHHHHGMASMTGGQQMGRDLYDDDDKDPKLKVTVNGTAYDVDVDVDKSHENPMG-
TILFGGGTGGAPAPAAGGAGAGKAGEGEIPA-
PLAGTVSKILVKEGDTVKAGQTVLVLEAMKME-
TEINAPTDGKVEKVLVKERDAVQGGQGLIKI-
GDLELIEGSSGSDPGRMVSKGEELFTGVVP-
ILVELDGDVNGHKFSVSGEGEGDATYG-
KLTLKFICTTGKLPVPWPTLVTTLTWGVQCFS-
RYPDHMKQHDFFKSAMPEGYVQERTIFFKD-
DGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNIL-
GHKLEYNYISHNVYITADKQKNGIKANFK-
IRHNIEDGSVQLADHYQQNTPIGDGPVLL-
PDNHYLSTQSALSKDPNEKRDHMLLEF-
VTAAGITLGMDELYKGGTGIMI-
VEGRTLNVAVSPASPPMLFKSADGKLQGID-
LELFSSYCQSRHCKLNITEYDWDGMLGAVASGQA-
DVAFSGISITDKRKKVIDFSEPYIINSLYLVSMANH-
KITLNNLNELNKYSIGYPRGMSQSOLIKNDL-
EPKGYYSLSKVLYPTYNETMADLKNGNLD-
LAFIEEPVYFYFKNKKKMPIESRYVFK-
NVEQLGIAFKKGSPVRDDFNLWLKEQGPQK-
ISGIVDSWMKLAGTGGMVSKGEELFTGVVP-
ILVELDGDVNGHKFSVSGEGEGDATY GK-
LTLKLICTTGKLPVPWPTLVTTLTGYGLQCFA-
RYPDHMKQHDFFKSAMPEGYVQERTIFFKD-
DGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNIL-
GHKLEYNYN SHNVYITADKQKNGIKANFKIRH-
NIEDGGVQLADHYQQNTPIGDGPVLLPDNHYL-
SYQSALSKDPNEKRDHMLLEFVTAAGITLGMDELYK*

Chapter 6

T-dependent B cell responses to *Plasmodium* induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein

6.1 Preface

As the hawk hunts mice or the blue whale hunts krill, so *Plasmodium falciparum* hunts humans. The parasite is the deadliest species of the genus *Plasmodium*, which causes the disease malaria. In 1740, the term ‘malaria’ was borrowed from the Italian for ‘bad air’;²⁹⁷ over a thousand years earlier it was called the ‘Roman fever’²⁹⁸ in Europe and was being treated effectively in China.²⁹⁹ Direct DNA evidence for human infection stretches back at least 4 000 years to ancient Egypt.³⁰⁰ The disease has had a significant impact on human evolution since the advent of agriculture 10 000 years ago led to a sudden rise in the population of the parasite.³⁰¹

Co-evolution has given rise to a number of human blood disorders that confer some resistance to the parasite, like sickle cell disease³⁰² and G6PD deficiency,³⁰³ as well as the loss of the sialic acid Neu5Gc³⁰⁴ which in most other mammals plays an important role in cell identification. *Plasmodium*, of course, quickly adapted, but the legacy of this evolutionary arms race may contribute to the prevalence of auto-immune disorders and cancers in humans compared to other mammals.³⁰⁴ The threat of malaria therefore represents an unusually sudden and dramatic selective pressure that has since characterised all of human health.

Malaria infected an estimated 228 million people in 2018, killing 405 thousand of them. Though this tropical disease is largely absent from developed countries, it is endemic in much

of the developing world, with six countries accounting for over half of the world's burden. Effective treatments are available, but drug resistance is a persistent problem and the existing RTS,S vaccine is only partially effective and not yet commercially available.^{305,306} Vaccine development has been hindered by the limited prevalence of the disease in wealthy markets, and seemingly by intrinsic features of the parasite that resist acquired immunity — individuals who recover from malaria are frequently re-infected by the same strain of parasite soon after recovery.³⁰⁷

6.1.1 Antibodies

Antibodies (figure 6.1) are large proteins responsible for identifying and binding to nearly arbitrary foreign molecules as part of the acquired immune response. When it encounters a foreign chemical signature, or epitope, the immune system matures antibodies to bind it in a process analogous to evolution. The mature antibody typically binds tightly and specifically to the epitope, allowing it to either directly inhibit the attacker or to recruit other elements of the immune system. Antibodies are produced by so-called B cells, which can either express them on their plasma membrane or secrete them in a soluble form. Differentiated B cells, which yield mature antibodies, are then retained by the immune system in case it encounters the same epitope again.

Vaccines work by presenting a known disease-associated epitope to the immune system and allowing it to mature antibodies without the threat of disease. The differentiated B cells are retained after the vaccine clears the body, granting acquired immunity to the disease. Understanding how antibodies are matured in response to a particular epitope is therefore essential to understanding why a vaccine is not completely effective.

Antibodies consist of four glycosylated peptide chains — two identical light chains and two identical heavy chains. Each light chain binds to one heavy chain, yielding two identical Fragment antigen-binding (Fab) regions that include the antibody's binding site at their tip. The opposite end of the heavy chain's Fab region is fused via a flexible linker, or hinge, to the heavy chain's Fragment crystallisable (Fc) region. In the Fc region, both heavy chains are bound together, so that the antibody forms a Y shape with the Fc region at the base and the two Fab regions as the two arms. The Fc region varies based on how the antibody is expressed or presented, but the Fab region always consists of four immunoglobulin domains, two on each chain. Within the Fab region therefore, both the light chain and heavy chain are of approximately equal size, about 220 residues.

Each immunoglobulin domain consists of two back-to-back beta sheets with a total of seven to nine strands in a Greek key motif. The immunoglobulin domain at the binding end of both

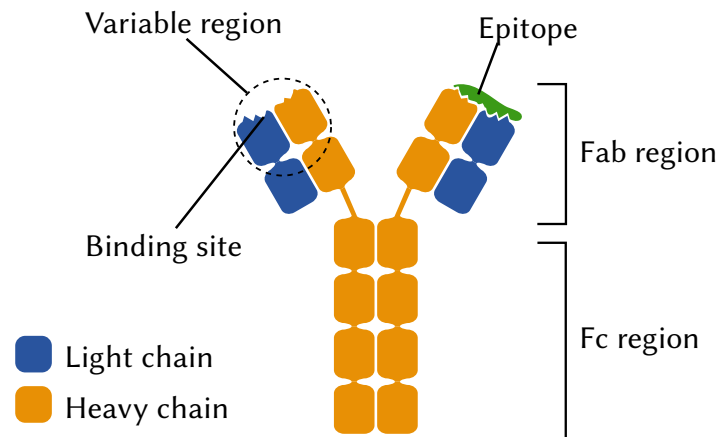


Figure 6.1: An antibody. The light chains are in blue, while the heavy chains are in yellow. A bound epitope is shown in green on the right while the identical Fab on the left is shown unbound. The variable region of the left Fab is circled. Individual immunoglobulin domains are shown as connected rounded rectangles.

the heavy and light chains is called the variable region, which is modified over the course of maturation to bind the epitope. The remaining two domains of the Fab region, one on each chain, are held constant during maturation.

In addition to their importance to immunity, antibodies have become a useful tool in synthetic biology. They offer an existing biological process for the development of binding proteins for arbitrary chemicals. The Fab region can be expressed alone as a binding protein, or fused to other proteins for additional functionality. Studying the maturation process has also been a useful model of both evolution and protein-substrate binding, and binding a target protein to an antibody is a proven strategy to stabilise proteins for crystallisation.

6.1.2 The Circumsporozoite NANP repeat region

A human is first infected with the *Plasmodium* parasite in the sporozoite stage of its life cycle. In this asymptomatic stage, the parasite travels from the mosquito bite site to the liver where it can develop further. The sporozoite is decorated with the circumsporozoite protein, an approximately 400 residue protein consisting of two domains joined by a long linker. The ~100 residue N-terminal domain is disordered and includes a heparan sulfate binding site, but is otherwise of unknown function, while the C-terminal domain is more structured and mostly consists of a thrombospondin-like type I repeat.^{308,309}

The remaining linker region primarily comprises a variable number (30 – 50) of repeats of the amino acid sequence NANP, interspersed with a few copies of similar sequences like NVDP. This long, unstructured linker region is highly conserved but plays an unknown role for the parasite. Nonetheless, it is known to be an important epitope for human immunity, and

the most well developed malaria vaccine consists largely of this repeat sequence presented to maximise the immune response.³¹⁰

In this study, we investigate how antibodies bind to the NANP sequence, and the dynamics of the sequence itself. We find that the high binding avidity is achieved by the ability of many relatively low-affinity antibodies to bind simultaneously. Our results were confirmed later by cryo-EM.³⁰⁹ Our investigation of the dynamics of the NANP repeat sequence sets the stage for further investigation of the sequence as a linker in synthetic fusion proteins.

6.2 Statement of contribution

I declare that the research presented in this chapter represents original work that I carried out during my candidature at the Australian National University, except for contributions to multi-author papers incorporated in the chapter where my contributions are specified in this Statement of Contribution.

6.2.1 Publication status

This manuscript has been published with the title *T-dependent B cell responses to Plasmodium induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein* in the journal **PLoS Pathogens** (2017, **13**(7):e1006469). The formatted article with supporting information is reproduced in this chapter.

6.2.2 Authorship and contribution

The manuscript was authored by Camilla R. Fisher, Henry J. Sutton, Joe A. Kaczmarek, Hayley A. McNamara, Ben Clifton, Joshua Mitchell (the author), Yeping Cai, Johanna N. Dups, Nicholas J. D'Arcy, Mandeep Singh, Aaron Chuah, Thomas S. Peat, Colin J. Jackson, and Ian A. Cockburn. CRF and HJS contributed equally to the work. The paper includes a MD modelling section in which we briefly investigate the reasonableness of a putative structure of the peptide under study. I designed, performed, analysed and wrote up this section, as well as constructing the multivalent model that appears in figure 4 and editing the paper. In addition, I contributed supplementary figures S2 and S3 and movies S8 and S9, all of which describe the MD work I performed.

RESEARCH ARTICLE

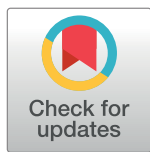
T-dependent B cell responses to *Plasmodium* induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein

Camilla R. Fisher¹*, Henry J. Sutton²*, Joe A. Kaczmariski¹, 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

Citation: Fisher CR, Sutton HJ, Kaczmariski JA, McNamara HA, Clifton B, Mitchell J, et al. (2017) T-dependent B cell responses to *Plasmodium* induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein. PLoS Pathog 13(7): e1006469. <https://doi.org/10.1371/journal.ppat.1006469>

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](https://creativecommons.org/licenses/by/4.0/), 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_{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

had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

determine the structure of CSP-binding antibodies at the atomic level. We use other techniques such as isothermal titration calorimetry and structural modeling to examine how this antibody interacts with the CSP molecule. Strikingly, we found that each CSP molecule 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 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 seminal studies that showed that immunization with irradiated sporozoites could induce sterile protection 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 original 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 effective *P. falciparum* sporozoite neutralizing antibodies identified in these early studies was 2A10 which can block sporozoite infectivity *in vitro* [7] and in *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 2A10 F_{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

Table 1. Thermodynamic parameters for interactions between 2A10 F_{AB}, 2A10 and antigens.

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

Parameters were determined by ITC at 25°C. Errors for n (Ag: F_{AB}), 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>

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 F_{AB}: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 F_{AB} 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 F_{AB} 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 F_{AB} domains to accumulate, which yields a complex with a high avidity dissociation constant in the low nM range.

Structural analysis of the (NANP)-repeat region and the 2A10 F_{AB}

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

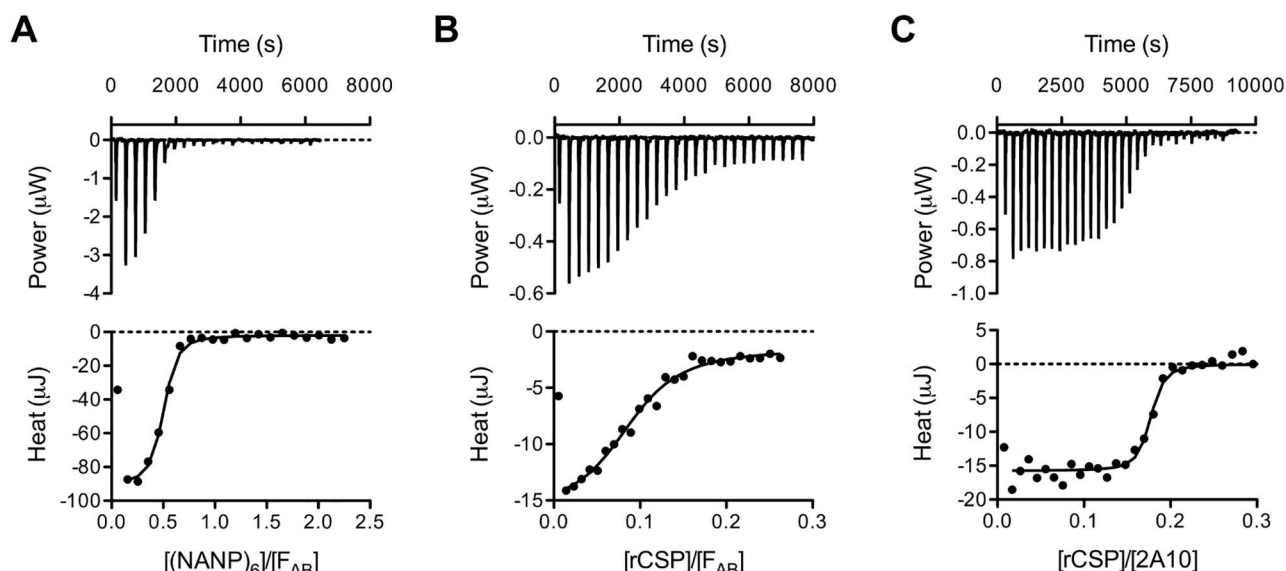


Fig 1. ITC data for interactions between 2A10 F_{AB} and antigens. (A) Titration of 2A10 F_{AB} with (NANP)₆. (B) Titration of 2A10 F_{AB} with rCSP. (C) Titration of 2A10 (complete antibody) with rCSP. The upper panels represent baseline-corrected power traces. By convention, negative power corresponds to exothermic binding. The lower panels represent the integrated heat data fitted to the independent binding sites model.

<https://doi.org/10.1371/journal.ppat.1006469.g001>

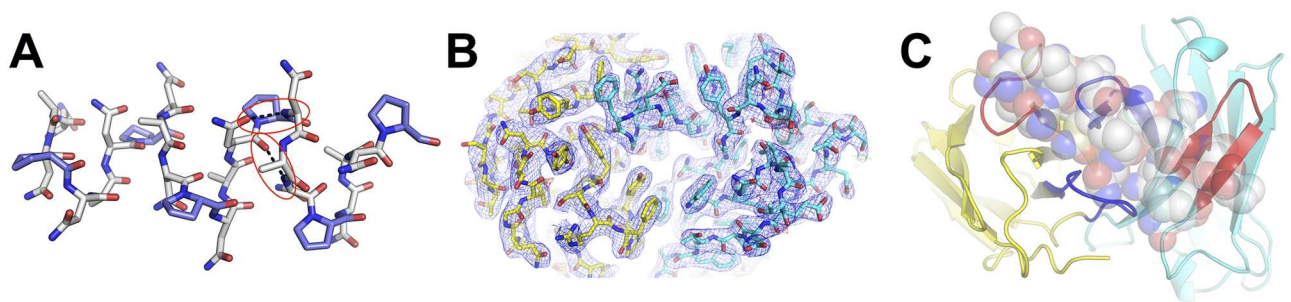


Fig 2. Structures of the (NANP)₆ peptide (A), the 2A10 F_{AB} fragment (B) and the model of the F_{AB} fragment-(NANP)₆ complex (C). (A) The calculated structure of the (NANP)₆ peptide is a helical structure containing the same hydrogen bonds between a carbonyl following the proline and the amide nitrogen of the alanine, and the carbonyl group of an asparagine and a backbone amide of asparagine 3 residues earlier (highlighted in red) that are observed in [29]. (B) Electron density (blue mesh; 2mF_o-dF_c at 1 σ) of the 2A10 F_{AB} fragment viewed from above the antigen-binding site. Light chain is shown as yellow sticks, heavy chain as cyan. (C) A calculated model of the (NANP)₆:2A10 F_{AB} fragment complex. The CDR2 regions of each chain are shown in red, the CDR3 regions of each chain are shown in blue.

<https://doi.org/10.1371/journal.ppat.1006469.g002>

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_{AB}. 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_{AB} and the (NANP)₆ 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_{AB} 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)₆ peptide into the high-solvent form of 2A10 F_{AB}, 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_{AB} 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)₆ peptide complex, the accurate structures of the 2A10 F_{AB} fragment, the (NANP)₆ 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_{AB} 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_{AB}:(NANP)₆ 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)₆ 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_{AB} 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)_n binding.

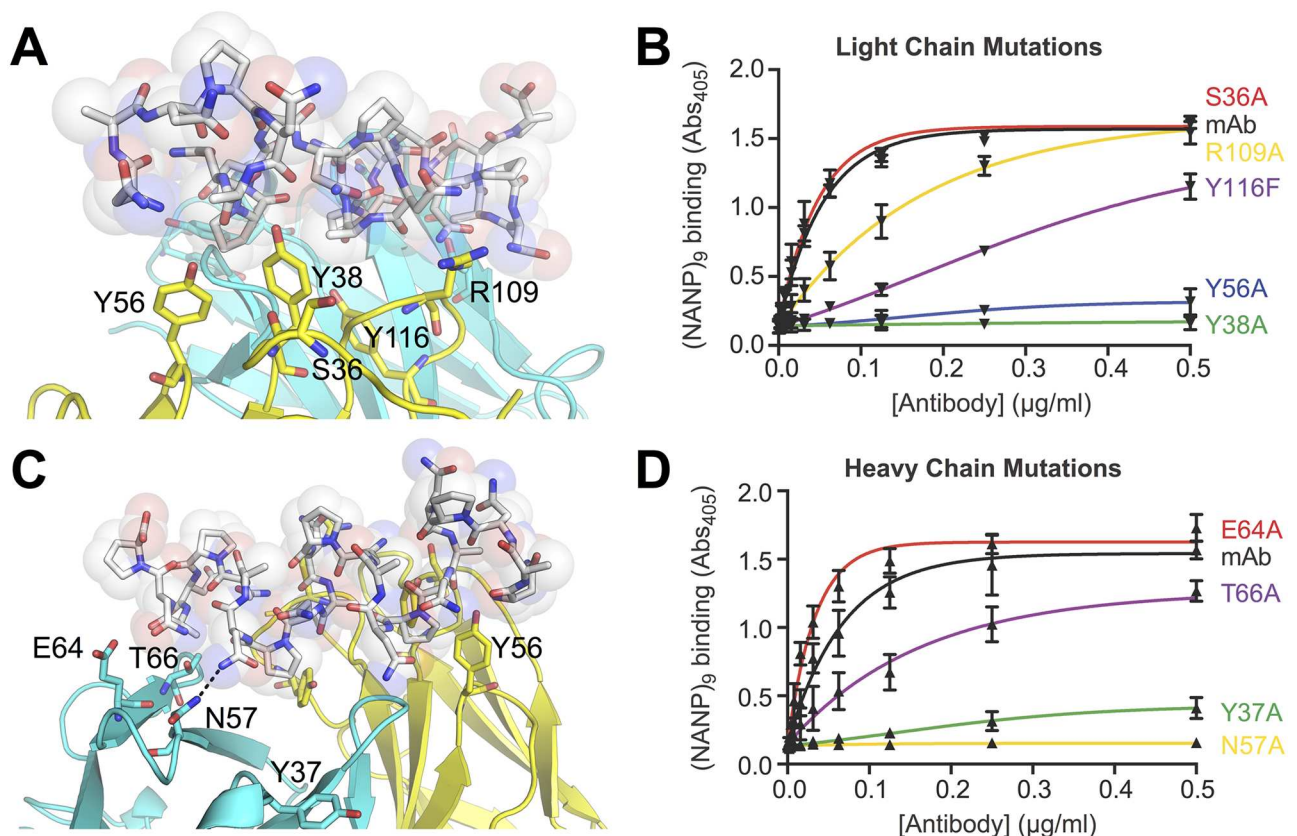


Fig 3. Detailed view of the (NANP)₆:2A10 F_{AB} interface and site directed mutagenesis. (A) A model of the light chain:(NANP)₆ interface. (B) ELISA results showing the effect of mutating light chain interface residues; error bars are based on technical replicates from one of two independent experiments. (C) A model of the heavy chain:(NANP)₆ interface. (D) ELISA results showing the effect of mutating heavy chain interface residues; error bars are based on technical replicates from one of two independent experiments.

<https://doi.org/10.1371/journal.ppat.1006469.g003>

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 location 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,

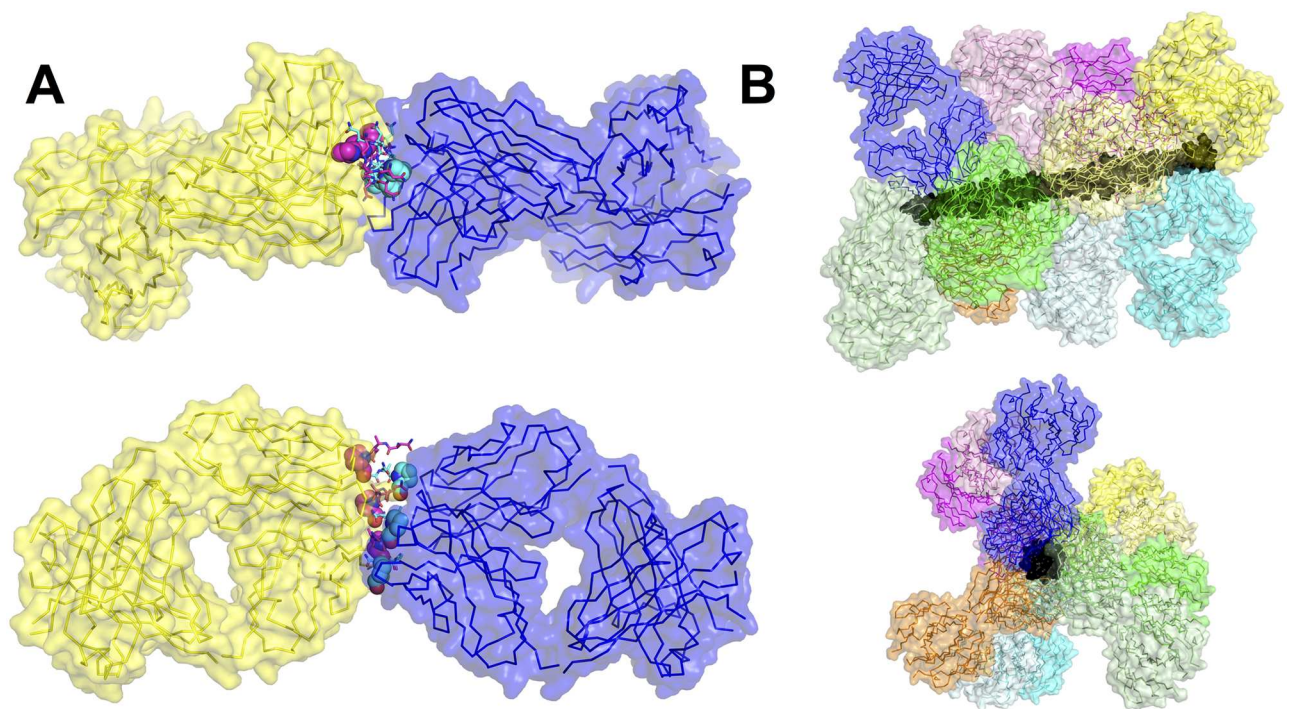


Fig 4. The multivalency of the NANP repeat region of the CSP protein. (A) An (NANP)₆ peptide results in the presentation of two symmetrical epitopes, formed by alternating repeats (cyan and magenta), allowing binding by two F_{AB} domains, in keeping with the stoichiometry observed by ITC. (B) The full 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_{AB} fragments, consistent with 4 antibodies bound by both F_{AB} domains, and two bound by a single F_{AB} domain.

<https://doi.org/10.1371/journal.ppat.1006469.g004>

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_{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_{AB} fragments, or six antibodies (containing 12 F_{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_{AB} fragments (Fig 4B). Six 2A10 antibodies can bind if two antibodies interact by a single F_{AB} domain and the other four interact with both F_{AB} domains. The observation that the F_{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_{AB} fragment binding, have more favorable binding enthalpy, i.e. the additional bonds formed between adjacent F_{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

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^{Pf}) [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^{Pf} or another line of *P. berghei* with a mutant CSP (*P. berghei* CSSM) that contains the endogenous (*P. berghei*) repeat region, which has a distinct repeat sequence (PPPPNPND)_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^{Pf} 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 naïve 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 GC B 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^{Pf} 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]. Interestingly 4 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.

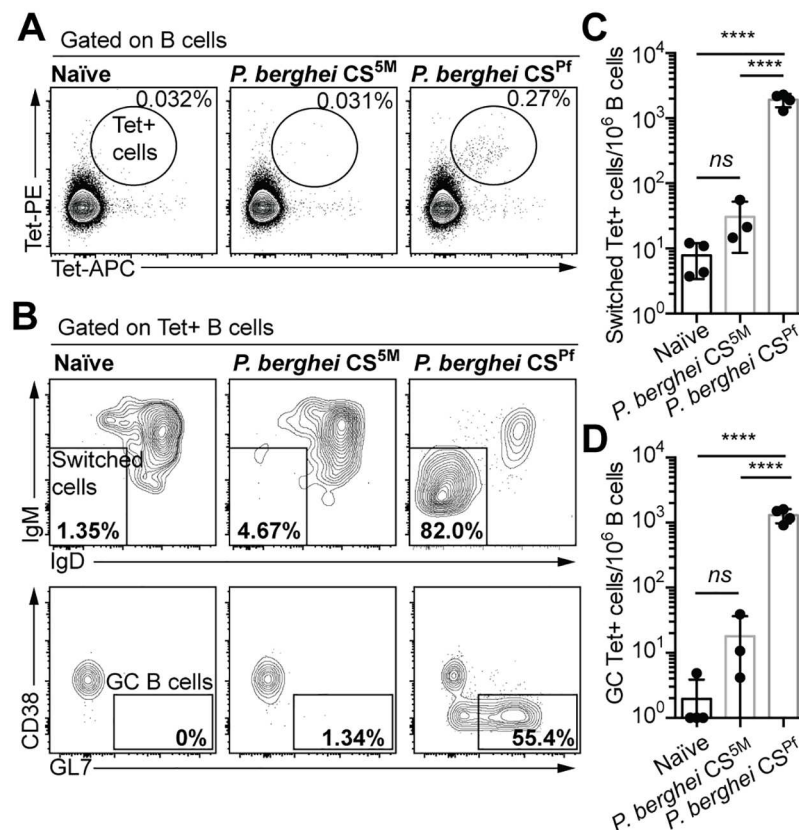


Fig 5. CSP-specific B cells enter the germinal center following sporozoite immunization. BALB/C mice were immunized with either 5×10^4 *P. berghei* CS^{5M} (expressing the endogenous *P. berghei* CSP repeat) or 5×10^4 *P. berghei* CS^{Pf} (expressing the circumsporozoite protein from *P. falciparum*) 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.

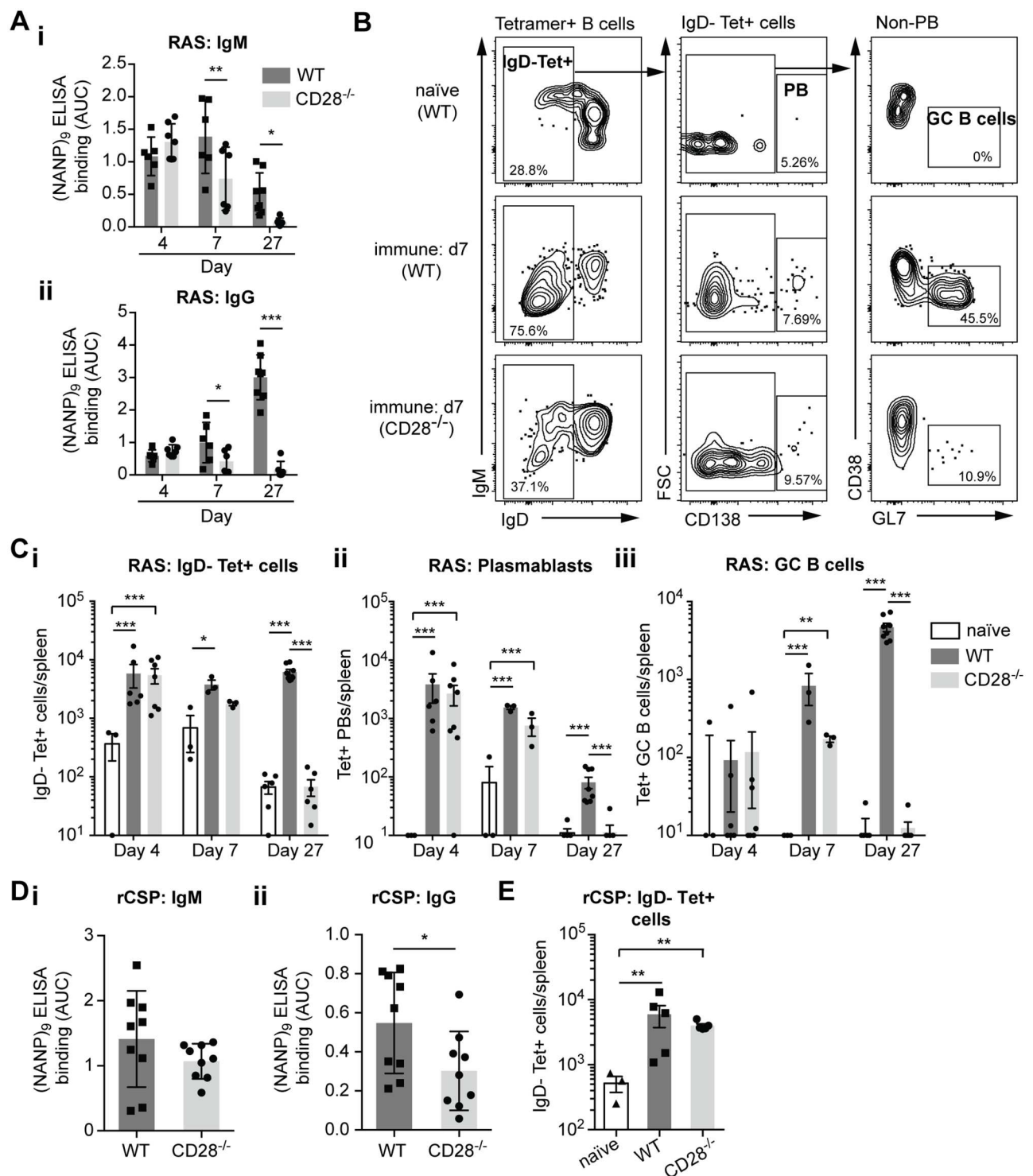


Fig 6. The B cell response to CSP has a T-independent component. CD28^{-/-} and control C57BL/6 mice were immunized with *P. berghei* CS^{Pf} radiation attenuated sporozoites (RAS) or rCSP in alum. Sera were taken and the spleens analyzed for antigen specific B cells using tetramers 4, 7 and 27 days post-immunization. (A) IgM and IgG (NANP)₉ 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⁻ (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.

<https://doi.org/10.1371/journal.ppat.1006469.g006>

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 IGHV1-20, 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

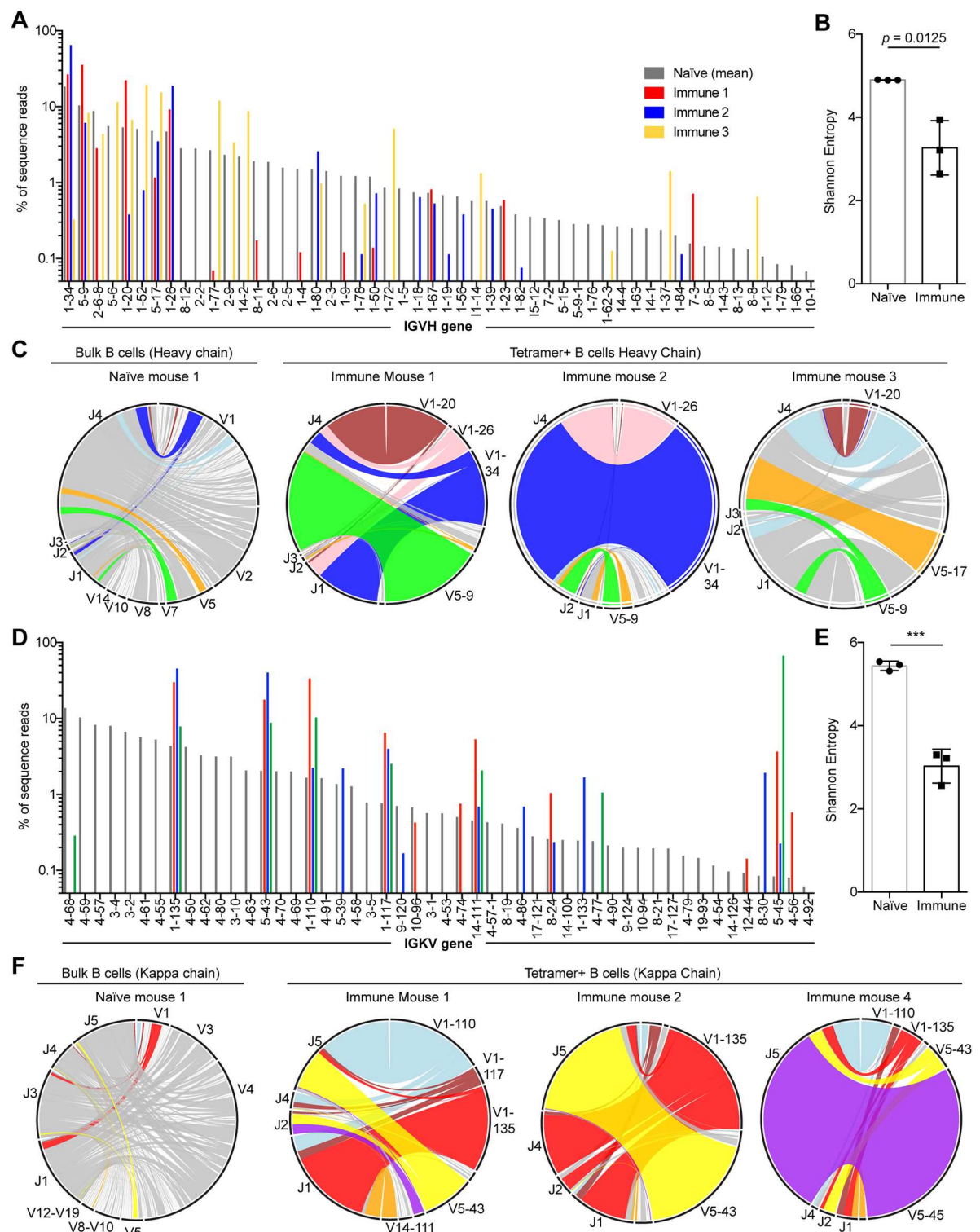


Fig 7. Limited diversity of (NANP)_n specific antibodies. BCR sequences were amplified from Tetramer⁺ cells sorted from BALB/C mice 35 days after immunization with live *P. berghei* CS^{Pf} sporozoites under CQ cover as well as bulk (B220⁺) B cells from naïve BALB/C mice (A) IGHV gene usage from among B cells from a representative naïve 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 naïve mice and 3 immune mice. (D) IGKV gene usage from among B cells from a representative naïve mouse (grey bars) and Tetramer⁺ cells from

immune mice (red, blue and green bars). (E) Shannon's diversity calculated for the diversity of IGKV region usage among bulk B cells and Tetramer⁺ cells. (F) Circos plots showing the IGKV-IGKJ pairings in a representative naïve mouse and 3 immune mice. Statistical analysis of Shannon's diversity index was by Student's T test.

<https://doi.org/10.1371/journal.ppat.1006469.g007>

examined each specific common kappa chain in turn (IGVK1-110; IGKV1-135; IGKV5-43/45) 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

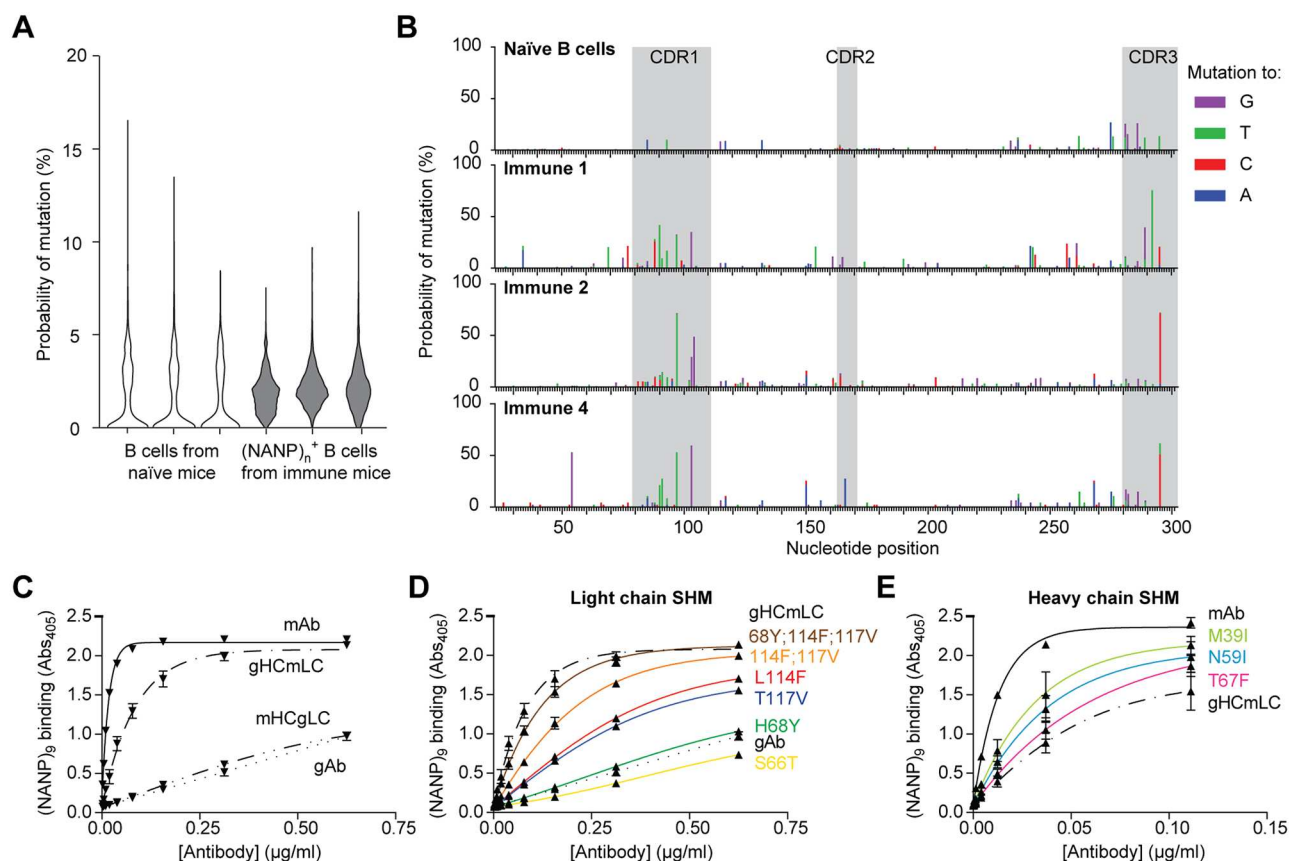


Fig 8. CSP-binding antibodies undergo somatic hypermutation and affinity maturation. (A) Violin plots showing the number of mutations per kappa chain read from bulk B cells from 3 individual naïve mice and sorted (NANP)_n specific B cells from sporozoite immunized mice (B) Skyscraper plots showing the 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)₉ 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)₉ 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)₉ measured by ELISA.

<https://doi.org/10.1371/journal.ppat.1006469.g008>

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

avidity, multivalent interaction. There is also additional enthalpic stabilization (per F_{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]. Moreover F_{AB} fragments alone are sufficient to block invasion [42,43]. However, it is well established 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-terminal 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. Alternatively since the repeat region is directly adjacent to the proteolytic cleavage site, anti-repeat antibodies 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 multivalency 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 sporozoite 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 induction 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 $(NANP)_n$ 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 $(NANP)_n$ 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 titres 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 5×10^4 *P. berghei* CS^{5M} sporozoites expressing mCherry [56] or $5 \times$

10^4 *P. berghei* CS^{Pf} 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, 30ug rCSP was emulsified in Imject 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 100ug 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 Nextera 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 supplemental 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, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$).

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.

(TIF)

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.

(TIF)

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

(TIF)

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* CS^{Pf} RAS, live *P. berghei* CS^{Pf} 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.

(TIF)

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.

(TIF)

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.

(TIF)

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.

Formal analysis: Camilla R. Fisher, Henry J. Sutton, Joe A. Kaczmarek, Ben Clifton, Joshua Mitchell, Mandeep Singh, Aaron Chuah.

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

Investigation: Camilla R. Fisher, Henry J. Sutton, Joe A. Kaczmarek, Hayley A. McNamara, Ben Clifton, Joshua Mitchell, Yeping Cai, Johanna N. Dups, Nicholas J. D'Arcy, Thomas S. Peat.

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.

Visualization: Camilla R. Fisher, Henry J. Sutton, Joe A. Kaczmarek, Ben Clifton, Joshua Mitchell, Colin J. Jackson, Ian A. Cockburn.

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

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

References

1. World Health Organization (2016) World Malaria Report 2016. Geneva: World Health Organization.
2. Casares S, Brumeanu TD, Richie TL (2010) The RTS,S malaria vaccine. *Vaccine* 28: 4880–4894. <https://doi.org/10.1016/j.vaccine.2010.05.033> PMID: 20553771
3. RTS,S Clinical Trials Partnership (2015) Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet* 386: 31–45. [https://doi.org/10.1016/S0140-6736\(15\)60721-8](https://doi.org/10.1016/S0140-6736(15)60721-8) PMID: 25913272
4. White MT, Bejon P, Olotu A, Griffin JT, Riley EM, et al. (2013) The relationship between RTS,S vaccine-induced antibodies, CD4(+) T cell responses and protection against *Plasmodium falciparum* infection. *PLoS One* 8: e61395. <https://doi.org/10.1371/journal.pone.0061395> PMID: 23613845
5. Nussenzweig RS, Vanderberg J, Most H, Orton C (1967) Protective immunity produced by the injection of x-irradiated sporozoites of *Plasmodium berghei*. *Nature* 216: 160–162. PMID: 6057225
6. Seder RA, Chang LJ, Enama ME, Zephir KL, Sarwar UN, et al. (2013) Protection against malaria by intravenous immunization with a nonreplicating sporozoite vaccine. *Science* 341: 1359–1365. <https://doi.org/10.1126/science.1241800> PMID: 23929949
7. Hollingdale MR, Nardin EH, Tharavani S, Schwartz AL, Nussenzweig RS (1984) Inhibition of entry of *Plasmodium falciparum* and *P. vivax* sporozoites into cultured cells; an in vitro assay of protective antibodies. *J Immunol* 132: 909–913. PMID: 6317752
8. Yoshida N, Nussenzweig RS, Potocnjak P, Nussenzweig V, Aikawa M (1980) Hybridoma produces protective antibodies directed against the sporozoite stage of malaria parasite. *Science* 207: 71–73. PMID: 6985745
9. Dame JB, Williams JL, McCutchan TF, Weber JL, Wirtz RA, et al. (1984) Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum*. *Science* 225: 593–599. PMID: 6204383
10. Coppi A, Natarajan R, Pradel G, Bennett BL, James ER, et al. (2011) The malaria circumsporozoite protein has two functional domains, each with distinct roles as sporozoites journey from mosquito to mammalian host. *J Exp Med* 208: 341–356. <https://doi.org/10.1084/jem.20101488> PMID: 21262960
11. Zavala F, Cochrane AH, Nardin EH, Nussenzweig RS, Nussenzweig V (1983) Circumsporozoite proteins of malaria parasites contain a single immunodominant region with two or more identical epitopes. *J Exp Med* 157: 1947–1957. PMID: 6189951
12. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498–511. <https://doi.org/10.1038/nature01097> PMID: 12368864
13. Zeeshan M, Alam MT, Vinayak S, Bora H, Tyagi RK, et al. (2012) Genetic variation in the *Plasmodium falciparum* circumsporozoite protein in India and its relevance to RTS,S malaria vaccine. *PLoS One* 7: e43430. <https://doi.org/10.1371/journal.pone.0043430> PMID: 22912873
14. Espinosa DA, Gutierrez GM, Rojas-Lopez M, Noe AR, Shi L, et al. (2015) Proteolytic Cleavage of the *Plasmodium falciparum* Circumsporozoite Protein Is a Target of Protective Antibodies. *J Infect Dis* 212: 1111–1119. <https://doi.org/10.1093/infdis/jiv154> PMID: 25762791
15. Persson C, Oliveira GA, Sultan AA, Bhanot P, Nussenzweig V, et al. (2002) Cutting edge: a new tool to evaluate human pre-erythrocytic malaria vaccines: rodent parasites bearing a hybrid *Plasmodium falciparum* circumsporozoite protein. *J Immunol* 169: 6681–6685. PMID: 12471098
16. Schofield L, Uadia P (1990) Lack of Ir-Gene Control in the Immune-Response to Malaria. 1. A Thymus-Independent Antibody-Response to the Repetitive Surface Protein of Sporozoites. *Journal of Immunology* 144: 2781–2788.
17. Defrance T, Taillardet M, Genestier L (2011) T cell-independent B cell memory. *Current Opinion in Immunology* 23: 330–336. <https://doi.org/10.1016/j.coi.2011.03.004> PMID: 21482090
18. Bachmann MF, Hengartner H, Zinkernagel RM (1995) T helper cell-independent neutralizing B cell response against vesicular stomatitis virus: role of antigen patterns in B cell induction? *Eur J Immunol* 25: 3445–3451. <https://doi.org/10.1002/eji.1830251236> PMID: 8566036
19. Schodel F, Peterson D, Zheng J, Jones JE, Hughes JL, et al. (1993) Structure of hepatitis B virus core and e-antigen. A single precore amino acid prevents nucleocapsid assembly. *J Biol Chem* 268: 1332–1337. PMID: 8419335

20. Anker R, Zavala F, Pollok BA (1990) VH and VL region structure of antibodies that recognize the (NANP)₃ dodecapeptide sequence in the circumsporozoite protein of *Plasmodium falciparum*. *Eur J Immunol* 20: 2757–2761. <https://doi.org/10.1002/eji.1830201233> PMID: 2125276
21. Foquet L, Hermesen CC, van Gemert GJ, Van Braeckel E, Weening KE, et al. (2014) Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent *Plasmodium falciparum* infection. *J Clin Invest* 124: 140–144. <https://doi.org/10.1172/JCI70349> PMID: 24292709
22. Chappel JA, Rogers WO, Hoffman SL, Kang AS (2004) Molecular dissection of the human antibody response to the structural repeat epitope of *Plasmodium falciparum* sporozoite from a protected donor. *Malar J* 3: 28. <https://doi.org/10.1186/1475-2875-3-28> PMID: 15283866
23. Zavala F, Tam JP, Hollingdale MR, Cochrane AH, Quakyi I, et al. (1985) Rationale for development of a synthetic vaccine against *Plasmodium falciparum* malaria. *Science* 228: 1436–1440. PMID: 2409595
24. Cerami C, Frevert U, Sinnis P, Takacs B, Clavijo P, et al. (1992) The basolateral domain of the hepatocyte plasma membrane bears receptors for the circumsporozoite protein of *Plasmodium falciparum* sporozoites. *Cell* 70: 1021–1033. PMID: 1326407
25. Braden BC, Poljak RJ (1995) Structural features of the reactions between antibodies and protein antigens. *FASEB J* 9: 9–16. PMID: 7821765
26. Plassmeyer ML, Reiter K, Shimp RL Jr., Kotova S, Smith PD, et al. (2009) Structure of the *Plasmodium falciparum* circumsporozoite protein, a leading malaria vaccine candidate. *J Biol Chem* 284: 26951–26963. <https://doi.org/10.1074/jbc.M109.013706> PMID: 19633296
27. Kelly SM, Jess TJ, Price NC (2005) How to study proteins by circular dichroism. *Biochim Biophys Acta* 1751: 119–139. <https://doi.org/10.1016/j.bbapap.2005.06.005> PMID: 16027053
28. Shen Y, Maupetit J, Derreumaux P, Tuffery P (2014) Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction. *J Chem Theory Comput* 10: 4745–4758. <https://doi.org/10.1021/ct500592m> PMID: 26588162
29. Ghasparian A, Moehle K, Linden A, Robinson JA (2006) Crystal structure of an NPNA-repeat motif from the circumsporozoite protein of the malaria parasite *Plasmodium falciparum*. *Chem Commun (Camb)*: 174–176.
30. Sela-Culang I, Alon S, Ofra Y (2012) A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J Immunol* 189: 4890–4899. <https://doi.org/10.4049/jimmunol.1201493> PMID: 23066154
31. Sircar A, Gray JJ (2010) SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol* 6: e1000644. <https://doi.org/10.1371/journal.pcbi.1000644> PMID: 20098500
32. Davis SJ, Davies EA, Tucknott MG, Jones EY, van der Merwe PA (1998) The role of charged residues mediating low affinity protein-protein recognition at the cell surface by CD2. *Proc Natl Acad Sci U S A* 95: 5490–5494. PMID: 9576909
33. Pape KA, Taylor JJ, Maul RW, Gearhart PJ, Jenkins MK (2011) Different B cell populations mediate early and late memory during an endogenous immune response. *Science* 331: 1203–1207. <https://doi.org/10.1126/science.1201730> PMID: 21310965
34. Keitany GJ, Kim KS, Krishnamurthy AT, Hondowicz BD, Hahn WO, et al. (2016) Blood Stage Malaria Disrupts Humoral Immunity to the Pre-erythrocytic Stage Circumsporozoite Protein. *Cell Rep* 17: 3193–3205. <https://doi.org/10.1016/j.celrep.2016.11.060> PMID: 28009289
35. de Vinuesa CG, Cook MC, Ball J, Drew M, Sunners Y, et al. (2000) Germinal centers without T cells. *J Exp Med* 191: 485–494. PMID: 10662794
36. Ferguson SE, Han S, Kelsoe G, Thompson CB (1996) CD28 is required for germinal center formation. *J Immunol* 156: 4576–4581. PMID: 8648099
37. Goronzy J, Weyand CM, Fathman CG (1986) Long-term humoral unresponsiveness in vivo, induced by treatment with monoclonal antibody against L3T4. *J Exp Med* 164: 911–925. PMID: 3091757
38. Schofield L (1990) The circumsporozoite protein of *Plasmodium*: a mechanism of immune evasion by the malaria parasite? *Bull World Health Organ* 68 Suppl: 66–73.
39. Arnaut R, Lee W, Cahill P, Honan T, Sparrow T, et al. (2011) High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* 6: e22365. <https://doi.org/10.1371/journal.pone.0022365> PMID: 21829618
40. Busse CE, Czogiel I, Braun P, Arndt PF, Wardemann H (2014) Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur J Immunol* 44: 597–603. <https://doi.org/10.1002/eji.201343917> PMID: 24114719
41. Brochet X, Lefranc MP, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 36: W503–508. <https://doi.org/10.1093/nar/gkn316> PMID: 18503082

42. Potocnjak P, Yoshida N, Nussenzweig RS, Nussenzweig V (1980) Monovalent fragments (Fab) of monoclonal antibodies to a sporozoite surface antigen (Pb44) protect mice against malarial infection. *J Exp Med* 151: 1504–1513. PMID: [6991628](#)
43. Stewart MJ, Nawrot RJ, Schulman S, Vanderberg JP (1986) *Plasmodium berghei* sporozoite invasion is blocked in vitro by sporozoite-immobilizing antibodies. *Infect Immun* 51: 859–864. PMID: [3512436](#)
44. Bouharoun-Tayoun H, Attanath P, Sabchareon A, Chongsuphajaisiddhi T, Druilhe P (1990) Antibodies that protect humans against *Plasmodium falciparum* blood stages do not on their own inhibit parasite growth and invasion in vitro, but act in cooperation with monocytes. *J Exp Med* 172: 1633–1641. PMID: [2258697](#)
45. Boyle MJ, Reiling L, Feng G, Langer C, Osier FH, et al. (2015) Human antibodies fix complement to inhibit *Plasmodium falciparum* invasion of erythrocytes and are associated with protection against malaria. *Immunity* 42: 580–590. <https://doi.org/10.1016/j.immuni.2015.02.012> PMID: [25786180](#)
46. White MT, Bejon P, Olotu A, Griffin JT, Bojang K, et al. (2014) A combined analysis of immunogenicity, antibody kinetics and vaccine efficacy from phase 2 trials of the RTS,S malaria vaccine. *BMC Med* 12: 117. <https://doi.org/10.1186/s12916-014-0117-2> PMID: [25012228](#)
47. Avent ND, Madgett TE, Lee ZE, Head DJ, Maddocks DG, et al. (2006) Molecular biology of Rh proteins and relevance to molecular medicine. *Expert Rev Mol Med* 8: 1–20.
48. Chang TY, Siegel DL (1998) Genetic and immunological properties of phage-displayed human anti-Rh (D) antibodies: implications for Rh(D) epitope topology. *Blood* 91: 3066–3078. PMID: [9531621](#)
49. Finne J, Leinonen M, Makela PH (1983) Antigenic similarities between brain components and bacteria causing meningitis. Implications for vaccine development and pathogenesis. *Lancet* 2: 355–357. PMID: [6135869](#)
50. North B, Lehmann A, Dunbrack RL (2011) A New Clustering of Antibody CDR Loop Conformations. *Journal of Molecular Biology* 406: 228–256. <https://doi.org/10.1016/j.jmb.2010.10.030> PMID: [21035459](#)
51. Stanfield RL, Wilson IA (2014) Antibody Structure. *Microbiology spectrum* 2.
52. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, et al. (2015) IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* 43: D413–422. <https://doi.org/10.1093/nar/gku1056> PMID: [25378316](#)
53. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, et al. (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* 184: 6986–6992. <https://doi.org/10.4049/jimmunol.1000445> PMID: [20495067](#)
54. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, et al. (2012) The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol* 188: 1333–1340. <https://doi.org/10.4049/jimmunol.1102097> PMID: [22205028](#)
55. Shahinian A, Pfeffer K, Lee KP, Kundig TM, Kishihara K, et al. (1993) Differential T cell costimulatory requirements in CD28-deficient mice. *Science* 261: 609–612. PMID: [7688139](#)
56. Cockburn IA, Tse SW, Zavala F (2014) CD8+ T cells eliminate liver stage *Plasmodium* parasites without detectable bystander effect. *Infect Immun* 82: 1460–1464. <https://doi.org/10.1128/IAI.01500-13> PMID: [24421043](#)
57. Bulheller BM, Hirst JD (2009) DichroCalc—circular and linear dichroism online. *Bioinformatics* 25: 539–540. <https://doi.org/10.1093/bioinformatics/btp016> PMID: [19129206](#)

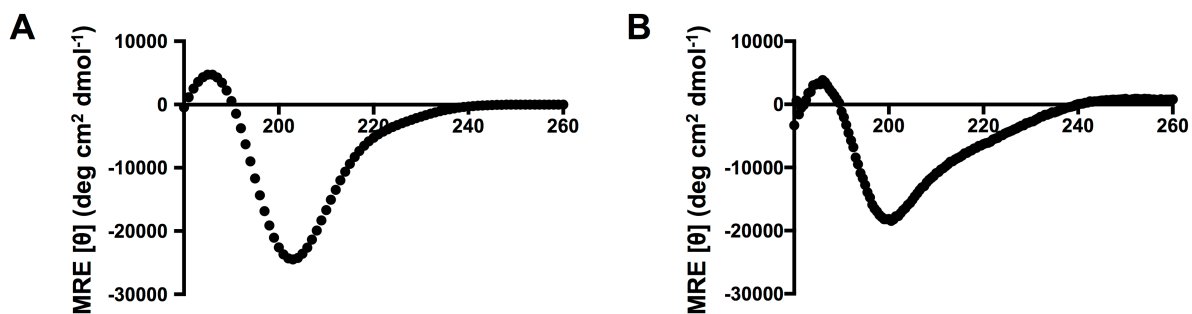


Figure 6.2: Theoretical (A) and experimental (B) CD spectra of the (NANP)₆ peptide. The computational prediction of the spectra (A) was performed using DichroCalc [57], the experimental spectra was measured at 222 nm at 25°C. A peak at 185 nm, minimum at 205 nm and shoulder between 215 and 240 nm are consistent with an intrinsically disordered, but not random coil, structure.

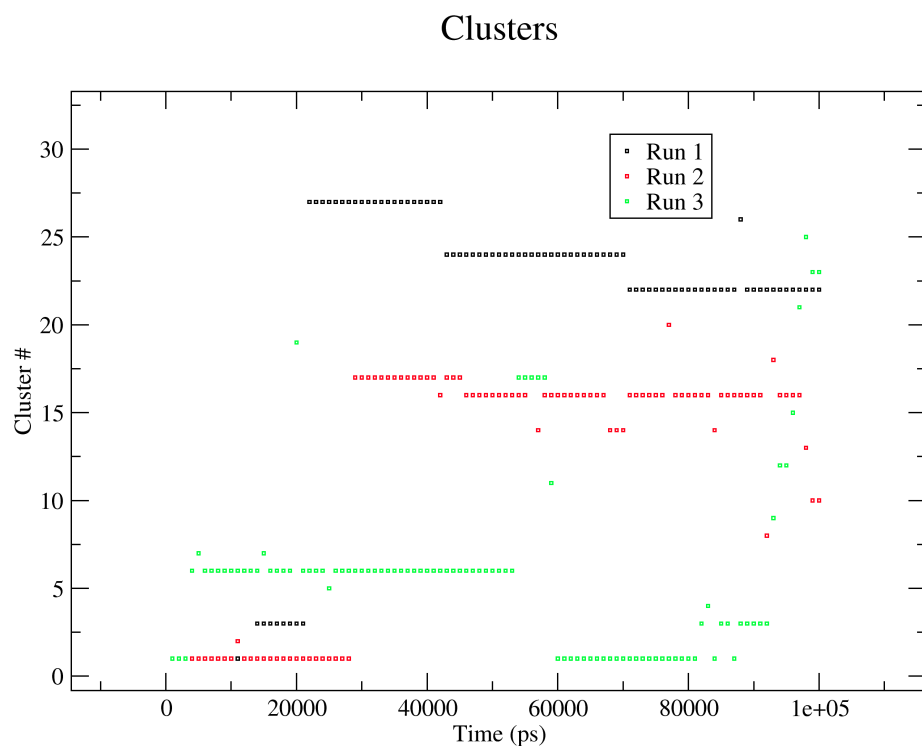


Figure 6.3: Cluster analysis for MD simulations of (NANP)₆ peptide. Conformations were clustered by concatenating the trajectory and performing a Jarvis-Patrick analysis. The clusters are sorted by their RMSD from the first cluster (starting geometry). As shown, Run 2 is stable in the starting geometry 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.

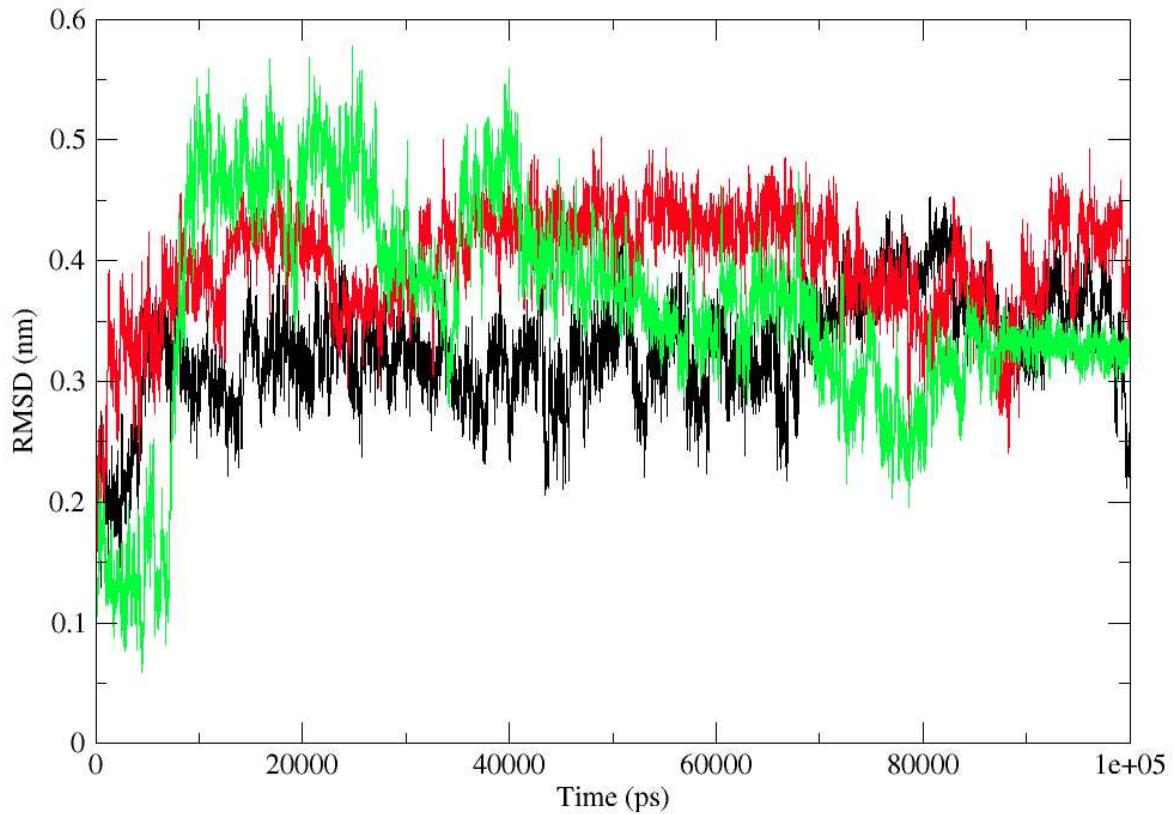


Figure 6.4: Molecular dynamics simulation of the (NANP)₆:Fab complex. Root mean square deviation (RMSD) of the (NANP)₆:Fab complex as a function of time. Independent simulations are shown in green, black and red.

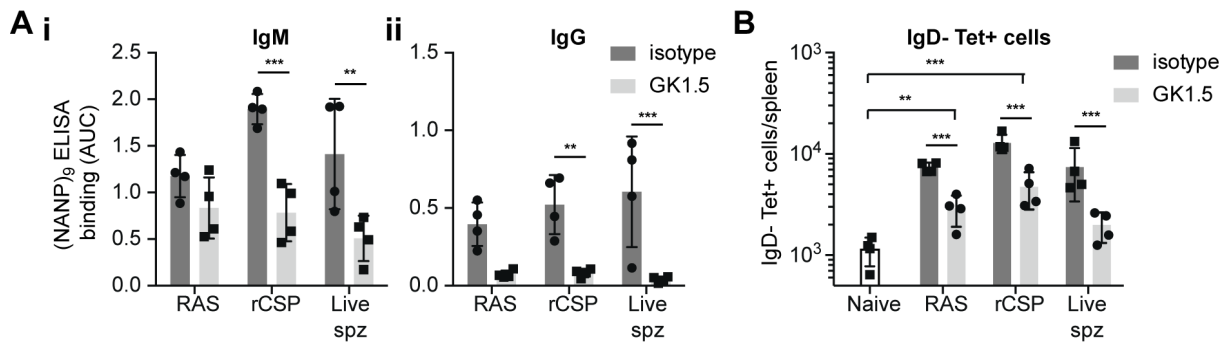


Figure 6.5: The B cell response to CSP has a T-independent component. Mice either treated with an anti-CD4 depleting antibody or an isotype control were immunized with either *P. berghei* CS^{Pf} RAS, live *P. berghei* CS^{Pf} under CQ cover or rCSP. (A) 4 days later the IgM and IgG response to the (NANP)_n 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.

	<----- FR1 - IMGT ----->																			
	1			5				10					15					20		
2A10 heavy chain	Q	I	Q	L	V	Q	S	G	P	E	L	K	K	P	G	E	T	V	K	
	cag	atc	cag	ttg	gtg	cag	tct	gga	cct	...	gag	ctg	aag	aag	cct	gga	gag	aca	gtc	aag
IGHV9-3*02	-----																			
2E7 heavy chain	-----																			
3D6 heavy chain	-----																			
	<----- CDR1 - IMGT ----->																			
				25				30					35					40		
2A10 heavy chain	I	S	C	K	A	S	G	Y	T	F	T	N	Y	G	I	N
	atc	tcc	tgc	aag	gct	tct	ggg	tat	acc	ttc	aca	aac	tat	gga	ata	aac
IGHV9-3*02	-----																			
2E7 heavy chain	-----																			
3D6 heavy chain	-----																			
	<----- FR2 - IMGT ----->																			
				45				50					55					60		
2A10 heavy chain	W	V	K	Q	A	P	G	K	G	L	K	W	M	G	W	I	N	T	I	...
	tgg	gtg	aag	cag	gct	cca	gga	aag	ggt	tta	aaa	tgg	atg	ggc	tgg	ata	aac	acc	atc	...
IGHV9-3*02	-----																			
2E7 heavy chain	-----																			
3D6 heavy chain	-----																			
	<----- IMGT ----->																			
				65				70					75					80		
2A10 heavy chain	T	E	E	P	T	F	A	E	E	F	T	G	R	F	A	F	S	L		
	...	act	gaa	gag	cca	acg	ttt	gct	gaa	gaa	ttc	acg	...	gga	cgg	ttt	gcc	ttc	tct	ttg
IGHV9-3*02	-----																			
2E7 heavy chain	-----																			
3D6 heavy chain	-----																			
	<----- FR3 - IMGT ----->																			
				85				90					95					100		
2A10 heavy chain	E	T	S	A	S	T	A	Y	L	Q	I	N	N	L	K	N	E	D	T	A
	gaa	acc	tct	gcc	agc	act	gcc	tat	ttg	cag	atc	aac	aac	ctc	aaa	aat	gag	gac	acg	gct
IGHV9-3*02	-----																			
2E7 heavy chain	-----																			
3D6 heavy chain	-----																			
	<----- CDR3 - IMGT ----->																			
				105				110					115							
2A10 heavy chain	T	Y	F	C	A	R	G	S	E	F	G	R	L	V	Y	W		
	aca	tat	ttc	tgt	gca	aga	gga	agt	gaa	ttt	ggg	cgt	ttg	gtc	tac	tgg		
IGHV9-3*02	-----																			
2E7 heavy chain	-----																			
3D6 heavy chain	-----																			

Figure 6.6: 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.

	<-----															FR1 - IMGT			----->																
	1															5			10			15			20										
2A10 light_chain	D	I	Q	M	T	Q	T	T	S	S	L	S	A	S	L	G	D	R	V	T															
	gat	atc	cag	atg	aca	cag	act	aca	tcc	tcc	ctg	tct	gcc	tct	ctg	gac	aga	gtc	acc	atc															
IGKV10-94*01	-----																																		
2E7 light chain	-----																																		
	----->															CDR1 - IMGT										-----<									
	25															30					35					40									
2A10_light_chain	I	S	C	S	A	S	Q	G	I					S	N	Y	L	N											
	gga	agt	tgc	agt	gca	agt	cag	ggc	att					agc	aat	tat	tta	aac											
IGKV10-94*01	-----																																		
2E7 light chain	-----																																		
	-----															FR2 - IMGT										----->					CDR2				
	45															50					55					60									
2A10_light_chain	W	Y	Q	Q	K	P	D	G	T	V	K	L	L	I	F	Y	T	...																	
	tgg	tat	cag	cag	aaa	cca	gat	gga	act	gtt	aaa	ctc	ctg	atc	ttt	tac	aca	...																	
IGKV10-94*01	-----															-----					-----					-----									
2E7 light chain	-----															-----					-----					-----									
	-----															FR3 - IMGT										----->					CDR3 - IMGT				
	65															70					75					80									
2A10_light_chain	...					S	T	L	Y	S	G	V	P	...					S	R	F	S	G	S	G										
	...					S	H																	
IGKV10-94*01	-----					-----					-----					-----					-----														
2E7 light chain	-----					-----					-----					-----					-----														
	-----															FR3 - IMGT										----->					CDR3 - IMGT				
	85															90					95					100									
2A10_light_chain	...					S	G	T	D	Y	S	L	T	I	S	N	L	E	P	E	D	I	A												
	...					tct	ggg	aca	gat	tat	tct	ctc	acc	atc	agc	aac	ctg	gaa	cct	gaa	gat	att	gcc												
IGKV10-94*01	-----																																		
2E7 light chain	-----																																		
	----->															CDR3 - IMGT										----->									
	105															110					115														
2A10_light_chain	T	Y	Y	C	Q	Q	Y	S	R					F	P	Y	V	F											
	act	tac	tat	tgt	cag	cag	tat	agt	agg	K					ttt	ccg	tac	gtg	ttc										
IGKV10-94*01	-----															-----					-----					-----									
2E7 light chain	-----															-----					-----					-----									

Figure 6.7: Alignment of 2A10 light 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.

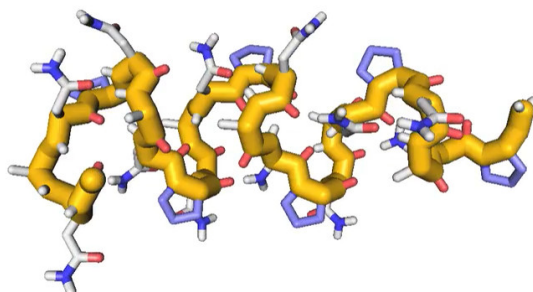


Figure 6.8: Molecular dynamics simulation of the solution structure of the (NANP)₆ peptide. Excerpt from (NANP)₆ run 3. The trajectory was fitted to minimize alpha-carbon RMSD and then passed through a low-pass filter with a filter length of 8 frames to reduce temporal aliasing.

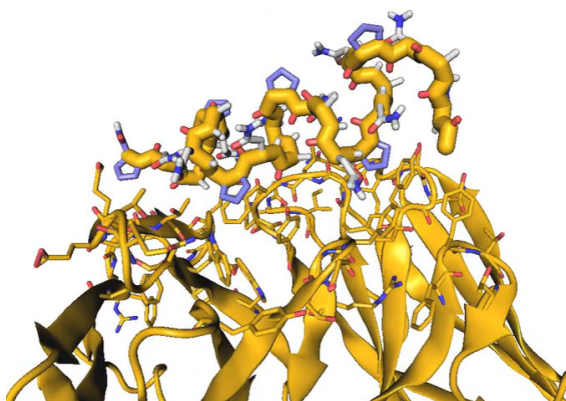


Figure 6.9: Molecular dynamics simulation of the interaction of the (NANP)_n repeat with the 2A10 Fab. Excerpt from 2A10:(NANP)₆ run 3. The trajectory was fitted to minimize alpha-carbon RMSD and then passed through a low-pass filter with a filter length 8 frames to reduce temporal aliasing.

Table 6.1: Data collection and refinement statistics for the crystal structures of 2A10 Fab presented in this work.

2A10 Fab 2.52 2A10	Fab 3.01	
PDB ID	5SZF	5T0Y
Data collection		
Space group	I4 ₁ 32	P432 ₁ 2
Cell dimensions		
a, b, c (Å)	204.21	231.68, 231.68, 81.78
α , β , γ (°)	90	90
Resolution (Å)	37.28–2.52	39.73–3.01
	(2.62–2.52)*	(3.12–3.01)*
R _{merge}	0.178 (1.836)*	0.315 (1.913)*
R _{pim}	0.028 (0.406)*	0.085 (0.517)*
CC _{1/2}	0.999 (0.620)*	0.991 (0.718)*
Completeness (%)	100 (100)*	99.9 (99.5)*
Redundancy	39.2 (22.0)*	14.6 (14.6)*
Refinement		
Resolution (Å)	37.28–2.52	39.73–3.01
	(2.61–2.52)*	(3.12–3.01)*
No. reflections	24 796 (2 440)*	44 596 (4 352)*
R _{work} / R _{free}	0.2251/0.2483	0.2248/0.2467
No. atoms		
Protein	3 288	9 856
Ligand/ion	20	90
Water	46	25
Wilson B-factor	46.80	60.66
R.m.s. deviations		
Bond lengths (Å)	0.003	0.002
Bond angles (°)	0.57	0.55
Ramachandran favored (%)	95	95
Ramachandran outliers (%)	0	0

* Values in parentheses are for highest-resolution shell.

Table 6.2: Heavy chain CDR sequences of CSP binding antibodies

Antibody	Species	Heavy chain	CDR1*	CDR1 grp	CDR2*	CDR2 grp	CDR3
2A10	Mouse	HV9-3*02	KASG Y TF TN Y GIN	H1-13-1	W I N T I . . TEEPT T	H2-10-1	ARGSEFGRLVY
PfNPNAI	Human	HV3-30-3	AASGFTF SSYAMH	H1-13-1	VISYD . . GSNKY	H2-10-2	DRDSSSYFDS
3D6	Mouse	HV9-2-1*01	KASGSPF PDSSMP	H1-13-1	WINTA . . TGEPT	H2-10-1	GGGGGPWFAY
2C11	Mouse	HV1-20 OR 37	KASGYSF TGSFMN	H1-13-1	RINPN . . DGYTF	H2-10-1	GKGNHGATDY
1E9	Mouse	HV1-20 OR 37	KASGYSF TGSFMN	H1-13-1	RILPY . . NGDTF	H2-10-1	GYVYDGGYATDY
	Mouse	HV5-9	AASGFTF SSYTMS	H1-13-1	TISSG . . GGNTY	H2-10-1	Variable
	Mouse	HV1-20	KASGYSF TGYFMN	H1-13-1	RINPY . . NGDTF	H2-10-1	Variable
	Mouse	HV1-26	KASGYTF TDYYMN	H1-13-1	DINPN . . NGGTS	H2-10-1	Variable
	Mouse	HV1-34	KASGYTF TDYYMH	H1-13-1	YIYPN . . NGGNG	H2-10-1	Variable

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

Table 6.3: Light chain CDR sequences of CSP binding immunoglobulins

Antibody	Species	Heavy chain	CDR1*	CDR1 grp	CDR2*	CDR2 grp	CDR3*	CDR3 grp
2A10	Mouse	KV10-94	SASQGI SN Y LN	L1-11-2	F Y TSTL Y S	L2-8-1	QQYS RFPYV	L3-9-cis7-1
PfNPNAI	Human	KV1-5	RASQSI SSWLA	L1-11-2	YDASSLES	L2-8-1	QQYNSYSGLT	L3-10-1
3D6	Mouse	KV6-20	KASENV VTYVS	L1-11-1	YRASNRYT	L2-8-1	GQGSSYPYT	L3-9-cis7-1
2C11	Mouse	KV5-43	RASQNI SNNLH	L1-11-1	TYASQSIG	L2-8-1	QQSNSWPLT	L3-9-cis7-1
1E9	Mouse	KV1-110	RSSQSLGHS . HGNTYLH	L1-16-1	YKVSNRFS	L2-8-1	SQSTQLRT	L3-8-1
High Throughput Sequencing	Mouse	KV1-117	RSSQSIVHS . NGNTYLE	L1-16-1	YKVSNRFS	L2-8-1	FQGSHPPTF	L3-9-cis7-1
	Mouse	KV1-110	RSSQSLVHS . NGNTYLH	L1-16-1	YKVSNRFS	L2-8-1	SQSTHLPLTF	L3-9-cis7-1
	Mouse	KV1-135	KSSQSLIDS . DGKTYLN	L1-16-1	YLVSKIDS	L2-8-1	WQGTDFPPTF	L3-9-cis7-1
	Mouse	KV5-43	RASQSI SNNLH	L1-11-1	KYASQSIG	L2-8-1	QQSNSWPLTF	L3-9-cis7-1
	Mouse	KV5-45	RASQSI SNYLH	L1-11-1	KYASQSIG	L2-8-1	QQSNSWPLTF	L3-9-cis7-1
	Mouse	KV14-111	KASQDI NSYLS	L1-11-1	YRANRLVD	L2-8-1	LQYGEFPPTF	L3-9-cis7-1

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

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 QuikChange 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 1ug/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)_n 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 than 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 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 were built based on 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_{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 *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_{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_{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_{AB} -(NANP)₆ titration, 50 μ M 2A10 F_{AB} was titrated with 1 \times 1.2 μ L, then 20 \times 2.0 μ L injections of 400 μ M (NANP)₆. For the F_{AB} -CSP titration, 8.1 μ M 2A10 F_{AB} was titrated with 1 \times 1.2 μ L, then 20 \times 2.0 μ L injections of 5.9 μ M CSP. For the 2A10-CSP titration, 8.8 μ M 2A10 was titrated with 1 \times 1.2 μ L, then 28 \times 1.5 μ L injections of 9.0 μ M 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_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 ΔH 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 $P4_32_12$ 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 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 $1000 \text{ kJ mol}^{-1} \text{ nm}^{-1}$. The position restraints were relaxed over a series of 5 further 1 ns equilibrations with restraints of 500, 100, 50, 10 and $0 \text{ kJ mol}^{-1} \text{ nm}^{-1}$. 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

1. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, et al. (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* 12: 380-381.
2. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, et al. (2015) tcR: an R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics* 16: 175.
3. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40: 658-674.
4. Oomen CJ, Hoogerhout P, Kuipers B, Vidarsson G, van Alphen L, et al. (2005) Crystal structure of an Anti-meningococcal subtype P1.4 PorA antibody provides basis for peptide-vaccine design. *J Mol Biol* 351: 1070-1080.
5. Braden BC, Goldbaum FA, Chen BX, Kirschner AN, Wilson SR, et al. (2000) X-ray crystal structure of an anti-Buckminsterfullerene antibody fab fragment: biomolecular recognition of C(60). *Proc Natl Acad Sci U S A* 97: 12193-12197.
6. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66: 486-501.
7. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, et al. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* 68: 352-367.
8. Shen Y, Maupetit J, Derreumaux P, Tuffery P (2014) Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction. *J Chem Theory Comput* 10: 4745-4758.
9. Bulheller BM, Hirst JD (2009) DichroCalc--circular and linear dichroism online. *Bioinformatics* 25: 539-540.
10. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res* 34: W310-314.
11. Sircar A, Gray JJ (2010) SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol* 6: e1000644.
12. Abraham MJ, Murtola T, Schulz R, Pall S, Smith JC, et al. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2: 19-25.
13. Schmid N, Eichenberger AP, Choutko A, Riniker S, Winger M, et al. (2011) Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur Biophys J* 40: 843-856.

Chapter 7

Multiscale Molecular Dynamics simulations of fusion proteins

7.1 Introduction

The construction of fusion proteins is a foundational technique in protein engineering. By combining the sequences of different proteins into a single construct, designers can co-localise, co-express, and even co-crystallise proteins simply and efficiently, can stabilise and solubilise problematic proteins, and can even produce sensors from inert components. Despite this, structural information about chimeric proteins is limited. Fusion proteins are generally very flexible and therefore resist crystallisation, and computer modelling of these proteins is in its infancy.³¹¹

While linkers are present in both natural and even the earliest synthetic fusion proteins, selecting a linker for a chimeric protein is typically somewhat ad-hoc. A typically glycine-rich repeat sequence is chosen, and a variety of constructs are expressed with varying lengths and compositions. Occasionally, an alanine-rich helix-forming peptide is chosen as a rigid linker, or a loop sequence is taken from another protein. Despite the absence of understanding or best practices, fusion proteins continue to be successfully produced, suggesting that for most applications the linker plays a modest role in the function of a fusion protein. However, the linker length and composition is known to have a profound impact on the function of sensors, where the precise dynamics and structure of the chimeric construct is supremely relevant. Despite this, linker design in sensors remains little more than guesswork.

To this end, in this chapter I attempt a detailed model of the ECFP-AYW fusion protein I developed in my Honours thesis.²⁶⁹ This involves a fusion of svECFP, the cyan fluorescent protein variant lacking surface-accessible cysteine residues described in section 3.2, and AYW, the glycine binding protein described in chapter 4. This fusion protein can be converted into an effective sensor by introducing a mutation to cysteine and labelling it with a maleimide dye (see section 3.2), but the linker has not been optimised or modelled.

Important recent work on the modelling of long linkers borrows concepts from polymer science to describe these peptides as either worm-like or Gaussian chains.²⁸¹ However, these techniques fail for linkers shorter than several dozen repeats. Fusion proteins generally consist of folded proteins

connected by relatively disordered linkers. The size of these proteins alone is prohibitive for atomistic sampling, and their flexibility and irregular proportions imply a need for an especially large box size. Coarse-graining with a force field like MARTINI improves the sampling situation, but at the cost of fidelity, especially for the disordered linker. Accurate computer modelling of disordered peptides has only recently started to become accessible, and their efficient sampling relies on a focussed approach.

By limiting our goals to a qualitative understanding of the fusion protein's dynamics, we can make use of a modified MARTINI force field to provide a reasonable description of the protein–protein interactions.^{126,312} However, even atomistic force fields are only now beginning to approach anything that could be described as accurate for disordered peptides.^{18,72} This suggests a multi-scale simulation of some sort, treating the protein–protein interactions with MARTINI and the linker with an atomistic force field.

Multi-scale approaches involve treating different parts of a simulation system at different levels of theory depending on how much detail is needed from the simulation. This allows computational expense to be focused on the parts of system that need them most without neglecting the context supplied by other parts of the environment. A thermostat (see section 1.4.6) might be considered a basic form of multi-scaling, in which the solute and a few layers of water are treated atomistically while the bulk solvent is treated as a heat bath that occasionally collides with the atomistic system. QM/MM methods,^{313,314} in which a quantum mechanical simulation of, say, an enzyme's active site is embedded in a larger molecular mechanical simulation, are a well-established multi-scale method, but one that demonstrates the enduring difficulty of an in-simulation approach.

The hardest part of any multi-scale simulation is coupling the different levels of detail together. The Mixing Martini¹²⁰ approach involves maintaining two decoupled simulation systems for the fine-grained region. Coarse-grained MARTINI virtual sites are placed at the centre of mass of the corresponding atomistic particles. Forces on the MARTINI virtual sites come from the bulk MARTINI system and are passed on to their constituent atoms, and likewise fine-grained forces apply to the atoms and therefore move the virtual sites around. The greatest difficulty with this approach is that it squanders one of the 10× speed-ups from MARTINI because the forces must be computed at every fine-grained step, while normally MARTINI forces can be calculated much less frequently. The CHARMM36/PRIMO hybrid approach³¹⁵ is similar, though it directly parametrizes interactions between coarse grained and atomistic interaction sites, and has similar detriments.

Rather than attempt an in-simulation multiscaling approach, in this chapter I attempt across-simulation multi-scaling. The interactions in a fusion protein can be broken down into 3 terms: protein–protein interactions, protein–linker interactions, and linker–linker interactions. If we can make some approximations about the protein–linker interactions, then the protein–protein and linker–linker interactions can be computed independently and at appropriate scales and combined in post-simulation processing. In essence, the protein–linker interactions are parametrised as a bias acting on some coarse-graining of the protein–protein system, and the protein–protein ensemble is re-weighted *post hoc* by this bias to produce the fusion protein ensemble.

This approach has the additional advantage of allowing the most expensive parts of the simulation to be recombined later in different fusion proteins. For example, testing 6 fusion proteins composed of the same two proteins connected by 6 different linkers with a traditional in-simulation approach would lead to 6 expensive simulations that all have identical protein–protein interactions. To compute a new fusion protein with the same linkers requires a redoubling of effort, despite conservation of the linker–linker interactions. In the method described here, all that needs repeating is the recombination of free energy landscapes.

To sample protein–protein interactions in this chapter, I use the Accelerated Weight Histogram (AWH) approach.²³³ This is a biasing enhanced sampling method similar to metadynamics (see section 1.4.7). A collective variable (or several) is computed from the simulation state and used to construct a (possibly many-dimensional) histogram of where the simulation spends its time, and this histogram is then used to bias the simulation to enhance sampling. The final trajectory can then be reweighted according to the final histogram to produce an unbiased ensemble. By contrast with metadynamics, AWH does not bias the simulation away from states it has visited, but rather biases the simulation towards a (typically flat) target distribution. In practice, this leads to AWH performing multiple runs over the entire volume of the collective variable space, rather than gradually exploring out from the starting location. It therefore converges quickly to a low-resolution energy landscape which is then refined with additional simulation time.

Fully describing the positions and orientations of two asymmetrical rigid bodies, such as the proteins in a simple fusion protein, requires a six dimensional coordinate; this can be visualised by fixing one body and specifying the position and orientation of the other with three cartesian coordinates and Euler angles respectively. With a typical biasing enhanced sampling method that explores collective variable space from the starting point outward, this would require a high-dimensional bias and be very inefficient because when the proteins are close together all six degrees of freedom are very slow. With AWH, after sampling at a close range the protein is brought back out, where it can tumble quickly. The peculiar convergence characteristics of AWH therefore suggest a single slow, biasable collective variable — the distance between the proteins.

7.2 Methods

Here, the approximations that we make are that the protein–linker interactions are limited only to the geometric restrictions the linker places on the termini of the folded proteins. That is, there are no interactions between the body of the linker and the proteins, and the bias may be framed in terms of a Z-matrix of the termini and caps of the proteins. We justify this by suggesting that an arbitrary repeat sequence is unlikely to have evolved specific interactions with any protein, and that the impact of any specific interactions that do exist by chance will be minimised by repetition along the linker, as any effect they have on one part of the linker will be diluted by repetition. The anticipated non-specific interactions amount to a crowding effect that can be corrected for simply if required.

Table 7.1: Simulations described in this chapter.

System	Atoms	Method	Temperatures	Time
EAAAK ₃	22 165	REST2/C22*	300 K – 600 K	8 × 500 ns
GGGGS ₃	22 138	REST2/C22*	300 K – 600 K	8 × 500 ns
NANP ₃	13 590	REST2/C22*	300 K – 600 K	8 × 500 ns
NANP	1 895	REST2/C22*	300 K – 600 K	8 × 200 ns
AYW	72 216	a99SB- <i>disp</i>	300 K	5 × 10 ns
svECFP	47 994	a99SB- <i>disp</i>	300 K	5 × 10 ns
AYW + svECFP	84 585	AWH/MARTINI	300 K	48 × 100 ns

C22* CHARMM22*⁶⁷

REST2 Replica Exchange with Solute Scaling^{224,316}

a99SB-*disp* MD with the a99SB-*disp* force field⁷²

AWH Accelerated Weight Histogram²³³ on distance between protein centres of mass

AYW Atu2422 with F77A, A100Y and L202W mutations (see chapter 4)

svECFP Enhanced Cyan Fluorescent Protein²⁷⁰ with C48S and C70V mutations (see chapter 3)

MARTINI MARTINI 2.2P¹¹⁴ with modifications^{108,117,118,126}

In brief, our strategy to prepare a free energy surface for a two-domain fusion protein is:

1. Simulate the linker using REST2 enhanced sampling and the C22* atomistic force field to prepare an ensemble of states of the linker in isolation. C22* is chosen for its fidelity to disordered peptides.
2. Establish a relaxed, atomistic initial structure for each domain to be used to construct a MARTINI model with elastic network. For many proteins, this would be a crystal structure; however, no crystal structure is available for AYW or svECFP, so mutations are introduced to the closest related structure and relaxed using the state-of-the-art for folded proteins, a99SB-*disp*.
3. Construct a MARTINI model with elastic network from the above domain simulations.
4. Simulate protein–protein interactions of the two MARTINI domains using AWH to prepare a weighted ensemble of states for the protein dimer without a linker.
5. Compute free energy landscapes of each linker in terms of geometric collective variables that can be computed from the linker simulation and the protein simulation.
6. Compute free energy landscapes of the fusion protein in any collective variable by reweighting the protein–protein ensemble according to the free energy landscapes computed from the linker simulations.

7.2.1 Linker simulations

First, I will describe simulations performed on the linker peptides themselves. Three linker repeat sequences were selected: EAAAK for its helical propensity, GGGGS for its flexibility, and NANP as an example of a natural linker. Each was simulated as trimers (eg, (NANP)₃), and NANP was additionally simulated as a monomer to investigate how the behaviour of a single repeat differs from that of a trimer.

Linker ensembles were sampled via REST2²²⁴ MD simulations, a Hamiltonian replica exchange method in which potential terms involving the linker are scaled such that higher ladders experience higher effective solute temperatures, but the solvent temperature remains fixed. This allows a small number of replicas to span a large temperature ladder, even with a large number of solvent molecules and a constant pressure ensemble.

The extended conformation of each linker was generated with the PeptideBuilder library.³¹⁷ These were capped with an N-terminal acetyl group and a C-terminal N-methyl group. CHARMM22*⁶⁷ topologies were generated for each with GROMACS' `pdb2gmx` tool. CHARMM22* was chosen for its faithful reproduction of helical propensities of small repeat peptides, especially at 300 K.^{18,72}

The extended peptides were placed in a rhombic dodecahedral box. The longest dimension of the peptide was aligned with the shortest dimension of the box, and a 0.6 nm buffer was placed between peptide and box. In this initial conformation, the distance between periodic images of the peptide is equal to the non-bonded cut-off distance of 1.2 nm. I note that this is contrary to the guideline given in the introduction that the image distance be twice the cutoff (see section 1.4.5); however, this initial configuration is the worst-case image distance and is extremely entropically unfavourable. Any rotation or conformational change would collapse the peptide or rotate it away from the shortest box dimension, and therefore increase the image distance. This guarantees that at constant box size the peptide remains beyond the cut-off distance.

In practice, buffer distances during production were much greater. After equilibration, the peptide had collapsed and rotated such that the image distance was greater than twice the cut-off distance, and excursions closer than this limit were never observed in any simulation. Thus, this practice was effective in producing small simulation boxes that follow the twice-cutoff rule. Also note that the twice-cutoff rule guarantees only that periodic images do not directly influence each other via the Lennard-Jones interaction; long-range effects from PME electrostatics and finite size effects provide no such guarantees.

The extended conformations were solvated with aqueous 0.15 M NaCl solution with GROMACS' `solvate` and `genion` tools. The solvent was pre-equilibrated at constant pressure in the target force field before solvation, as the pre-equilibrated SPC water boxes distributed with GROMACS were found to have a substantially different density to TIP3P or a99SB-*disp* water. Solute concentration was calculated relative to the concentration of water and not based on the box size. The solvated system was energy minimised until the maximum force was less than 1000 KJ/mol/nm, sufficient for stable simulation with a 2 fs time step. A first round of equilibration with an unmodified potential was carried out for 1 ns with the Berendsen barostat ($\tau_p = 0.5$ ps) to equilibrate the box size and pressure. The

equilibration otherwise followed the production simulation parameters below. In each equilibration, the box expanded rather than contracted while the peptide itself contracted, so that the distance between periodic images grew, consistent with our box size guarantees above.

A second equilibration was performed for a further 1 ns with identical parameters but with the full REST2 ladder. Effective solute temperatures were chosen to fit 8 replicas from 300 K to 600 K with a geometric progression — 300 K (unmodified), 331.23 K, 365.70 K, 403.77 K, 445.80 K, 492.20 K, 543.43 K, 600.00 K. Exchange was not attempted.

The production run consisted of a 500 ns run with exchange attempted every 100 steps.^{215,216} Peptide and solvents were separately treated with the CSV¹⁸⁶ thermostat at 300 K and a coupling constant of 1.0 ps. Note that the thermostat was set at 300 K for all replicas, and higher temperature rungs come about purely because of potential scaling.²²⁴ Pressure was maintained with the Parrinello-Rahman²⁰² barostat with a coupling time of 12.0 ps. Electrostatics were treated with PME and the LJ forces were switched to zero from 1.0 nm to the cut-off at 1.2 nm consistent with their parametrisation.⁶⁷ Centre of mass motion was removed every 2 ps and all bonds were modelled with constraints.

7.2.2 Fluorophore parametrisation

One of the proteins involved in the target fusion protein is svECFP, a variant of the Enhanced Cyan Fluorescent Protein²⁷⁰ with surface cysteine residues mutated to serine or valine. Both ECFP and svECFP are in the GFP family, which is characterised by a beta-barrel fold enclosing a solvent-protected loop that undergoes an auto-catalytic reaction to produce a fluorophore.

The ECFP fluorophore has not to my knowledge previously been parametrised as an amino acid building block. I require parameters both in an atomistic force field, for relaxing mutations introduced to ECFP, and in the MARTINI force field for the protein-protein simulations. While this was not found to be important in previous simulations of ECFP, my goals with these simulations was to exceed previous simulation quality and the cost of this parametrisation was relatively low. I chose the a99SB-*disp* force field⁷² for Amber's superior support for residue parametrisation compared to CGenFF, and for the force field's excellent treatment of folded proteins relative to GROMOS, whose ATB would likely produce superior fluorophore parameters.

Amber/GAFF2

The fluorophore structure and the surrounding residues (residues 64-68 inclusive) were taken from the crystal structure of ECFP (PDB 2WSN³¹⁸). Atoms were removed from the surrounding residues to yield a structure of the fluorophore capped by acetyl- and N-methyl groups, and hydrogen atoms were added with Avogadro 1.20's 'Add Hydrogens for pH' function.³¹⁹ AM1-bcc charges, Amber atom types, and bonded interactions were computed with the Ambertools 17.0³²⁰ programs Antechamber and prepren. GAFF2 atom types were used where Amber atom types were unavailable.^{150,321} Bonded interactions were taken from the a99-*disp* force field where available, or otherwise from GAFF2.

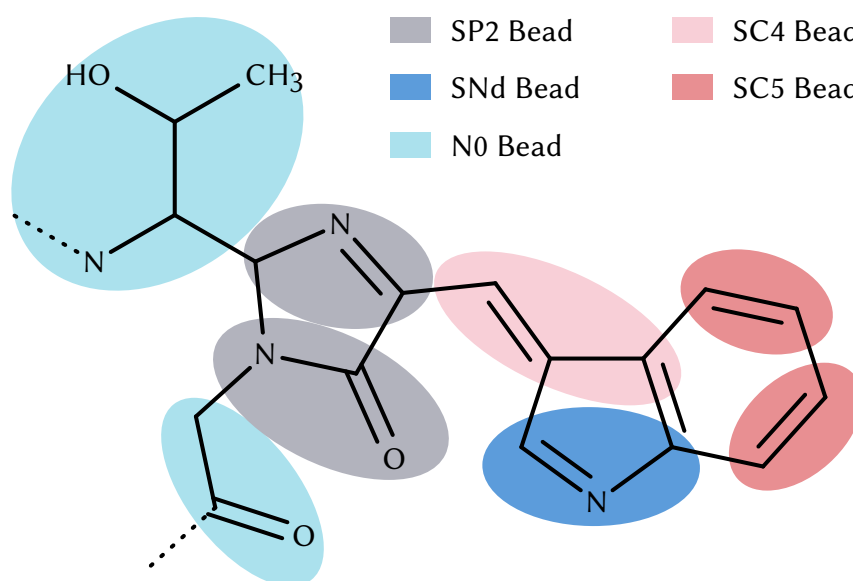


Figure 7.1: Mapping of atoms in the ECFP fluorophore to MARTINI beads.

The parameters were then checked by simulation. Acetyl/N-methyl groups were added to the termini as neutral capping moieties and the molecule was solvated in 363 water molecules in a rhombic dodecahedral box. A 1 ns equilibration with a 2 fs time-step and the Berendsen barostat was followed by a 50 ns production run using the parameters for an a99SB-*disp* production simulation described below (7.2.3). The resulting trajectory was checked visually for any unrealistic or absent motions.

MARTINI

An initial atom-bead mapping was produced by `auto_martini`³²² and then tweaked for consistency with existing amino acid parameters. The resulting mapping is presented in figure 7.1. The bonded interactions were then optimised against the reference simulation performed above (7.2.2) with PyCG-TOOL³²³ with the constraint threshold set to 100 000 KJ/mol/nm² and validated via a brief simulation with explicit solvent. This consisted first of a 1 ns equilibration with the Berendsen barostat and a 10 fs time step, then a 10 ns equilibration with a 20 fs time step and the Berendsen barostat, and finally a 100 ns simulation with a 20 fs time step matching the standard NEW-RF MARTINI parameters.¹¹⁸

7.2.3 Mutated structure derivation

Both protein domains in the target fusion protein do not have crystal structures available, but are mutants of proteins that do. svECFP involves introducing two mutations to ECFP, and AYW is a triple-mutant Atu2422. Instead of generate MARTINI models directly from the wild type crystal structures, I elected to introduce these mutations computationally and generate relaxed structures for further simulation. Relaxation followed a method inspired by Heo et al.¹⁰ Crystal structures of the wild type Atu2422 and ECFP were taken from the PDB (IDs 3IP5³²⁴ and 2WSN³¹⁸ respectively). Mutations and

neutral acetyl/N-methyl caps were introduced to the crystal structures in PyMOL. The C-terminal N-methyl group was not added to AYW as no linker was to be added to this terminus and the crystal structure included the entire protein.

Topologies were generated for the a99SB-*disp* force field with `pdb2gmx`. Each protein was placed in a rhombic dodecahedral box with a 1.2 nm buffer region between the protein and the edge of the box. The systems were solvated with the GROMACS `solvate` tool and salt was added to a concentration of 0.15 M with `genion`. Solute concentration was calculated relative to the concentration of water and not based on the box size leading to a total of 41 Cl^- ions, 41 Na^+ ions, and 16 745 H_2O molecules in the AYW system and 27 Cl^- ions, 34 Na^+ ions, and 11 095 H_2O molecules for svECFP. Both systems were energy minimised until they reached a maximum force less than 1000 KJ/mol/nm, suitable for integration with a 2 fs time step.

Five replica simulations of each system were carried forward from the energy minimisation. Each was subjected to a 100 ps equilibration with the Berendsen barostat ($\tau_T = 0.5$ ps) and 1000 KJ/mol/nm² position restraints on all heavy atoms. Production runs consisted of 10 ns runs with α -carbons restrained to the crystal structure positions with a harmonic flat-bottom potential ($r = 0.4$ nm, $k = 100$ KJ/mol/nm²)¹⁰ and the Parrinello-Rahman barostat ($\tau_P = 12.0$ ps). All simulations used a 2 fs time step, constraints on bonds involving hydrogen atoms, a straight LJ cut-off at 1.2 nm, and PME electrostatics. The centre of mass motion was removed every 2 ps. The frame from the 5 replicas with the least potential energy, ignoring the contribution of the restraints, was carried forward as the MARTINI starting structure.

7.2.4 Protein–Protein simulations

With relaxed structures of both mutant domains of the fusion protein in hand, MARTINI simulations of their protein–protein interactions could begin. MARTINI coordinates and topologies were generated from the relaxed mutant structures with MARTINIZE v2.6_3 running on Python 3.¹⁰⁷ MARTINIZE was modified to work with acetyl/N-methyl termini caps, which always use the N0 bead, and to work correctly with multiple chains. The ElnDyn 2.2p force field^{108,114} with refPol water¹¹⁷ was rescaled for better protein–protein interactions¹²⁶ with $\alpha = 0.3$ for use in this project. The svECFP fluorophore was changed to GLY-GLY before being passed through MARTINIZE to generate the elastic network and backbone bead positions, and the bespoke fluorophore parameters were added later by hand. The svECFP system was then energy minimised *in vacuo* to machine precision.

AWH simulations with a shared bias perform best when initial conformations are distributed evenly along the target range of the AWH reaction coordinate. As such, 48 conformations of the heterodimer were generated by keeping AYW fixed and placing svECFP in a random orientation and displacement in the range 5.5 – 8.0 nm away. The dimer system was solvated in a rhombic dodecahedral box with 26 194 polarisable water particles, each representing four water molecules, 280 Cl^- ions, and 298 Na^+ ions, corresponding to a concentration of 0.15 M.

Each system was energy minimised to machine precision, which was not sufficient to permit a

10 fs time step. An initial five 1 fs steps of equilibration with the Berendsen barostat ($\tau_p = 2.0$ ps) was sufficient to permit further equilibration for 1 ns with a 10 fs time step, the Berendsen barostat ($\tau_p = 2.0$ ps) and a flat bottom potential ($k = 1000$ KJ/mol/nm²) between the centres of masses of the proteins. This initial five-step equilibration was required to automatically generate equilibrated structures across the target range of the AWH reaction coordinate. While it is generally preferable to correct structural issues by hand, this is not only labourious when performed on up to 48 conformations but would most likely involve regenerating starting orientations and displacements, thus disturbing the uniform distribution across the reaction coordinate. As only 5 steps at the smaller time step were necessary, and both clashes and voids are expected results of generating dimer conformations in close proximity, this brief initial equilibration was deemed both harmless and useful.

Production simulations consisted of 48×100 ns replicas with a 10 fs time-step. The standard NEW-RF¹¹⁸ parameter set was used for all MARTINI simulations except as noted. The CSV thermostat at 300 K with a coupling time of 1.0 ps was separately applied to the protein and solvent, and a Parrinello-Rahman barostat was applied with a coupling time of 24.0 ps. The Accelerated Weight Histogram²³³ method was applied to the distance between the centres of mass of the two proteins in the range 0.0 nm to 8.0 nm to enhance sampling. The AWH bias was shared across all 48 simulations and the histogram was equilibrated before entering the initial stage. The potential was constructed from convolved umbrellas with force constants of 1000 KJ/mol/nm² and the initial diffusion constant was 10^{-4} nm²/ps. The target distribution $\rho(\lambda)$ approximated a uniform distribution up to a free energy cut-off of $F_{\text{cut}} = 50.0$ KJ/mol, where it smoothly switches to a Boltzmann distribution in the free energy $F(\lambda)$:

$$\rho(\lambda) \propto \frac{1}{1 + \exp(F(\lambda) - F_{\text{cut}})}$$

7.2.5 Free energy calculation

5000 snapshots from each linker simulation were taken. Four-dimensional histograms were constructed for each linker simulation along four geometric co-ordinates that could be computed both from the protein–protein simulations and the linker simulations (see table 7.2). These histograms were constructed with 50 equally spaced bins along the primary co-ordinate (r), and 10 equally spaced bins along other co-ordinates (α , β , and γ). These linker histograms are used as a post-processing bias for the protein–protein trajectories. The use of a free energy estimator such as MBAR³²⁵ and equally populated rather than equally spaced bins would improve the construction of this bias, as well as permit the inclusion of data from other replicas, but was deemed unnecessary for this proof of concept.

It then remains to reweight the coarse-grained protein–protein trajectory to remove the effect of the AWH bias³²⁶ and add in the effect of the linker. To this end, the trajectory data was binned along a target collective variable, and a weighted histogram was constructed in which a frame at time t contributes a weight $w(t)$ to its bin:

$$w(t) = \exp \left(U_{\text{AWH}}(t) - U_{\text{Linker}}(r(t), \alpha(t), \beta(t), \gamma(t)) \right) \quad (7.1)$$

Table 7.2: Collective variables used for PMF. Atom identities are found in table 7.3.

Definition	Description
$r = \ \mathbf{r}_i - \mathbf{r}_l\ $	Protein–Protein inter-terminal distance
$\alpha = \angle ijl$	Angle between end of linker and peptide bond at start of linker
$\beta = \angle ikl$	Angle between start of linker and peptide bond at end of linker
$\gamma = \text{dih}(i, j, k, l)$	Dihedral angle between terminal peptide bonds

Table 7.3: Atoms used to compute collective variables in table 7.2. See also figure 7.4.

Atom	Protein–Protein (CG)	Linker	System
i	C-terminal backbone	ACE cap CH ₃	Last α -carbon of protein A
j	NME cap	N-terminal α -carbon	First α -carbon of linker
k	ACE cap	C-terminal α -carbon	Last α -carbon of linker
l	N-terminal backbone	NME cap CH ₃	First α -carbon of protein B

Where $U_{\text{AWH}}(t)$ is the energy of the bias at time t and $U_{\text{Linker}}(r, \alpha, \beta, \gamma)$ is the free energy of the linker at that co-ordinate. The estimated relative free energy $\hat{\phi}$ of bin b is then computed by summing over its weights (see section 1.2.3):

$$\hat{\phi}(b) = -k_B T \ln \sum_{t \in b} w(t)$$

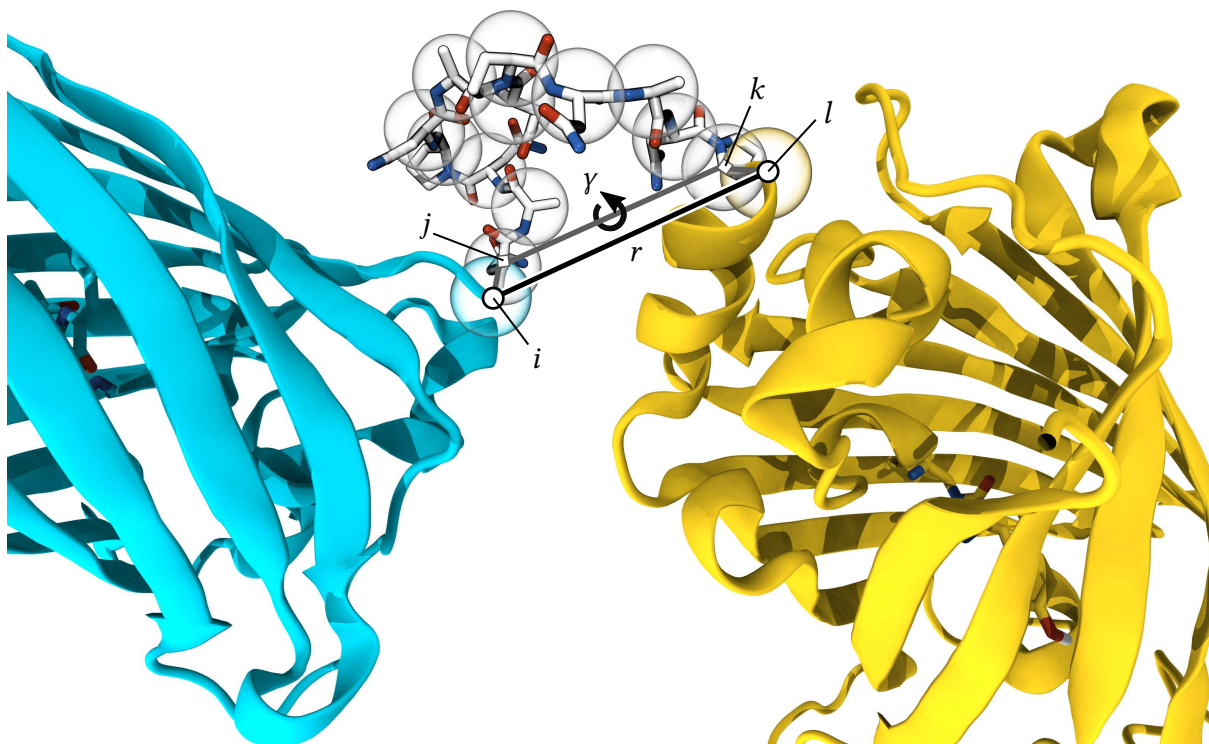


Figure 7.2: Collective variables used for PMF. Here, the two domains represented as cartoons are the fluorescent proteins ECFP and Venus. The linker, represented in white, is $(\text{NANP})_3$. This frame was chosen from the $(\text{NANP})_3$ ensemble for visualisation purposes only and may not represent a state in any ensemble. Backbone beads corresponding to the linker are shown as bubbles. See also table 7.2 and table 7.3. The torsion γ and end-to-end distance r are also shown in grey and black, respectively. The angles $\alpha = \angle ij l$ and $\beta = \angle ik l$ are not shown explicitly. Note that atoms i, j, k , and l are modelled in both the atomistic linker simulations and the coarse grained protein–protein simulation.

7.3 Results

7.3.1 Linker simulations

Linkers were simulated in a Hamiltonian replica exchange scheme (see section 7.2.1) with the CHARMM22* force field. Linkers were simulated as trimers of their repeat sequences to capture some of the effects of context on structure; though this effect may be small compared to the impact of the proteins, it is at least computationally inexpensive to include and transferrable across fusions. The $(\text{NANP})_n$ repeat was also simulated alone for comparison. By analysing the end-to-end distance r of each repeat separately, the effect of context on the structure of the linker can be investigated. As shown for $(\text{NANP})_n$ (figure 7.3, left), each free energy landscape in r is largely independent of context, featuring a densely populated extended state at 1.5 nm and an entropically favoured collapsed state. Though the population of these two states varies with context, these variations are negligible compared to the variations between linker compositions. This is helpful moving forward as it allows aggregation of sampling of the trimer and potentially the extrapolation of results to longer repeats.

While the context of the repeat seems to have a minimal impact on the dynamics of its end-to-

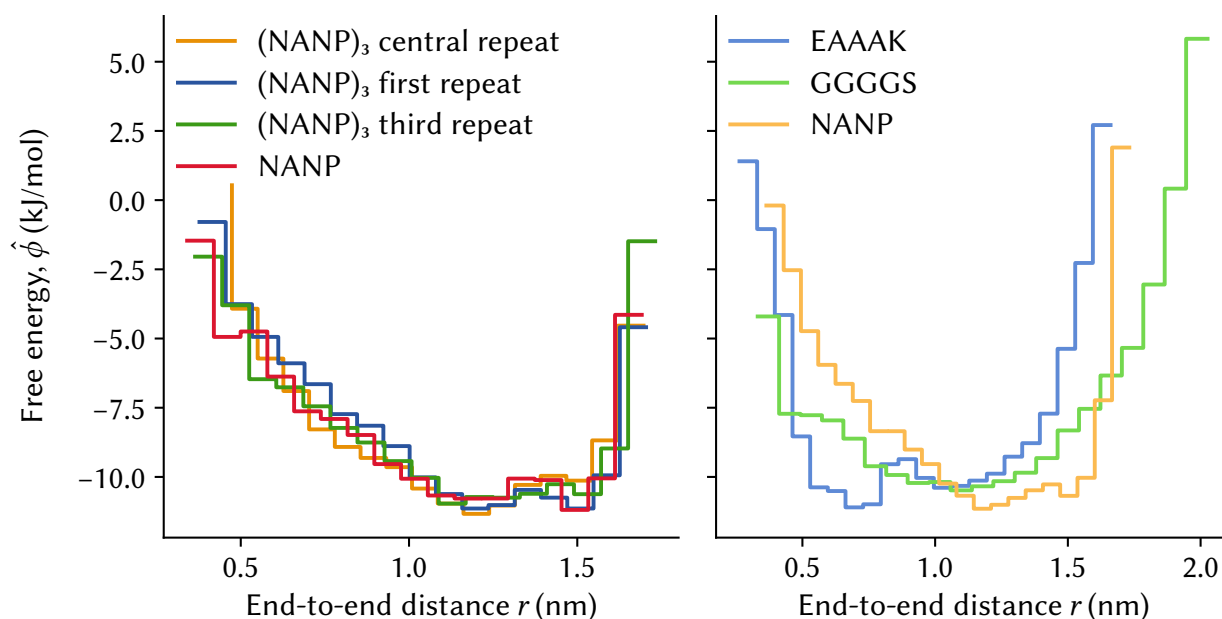


Figure 7.3: Single repeat linkers in different contexts. A Jacobian correction has been applied.

end distance, the composition is critically important (figure 7.3, right). The flexible (GGGGS)_n linker lives up to its name, exhibiting a broad, featureless free energy landscape, except for a notable dip corresponding to the α -helix around 0.6 nm. By contrast, the rigid, helical linker (EAAAK)_n features a peak at approximately 0.7 nm, consistent with an α -helix of 5 residues. The (NANP)_n linker prefers an extended conformation thanks to its proline content, and also exhibits a broad collapsed state. Notably, while (EAAAK)_n is rigid, with a distinct preferred configuration, single repeats may not serve to separate domains in space, as the helical configuration is quite compact. By contrast, NANP is similarly rigid but with a much more extended preferred state. This may help explain its conservation in the circumsporozoite protein, in addition to its immunological properties. Further, it may be helpful for designed fusion proteins that need to fine tune the distance between domains without simply introducing a long α -helix.

The stark differences between linker compositions are softened when they are simulated as trimers (figure 7.4, upper left). The glycine-rich flexible linker prefers a high-entropy collapsed state, while the proline-containing (NANP)₃ linker has a remarkably flat energy landscape, particularly between 2 and 3 nm. The helical (EAAAK)₃ linker features a notable, though subtle, valley at approximately 2.3 nm end-to-end distance, consistent with an alpha helix of 15 residues. While the precise size of this peak is likely force-field dependent^{18,72} and may not accurately reflect an experimental free energy landscape, the generally extended end-to-end distance of the helical linker reflects substantial but not overwhelming helical structure in the ensemble. The overall DSSP³²⁷ per-residue helical content for the (EAAAK)₃ simulation was about 20% and was highly context-specific, with central residues spending as much as 36% of the simulation as a helix.

While the end-to-end distance r is a convenient and conceptually simple collective variable for the

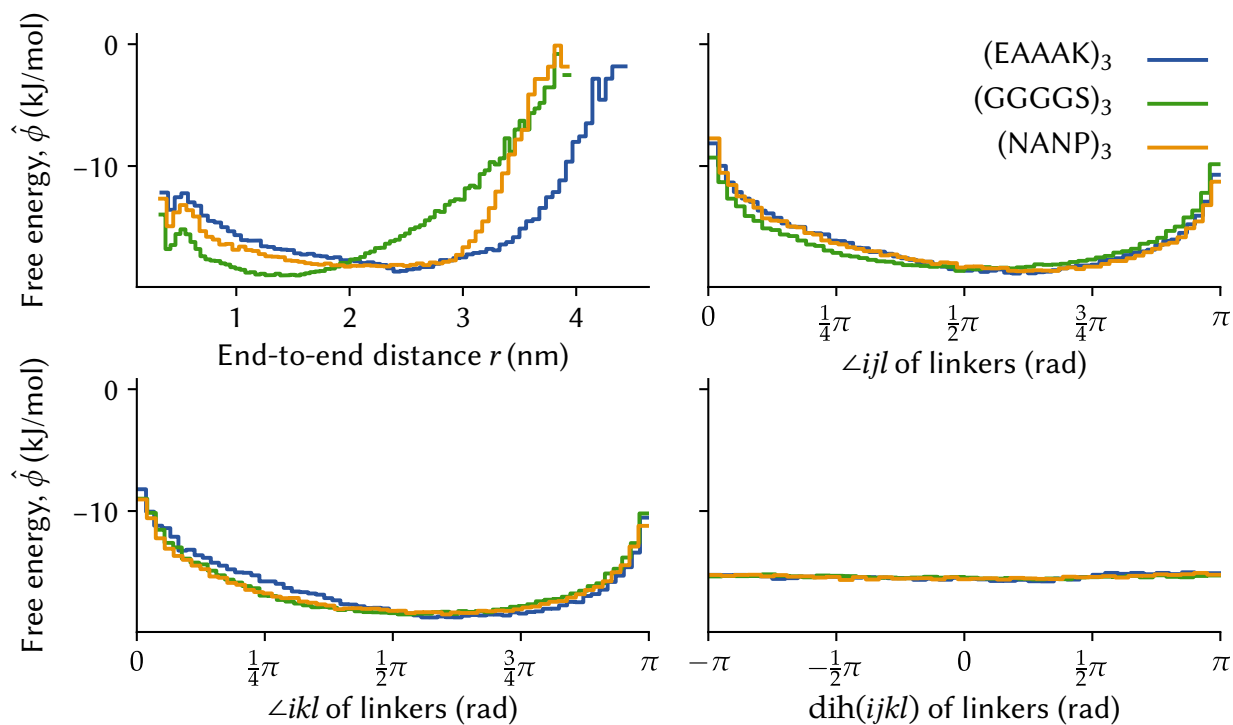


Figure 7.4: Internal co-ordinates of trimeric linkers

linker, other degrees of freedom may be significant in constraining the fusion protein domains. To this end, the entire linker may be considered as a single chemical bond, in which case internal coordinate representations may be used to completely describe the positions and orientations of both domains. In this scheme, the acetyl and N-methyl caps on the linker respectively correspond with the C- and N-terminal amino acid residues of the domains, the N-methyl cap of the N-terminal domain corresponds with the N-terminal amino acid residue of the linker, and the acetyl cap of the C-terminal domain corresponds with the C-terminal amino acid residue of the linker (table 7.3). Thus, the terminal peptide bonds of the linker model represent the terminal peptide bonds of the domain models. The end-to-end distance r then represents a virtual bond length for the linker, while bond angle terms and a torsion term (α , β , and γ , see table 7.2 and figure 7.2) represent their relative orientation.

When considering only a single repeat, correlations between these four variables are very clear (data not shown), and so the 4-dimensional free energy surface was used to model the linker in the fusion protein reweighting procedure. However, for the longer triple repeats, correlations were insignificant. In addition, only the end-to-end distance r depended substantially on linker composition (figure 7.4). The linker angles α and β were similar for all three linkers, and are probably largely entropic in origin as they reflect the relative scarcity of peptide conformations with parallel terminal bonds compared to intermediate angles. The linker torsion γ was extremely flat, with less than a kJ/mol's variation along the entire range.

Note that this scheme is degenerate with respect to torsions around the peptide bonds connecting the linker to the domains, as these torsions are not well defined in the MARTINI model or the N-methyl linker. In a model of the entire fusion protein, these would be represented by $\text{dih}(j, k, l, m)$ and

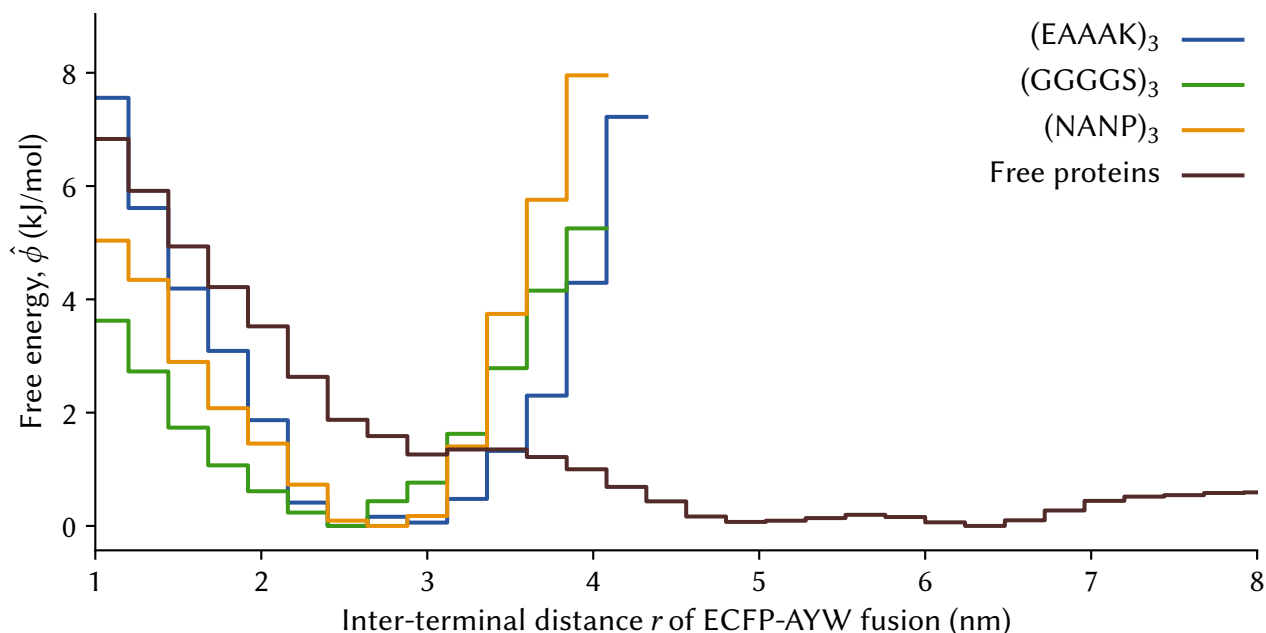


Figure 7.5: Inter-terminal distance of AlexaFluor 532-labelled ECFP-AYW fusion protein

$\text{dih}(h, i, j, k)$, where h represents some fixed atom in the N-terminal domain and m some fixed atom in the C-terminal domain. Both h and m exist in the MARTINI model, where they could be represented by beads further up the respective chains, and a candidate for h exists in the atomistic linker model for the acetyl cap, where it corresponds to the carbonyl oxygen. However, these candidates do not correspond with each other, and are absent altogether for the atomistic N-methyl cap. This could be remedied, albeit with significant loss of generality, by extending the linker models to include the first residue of each domain, or by extending the domain models to include the first residues of the linker. However, this would eliminate the combinatorial advantage of being able to mix and match linker and protein–protein simulations. In addition, similarly to the other orientational degrees of freedom (figure 7.4), the effect of this torsion is likely to be very small, as it is amortized over torsions all along the linker.

7.3.2 Multiscale technique

The four-dimensional free energy surface of the linker was used to reweight AWH simulations sampling along the distance between centres of mass of ECFP and AYW, proteins taken from the sensor GlyFS (chapter 4). Initially, the inter-terminal distance r was taken from the reweighted ensemble (figure 7.5). The free proteins, that is, ECFP and AYW reweighted only to remove the biasing effect of AWH and not with any linker, exhibit clashes at close range, valleys at 5 and 6.5 nm distance, and are then flat out to a distance of about 12 nm where finite size effects come into play (not shown).

Including the effect of the trimeric linkers constrains the end-to-end distance to 4 nm, as expected. The difference between the linkers on the full fusion construct is subtle but mirrors the behaviour

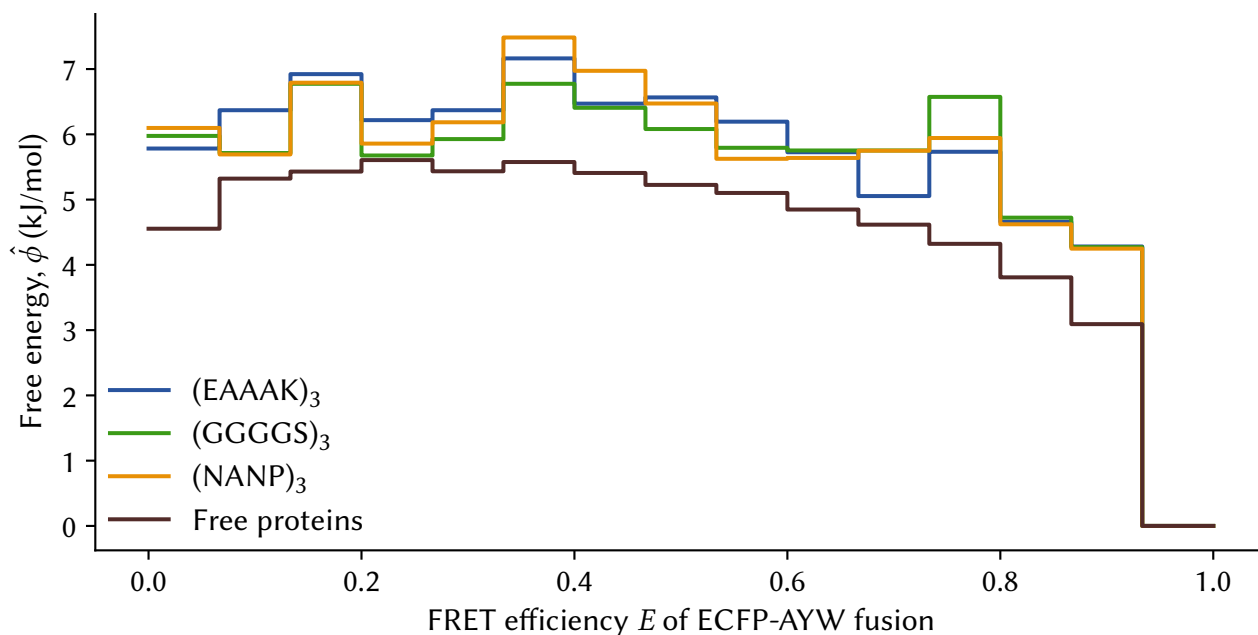


Figure 7.6: FRET efficiency of ECFP-AYW fusion

of the trimers themselves. The rigid linker (EAAAK)₃ prefers slightly extended conformations, while the flexible linker (GGGS)₃ prefers collapsed ones. However, owing to the high dimensionality of the linker free energy surface and the unsophisticated binning procedure chosen, large bins were required to ensure that all bins were populated and the resulting free energy surfaces are concomitantly low-resolution. A better choice of secondary collective variables and a variable-width binning procedure would alleviate these issues. Finally, the use of a free energy estimator like MBAR³²⁵ would enable both the use of linker sampling from the entire ladder and the estimation of uncertainties.

To demonstrate that arbitrary quantities can be computed from the reweighted ensemble, the theoretical FRET efficiency from the ECFP fluorophore to a putatively AlexaFluor 532-labelled residue Asn177 of AYW was computed (figure 7.6).²⁶⁹ The resulting free energy surface demonstrates that the addition of a linker improves FRET efficiency of the putative sensor in the closed state. However, the details required to compare different linkers are obscured by the low resolution of the landscape, which is limited by the sampling performed in the protein–protein simulation. This demonstrates that any collective variable can be obtained by this reweighting approach with appropriate sampling and bin selection.

7.4 Discussion and future work

While notable valleys exist for all of the simulated linker peptides, their shallow depths of only a few kJ/mol are surprisingly small given the dramatic changes in sensor dynamic range described in chapter 4. However, they are clearly sufficient to change the average inter-terminal distance for this fusion protein by a few nanometres, as can be seen in figure 7.6. It may be that the shape of the higher-dimensional

free energy landscape that plays a role as well. Indeed, the helical (EAAAK)₃ linker should not be thought of as simply forming a helix, but instead should be considered composed of residues that have a propensity towards helical structure. In other words, (EAAAK)_n does not form a single long helix, but short lengths of its residues tend to dynamically fold and unfold into short helices. In the case of (EAAAK)₁, this favours a collapsed state, as a short helix is very compact and stretches almost the full length of the linker. By contrast, (EAAAK)₃ favours extended conformations composed of short helices separated by short coils. This is conceptually consistent with previous modelling of long linkers as wormlike or Gaussian chains.²⁸¹ The fully helical state, which is indeed present in the (EAAAK)₃ ensemble, may however be stabilised by interactions with domains of the fusion protein.

Interactions between the linker and the domains is, importantly, missing from this analysis. While this means that predictions from this method are likely to be sufficient for qualitative comparison, they are not expected to be quantitative. It is difficult to imagine an efficient method that captures linker–protein interactions; the most obvious approach would be an in-simulation multi-scaling approach, treating the linker atomistically and the proteins as coarse-grained.^{120,315} However, not only would this involve slow sampling across all six degrees of freedom defining the position and orientation of the two proteins, it would also forgo much of the speedup MARTINI provides and only provide coarse grained interactions between the linker and protein.

The approach described here only requires one explicit collective variable — the protein–protein distance — as sampling across the others happens quickly when the proteins tumble at larger distances. In addition, this approach allows linker simulations to be reused for new fusion proteins, whereas a more detailed approach would require a new simulation of the entire fusion for every variant. As the state of the art currently neglects most of even the linker–linker interactions, this is a substantial improvement. However, if a highly accurate view of the fusion protein is required, a more detailed approach is most likely required.

One general effect on the linker is that it is, in essence, crowded by the surrounding domains.³²⁸ This crowding-like effect could be modelled explicitly by adding a large Lennard-Jones sphere to the termini of the linker models,³²⁹ or implicitly by adding a bespoke correction term to the exponent in equation 7.1. These large Lennard-Jones spheres were considered while planning the simulations in this chapter, but their inclusion would require much larger linker simulation boxes and would come with a concomitant increase in computational expense. As the goal of this chapter was to efficiently simulate fusion proteins, this approach rather defeats the purpose. A correction term would be more efficient, but would require substantial parametrisation that is beyond the scope of this chapter. Future work might involve calibrating such a correction against experimental data.

Molecular dynamics has proven difficult to apply to protein–protein interactions.³³⁰ Not only are atomistic force fields optimised for the dynamics of individual proteins, and not for that of complexes,⁸⁴ but the inherently high dimensionality of the problem resists enhanced sampling approaches. To completely describe the relative positions of two rigid bodies, six variables are required; three to describe displacement, and three for orientation. While some of these degrees of freedom may be redundant

for a particular problem, traditional biased enhanced sampling methods (see 1.4.7) like metadynamics cannot tolerate slow unbiased degrees of freedom that are correlated with the accelerated variable. In protein–protein interactions, all of the 6 variables are highly correlated at short ranges.

Traditional bias methods fail with high-dimensionality problems like this because they must fully explore one region of the 6D hyper-space before they can move on to the next. Thus, for example, they may provide no information at all about interactions on one side of a protein until the other side is completely explored. By contrast, AWH performs sweeps over the entire target space, iteratively improving its estimate of the entire free energy surface as it goes. For protein–protein interactions, this is ideal as the pathological case of many highly correlated slow degrees of freedom only exists when the two proteins are in close proximity. Thus, by using a single distance between the two proteins as a collective variable, the proteins are brought into close contact in a particular orientation before being brought apart, allowed to tumble while still under the influence of mutual long-range interactions, and then brought back together. In addition, AWH is trivially adapted to parallel simulation with multiple replica schemes. Taken together, this allows an enormous amount of sampling to be acquired quickly.

With the methodology described here, uncertainties in the resulting free energies could not be easily estimated. While the 48 replicas used in the protein–protein simulations would suggest a jack-knifing or bootstrapping approach, these replicas shared a bias and cannot therefore be assumed independent. The use of independent trajectories would support error estimation, as well as simplifying execution of the simulations themselves and allowing even greater sampling. However, the simplistic approach used here serves to demonstrate the technique.

The great advantage of the approach taken here is that linkers and domain pairs can be simulated in isolation, and then combined in a computationally cheap post-processing step. This allows a library of linkers to be built up, and then an entirely new fusion protein to be modelled with only one additional simulation of the two domains. Thus, the cost of comparing a range of fusion protein candidates is linear ($\mathcal{O}(l + p)$) in the number of domain pairs p and in the number of linkers l . This is an important consideration when designing MD techniques for protein engineering, as the alternative is to simply construct all of the fusion protein candidates in the lab. This can be a favourable alternative, especially when individual simulations are expensive, their accuracy is in doubt, or available computational resources are limited.

Existing multi-scale techniques for combining atomistic simulations with MARTINI could have been used here,^{120,331–333} but they do not provide the above linear scaling and a new, full-cost simulation must be done for every protein–linker–protein configuration. A Hamiltonian replica exchange simulation similar to that described by Liu et al.³³³ but with an additional bias-enhanced sampling method was considered, but the ability to simulate a given linker only once was considered valuable. However, a virtual-site method such as that described by Wassenaar et al.¹²⁰ would make an ideal reference simulation against which to parametrise a correction bias, as it would permit direct comparison of our linear approach to a non-linear approach with identical force fields.

Finally, for this method to be a useful design technique, it must not only be efficient but also accurate.

Thus the technique must accurately reproduce the true value given the force field, and also the force field must be an accurate model of the true interactions. While this chapter demonstrates that computing reasonable values for these systems is possible, it must be compared to an experiment as a benchmark. A semisynthetic variant of GlyFS (see chapters 3 and 4) described in my Honours thesis²⁶⁹ was chosen to facilitate these tests. It has the advantage of being a simple two-domain sensor whose performance, like GlyFS, is likely to be linker-dependent, though these experiments have not yet been carried out. Predicted FRET efficiencies may be measured via fluorescence, and compared between open and closed states. In addition, fluorescence permits single molecule experiments to validate the entire free energy landscape, allowing isolation of incorrect predictions.³³⁴

7.4.1 Combining force fields

Combining parameters from different force field families into a single simulation is generally unwise, as each family uses its own functional forms and derivation strategies (see section 1.4.1). For example, Amber and CHARMM charges are both derived from quantum chemical calculations, but Amber charges are fitted to reproduce the calculation's electrostatic potential, while CHARMM charges reproduce electrostatic binding energies of dimers.⁶¹ These procedures produce different charges that are nonetheless internally consistent. The resulting charges are proven in the context of their respective force fields electrostatic and Lennard-Jones potentials, but there is no reason to expect them to perform with any accuracy in a completely different environment.

Despite this, this chapter involves combining results from three different force fields. Crucially, this does not involve combination of parameters; instead, each force field is used to generate an trajectory, and complex properties are computed from the trajectories. Since the trajectories are the intended product of the force fields, this approach is no more problematic than combining models produced from crystallographic and cryo-EM data to derive the structure of a protein complex. Instead of parameters being moved outside of their optimal context to produce spurious results, each force field is used where it is the best tool for the job.

Rather than being detrimental, this strategy can improve the quality of results; it means that highly accurate but computationally expensive force fields can be used where their expense is justified. Extending this strategy and introducing even more force fields could improve results further; one might imagine combining trajectories of the same system from multiple force fields to average out opposing biases from different parameterisation strategies, or using force fields that are highly optimised to a particular chemical context.

7.5 Conclusion

A multi-scale method for modelling the dynamics of fusion proteins in a piecemeal fashion that lends itself to combinatorial selection of linkers for extended fusion protein systems is described, and a

simplistic version implemented and demonstrated on a real biosensor platform. With the completion of additional simulations of relevant linkers and therefore the provision of accurate, high-dimensional free energy surfaces, true rational design of fusion proteins may soon be generally available. In addition, comprehensive sampling of several popular linker sequences is performed and described, and the application of AWH to protein–protein interactions is attempted.

Chapter 8

Conclusions and future work

This work uses FRET biosensors as a platform for the design of computational techniques for protein engineering. FRET biosensors leverage a binding-associated motion from a analyte-specific binding protein to correlate the distance between two fluorophores to the concentration of the analyte. The fluorophores are carefully chosen so that the emission spectrum of the first (the donor) overlaps the excitation spectrum of the second (the acceptor). At close distances, the donor's excitation energy is efficiently transferred to the acceptor via FRET, while at greater distances this transfer is dramatically less efficient. Thus, after excitation of the donor by the experimenter, the concentration of the analyte is reported via the ratio of emission peaks of the two fluorophores.

FRET biosensors are simple to design construct, but their quality is strongly dependent on a multitude of details of their components. In particular, the positions and dynamics of the fluorophores, and their coupling with the binding domain, determine the sensitivity and resolution of the sensor. These details depend in turn on complex interactions between the components and the linkers that connect them. This sensitivity suggests that improvements in modelling fusion proteins will be highly applicable to sensors. As a practical side effect of this choice, this thesis also describes the design of a number of new sensors.

8.1 Conclusions

8.1.1 Rangefinder is a rapid semisynthetic sensor design method

Fusing a fluorescent protein with a periplasmic binding protein and then chemically labelling a carefully chosen site on the binding protein produces a so-called semisynthetic sensor. These sensors are easy to produce with a simple computational model and produce bright, high dynamic range signals. The Rangefinder program, written and conceived of by the author, requires only structures of the binding protein, is tolerant of simple homology models, and produces multiple high-quality candidate labelling sites instantaneously. Though semisynthetic sensors cannot be genetically encoded, their brightness, high dynamic range and small size make them appropriate for *in vivo* and *ex vivo* use, and their simple construction lends themselves to evaluation of techniques for modelling fusion proteins more generally.

8.1.2 The domains of a fusion protein are highly dynamic

By performing unprecedented MD simulations on fusion proteins, this thesis demonstrates that the domains of a synthetic fusion protein, lacking specific interactions between them, tend to distribute themselves widely across their available configuration space. In addition to rotational averaging described in the literature,²⁷¹ this wide distribution covers a translational aspect, and the orientational and translational aspects can display complex correlations. Protein engineers should not expect two domains to form an ordered complex, even with a short linker or no linker at all, and should likewise not expect two domains to remain separated by a linker, even a helical or ‘rigid’ linker. Rather, the fusion protein will adopt a disordered ensemble of all available states, unless significant effort is directed towards localising it. This effort may include the careful selection of a linker, but will also be enhanced by engineering the domains themselves to associate favourably.^{335,336}

8.1.3 Linker length and composition has a large effect on sensor performance

In the case of the glycine sensor GlyFS, modest improvements to dynamic range were obtained through literature practices such as linker truncation and fluorophore reorientation. However, a satisfactory sensor was only obtained by dramatically lengthening one linker and altering its composition to favour helicity, changes that were inspired by MD simulations performed by the author. These changes to one linker improved dynamic range more than three-fold, and even minor changes to this linker designed to make it more flexible or longer were detrimental. The sensitivity of this result and the difficulty in explaining it in a satisfying manner suggest that there are yet insights to be had in modelling fusion proteins.

8.1.4 The design of sensors for a multitude of target molecules

This thesis describes the construction of soluble, fluorescence-based FRET biosensors for maltose, N-acetylneuraminic acid, arginine, glycine, and D-serine, all using techniques or insights developed by the author. These sensors have affinities and dynamic ranges appropriate for use *in vivo*.

8.1.5 The NANP repeat linker of the *Plasmodium* circumsporozoite protein produces a multivalent immune response

The circumsporozoite protein of the malaria parasite is both a crucial target for vaccines and a tantalising example of a natural fusion protein. Like synthetic fusion proteins, it is composed of two domains connected by a long repeat sequence. This linker sequence primarily consists of an (NANP)_n repeat, with the occasional (NVDP) sequence inserted throughout. It appears that this linker sequence has an

important role in protecting the parasite against the host's immune system by presenting an enticing decoy target to the immune system while limiting the scope of the immune response.

8.1.6 A REST2 regime efficiently samples disordered peptides

REST2 is a Hamiltonian replica exchange scheme for condensed phase simulations in which the potential energy of the solute is scaled at higher rungs of the exchange ladder but the potential energy of the solvent is held fixed. This yields a scheme similar to traditional replica exchange, but requiring many fewer replicas as the vast majority of the system, the solvent, is not scaled. This is ideal for systems that vary in size dramatically, as the box size and number of solvent molecules can be increased without incurring the cost of additional replicas. This helps to minimise finite size effects. A peptide of 15 amino acids could be simulated in a way that guarantees that periodic images do not come within the cut-off distance of each other, even when initiated in an extended conformation.

8.1.7 Linker peptides have characteristic free energy landscapes

The three linker sequence studied in chapter 7 — $(\text{NANP})_n$, $(\text{EAAAK})_n$, and $(\text{GGGS})_n$ — each possess remarkably different free energy landscapes. This is most evident when $n = 1$, where the so-called 'rigid' linker (EAAAK) favours a relatively collapsed state, with a substantial peak corresponding to an α -helix. By contrast, (NANP) favours an extended state, and (GGGS) has a broad, featureless landscape. Notably, these general shapes are recapitulated regardless of the context of the repeat despite their composition dependence. When $n = 3$, these features are largely averaged out, but each landscape remains characteristic of its peptide sequence. Thus, the length and composition of a linker interact in complex ways, which accounts for many researchers ability to produce sensors by varying only one of the two features. However, especially for very short linkers, the ability to choose a linker with a peak at the target distance of even only a handful of kilojoules per mole may assist in stabilising a target conformation when combined with the introduction of specific and targeted contacts between the domains.

8.1.8 AWH is well-suited to sampling protein–protein interactions

The interactions between domains of a fusion protein are difficult to model owing to their high intrinsic dimensionality and large barriers to unbinding. A multiple replica accelerated weight histogram approach, described in this thesis, can efficiently sample both orientational and translation degrees of freedom with a single collective variable, and converges all parts of the free energy surface uniformly rather than fully exploring one region before moving on to the next. Thus, it can be used to generate approximate solutions simply by terminating the simulation early.

8.1.9 A reweighting method can combine data from different resolutions of simulation

The free energy is a universal language of sampling. As it is simply the negative log likelihood, it can be compared across resolutions or force fields. By simulating different parts of a system independently, optimisations can be tailored to each component. The components can then be recombined by computing their free energy surfaces in a collective variable system shared by each simulation, and including any generalised corrections necessary to account for interaction energies that are missing from each simulation. Because these interaction energies are implicit or absent, this technique will not achieve the levels of accuracy available to more complex multi-scale approaches, but it has the great advantage of scaling linearly with the number of components rather than with the number of combinations of components.

This approach is demonstrated on a fusion protein by simulating the linker with an accurate, atomistic force field, and simulating the domain–domain interactions with a modified MARTINI force field. Distances between the domains and a putative FRET efficiency are computed for these systems. This application is ideal because new linkers and domain pairs may be incorporated with a linear amount of new simulation, rather than quadratic. In addition, the missing energy cross-terms only arise from non-bonded interactions between the linker and the domains, which is likely to average out for a given coordinate in collective variable space, and which when significant likely arises from a generalisable crowding effect.

8.2 Future work

8.2.1 Multi-state modelling of sensors

Sensors built on a conformational change of a binding protein often require intensive re-engineering of the binding core for satisfactory affinity and stability. Rational and computational approaches to this, like that described in chapter 4, typically only consider a single state, such as the closed bound state, when engineering mutations. However, at least four states are relevant, as described in the preface to chapter 5: (1) the bound closed state, (2) the bound open state, (3) the unbound closed state, and (4) the unbound open state. In that work, the critically important bound closed and unbound open states were explicitly considered. This practice should be extended to all four states, so that for example the unbound closed state is not stabilised alongside the bound closed state. Given the low computational cost of techniques like PROSS used in chapter 5, automated comparison of these states is a valuable target for future work.

8.2.2 Improvements to fusion protein modelling technique

In chapter 7, a computational method for efficiently and combinatorially simulating variants of fusion protein is proposed and demonstrated. Several improvements to the technique are proposed in that chapter. Firstly, a statistical free energy estimator, most likely MBAR,³²⁵ could be used in place of a simple histogram. At the cost of the increased complexity of evaluating each analysed frame at every rung of the exchange ladder, MBAR would provide uncertainties and allow the data from every rung to be incorporated into free energies. The AWH domain–domain simulations described in chapter 7 have several potential improvements. Rather than conducting replicas that build a shared bias, each simulation should be performed with its own bias. This eliminates communication between processes and is thus computationally favourable, and when combined with the randomly selected starting state already implemented allows replicas to be statistically independent. Longer simulation and additional replicas would improve statistics and could be made affordable by reducing the size of the system to suit the size of the linkers being examined.

8.2.3 Assess size of error induced by ignoring non-bonded linker–domain interactions

A key assumption underlying the efficient simulation of fusion proteins described in chapter 7 is that non-bonded interaction energies between the linker and the domains cancels out when averaged over the PMF, or otherwise that these energies are negligible. To be confident about this approximation, the described reweighting approach should be compared to a multiscaling approach that shares the same systems and force fields but includes these interaction terms explicitly. A virtual site multiscaling approach like that described by Wassenaar et al.¹²⁰ is ideally suited to this, as it is compatible with the AWH schema and allows coupling between force fields of different scales.

8.2.4 Compare fusion protein modelling technique to experiment

The ultimate judge of a computational method is its comparison to experiment, and these experiments must be performed to fully assess the technique described in chapter 7. The system studied in that chapter was designed for comparison between experiment and computation. It features a two domain fluorescent sensor, meaning that additional domains featured in most FRET sensors are absent and need not be modelled. By chemically labelling the binding protein as described in chapter 3, a two domain protein can be experimentally evaluated via fluorescence without any potential complications arising from symmetry between the two domains. In addition to comparing dynamic ranges and FRET efficiencies for sensors with varying linkers, fluorescence enables the possibility of single molecule experiments to precisely evaluate the entire free energy surface experimentally and therefore isolate the sources of any discovered errors.

References

- ¹J. A. Anderson, C. D. Lorenz and A. Travesset, ‘General purpose molecular dynamics simulations fully implemented on graphics processing units’, *Journal of Computational Physics* **227**, 5342–5359 (2008).
- ²S. Le Grand, A. W. Götz and R. C. Walker, ‘SPFP: Speed without compromise — a mixed precision model for GPU accelerated molecular dynamics simulations’, *Computer Physics Communications* **184**, 374–380 (2013).
- ³C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B. L. de Groot and H. Grubmüller, ‘More bang for your buck: Improved use of GPU nodes for GROMACS 2018’, *arXiv. Preprint*. <https://arxiv.org/abs/1903.05918> (2019).
- ⁴D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L.-S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y.-H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. Ben Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang and C. Young, ‘Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer’, in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC ’14* (2014), pp. 41–53.
- ⁵T. Fröhlking, M. Bernetti, N. Calonaci and G. Bussi, ‘Towards empirical force fields that match experimental observables’, *arXiv. Preprint*. <https://arxiv.org/abs/2004.01630> (2020).
- ⁶K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, ‘How fast-folding proteins fold’, *Science* **334**, 517–520 (2011).
- ⁷Y. Miao, F. Feixas, C. Eun and J. A. McCammon, ‘Accelerated molecular dynamics simulations of protein folding’, *Journal of Computational Chemistry* **36**, 1536–1549 (2015).
- ⁸H. Nguyen, J. Maier, H. Huang, V. Perrone and C. Simmerling, ‘Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent’, *Journal of the American Chemical Society* **136**, 13959–13962 (2014).

- ⁹A. Raval, S. Piana, M. P. Eastwood, R. O. Dror and D. E. Shaw, 'Refinement of protein structure homology models via long, all-atom molecular dynamics simulations.', *Proteins* **80**, 2071–9 (2012).
- ¹⁰L. Heo, C. F. Arbour and M. Feig, 'Driven to near-experimental accuracy by refinement via molecular dynamics simulations', *Proteins: Structure, Function, and Bioinformatics* **87**, 1263–1275 (2019).
- ¹¹R. J. Read, M. D. Sammito, A. Kryshtafovych and T. I. Croll, 'Evaluation of model refinement in CASP13', *Proteins: Structure, Function, and Bioinformatics* **87**, 1249–1262 (2019).
- ¹²G. de M. Seabra, R. C. Walker and A. E. Roitberg, 'Are current semiempirical methods better than force fields? a study from the thermodynamics perspective', *The Journal of Physical Chemistry A* **113**, 11938–11948 (2009).
- ¹³S. Can, S. Lacey, M. Gur, A. P. Carter and A. Yildiz, 'Directionality of dynein is controlled by the angle and length of its stalk', *Nature* **566**, 407–410 (2019).
- ¹⁴J. D. Durrant, S. E. Kochanek, L. Casalino, P. U. Jeong, A. C. Dommer and R. E. Amaro, 'Mesoscale all-atom influenza virus simulations suggest new substrate binding mechanism', *ACS Central Science*, 10.1021/acscentsci.9b01071 (2020).
- ¹⁵W. Pezeshkian, M. König, T. A. Wassenaar and S. J. Marrink, 'Backmapping triangulated surfaces to coarse-grained membrane models', *Nature Communications* **11**, 10.1038/s41467-020-16094-y (2020).
- ¹⁶E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel and R. A. Friesner, 'OPLS3: A force field providing broad coverage of drug-like small molecules and proteins', *Journal of Chemical Theory and Computation* **12**, 281–296 (2016).
- ¹⁷K. A. Beauchamp, Y.-S. Lin, R. Das and V. S. Pande, 'Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements', *Journal of Chemical Theory and Computation* **8**, 1409–1414 (2012).
- ¹⁸S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot and H. Grubmüller, 'Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment', *Journal of Chemical Theory and Computation* **11**, 5513–5524 (2015).
- ¹⁹J. Wong-ekkabut and M. Karttunen, 'The good, the bad and the user in soft matter simulations', *Biochimica et Biophysica Acta (BBA) - Biomembranes, Biosimulations of Lipid Membranes Coupled to Experiments* **1858**, 2529–2538 (2016).

- ²⁰V. V. Gapsys and B. B. L. de Groot, 'Comment on 'Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size'', *eLife* **8**, e44718 (2019).
- ²¹V. Agarwal, Y. Xue, B. Reif and N. R. Skrynnikov, 'Protein side-chain dynamics as observed by solution- and solid-state NMR spectroscopy: A similarity revealed', *Journal of the American Chemical Society* **130**, 16611–16621 (2008).
- ²²J. S. Fraser, H. van den Bedem, A. J. Samelson, P. T. Lang, J. M. Holton, N. Echols and T. Alber, 'Accessing protein conformational ensembles using room-temperature X-ray crystallography', *Proceedings of the National Academy of Sciences* **108**, 16247–16252 (2011).
- ²³D. A. Keedy, H. van den Bedem, D. A. Sivak, G. A. Petsko, D. Ringe, M. A. Wilson and J. S. Fraser, 'Crystal cryocooling distorts conformational heterogeneity in a model Michaelis complex of DHFR', *Structure* **22**, 899–910 (2014).
- ²⁴J. S. Fraser, K. Lindorff-Larsen and M. Bonomi, 'What will computational modeling approaches have to say in the era of atomistic Cryo-EM data?', *Journal of Chemical Information and Modeling*, 10.1021/acs.jcim.0c00123 (2020).
- ²⁵S. M. C. Gobeil, M. C. C. J. C. Ebert, J. Park, D. Gagné, N. Doucet, A. M. Berghuis, J. Pleiss and J. N. Pelletier, 'The structural dynamics of engineered β -lactamases vary broadly on three timescales yet sustain native function', *Scientific Reports* **9**, 10.1038/s41598-019-42866-8 (2019).
- ²⁶H. van den Bedem and J. S. Fraser, 'Integrative, dynamic structural biology at atomic resolution—it's about time', *Nature Methods* **12**, 307–318 (2015).
- ²⁷S. Bottaro and K. Lindorff-Larsen, 'Biophysical experiments and biomolecular simulations: A perfect match?', *Science* **361**, 355–360 (2018).
- ²⁸S. S. Chen, R. R. P. Wiewiora, F. F. Meng, N. N. Babault, A. A. Ma, W. W. Yu, K. K. Qian, H. H. Hu, H. H. Zou, J. J. Wang, S. S. Fan, G. G. Blum, F. F. Pittella-Silva, K. K. A. Beauchamp, W. W. Tempel, H. H. Jiang, K. K. Chen, R. R. J. Skene, Y. Y. G. Zheng, P. P. J. Brown, J. J. Jin, C. C. Luo, J. J. D. Chodera and M. M. Luo, 'The dynamic conformational landscape of the protein methyltransferase SETD8', *eLife* **8**, e45403 (2019).
- ²⁹S. Honda, K. Yamasaki, Y. Sawada and H. Morii, '10 residue folded peptide designed by segment statistics', *Structure* **12**, 1507–1518 (2004).
- ³⁰S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, 'Coarse-grained protein models and their applications', *Chemical Reviews* **116**, 7898–7936 (2016).

- ³¹A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Židek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis, 'Improved protein structure prediction using potentials from deep learning', *Nature*, 1–5 (2020).
- ³²M. Karplus, 'Behind the folding funnel diagram', *Nature Chemical Biology* **7**, 401–404 (2011).
- ³³R. R. F. Alford, A. A. Leaver-Fay, J. J. R. Jeliazkov, M. M. J. O'Meara, F. F. P. DiMaio, H. H. Park, M. M. V. Shapovalov, P. P. D. Renfrew, V. V. K. Mulligan, K. K. Kappel, J. J. W. Labonte, M. M. S. Pacella, R. R. Bonneau, P. P. Bradley, R. R. L. Dunbrack, R. R. Das, D. D. Baker, B. B. Kuhlman, T. T. Kortemme and J. J. J. Gray, 'The Rosetta all-atom energy function for macromolecular modeling and design.', *Journal of chemical theory and computation* **13**, 3031–3048 (2017).
- ³⁴N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 'Equation of state calculations by fast computing machines', *The Journal of Chemical Physics* **21**, 1087–1092 (1953).
- ³⁵J. Cheng, M.-H. Choe, A. Elofsson, K.-S. Han, J. Hou, A. H. A. Maghrabi, L. J. McGuffin, D. Menéndez-Hurtado, K. Olechnovič, T. Schwede, G. Studer, K. Uziela, Č. Venclovas and B. Wallner, 'Estimation of model accuracy in CASP13', *Proteins: Structure, Function, and Bioinformatics*, 10.1002/prot.25767 (2019).
- ³⁶F. Noé, S. Olsson, J. Köhler and H. Wu, 'Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning', *Science* **365**, eaaw1147 (2019).
- ³⁷L. Monticelli, C. Simões, L. Belvisi and G. Colombo, 'Assessing the influence of electrostatic schemes on molecular dynamics simulations of secondary structure forming peptides', *Journal of Physics: Condensed Matter* **18**, S329–S345 (2006).
- ³⁸U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, 'A smooth particle mesh Ewald method', *The Journal of Chemical Physics* **103**, 8577–8593 (1995).
- ³⁹Y. Shan, J. L. Klepeis, M. P. Eastwood, R. O. Dror and D. E. Shaw, 'Gaussian split ewald: a fast ewald mesh method for molecular simulation', *The Journal of Chemical Physics* **122**, 054101 (2005).
- ⁴⁰J. Barker and R. Watts, 'Monte carlo studies of the dielectric properties of water-like models', *Molecular Physics* **26**, 789–792 (1973).
- ⁴¹B. Leimkuhler and C. Matthews, 'Robust and efficient configurational molecular sampling via Langevin dynamics', *The Journal of Chemical Physics* **138**, 174102 (2013).

- ⁴²D. A. Sivak, J. D. Chodera and G. E. Crooks, 'Using nonequilibrium fluctuation theorems to understand and correct errors in equilibrium and nonequilibrium simulations of discrete Langevin dynamics', *Physical Review X* **3**, 10.1103/physrevx.3.011007 (2013).
- ⁴³J. J. Fass, D. D. A. Sivak, G. G. E. Crooks, K. K. A. Beauchamp, B. B. Leimkuhler and J. J. D. Chodera, 'Quantifying configuration-sampling error in langevin simulations of complex molecular systems.', *Entropy (Basel, Switzerland)* **20**, 318 (2018).
- ⁴⁴P. P. V. Coveney and S. S. Wan, 'On the calculation of equilibrium thermodynamic properties from molecular dynamics.', *Physical chemistry chemical physics : PCCP* **18**, 30236–30240 (2016).
- ⁴⁵A. A. Vitalis and R. R. V. Pappu, 'Methods for Monte Carlo simulations of biomacromolecules.', *Annual reports in computational chemistry* **5**, 49–76 (2009).
- ⁴⁶B. F. E. Curchod and T. J. Martínez, 'Ab initio nonadiabatic quantum molecular dynamics', *Chemical Reviews* **118**, 3305–3336 (2018).
- ⁴⁷M. Winger, D. Trzesniak, R. Baron and W. F. van Gunsteren, 'On using a too large integration time step in molecular dynamics simulations of coarse-grained molecular models', *Phys. Chem. Chem. Phys.* **11**, 1934–1941 (2009).
- ⁴⁸S. Duane, A. Kennedy, B. J. Pendleton and D. Roweth, 'Hybrid Monte Carlo', *Physics Letters B* **195**, 216–222 (1987).
- ⁴⁹C. R. Sweet, S. S. Hampton, R. D. Skeel and J. A. Izaguirre, 'A separable shadow Hamiltonian hybrid Monte Carlo method', *The Journal of Chemical Physics* **131**, 174106 (2009).
- ⁵⁰C. Grebogi, S. M. Hammel, J. A. Yorke and T. Sauer, 'Shadowing of physical trajectories in chaotic dynamics: Containment and refinement', *Physical Review Letters* **65**, 1527–1530 (1990).
- ⁵¹W. B. Hayes, 'Shadowing high-dimensional Hamiltonian systems: The gravitational N-body problem', *Physical Review Letters* **90**, 10.1103/physrevlett.90.054104 (2003).
- ⁵²F. Hoffmann, F. A. A. Mulder and L. V. Schäfer, 'Accurate methyl group dynamics in protein simulations with AMBER force fields', *The Journal of Physical Chemistry B* **122**, 5038–5048 (2018).
- ⁵³F. Hoffmann, F. A. A. Mulder and L. V. Schäfer, 'Predicting NMR relaxation of proteins from molecular dynamics simulations with accurate methyl rotation barriers', ChemRxiv. Preprint. <https://doi.org/10.26434/chemrxiv.8982338.v4>, 2020.

- ⁵⁴S. S. D. Lotz and A. A. Dickson, 'Unbiased molecular dynamics of 11 min timescale drug unbinding reveals transition state stabilizing interactions.', *Journal of the American Chemical Society* **140**, 618–628 (2018).
- ⁵⁵R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand and R. C. Walker, 'Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald', *Journal of Chemical Theory and Computation* **9**, 3878–3888 (2013).
- ⁵⁶B. Knapp, L. Ospina and C. M. Deane, 'Avoiding false positive conclusions in molecular simulation: The importance of replicas', *Journal of Chemical Theory and Computation* **14**, 6127–6138 (2018).
- ⁵⁷K. Sikic, S. Tomic and O. Carugo, 'Systematic comparison of crystal and NMR protein structures deposited in the Protein Data Bank', *The Open Biochemistry Journal* **4**, 83–95 (2010).
- ⁵⁸J. J. D. Chodera, 'A simple method for automated equilibration detection in molecular simulations.', *Journal of chemical theory and computation* **12**, 1799–805 (2016).
- ⁵⁹H. Paliwal and M. R. Shirts, 'A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods', *Journal of Chemical Theory and Computation* **7**, 4115–4134 (2011).
- ⁶⁰A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius and D. M. Zuckerman, 'Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0]', *Living Journal of Computational Molecular Science* **1**, 10.33011/livecoms.1.1.5067 (2019).
- ⁶¹S. Riniker, 'Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview', *Journal of Chemical Information and Modeling* **58**, 565–578 (2018).
- ⁶²J. A. Lemkul, 'Pairwise-additive and polarizable atomistic force fields for molecular dynamics simulations of proteins', in *Computational approaches for understanding dynamical systems: protein folding and assembly* (Elsevier, 2020), pp. 1–71.
- ⁶³A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin and M. Karplus, 'All-atom empirical potential for molecular modeling and dynamics studies of proteins', *The Journal of Physical Chemistry B* **102**, 3586–3616 (1998).

- ⁶⁴A. D. Mackerell, 'Empirical force fields for biological macromolecules: Overview and issues', [Journal of Computational Chemistry](#) **25**, 1584–1604 (2004).
- ⁶⁵J. J. Huang and A. A. D. MacKerell, 'CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data.', [Journal of computational chemistry](#) **34**, 2135–45 (2013).
- ⁶⁶J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller and A. D. MacKerell, 'CHARMM36m: an improved force field for folded and intrinsically disordered proteins', [Nature Methods](#) **14**, 71–73 (2016).
- ⁶⁷S. Piana, K. Lindorff-Larsen and D. E. Shaw, 'How robust are protein folding simulations with respect to force field parameterization?', [Biophysical Journal](#) **100**, L47–L49 (2011).
- ⁶⁸K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror and D. E. Shaw, 'Systematic validation of protein force fields against experimental data.', [PLoS ONE](#) **7**, e32131 (2012).
- ⁶⁹A. M. Fluitt and J. J. de Pablo, 'An analysis of biomolecular force fields for simulations of polyglutamine in solution', [Biophysical Journal](#) **109**, 1009–1018 (2015).
- ⁷⁰M. Carballo-Pacheco and B. Strodel, 'Comparison of force fields for Alzheimer's A β_{42} : A case study for intrinsically disordered proteins', [Protein Science](#) **26**, 174–185 (2017).
- ⁷¹D. Petrović, X. Wang and B. Strodel, 'How accurately do force fields represent protein side chain ensembles?', [Proteins: Structure, Function, and Bioinformatics](#) **86**, 935–944 (2018).
- ⁷²P. Robustelli, S. Piana and D. E. Shaw, 'Developing a molecular dynamics force field for both folded and disordered protein states', [Proceedings of the National Academy of Sciences](#) **115**, E4758–E4766 (2018).
- ⁷³M. U. Rahman, A. U. Rehman, H. Liu and H.-F. Chen, 'Comparison and evaluation of force fields for intrinsically disordered proteins', [Journal of Chemical Information and Modeling](#) **60**, PMID: 32816485, 4912–4923 (2020).
- ⁷⁴J. Wang, P. Cieplak and P. A. Kollman, 'How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?', [Journal of Computational Chemistry](#) **21**, 1049–1074 (2000).
- ⁷⁵Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. Kollman, 'A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations', [Journal of Computational Chemistry](#) **24**, 1999–2012 (2003).

- ⁷⁶V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, 'Comparison of multiple Amber force fields and development of improved protein backbone parameters', *Proteins: Structure, Function, and Bioinformatics* **65**, 712–725 (2006).
- ⁷⁷J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, 'ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB', *Journal of Chemical Theory and Computation* **11**, 3696–3713 (2015).
- ⁷⁸K. T. Debiec, D. S. Cerutti, L. R. Baker, A. M. Gronenborn, D. A. Case and L. T. Chong, 'Further along the road less traveled: AMBER ff15ipq, an original protein force field built on a self-consistent physical model', *Journal of Chemical Theory and Computation* **12**, 3926–3947 (2016).
- ⁷⁹L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martínez and V. S. Pande, 'Building a more predictive protein force field: A systematic and reproducible route to AMBER-FB15', *The Journal of Physical Chemistry B* **121**, 4023–4039 (2017).
- ⁸⁰R. B. Best and G. Hummer, 'Optimized molecular dynamics force fields applied to the helix–coil transition of polypeptides', *The Journal of Physical Chemistry B* **113**, 9004–9015 (2009).
- ⁸¹K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, 'Improved side-chain torsion potentials for the Amber ff99SB protein force field', *Proteins: Structure, Function, and Bioinformatics*, NA–NA (2010).
- ⁸²R. B. Best, D. de Sancho and J. Mittal, 'Residue-specific α -helix propensities from molecular simulation', *Biophysical Journal* **102**, 1462–1467 (2012).
- ⁸³S. Piana, A. G. Donchev, P. Robustelli and D. E. Shaw, 'Water dispersion interactions strongly influence simulated structural properties of disordered protein states', *The Journal of Physical Chemistry B* **119**, 5113–5123 (2015).
- ⁸⁴S. Piana, P. Robustelli, D. Tan, S. Chen and D. E. Shaw, 'Development of a force field for the simulation of single-chain proteins and protein–protein complexes', *Journal of Chemical Theory and Computation*, 10.1021/acs.jctc.9b00251 (2020).
- ⁸⁵R. B. Best and J. Mittal, 'Protein simulations with an optimized water model: Cooperative helix formation and temperature-induced unfolded state collapse', *The Journal of Physical Chemistry B* **114**, 14916–14923 (2010).
- ⁸⁶R. B. Best, W. Zheng and J. Mittal, 'Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association', *Journal of Chemical Theory and Computation* **10**, 5113–5124 (2014).

- ⁸⁷D. Song, W. Wang, W. Ye, D. Ji, R. Luo and H.-F. Chen, 'ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins', *Chemical Biology & Drug Design* **89**, 5–15 (2016).
- ⁸⁸D. Song, R. Luo and H.-F. Chen, 'The IDP-Specific force field ff14IDPSFF improves the conformer sampling of intrinsically disordered proteins', *Journal of Chemical Information and Modeling* **57**, 1166–1178 (2017).
- ⁸⁹D. S. Cerutti, W. C. Swope, J. E. Rice and D. A. Case, 'Ff14ipq: a self-consistent force field for condensed-phase simulations of proteins', *Journal of Chemical Theory and Computation* **10**, 4515–4534 (2014).
- ⁹⁰D.-W. Li and R. Brüschweiler, 'NMR-Based protein potentials', *Angewandte Chemie International Edition* **49**, 6778–6780 (2010).
- ⁹¹G. A. Khoury, J. Smadbeck, P. Tamamis, A. C. Vandris, C. A. Kieslich and C. A. Floudas, 'Force-field_NCAA: Ab initio charge parameters to aid in the discovery and design of therapeutic proteins and peptides with unnatural amino acids and their application to complement inhibitors of the compstatin family', *ACS Synthetic Biology* **3**, 855–869 (2014).
- ⁹²G. A. Khoury, J. P. Thompson, J. Smadbeck, C. A. Kieslich and C. A. Floudas, 'Force-field_PTM: Ab initio charge and AMBER forcefield parameters for frequently occurring post-translational modifications', *Journal of Chemical Theory and Computation* **9**, 5653–5674 (2013).
- ⁹³W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, 'Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids', *Journal of the American Chemical Society* **118**, 11225–11236 (1996).
- ⁹⁴M. M. Reif, P. H. Hünenberger and C. Oostenbrink, 'New interaction parameters for charged amino acid side chains in the gromos force field', *Journal of Chemical Theory and Computation* **8**, PMID: 26593015, 3705–3723 (2012).
- ⁹⁵N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark and W. F. Van Gunsteren, 'Definition and testing of the GROMOS force-field versions 54A7 and 54B7', *European Biophysics Journal* **40**, 843–856 (2011).
- ⁹⁶C. T. Andrews and A. H. Elcock, 'Molecular dynamics simulations of highly crowded amino acid solutions: comparisons of eight different force field combinations with experiment and with each other', *Journal of Chemical Theory and Computation* **9**, 4585–4602 (2013).
- ⁹⁷A. A. Sandoval-Perez, K. K. Pluhackova and R. R. A. Böckmann, 'Critical comparison of biomembrane force fields: Protein-lipid interactions at the membrane interface.', *Journal of chemical theory and computation* **13**, 2310–2321 (2017).

- ⁹⁸A. Plazinska and W. Plazinski, 'Comparison of carbohydrate force fields in Molecular Dynamics simulations of protein-carbohydrate complexes', *Journal of Chemical Theory and Computation* **17**, PMID: 33703894, 2575–2585 (2021).
- ⁹⁹E. A. Cino, W.-Y. Choy and M. Karttunen, 'Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations', *Journal of Chemical Theory and Computation* **8**, 2725–2740 (2012).
- ¹⁰⁰S. J. Marrink and D. P. Tieleman, 'Perspective on the Martini model.', *Chemical Society reviews* **42**, 6801–22 (2013).
- ¹⁰¹K. H. Kanekal and T. Bereau, 'Resolution limit of data-driven coarse-grained models spanning chemical space', *The Journal of Chemical Physics* **151**, 164106 (2019).
- ¹⁰²R. Alessandri, P. C. T. Souza, S. Thallmair, M. N. Melo, A. H. de Vries and S. J. Marrink, 'Pitfalls of the Martini model', *Journal of Chemical Theory and Computation* **15**, 5448–5460 (2019).
- ¹⁰³S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. de Vries, 'The MARTINI force field: coarse grained model for biomolecular simulations.', *The Journal of Physical Chemistry B* **111**, 7812–24 (2007).
- ¹⁰⁴C. A. López, Z. Sovova, F. J. van Eerden, A. H. de Vries and S. J. Marrink, 'Martini force field parameters for glycolipids', *Journal of Chemical Theory and Computation* **9**, 1694–1708 (2013).
- ¹⁰⁵T. A. Wassenaar, H. I. Ingólfsson, R. A. Böckmann, D. P. Tieleman and S. J. Marrink, 'Computational Lipidomics with insane : A Versatile Tool for Generating Custom Membranes for Molecular Simulations', *Journal of Chemical Theory and Computation* **11**, 2144–2155 (2015).
- ¹⁰⁶S. J. Marrink, 'Computational modeling of realistic cell membranes', *Biophysical Journal* **114**, 367a (2018).
- ¹⁰⁷L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S.-J. Marrink, 'The MARTINI coarse-grained force field: Extension to proteins', *Journal of Chemical Theory and Computation* **4**, 819–834 (2008).
- ¹⁰⁸X. Periole, M. Cavalli, S.-J. Marrink and M. a. Ceruso, 'Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition', *Journal of Chemical Theory and Computation* **5**, 2531–2543 (2009).
- ¹⁰⁹D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman and S. J. Marrink, 'Improved parameters for the Martini coarse-grained protein force field', *Journal of Chemical Theory and Computation* **9**, 687–697 (2013).

- ¹¹⁰H. M. Khan, P. C. Telles de Souza, S. Thallmair, J. Barnoud, A. H. De Vries, S. J. Marrink and N. Reuter, 'Capturing choline-aromatics cation- π interactions in the MARTINI force field', *Journal of Chemical Theory and Computation*, [10.1021/acs.jctc.9b01194](https://doi.org/10.1021/acs.jctc.9b01194) (2020).
- ¹¹¹C. A. López, A. J. Rzepiela, A. H. de Vries, L. Dijkhuizen, P. H. Hünenberger and S. J. Marrink, 'Martini coarse-grained force field: Extension to carbohydrates', *Journal of Chemical Theory and Computation* **5**, 3195–3210 (2009).
- ¹¹²J. J. Uusitalo, H. I. Ingólfsson, P. Akhshi, D. P. Tieleman and S. J. Marrink, 'Martini coarse-grained force field: Extension to DNA', *Journal of Chemical Theory and Computation* **11**, 3932–3945 (2015).
- ¹¹³J. J. Uusitalo, H. I. Ingólfsson, S. J. Marrink and I. Faustino, 'Martini coarse-grained force field: Extension to RNA', *Biophysical Journal* **113**, 246–256 (2017).
- ¹¹⁴S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta and S. J. Marrink, 'Polarizable water model for the coarse-grained MARTINI force field', *PLoS Computational Biology* **6**, edited by M. Levitt, [e1000810](https://doi.org/10.1371/journal.pcbi.1000810) (2010).
- ¹¹⁵Z. Wu, Q. Cui and A. Yethiraj, 'A new coarse-grained model for water: The importance of electrostatic interactions', *The Journal of Physical Chemistry B* **114**, 10524–10529 (2010).
- ¹¹⁶S. Riniker and W. F. van Gunsteren, 'A simple, efficient polarizable coarse-grained water model for molecular dynamics simulations', *The Journal of Chemical Physics* **134**, 084110 (2011).
- ¹¹⁷J. Michalowsky, L. V. Schäfer, C. Holm and J. Smiatek, 'A refined polarizable water model for the coarse-grained MARTINI force field with long-range electrostatic interactions', *The Journal of Chemical Physics* **146**, 054501 (2017).
- ¹¹⁸D. H. de Jong, S. Baoukina, H. I. Ingólfsson and S. J. Marrink, 'Martini straight: Boosting performance using a shorter cutoff and GPUs', *Computer Physics Communications*, [10.1016/j.cpc.2015.09.014](https://doi.org/10.1016/j.cpc.2015.09.014) (2015).
- ¹¹⁹C. Arnarez, J. J. Uusitalo, M. F. Masman, H. I. Ingólfsson, D. H. de Jong, M. N. Melo, X. Periole, A. H. de Vries and S. J. Marrink, 'Dry Martini, a coarse-grained force field for lipid membrane simulations with implicit solvent', *Journal of Chemical Theory and Computation* **11**, 260–275 (2014).
- ¹²⁰T. A. Wassenaar, H. I. Ingólfsson, M. Prieß, S. J. Marrink and L. V. Schäfer, 'Mixing MARTINI: Electrostatic Coupling in Hybrid Atomistic–Coarse-Grained Biomolecular Simulations', *The Journal of Physical Chemistry B* **117**, 3516–3530 (2013).

- ¹²¹J. Zavadlav, S. J. Marrink and M. Praprotnik, 'SWINGER: A clustering algorithm for concurrent coupling of atomistic and supramolecular liquids', [Interface Focus](#) **9**, 20180075 (2019).
- ¹²²S. Thallmair, P. A. Vainikka and S. J. Marrink, 'Lipid fingerprints and cofactor dynamics of light-harvesting complex ii in different membranes', [Biophysical Journal](#) **116**, 1446–1455 (2019).
- ¹²³J. K. Marzinek, D. A. Holdbrook, R. G. Huber, C. Verma and P. J. Bond, 'Pushing the envelope: Dengue viral membrane coaxed into shape by molecular simulations', [Structure](#) **24**, 1410–1420 (2016).
- ¹²⁴K. K. Sharma, X.-X. Lim, S. N. Tantirimudalige, A. Gupta, J. K. Marzinek, D. Holdbrook, X. Y. E. Lim, P. J. Bond, G. S. Anand and T. Wohland, 'Infectivity of Dengue virus serotypes 1 and 2 is correlated with E-protein intrinsic dynamics but not to envelope conformations', [Structure](#) **27**, 618–630.e4 (2019).
- ¹²⁵A. B. Poma, M. Cieplak and P. E. Theodorakis, 'Combining the MARTINI and Structure-Based Coarse-Grained Approaches for the Molecular Dynamics Studies of Conformational Transitions in Proteins', [Journal of Chemical Theory and Computation](#) **13**, 1366–1374 (2017).
- ¹²⁶A. C. Stark, C. T. Andrews and A. H. Elcock, 'Toward Optimized Potential Functions for Protein–Protein Interactions in Aqueous Solutions: Osmotic Second Virial Coefficient Calculations Using the MARTINI Coarse-Grained Force Field', [Journal of Chemical Theory and Computation](#) **9**, 4176–4185 (2013).
- ¹²⁷M. Javanainen, H. Martinez-Seara and I. Vattulainen, 'Excessive aggregation of membrane proteins in the Martini model', [PLOS ONE](#) **12**, edited by E. Papaleo, e0187936 (2017).
- ¹²⁸J. Liu, L. Qiu, R. Alessandri, X. Qiu, G. Portale, J. Dong, W. Talsma, G. Ye, A. A. Sengrian, P. C. T. Souza, M. A. Loi, R. C. Chiechi, S. J. Marrink, J. C. Hummelen and L. J. A. Koster, 'Enhancing molecular n-type doping of donor-acceptor copolymers by tailoring side chains', [Advanced Materials](#) **30**, 1704630 (2018).
- ¹²⁹P. C. T. Souza, S. Thallmair, S. J. Marrink and R. Mera-Adasme, 'An allosteric pathway in copper, zinc superoxide dismutase unravels the molecular mechanism of the G93A amyotrophic lateral sclerosis-linked mutation', [The Journal of Physical Chemistry Letters](#) **10**, 7740–7744 (2019).
- ¹³⁰A. Wang, Z. Zhang and G. Li, 'Higher accuracy achieved in the simulations of protein structure refinement, protein folding, and intrinsically disordered proteins using polarizable force fields', [The Journal of Physical Chemistry Letters](#) **9**, 7110–7116 (2018).

- ¹³¹S. Decherchi, M. Masetti, I. Vyalov and W. Rocchia, 'Implicit solvent methods for free energy estimation', *European Journal of Medicinal Chemistry* **91**, 27–42 (2015).
- ¹³²A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand and R. C. Walker, 'Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born', *Journal of Chemical Theory and Computation* **8**, 1542–1555 (2012).
- ¹³³G. Copie, F. Cleri, R. Blossey and M. F. Lensink, 'On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes', *Scientific Reports* **6**, 10.1038/srep38259 (2016).
- ¹³⁴T. T. Nguyen, M. H. Viet and M. S. Li, 'Effects of water models on binding affinity: Evidence from all-atom simulation of binding of Tamiflu to A/H5N1 neuraminidase', *The Scientific World Journal* **2014**, 1–14 (2014).
- ¹³⁵W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, 'Comparison of simple potential functions for simulating liquid water', *The Journal of Chemical Physics* **79**, 926–935 (1983).
- ¹³⁶F. Sajadi and C. N. Rowley, 'Simulations of lipid bilayers using the CHARMM36 force field with the TIP3P-FB and TIP4P-FB water models', *PeerJ* **6**, e5472 (2018).
- ¹³⁷H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, 'The missing term in effective pair potentials', *The Journal of Physical Chemistry* **91**, 6269–6271 (1987).
- ¹³⁸H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, 'Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew', *The Journal of Chemical Physics* **120**, 9665–9678 (2004).
- ¹³⁹D. J. Price and C. L. Brooks, 'A modified TIP3P water potential for simulation with ewald summation', *The Journal of Chemical Physics* **121**, 10096–10103 (2004).
- ¹⁴⁰D. S. Cerutti, P. L. Freddolino, R. E. Duke and D. A. Case, 'Simulations of a protein crystal with a high resolution X-ray structure: evaluation of force fields and water models', *The Journal of Physical Chemistry B* **114**, 12811–12824 (2010).
- ¹⁴¹J. L. F. Abascal and C. Vega, 'A general purpose model for the condensed phases of water: TIP4P/2005', *The Journal of Chemical Physics* **123**, 234505 (2005).
- ¹⁴²D. R. Nutt and J. C. Smith, 'Molecular dynamics simulations of proteins: Can the explicit water model be varied?', *Journal of Chemical Theory and Computation* **3**, 1550–1560 (2007).
- ¹⁴³D. C. Glass, M. Krishnan, D. R. Nutt and J. C. Smith, 'Temperature dependence of protein dynamics simulated with three different water models', *Journal of Chemical Theory and Computation* **6**, 1390–1400 (2010).

- ¹⁴⁴A. Saxena and D. Sept, 'Multisite ion models that improve coordination and free energy calculations in molecular dynamics simulations', *Journal of Chemical Theory and Computation* **9**, 3538–3542 (2013).
- ¹⁴⁵P. Li, L. F. Song and K. M. Merz, 'Systematic parameterization of monovalent ions employing the nonbonded model', *Journal of Chemical Theory and Computation* **11**, 1645–1657 (2015).
- ¹⁴⁶Y. Jiang, H. Zhang and T. Tan, 'Rational design of methodology-independent metal parameters using a nonbonded dummy model', *Journal of Chemical Theory and Computation* **12**, 3250–3260 (2016).
- ¹⁴⁷F. Duarte, P. Bauer, A. Barrozo, B. A. Amrein, M. Purg, J. Åqvist and S. C. L. Kamerlin, 'Force field independent metal parameters using a nonbonded dummy model', *The Journal of Physical Chemistry B* **118**, 4351–4362 (2014).
- ¹⁴⁸P. Li, L. F. Song and K. M. Merz, 'Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water', *The Journal of Physical Chemistry B* **119**, 883–895 (2014).
- ¹⁴⁹K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. Mackerell, 'CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields', *Journal of Computational Chemistry*, NA–NA (2009).
- ¹⁵⁰J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, 'Development and testing of a general AMBER force field', *Journal of Computational Chemistry* **25**, 1157–1174 (2004).
- ¹⁵¹C. Caleman, P. J. van Maaren, M. Hong, J. S. Hub, L. T. Costa and D. van der Spoel, 'Force field benchmark of organic liquids: Density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant', *Journal of Chemical Theory and Computation* **8**, 61–74 (2011).
- ¹⁵²H. S. Muddana and M. K. Gilson, 'Prediction of SAMPL3 host–guest binding affinities: Evaluating the accuracy of generalized force-fields', *Journal of Computer-Aided Molecular Design* **26**, 517–525 (2012).
- ¹⁵³V. Kumar, K. S. Rane, S. Wierzbowski, M. Shaik and J. R. Errington, 'Evaluation of the performance of GAFF and CGenFF in the prediction of liquid–vapor saturation properties of naphthalene derivatives', *Industrial & Engineering Chemistry Research* **53**, 16072–16081 (2014).

- ¹⁵⁴B. A. C. Horta, P. T. Merz, P. F. J. Fuchs, J. Dolenc, S. Riniker and P. H. Hünenberger, ‘A GROMOS-compatible force field for small organic molecules in the condensed phase: the 2016H66 parameter set’, *Journal of Chemical Theory and Computation* **12**, PMID: 27248705, 3825–3850 (2016).
- ¹⁵⁵A. A. K. Malde, L. L. Zuo, M. M. Breeze, M. M. Stroet, D. D. Poger, P. P. C. Nair, C. C. Oostenbrink and A. A. E. Mark, ‘An automated force field topology builder (ATB) and repository: Version 1.0.’, *Journal of Chemical Theory and Computation* **7**, 4026–37 (2011).
- ¹⁵⁶M. M. Stroet, B. B. Caron, K. K. M. Visscher, D. D. P. Geerke, A. A. K. Malde and A. A. E. Mark, ‘Automated topology builder version 3.0: Prediction of solvation free enthalpies in water and hexane.’, *Journal of chemical theory and computation* **14**, 5834–5845 (2018).
- ¹⁵⁷D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, M. R. Shirts, M. K. Gilson and P. K. Eastman, ‘Open Force Field Consortium: Escaping atom types using direct chemical perception with SMIRNOFF v0.1’, *bioRxiv*, 10.1101/286542 (2018).
- ¹⁵⁸D. Slochow, N. Henriksen, L.-P. Wang, J. Chodera, D. Mobley and M. Gilson, ‘Binding thermodynamics of host-guest systems with SMIRNOFF99Frosst 1.0.5 from the Open Force Field Initiative’, ChemRxiv. Preprint. <https://doi.org/10.26434/chemrxiv.9159872.v2>, 2019.
- ¹⁵⁹J. J. Galano-Frutos and J. Sancho, ‘Accurate calculation of Barnase and SNase folding energetics using short molecular dynamics simulations and an atomistic model of the unfolded ensemble: Evaluation of force fields and water models’, *Journal of Chemical Information and Modeling* **59**, 4350–4360 (2019).
- ¹⁶⁰P. P. T. Merz and M. M. R. Shirts, ‘Testing for physical validity in molecular simulations.’, *PloS one* **13**, e0202764 (2018).
- ¹⁶¹D. E. Shaw, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Lerardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. M. Deneroff, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S. C. Wang, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson and K. J. Bowers, ‘Anton, a special-purpose machine for molecular dynamics simulation’, *Communications of the ACM* **51**, 91 (2008).
- ¹⁶²H. H. Loeffler, S. Bosisio, G. Duarte Ramos Matos, D. Suh, B. Roux, D. L. Mobley and J. Michel, ‘Reproducibility of free energy calculations across different molecular simulation software packages’, *Journal of Chemical Theory and Computation* **14**, 5567–5582 (2018).

- ¹⁶³A. Rizzi, T. Jensen, D. R. Slochower, M. Aldeghi, V. Gapsys, D. Ntekoumes, S. Bosisio, M. Papadourakis, N. M. Henriksen, B. L. de Groot, Z. Cournia, A. Dickson, J. Michel, M. K. Gilson, M. R. Shirts, D. L. Mobley and J. D. Chodera, 'The SAMPL6 SAMPLing challenge: Assessing the reliability and efficiency of binding free energy calculations', [bioRxiv](#), **10.1101/795005** (2019).
- ¹⁶⁴H. Berendsen, D. van der Spoel and R. van Drunen, 'GROMACS: A message-passing parallel molecular dynamics implementation', [Computer Physics Communications](#) **91**, 43–56 (1995).
- ¹⁶⁵E. Lindahl, B. Hess and D. van der Spoel, 'GROMACS 3.0: a package for molecular simulation and trajectory analysis', [Journal of Molecular Modeling](#) **7**, 306–317 (2001).
- ¹⁶⁶D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, 'GROMACS: fast, flexible, and free.', [Journal of computational chemistry](#) **26**, 1701–18 (2005).
- ¹⁶⁷B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, 'GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation', [Journal of Chemical Theory and Computation](#) **4**, 435–447 (2008).
- ¹⁶⁸S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, 'GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.', [Bioinformatics \(Oxford, England\)](#) **29**, 845–54 (2013).
- ¹⁶⁹M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, 'GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers', [SoftwareX](#) **1-2**, 19–25 (2015).
- ¹⁷⁰M. E. Irrgang, J. M. Hays and P. M. Kasson, 'gmxml: a high-level interface for advanced control and extension of molecular dynamics simulations', [Bioinformatics](#) **34**, edited by A. Valencia, 3945–3947 (2018).
- ¹⁷¹D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, 'The Amber biomolecular simulation programs', [Journal of Computational Chemistry](#) **26**, 1668–1688 (2005).
- ¹⁷²R. Salomon-Ferrer, D. A. Case and R. C. Walker, 'An overview of the Amber biomolecular simulation package', [Wiley Interdisciplinary Reviews: Computational Molecular Science](#) **3**, 198–210 (2012).
- ¹⁷³P. Eastman and V. Pande, 'OpenMM: A hardware-independent framework for molecular simulations', [Computing in Science & Engineering](#) **12**, 34–39 (2010).

- ¹⁷⁴P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts and V. S. Pande, 'OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation', [Journal of Chemical Theory and Computation](#) **9**, 461–469 (2012).
- ¹⁷⁵P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, 'OpenMM 7: Rapid development of high performance algorithms for molecular dynamics', [PLOS Computational Biology](#) **13**, edited by R. Gentleman, e1005659 (2017).
- ¹⁷⁶J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, 'Scalable molecular dynamics with NAMD', [Journal of Computational Chemistry](#) **26**, 1781–1802 (2005).
- ¹⁷⁷K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolosvary, M. A. Moraes, F. D. Sacerdoti and et al., 'Scalable algorithms for molecular dynamics simulations on commodity clusters', in [Proceedings of the 2006 acm/ieee conference on supercomputing](#), SC '06 (2006), 84–es.
- ¹⁷⁸J. Jung, C. Kobayashi, K. Kasahara, C. Tan, A. Kuroda, K. Minami, S. Ishiduki, T. Nishiki, H. Inoue, Y. Ishikawa, M. Feig and Y. Sugita, 'New parallel computing algorithm of molecular dynamics for extremely huge scale biological systems', [Journal of Computational Chemistry](#), 10.1002/jcc.26450 (2020).
- ¹⁷⁹M. M. Reif and C. Oostenbrink, 'Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation', [Journal of Computational Chemistry](#) **35**, 227–243 (2014).
- ¹⁸⁰G. J. Rocklin, D. L. Mobley, K. A. Dill and P. H. Hünenberger, 'Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects', [The Journal of Chemical Physics](#) **139**, 184103 (2013).
- ¹⁸¹A. de Ruiter and C. Oostenbrink, 'Advances in the calculation of binding free energies', [Current Opinion in Structural Biology](#) **61**, Theory and Simulation • Macromolecular Assemblies, 207–212 (2020).
- ¹⁸²M. Lingenheil, R. Denschlag, R. Reichold and P. Tavan, 'The "hot-solvent/cold-solute" problem revisited', [Journal of Chemical Theory and Computation](#) **4**, 1293–1306 (2008).

- ¹⁸³E. Braun, S. M. Moosavi and B. Smit, ‘Anomalous effects of velocity rescaling algorithms: The flying ice cube effect revisited’, *Journal of Chemical Theory and Computation* **14**, 5262–5272 (2018).
- ¹⁸⁴P. Langevin, ‘Sur la théorie du mouvement brownien’, *Compt. Rendus* **146**, 530–533 (1908).
- ¹⁸⁵D. S. Lemons and A. Gythiel, ‘Paul Langevin’s 1908 paper “On the Theory of Brownian Motion” [“Sur la théorie du mouvement brownien,” C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]’, *American Journal of Physics* **65**, 1079–1081 (1997).
- ¹⁸⁶G. Bussi, D. Donadio and M. Parrinello, ‘Canonical sampling through velocity rescaling’, *Journal of Chemical Physics* **126**, 10.1063/1.2408420 (2007).
- ¹⁸⁷G. Bussi and M. Parrinello, ‘Stochastic thermostats: Comparison of local and global schemes’, *Computer Physics Communications* **179**, 26–29 (2008).
- ¹⁸⁸G. Bussi, T. Zykova-Timan and M. Parrinello, ‘Isothermal-isobaric molecular dynamics using stochastic velocity rescaling’, *The Journal of Chemical Physics* **130**, 074101 (2009).
- ¹⁸⁹J. E. Basconi and M. R. Shirts, ‘Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations’, *Journal of Chemical Theory and Computation* **9**, 2887–2899 (2013).
- ¹⁹⁰S. Nosé, ‘A molecular dynamics method for simulations in the canonical ensemble’, *Molecular Physics* **52**, 255–268 (1984).
- ¹⁹¹W. G. Hoover, ‘Canonical dynamics: Equilibrium phase-space distributions’, *Physical Review A* **31**, 1695–1697 (1985).
- ¹⁹²G. J. Martyna, M. L. Klein and M. Tuckerman, ‘Nosé–hoover chains: The canonical ensemble via continuous dynamics’, *The Journal of Chemical Physics* **97**, 2635–2643 (1992).
- ¹⁹³B. Cooke and S. C. Schmidler, ‘Preserving the boltzmann ensemble in replica-exchange molecular dynamics’, *The Journal of Chemical Physics* **129**, 164112 (2008).
- ¹⁹⁴H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, ‘Molecular dynamics with coupling to an external bath’, *The Journal of Chemical Physics* **81**, 3684–3690 (1984).
- ¹⁹⁵E. Rosta, N.-V. Buchete and G. Hummer, ‘Thermostat artifacts in replica exchange molecular dynamics simulations’, *Journal of Chemical Theory and Computation* **5**, 1393–1399 (2009).
- ¹⁹⁶H. C. Andersen, ‘Molecular dynamics simulations at constant pressure and/or temperature’, *The Journal of Chemical Physics* **72**, 2384–2393 (1980).

- ¹⁹⁷C. A. Fuzo and L. Degève, 'Effect of the thermostat in the molecular dynamics simulation on the folding of the model protein chignolin', *Journal of Molecular Modeling* **18**, 2785–2794 (2012).
- ¹⁹⁸F. Jin, T. Neuhaus, K. Michielsen, S. Miyashita, M. A. Novotny, M. I. Katsnelson and H. De Raedt, 'Equilibration and thermalization of classical systems', *New Journal of Physics* **15**, 033009 (2013).
- ¹⁹⁹K.-H. Chow and D. M. Ferguson, 'Isothermal-isobaric molecular dynamics simulations with Monte Carlo volume sampling', *Computer Physics Communications* **91**, 283–289 (1995).
- ²⁰⁰J. Åqvist, P. Wennerström, M. Nervall, S. Bjelic and B. O. Brandsdal, 'Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm', *Chemical Physics Letters* **384**, 288–294 (2004).
- ²⁰¹S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, 'Constant pressure molecular dynamics simulation: The Langevin piston method', *The Journal of Chemical Physics* **103**, 4613–4621 (1995).
- ²⁰²M. Parrinello and A. Rahman, 'Polymorphic transitions in single crystals: A new molecular dynamics method', *Journal of Applied Physics* **52**, 7182–7190 (1981).
- ²⁰³M. Tuckerman, B. J. Berne and G. J. Martyna, 'Reversible multiple time scale molecular dynamics', *The Journal of Chemical Physics* **97**, 1990–2001 (1992).
- ²⁰⁴G. J. Martyna, M. E. Tuckerman, D. J. Tobias and M. L. Klein, 'Explicit reversible integrators for extended systems dynamics', *Molecular Physics* **87**, 1117–1157 (1996).
- ²⁰⁵M. R. Shirts, 'Simple quantitative tests to validate sampling from thermodynamic ensembles', *Journal of Chemical Theory and Computation* **9**, 909–926 (2013).
- ²⁰⁶S. Rogge, L. Vanduyfhuys, A. Ghysels, M. Waroquier, T. Verstraelen, G. Maurin and V. Van Speybroeck, 'A comparison of barostats for the mechanical characterization of metal–organic frameworks', *Journal of Chemical Theory and Computation* **11**, 5583–5597 (2015).
- ²⁰⁷M. Bernetti and G. Bussi, *Pressure control using stochastic cell rescaling*, arXiv. Preprint. <https://arxiv.org/abs/2006.09250>, 2020.
- ²⁰⁸Y. I. Yang, Q. Shao, J. Zhang, L. Yang and Y. Q. Gao, 'Enhanced sampling in molecular dynamics', *The Journal of Chemical Physics* **151**, 070902 (2019).
- ²⁰⁹S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, 'Optimization by simulated annealing', *Science* **220**, 671–680 (1983).

- ²¹⁰T. Mori and Y. Okamoto, 'Folding simulations of gramicidin A into the β -helix conformations: Simulated annealing molecular dynamics study', *The Journal of Chemical Physics* **131**, 165103 (2009).
- ²¹¹K. Hukushima and K. Nemoto, 'Exchange Monte Carlo method and application to spin glass simulations', *Journal of the Physical Society of Japan* **65**, 1604–1608 (1996).
- ²¹²Y. Sugita and Y. Okamoto, 'Replica-exchange molecular dynamics method for protein folding', *Chemical Physics Letters* **314**, 141–151 (1999).
- ²¹³J. D. Chodera and M. R. Shirts, 'Replica exchange and expanded ensemble simulations as Gibbs sampling: Simple improvements for enhanced mixing', *The Journal of Chemical Physics* **135**, 194110 (2011).
- ²¹⁴H. Nymeyer, 'How efficient is replica exchange molecular dynamics? An analytic approach', *Journal of Chemical Theory and Computation* **4**, 626–636 (2008).
- ²¹⁵D. J. Sindhikara, D. J. Emerson and A. E. Roitberg, 'Exchange often and properly in replica exchange molecular dynamics', *Journal of Chemical Theory and Computation* **6**, 2804–2808 (2010).
- ²¹⁶D. Sindhikara, Y. Meng and A. E. Roitberg, 'Exchange frequency in replica exchange molecular dynamics', *The Journal of Chemical Physics* **128**, 024103 (2008).
- ²¹⁷R. Iwai, K. Kasahara and T. Takahashi, 'Influence of various parameters in the replica-exchange molecular dynamics method: Number of replicas, replica-exchange frequency, and thermostat coupling time constant', *Biophysics and Physicobiology* **15**, 165–172 (2018).
- ²¹⁸M. J. Abraham and J. E. Gready, 'Ensuring mixing efficiency of replica-exchange molecular dynamics simulations', *Journal of Chemical Theory and Computation* **4**, 1119–1128 (2008).
- ²¹⁹D. Gront and A. Kolinski, 'Efficient scheme for optimization of parallel tempering Monte Carlo method', *Journal of Physics: Condensed Matter* **19**, 036225 (2007).
- ²²⁰D. Gront, 'Optimal temperature set for replica exchange sampling', *Biophysical Journal* **112**, 46a (2017).
- ²²¹A. Patriksson and D. van der Spoel, 'A temperature predictor for parallel tempering simulations', *Physical Chemistry Chemical Physics* **10**, 2073 (2008).
- ²²²R. Qi, G. Wei, B. Ma and R. Nussinov, 'Replica Exchange Molecular Dynamics: a practical application protocol with solutions to common problems and a peptide aggregation and self-assembly example', in *Methods in molecular biology* (Springer New York, 2018), pp. 101–119.

- ²²³P. Liu, B. Kim, R. A. Friesner and B. J. Berne, 'Replica exchange with solute tempering: A method for sampling biological systems in explicit water', [Proceedings of the National Academy of Sciences](#) **102**, 13749–13754 (2005).
- ²²⁴L. Wang, R. A. Friesner and B. J. Berne, 'Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2)', [The Journal of Physical Chemistry B](#) **115**, 9431–9438 (2011).
- ²²⁵A. Gil-Ley and G. Bussi, 'Enhanced conformational sampling using replica exchange with collective-variable tempering', [Journal of Chemical Theory and Computation](#) **11**, 1077–1085 (2015).
- ²²⁶P. Shaffer, O. Valsson and M. Parrinello, 'Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin.', [Proceedings of the National Academy of Sciences of the United States of America](#) **113**, 1150–5 (2016).
- ²²⁷G. Torrie and J. Valleau, 'Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling', [Journal of Computational Physics](#) **23**, 187–199 (1977).
- ²²⁸T. Huber, A. E. Torda and W. F. van Gunsteren, 'Local elevation: a method for improving the searching properties of molecular dynamics simulation.', [Journal of Computer-Aided Molecular Design](#) **8**, 695–708 (1994).
- ²²⁹H. Grubmüller, 'Predicting slow structural transitions in macromolecular systems: Conformational flooding', [Physical Review E](#) **52**, 2893–2906 (1995).
- ²³⁰E. Darve and A. Pohorille, 'Calculating free energies using average force', [The Journal of Chemical Physics](#) **115**, 9169–9183 (2001).
- ²³¹A. Laio and M. Parrinello, 'Escaping free-energy minima', [Proceedings of the National Academy of Sciences](#) **99**, 12562–12566 (2002).
- ²³²A. Barducci, G. Bussi and M. Parrinello, 'Well-tempered metadynamics: A smoothly converging and tunable free-energy method', [Physical Review Letters](#) **100**, 10.1103/physrevlett.100.020603 (2008).
- ²³³V. Lindahl, J. Lidmar and B. Hess, 'Accelerated weight histogram method for exploring free energy landscapes', [The Journal of Chemical Physics](#) **141**, 044110 (2014).
- ²³⁴O. Valsson and M. Parrinello, 'Variational approach to enhanced sampling and free energy calculations', [Physical Review Letters](#) **113**, 090601 (2014).

- ²³⁵G. Bussi, F. L. Gervasio, A. Laio and M. Parrinello, 'Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics', *Journal of the American Chemical Society* **128**, 13435–13441 (2006).
- ²³⁶M. M. Sultan and V. S. Pande, 'tICA-Metadynamics: Accelerating metadynamics by using kinetically selected collective variables', *Journal of Chemical Theory and Computation* **13**, 2440–2447 (2017).
- ²³⁷Z. F. Brotzakis and M. Parrinello, 'Enhanced sampling of protein conformational transitions via dynamically optimized collective variables', *Journal of Chemical Theory and Computation* **15**, 1393–1398 (2018).
- ²³⁸L. Sutto, S. Marsili and F. L. Gervasio, 'New advances in metadynamics', *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**, 771–779 (2018).
- ²³⁹Y.-Y. Zhang, H. Niu, G. Piccini, D. Mendels and M. Parrinello, 'Improving collective variables: The case of crystallization', *The Journal of Chemical Physics* **150**, 094509 (2019).
- ²⁴⁰J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille and C. Chipot, 'The adaptive biasing force method: Everything you always wanted to know but were afraid to ask', *The Journal of Physical Chemistry B* **119**, 1129–1151 (2014).
- ²⁴¹A. Dickson and C. L. Brooks, 'Wexplore: Hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm', *The Journal of Physical Chemistry B* **118**, 3532–3542 (2014).
- ²⁴²B. W. Zhang, D. Jasnow and D. M. Zuckerman, 'The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures', *The Journal of Chemical Physics* **132**, 054107 (2010).
- ²⁴³G. Huber and S. Kim, 'Weighted-ensemble Brownian dynamics simulations for protein association reactions', *Biophysical Journal* **70**, 97–110 (1996).
- ²⁴⁴D. Bhatt, B. W. Zhang and D. M. Zuckerman, 'Steady-state simulations using weighted ensemble path sampling', *The Journal of Chemical Physics* **133**, 10.1063/1.3456985 (2010).
- ²⁴⁵V. S. Pande, K. Beauchamp and G. R. Bowman, 'Everything you wanted to know about Markov State Models but were afraid to ask', *Methods* **52**, 99–105 (2010).
- ²⁴⁶D. M. Zuckerman and L. T. Chong, 'Weighted ensemble simulation: Review of methodology, applications, and software', *Annual Review of Biophysics* **46**, 43–57 (2017).
- ²⁴⁷M. Fernández-Suárez and A. Y. Ting, 'Fluorescent probes for super-resolution imaging in living cells', *Nature Reviews Molecular Cell Biology* **9**, 929–943 (2008).

- ²⁴⁸G. G. Yang, F. F. Pan, C. C. N. Parkhurst, J. J. Grutzendler and W. W.-B. Gan, 'Thinned-skull cranial window technique for long-term imaging of the cortex in live mice.', [Nature protocols](#) **5**, 201–8 (2010).
- ²⁴⁹C. E. Tinberg and S. D. Khare, 'Improving Binding Affinity and Selectivity of Computationally Designed Ligand-Binding Proteins Using Experiments', in [Methods in Molecular Biology](#) (Springer New York, 2016), pp. 155–171.
- ²⁵⁰C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson, B. L. Stoddard and D. Baker, 'Computational design of ligand-binding proteins with high affinity and selectivity.', [Nature](#) **501**, 212–6 (2013).
- ²⁵¹M. Goldsmith and D. S. Tawfik, 'Directed enzyme evolution: beyond the low-hanging fruit', [Current Opinion in Structural Biology](#) **22**, 406–412 (2012).
- ²⁵²L. M. Costantini, M. Baloban, M. L. Markwardt, M. Rizzo, F. Guo, V. V. Verkhusha and E. L. Snapp, 'A palette of fluorescent proteins optimized for diverse cellular environments', [Nature Communications](#) **6**, 10.1038/ncomms8670 (2015).
- ²⁵³R. N. Day and M. W. Davidson, 'The fluorescent protein palette: tools for cellular imaging', [Chemical Society Reviews](#) **38**, 2887 (2009).
- ²⁵⁴S. Okada, K. Ota and T. Ito, 'Circular permutation of ligand-binding module improves dynamic range of genetically encoded FRET-based nanosensor', [Protein Science](#) **18**, 2518–2527 (2009).
- ²⁵⁵A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik and S. J. Fleishman, 'Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability', [Molecular Cell](#) **63**, 337–346 (2016).
- ²⁵⁶K. Tyagarajan, E. Pretzer and J. E. Wiktorowicz, 'Thiol-reactive dyes for fluorescence labeling of proteomic samples', [ELECTROPHORESIS](#) **24**, 2348–2358 (2003).
- ²⁵⁷R. J. Pounder, M. J. Stanford, P. Brooks, S. P. Richards and A. P. Dove, 'Metal free thiol–maleimide 'Click' reaction as a mild functionalisation strategy for degradable polymers', [Chemical Communications](#), 5158 (2008).
- ²⁵⁸A. Keppler, S. Gendreizig, T. Gronemeyer, H. Pick, H. Vogel and K. Johnsson, 'A general method for the covalent labeling of fusion proteins with small molecules in vivo', [Nature Biotechnology](#) **21**, 86–89 (2002).
- ²⁵⁹A. Gautier, A. Juillerat, C. Heinis, I. R. Corrêa, M. Kindermann, F. Beaufils and K. Johnsson, 'An engineered protein tag for multiprotein labeling in living cells', [Chemistry & Biology](#) **15**, 128–136 (2008).

- ²⁶⁰S. Watanabe, S. Mizukami, Y. Hori and K. Kikuchi, 'Multicolor protein labeling in living cells using mutant β -Lactamase-Tag technology', *Bioconjugate Chemistry* **21**, 2320–2326 (2010).
- ²⁶¹D. Summerer, S. Chen, N. Wu, A. Deiters, J. W. Chin and P. G. Schultz, 'A genetically encoded fluorescent amino acid', *Proceedings of the National Academy of Sciences* **103**, 9785–9789 (2006).
- ²⁶²A. J. de Graaf, M. Kooijman, W. E. Hennink and E. Mastrobattista, 'Nonnatural Amino Acids for Site-Specific Protein Conjugation', *Bioconjugate Chemistry* **20**, 1281–1295 (2009).
- ²⁶³K. K. J. Lee, D. D. Kang and H. H.-S. Park, 'Site-specific labeling of proteins using unnatural amino acids.', *Molecules and cells* **42**, 386–396 (2019).
- ²⁶⁴M. A. Brun, K.-T. Tan, E. Nakata, M. J. Hinner and K. Johnsson, 'Semisynthetic fluorescent sensor proteins based on self-labeling protein tags', *Journal of the American Chemical Society* **131**, 5873–5884 (2009).
- ²⁶⁵M. A. Brun, R. Griss, L. Reymond, K.-T. Tan, J. Piguet, R. J. Peters, H. Vogel and K. Johnsson, 'Semisynthesis of Fluorescent Metabolite Sensors on Cell Surfaces', *Journal of the American Chemical Society* **133**, 16235–16242 (2011).
- ²⁶⁶M. A. Brun, K.-T. Tan, R. Griss, A. Kielkowska, L. Reymond and K. Johnsson, 'A Semisynthetic Fluorescent Sensor Protein for Glutamate', *Journal of the American Chemical Society* **134**, 7676–7678 (2012).
- ²⁶⁷A. Masharina, L. Reymond, D. Maurel, K. Umezawa and K. Johnsson, 'A fluorescent sensor for GABA and synthetic GABA(B) receptor ligands.', *Journal of the American Chemical Society* **134**, 19026–34 (2012).
- ²⁶⁸O. Sallin, L. Reymond, C. Gondrand, F. Raith, B. Koch and K. Johnsson, 'Semisynthetic biosensors for mapping cellular concentrations of nicotinamide adenine dinucleotides', *eLife* **7**, 10.7554/elife.32638 (2018).
- ²⁶⁹J. A. Mitchell, 'Design of GABA and Glycine FRET Biosensors by Paired Use of Synthetic and Genetically Encoded Fluorophores', Honours thesis (The Australian National University, 2015).
- ²⁷⁰M. A. Rizzo, G. H. Springer, B. Granada and D. W. Piston, 'An improved cyan fluorescent protein variant useful for FRET', *Nature Biotechnology* **22**, 445–449 (2004).
- ²⁷¹K. Deuschle, S. Okumoto, M. Fehr, L. L. Looger, L. Kozhukh and W. B. Frommer, 'Construction and optimization of a family of genetically encoded metabolite sensors by semirational protein engineering', *Protein Science* **14**, 2304–2314 (2005).

- ²⁷²R. M. Clegg, 'Förster resonance energy transfer—FRET what is it, why do it, and how it's done', in *FRET and Flim Techniques*, Vol. 33, Series Title: Laboratory Techniques in Biochemistry and Molecular Biology (Elsevier, 2009), pp. 1–57.
- ²⁷³E. Sisamakis, A. Valeri, S. Kalinin, P. J. Rothwell and C. A. Seidel, 'Accurate single-molecule FRET studies using multiparameter fluorescence detection', in *Methods in enzymology* (Elsevier, 2010), pp. 455–514.
- ²⁷⁴R. S. Knox and H. van Amerongen, 'Refractive index dependence of the Förster resonance excitation transfer rate', *The Journal of Physical Chemistry B* **106**, 5289–5293 (2002).
- ²⁷⁵D. W. Piston and G.-J. Kremers, 'Fluorescent protein FRET: the good, the bad and the ugly', *Trends in Biochemical Sciences* **32**, 407–414 (2007).
- ²⁷⁶B. van der Meer, 'Kappa-squared: from nuisance to new sense', *Reviews in Molecular Biotechnology* **82**, 181–196 (2002).
- ²⁷⁷R. E. Dale and J. Eisinger, 'Intramolecular energy transfer and molecular conformation.', *Proceedings of the National Academy of Sciences* **73**, 271–273 (1976).
- ²⁷⁸R. Y. Tsien, 'The green fluorescent protein', *Annual Review of Biochemistry* **67**, 509–544 (1998).
- ²⁷⁹T. D. Craggs, 'Green fluorescent protein: structure, folding and chromophore maturation.', *Chemical Society reviews* **38**, 2865–75 (2009).
- ²⁸⁰X. Chen, J. L. Zaro and W.-C. Shen, 'Fusion protein linkers: Property, design and functionality', *Advanced Drug Delivery Reviews* **65**, 1357–1369 (2013).
- ²⁸¹M. van Rosmalen, M. Krom and M. Merks, 'Tuning the flexibility of glycine-serine linkers to allow rational design of multidomain proteins', *Biochemistry* **56**, 6565–6574 (2017).
- ²⁸²G. Li, Z. Huang, C. Zhang, B. J. Dong, R. H. Guo, H. W. Yue, L. T. Yan and X. H. Xing, 'Construction of a linker library with widely controllable flexibility for fusion protein design', *Applied Microbiology and Biotechnology* **100**, 215–225 (2016).
- ²⁸³R. Arai, H. Ueda, A. Kitayama, N. Kamiya and T. Nagamune, 'Design of the linkers which effectively separate domains of a bifunctional fusion protein', *Protein Engineering, Design and Selection* **14**, 529–532 (2001).
- ²⁸⁴B. Liu, C. Åberg, F. J. van Eerden, S. J. Marrink, B. Poolman and A. J. Boersma, 'Design and properties of genetically encoded probes for sensing macromolecular crowding', *Biophysical Journal* **112**, 1929–1939 (2017).

- ²⁸⁵E. M. W. M. van Dongen, T. H. Evers, L. M. Dekkers, E. W. Meijer, L. W. J. Klomp and M. Merkx, 'Variation of linker length in ratiometric fluorescent sensor proteins allows rational tuning of Zn(II) affinity in the picomolar to femtomolar range', *Journal of the American Chemical Society* **129**, 3494–3495 (2007).
- ²⁸⁶A. Jung, J. E. Garcia, E. Kim, B.-J. Yoon and B. J. Baker, 'Linker length and fusion site composition improve the optical signal of genetically encoded fluorescent voltage sensors', *Neurophotonics* **2**, 021012 (2015).
- ²⁸⁷J. Zhang, J. Yun, Z. Shang, X. Zhang and B. Pan, 'Design and optimization of a linker for fusion protein construction', *Progress in Natural Science* **19**, 1197–1200 (2009).
- ²⁸⁸N. Amet, H.-F. Lee and W.-C. Shen, 'Insertion of the designed helical linker led to increased expression of Tf-based fusion proteins', *Pharmaceutical Research* **26**, 523 (2008).
- ²⁸⁹Z. Huang, G. Li, C. Zhang and X.-H. Xing, 'A study on the effects of linker flexibility on acid phosphatase PhoC-GFP fusion protein using a novel linker library', *Enzyme and Microbial Technology* **83**, 1–6 (2016).
- ²⁹⁰S. Shamriz, H. Ofoghi and N. Moazami, 'Effect of linker length and residues on the structure and stability of a fusion protein with malaria vaccine application', *Computers in Biology and Medicine* **76**, 24–29 (2016).
- ²⁹¹V. P. R. Chichili, V. Kumar and J. Sivaraman, 'Linkers in the structural biology of protein–protein interactions', *Protein Science* **22**, 153–167.
- ²⁹²R. a George and J. Heringa, 'An analysis of protein domain linkers: their classification and role in protein folding', *Protein Engineering Design and Selection* **15**, 871–879 (2002).
- ²⁹³S. Banjade, Q. Wu, A. Mittal, W. B. Peeples, R. V. Pappu and M. K. Rosen, 'Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck', *Proceedings of the National Academy of Sciences* **112**, E6426–35 (2015).
- ²⁹⁴E. Persson, J. J. Madsen and O. H. Olsen, 'The length of the linker between the epidermal growth factor-like domains in factor VIIa is critical for a productive interaction with tissue factor', *Protein Science* **23**, 1717–1727 (2014).
- ²⁹⁵E. Papaleo, G. Saladino, M. Lambrugh, K. Lindorff-Larsen, F. L. Gervasio and R. Nussinov, 'The role of protein loops and linkers in conformational dynamics and allostery', *Chemical Reviews* **116**, 6391–6423 (2016).
- ²⁹⁶J. S. Klein, S. Jiang, R. P. Galimidi, J. R. Keeffe and P. J. Bjorkman, 'Design and characterization of structured protein linkers with differing flexibilities', *Protein Engineering Design and Selection* **27**, 325–330 (2014).

- ²⁹⁷E. Hempelmann and K. Krafts, 'Bad air, amulets and mosquitoes: 2,000 years of changing perspectives on malaria', *Malaria Journal* **12**, 10.1186/1475-2875-12-232 (2013).
- ²⁹⁸R. Sallares, *Malaria and Rome : a history of malaria in ancient Italy* (Oxford University Press, Oxford New York, 2002).
- ²⁹⁹C. W. Wright, P. A. Linley, R. Brun, S. Wittlin and E. Hsu, 'Ancient Chinese methods are remarkably effective for the preparation of artemisinin-rich extracts of qing hao with potent antimalarial activity', *Molecules* **15**, 804–812 (2010).
- ³⁰⁰A. A. G. Nerlich, B. B. Schraut, S. S. Dittrich, T. T. Jelinek and A. A. R. Zink, 'Plasmodium falciparum in ancient Egypt.', *Emerging infectious diseases* **14**, 1317–9 (2008).
- ³⁰¹K. Harper and G. Armelagos, 'The changing disease-scape in the third epidemiological transition', *International Journal of Environmental Research and Public Health* **7**, 675–697 (2010).
- ³⁰²L. L. Luzzatto, 'Sickle cell anaemia and malaria.', *Mediterranean journal of hematology and infectious diseases* **4**, e2012065 (2012).
- ³⁰³E. C. Mbanefo, A. M. Ahmed, A. Titouna, A. Elmaraezy, N. T. H. Trang, N. Phuoc Long, N. Hoang Anh, T. Diem Nghi, B. The Hung, M. Van Hieu, N. Ky Anh, N. T. Huy and K. Hirayama, 'Association of glucose-6-phosphate dehydrogenase deficiency and malaria: a systematic review and meta-analysis', *Scientific Reports* **7**, 10.1038/srep45963 (2017).
- ³⁰⁴M. O. Altman and P. Gagneux, 'Absence of Neu5Gc and presence of anti-Neu5Gc antibodies in humans — an evolutionary perspective', *Frontiers in Immunology* **10**, 10.3389/fimmu.2019.00789 (2019).
- ³⁰⁵*World malaria report 2019*, Licence: CC BY-NC-SA 3.0 IGO (World Health Organisation, Geneva, 2019).
- ³⁰⁶C. Keating, 'The history of the RTS,S/AS01 malaria vaccine trial', *The Lancet* **395**, 1336–1337 (2020).
- ³⁰⁷B. Greenwood and G. Targett, 'The mysteries of immunity to malaria', *The Lancet* **377**, 1729–1730 (2011).
- ³⁰⁸J. B. Ancsin and R. Kisilevsky, 'A binding site for highly sulfated heparan sulfate is identified in the N terminus of the circumsporozoite protein', *Journal of Biological Chemistry* **279**, 21824–21832 (2004).
- ³⁰⁹D. Oyen, J. L. Torres, C. A. Cottrell, C. Richter King, I. A. Wilson and A. B. Ward, 'Cryo-EM structure of *P. falciparum* circumsporozoite protein with a vaccine-elicited antibody is stabilized by somatically mutated inter-Fab contacts', *Science Advances* **4**, eaau8529 (2018).

- ³¹⁰S. Casares, T.-D. Brumeanu and T. L. Richie, 'The RTS,S malaria vaccine', *Vaccine* **28**, 4880–4894 (2010).
- ³¹¹K. Yu, C. Liu, B.-G. Kim and D.-Y. Lee, 'Synthetic fusion protein design and applications', *Biotechnology Advances* **33**, 155–164 (2015).
- ³¹²A. May, R. Pool, E. van Dijk, J. Bijlard, S. Abeln, J. Heringa and K. A. Feenstra, 'Coarse-grained versus atomistic simulations: Realistic interaction free energies for real proteins', *Bioinformatics* **30**, 326–334 (2013).
- ³¹³R. P. Magalhães, H. S. Fernandes and S. F. Sousa, 'Modelling enzymatic mechanisms with qm/mm approaches: current status and future challenges', *Israel Journal of Chemistry* **60**, 655–666 (2020).
- ³¹⁴V. Vennelakanti, A. Nazemi, R. Mehmood, A. H. Steeves and H. J. Kulik, 'Harder, better, faster, stronger: large-scale qm and qm/mm for predictive modeling in enzymes and proteins', *Current Opinion in Structural Biology* **72**, 9–17 (2022).
- ³¹⁵P. Kar and M. Feig, 'Hybrid all-atom/coarse-grained simulations of proteins by direct coupling of charmm and primo force fields', *Journal of Chemical Theory and Computation* **13**, PMID: 28992696, 5753–5765 (2017).
- ³¹⁶G. Bussi, 'Hamiltonian replica exchange in GROMACS: a flexible implementation', *Molecular Physics* **112**, 379–384 (2013).
- ³¹⁷M. Z. Tien, D. K. Sydykova, A. G. Meyer and C. O. Wilke, 'Peptidebuilder: a simple Python library to generate model peptides', *PeerJ* **1**, e80 (2013).
- ³¹⁸M. M. Lelimousin, M. M. Noirclerc-Savoye, C. C. Lazareno-Saez, B. B. Paetzold, S. S. Le Vot, R. R. Chazal, P. P. Macheboeuf, M. M. J. Field, D. D. Bourgeois and A. A. Royant, 'Intrinsic dynamics in ECFP and Cerulean control fluorescence quantum yield.', *Biochemistry* **48**, 10038–46 (2009).
- ³¹⁹M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek and G. R. Hutchison, 'Avogadro: an advanced semantic chemical editor, visualization, and analysis platform', *Journal of Cheminformatics* **4**, 10.1186/1758-2946-4-17 (2012).
- ³²⁰J. Wang, W. Wang, P. A. Kollman and D. A. Case, 'Automatic atom type and bond type perception in molecular mechanical calculations', *Journal of Molecular Graphics and Modelling* **25**, 247–260 (2006).

- ³²¹D. Case, R. Betz, D. Cerutti, T. Cheatham III, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, D. Roe, A. Roitberg, C. Sagui, C. Simmeling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao and P. Kollman, *Amber 2016*, University of California and San Francisco, 2016.
- ³²²T. Bereau and K. Kremer, 'Automated parametrization of the coarse-grained martini force field for small organic molecules', *Journal of Chemical Theory and Computation* **11**, 2783–2791 (2015).
- ³²³J. A. Graham, J. W. Essex and S. Khalid, 'PyCGTOOL: automated generation of coarse-grained molecular dynamics models from atomistic trajectories', *Journal of Chemical Information and Modeling* **57**, 650–656 (2017).
- ³²⁴S. Planamente, A. Vigouroux, S. Mondy, M. Nicaise, D. Faure and S. Moréra, 'A conserved mechanism of GABA binding and antagonism is revealed by structure-function analysis of the periplasmic binding protein Atu2422 in *Agrobacterium tumefaciens*', *Journal of Biological Chemistry* **285**, 30294–30303 (2010).
- ³²⁵M. R. Shirts and J. D. Chodera, 'Statistically optimal analysis of samples from multiple equilibrium states', *The Journal of Chemical Physics* **129**, 124105 (2008).
- ³²⁶V. Lindahl, A. Villa and B. Hess, 'Sequence dependency of canonical base pair opening in the DNA double helix', *PLOS Computational Biology* **13**, edited by A. MacKerell, e1005463 (2017).
- ³²⁷W. Kabsch and C. Sander, 'Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers* **22**, 2577–2637 (1983).
- ³²⁸H.-X. Zhou, G. Rivas and A. P. Minton, 'Macromolecular crowding and confinement: Biochemical, biophysical, and potential physiological consequences', *Annual Review of Biophysics* **37**, 375–397 (2008).
- ³²⁹N. Ostrowska, M. Feig and J. Trylska, 'Modeling crowded environment in molecular simulations', *Frontiers in Molecular Biosciences* **6**, 10.3389/fmolb.2019.00086 (2019).
- ³³⁰L. A. Abriata and M. Dal Peraro, 'Assessing the potential of atomistic molecular dynamics simulations to probe reversible protein-protein recognition and binding', *Scientific Reports* **5**, 10.1038/srep10549 (2015).
- ³³¹G. S. Ayton, W. G. Noid and G. A. Voth, 'Multiscale modeling of biomolecular systems: In serial and in parallel', *Current Opinion in Structural Biology* **17**, 192–198 (2007).

- ³³²A. J. Rzepiela, M. Louhivuori, C. Peter and S. J. Marrink, 'Hybrid simulations: Combining atomistic and coarse-grained force fields using virtual sites', *Physical Chemistry Chemical Physics* **13**, 10437 (2011).
- ³³³Y. Liu, W. Pezeshkian, J. Barnoud, A. H. de Vries and S. J. Marrink, 'Coupling coarse-grained to fine-grained models via Hamiltonian replica exchange', *Journal of Chemical Theory and Computation* **16**, 5313–5322 (2020).
- ³³⁴N. A. Tanner, S. M. Hamdan, S. Jergic, K. V. Loscha, P. M. Schaeffer, N. E. Dixon and A. M. van Oijen, 'Single-molecule studies of fork dynamics in Escherichia coli DNA replication', *Nature Structural & Molecular Biology* **15**, 170–176 (2008).
- ³³⁵M. Merks, M. V. Golynskiy, L. H. Lindenburg and J. L. Vinkenburg, 'Rational design of FRET sensor proteins based on mutually exclusive domain interactions', *Biochemical Society Transactions* **41**, 1201–1205 (2013).
- ³³⁶L. L. H. Lindenburg, M. M. Malisauskas, T. T. Sips, L. L. van Oppen, S. S. P. W. Wijnands, S. S. F. J. van de Graaf and M. M. Merks, 'Quantifying stickiness: Thermodynamic characterization of intramolecular domain interactions to guide the design of Förster resonance energy transfer sensors.', *Biochemistry* **53**, 6370–81 (2014).