



Australian
National
University

Biophysical principles for the design of photoluminescent proteins

Edan Habel

December 2025

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

© Copyright by 2025
All Rights Reserved

Declaration

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. The content of my thesis is the result of work I have carried out since the commencement of my candidature (February 2022 to December 2025) and has not been previously submitted for another degree or diploma in any university or tertiary institution.

Edan Isaac Habel

December 2025

Acknowledgements

A PhD demands self-direction, independence, and ownership of a small part of a much larger corpus of collective knowledge. Nonetheless, this thesis is a product of continuous professional, academic, financial, and social support from a large community of people.

To Thomas Huber, you've nurtured my curiosity at every turn, guiding me to find quantitative answers to qualitative oddities. I am grateful to complete this work under your supervision, projects which come from your encouragement to follow my nose.

To Gottfried Otting, your relentless commitment to the advancement of pure scientific knowledge is infectious. Your input over these years has made this research grounded in meaningful experiments.

To Rebecca Frkic, thank you for teaching me everything I needed to take my proteins from an Eppendorf to the PDB. Your guidance meant I developed skills that I wouldn't have without you.

To Dan Preston, your approach to supramolecular chemistry has framed how I see protein structures. Thank you for your insight and guiding me through the university system.

To the many academics at ANU, thank you for your collaboration, offering advice, sharing your resources, and letting me pick your brains. Namely: Colin Jackson, Lara Malins, Alison Rodger, Nick Cox, John Carver, Nick White, Jamie Hicks, Adam Perriman, Valeska Ting and Elmars Krausz.

Thank you to the collaborators who worked with me on these projects, Haocheng Qianzhu, Elwy Abdelkader, Adarshi Welegedara, and Martyna Judd. Thank you for the education, the samples, the work, and the good company. We learnt some great things together.

Thank you to the people who kept the machines alive: Doug, Eiri, Joe, and Anitha. I wouldn't have data without you. An extra thank you to Doug, who worked through the Keeler textbook with me in my first year.

To my colleagues: Rosie, Nathan, Andrew, Sacha, Barnabas, and Sven. Thank you for the camaraderie these past few years. You're all fantastic scientists, and, I hope, lifelong friends.

To the students I taught in Honours and undergraduate projects: Caitlin, Becca, Tim, Aidan, Luca, and Ryan. Thank you for all your questions, I had to learn a lot to answer them.

To Sue and Sam at Blue Olive: thanks for making sure I had lunch, lending me black tie accoutrements, and giving me a great start to each day.

Thank you to the Australian taxpayer, who supported this work through ARC grants and the RTP scholarship. Publicly funded research is for the benefit of all of us.

Thank you to my parents, who funded my early education. To my mum, Sylvia, thank you for supporting my independence every step of the way. I've never had an interest, no matter how fleeting, that you haven't encouraged me to pursue.

To Nonna Dora and Nonno Luigi, you moved to Australia over 50 years ago and worked hard for your family ever since. Thanks to that, I got to spend years focused on my education, a privilege not afforded to everyone.

To my brother, Jonah and my sister, Mia. Thank you for the long phone calls.

To my aunties: Tania, Josie, Sonia, Lisa; uncles: Piers, Aaron, Chad; and cousins: Jaede, Luca, Charlie, Sascha, Ari, and little Sigrun. You've all shown interest in my work and made me feel missed while I've been at ANU. I am truly spoilt to have every one of you.

To Vanessa and Andrew, thank you for treating me like a member of your family. Thanks to James, Kath and Arun for being a bright addition to my life.

To Laura, you're the best. We've done this together, and now we can do other things. Thank you.

Abstract

Photoluminescence is the emission of light following optical excitation. As biological molecules typically have weak emission, photoluminescent compounds provide excellent analytical utility for interrogating biological systems. While synthetic dyes and labels are one way to provide the analytic signal for such interrogations, photoluminescent proteins offer the advantage of being categorically biocompatible. Photoluminescent proteins are also ideal targets for application-specific design, as their targeting to specific cellular loci, binding affinity for target analytes, or response to different environments may be modified while having minimal or predictable effect on the optical readout. This thesis analyses the biophysical underpinnings of how photoluminescent proteins may be designed to have the desired properties.

Following a general introduction in Chapter 1, Chapter 2 explores the design of phosphorescent proteins, pairing genetically encoded antennae and lanthanide binding sites for lanthanide (III) ions. This work presents a novel, stable, and high-yield lanthanide binding protein which exhibits nanomolar binding affinity. Chapter 3 characterises the tryptophan orientations within this designed protein with experimental electron-nuclear double resonance (ENDOR) spectroscopy. Analysis of ENDOR spectra reveals rotameric conformations that are absent from computationally predicted structures. Chapter 4 undergoes an in-depth analysis of the steric accommodation of 4-substituted tryptophans in maltose binding protein. This work shows the side-chain repacking that takes place to maintain the native global fold. Chapter 5 assesses the steric accommodation of benzadiazole-containing non-canonical amino acids, showing that the bulkier side chain is surprisingly isosteric for the canonical tryptophan. This work also presents experimental evidence for the expansion of the classical fluorescent protein catalysis to include a secondary amine between the aromatic side chain of the chromophore tripeptide and the backbone.

Finally, I conclude this thesis with a summary of results, and areas of focus that would expand this current work.

Abbreviations

4CNW	4-cyanotryptophan
4FW	4-fluorotryptophan
5FW	5-fluorotryptophan
6AW	6-azatryptophan
6FW	6-fluorotryptophan
7FW	7-fluorotryptophan
aaRS	Aminoacyl-tRNA synthetase
AF3	AlphaFold3
ATP	Adenosine triphosphate
BMRB	Biological Magnetic Resonance Bank
BODIPY	Boron-dipyrromethene
BSA	Buried surface area
CIF	Crystallographic information file
DFT	Density functional theory
DTT	Dithiothreitol
<i>E. coli</i>	<i>Escherichia coli</i>
EBI	European Bioinformatics Institute
EDTA	Ethylenediaminetetraacetic acid
EF	EF-hand motif
eLBOW	electronic Ligand Builder and Optimization Workbench
EMBL	European Molecular Biology Laboratory
ENDOR	Electron nuclear double resonance
EPR	Electron paramagnetic resonance
FP	Fluorescent Protein
FRET	Förster resonance energy transfer
GCE	Genetic code expansion
GFP	Green fluorescent protein
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl β -D-1-thiogalactopyranoside
ISC	Intersystem crossing
K_D	Dissociation constant
λ	Lambda (wavelength)

LBT	Lanthanide-binding tag
LRET	Luminescence resonance energy transfer
MBP	Maltose-binding protein
MES	2-(<i>N</i> -morpholino)ethanesulfonic acid
<i>M. extorquens</i>	<i>Methylobacterium extorquens</i>
MOPS	3-(<i>N</i> -morpholino)propanesulfonic acid
MR	Molecular replacement
MWCO	Molecular weight cut-off
NBD	Nitrobenzoxadiazole
NBDO	<i>N</i> -(7-nitro-2,1,3-benzoxadiazol-4-yl)-L-2,3-diaminopropionic acid
NBDSe	<i>N</i> -(7-nitro-2,1,3-benzoselenadiazol-4-yl)-L-2,3-diaminopropionic acid
ncAA	non-canonical amino acid
NCS	Non-crystallographic symmetry
NMR	Nuclear magnetic resonance
PBS	Phosphate-buffered saline
PDB	Protein Data Bank
PEG	Polyethylene Glycol
RFdiffusion	Rosetta Fold diffusion
RFP	Red fluorescent protein
RMSD	Root mean square deviation
S_1	Excited singlet state
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
T_1	Excited triplet state
TAG	Amber stop codon
TEV	Tobacco etch virus
tRNA	transfer RNA
UV	Ultraviolet
vdW	van der Waals
WT	Wild type

List of publications

The work in this thesis has resulted in the following publications/manuscripts which contain work from chapters in this text:

Habel, E., Qianzhu, H., Abdelkader, E. H., Otting, G., Huber, T. “Biosynthetic lanthanide-luminescent mini-proteins using genetic code expansion”, *Journal of the American Chemical Society*, **2026**, 148, 14391–14399, <https://doi.org/10.1021/jacs.6c01416>

Habel, E., Judd, M., Huber, T., Cox, N. “Mapping conformational heterogeneity: ^{19}F - Gd^{3+} ENDOR reveals hidden rotamer conformations”, **2025**, *under review*

Qianzhu, H., **Habel, E.**, Butler, S. M., Abdelkader, E. H., Paul, N., Zhou, X. Otting, G. Joliffe, K. A., Huber, T. “Site-specific FLIM probes via genetic encoding of environmentally responsive benzodiazoles”, **2025**, *in preparation*

Qianzhu, H., Abdelkader, E. H., Welegedara, A. P., **Habel, E.**, Paul, N., Frkic, R. L., Jackson, C. J., Huber, T., Otting, G. “Rendering proteins fluorescent inconspicuously: Genetically encoded 4-cyanotryptophan conserves their structure and enables the detection of ligand binding sites”, *Angewandte Chemie International Edition*, **2025**, 65, e202421000, <https://doi.org/10.1002/chem.202502645>

Additional publications/manuscripts that resulted as part of research in parallel to the work in this thesis:

Judd, M., Abdelkader, E. H., Qianzhu, H., **Habel, E.**, Lindsay, C., Brancewicz, A., Huber, T., Otting, G., Cox, N. “Characterising side chain conformations in proteins using ^{19}F ENDOR and new fluorinated amino acid spin labels”, *Chemistry – A European Journal*, **2025**, e02647, <https://doi.org/10.1002/chem.202502647>

Ambros, A. M., Qianzhu, H., Tiemann, M., **Habel, E.**, Denzinger, K., Zupan, H., Accorsi, M., Keller, B. G., Wolber, G., Huber, T., Rademann, J. “Genetic encoding of pentafluorophosphato-phenylalanine provides PF₅-proteins as phosphoprotein mimetics”, *Angewandte Chemie International Edition*, **2025**, e18789, <https://doi.org/10.1002/anie.202518789>

Habel, E., Vishwakarma, P., Huber, T., Rodger, A. “Advances in polarised light approaches to understanding interactions between biomolecules”, *Biochemical Society Transactions*, **2026**, 54, 349–361, <https://doi.org/10.1042/BST20253138>

Gare, C. L., Schwartz, B. D., Frkic, R. L., **Habel, E.**, Hueneburg, T., Prasad, K. M., Tailhades, J., Gardiner, M. G., Cryle, M. J., Huber, T., Jackson, C. J., McGuire, H. M., White, A. M., Malins, L. R. “Single-atom replacement unlocks iso-pseudoprolines as minimalist proline bioisosteres for peptide detection and late-stage functionalisation”, **2026**, *under review*

Table of Contents

Declaration	ii
Acknowledgements	iii
Abstract	v
Abbreviations	vi
List of publications	viii
Table of Contents	x
Chapter 1 : Introduction	1
1.1 Photoluminescence	2
1.2 Fluorescent proteins	4
1.3 Fluorescent labels and non-canonical amino acids	5
1.4 Lanthanide-luminescent proteins	7
1.5 Contemporary protein design	10
1.6 Research outline	13
1.7 References	15
Chapter 2 : Biosynthetic lanthanide-luminescent mini-proteins using genetic code expansion	21
2.1 Introduction	22
2.2 Photoluminescent properties of LBTs with ncAAs	25
2.3 Computational design of novel lanthanide-luminescent mini-proteins	30
2.4 Enhanced brightness and more robust binding affinity with an ncAA antenna	33
2.4.1 New lanthanide-luminescent mini-protein shows remarkable selectivity and in vivo sensitisation	36
2.5 Conclusion	40
2.6 Methods	41
2.6.1 Expression and purification of LBT and EF2 proteins	41
2.6.2 Luminescence measurements	42
2.7 References	43
Chapter 3 : Mapping conformational heterogeneity: ^{19}F-Gd^{3+} ENDOR reveals hidden rotamer conformations.....	45
3.1 Introduction	46
3.2 Fitting of protein or ENDOR frame parametrised models to ^{19}F - Gd^{3+} ENDOR spectra	49
3.3 Uncovering hidden rotamer conformations from simulated ^{19}F - Gd^{3+} distributions using Tikhonov regularisation	54

3.4	Conclusion	59
3.5	Methods	60
3.5.1	Protein expression and purification.....	60
3.5.2	ENDOR sample preparation	61
3.5.3	EPR, ENDOR measurements, and lineshape calculations.....	61
3.6	References	62
Chapter 4 : Structural accommodation of 4-substituted tryptophans in maltose-binding protein		64
4.1	4-substituted tryptophans	65
4.2	Maltose-binding protein.....	67
4.2.1	Generating the MBP rotamer dataset	68
4.3	Homogeneity of global folds	69
4.4	Trp10 substitution	70
4.5	Trp62 substitution	73
4.6	Trp 340 substitution.....	76
4.7	Conclusion	82
4.8	Methods	83
4.8.1	Cloning, expression and purification.....	83
4.8.2	Size-exclusion chromatography.....	84
4.8.3	Crystallisation	84
4.8.4	Structure determination and refinement	84
4.9	Crystallographic statistics	86
4.10	References	89
Chapter 5 : Minimalist fluorescent reporters and novel autocatalytic chromophore chemistry with NBD amino acids.....		93
5.1	Introduction	94
5.1.1	Nitrobenzodiazole fluorophores	94
5.1.2	Fluorescent protein chromophore formation.....	95
5.2	Structural analysis of a nanobody functionalised with NBD-fluorophores ..	97
5.2.1	Tryptophan-NBD pseudoisosterism	97
5.2.2	Crystal structure of ArcNL–H11 W37 NBDO	98
5.2.3	Crystal structure of ArcNL–H11 W37 NBDSe	101
5.3	NowGFP with NBD chromophores	103
5.3.1	Crystal structure of NowGFP NBDO	104
5.4	Crystal structure of NowGFP NBDSe	107
5.5	Restoring NBD chromophore formation in a related scaffold	109

5.5.1	Crystal structures of mCerulean D148G NBDO and mCerulean D148G/V61K NBDO	111
5.5.2	Novel chromophore chemistry with NBD amino acids	113
5.6	Conclusion	115
5.7	Methods	116
5.7.1	Development of translation system to incorporate NBD amino acids ..	116
5.7.2	Cloning, expression and protein purification	116
5.7.3	Size-exclusion chromatography.....	117
5.7.4	Protein crystallisation	117
5.7.5	Structure determination and refinement	118
5.7.6	UV-Vis and fluorescence	119
5.8	Crystallographic statistics	120
5.9	References	123
Chapter 6 : Conclusion and future directions		128
6.1	Conclusion and future directions	129
Appendix A : Thesis supporting information		132
A.1	Supporting information for Chapter 2	134
A.1.1	Intact protein mass spectrometry	134
A.1.2	Lanthanide titration experiments	138
A.1.3	Integrated lifetime measurements	140
A.1.4	Spectral overlap calculations	144
A.1.5	Sequence design of EF2-derived mini-proteins.....	148
A.1.6	NMR measurements of ¹³ C/ ¹⁵ N-labeled RF2	149
A.1.7	Photoluminescence properties of RF2 and RF2 6AW with Tb ³⁺ , Eu ³⁺ , Dy ³⁺ and Sm ³⁺	152
A.1.8	q-value determination.....	155
A.1.9	Supplementary tables	156
A.2	Supporting information for Chapter 3	158
A.2.1	Protein sequence and expression yield	158
A.2.2	Intact protein mass spectrometry	158
A.2.3	Rotamer distributions	160
A.2.4	Model generation and clash score calculation.....	161
A.3	References	164

Chapter 1: Introduction

1.1 Photoluminescence

Photoluminescent reporters enable real-time visualisation of protein localisation, dynamics and interactions.¹ Photoluminescent biotechnology has revolutionised the understanding of cellular processes, from tracking molecular machines and organelle dynamics to mapping complex signalling networks in living systems.^{2,3} Applications span research in cell biology,⁴ neuroscience⁵ and developmental biology⁶ through to clinical diagnostics,⁷ drug discovery⁸ and therapeutic monitoring,⁹ making photoluminescence indispensable across modern life sciences. The ideal probe reports on protein localisation, activity, and local environment while minimising perturbation to its native structure.

Photoluminescence is the radiative decay of a molecule's electronically excited state. After excitation, chromophores relax radiatively (fluorescence ($S_1 \rightarrow S_0$); phosphorescence ($T_1 \rightarrow S_0$)) or non-radiatively via internal conversion, vibrational dissipation, or energy transfer to other compounds (Figure 1.1), all of which are strongly influenced by the chromophore's local environment. Brightness and excitation/emission wavelength are the primary design variables. These can be manipulated by understanding the protein structure surrounding a luminophore. Brightness increases with higher absorption efficiency and reduced non-radiative decay, which competes with emission and shortens lifetimes. In practice, this requires suppressing undesired internal conversion, vibrational coupling to nearby molecules and energy transfer to nearby quenchers. Greater brightness improves signal-to-noise, so less probe is needed. Emission wavelength reflects the transition energy gap and is tuned by modulating the chromophore structure and its environment. Red-shifted wavelengths reduce phototoxicity,^{10,11} autofluorescence¹² and scattering while improving imaging depth.¹³

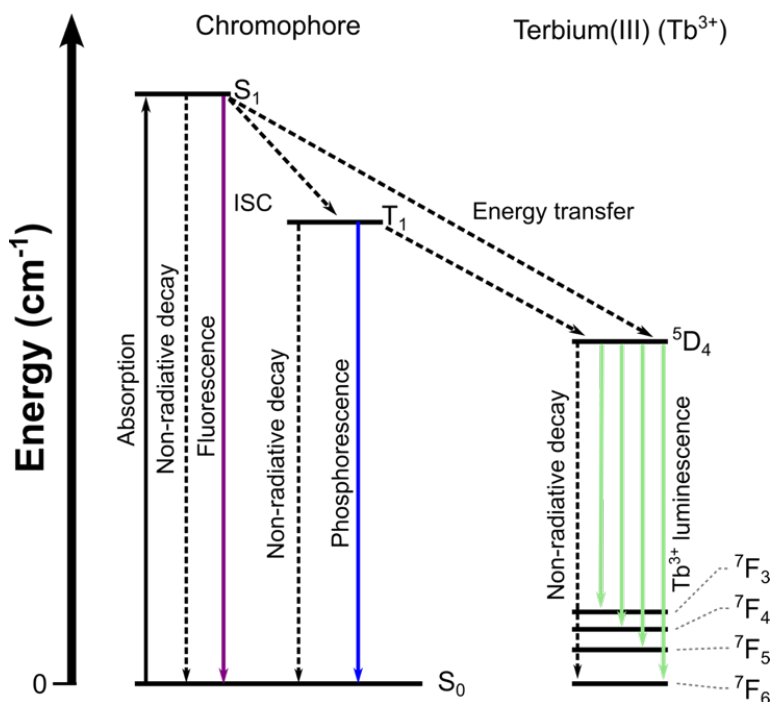


Figure 1.1 – Simplified Jablonski diagram. Photoexcitation populates S₁. From S₁, the chromophore relaxes by fluorescence (S₁→ S₀) or by non-radiative decay, including intersystem crossing (ISC) to T₁, energy can also be transferred from S₁ to the Ln³⁺. From T₁, relaxation proceeds by phosphorescence (T₁→ S₀), non-radiative decay, or triplet-to-Ln³⁺ energy transfer. For Tb³⁺, the commonly observed bands are ⁵D₄→⁷F₆ ≈ 488–495 nm, ⁵D₄→⁷F₅ ≈ 541–547 nm, ⁵D₄→⁷F₄ ≈ 583–587 nm, and ⁵D₄→⁷F₃ ≈ 619–624 nm.

Current photoluminescent protein strategies have limitations. Fusion of the target protein with a fluorescent protein introduces a large (~27 kDa) domain that can perturb protein function and localisation.¹⁴ The conjugation of small-molecule chromophores offers small-molecule labelling but necessitates post-translational covalent modification, limiting their use in live cells and complicating experimental workflows. Genetic code expansion (GCE) enables site-specific incorporation of fluorescent non-canonical amino acid (ncAA) sidechains, but the principles surrounding the optimal position to introduce these moieties remain unclear. Lanthanide-luminescent proteins offer an alternative strategy with long-lived luminescence but require fine-tuning of the lanthanide coordination environment and must compete with common biological chelates for metal binding. Contemporary protein design tools, including AlphaFold and RFdiffusion, provide high-confidence structural predictions but cannot directly predict photoluminescent properties, which depend on ligands, ncAAs, chromophore maturation and conformational dynamics, which are not explicitly encoded in primary

structure. Despite the diversity of existing reporters, it remains challenging to predict how a protein environment will tune brightness, lifetime and spectral properties of a protein with an embedded luminophore. This thesis addresses this gap by combining structural and biophysical measurements to investigate the principles for rationally designing photoluminescent proteins with desired properties.

1.2 Fluorescent proteins

The most widely used fluorescent proteins (FPs) house an endogenous chromophore formed autocatalytically from a tripeptide inside an 11-stranded β -barrel.¹⁵ The chromophore maturation pathway proceeds via cyclisation, dehydration and oxidation, with conserved active-site residues steering the reaction. FP chromophores are genetically encoded, requiring only canonical amino acids. Canonical green-fluorescent protein (GFP) chromophores feature a serine-tyrosine-glycine (Ser-Tyr-Gly) tripeptide.¹⁶ Early diversification shifted colour by editing the triad (e.g., Ser $\rightarrow$ Thr, Tyr $\rightarrow$ Trp), and modifying the residues in the chromophore pocket, yielding the blue, cyan, yellow and red families.^{17–19} These variants use the same fundamental chromophore maturation chemistry with different pocket environments and aromatic amino acid sidechains. Despite the wealth of structures available, it remains difficult to predict *a priori* how specific sidechain substitutions will alter photoluminescent properties.

Spectral range in FPs is controlled by the chromophore structure, the surrounding chemical environment, and chromophore conjugation extension. Chromoproteins from Anthozoan organisms have extended conjugation from additional dehydration of the core tripeptide driven by modifications of the surrounding residues.^{20,21} This red-shifts absorption and increases extinction coefficients, but significantly reduces fluorescence intensity.^{22,23} Small structural changes around the chromophore can switch between emissive and dark states. One such Anthozoan FP tripeptide contains histidine (His) in the canonical Gly position, undergoes photocatalytic dehydrogenation of the alkane to an alkene between the His C $_{\alpha}$ and C $_{\beta}$,²⁴ which allowed the tracking of protein movement

in microscopy, and an improvement in the signal-to-noise.¹⁴ Deviation from standard FP tripeptide chemistry enables the engineering of alternative chromophore applications.

Most FP designs tune the chromophore electronic environment rather than the chromophore chemistry. This optimises brightness, durability and spectral properties. Recent FPs selected from random mutagenesis libraries feature low-energy (>500 nm) excitation and high brightness.^{25,26} Far-red and near infra-red wavelengths enable deep-tissue imaging,²⁷ and bright photoswitchable FPs enable high-resolution microscopy that overcomes the diffraction limit of visible light.^{28,29} These examples show that local FP chromophore environment can be steered towards having desirable properties. FPs have become ubiquitous in biological research since the 1990s. Despite their utility, FPs' β -barrel can perturb protein function and localisation as fusion proteins. This motivates alternate strategies for introducing photoluminescence that decouple the reporter from the native fold.

1.3 Fluorescent labels and non-canonical amino acids

Small-molecule dyes provide compact, bright chromophores with tuneable photophysical properties. However, conventional conjugation chemistry lacks site-specificity. Dyes typically react with nucleophilic sidechains like cysteine or lysine, yielding heterogeneous labelling even when targeting unique chemical environments.^{30–33} Post-translational modification limits live-cell applications. Small-molecule chromophores are also acutely sensitive to their local environment, which enables sensing applications but may cause unpredictable behaviour when the goal is a stable, uniform signal.

Genetic code expansion (GCE) offers a powerful solution to achieve site-specific labelling. This technology uses an orthogonal aminoacyl synthetase (aaRS)/tRNA pair, typically derived from archaeal species,³⁴ that does not cross-react with the host translational machinery.³⁵ The orthogonal aaRS is evolved to charge its cognate tRNA with the desired ncAA, while the tRNA recognises a repurposed codon, most commonly the amber stop (UAG),³⁶ allowing site-specific ncAA incorporation at genetically encoded

positions in the target protein.³⁷ Once established, GCE requires only the co-expression of the orthogonal aaRS/tRNA pair and supplementation of growth media with the ncAA, making it straightforward to synthesise proteins with ncAAs *in vivo*.³⁸

GCE enables two complementary strategies for introducing fluorescent labels into proteins. First, ncAAs bearing bio-orthogonal reactive groups (such as azides, alkynes, or strained alkenes) can serve as handles for conjugation with fluorescent dyes via click chemistry or related reactions.^{33,39,40} This approach combines the site-specificity of GCE with the photophysical properties of optimised synthetic fluorophores, though it still requires post-translational modification steps. Second, ncAAs with intrinsically fluorescent sidechains can be directly incorporated during translation, eliminating the need for any post-translational labelling.⁴¹ The fluorescent ncAAs provide immediate, site-specific fluorescence without additional chemical steps, making them particularly suitable for live-cell applications.

Sites are generally chosen on the basis of solvent accessibility, distance from active sites, and predicted side-chain packing, to minimise structural perturbation while reporting on protein localisation or activity.^{41–45} The local environment of an incorporated fluorescent ncAA is important as it mirrors classical small-molecule chromophores; polarity, rigidity and specific quenchers modulate photophysical properties. Here, the “solvent” includes a heterogeneous protein matrix. Structure prediction software guides site-selection in proteins lacking experimental models, though local structural response to ncAA incorporation must be determined case-by-case. High-resolution structures are available for few fluorescent ncAAs, limiting the ability to generalise design principles.^{46–48} As a result, the consequences of incorporating a given fluorescent ncAA on emission wavelength, lifetime and quantum yield cannot be pre-empted from sequence alone. Fluorescent proteins, small-molecule dyes and fluorescent ncAAs all rely on organic chromophores with nanosecond-timescale emission. Lanthanide-based luminophores offer complementary properties: long emission lifetimes enable time-gated detection that eliminates short-lived autofluorescence.

1.4 Lanthanide-luminescent proteins

Lanthanides are relatively abundant in the Earth's crust, despite being termed rare earth elements. Cerium, for example, is more abundant than copper.⁴⁹ Lanthanide(III) (Ln^{3+}) ions have unique optical properties, arising from their numerous electron orbitals with unpaired electrons, and most Ln^{3+} ions can luminesce under appropriate conditions.^{50,51} Laporte-forbidden $f-f$ transitions yield slow radiative decay of excited lanthanide ions, ranging from a few microseconds to tens of milliseconds. This enables time-gated detection that eliminates short-lived autofluorescence.^{52–54} Unlike organic fluorophores, lanthanide emission spectra are relatively insensitive to environmental conditions, providing stable, sharp and predictable signals which are often in the visible light range. However, Ln^{3+} ions suffer from low extinction coefficients ($1\text{--}10\text{ M}^{-1}\text{cm}^{-1}$) and are efficiently quenched by water molecules. Engineering a lanthanide-luminescent protein requires a nearby chromophore antenna to compensate for low extinction and the exclusion of coordinating water. Efficient sensitisation typically proceeds via population of the antenna S_1 or T_1 states, followed by energy transfer to the Ln^{3+} 4f manifold; optimal antenna performance requires the excited-state energy to lie $\sim 2500\text{--}4000\text{ cm}^{-1}$ above the accepting Ln^{3+} level to minimise back-transfer.^{50,51,55,56} Ln^{3+} 4f orbitals are shielded by filled $5s^25p^6$ subshells, making coordination primarily electrostatic. Ln^{3+} ions prefer hard ligands; together with electrostatic compensation, carboxylic acids are prime candidates for strong coordination. Binding sites for Ln^{3+} ions contain acidic residues such as aspartate (Asp) and glutamate (Glu), fine-tuned with amide-containing sidechains such as asparagine (Asn) and glutamine (Gln).

Naturally evolved chelating proteins and metalloenzymes have been discovered in the past decade, many in *Methylobacterium extorquens* (*M. extorquens*). The organism seems to have evolved to use trace lanthanum(III) found on plant leaves to trigger methanol metabolism, rationalising *M. extorquens* as the origin of many lanthanide-binding protein sequences.⁵⁷ The first to be structurally characterised was XoxF, is a methanol dehydrogenase which coordinates lanthanides via 3 acidic residues, 1 Asn, and a pyrroloquinoline quinone (PQQ) cofactor (Figure 1.2D).⁵⁸ In this protein, the

lanthanide preorganises the ligand binding site, and acts as a hard Lewis acid to activate PQQ for alcohol oxidation. Lanthanides play a similar role in other lanthanide-dependent enzymes.⁵⁹ Landiscernin (LanD) from *M. extorquens*, contains four localised Glu residues on the loop between two stable α -helices. This protein binds Ln^{3+} ions through metal-centred dimerization; mutations (Glu $\rightarrow$ Gln, Glu $\rightarrow$ Ala) improve selectivity for lighter lanthanides 100-fold.⁶⁰ Lanmodulin (LanM), is a picomolar-affinity calmodulin homologue containing multiple Ln^{3+} binding sites. Each motif includes a threonine (Thr) backbone carbonyl coordinated to the lanthanide (Figure 1.2B).^{61,62} Like calmodulin, the tertiary fold of LanM is facilitated by Ln^{3+} , which brings together its termini. A fusion protein, LanTERN brings together enhanced cyan fluorescent protein (eCFP) and Citrine (a yellow fluorescent protein variant) upon lanthanide binding, which could be measured spectroscopically using Förster resonance energy transfer (FRET). This protein acts as a fluorescent sensor for lanthanides with apparent K_D s of 10–50 pM.⁶³ Wang and co-workers reported in 2025 the evolution of LanM with acridone as an antenna, which facilitated the tracking of protease cleavage using time-resolved experiments.⁶⁴ Collectively, these natural systems illustrate that precise Ln^{3+} coordination geometry and binding kinetics can be applied for diverse applications.

Few natural lanthanide-binding proteins have been characterised in detail,^{65,66} motivating the design of novel lanthanide-binding proteins and peptides. Most designed lanthanide-binding proteins derive from EF-hand motifs, which bind Ca^{2+} via multiple negatively charged residues. The EF-hand motif was employed in lanthanide binding tags (LBTs), which are short peptides (15–17 amino acids) optimised for terbium(III) (Tb^{3+}) binding and luminescence.^{67,68} LBTs coordinatively saturate the bound Ln^{3+} ion, and bind with high affinity ($K_D = 10\text{--}150$ nM). These peptides position a tryptophan (Trp) residue such that its backbone carbonyl coordinates Ln^{3+} , holding the indole nearby (Figure 1.2A).⁶⁹ The luminescence of LBT- Tb^{3+} has been applied to measure the binding of Src kinases to phosphopeptides, using luminescence resonance energy transfer (LRET) between the LBT and a BODIPY ligand.⁷⁰ This same LRET pair has also been used to measure the distance between the LBT and BODIPY in a membrane-bound oligosaccharyltransferase, giving useful indications of conformational changes upon glycan binding on structurally

uncharacterised complexes.⁷¹ Imperiali and co-workers later synthesised the LBT with acridone and carbostyryl 124, and showed the complex was also capable of efficiently exciting europium(III) (Eu^{3+}) with these non-canonical antennae.⁷²

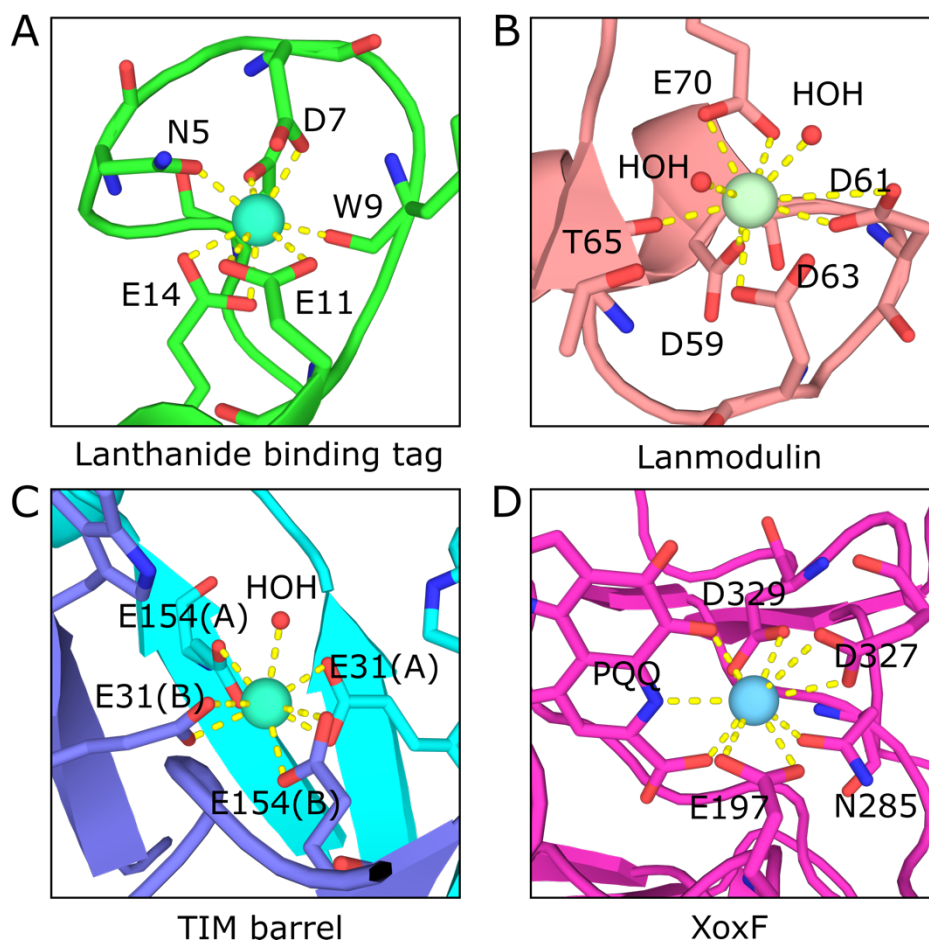


Figure 1.2 – Different binding environments for Ln^{3+} ions. (A) Lanthanide binding tag with Tb^{3+} (PDB ID: 1TJB).⁶⁷ (B) Lanmodulin EF-hand 2 with neodymium(III) (PDB ID: 8FNS).⁶² (C) Engineered TIM barrel homodimer TFD-EE N6W with Tb^{3+} (PDB ID: 6ZV9).⁷⁵ (D) Methanol dehydrogenase XoxF with lanthanum(III) (PDB ID: 6DAM).⁵⁸

Non-EF-hand designs use symmetric arrangements to pre-organise negatively charged binding pockets. C_3 -symmetric coiled-coil proteins, designed around known coiled-coil motifs, have Asp and Asn residues strategically placed for coordination, and a nearby Trp for excitation.⁷³ These proteins effectively bound Tb^{3+} ions and excluded water from the first coordination sphere. A *de novo* designed TIM-barrel fold, reported in 2020, is a C_2 -symmetric dimer using Glu residues preorganised to mimic EDTA-like binding geometry, and includes two nearby Trp residues to excite a bound Tb^{3+} ion (Figure 1.2C).⁷⁴ A follow-

up study engineered a photocatalyst containing cerium(III), which catalysed the breaking of a C–C bond following visible light excitation.⁷⁵

Designing lanthanide-luminescent proteins requires coordinatively saturating Ln^{3+} with high affinity binding sites, while positioning an antenna for efficient energy transfer. In practical terms, efficient sensitisation requires minimising non-radiative quenching (inner sphere water, high-frequency vibrations) while maintaining sub-nanometre antenna– Ln^{3+} distances that support rapid energy transfer. Sensitisation efficiency depends on spectral overlap between antenna emission and Ln^{3+} absorption, and on antenna– Ln^{3+} distance with the transfer rate decaying steeply as distance increases. The choice of antenna also influences back-transfer from the excited Ln^{3+} ion to the antenna, influencing Ln^{3+} radiative emission. Coordinating Ln^{3+} ions via the Trp backbone carbonyl positions the indole nearby but does not guarantee optimal distance. In EF-hands, hydrophobic Trp sidechains may orient away from the negatively charged Ln^{3+} site. This thesis addresses balancing strong coordination, solvent exclusion and antenna distance while preserving favourable energy-level alignment.

1.5 Contemporary protein design

The past decade has seen a revolution in structure prediction and design. Rosetta's physics-based energy functions enabled the first designed protein with topologies not observed in nature,^{76,78} but could not predict tertiary structure from sequence alone. AlphaFold's transformer architectures, trained on >170,000 structures available on the PDB, then achieved near-experimental accuracy for many proteins by interpreting co-evolutionary contacts from multiple sequence alignments.⁷⁹ RFdiffusion generates novel folds from geometric constraints using diffusion models.⁸⁰ The 2024 Nobel Prize in Chemistry, awarded to the authors of AlphaFold and RFdiffusion, marks the point where high-accuracy prediction and design of protein structures became practically accessible. However, while these tools are exceptional for predicting static structure from sequence, they remain limited in contexts where structure alone is insufficient to determine function. The extant popularity of sequence design approaches outside

explicit 3D structure optimisation—ancestral protein reconstruction,⁸¹ directed evolution,⁸² targeted mutagenesis libraries⁸³—is evidence of this. Protein engineering is nonetheless significantly enhanced by using accessible, high-confidence models before experimental validation. These tools, however, operate primarily on canonical amino acid sequences. For photoluminescent proteins, critical design parameters, including conformational dynamics, post-translational modifications, and cofactors, are not captured in the primary sequence of canonical amino acids. Consequently, current structure-prediction tools are excellent for generating static backbone models but have limited predictive power for the photophysical properties of proteins containing chromophores, fluorescent ncAAs or lanthanide centres.

Conformational dynamics are critical when designing photoluminescent proteins as non-radiative pathways are often intricately related to dynamics. For a given sequence, AlphaFold generates a small set of structures,⁷⁹ yet, proteins exist as conformational ensembles. For photoluminescent proteins, these ensembles include sidechain rotamer exchanges, loop ‘breathing’ around chromophores, and fluctuations in metal coordination, all of which can open or close non-radiative pathways. Low-populated, transient states may dominate radiative decay but be invisible to single-structure predictors, further decoupling predicted structure from actual photophysical behaviour. While AlphaFold's pLDDT scores indicate uncertainty for given regions,⁸⁴ but experimental characterisation remains essential for determining conformational dynamics. Ensemble generation methods (AlphaFold-based⁸⁵ or BioEmu)⁸⁶ allow for the inference conformational dynamics and minor populations, but remain limited to proteins containing only canonical amino acids. Photoluminescent proteins should be designed with physiological conditions in mind, meaning they will often be at non-cryogenic temperatures, surrounded by solvent, and interacting transiently with salt ions. For optimal emission, light-emitting moieties benefit from precise geometric constraints and limited non-radiative decay pathways.⁸⁷ Predicting whether a designed protein will achieve the geometry required for favourable photophysical properties under dynamic, physiological conditions remains an open challenge.

AlphaFold3 and RoseTTAFold All-Atom have extended prediction to select post-translational modifications, metal ions and organic ligands.^{28,88} However, coverage remains limited to well-represented modifications in training data. Chromophores, most ncAAs, and elements beyond period 4 are relatively scarce on the PDB. Interestingly, AlphaFold2 can distinguish chromophore-forming from non-chromophore-forming GFP-like proteins based on predicted α -helix geometry,⁸⁹ and researchers have extended AlphaFold2 to include unseen ncAAs by training on large small-molecule datasets.⁹⁰ This suggests high-confidence structure prediction for proteins with fluorescent moieties may be possible with existing data, but tools with comparable confidence to AlphaFold3 are not yet available.

Designing lanthanide-luminescent proteins requires scaffolds with preorganised residues that create specific binding sites. Large symmetric protein scaffolds are particularly well-suited for this purpose. High-confidence interfaces can be amplified by symmetry operators, and the same symmetry operator can build a binding pocket, coupling the design of a multimer to the target binding pocket. The aforementioned *de novo* TIM-barrel fold by Zeymer and co-workers used C_2 -symmetry constraints in Rosetta to design linkers between subunits, creating a dimeric lanthanide-binding scaffold at the symmetry axis.^{74,75} For other metals, target coordination geometries have been achieved through similar strategies. A C_3 -symmetric protein scaffold with bipyridine-ncAAs was designed to bind Fe^{2+} in a specific chiral octahedral coordination geometry.⁹¹ RFdiffusion was used to engineer C_4 -symmetric assemblies coordinating Ni^{2+} with square-planar geometry.⁸⁰ Computational redesign transformed a multi-chain C_3 -symmetric homotrimer scaffold into a single chain pseudosymmetric protein that retains the tetrahedral His_3/H_2O-Zn^{2+} coordination motif, but with an asymmetric binding pocket for alternate substrates.⁹² While a reliable strategy for engineering novel metal binding scaffolds, the resultant assemblies are necessarily large, typically between 30 and 100 kDa, in order to accommodate the protein-protein interfaces required for the scaffold. Large size may be an inherent feature of symmetry-based design strategies, but smaller, asymmetric metalloproteins would be more practical as fusion partners and more informative for grafting metal-binding motifs into existing proteins.

1.6 Research outline

Photoluminescent proteins have diverse biophysical origins, and contemporary protein design tools provide high-confidence structural predictions. However, connecting structure to photophysical properties remains challenging. Computational tools cannot predict the structure of matured FP chromophores, how ncAAs will influence local protein structure, or whether metal-binding sites will achieve optimal antenna geometry under physiological conditions. This thesis investigates structural and biophysical principles governing photoluminescent protein design through three complementary systems: lanthanide-luminescent proteins, fluorescent ncAAs, and fluorescent proteins.

The central questions addressed in this thesis are: (i) how can lanthanide-binding proteins be designed with defined coordination environments and efficient antennae; (ii) how do fluorescent ncAAs alter local protein structure and dynamics; and (iii) how do these structural changes translate into photophysical behaviour?

Chapter 2 explores lanthanide-luminescent peptides and proteins. Using an LBT as a model complex, ncAAs are tested as antennae. The tryptophan analogue 6-azatryptophan (6AW) was capable of exciting both Tb^{3+} and Eu^{3+} . *De novo* design produces lanthanide-binding mini-proteins with high affinity ($K_D \approx 15$ nM) and selectivity. A mini-protein/antenna pair (RF2/6AW) shows enhanced sensitisation over LBT in biological contexts, illustrating how protein design and antenna can be co-optimised for Ln^{3+} emission in complex environments.

Chapter 3 characterises the Trp orientations in RF2, combining rotamer simulations of computationally predicted structures with ^{19}F - Gd^{3+} ENDOR to obtain structural information on a dynamic sidechain. Analysis of ENDOR spectra reveals rotamers absent from predicted structures. This establishes ^{19}F - Gd^{3+} ENDOR as a method for characterising dynamic sidechains.

Chapter 4 presents high-resolution X-ray crystal structures of maltose-binding protein with 4-F-Trp and 4-CN-Trp. These structures reveal how 4-substitutions affect Trp rotamer preferences. This identifies structural determinants of substitution tolerance.

Chapter 5 presents structures of proteins containing two fluorescent ncAAs (nitrobenzoxadiazol-L-diaminopropionic acid (NBDO) and nitrobenzoselenadiazol-L-diaminopropionic acid (NBDS_e)) incorporated via GCE. Despite their larger size, NBDO and NBDS_e are accommodated at Trp sites. Solvent exclusion significantly enhances fluorescence. In NowGFP, the FP chromophore maturation pathway proceeds despite the exocyclic secondary amine linker. This amine-substituted methine linker has not yet been reported in proteins.

Chapter 6 summarises the biophysical insights derived from the work in the previous chapters and proposes future work to build on the principles explored in this thesis. Together, the structural and spectroscopic data across Chapters 2–5 provide an experimentally grounded framework for understanding the biophysical principles dictating the photoluminescent properties of proteins. This work supports the rational engineering of photoluminescent proteins.

1.7 References

1. Lichtman, J. W.; Conchello, J. A. Fluorescence microscopy. *Nat. Methods* **2005**, *2*, 910–919. <https://doi.org/10.1038/nmeth817>
2. Shroff, H.; Testa, I.; Jug, F.; Manley, S. Live-cell imaging powered by computation. *Nat. Rev. Mol. Cell Biol.* **2024**, *25*, 443–463. <https://doi.org/10.1038/s41580-024-00702-6>
3. Nicoli, A. M.; Frei, M. S. Synthetic fluorophores for live-cell fluorescence microscopy and biosensing. *Nat. Chem. Biol.* **2025**. <https://doi.org/10.1038/s41589-025-02038-4>
4. D'Este, E.; Lukinavicius, G.; Lincoln, R.; Opazo, F.; Fornasiero, E. F. Advancing cell biology with nanoscale fluorescence imaging: Essential practical considerations. *Trends Cell Biol.* **2024**, *34*, 671–684. <https://doi.org/10.1016/j.tcb.2023.12.001>
5. Inavalli, V.; Puente Munoz, V.; Draffin, J. E.; Tonnesen, J. Fluorescence microscopy shadow imaging for neuroscience. *Front. Cell Neurosci.* **2024**, *18*, 1330100. <https://doi.org/10.3389/fncel.2024.1330100>
6. Veerapathiran, S.; Wohland, T. Fluorescence techniques in developmental biology. *J. Biosci.* **2018**, *43*, 541–553. <https://doi.org/10.1007/s12038-018-9768-z>
7. Wang, X.; Ding, Q.; Groleau, R. R.; Wu, L.; Mao, Y.; Che, F.; Kotova, O.; Scanlan, E. M.; Lewis, S. E.; Li, P.; Tang, B.; James, T. D.; Gunnlaugsson, T. Fluorescent probes for disease diagnosis. *Chem. Rev.* **2024**, *124*, 7106–7164. <https://doi.org/10.1021/acs.chemrev.3c00776>
8. Roostalu, U.; Hansen, H. H.; Hecksher-Sorensen, J. 3D light-sheet fluorescence microscopy in preclinical and clinical drug discovery. *Drug Discov. Today* **2024**, *29*, 104196. <https://doi.org/10.1016/j.drudis.2024.104196>
9. Astner, S.; Dietterle, S.; Otberg, N.; Rowert-Huber, H. J.; Stockfleth, E.; Lademann, J. Clinical applicability of *in vivo* fluorescence confocal microscopy for noninvasive diagnosis and therapeutic monitoring of nonmelanoma skin cancer. *J. Biomed. Opt.* **2008**, *13*, 014003. <https://doi.org/10.1117/1.2837411>
10. Ludvikova, L.; Simon, E.; Deygas, M.; Panier, T.; Plamont, M. A.; Ollion, J.; Tebo, A.; Piel, M.; Jullien, L.; Robert, L.; Le Saux, T.; Espagne, A. Near-infrared co-illumination of fluorescent proteins reduces photobleaching and phototoxicity. *Nat. Biotechnol.* **2024**, *42*, 872–876. <https://doi.org/10.1038/s41587-023-01893-7>
11. Debarre, D.; Olivier, N.; Supatto, W.; Beaupaire, E. Mitigating phototoxicity during multiphoton microscopy of live *Drosophila* embryos in the 1.0–1.2 μm wavelength range. *PLoS One* **2014**, *9*, e104250. <https://doi.org/10.1371/journal.pone.0104250>
12. Jun, Y. W.; Kim, H. R.; Reo, Y. J.; Dai, M.; Ahn, K. H. Addressing the autofluorescence issue in deep tissue imaging by two-photon microscopy: The significance of far-red emitting dyes. *Chem. Sci.* **2017**, *8*, 7696–7704. <https://doi.org/10.1039/c7sc03362a>
13. Jacques, S. L. Optical properties of biological tissues: A review. *Phys. Med. Biol.* **2013**, *58*, R37–61. <https://doi.org/10.1088/0031-9155/58/11/R37>
14. Chudakov, D. M.; Matz, M. V.; Lukyanov, S.; Lukyanov, K. A. Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol. Rev.* **2010**, *90*, 1103–1163. <https://doi.org/10.1152/physrev.00038.2009>
15. Örmö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science* **1996**, *273*, 1392–1395. <https://doi.org/10.1126/science.273.5280.1392>
16. Tsien, R. Y. The green fluorescent protein. *Annu. Rev. Biochem.* **1998**, *67*, 509–544. <https://doi.org/10.1146/annurev.biochem.67.1.509>
17. Ai, H. W.; Shaner, N. C.; Cheng, Z.; Tsien, R. Y.; Campbell, R. E. Exploration of new chromophore structures leads to the identification of improved blue fluorescent proteins. *Biochemistry* **2007**, *46*, 5904–5910. <https://doi.org/10.1021/bi700199g>
18. Chudakov, D. M.; Verkhusha, V. V.; Staroverov, D. B.; Souslova, E. A.; Lukyanov, S.; Lukyanov, K. A. Photoswitchable cyan fluorescent protein for protein tracking. *Nat. Biotechnol.* **2004**, *22*, 1435–1439. <https://doi.org/10.1038/nbt1025>

19. Shaner, N. C.; Campbell, R. E.; Steinbach, P. A.; Giepmans, B. N.; Palmer, A. E.; Tsien, R. Y. Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nat. Biotechnol.* **2004**, *22*, 1567–1572. <https://doi.org/10.1038/nbt1037>
20. Matz, M. V.; Fradkov, A. F.; Labas, Y. A.; Savitsky, A. P.; Zaraisky, A. G.; Markelov, M. L.; Lukyanov, S. A. Fluorescent proteins from nonbioluminescent Anthozoa species. *Nat. Biotechnol.* **1999**, *17*, 969–973. <https://doi.org/10.1038/13657>
21. Verkhusha, V. V.; Lukyanov, K. A. The molecular properties and applications of Anthozoa fluorescent proteins and chromoproteins. *Nat. Biotechnol.* **2004**, *22*, 289–296. <https://doi.org/10.1038/nbt943>
22. Ahmed, F. H.; Caputo, A. T.; French, N. G.; Peat, T. S.; Whitfield, J.; Warden, A. C.; Newman, J.; Scott, C. Over the rainbow: structural characterization of the chromoproteins gfasPurple, amilCP, spisPink and eforRed. *Acta Crystallogr. D Struct. Biol.* **2022**, *78*, 599–612. <https://doi.org/10.1107/S2059798322002625>
23. Yang, X.; Yang, N.; Wu, S.; He, S.; Jiang, W.; Zhong, W.; Du, J.; Zhang, G.; Wang, X.; Yang, S. Establishment of a visible reporter system in *Zymomonas mobilis* through random mutagenesis and rational design of a chromoprotein eforRed. *ACS Synth. Biol.* **2025**, *14*, 2548–2558. <https://doi.org/10.1021/acssynbio.4c00872>
24. Mizuno, H.; Mal, T. K.; Tong, K. I.; Ando, R.; Furuta, T.; Ikura, M.; Miyawaki, A. Photo-induced peptide cleavage in the green-to-red conversion of a fluorescent protein. *Mol. Cell* **2003**, *12*, 1051–1058. [https://doi.org/10.1016/s1097-2765\(03\)00393-9](https://doi.org/10.1016/s1097-2765(03)00393-9)
25. Gadella, T. W. J., Jr.; van Weeren, L.; Stouthamer, J.; Hink, M. A.; Wolters, A. H. G.; Giepmans, B. N. G.; Aumonier, S.; Dupuy, J.; Royant, A. mScarlet3: A brilliant and fast-maturing red fluorescent protein. *Nat. Methods* **2023**, *20*, 541–545. <https://doi.org/10.1038/s41592-023-01809-y>
26. Xiong, H.; Chang, Q.; Ding, J.; Wang, S.; Zhang, W.; Li, Y.; Wu, Y.; Lin, P.; Yang, C.; Liu, M.; Fang, G.; Yang, Y.; Xie, J.; Qi, D.; Jiang, T.; Fu, W.; Hu, F.; Chen, Y.; Yue, R.; Li, Y.; Cui, Y.; Li, M.; Fan, S.; Yang, Y.; Xu, Y.; Li, D.; Zhang, F.; Zhao, H.; Wu, C.; Zheng, Q.; Piatkevich, K. D.; Fu, Z. A highly stable monomeric red fluorescent protein for advanced microscopy. *Nat. Methods* **2025**, *22*, 1288–1298. <https://doi.org/10.1038/s41592-025-02676-5>
27. Zhang, H.; Papadaki, S.; Sun, X.; Wang, X.; Drobizhev, M.; Yao, L.; Rehbock, M.; Koster, R. W.; Wu, L.; Namikawa, K.; Piatkevich, K. D. Quantitative assessment of near-infrared fluorescent proteins. *Nat. Methods* **2023**, *20*, 1605–1616. <https://doi.org/10.1038/s41592-023-01975-z>
28. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C. C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Zengulyte, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlín, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Zidek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
29. Fatima, A.; He, Y.; Iuliano, J. N.; Greetham, G. M.; Malakar, P.; Hall, C.; Woroniecka, H. A.; Richardson, B. C.; French, J. B.; Lukacs, A.; Tonge, P. J.; Meech, S. R. Ultrafast photophysics of a positive reversibly switchable fluorescent protein. *Chem. Sci.* **2025**, *16*, 16955–16969. <https://doi.org/10.1039/d5sc04491j>
30. Ren, H.; Xiao, F.; Zhan, K.; Kim, Y. P.; Xie, H.; Xia, Z.; Rao, J. A biocompatible condensation reaction for the labeling of terminal cysteine residues on proteins. *Angew. Chem., Int. Ed.* **2009**, *48*, 9658–9662. <https://doi.org/10.1002/anie.200903627>
31. Gunnoo, S. B.; Madder, A. Chemical protein modification through cysteine. *ChemBioChem* **2016**, *17*, 529–553. <https://doi.org/10.1002/cbic.201500667>
32. Grimm, J. B.; Lavis, L. D. Caveat fluorophore: An insiders' guide to small-molecule fluorescent labels. *Nat. Methods* **2022**, *19*, 149–158. <https://doi.org/10.1038/s41592-021-01338-6>
33. Jing, C.; Cornish, V. W. Chemical tags for labeling proteins inside living cells. *Acc. Chem. Res.* **2011**, *44*, 784–792. <https://doi.org/10.1021/ar200099f>
34. Koch, N. G.; Budisa, N. Evolution of pyrrolysyl-tRNA synthetase: From methanogenesis to genetic code expansion. *Chem. Rev.* **2024**, *124*, 9580–9608. <https://doi.org/10.1021/acs.chemrev.4c00031>

35. Krahn, N.; Tharp, J. M.; Crnkovic, A.; Soll, D. Engineering aminoacyl-tRNA synthetases for use in synthetic biology. *Enzymes* **2020**, *48*, 351–395. <https://doi.org/10.1016/bs.enz.2020.06.004>
36. Wang, L.; Schultz, P. G. A general approach for the generation of orthogonal tRNAs. *Chem. Biol.* **2001**, *8*, 883–890. [https://doi.org/10.1016/s1074-5521\(01\)00063-1](https://doi.org/10.1016/s1074-5521(01)00063-1)
37. Wals, K.; Ovaa, H. Unnatural amino acid incorporation in *E. coli*: Current and future applications in the design of therapeutic proteins. *Front. Chem.* **2014**, *2*, 15. <https://doi.org/10.3389/fchem.2014.00015>
38. Huang, Y.; Zhang, P.; Wang, H.; Chen, Y.; Liu, T.; Luo, X. Genetic code expansion: Recent developments and emerging applications. *Chem. Rev.* **2025**, *125*, 523–598. <https://doi.org/10.1021/acs.chemrev.4c00216>
39. Abdelkader, E. H.; Qianzhu, H.; Otting, G.; Huber, T. Biosynthesis and genetic encoding of activated nitriles for fast protein conjugation and tunable fluorogenic labeling. *Chem* **2025**, *11*, 102385. <https://doi.org/10.1016/j.chempr.2024.12.003>
40. Abdelkader, E. H.; Qianzhu, H.; George, J.; Frkic, R. L.; Jackson, C. J.; Nitsche, C.; Otting, G.; Huber, T. Genetic encoding of cyanopyridylalanine for in-cell protein macrocyclization by the nitrile–aminothiol click reaction. *Angew. Chem., Int. Ed.* **2022**, *61*, e202114154. <https://doi.org/10.1002/anie.202114154>
41. Lee, S.; Kim, J.; Koh, M. Recent advances in fluorescence imaging by genetically encoded non-canonical amino acids. *J. Mol. Biol.* **2022**, *434*, 167248. <https://doi.org/10.1016/j.jmb.2021.167248>
42. Katoh, T.; Suga, H. *In vitro* genetic code reprogramming for the expansion of usable noncanonical amino acids. *Annu. Rev. Biochem.* **2022**, *91*, 221–243. <https://doi.org/10.1146/annurev-biochem-040320-103817>
43. Dangerfield, T. L.; Johnson, K. A. Optimized incorporation of an unnatural fluorescent amino acid affords measurement of conformational dynamics governing high-fidelity DNA replication. *J. Biol. Chem.* **2020**, *295*, 17265–17280. <https://doi.org/10.1074/jbc.RA120.015557>
44. Zhao, C.; Fernandez, A.; Avlonitis, N.; Vande Velde, G.; Bradley, M.; Read, N. D.; Vendrell, M. Searching for the optimal fluorophore to label antimicrobial peptides. *ACS Comb. Sci.* **2016**, *18*, 689–696. <https://doi.org/10.1021/acscombsci.6b00081>
45. Talukder, P.; Chen, S.; Roy, B.; Yakovchuk, P.; Spiering, M. M.; Alam, M. P.; Madathil, M. M.; Bhattacharya, C.; Benkovic, S. J.; Hecht, S. M. Cyanotryptophans as novel fluorescent probes for studying protein conformational changes and DNA-protein interaction. *Biochemistry* **2015**, *54*, 7457–7469. <https://doi.org/10.1021/acs.biochem.5b01085>
46. Gleason, P. R.; Kolbaba-Kartchner, B.; Henderson, J. N.; Stahl, E. P.; Simmons, C. R.; Mills, J. H. Structural origins of altered spectroscopic properties upon ligand binding in proteins containing a fluorescent noncanonical amino acid. *Biochemistry* **2021**, *60*, 2577–2585. <https://doi.org/10.1021/acs.biochem.1c00291>
47. Huang, Y.; Zhang, P.; Wang, H.; Chen, Y.; Liu, T.; Luo, X. Genetic code expansion: Recent developments and emerging applications. *Chem. Rev.* **2025**, *125*, 523–598. <https://doi.org/10.1021/acs.chemrev.4c00216>
48. Henderson, J. N.; Simmons, C. R.; Fahmi, N. E.; Jeffs, J. W.; Borges, C. R.; Mills, J. H. Structural insights into how protein environments tune the spectroscopic properties of a noncanonical amino acid fluorophore. *Biochemistry* **2020**, *59*, 3401–3410. <https://doi.org/10.1021/acs.biochem.0c00474>
49. Goodwin, C. A. P.; Corbey, J. F. Ligand–metal complementarity in rare-earth and actinide chemistry. *Inorg. Chem.* **2024**, *63*, 9355–9362. <https://doi.org/10.1021/acs.inorgchem.4c01504>
50. Bünzli, J.-C. G. Lanthanide luminescence for biomedical analyses and imaging. *Chem. Rev.* **2010**, *110*, 2729–2755. <https://doi.org/10.1021/cr900362e>
51. Bünzli, J.-C. G. On the design of highly luminescent lanthanide complexes. *Coord. Chem. Rev.* **2015**, *293*, 19–47. <https://doi.org/10.1016/j.ccr.2014.10.013>
52. Falcone, E.; Mathieu, E.; Hureau, C. Lanthanide(III)-binding peptides and proteins: Coordination properties and applications. *Chem. Soc. Rev.* **2025**, *54*, 9245–9288. <https://doi.org/10.1039/d4cs01148a>
53. Alexander, C.; Guo, Z.; Glover, P. B.; Faulkner, S.; Pikramenou, Z. Luminescent lanthanides in biorelated applications: From molecules to nanoparticles and diagnostic probes to therapeutics. *Chem. Rev.* **2025**, *125*, 2269–2370. <https://doi.org/10.1021/acs.chemrev.4c00615>
54. Hemmila, I.; Laitala, V. Progress in lanthanides as luminescent probes. *J. Fluoresc.* **2005**, *15*, 529–542. <https://doi.org/10.1007/s10895-005-2826-6>
55. Bodman, S. E.; Butler, S. J. Advances in anion binding and sensing using luminescent lanthanide complexes. *Chem. Sci.* **2021**, *12*, 2716–2734. <https://doi.org/10.1039/d0sc05419d>

56. Parker, D.; Fradgley, J. D.; Wong, K. L. The design of responsive luminescent lanthanide probes and sensors. *Chem. Soc. Rev.* **2021**, *50*, 8193–8213. <https://doi.org/10.1039/d1cs00310k>
57. Ochsner, A. M.; Hemmerle, L.; Vonderach, T.; Nussli, R.; Bortfeld-Miller, M.; Hattendorf, B.; Vorholt, J. A. Use of rare-earth elements in the phyllosphere colonizer *Methylobacterium extorquens* PA1. *Mol. Microbiol.* **2019**, *111*, 1152–1166. <https://doi.org/10.1111/mmi.14208>
58. Deng, Y. W.; Ro, S. Y.; Rosenzweig, A. C. Structure and function of the lanthanide-dependent methanol dehydrogenase XoxF from the methanotroph *Methylomicrobium buryatense* 5GB1C. *J. Biol. Inorg. Chem.* **2018**, *23*, 1037–1047. <https://doi.org/10.1007/s00775-018-1604-2>
59. Wehrmann, M.; Billard, P.; Martin-Meriadec, A.; Zegeye, A.; Klebensberger, J. Functional role of lanthanides in enzymatic activity and transcriptional regulation of pyrroloquinoline quinone-dependent alcohol dehydrogenases in *Pseudomonas putida* KT2440. *mBio* **2017**, *8*, e00570-17. <https://doi.org/10.1128/mBio.00570-17>
60. Larrinaga, W. B.; Jung, J. J.; Lin, C. Y.; Boal, A. K.; Cotruvo, J. A., Jr. Modulating metal-centered dimerization of a lanthanide chaperone protein for separation of light lanthanides. *Proc. Natl. Acad. Sci. U.S.A.* **2024**, *121*, e2410926121. <https://doi.org/10.1073/pnas.2410926121>
61. Cotruvo, J. A., Jr.; Featherston, E. R.; Mattocks, J. A.; Ho, J. V.; Laremore, T. N. Lanmodulin: A highly selective lanthanide-binding protein from a lanthanide-utilizing bacterium. *J. Am. Chem. Soc.* **2018**, *140*, 15056–15061. <https://doi.org/10.1021/jacs.8b09842>
62. Cook, E. C.; Featherston, E. R.; Showalter, S. A.; Cotruvo, J. A. Structural basis for rare earth element recognition by lanmodulin. *Biochemistry* **2019**, *58*, 120–125. <https://doi.org/10.1021/acs.biochem.8b01019>
63. Jones, E. M.; Su, Y.; Sander, C.; Justman, Q. A.; Springer, M.; Silver, P. A. LanTERN: A fluorescent sensor that specifically responds to lanthanides. *ACS Synth. Biol.* **2024**, *13*, 958–962. <https://doi.org/10.1021/acssynbio.3c00600>
64. Wang, J.; Liu, X.; Li, K.; Shi, T.; Xu, Q.; Peng, T.; Huang, Q.; Gao, Z.; Zhou, H.; Lu, W.; Wang, J. Design and evolution of a phosphorescent protein via the proximal encoding of lanthanide and the antenna chromophore. *J. Am. Chem. Soc.* **2025**, *147*, 15205–15215. <https://doi.org/10.1021/jacs.5c00199>
65. Larrinaga, W. B.; Cotruvo, J. A., Jr. A tale of two dimers: Lanthanide recognition at biomolecular interfaces. *Trends Biochem. Sci.* **2025**, e20251022. <https://doi.org/10.1016/j.tibs.2025.08.008>
66. Cotruvo, J. A., Jr. The chemistry of lanthanides in biology: Recent discoveries, emerging principles, and technological applications. *ACS Cent. Sci.* **2019**, *5*, 1496–1506. <https://doi.org/10.1021/acscentsci.9b00642>
67. Martin, L. L.; Sculimbrene, B. R.; Nitz, M.; Imperiali, B. Rapid combinatorial screening of peptide libraries for the selection of lanthanide-binding tags (LBTs). *QSAR Comb. Sci.* **2005**, *24*, 1149–1157. <https://doi.org/10.1002/qsar.200540007>
68. Nitz, M.; Franz, K. J.; Maglathlin, R. L.; Imperiali, B. A powerful combinatorial screen to identify high-affinity terbium(III)-binding peptides. *ChemBioChem* **2003**, *4*, 272–276. <https://doi.org/10.1002/cbic.200390047>
69. Nitz, M.; Sherawat, M.; Franz, K. J.; Peisach, E.; Allen, K. N.; Imperiali, B. Structural origin of the high affinity of a chemically evolved lanthanide-binding peptide. *Angew. Chem., Int. Ed.* **2004**, *43*, 3682–3685. <https://doi.org/10.1002/anie.200460028>
70. Sculimbrene, B. R.; Imperiali, B. Lanthanide-binding tags as luminescent probes for studying protein interactions. *J. Am. Chem. Soc.* **2006**, *128*, 7346–7352. <https://doi.org/10.1021/ja061188a>
71. Musial-Siwiek, M.; Jaffee, M. B.; Imperiali, B. Probing polytopic membrane protein-substrate interactions by luminescence resonance energy transfer. *J. Am. Chem. Soc.* **2016**, *138*, 3806–3812. <https://doi.org/10.1021/jacs.5b13426>
72. Reynolds, A. M.; Sculimbrene, B. R.; Imperiali, B. Lanthanide-binding tags with unnatural amino acids: sensitizing Tb³⁺ and Eu³⁺ luminescence at longer wavelengths. *Bioconjugate Chem.* **2008**, *19*, 588–591. <https://doi.org/10.1021/bc700426c>
73. Berwick, M. R.; Lewis, D. J.; Jones, A. W.; Parslow, R. A.; Dafforn, T. R.; Cooper, H. J.; Wilkie, J.; Pikramenou, Z.; Britton, M. M.; Peacock, A. F. De novo design of Ln(III) coiled coils for imaging applications. *J. Am. Chem. Soc.* **2014**, *136*, 1166–1169. <https://doi.org/10.1021/ja408741h>
74. Caldwell, S. J.; Haydon, I. C.; Piperidou, N.; Huang, P. S.; Bick, M. J.; Sjostrom, H. S.; Hilvert, D.; Baker, D.; Zeymer, C. Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity

- designed by symmetric domain fusion. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 30362–30369. <https://doi.org/10.1073/pnas.2008535117>
75. Klein, A. S.; Leiss-Maier, F.; Muhlhofer, R.; Boesen, B.; Mustafa, G.; Kugler, H.; Zeymer, C. A *de novo* metalloenzyme for cerium photoredox catalysis. *J. Am. Chem. Soc.* **2024**, *146*, 25976–25985. <https://doi.org/10.1021/jacs.4c04618>
76. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209–225. <https://doi.org/10.1006/jmbi.1997.0959>
77. Park, H.; Bradley, P.; Greisen, P., Jr.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **2016**, *12*, 6201–6212. <https://doi.org/10.1021/acs.jctc.6b00819>
78. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364–1368. <https://doi.org/10.1126/science.1089427>
79. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
80. Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100. <https://doi.org/10.1038/s41586-023-06415-8>
81. Spence, M. A.; Kaczmarek, J. A.; Saunders, J. W.; Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **2021**, *69*, 131–141. <https://doi.org/10.1016/j.sbi.2021.04.001>
82. Selles Vidal, L.; Isalan, M.; Heap, J. T.; Ledesma-Amaro, R. A primer to directed evolution: Current methodologies and future directions. *RSC Chem. Biol.* **2023**, *4*, 271–291. <https://doi.org/10.1039/d2cb00231k>
83. Tan, L.; Zheng, Z.; Xu, Y.; Kong, W.; Dai, Z.; Qin, X.; Liu, T.; Tang, H. Efficient selection scheme for incorporating noncanonical amino acids into proteins in *Saccharomyces cerevisiae*. *Front. Bioeng. Biotechnol.* **2020**, *8*, 569191. <https://doi.org/10.3389/fbioe.2020.569191>
84. Terwilliger, T. C.; Liebschner, D.; Croll, T. I.; Williams, C. J.; McCoy, A. J.; Poon, B. K.; Afonine, P. V.; Oeffner, R. D.; Richardson, J. S.; Read, R. J.; Adams, P. D. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat. Methods* **2024**, *21*, 110–116. <https://doi.org/10.1038/s41592-023-02087-4>
85. Del Alamo, D.; Sala, D.; McHaourab, H. S.; Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **2022**, *11*, e75751. <https://doi.org/10.7554/eLife.75751>
86. Lewis, S.; Hempel, T.; Jimenez-Luna, J.; Gastegger, M.; Xie, Y.; Foong, A. Y. K.; Satorras, V. G.; Abidin, O.; Veeling, B. S.; Zaporozhets, I.; Chen, Y.; Yang, S.; Foster, A. E.; Schneuing, A.; Nigam, J.; Barbero, F.; Stimper, V.; Campbell, A.; Yim, J.; Lienen, M.; Shi, Y.; Zheng, S.; Schulz, H.; Munir, U.; Sordillo, R.; Tomioka, R.; Clementi, C.; Noe, F. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* **2025**, *389*, eadv9817. <https://doi.org/10.1126/science.adv9817>
87. Lee, M. H.; Kim, J. S.; Sessler, J. L. Small molecule-based ratiometric fluorescence probes for cations, anions, and biomolecules. *Chem. Soc. Rev.* **2015**, *44*, 4185–4191. <https://doi.org/10.1039/c4cs00280f>
88. Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R.; McHugh, R.; Vafeados, D.; Li, X.; Sutherland, G. A.; Hitchcock, A.; Hunter, C. N.; Kang, A.; Brackenbrough, E.; Bera, A. K.; Baek, M.; DiMaio, F.; Baker, D. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **2024**, *384*, eadl2528. <https://doi.org/10.1126/science.adl2528>

89. Hartley, S. M.; Tiernan, K. A.; Ahmetaj, G.; Cretu, A.; Zhuang, Y.; Zimmer, M. AlphaFold2 and RoseTTAFold predict posttranslational modifications: Chromophore formation in GFP-like proteins. *PLoS One* **2022**, *17*, e0267560. <https://doi.org/10.1371/journal.pone.0267560>
90. Lee, J. S.; Kim, P. M. Design of peptides with non-canonical amino acids using flow matching. *bioRxiv* **2025**. <https://doi.org/10.1101/2025.07.31.667780>
91. Mills, J. H.; Sheffler, W.; Ener, M. E.; Almhjell, P. J.; Oberdorfer, G.; Pereira, J. H.; Parmeggiani, F.; Sankaran, B.; Zwart, P. H.; Baker, D. Computational design of a homotrimeric metalloprotein with a trisbipyridyl core. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 15012–15017. <https://doi.org/10.1073/pnas.1600188113>
92. Hoffnagle, A. M.; Srisantitham, S.; Neeley, M.; Tsai, C. Y.; Tezcan, F. A. A *de novo* designed protein with versatile metal binding and tunable hydrolytic activity. *Biochemistry* **2025**, *64*, 3261–3271. <https://doi.org/10.1021/acs.biochem.5c00259>

Chapter 2: Biosynthetic lanthanide-luminescent mini-proteins using genetic code expansion

2.1 Introduction

Lanthanide ions are promising luminescent probes: they exhibit long luminescence lifetimes (0.5–3 ms), large pseudo-Stokes shifts (150–300 nm) and narrow emission bands.¹ These photophysical properties are attractive for time-resolved fluorescence imaging as used with great benefit in immunoassays.^{2,3} In biological systems, however, free lanthanides readily precipitate with phosphates and their toxicity warrants strong chelators. Most critically, the molar extinction coefficients of lanthanide ions are extremely low, typically ranging between 0.1–10 M⁻¹cm⁻¹, which makes direct excitation inefficient. Fortunately, the lanthanide luminescence can be enhanced indirectly through nearby organic chromophores. This phenomenon, known as the antenna effect, involves the photoexcitation of the organic chromophore followed by energy transfer to the lanthanide ion, overcoming the problem of the lanthanide's low intrinsic absorptivity.

Choosing an optimal antenna is difficult because the energy transfer mechanism is complex. After excitation of the chromophore to the singlet state (S_1), lanthanide(III) (Ln^{3+}) excitation can occur via multiple pathways.¹ The canonical pathway involves intersystem crossing (ISC) to the long-lived triplet state (T_1), which in turn excites the Ln^{3+} ion.¹ Sensitisation therefore depends on the ISC efficiency as well as the T_1 - Ln^{3+} spectral overlap, rather than the more easily measurable absorbance and fluorescence properties of the antenna. In addition, energy transfer is also possible directly from the S_1 state to the Ln^{3+} ion, although this mechanism is usually disadvantaged by its much shorter (nanosecond) lifetime.¹ Nonetheless, direct S_1 sensitisation can be significant when there is strong spectral overlap, which has been explicitly modelled in protein- Ln^{3+} complexes.⁴ Crucially, both mechanisms rely on the antenna being in close spatial proximity to the Ln^{3+} ion and overlapping spectral properties. Additionally, Ln^{3+} emission is readily quenched by nearby water molecules, necessitating the exclusion of inner-sphere waters to achieve appreciable emission, even when other factors are favourable.⁵ Given these intertwined variables, predicting an antenna's performance *a priori* is challenging, prompting the empirical strategy taken in this chapter.

Peptidic lanthanide-binding tags (LBTs), developed by Imperiali and co-workers,⁶⁻¹² are biological lanthanide chelators with high specificity and affinity to lanthanide ions. LBTs are derived from the EF-hand motif, which is a helix-loop-helix structural domain responsible for Ca^{2+} coordination in proteins.^{6,7} They were specifically designed to include a tryptophan (Trp) residue as the antenna for terbium(III) (Tb^{3+}) excitation, with the backbone carbonyl oxygen of the Trp coordinating the lanthanide, positioning the Trp indole ring close to the Tb^{3+} ion to efficiently photo-sensitise the Tb^{3+} luminescence. Additional carboxyl groups from acidic sidechains and carbonyls from amide-containing sidechains prevent the direct access of water to the lanthanide ion, enhancing the luminescence lifetimes and integrated emission intensities. Among the canonical amino acids, Trp has the highest molar absorption coefficient ($\sim 5500 \text{ M}^{-1}\text{cm}^{-1}$ at 280 nm) and overall provides the best antenna functionality, although Tyrosine (Tyr) at the same position has also been established to sensitise Tb^{3+} luminescence. Other Ln^{3+} ions are more strongly quenched by inner-sphere water than Tb^{3+} and thus their complexes have weaker emission with Trp or Tyr antennae. When the site is well dehydrated ($q \approx 0$), Trp and Tyr can still sensitise these ions, the remaining differences reflect energy-level matching rather than water quenching.

Over the last decade, there has been increasing interest in biosynthetic molecules to detect, isolate and utilise lanthanides. In 2018, lanmodulin, a protein from a lanthanide-utilizing bacterium, was identified to have picomolar affinity for lanthanide ions.¹³ Lanmodulin, which is structurally analogous to calmodulin, presents a lanthanide chelating environment similar to the LBTs with one residue coordinating the Ln^{3+} ion by a threonine (Thr) backbone carbonyl oxygen. Featherston *et al.* reported that mutation of this amino acid to Trp sensitises a Tb^{3+} ion for excitation at 280 nm.¹⁴ Further studies with luminescent lanthanide-binding proteins and peptides have been reviewed by Pazos and co-workers.¹⁵ However, few attempts have been made to generate new lanthanide-binding peptides or proteins explicitly for optimal luminescence properties.

Substituting canonical amino acids with chemically synthesised antennae has been shown to allow for different lanthanide ions to be excited. Reynolds *et al.* used solid-phase peptide synthesis to substitute a tryptophan with a non-canonical amino acid

(ncAA) containing either an acridone or carbostyryl sidechain. This enhanced the excitation of europium(III) (Eu^{3+}) luminescence and, with acridone, the luminosity of Tb^{3+} .⁹ Other peptide-based lanthanide-luminescent probes were synthetically modified to include a fluorescent moiety,¹⁶⁻¹⁹ although this approach tends to increase the distance between the antenna and the lanthanide ion. These peptides require specific chemical synthesis, and none can be produced *in vivo*.

Genetic code expansion (GCE) enables the site-specific installation of spectroscopically active ncAAs during protein biosynthesis in *E. coli* by using orthogonal aminoacyl-tRNA synthetase/tRNA pairs with amber stop suppression.²⁰ The repertoire of genetically encoded tryptophan analogues has expanded significantly in recent years, opening a modular route to incorporate antennae directly in recombinantly produced polypeptides.²¹⁻²⁴ The work in this chapter systematically explores the performance of these ncAAs (Figure 1) as substitutes of Trp at the lanthanide-proximal site.

The ncAAs feature excitation wavelengths in the range 260–350 nm, which proved particularly beneficial for Eu^{3+} sensitisation. This chapter analyses the relationship between the antennae and the photo-properties of the Ln^{3+} complexes. Building on the experience gained with LBTs, computational design led to the synthesis of three new lanthanide-binding mini-proteins. Two of these showed increased luminosity compared to their LBT counterparts and one displayed low-nanomolar binding affinity with Ln^{3+} ions. The best performing antenna was successfully combined with the most tightly binding lanthanide-binding protein, indicating that the luminescence enhancement achieved by the ncAAs is generalisable to other protein scaffolds, providing a basis for new *in vivo* probes.

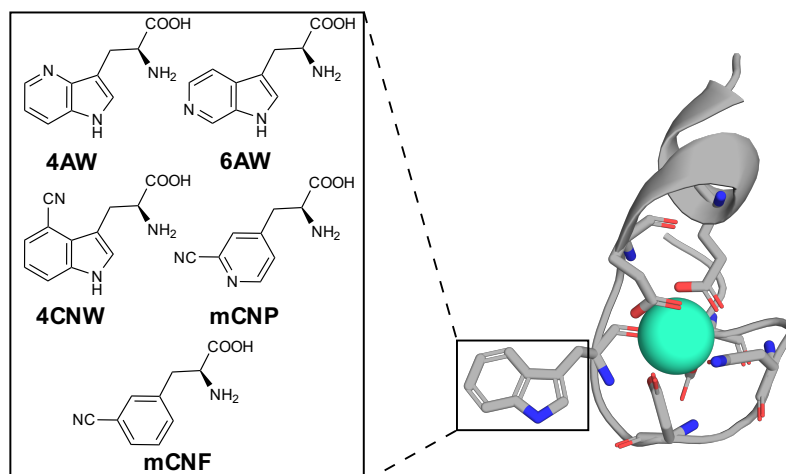


Figure 2.1 – Binding geometry of the LBT with a Tb^{3+} ion bound. The chemical structures are shown for the ncAAs used to substitute the antenna tryptophan, including 4-azatryptophan (4AW), 6-azatryptophan (6AW), 4-cyanotryptophan (4CNW), 3-cyano-4-pyridylalanine (mCNP), and 3-cyanophenylalanine (mCNF). In the LBT crystal structure (PDB ID: 1TJB)⁷ The Tb^{3+} ion has a coordination number of eight, satisfied by two bidentate acidic residues, one monodentate acidic residue, the backbone carbonyl of Trp and the oxygen of an asparagine sidechain.

This chapter first benchmarks a panel of genetically encoded ncAAs in a well-characterised LBT scaffold, quantifying their effects on brightness and lanthanide affinity. Then, using computational design around an EF-hand of lanmodulin, new lanthanide-binding mini-proteins are generated, with RF2 standing out as a scaffold with improved Tb^{3+} luminosity and nanomolar affinity. Finally, this chapter combines RF2 with the best-performing antenna and shows that this construct retains high affinity and supports the detection of Eu^{3+} and Tb^{3+} in cell cultures. Together, these results define a practical approach for designing genetically encoded lanthanide-luminescent probes.

2.2 Photoluminescent properties of LBTs with ncAAs

The LBT with the sequence GFIDTNNDGWIEGDELLEG serves as a benchmark scaffold, due to strong Tb^{3+} binding affinity and specificity, low intrinsic fluorescence and documented luminescence.⁸ The Trp at position 10 marks the antenna position, which was modified by ncAAs. The N-terminal glycine is a remnant of the N-terminal TEV cleavage site used to liberate the peptide following production as a fusion with the solubilizing NT* domain derived from spider silk protein.²⁵

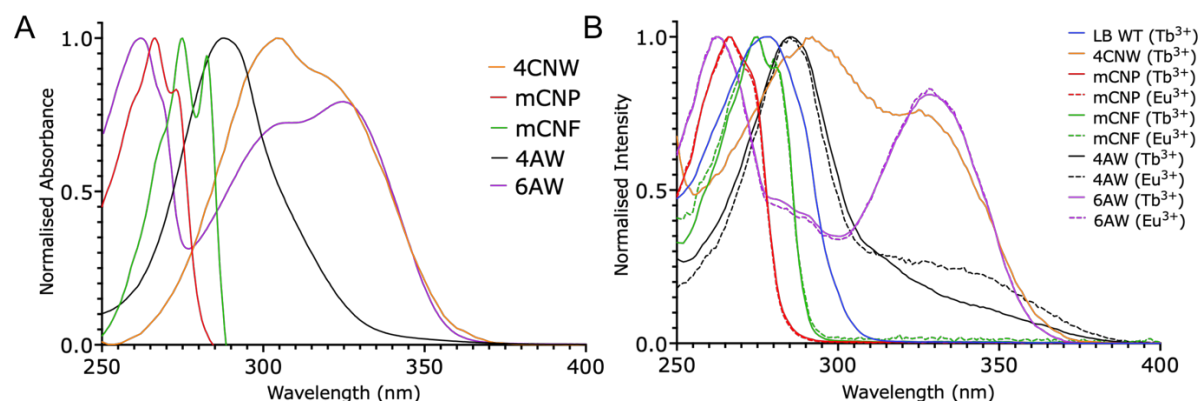


Figure 2.2 – Spectra of non-canonical amino acids. (A) UV-vis of 6AW, 4AW, mCNP, mCNF, 4CNW. Extinction coefficients were measured at 250 μ M amino acid concentration in phosphate buffered saline (PBS) using a 1 cm path length. Extinction coefficients in Appendix Table A.2. Each spectrum is normalised to the maximum absorbance. (B) Excitation scans of the LBT complexes with Tb^{3+} and Eu^{3+} . Emission at 545 nm or 615 nm was integrated for 5000 μ s after a delay of 100 μ s, using an excitation slit width of 5 nm and an emission slit width of 10 nm. Protein and lanthanide chloride were at 1 μ M and 2 μ M, respectively. The samples were prepared in 20 mM MOPS pH 7.0, 50 mM KCl. Each spectrum is normalised to the maximum intensity.

Following incorporation of different ncAAs into the LBT, the link between ncAA fluorescence and luminescence sensitisation was interrogated. Comparing the UV-vis spectra of the free ncAA and the excitation spectra of the ncAA incorporated into LBT with Tb^{3+} and Eu^{3+} clearly illustrates that the absorbance wavelengths of the antennae translate into the excitation wavelengths for the LBT- Ln^{3+} complexes (Figure 2.2). As the relative peak intensities within the emission spectra of both Tb^{3+} - and Eu^{3+} - complexes were consistent across excitation wavelengths, the intensity at the emission maximum wavelength was used as the single-wavelength comparator between samples, which corresponds to 545 nm for Tb^{3+} and 615 nm for Eu^{3+} . Using the integrated emission lifetimes allows for the comparison of luminescence intensities across different ncAA- Ln^{3+} pairs and considers the exclusion of waters from the inner-coordination sphere of the Ln^{3+} , which significantly alters the emissive lifetime. The apparent K_D gives an indication of the impact of ncAA incorporation on the complex's ability to bind the Ln^{3+} ion, and the titrations are shown in Appendix A.1.2

Unexpectedly, the highly fluorescent 4CNW (fluorescence $\phi > 0.8$ compared with ca. 0.15 in Trp)²⁶ only minimally enhanced the luminescence of Eu^{3+} and delivered weaker integrated Tb^{3+} emission intensities. In contrast, the relatively less fluorescent 4AW ($\phi \approx$

0.01) revealed brighter luminescence intensity with Tb^{3+} than tryptophan and excited Eu^{3+} . The best performance was achieved with 6AW, which features an increased fluorescence quantum yield.²⁴ Compared with Trp, 6AW enhanced the Tb^{3+} luminescence 50% more and excited the Eu^{3+} luminescence more effectively than any of the other antennae studied.

Table 2.1 - Luminescence emission intensities and apparent K_D s of the LBT-based Eu^{3+} and Tb^{3+} complexes

Antenna	λ_{ex} max (nm)	Quantum yield of ncAA f_{S1}	Extinction coefficient at λ_{ex} max ($\text{M}^{-1} \text{cm}^{-1}$)	Intensity Tb^b	Intensity Eu^a	Apparent K_D (Tb^{3+}) ^c (nM)	Apparent K_D (Eu^{3+}) ^c (nM)
Trp	280	0.14 ^d	5422	682	5	190 ± 40	N.D.
mCNP	267	–	1187	630	361	1300 ± 90	490 ± 30
mCNF	275	–	843	423	50	1400 ± 100	940 ± 100
4AW	285	0.01 ^d	6272	728	86	1100 ± 50	580 ± 30
4CNW	290	0.80 ^e	6041	323	5	1000 ± 40	N.D.
6AW	260	0.41 ^d	2108	1072	378	800 ± 20	830 ± 20

^a N.D. = not determined. ^b Integrated emission intensity for the complex with lanthanide ion measured at the excitation maximum and 545 nm for Tb^{3+} and 615 nm for Eu^{3+} with the same instrument settings (see Appendix A.1.3). Samples contained 500 nM Ln^{3+} ion with an excess of LBT in 20 mM MOPS pH 7.0, 50 mM KCl. Exact values are reported in Table A4. ^c Apparent K_D values determined by titrating Ln^{3+} (added as LnCl_3) into a solution of 500 nM protein in 20 mM MOPS pH 7.0, 50 mM KCl. The K_D values and error ranges were determined from a non-linear fit to the Hill equation for three technical replicates (see Appendix A.1.2). ^d From Abdelkader *et al.*,²⁴ measured in PBS. ^e From Qianzhu *et al.*,²⁶ measured in PBS.

Non-canonical antennae at position 10 in the LBT consistently reduced the binding affinity 2–8-fold (Table 2.1 and Figure 2.3). Given that the amino acid sequence of LBT was selected from a large peptide library using Tb^{3+} luminescence,⁸ it is unsurprising that even minor sidechain modifications can reduce the ion affinity. In contrast, acridone and carbostyryl sidechains installed in a similar LBT have been reported to perturb binding less (K_D = 34 nM vs. 19 nM for wild type).⁹ In these experiments, the apparent K_D values differed up to 2.5-fold between Tb^{3+} and Eu^{3+} , indicating extreme sensitivity of the lanthanide affinity towards optimal coordination geometry. While the nominal ionic radius differs by only 0.02 Å between Tb^{3+} and Eu^{3+} ,²⁷ the observed effect may result from the non-isotropic electron densities of lanthanide ions.^{28,29}

Substituting the antenna with the less fluorescent mCNP also greatly enhanced the Eu^{3+} luminescence, while the enhancement of the Tb^{3+} emission was minimally altered. This suggested that the pyridine nitrogen of the mCNP amino acid directly coordinates the lanthanide ion, minimising the distance between the antenna and the Ln^{3+} ion. To test this, mCNF, which exhibits a fluorescence intensity similar to mCNP but lacks the pyridine moiety, was incorporated into the LBT. mCNF still enhanced the luminescence of Eu^{3+} but less than mCNP, suggesting that coordination plays a significant role. The dissociation constants in Table 2.1 support this interpretation for Eu^{3+} , but not for Tb^{3+} . Nonetheless, mCNP lends itself to rapid chemical diversification using the biocompatible nitrile-aminothiol arcNAT click reaction to generate a range of different fluorophores,^{30,31} which invites future exploration.

The two most highly fluorescent antennae produced very different outcomes: 6AW was the most effective sensitiser for both Tb^{3+} and Eu^{3+} complexes, while 4CNW, despite its high intrinsic fluorescence, sensitised the luminescence of Tb^{3+} only modestly and hardly sensitised Eu^{3+} . Measuring the spectral overlap between the chromophore S_1 emission and Ln^{3+} absorption showed no simple correlation with emission intensity (Figure 2.3).

Coordination with hydration water efficiently quenches Ln^{3+} luminescence.³² As the luminescence lifetimes of the Eu^{3+} complexes were consistently 1.4 ms, the number of coordinated water molecules does not change between complexes with different ncAAs (Appendix Table A.1). The exception to this was LBT 4CNW, which showed a Eu^{3+} lifetime of 0.7 ms. In contrast, the Tb^{3+} luminescence lifetimes of the complexes with tryptophan analogues substituted at the 4-position (4CNW and 4AW) were significantly shorter than for the complex of the wild-type peptide or the LBT made with other ncAAs (~1.5 compared to ~2.7 ms), suggesting a variation in the number of hydration water molecules coordinating the Tb^{3+} ion. Chemical changes at position 4 of the Trp antenna thus appear to alter the coordination geometry of the Ln^{3+} ion.

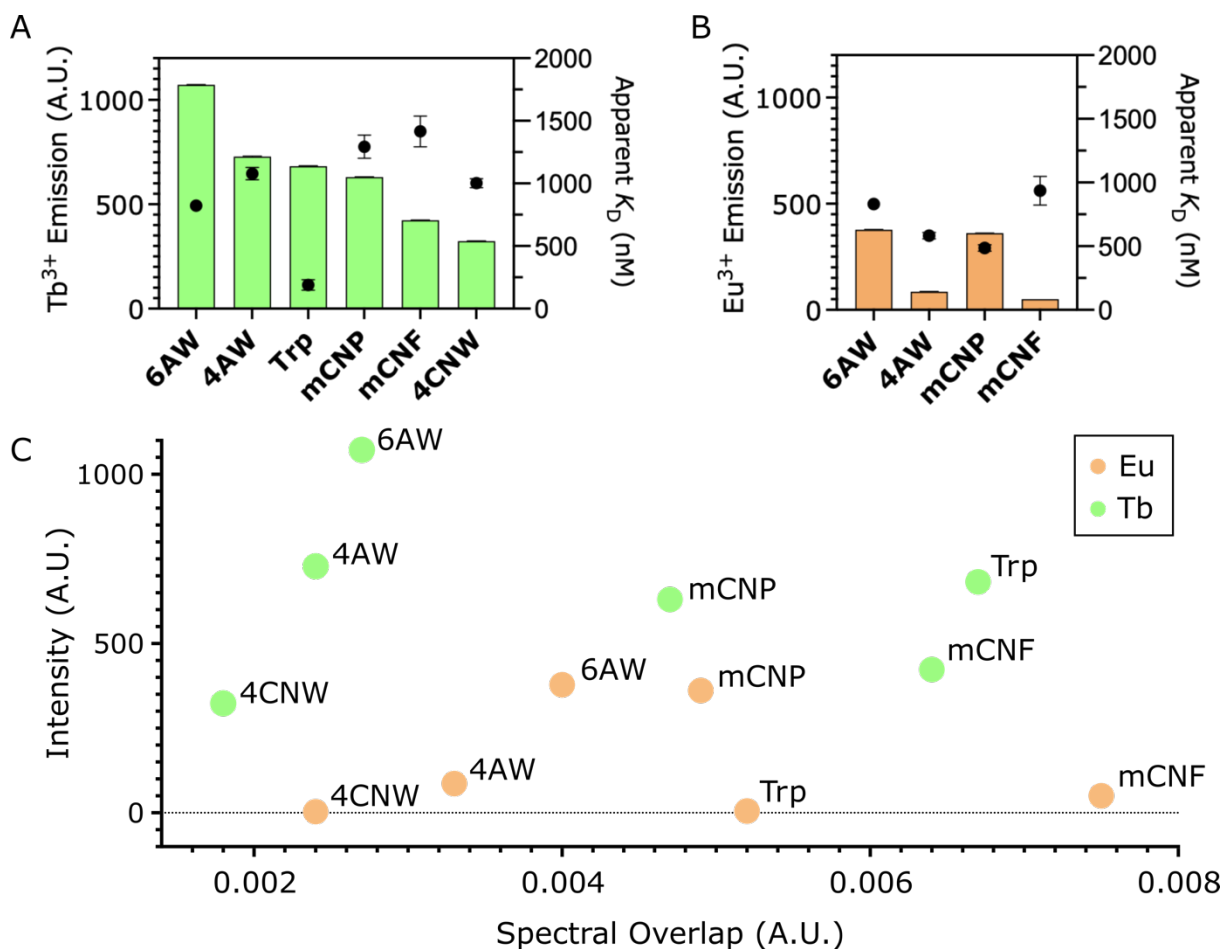


Figure 2.3 - Luminescence properties of LBT variants. The histograms show the integrated emission intensities following excitation at the absorption maximum, which was 265 nm for 6AW, 285 nm for 4AW, 280 nm for Trp, 267 nm for mCNP, 275 nm for mCNF and 290 nm for 4CNW. The apparent binding affinities for these complexes are indicated by black points which correlate to the right y-axis. (A) Tb³⁺ emission at 545 nm. The K_D value of the reference LBT is greater than reported,⁶ attributed to different buffer conditions. Error bars for apparent K_D measurements are standard deviations from three replicate experiments. (B) Eu³⁺ emission at 615 nm. The signal intensity of LBT and LBT 4CNW was too weak to measure the affinity. Error bars for apparent K_D measurements are standard deviations from three replicate experiments. (C) Plot of observed luminescence intensity versus spectral overlap for the series of LBT complexes. Spectral overlap was calculated as the integral of the scalar product between normalised absorption and emission spectra, with each spectrum normalised to unity.

Table 2.1 and Figure 2.3 show the absence of any simple relationship between the integrated emission intensity and the extinction coefficient or fluorescence quantum yield of the antenna. There was also no correlation with the S_1 -Ln³⁺ spectral overlap. These observations are consistent with a model in which T_1 energy, ISC efficiency, and local non-radiative decay pathways dominate sensitisation efficiency, and cannot be predicted from the performed measurements alone. This underscores the complexity of the antenna effect in a peptide context and its dependence not only on chromophore brightness but also geometric and electronic compatibility with the lanthanide. The

observation that the same antenna performed differently for Eu^{3+} and Tb^{3+} compromises any simple generalisation across lanthanides. Furthermore, all ncAAs reduced the binding affinity at least two-fold. This motivates the design of a robust protein scaffold that tolerates antenna modifications without trade-offs in lanthanide binding affinity.

2.3 Computational design of novel lanthanide-luminescent mini-proteins

Coordinates of the 12 residues of EF2 (DPDKDGTLDKE) were extracted from the NMR structure (PDB ID: 6MI5).³³ After mutating Thr7 to Trp, the Rosetta energy minimised peptide was used as seed structure input to RFDiffusion, diffusing a 70 or 50 amino acid protein with both N- and C-terminal extensions to the lanthanide ion binding motif. The generated backbone structures were used as input into ProteinMPNN, yielding eight sequences for each backbone. A Tb^{3+} ion was placed for coordination in the protein and threaded with the ProteinMPNN sequences in Rosetta using the *ref2015* score function.³⁴ The full details of the design process are in Appendix A.1.5.

Given the absence of explicit activity or binding considerations in RFDiffusion and ProteinMPNN, the scoring only accounted for a known binding geometry. Final structures were visually assessed, considering secondary structure diversity, solvent exposure of the Ln^{3+} ion, and plausible Ln^{3+} -O distances. Three proteins with unique topologies and the best scores across all evaluation criteria were chosen for experimental validation and expressed with an N-terminal His-tag and TEV cleavage site. For the rest of this chapter, these proteins are referred to as RF1, RF2, and RF3.

The three proteins have less than 35% sequence identity to lanmodulin. Using the AlphaFold3 webserver,³⁵ each sequence predicted high-confidence structures with a calcium ion bound at the desired site. Each protein was recombinantly expressed in *Escherichia coli* (*E. coli*), using IPTG induction of the gene under a T7 promoter as described above. RF1, RF2, and RF3 were isolated by immobilised metal ion chromatography (IMAC), with yields of 10 mg, 400 mg, and 15 mg per 1 L of cell culture, respectively. RF1 and RF3 not only expressed in much lower yields but were also more

prone to precipitation at room temperature compared with the highly expressing RF2 protein.

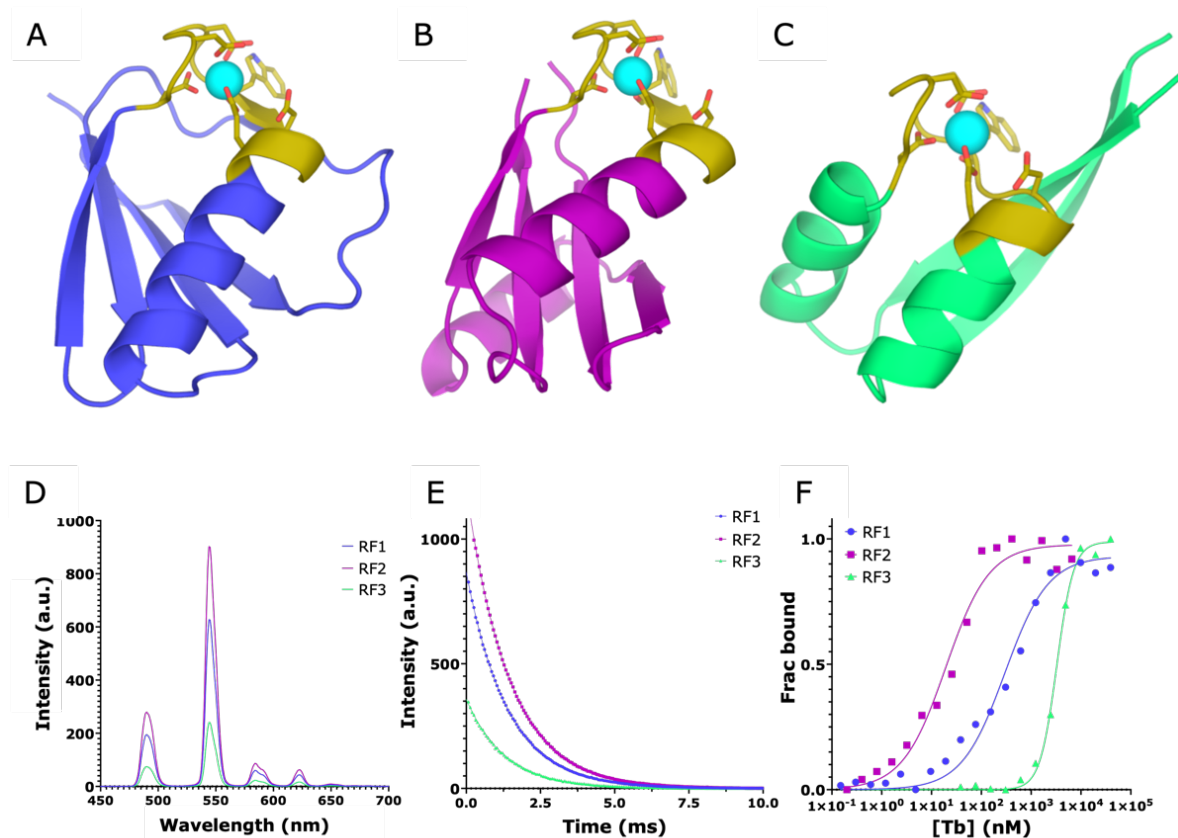


Figure 2.4 - Structure and characterisation of lanthanide-binding mini-proteins RF1-3. (A-C) AlphaFold3 models of RF1 (dark blue), RF2 (purple), and RF3 (green). The EF-hand motif is coloured gold with coordinating residues shown in stick representation. Tb³⁺ ions shown as cyan spheres. (D) Emission spectra of RF1-3 (1 μ M protein, 2 μ M TbCl₃, 20 mM MOPS pH 7.0, 50 mM KCl, λ_{ex} = 280 nm, 100 μ s delay, 1 ms integration time). (E) Luminescence decay of RF1-3 (1 μ M protein, 2 μ M TbCl₃, 20 mM MOPS pH 7.0, 50 mM KCl). Lifetimes fitted to single exponential decay (Appendix A.1.3): RF1 τ = 1.39 ms, RF2 τ = 1.45 ms, RF3 τ = 1.30 ms. (F) Apparent K_D determination by titrating TbCl₃ into 500 nM protein in 20 mM MOPS pH 7.0, 50 mM KCl. Points are averages of three technical replicates. K_D values determined from non-linear fit to Hill equation (Appendix A.1.2): RF2 K_D = 19 $\pm$ 3 nM, RF1 K_D = 320 $\pm$ 50 nM, RF3 K_D = 3300 $\pm$ 130 nM.

All three designed proteins (RF1, RF2 and RF3) produced a significant luminescence response when incubated with Tb³⁺ (Figure 2.4). Relative to the reference LBT, the luminescence intensities for RF1 and RF2 represented a 1.8-fold and 2.5-fold enhancement, respectively. In contrast, RF3 showed a diminished intensity of about 70 % compared to LBT. The Tb³⁺ affinities followed a similar trend, with RF2 showing the strongest binding affinity at K_D = 19 $\pm$ 3 nM. RF2 also had the longest luminescence lifetime (1.45 ms). All were above 1.2 ms, suggesting effective exclusion of water from the inner coordination sphere of the Tb³⁺ complex.³² However, the values are shorter than 2.5

ms, which is the typical lifetime of a fully shielded Tb^{3+} complex, implying water is not completely excluded. In summary, all data pointed to RF2 as the most attractive variant, and it was selected for further study.

Figure 2.5A and B show the AlphaFold3 model of RF2 with a Ca^{2+} ion (Figure 3A). The protein contains two α -helices and one β -sheet comprised of four β -strands. The EF-hand is positioned between β -strand 3 and α -helix 1. Trp42 is placed in the antenna position with the backbone carbonyl coordinated to the ion. A second aromatic amino acid, Tyr5, almost contacts Trp42 (Figure 3B). The presence of Tyr5 may be beneficial, as Tyr excited at 280 nm can also excite lanthanide luminescence. In principle, the more distant Trp50 may also enhance the antenna effect. These two amino acids introduced by the sequence design of ProteinMPNN may help to produce a stable protein and were therefore retained.

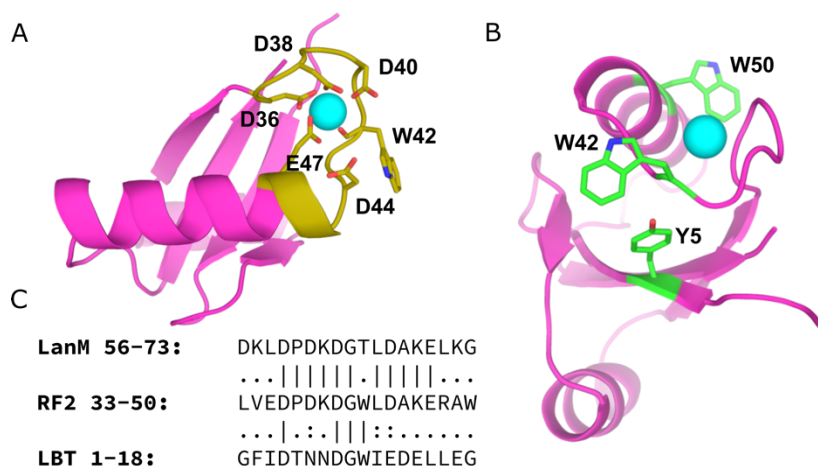


Figure 2.5 - AlphaFold3 model of RF2. (A) RF2 model (magenta) with the EF-hand highlighted in gold, the Tb^{3+} ion in cyan, and the Tb^{3+} coordinating residues shown in stick representation. (B) Same as (A), but showing RF2 in a different orientation and showing the aromatic residues (green) instead of highlighting the EF hand. (C) Pairwise sequence alignment of lanmodulin EF2, RF2, and LBT with their flanking residues.

Using 2D and 3D NMR experiments recorded on an 800 MHz spectrometer, the backbone chemical shifts (C , C_α , C_β , N , H , H_α , H_β) of most of the 71 amino acids were assigned in the protein complexed with yttrium (Y^{3+} ; Appendix Figures S12 and S13). Resonances of residues 10–14 could not be assigned which suggests that this loop may be too dynamic to produce distinct backbone resonances. As these residues are located at the termini of the β -strands distal to the lanthanide-binding region, their conformational dynamics are

not expected to perturb lanthanide coordination or antenna geometry. The backbone dihedral angles ϕ and ψ derived from the chemical shifts by the program TALOS+³⁶ closely aligned with the AlphaFold3 model (Figure 2.6). The region of the EF-hand (residues 34–49) makes expected inter-residual contacts between backbone protons. While these NMR experiments do not fully assign a 3-dimensional structure to the novel protein, they confirm that the structural prediction for a lanthanide-binding protein using Ca^{2+} as a proxy for Ln^{3+} correlates with experiments.

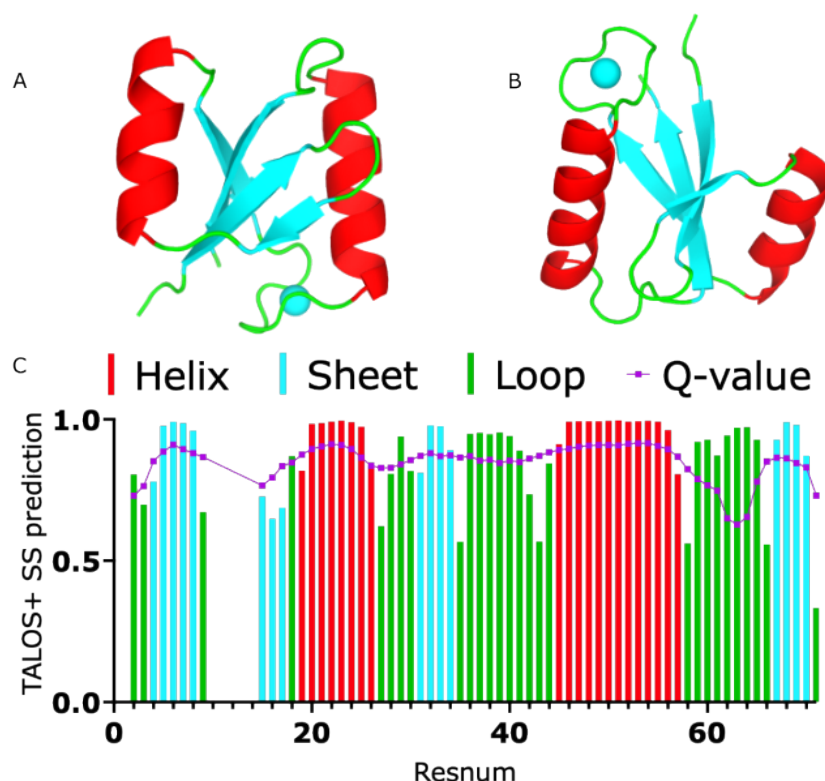


Figure 2.6 - Comparison of the secondary structure of the AlphaFold3 model and the prediction made from the assigned chemical shifts. (A) and (B) Two different orientations of the AlphaFold3 model of RF2 generated with a Ca^{2+} ion (cyan). The cartoon representation is coloured by secondary structure: red-helix, cyan-sheet, green-loop. (C) Output from the TALOS+ webserver.³⁶ Coloured bars indicate the secondary structure prediction associated with the residue number. The Q-value (purple) indicates the confidence for that prediction.

2.4 Enhanced brightness and more robust binding affinity with an ncAA antenna

To test whether the enhanced sensitisation observed for LBT 6AW was generalisable to another protein, Trp42 in RF2 was substituted with 6AW. Replacing Trp42 with 6AW yielded 56 mg RF2 6AW per 1 L of cell culture. Unlike LBT 6AW, RF2 6AW ($\lambda_{\text{ex}} = 260 \text{ nm}$)

showed no increase in Tb^{3+} luminescence relative to unmodified RF2 ($\lambda_{\text{ex}} = 280 \text{ nm}$). This may be because RF2- Tb^{3+} features additional sensitisation by Tyr5 and possibly Trp50, which are excited to a lesser extent at the absorption maximum of 6AW. In contrast, RF2 6AW showed a significant enhancement of Eu^{3+} luminescence, with the excitation spectrum showing increased excitation around 280 nm, relative to LBT 6AW (Figure 2.7). As the emission maxima for Tyr and Trp are at 304 nm and 350 nm, respectively, and 6AW possesses a second broad excitation band between 300 nm and 360 nm (Figure 2.2), it is possible that a non-radiative energy transfer from the nearby Tyr and Trp residues to 6AW can add to the enhancement of the Eu^{3+} emission. A mutant with 6AW in position 50 and Trp retained at position 42 showed a decrease in Tb^{3+} excitation, and a minimal increase in Eu^{3+} excitation, suggesting that there is some sensitisation from position 50 (Table 2.2).

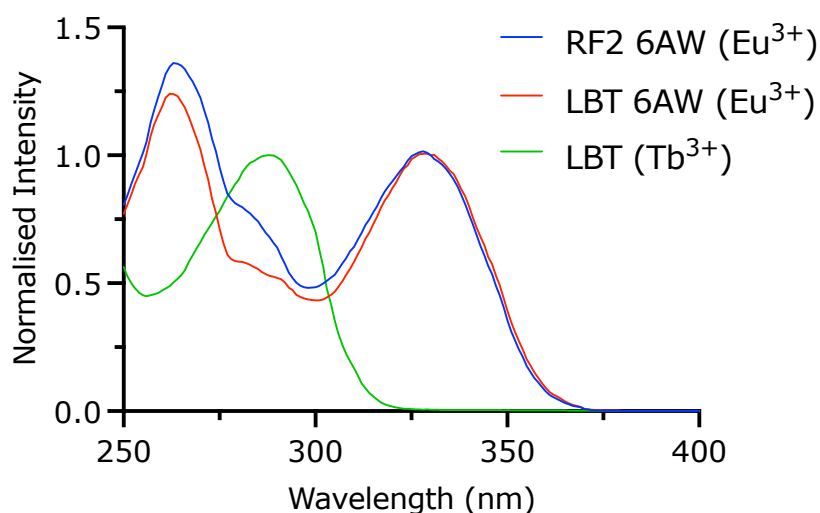


Figure 2.7 - Comparison of excitation spectra for RF2 6AW/ Eu^{3+} , LBT 6AW/ Eu^{3+} and LBT/ Tb^{3+} . Emission at 545 nm or 615 nm was integrated for 1000 μs after a delay of 100 μs . Protein and LnCl_3 were at 1 μM and 2 μM , respectively. RF2 6AW/ Eu^{3+} (blue) and LBT 6AW/ Eu^{3+} have been normalised to the intensity of 330 nm, where the contribution of canonical amino acids to sensitisation is negligible. LBT/ Tb^{3+} has been normalised to the intensity at 280 nm.

The changes in binding affinity of RF2 6AW with Tb^{3+} and Eu^{3+} compared to RF2 showed that the ncAA affects the Ln^{3+} binding affinity by less than an order of magnitude (Table 2.2). 6AW at position 50 decreases the binding affinity only 2-fold, whereas 6AW at position 42 decreases the affinity almost 10-fold. As in the case of the LBT, the binding affinities differed between Eu^{3+} and Tb^{3+} (Table 2.2). Nonetheless, even the reduced binding affinities are in the nanomolar range, making the RF2 6AW construct suitable for a wide range of sensing applications.

Table 2.2 - Luminescence emission intensities and apparent K_D s of the LBT-based Eu^{3+} and Tb^{3+} complexes^a

Positions	Intensity Tb^b	Intensity Eu^b	K_D (Tb^{3+}) ^c	K_D (Eu^{3+}) ^c
Trp42/Trp50	1784	<1	19 ± 3 nM	N.D. ^a
Trp42/6AW50	1507	13	37 ± 5 nM	N.D. ^a
6AW42/Trp50	1700	506	180 ± 12 nM	150 ± 12 nM

^a N.D = not determined. ^b Integrated emission intensity for complex with lanthanide ion measured at the excitation maximum for the antenna and observed at emission wavelengths of 545 nm for Tb^{3+} and 615 nm for Eu^{3+} . Samples contained 500 nM Ln^{3+} ion with an excess of RF2. Variants were excited at their excitation maxima, 280 nm for Trp42/Trp50 and Trp42/6AW50, and 260 nm for 6AW42/Trp50. ^c Apparent K_D were measured by titration of Ln^{3+} into a solution of 500 nM of protein in 20 mM MOPS pH 7.0, 50 mM KCl. K_D values and error ranges were determined from a non-linear fit to three technical replicates.

Considering the excellent enhancement of Eu^{3+} luminescence of RF2 6AW and its nanomolar binding affinities to Eu^{3+} and Tb^{3+} , emission spectra with other trivalent lanthanide ions were recorded, including cerium, praseodymium, neodymium, samarium (Sm^{3+}), gadolinium, dysprosium (Dy^{3+}), erbium, thulium, and ytterbium. There were weak but significant emission peaks with Sm^{3+} and Dy^{3+} , which overlapped with the long-wavelength end of the emission peak from 6AW (Appendix Figure A.24). While the Sm^{3+} emission was too weak for affinity titrations, the K_D value for Dy^{3+} (Appendix Figure A.14) as 290 ± 30 nM, could be determined which corresponds to an only slightly weaker affinity than for Eu^{3+} or Tb^{3+} .

The number of water molecules (q) coordinated to the lanthanide ions can be determined using luminescence lifetime measurements in D_2O and the modified Horrocks equation (Appendix A.1.8).³² Empirical correlation coefficients of luminescence quenching by H_2O have been published for Tb^{3+} and Eu^{3+} ,³² as well as for Dy^{3+} and Sm^{3+} .³⁷ All four D_2O titrations showed a clear linear response to the percentage of D_2O (Figure 2.8), yielding q -values of 1.07, 1.06, 1.30, and 1.31 for Tb^{3+} , Eu^{3+} , Sm^{3+} , and Dy^{3+} , respectively. In view of the inherent uncertainty of this equation of ± 0.3 for Tb^{3+} and Eu^{3+} , and ± 0.5 for Dy^{3+} and Sm^{3+} ,³⁷ this indicates that the Ln^{3+} ion in the RF2 complex consistently coordinates a single water molecule. The q -values near 1 across Tb^{3+} , Eu^{3+} , Sm^{3+} , and Dy^{3+} indicate similar inner-sphere hydration. Brightness and apparent affinity therefore reflect antenna identity and its local geometry, rather than changes in its hydration state. Thus, a single EF-hand scaffold bearing 6AW at the primary antenna position supports $q \approx 1$ and

measurable emission for Eu^{3+} , Tb^{3+} , Sm^{3+} and Dy^{3+} , indicating the same protein framework can be used to interrogate multiple lanthanides without re-optimising the binding site.

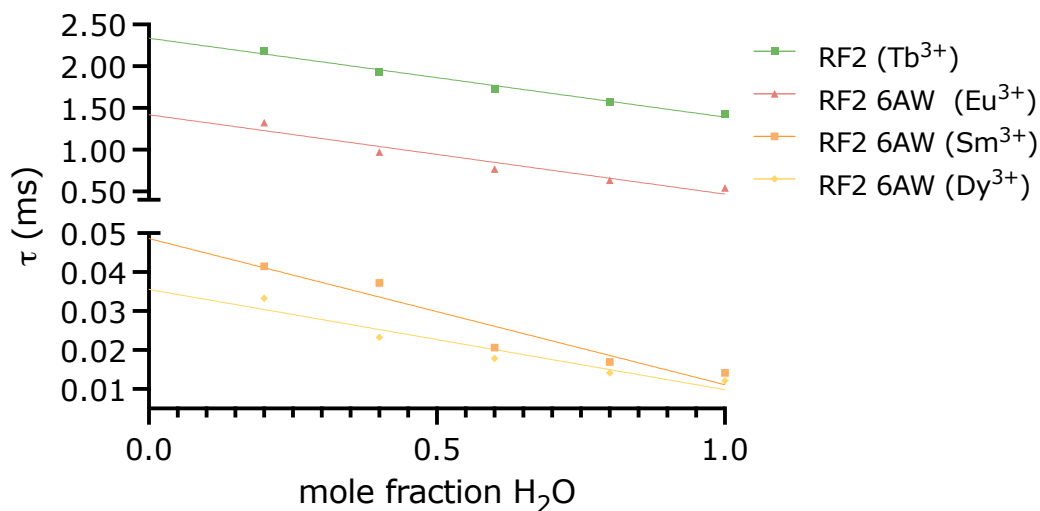


Figure 2.8 - D_2O titrations of RF2/ Ln^{3+} complexes. Emission at 545 nm or 615 nm was integrated for 1000 μs after a delay of 100 μs . Protein and LnCl_3 were at 1 μM and 2 μM , respectively. RF2 6AW/ Eu^{3+} (blue) and LBT 6AW/ Eu^{3+} have been normalised to the intensity of 330 nm, where the contribution of canonical amino acids to sensitisation is negligible. LBT/ Tb^{3+} has been normalised to the intensity at 280 nm.

Two recent studies illustrate alternative strategies in lanthanide-luminescent protein design. A 2023 study grafted a Tb^{3+} -binding EF-hand into a *de novo* immunoglobulin-like scaffold using ProteinMPNN, yielding a Tb^{3+} affinity of 8–12 μM by time-resolved luminescence.³⁸ This establishes a *de novo* lanthanide-luminescent protein, but the moderate affinity limits its use under dilute or competitive conditions. More recent work focused on antenna modification, identifying acridone as a strong Eu^{3+} antenna if installed by GCE at three EF-hand positions of LanM.³⁹ Besides involving multiple antenna sites and four additional mutations, the construct led to Ln^{3+} stoichiometries greater than 1:1 of Ln^{3+} :protein, which compromises the analysis of antenna performance and lanthanide binding.

2.4.1 New lanthanide-luminescent mini-protein shows remarkable selectivity and *in vivo* sensitisation

Lanthanide ions have high charge density and bind promiscuously to anionic cellular components, including phospholipids,⁴⁰ nucleic acids,⁴¹ metal-chelating proteins,⁴² and

organic phosphates.⁴³ LanM has been used as an *in vivo* reporter for Ln^{3+} cellular uptake using a Förster resonance energy transfer (FRET) pair of fluorescent proteins fused to its N- and C-termini.⁴⁴ This approach relies on the fluorescent protein FRET, not the antenna effect. It has been reported that intracellular Y^{3+} reaches 3.4 mM in *E. coli* cells in the presence of 20 mM extracellular Y^{3+} , measured by inductively coupled plasma-mass spectrometry.⁴⁵ However, a more recent report illustrates that *E. coli* cells suffer membrane damage in millimolar concentrations,⁴⁶ suggesting this value may reflect the concentration of membrane-bound Y^{3+} in samples of lysed cells. Although short incubation periods (<30 minutes) and isotonic buffer washes were employed, luminescence contributions from cytosolic components released by lysed cells cannot be excluded, meaning the observed signal may not reflect exclusively intracellular Ln^{3+} concentration. In addition, nonspecific binding of Ln^{3+} to DNA produces UV-excited luminescence,⁴⁷ confounding the luminescence readout. In the face of these challenges, the RF2 scaffold for maintaining its luminescence signal in the face of competing cellular ligands and metal ions.

To investigate RF2 as a general intracellular lanthanide sensor, the decrease in observable luminescence was measured in the presence of a competitor.⁴⁸ Specifically, the binding of various metal ions was evaluated by measuring the change in the sensitised luminescence of a solution containing the RF2/ Tb^{3+} complex after the addition of an equimolar amount of the competing metal ion. The remaining fraction of initial Tb^{3+} luminescence was used to determine the relative affinity of the competitor, with a fraction of ca. 0.5 indicating a binding affinity comparable to that of Tb^{3+} , a fraction <0.5 indicating a stronger affinity, and a fraction >0.5 weaker binding.

The results for competing lanthanide ions were consistent with expectations based on previous titration experiments with RF2 6AW. The addition of Eu^{3+} resulted in a luminescence fraction of <0.5, while Dy^{3+} yielded a fraction >0.5 (Figure 2.9). Along with many of the Ln^{3+} metals, Ca^{2+} and Mg^{2+} , which are hard cations like Ln^{3+} ions, and Zn^{2+} and Co^{2+} , which typically prefer coordination environments with softer Lewis bases, were also tested. None of the divalent metal ions decreased the Tb^{3+} luminescence, demonstrating their inability to compete for the Ln^{3+} binding site. Co^{2+} and Zn^{2+} increased the emission

intensity (Figure 2.9), which may indicate the presence of a second metal binding site in RF2 that promotes the luminescence brightness by an allosteric effect. Defining this site and its structural basis lies beyond the scope of this thesis but represents an avenue for future work. The strong preference for Ln^{3+} ions relative to the M^{2+} ions may be due to the difference in charge density or due to the subtle non-spherical nature of the Ln^{3+} ions fitting the designed pocket better.

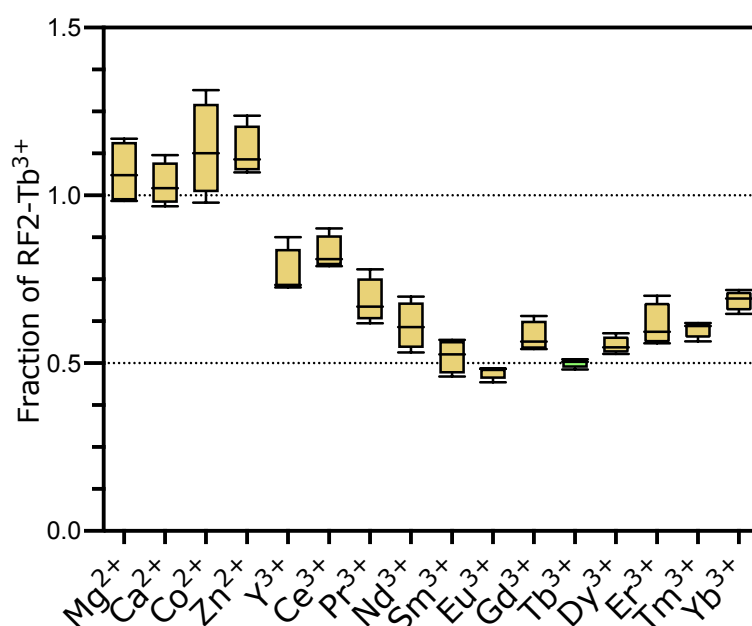


Figure 2.9 – Competition assay with RF2/ Tb^{3+} . Fractional change in the luminescence intensity (gold) of RF2 (1 μM) samples with Tb^{3+} (2 μM) and competitor (2 μM) in 20 mM MOPS pH 7.0, 50 mM KCl. This indicates relative binding tendencies but does not yield apparent K_D values for the competitors. Each data point represents the ratio of integrated emission intensity (collected at 545 nm emission with a 100 μs delay and 1 ms integration window, following excitation at 280 nm) normalised to the intensity of a sample containing only Tb^{3+} . The green box indicates 50% of the intensity of the RF2/ Tb^{3+} complex. Boxes indicate the mean and range of four technical replicates.

The competition assay gives rise to a parabolic trend in affinity across the lanthanide series, which is strongest for Eu^{3+} and Tb^{3+} , with diminishing affinities towards both the heavier and lighter elements (Figure 2.9). The parabolic dependence on lanthanide identity likely reflects the interplay between lanthanide contraction and the fixed EF-hand geometry. A similar curve has been reported for LBT,⁷ and a deviation for lighter rare-earth elements has also been reported for lanmodulin.⁴⁸ The sensitivity to lanthanide identity afforded by the RF2 mini-protein suggests the potential to design proteins with refined selectivity. Small changes in EF-hand geometry could be exploited to tune preference for specific Ln^{3+} ions for separation and sensing applications.

In view of the selectivity of RF2 for Ln^{3+} over metal ion competitors found in biological matrices, RF2 was then tested for its utility in detecting Ln^{3+} ions in cultures of *E. coli* cells. A culture of *E. coli* B95 cells was transformed with the dual-plasmid system. One plasmid encoded either the reference LBT, amber-LBT, or RF2 gene with an amber stop codon at either position 42 or 50. The second carried the orthogonal tRNA/tRNA synthetase pair for the incorporation of 6AW. The cells were grown to OD_{600} of 0.6 and protein expression was induced with 1 mM IPTG. After overnight expression, cells were resuspended in a solution of 150 mM NaCl containing 20 mM Ln^{3+} , and measured at pH ~6.0. High Ln^{3+} concentrations (20 mM) and unbuffered media are non-physiological, but this stringent proof-of-principle demonstrates clean Ln^{3+} detection in a complex mixture. Future applications will require optimisation for lower Ln^{3+} concentrations in buffered media.

The presence of the Tb^{3+} - and Eu^{3+} -complexes in the cells was readily detected (Figure 2.10). Excitation at 280 nm and 330 nm corresponds to the excitation maximum for tryptophan and the second excitation peak for 6AW, respectively. The luminescence of native cells with Eu^{3+} and Tb^{3+} excited at 330 nm featured a notably high background fluorescence due to the lanthanides binding to components in the cytosol that also sensitise their luminescence. Nonetheless, excitation at 280 nm produced strong sensitisation in RF2 6AW, exceeding the signal of the other variants ten-fold (Figure 2.10). This high signal-to-background enables single-wavelength, plate-based measurements in whole cells to detect the presence of Eu^{3+} in cell cultures. RF2 6AW is a promising starting point for genetically encoded lanthanide sensors in cells. Practical imaging or sensing will require balancing of metal dosage, toxicity and background luminescence.

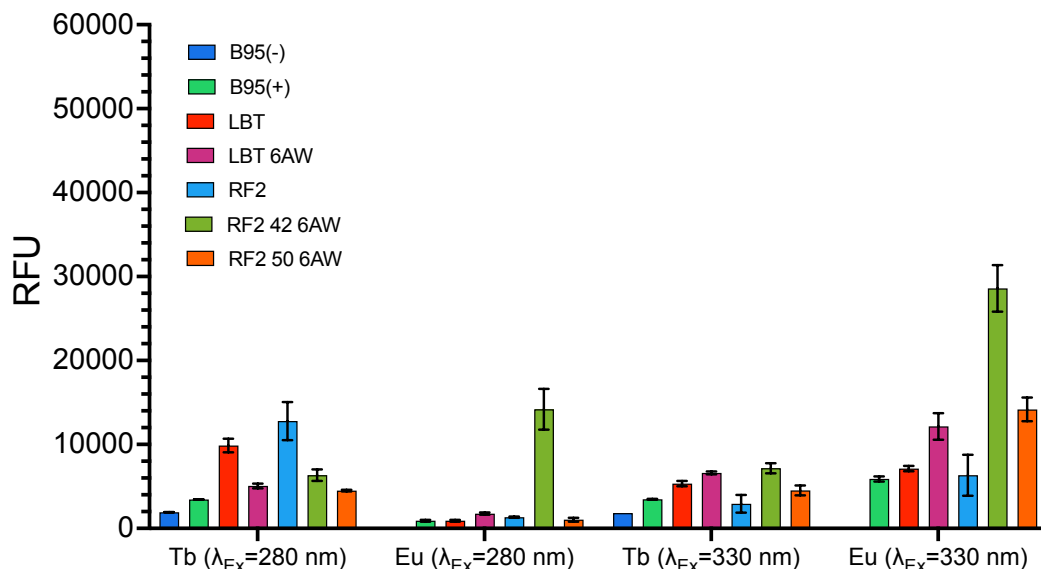


Figure 2.10 – Sensitisation of lanthanide luminescence in B95 cell cultures. Untransformed B95 cells with and without 6AW (B95(+), B95(-)), transformed B95 cells with a WT LBT or RF2 plasmid, or the two-plasmid expression system for installing 6AW were grown to $OD_{600} = 0.6$ – 0.8 in LB media and supplemented with 1 mM IPTG. Cultures with cells containing the G1(6AW)RS plasmid, as well as the B95(+) cells were supplemented with 1 mM 6AW. Cells were cultured overnight at RT and resuspended to $OD_{600} = 1.0$ in 150 mM NaCl. The standardised cell resuspensions were then resuspended again in 150 mM NaCl containing 20 mM $EuCl_3$ or 20 mM $TbCl_3$. Readings are integrated emission intensities over 1000 μs following a 600 μs delay. Emission intensities were $\lambda_{Em} = 545$ nm for Tb^{3+} , and $\lambda_{Em} = 615$ nm for Eu^{3+} . Datasets for Eu^{3+} - and Tb^{3+} -incubated cells were recorded with different gain settings for the detector. Therefore, the values between them are not directly comparable.

2.5 Conclusion

This chapter presents methods for designing genetically encoded, bright, lanthanide-luminescent complexes in *E. coli* cells, leveraging GCE using selected aaRS:tRNA pairs. The investigation of various antenna amino acids showed that an empirical approach is required for selecting good antennae for a range of lanthanides, because sensitisation efficiency cannot be reliably predicted from the photophysical properties of the sensitising moiety alone.

The computational designs yielded three proteins that all bind and excite Tb^{3+} ions, two of which outperform the reference LBT as a sensitising complex. The lead mini-protein, RF2, has high expression levels, strong and specific binding to lanthanides, as well as enhanced luminosity. These peptides and proteins can be recombinantly produced and excited by a range of wavelengths (260–350 nm), and retain their binding behaviours in the complex environment of a cell culture.

These results establish several practical design considerations for biosynthetic lanthanide-luminescent complexes. First, antenna performance cannot be predicted from extinction coefficients, fluorescence quantum yields, or spectral overlap. Empirical screening via GCE is necessary. Second, *de novo* EF-hand scaffolds like RF2 can be designed using RFdiffusion, which achieve low-nanomolar Tb^{3+} affinity, high specificity and $q \approx 1$, providing a robust framework for antenna modification. Third, 6AW, the optimal antenna identified in LBT, generalises to novel proteins. RF2 6AW retains Tb^{3+} and Eu^{3+} sensing in *E. coli* cultures, demonstrating potential for *in vivo* lanthanide sensitisation.

2.6 Methods

All buffers used for lanthanide luminescence measurements were treated with Chelex 100 resin (BioRad). Metal stocks were made in 10 mL volumetric flasks using water treated with Chelex 100 resin, and a small amount of concentrated HCl to prevent hydroxide formation. Buffers used for experiments with lanthanides were also treated with Chelex 100 resin. Concentrations of rare earth element stocks were validated using complexometric titration with EDTA and Xylenol Orange (Sigma). For experiments with protein- Ln^{3+} complexes, buffer conditions were used that have no or minimal affinity to Ln^{3+} ions.⁴⁹

2.6.1 Expression and purification of LBT and EF2 proteins

The gene for the fusion protein was ordered in a low-copy number pCDF plasmid under the control of a T7 promoter, either as the wild-type sequence (LBT-Trp) or with the tryptophan codon as an amber codon (LBT-Amb). The genes for the EF2 variants RF1, RF2, and RF3, as well as the RF2 amber mutants (RF2-Amb), were also cloned into the pCDF vector under control of the T7 promoter. High-copy number pRSF plasmids were used to encode the genes of the G1 tRNA synthetase mutants under the T7 promoter and the paired G1-tRNA under the *proK* promoter. This two-plasmid system has previously been established as an optimised system²² to genetically encode ncAAs in *E. coli* B-95. $\Delta\text{A}\Delta\text{fabR}$ cells.⁵⁰ For LBT-Amb, and RF2-Amb, the two plasmids were transformed into cells concurrently. For LBT-Trp, RF1, RF2 and RF3, the pRSF plasmid was omitted.

Transformed cells were grown in LB media supplemented with 1 mM ncAA at 37 °C, and protein expression was induced with 1 mM IPTG at an OD₆₀₀ of 0.6. After 16–18 hours of expression, the cells were concentrated using centrifugation, and resuspended in binding buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl). Resuspended cells were then subjected to sonication, and fusion protein was isolated from the cells using immobilised metal-ion affinity chromatography (IMAC) using imidazole as the eluent (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 500 mM imidazole). This typically yielded 20–30 mg L⁻¹ of fusion protein. Incorporation of the ncAA was confirmed by mass spectrometry (see Appendix A.1.1). The fusion protein was then concentrated to 1 mL in TEV buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl) and cleaved at 4 °C overnight after the addition of TEV protease at a 50:1 ratio of protein:TEV and 10 mM DTT. The resultant LBT-Amb fusion proteins were then isolated via size-exclusion chromatography (SEC) in volatile buffer (200 mM ammonium bicarbonate pH 8.5). Fractions containing the peptide were then isolated and lyophilised before luminescence analysis. For the EF2 proteins, IMAC elution fractions were dialyzed into TEV buffer and cleaved at RT overnight after the addition of TEV protease at a 1:1 ratio of protein:TEV and 2 mM β-mercaptoethanol. For RF2 6AW, incorporation of the ncAA was confirmed by mass spectrometry (see Appendix A.1.1). The EF2 proteins were then isolated by SEC in buffer D.

¹⁵N/¹³C-labeled RF2 was produced in a bioreactor (Infors-HT, Switzerland) at 300 mL media volume using a protocol by Klopp *et al.*⁵¹ The protein was isolated as above with a yield of 100 mg purified protein per litre cell culture.

2.6.2 Luminescence measurements

Luminescence measurements were performed on a Cary Eclipse fluorescence spectrophotometer (Agilent, USA) equipped with a phosphorescence module. Precise instrument settings for fluorescence measurements are indicated with the associated spectra. Time-resolved emission and excitation spectra were recorded with a 100 μs lag time, followed by a 1 ms integration time.

2.7 References

1. J.-C. G. Bünzli, *Coord. Chem. Rev.*, 2015, **293**, 19–47.
2. J.-C. G. Bünzli, *Chem. Rev.*, 2010, **110**, 2729–2755.
3. S. E. Bodman and S. J. Butler, *Chem. Sci.*, 2021, **12**, 2716–2734.
4. J. Bruno, W. D. Horrocks, Jr. and R. J. Zauhar, *Biochemistry*, 1992, **31**, 7016–7026.
5. P. Mukherjee, C. M. Shade, A. M. Yingling, D. N. Lamont, D. H. Waldeck and S. Petoud, *J. Phys. Chem. A*, 2011, **115**, 4031–4041.
6. M. Nitz, K. J. Franz, R. L. Maglathlin and B. Imperiali, *ChemBioChem*, 2003, **4**, 272–276.
7. M. Nitz, M. Sherawat, K. J. Franz, E. Peisach, K. N. Allen and B. Imperiali, *Angew. Chem. Int. Ed.*, 2004, **43**, 3682–3685.
8. L. L. Martin, B. R. Sculimbrene, M. Nitz and B. Imperiali, *QSAR Comb. Sci.*, 2005, **24**, 1149–1157.
9. A. M. Reynolds, B. R. Sculimbrene and B. Imperiali, *Bioconjugate Chem.*, 2008, **19**, 588–591.
10. B. R. Sculimbrene and B. Imperiali, *J. Am. Chem. Soc.*, 2006, **128**, 7346–7352.
11. K. Barthelmes, A. M. Reynolds, E. Peisach, H. R. Jonker, N. J. DeNunzio, K. N. Allen, B. Imperiali and H. Schwalbe, *J. Am. Chem. Soc.*, 2011, **133**, 808–819.
12. T. W. Victor, K. H. O'Toole, L. M. Easthon, M. Ge, R. J. Smith, X. Huang, H. Yan, Y. S. Chu, S. Chen, D. Gursoy, M. Ralle, B. Imperiali, K. N. Allen and L. M. Miller, *J. Am. Chem. Soc.*, 2020, **142**, 2145–2149.
13. J. A. Cotruvo, Jr., E. R. Featherston, J. A. Mattocks, J. V. Ho and T. N. Laremore, *J. Am. Chem. Soc.*, 2018, **140**, 15056–15061.
14. E. R. Featherston, E. J. Issertell and J. A. Cotruvo, *J. Am. Chem. Soc.*, 2021, **143**, 14287–14299.
15. R. Sánchez-Fernández, I. Obregón-Gómez, A. Sarmiento, M. E. Vázquez and E. Pazos, *Chem. Commun.*, 2024, **60**, 12650–12661.
16. C. S. Bonnet, M. Devocelle and T. Gunnlaugsson, *Org. Biomol. Chem.*, 2012, **10**, 126–133.
17. L. Ancel, C. Gateau, C. Lebrun and P. Delangle, *Inorg. Chem.*, 2013, **52**, 552–554.
18. L. Ancel, A. Niedzwiecka, C. Lebrun, C. Gateau and P. Delangle, *C. R. Chim.*, 2013, **16**, 515–523.
19. Y.-H. Yeung, H.-F. Chau, H.-Y. Kai, W. Zhou, K. H.-Y. Chan, W. Thor, L. J. Charbonnière, F. Zhang, Y. Fan, Y. Wu and K.-L. Wong, *Adv. Opt. Mater.*, 2024, **12**, 2302070.
20. L. Wang, A. Brock, B. Herberich and P. G. Schultz, *Science*, 2001, **292**, 498–500.
21. E. H. Abdelkader, H. Qianzhu, T. Huber and G. Otting, *ACS Sens.*, 2023, **8**, 4402–4406.
22. E. H. Abdelkader, H. Qianzhu, Y. J. Tan, L. A. Adams, T. Huber and G. Otting, *J. Am. Chem. Soc.*, 2021, **143**, 1133–1143.
23. H. Qianzhu, E. H. Abdelkader, G. Otting and T. Huber, *J. Am. Chem. Soc.*, 2024, **146**, 13641–13650.
24. E. H. Abdelkader, H. Qianzhu, T. Huber and G. Otting, *submitted for publication*.
25. E. H. Abdelkader and G. Otting, *J. Biotechnol.*, 2021, **325**, 145–151.
26. H. Qianzhu, E. H. Abdelkader, A. P. Welegedara, E. Habel, N. Paul, R. L. Frkic, C. J. Jackson, T. Huber and G. Otting, *Angew. Chem. Int. Ed.*, 2025, **64**, e202421000.
27. R. D. Shannon, *Acta Crystallogr., Sect. A*, 1976, **32**, 751–767.
28. L. Valencia, J. Martinez, A. Macias, R. Bastida, R. A. Carvalho and C. F. Geraldès, *Inorg. Chem.*, 2002, **41**, 5300–5312.
29. S.-D. Jiang and S.-X. Qin, *Inorg. Chem. Front.*, 2015, **2**, 613–619.
30. E. H. Abdelkader, H. Qianzhu, J. George, R. L. Frkic, C. J. Jackson, C. Nitsche, G. Otting and T. Huber, *Angew. Chem. Int. Ed.*, 2022, **61**, e202114154.
31. E. H. Abdelkader, H. Qianzhu, G. Otting and T. Huber, *Chem*, 2025, **11**, 102385.
32. A. Beeby, I. M. Clarkson, R. S. Dickins, S. Faulkner, D. Parker, L. Royle, A. S. de Sousa, J. A. G. Williams and M. Woods, *J. Chem. Soc., Perkin Trans. 2*, 1999, DOI: 10.1039/A808692C, 493–504.
33. E. C. Cook, E. R. Featherston, S. A. Showalter and J. A. Cotruvo, *Biochemistry*, 2019, **58**, 120–125.
34. H. Park, P. Bradley, P. Greisen, Jr., Y. Liu, V. K. Mulligan, D. E. Kim, D. Baker and F. DiMaio, *J. Chem. Theory Comput.*, 2016, **12**, 6201–6212.
35. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C. C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Zemgulyte, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Zidek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **630**, 493–500.
36. Y. Shen, F. Delaglio, G. Cornilescu and A. Bax, *J. Biomol. NMR*, 2009, **44**, 213–223.
37. T. Kimura and Y. Kato, *J. Alloys Compd.*, 1995, **225**, 284–287.
38. J. Roel-Touris, M. Nadal and E. Marcos, *Nat. Commun.*, 2023, **14**, 5939.
39. J. Wang, X. Liu, K. Li, T. Shi, Q. Xu, T. Peng, Q. Huang, Z. Gao, H. Zhou, W. Lu and J. Wang, *J. Am. Chem. Soc.*, 2025, **147**, 15205–15215.
40. J. Sun and M. Petersheim, *Biochim. Biophys. Acta, Biomembr.*, 1990, **1024**, 159–166.

41. M. Komiyama, N. Takeda and H. Shigekawa, *Chem. Commun.*, 1999, DOI: 10.1039/A901621J, 1443–1451.
42. C. Chen, P. Zhang and Z. Chai, *Anal. Chim. Acta*, 2001, **439**, 19–27.
43. C. D. Eads, P. Mulqueen, W. D. Horrocks, Jr. and J. J. Villafranca, *J. Biol. Chem.*, 1984, **259**, 9379–9383.
44. J. A. Mattocks, J. V. Ho and J. A. Cotruvo, Jr., *J Am Chem Soc*, 2019, **141**, 2857–2861.
45. Q. Wu, X. L. Liu, Z. F. Chai, K. Cheng, G. H. Xu, L. Jiang, M. L. Liu and C. G. Li, *Chem. Commun.*, 2022, **58**, 8230–8233.
46. H. S. Zurier, R. Farinato, K. H. Kucharzyk and S. Banta, *Appl Environ Microbiol*, 2025, **91**, e0245024.
47. R. Jastrzab, M. Nowak, M. Skrobanska, A. Tolinska, M. Zabiszak, M. Gabryel, L. Marciniak and M. T. Kaczmarek, *Coord. Chem. Rev.*, 2019, **382**, 145–159.
48. H. Singer, B. Drobot, C. Zeymer, R. Steudtner and L. J. Daumann, *Chem. Sci.*, 2021, **12**, 15581–15587.
49. P. Mandal, J. Kretzschmar and B. Drobot, *J. Biol. Inorg. Chem.*, 2022, **27**, 249–260.
50. T. Mukai, H. Hoshi, K. Ohtake, M. Takahashi, A. Yamaguchi, A. Hayashi, S. Yokoyama and K. Sakamoto, *Sci. Rep.*, 2015, **5**, 9699.
51. J. Klopp, A. Winterhalter, R. Gebleux, D. Scherer-Becker, C. Ostermeier and A. D. Gossert, *J. Biomol. NMR*, 2018, **71**, 247–262.

**Chapter 3: Mapping conformational heterogeneity: ^{19}F - Gd^{3+}
ENDOR reveals minor rotamer conformations**

3.1 Introduction

RFdiffusion¹ and AlphaFold² have transformed structural biology but predict static protein structures. Proteins are dynamic, and conformational heterogeneity is intrinsic to proteins in solution.³ Static structures capture single conformations but do not represent thermally populated states. In solution, sidechains sample multiple conformations that determine protein function, while backbones are more likely to be restrained by secondary structure which satisfies many hydrogen bonds.⁴ Although preferred rotamer orientations are encoded in backbone geometry,⁸ this is not deterministic. In protein design, knowing not only the major sidechain conformation but also its minor conformations is advantageous, because design and function may need to favour these minor states. Sidechain rotamers determine packing,⁵ hydrogen-bond networks,⁶ and preorganisation of catalytic or binding sites.⁷ In lanthanide-luminescent proteins, antenna residue distance and orientation is a critical variable for excitation transfer efficiency to bound metal centres. Minor rotamers that position chromophores optimally can dominate excitation, despite low populations.

Characterising minor conformations requires an understanding of which sidechain conformations are physically plausible. Rotamers of tryptophan are constrained by sterics, (χ_1, χ_2) combinations are preferred when they minimise steric clashes between atoms in the tryptophan (Trp) and other atoms in the protein. The Top8000 rotamer library (Figure 3.1) illustrates an empirical energy landscape for Trp (χ_1, χ_2) angles, and represents a curated, non-redundant set of high-resolution PDB structures.⁹ In X-ray crystallography, this distribution is used to validate structures, and outliers help identify misinterpretation of the data. This chapter applies a similar philosophy to the interpretation of experimental distance data; using rotamer distributions built upon AlphaFold models guides the interpretation of complex ¹⁹F electron nuclear double resonance (ENDOR) spectra, uncovering minor rotamer species.

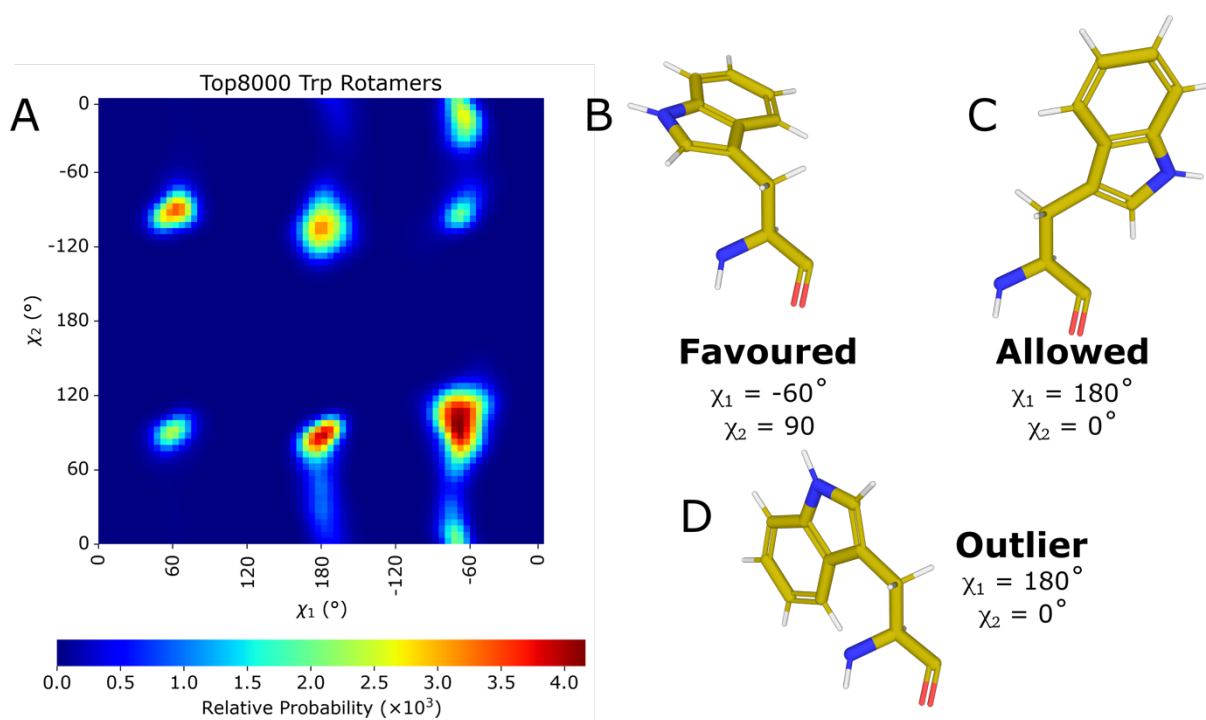


Figure 3.1 – Top8000 rotamer distribution. (A) Heatmap of the probabilities from the Top8000 rotamer distribution.⁹ The two dihedral angles χ_1 and χ_2 define the orientation of the Trp indole relative to the protein backbone: χ_1 describes rotation around the C_α - C_β bond, while χ_2 describes rotation around the C_β - C_γ bond. The distribution reveals that specific conformations of (χ_1, χ_2) are energetically favoured. Probability peaks correspond to rotamers that minimise atomic overlap between the bulky indole sidechain and the protein backbone or neighbouring residues. Large blue regions represent conformations that are sterically forbidden due to atomic clashes. (B) Stick representation of a “favoured” rotamer, corresponding to a high probability region in the heatmap. (C) An “allowed” rotamer, representing a moderately populated conformation with acceptable but sub-optimal geometry. (D) An “outlier” rotamer, corresponding to a low-probability region that is rarely observed in experimental structures.

Traditional structural methods can access sidechain heterogeneity, but have practical limitations. Crystallography and cryo-EM yield rotamer ensembles only after full model building and refinement; ensemble refinement is non-trivial and requires iteratively refitting entire protein models to experimental data.¹⁰ Protein NMR spectroscopy can directly report rotamer populations using assigned chemical shifts, but is limited to proteins under 40 kDa.¹¹ Methods that directly probe sidechain conformational distributions without global structure refinement would therefore complement existing approaches.

High-field (>1.2 T, >34 GHz) ^{19}F ENDOR spectroscopy combined with genetic encoding of fluorinated non-canonical amino acids (ncAAs) provides distance measurements in

proteins under physiological conditions. ^{19}F ENDOR exploits the high sensitivity of electron paramagnetic resonance (EPR) detection at high magnetic fields to measure magnetic coupling between an introduced paramagnetic centre and nearby ^{19}F nuclei,¹² yielding distances between 5–20 Å in proteins labelled with high-spin paramagnetic centres such as gadolinium (Gd^{3+}).^{12,13} An ENDOR spectrum is a population-weighted measurement of the ^{19}F – Gd^{3+} distance distribution in the sample. The spectrum is the sum of the splittings ($\sim 1/r^3$) and the intensities ($\sim 1/r^6$) for an ensemble of ^{19}F – Gd^{3+} distances (Figure 3.2A). A recent report showed that in metalloproteins with rigid Gd^{3+} sites, the orientation of buried sidechains could be determined by triangulating ^{19}F ENDOR datasets from systematically fluorinated amino acids, giving conformations consistent with reported crystal structures.¹⁴

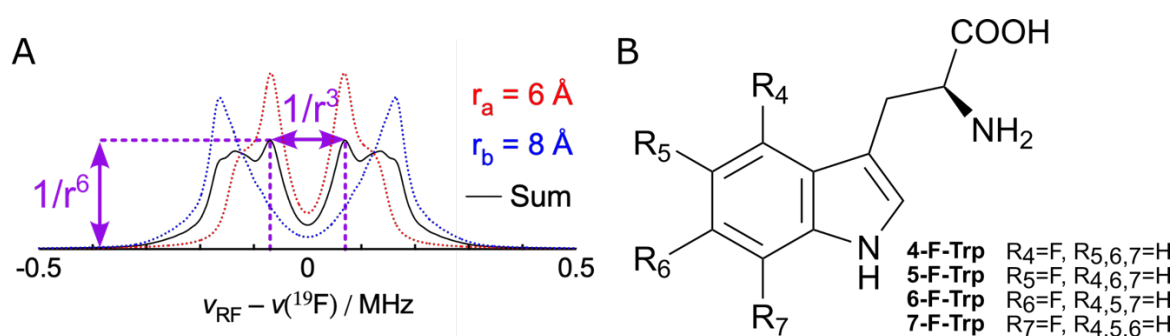


Figure 3.2 – Simulated ^{19}F – Gd^{3+} ENDOR spectra and the Trp variants used in this Chapter. (A) Relationship between ^{19}F – Gd^{3+} distance and ENDOR spectral features: dipolar coupling manifests as splitting which scales as $1/r^3$ (horizontal) while signal intensity scales as $1/r^6$ (vertical). Spectra from samples with multiple ^{19}F – Gd^{3+} distances are population-weighted sums where $\rho_a = \rho_b$. (B) Chemical structures of the four monofluorinated tryptophan variants (4-F-Trp, 5-F-Trp, 6-F-Trp, 7-F-Trp) used as ncAAs.

Here, conformational analysis by ^{19}F ENDOR is extended from rigid, buried sidechains to solvent-exposed residues with dynamic conformations. The lanthanide-binding mini-protein RF2 developed in Chapter 2 is an ideal candidate for developing this technique, as the site of the paramagnetic probe is fixed and conformationally unambiguous, while the Trp is solvent exposed and dynamic. RF2 is a *de novo* designed lanthanide-binding mini-protein where Trp42 transfers photoexcitation to bound lanthanides (Ln^{3+}) via the antenna effect. Energy transfer via the antenna effect depends on Trp42– Ln^{3+} proximity and orientation, making rotamer populations functionally consequential. Using Förster formalism, Harrocks and co-workers validated $R_0 \approx 3.4 \text{ Å}$ for Trp– Tb^{3+} pairs in Tb^{3+} -bound

parvalbumin and thermolysin.^{15,16} Förster efficiency scales as $E = 1/(1+(R/R_0)^6)$.¹⁷ Therefore, a rotamer flip that decreases R from 10 Å → 7 Å increases E from ~0.002 → ~0.013, an 8.5-fold enhancement, illustrating how changes in rotamer geometry can significantly affect antenna performance.

The solvent-exposed Trp42 is positioned with its backbone carbonyl oxygen coordinated to the Ln³⁺ ion and the indole sidechain exposed to solvent, thus a characterisation of its rotamer will inform future design iterations. Using a comprehensive set of mono-fluorinated tryptophan residues (n -F-Trp, $n = 4,5,6,7$) (Figure 3.2B), this chapter infers common indole geometries using the distance correlations between the ¹⁹F–Gd³⁺ distances.

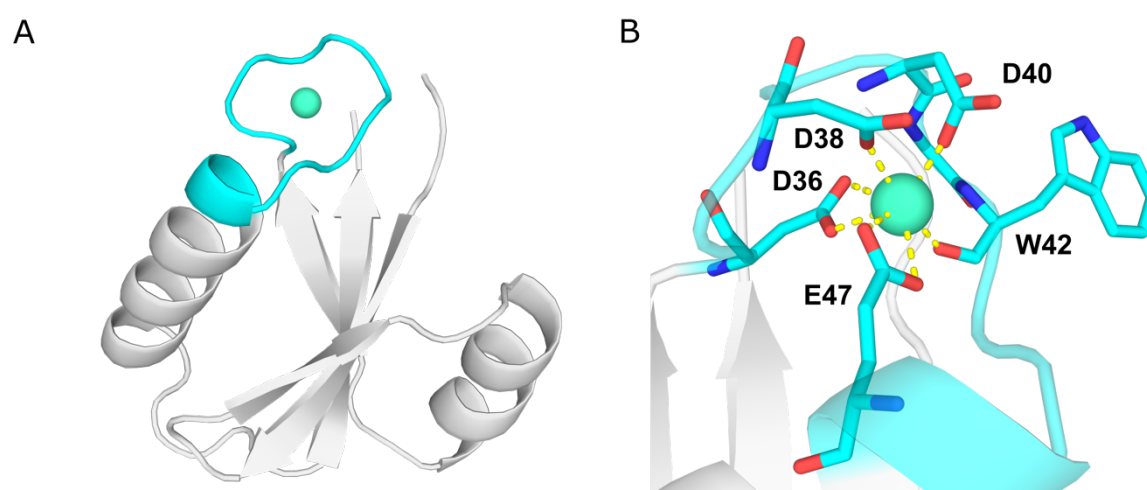


Figure 3.3 – Overall structure of RF2 and the lanthanide-binding region (A) Predicted structure of RF2 shown as a cartoon representation, with the lanthanide-binding region (residues 36–47) coloured cyan and the Gd³⁺ ion represented as a green sphere. (B) Close-up of the predicted coordination environment of Gd³⁺, with coordinating residues illustrated in the stick representation. In this predicted structure, seven oxygen atoms are coordinated to the lanthanide, as indicated by the dashed lines.

3.2 Fitting of protein or ENDOR frame parametrised models to ¹⁹F–Gd³⁺ ENDOR spectra

RF2 is highly soluble and attempts to crystallise the protein were unsuccessful. The AlphaFold-predicted tertiary structure was confirmed by backbone NMR chemical shift

assignment using $^{15}\text{N}/^{13}\text{C}$ enriched samples. The rigid Ln^{3+} binding site in this validated structure provides a fixed reference point for $^{19}\text{F}\text{--}\text{Gd}^{3+}$ distance measurements.

Loading Gd^{3+} into the metal binding site of samples of RF2, where Trp42 was replaced with four monofluorotryptophan variants (4-, 5-, 6- and 7-FTrp), produces four samples with a Gd^{3+} paramagnetic centre near a ^{19}F nucleus. The four ncAAs were incorporated during protein biosynthesis in *E. coli* as described in §3.5.¹⁸ Previous NMR studies showed rotamer configurations are conserved across fluorotryptophan isomers, but populations vary.^{Qianzhu, 2024 #97} The four FTrp isomers together provide four distinct $^{19}\text{F}\text{--}\text{Gd}^{3+}$ distance distributions for a given rotamer. A single $^{19}\text{F}\text{--}\text{Gd}^{3+}$ distance maps to multiple (χ_1, χ_2) solutions due to geometrical degeneracy; four fluorine positions reduce this degeneracy.

Following the measurement of ^{19}F ENDOR spectra for the four fluorotryptophan variants, the $^{19}\text{F}\text{--}\text{Gd}^{3+}$ distance distributions $\{r_{\text{F}}\} = \{r_{\text{F4}}, r_{\text{F5}}, r_{\text{F6}}, r_{\text{F7}}\}$ must be fit to some model to provide structural information. Fitting orientations from $\{r_{\text{F}}\}$ is an ill-posed problem; many distinct (χ_1, χ_2) combinations can yield similar distance sets. Two initial approaches constrain the system differently: i) the protein frame, parameterised in terms of the sidechain rotation angles $\{\chi_1, \chi_2\}$ (Figure 3.4A,B); and ii) the ENDOR frame, parametrised in terms of the $^{19}\text{F}\text{--}\text{Gd}^{3+}$ vector/Trp ring plane, described by the inter-spin distance (r_{c}) and the Euler angles (ϕ, θ) (Figure 3.4C,D).

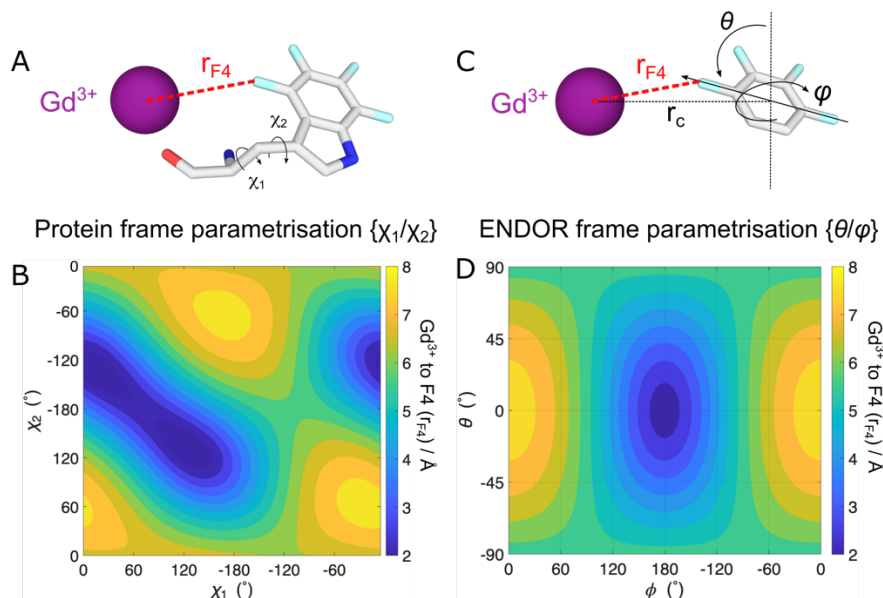


Figure 3.4 – Initial frame parametrisation for fitting ^{19}F - Gd^{3+} ENDOR spectra. (A) Protein-frame parametrisation showing sidechain dihedral angles χ_1 and χ_2 . (B) Contour plot of r_{F4} as a function of (χ_1, χ_2) demonstrates how fluorine position varies across rotamer space (colour scale in Å). (C) ENDOR frame parametrisation using the Gd^{3+} →ring-centre distance, r_c , azimuthal angle ϕ , and tilt angle θ (where $\phi = 180^\circ$, $\theta = 0^\circ$ has the ring vector aligned with r_c). (D) Contour plot of r_{F4} as a function of (ϕ, θ) at fixed $r_c = 5$ Å (colour bar in Å).

Fitting the distances from data generated from the protein model has the advantage of simple mapping of the distance measurements onto the protein model. We simulated the ^{19}F -ENDOR data using rotamers generated from AlphaFold models. After first sampling the $\{\chi_1, \chi_2\}$ rotamer space in the protein frame and calculating $\{r_F\}$ for each rotamer, each (χ_1, χ_2) was scored by a per-atom clash value. A previous approach simply used a crystal structure, and scaled van der Waals (vdW) radii by a factor of 0.8 and discarded rotamers with overlap ($d_{ij} < \sum \text{vdW}_{ij}$).¹⁴ Without a crystal structure, clash scores and $\{r_F\}$ values were averaged across five AlphaFold models as the simple mean. The clash-weighted rotamer library generation is described extensively in Appendix A.2.3. The clash-filtered library provides physically plausible sampling of distances, eliminating non-physical distances < 4 Å (Figure 3.5A) but cannot predict relative populations. The clash filtered library alone produced poor *EasySpin* ENDOR simulation fits (Figure 3.5B). Dipolar splittings were somewhat reproduced, but short distances were not correctly represented resulting in poor fits to the shoulder regions of the ENDOR lineshapes. This indicated that simply sampling clash-allowed $\{\chi_1, \chi_2\}$ space is insufficient to capture the true distance distributions.

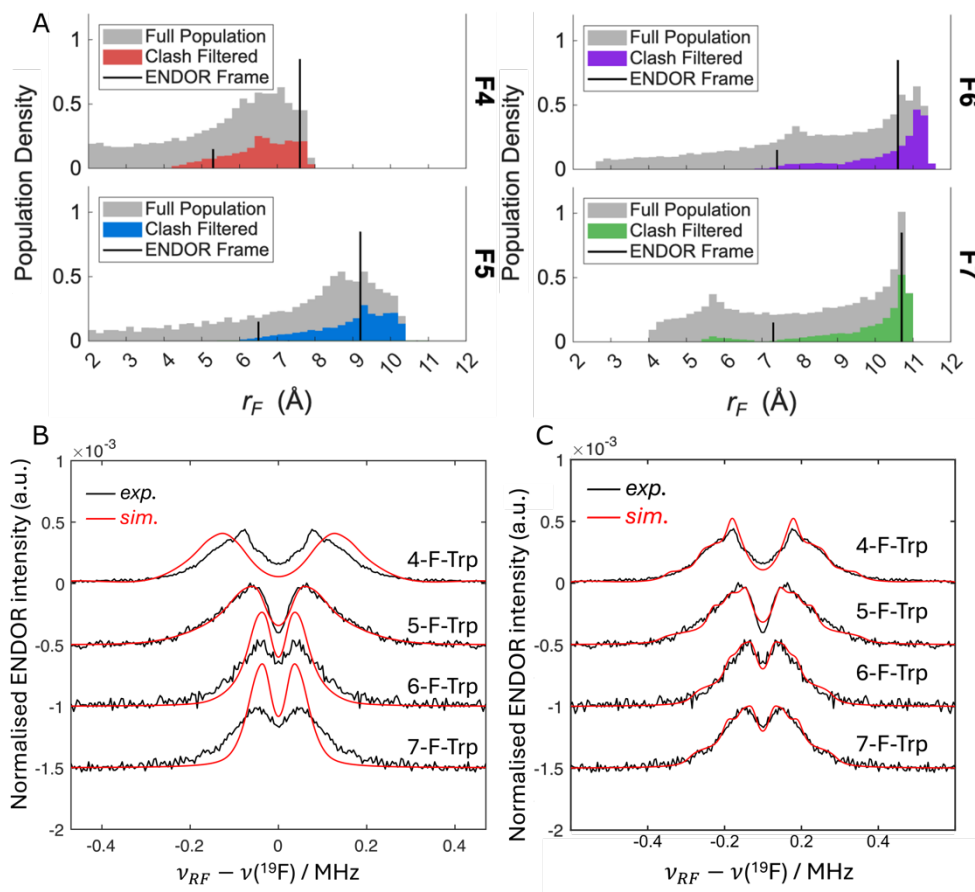


Figure 3.5 – Initial ^{19}F - Gd^{3+} ENDOR simulation using protein and ENDOR frame parametrisation. (A) Distance distributions for each fluorine positions before (grey) and after (coloured) clash filtering using a 0.4 Å cut-off. Black vertical lines represent the distances determined from ENDOR frame fits. (B) Experimental ^{19}F ENDOR spectra of RF2 variants (black trace) against simulated spectra using the clash-masked rotamer libraries from (A). (C) Experimental ^{19}F ENDOR spectra of RF2 variants (black trace) against simulated spectra using the ENDOR frame fit distances marked in (A).

In parallel, the data was fit directly in the ENDOR frame by simulating discrete ring orientations defined by the parameters $\{r_c, \phi, \theta\}$ outlined in Figure 3.4, which allows computationally efficient fitting.¹⁴ Unlike the protein frame, the ENDOR frame is agnostic about conformation of the whole protein and produces distributions that do not directly map to sidechain conformations in $\{\chi_1, \chi_2\}$ parameter space. The experimental data fit poorly to a single (r_c, ϕ, θ) , and the fit improved by including two discrete ring orientations (Figure 3.5C). This involves a seven-parameter fit: (r_c, ϕ, θ) for orientation one, (r'_c, ϕ', θ') for orientation two, and their relative population $p_1 = 1 - p_2$. The $\{r_F\}$ distance components and weights predicted by this two-orientation fit correlates with the peaks in the clash-filtered protein-frame predictions above, indicating that both approaches produce sets of physically plausible conformations. However, fitting more than two

discrete conformations would, no doubt, improve the fit quality but risks overfitting rather than capturing useful geometric information. It also becomes computationally expensive: three conformers require 11 parameters (3×3 orientation parameters + 2 population weights). These results were unexpected, given these approaches converged in previous work for buried phenylalanine.¹⁴

Simulations illustrate that Gaussian conformational distributions in $\{\chi_1, \chi_2\}$ space of rotamers produce distance distributions that cannot be simply approximated by Gaussians with shared variables. Using the clash-naïve protein parametrisation above, 500 points were sampled around two (χ_1, χ_2) pairs with a standard deviation (SD) of 5° . This could then produce distance distributions for each fluorine that correlated to an identical rotamer distribution. Figure 3.6 shows that the two rotamers represented by Gaussians (χ_1, χ_2) space cannot be approximated by Gaussians with shared parameters in distance space. Notably, F7 produces a distance distribution that would best be approximated by a single gaussian, whereas F4, F5, and F6 are two distinct populations. Furthermore, the rotamers are simulated at populations of 1:1, whereas the ratios by peak height in F4, F5 and F6 are closer to 1:5. Furthermore, the protein frame approach using clash-filtered distance distributions presupposes that sampled indole conformations have shared populations across RF2 FTrp variants. Where only a single, buried rotamer is concerned, this may be true, but for dynamic FTrp sidechains, the literature suggests that different conformations have different relative populations.^{18–22} This motivates a generalisable approach that efficiently fits distance distributions from a set of physically plausible conformations.

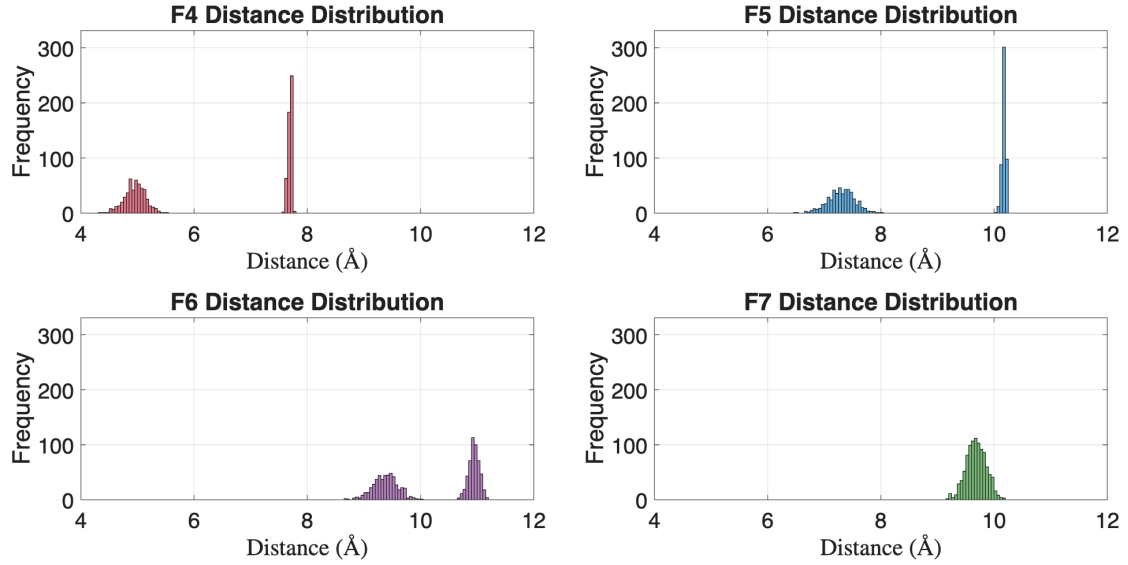


Figure 3.6 – Simulated F-distance distributions from two (χ_1, χ_2) rotamers. In $\{\chi_1, \chi_2\}$ dihedral space, 500 points were sampled with a 5° SD around $(\chi_1 = 180^\circ, \chi_2 = -60^\circ)$ and $(\chi_1 = -60^\circ, \chi_2 = -60^\circ)$. For each sampled (χ_1, χ_2) pair, the corresponding $(r_{F4}, r_{F5}, r_{F6}, r_{F7})$ was determined using the models in Appendix A.2.3. These values were then plotted in the above histogram using $0.05 \text{ \AA}$ binning.

3.3 Uncovering minor rotamer conformations from simulated $^{19}\text{F}\text{--}\text{Gd}^{3+}$ distributions using Tikhonov regularisation

The forward mapping from distance distributions $\{r_F\}$ to $S(\nu)$ is injective, but the reverse is not. Recovering $\{r_F\}$ from experimental spectra is therefore an ill-posed inverse problem. Using Tikhonov regularisation stabilises this inverse problem, combining least-squares fitting of the measured spectrum with a penalty that favours smooth, physically realistic distance distributions.

The ENDOR spectrum $S(\nu)$ is related to $\{r_F\}$ by Eq. 1:

$$S(\nu) = \int_{r_{\min}}^{r_{\max}} K(r_F, \nu) P(r_F) dr_F \quad 1$$

Where $P(r_F)$ is the $^{19}\text{F}\text{--}\text{Gd}^{3+}$ distance distribution for a given F_n , and $K(r_F, \nu)$ is the ENDOR kernel, an array of ENDOR lineshapes for all possible distances r_{F_n} in the range $[r_{\min}, r_{\max}]$. Tikhonov regularisation estimates a solution $P(r_F)$ for $S(\nu)$ by combining ordinary least-squares fitting with a P -dependent penalty term:

$$P_{fit} = \arg \min_{P \geq 0} (\|S - KP\|^2 + \alpha^2 \|\Gamma P\|^2) \quad 2$$

The first term enforces the least squares minimisation, while the second term contains the regularisation operator Γ (typically the second derivative) and regularisation parameter α that together ensure the solution is smooth and realistic.

Pribitzer *et al.* applied Tikhonov regularisation to ENDOR data of vanadyl porphyrin complexes with distributed isotropic and dipolar hyperfine components.²³ Instead of imposing a smoothness criterion via $\|\Gamma P\|$, they regularised against a prior distribution $\tilde{P}$, derived from DFT calculations. This adapted approach has two modifications: (i) the Kernel $K(r, \nu)$ was sampled directly from the clash-filtered rotamer library using non-uniform discretisation of distances, making $K(r, \nu)$ more densely sampled at more probable distances and excludes conformations with a clash score above a certain threshold; and (ii) both the second-derivative and prior regularisation operators, and the prior distribution $\tilde{P}$ was the clash-filtered library were tested. Figure 3.7A–D shows the regularised distance distribution solutions and regularised ^{19}F ENDOR spectrum simulations to fit the experimental data. These distributions derive from discretised rotamers in from Figure 3.5 and refine the distance populations (Figure 3.7E–H).

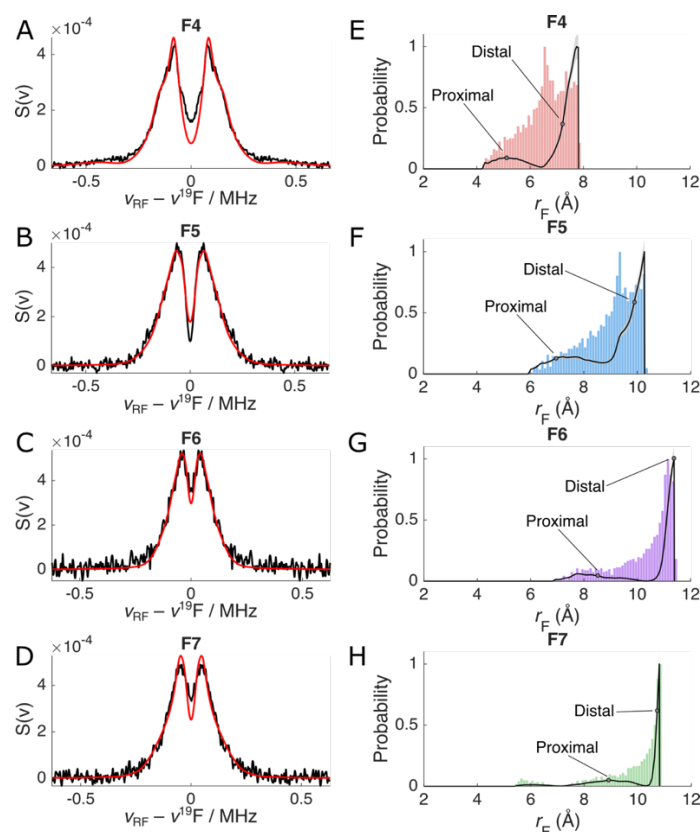


Figure 3.7 – Tikhonov regularisation fits distance distributions to experimental spectra. (A–D) Experimental ^{19}F ENDOR data (black) with *Easyspin* simulations using regularised distance distributions (red). $S(t)$ is the ENDOR echo intensity as a function of radiofrequency offset from the ^{19}F Larmor frequency. (E–H) Regularised distance distribution from Tikhonov regularisation (black line) compared to distance distribution from clash-filtered library (coloured histogram). Annotations represent consensus rotamers from Figure 3.10.

The mapping of distances to corresponding (χ_1, χ_2) pairs is retained in the non-uniform discretisation, meaning the 2D probability densities of $\{\chi_1, \chi_2\}$ can be reconstructed from the same regularisation. Plotting the product of these probability densities reveals that the geometric degeneracy between (χ_1, χ_2) is not distinguished in this approach. Figure 3.8 illustrates that multiple (χ_1, χ_2) pairs have indistinguishable probabilities. These solutions are related by $\sim 90^\circ$ rotations in χ_1 , and $\sim 180^\circ$ rotations in χ_2 . While these geometrically degenerate pairs may reflect legitimate variability in the system, the indifferentiable confidence makes assignment of rotamers from distances using this methodology difficult.

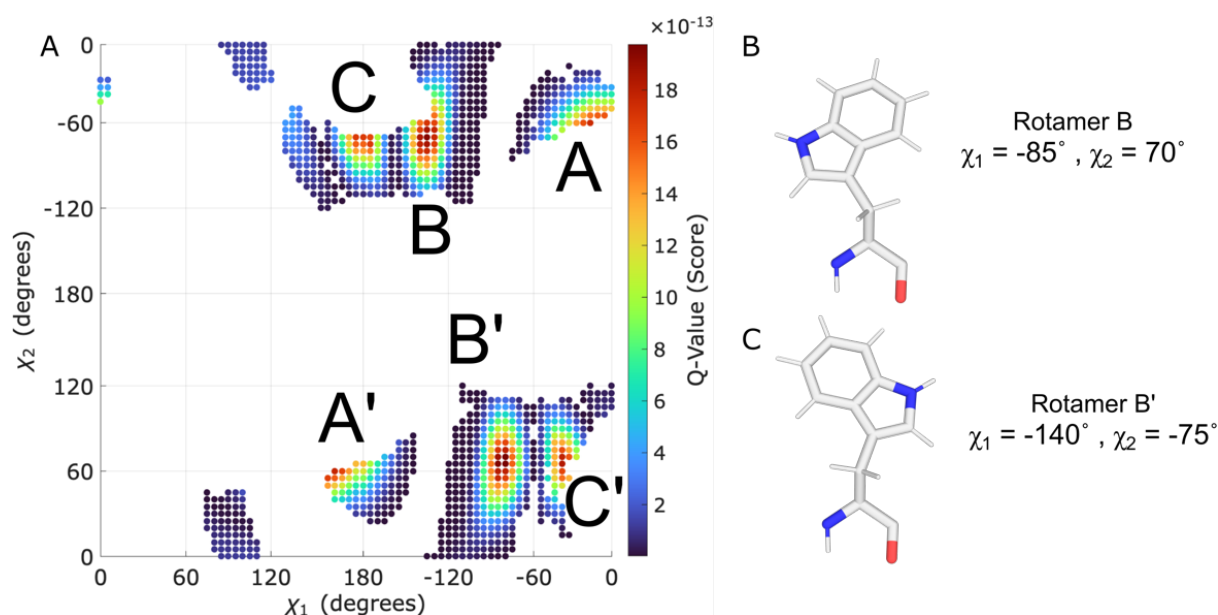


Figure 3.8 – Geometric degeneracy in rotamer distributions. (A) Q-value plots of rotamer fits from Tikhonov regularisation. The 2D (χ_1, χ_2) distribution matrices from four independent Tikhonov fits to ^{19}F - Gd^{3+} ENDOR spectra are normalised by area. Q-value at each (χ_1, χ_2) is the product of these four normalised probability distributions. Geometric degeneracy produces symmetry-related peaks of equivalent confidence, labelled (A, A'), (B, B'), (C, C'). Values (χ_1, χ_2) pairs with Q-values $< 1 \times 10^{-13}$ are not plotted. (B–C) Stick illustration of two geometrically degenerate rotamers from (A). Both illustrations have identical backbone dihedrals, and (χ_1, χ_2) values were modified to match the coordinates of local Q-value maxima in (A).

To resolve this degeneracy, a consensus approach factoring in the relative probabilities of each (χ_1, χ_2) pair in each Tikhonov regularisation. A frequently populated distance may arise from many different rotamers, not a single predominant conformation. The Tikhonov-regularised (χ_1, χ_2) probability distributions reflect which physically plausible conformations best explain the experimental data for each fluorine. In the consensus approach, 2D probabilities are normalised and rotamers for each (χ_1, χ_2) that exceeds a 30% threshold are considered a “match”. Consensus rotamers are those whose match count = 4. This stringency yields only 3 rotamers satisfying all four fluorotryptophan variants.

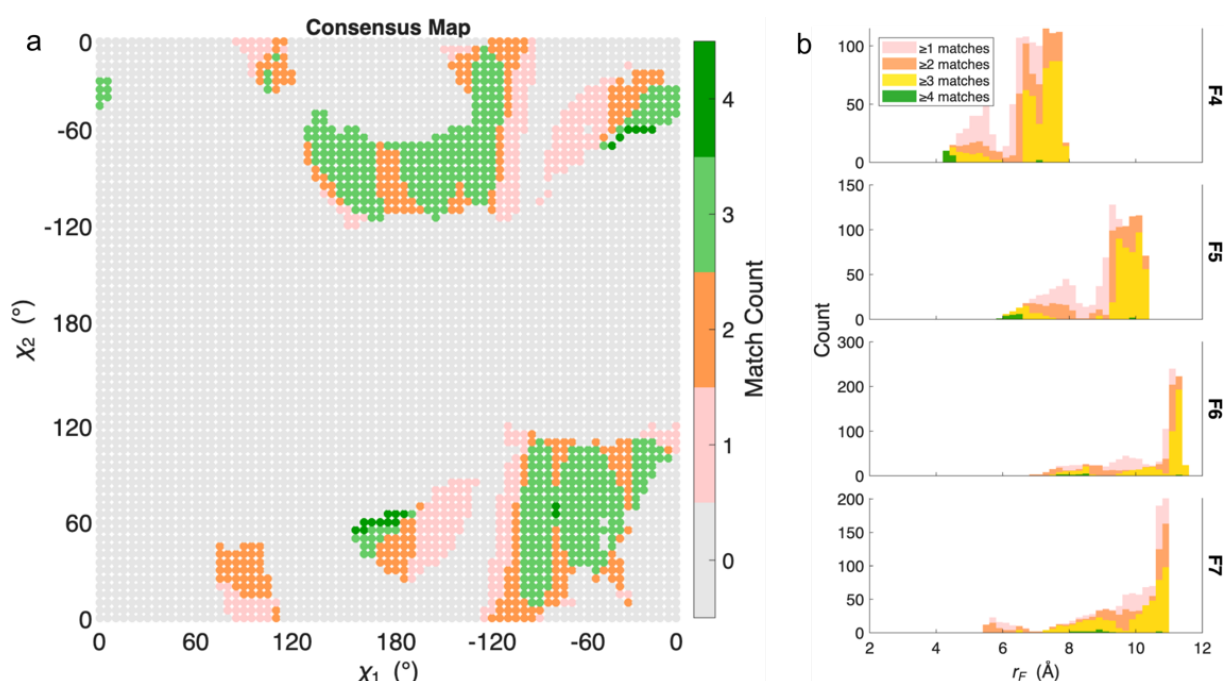


Figure 3.9 – Consensus approach to filtering Trp rotamers from Tikhonov regularisation. Each 2D probability density is normalised to its maximum: $P_{norm}(\chi_1, \chi_2) = P(\chi_1, \chi_2) / \max[P(\chi_1, \chi_2)]$. A rotamer passes threshold for a given fluorine if $P_{norm}(\chi_1, \chi_2) \geq 0.30$ (30% of maximum). The 30% threshold captures all minor peaks in (χ_1, χ_2) distributions while excluding broad, diffuse regions representing overlapping conformations. (A) Consensus map, match counts (0–4) at each (χ_1, χ_2) pair. Three rotamers achieve 4 matches (dark green) (b) Model distance distributions with ≥ 1 , ≥ 2 , ≥ 3 or 4 matches, coloured pink, orange, yellow and green, respectively.

Figure 3.10 shows the rotamers identified as consensus conformations. Three consensus rotamers were identified: one distal rotamer Figure 3.10A with orientation similar to the AlphaFold predictions Figure 3.10D, and two proximal rotamers with shorter, less common distances to the lanthanide site Figure 3.10B–C. The ^{19}F – Gd^{3+} distances for each rotamer are then compared the distance-space % of max in Figure 3.7,

meaning evidence supporting the population of each consensus rotamer in the model can be compared with experimental values. The distance r_c between the centre of the six-membered ring and the Gd^{3+} ion is 8.6 Å, 6.4 Å and 6.6 Å for the distal, proximal 1 and proximal 2 rotamers, respectively, indicating that the proximal rotamers represent a ~ 2 Å decrease in $^{19}\text{F}\text{--}\text{Gd}^{3+}$ distance. The AlphaFold model ring distance of 8.6 Å is consistent with the distal rotamer, and although the (χ_1, χ_2) values differ, both conformations fall within the same rotamer region of the (χ_1, χ_2) plot.

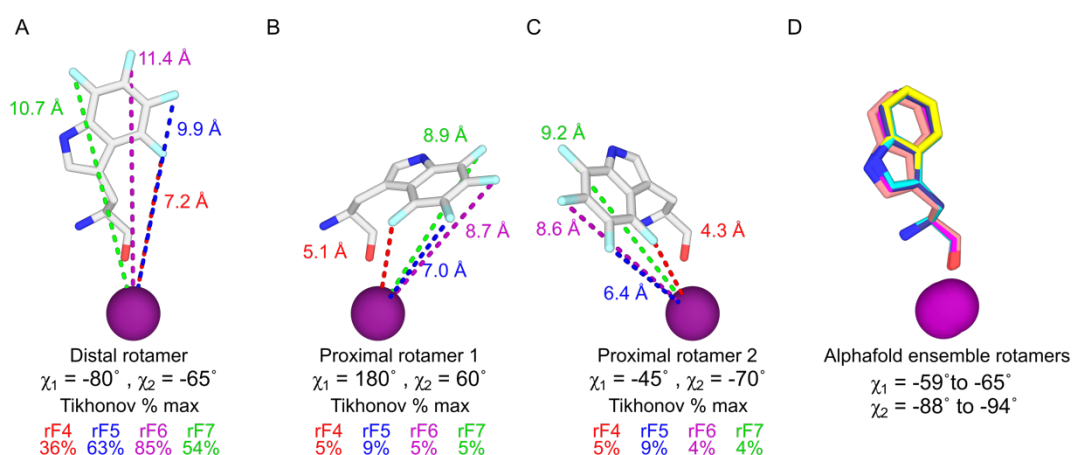


Figure 3.10 – Consensus rotamers from remapping of Tikhonov regularisation distances to rotamer conformers. Conformations are annotated with $\bar{r}_{\text{Fn}}$ values from the AlphaFold ensemble, and % of maximum $\bar{r}_{\text{Fn}}$ values in the Tikhonov distributions. (A–C) Fluorotryptophan rotamers illustrated in the stick representation. Gd^{3+} ion represented as a purple sphere. Annotated (χ_1, χ_2) values and distance-space likelihoods as % of maximum (F4–F7), within Tikhonov $P(r_F)$. (D) AlphaFold ensemble rotamers, coloured by model number: model_0, dark blue; model_1, cyan; model_2, magenta; model_3, yellow; model_4, salmon. Annotated (χ_1, χ_2) represents the range across five models.

The diversity of rotamers for each fluorine reflects the solution ensemble captured by snap-freezing small (4 μL) room-temperature samples in liquid N_2 and then cooling to 10 K for ENDOR. The modelling indicates that the same (χ_1, χ_2) sampling cannot be assumed for all four fluorotryptophan variants suggesting that fluorination changes the population densities. A simultaneous regularisation of all four data sets would help answer this question, which requires the intensities of the ENDOR spectra to be directly comparable to encompass a common solution with a single kernel. However, expression at 25°C rather than 18°C resulted in a large proportion of glutamine suppression at position 42, (Appendix Figures A.29–32) preventing quantitative interpretation of the ENDOR intensities across samples. Time-domain ENDOR could improve confidence by

producing narrower linewidths and better resolving dipolar coupling components,²⁴ but requires a high-power RF amplifier.

3.4 Conclusion

¹⁹F–Gd³⁺ ENDOR reveals Trp42 in RF2 reveals at least three discrete rotamers: one distal conformation ($\chi_1 = -80^\circ$, $\chi_2 = -65^\circ$) matching AlphaFold predictions, and two proximal conformations ($\chi_1, \chi_2 = 180^\circ, 60^\circ$ or $\chi_1, \chi_2 = -45^\circ, -70^\circ$) positioning the indole closer to the Gd³⁺ site. The proximal rotamers are less populated but place Trp42 in more photoactive geometries for excitation via the antenna effect. AlphaFold predicts only the major, less photoactive conformation, demonstrating that computational models miss functionally important minor populations.

For RF2 optimisation, stabilizing the proximal rotamers (Trp–Gd³⁺ $\approx 6\text{--}7$ Å vs Trp–Gd³⁺ ~ 9 Å) should enhance luminescence quantum yield. Mutagenesis to restrict (χ_1, χ_2) sampling or backbone redesign to favour proximal geometries would test whether the minor conformations dominate photoexcitation transfer despite low populations. This is a narrow example of potential applications: catalytic residues, binding sites, and chromophores often derive function from minor conformations that ensemble-averaged structures obscure.

Computational models may predict major conformations but can miss functionally important minor populations. Notably, the AlphaFold model provided the geometric scaffold that enabled identification of this minor population. This contrasts the typical workflow where predicted structures serve as an initial model to be superseded by experimental structure determination. For proteins that resist crystallisation, this approach provides design-relevant conformational information that would otherwise be inaccessible. The hybrid approach—genetic ncAA incorporation, ¹⁹F–Gd³⁺ ENDOR, and Tikhonov regularization against clash-filtered rotamer libraries—maps dynamic sidechain conformations within 5–20 Å of paramagnetic centres. The method generalises to any fluorinated sidechain (buried or exposed) and requires only appropriate geometric parametrisation of the kernel $K(r, \nu)$.

This work establishes that ^{19}F - Gd^{3+} ENDOR can determine conformational heterogeneity in solution for dynamic, solvent-exposed sidechains—extending the technique beyond the rigid, buried residues previously characterised. The consensus filtering approach (Figure 1.8) used in rotamer assignment is robust in the face of questions surrounding conformational proportions across different samples, providing a systematic framework for interpreting multi-label ENDOR datasets.

3.5 Methods

3.5.1 Protein expression and purification

The protein variants were produced following the protocol using fluorindoles from Qianzhu *et al.* with minor modifications.¹⁸ The pCDF plasmid containing RF2-W42Z (“Z” denotes amber incorporation site) was co-transformed with a pRSF plasmid containing either G1F4W27 or G1F7W47 into *E. coli* B95 B-95. $\Delta\text{A}\Delta\text{fabR}$ cells.²⁵ Transformed cells were grown at 37 °C in LB medium containing 25 mg/L kanamycin and 25 mg/mL spectinomycin. 500 μL of the overnight culture was used to inoculate 50 mL LB medium, supplemented with same concentration of antibiotics. Cells were grown at 37 °C to an OD_{600} of 0.5, at which fluorindole was added to a final concentration of 0.5 mM. For 4-fluorindole, the final concentration was 2 mM, and the media was supplemented with L-Tryptophan to a final concentration of 4 mM. The cells were grown for an additional 30 minutes at 37 °C. The temperature was then reduced to 25 °C and protein expression was induced by the addition of 1 mM IPTG.

After expression for 16 hours, the fusion protein was isolated from cell lysate using immobilised metal-ion affinity chromatography (IMAC). The fusion protein was then dialysed into TEV buffer (Tris-HCl pH 8.0, 300 mM NaCl) and incubated for 16 hours at 4 °C with mild shaking after the addition of TEV protease at 50:1 ratio of protein:TEV and 10 mM DTT. The resultant sample was then subjected to reverse IMAC, where the sample flow through and column wash are retained. Incorporation of fluorotryptophan was confirmed by mass spectrometry (Appendix A.2).

3.5.2 ENDOR sample preparation

After reverse-IMAC, protein samples were dialysed into 20 mM MOPS pH 7.0, 50 mM KCl in 100% D₂O. 10 μ L samples were prepared at 200 μ M protein concentration, with 100 μ M Gd³⁺ (from 1 mM stock in 100% D₂O) and 1 μ L d₈-glycerol. After preparation, samples were incubated for at least 1 hour at 25 °C. 4 μ L of sample was then transferred into W-band tubes. Before collection, samples were snap frozen in liquid nitrogen.

3.5.3 EPR, ENDOR measurements, and lineshape calculations

The EPR and ENDOR spectroscopy, Tikhonov regularisation and ENDOR spectrum simulation were performed by collaborator Martyna Judd. All EPR and ¹⁹F ENDOR measurements were performed at 10 K on a modified W band 94 GHz E680 Bruker spectrometer equipped with a TE₀₁₁-mode EPR/ENDOR resonator in a helium gas flow cryostat (Oxford Instruments). Echo-detected field sweep spectra were collected to characterise the EPR lineshape using the standard Hahn echo sequence $\pi/2-\tau-\pi-\tau$ -echo, with $\pi/2 = 20$ ns, $\pi = 40$ ns and $\tau = 300$ ns, recording the integrated echo intensity as a function of field position. ¹⁹F Mims ENDOR spectra were acquired at the field maximum of the Gd³⁺ EPR-spectrum, using the Mims ENDOR pulse sequence²⁶ $\pi/2-\tau-\pi/2-\tau_{\pi,RF}-\pi/2-\tau$ -echo. The $\pi/2$ microwave pulse length used was 12 ns, and the RF pulse length of $\tau_{\pi,RF} = 30$ μ s for all measurements at full power (100 W amplifier). ENDOR line shapes were calculated using *Easyspin* ENDOR *saffron* simulations.

3.6 References

1. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C. C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Zemgulyte, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlín, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Zidek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
2. Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100. <https://doi.org/10.1038/s41586-023-06415-8>
3. Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature* **2007**, *450*, 964–972. <https://doi.org/10.1038/nature06522>
4. Pritchard, R. B.; Hansen, D. F. Characterising side chains in large proteins by protonless (13)C-detected NMR spectroscopy. *Nat. Commun.* **2019**, *10*, 1747. <https://doi.org/10.1038/s41467-019-09743-4>
5. Dahiyat, B. I.; Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10172–10177. <https://doi.org/10.1073/pnas.94.19.10172>
6. Maguire, J. B.; Boyken, S. E.; Baker, D.; Kuhlman, B. Rapid sampling of hydrogen bond networks for computational protein design. *J. Chem. Theory Comput.* **2018**, *14*, 2751–2760. <https://doi.org/10.1021/acs.jctc.8b00033>
7. Fecker, T.; Galaz-Davison, P.; Engelberger, F.; Narui, Y.; Sotomayor, M.; Parra, L. P.; Ramirez-Sarmiento, C. A. Active site flexibility as a hallmark for efficient PET degradation by *I. sakaiensis* PETase. *Biophys. J.* **2018**, *114*, 1302–1312. <https://doi.org/10.1016/j.bpj.2018.02.005>
8. Shapovalov, M. V.; Dunbrack, R. L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19*, 844–858. <https://doi.org/10.1016/j.str.2011.03.019>
9. Hintze, B. J.; Lewis, S. M.; Richardson, J. S.; Richardson, D. C. MolProbity's ultimate rotamer-library distributions for model validation. *Proteins* **2016**, *84*, 1177–1189. <https://doi.org/10.1002/prot.25039>
10. Burnley, B. T.; Afonine, P. V.; Adams, P. D.; Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **2012**, *1*, e00311. <https://doi.org/10.7554/eLife.00311>
11. Wider, G. Technical aspects of NMR spectroscopy with biological macromolecules and studies of hydration in solution. *Prog. Nucl. Magn. Reson. Spectrosc.* **1998**, *32*, 193–275. [https://doi.org/10.1016/S0079-6565\(98\)00014-4](https://doi.org/10.1016/S0079-6565(98)00014-4)
12. Judd, M.; Qi, M.; Abdelkader, E. H.; Qianzhu, H.; Savitsky, A.; Huber, T.; Harmer, J. R.; Godt, A.; Otting, G.; Cox, N. Measuring nanometer distances in proteins and rigid rulers between ¹⁹F and Gd³⁺ by integration of ¹⁹F-ENDOR signal intensities. *J. Am. Chem. Soc.* **2025**, *147*, 16826–16835. <https://doi.org/10.1021/jacs.4c13530>
13. Seal, M.; Zhu, W.; Dalaloyan, A.; Feintuch, A.; Bogdanov, A.; Frydman, V.; Su, X. C.; Gronenborn, A. M.; Goldfarb, D. Gd^{III}-¹⁹F distance measurements for proteins in cells by electron-nuclear double resonance. *Angew. Chem., Int. Ed.* **2023**, *62*, e202218780. <https://doi.org/10.1002/anie.202218780>
14. Judd, M.; Abdelkader, E. H.; Qianzhu, H.; Habel, E.; Lindsay, C.; Brancewicz, A.; Huber, T.; Otting, G.; Cox, N. Characterising side chain conformations in proteins using ¹⁹F-ENDOR and new fluorinated amino acid spin labels. *Chem. Eur. J.* **2025**, e02647. <https://doi.org/10.1002/chem.202502647>
15. Horrocks, W. D.; Peter Snyder, A. Measurement of distance between fluorescent amino acid residues and metal ion binding sites. Quantitation of energy transfer between tryptophan and terbium(III) or europium(III) in thermolysin. *Biochem. Biophys. Res. Commun.* **1981**, *100*, 111–117. [https://doi.org/10.1016/S0006-291X\(81\)80070-8](https://doi.org/10.1016/S0006-291X(81)80070-8)
16. Horrocks, W. D., Jr.; Collier, W. E. Lanthanide ion luminescence probes. Measurement of distance between intrinsic protein fluorophores and bound metal ions: quantitation of energy transfer between

- tryptophan and terbium(III) or europium(III) in the calcium-binding protein parvalbumin. *J. Am. Chem. Soc.* **1981**, *103*, 2856–2862. <https://doi.org/10.1021/ja00400a061>
17. Stryer, L.; Haugland, R. P. Energy transfer: a spectroscopic ruler. *Proc. Natl. Acad. Sci. U.S.A.* **1967**, *58*, 719–726. <https://doi.org/10.1073/pnas.58.2.719>
 18. Qianzhu, H.; Abdelkader, E. H.; Otting, G.; Huber, T. Genetic encoding of fluoro-L-tryptophans for site-specific detection of conformational heterogeneity in proteins by NMR spectroscopy. *J. Am. Chem. Soc.* **2024**, *146*, 13641–13650. <https://doi.org/10.1021/jacs.4c03743>
 19. Treiber-Kleinke, C.; Berger, A. A.; Adrian, L.; Budisa, N.; Koks, B. *Escherichia coli* adapts metabolically to 6- and 7-fluoroindole, enabling proteome-wide fluorotryptophan substitution. *Front. Synth. Biol.* **2023**, *1*, 1345634. <https://doi.org/10.3389/fsybi.2023.1345634>
 20. Reiter, D.; Mutlu, F.; Ebert, D.; Noutio, M.; Heberle, J.; Schubert, M.; Koks, B. Investigating the site-specific impact of fluorine substitution on aromatic interactions in a tryptophan zipper peptide. *Chem. Eur. J.* **2025**, *31*, e202501263. <https://doi.org/10.1002/chem.202501263>
 21. Qianzhu, H.; Abdelkader, E. H.; Herath, I. D.; Otting, G.; Huber, T. Site-specific incorporation of 7-fluoro-L-tryptophan into proteins by genetic encoding to monitor ligand binding by ^{19}F NMR spectroscopy. *ACS Sens.* **2022**, *7*, 44–49. <https://doi.org/10.1021/acssensors.1c02467>
 22. Pan, B.; Guo, C.; Liu, D.; Wüthrich, K. Fluorine-19 labeling of the tryptophan residues in the G protein-coupled receptor NK1R using the 5-fluoroindole precursor in *Pichia pastoris* expression. *J. Biomol. NMR* **2024**, *78*, 133–138. <https://doi.org/10.1007/s10858-024-00439-6>
 23. Pribitzer, S.; Mannikko, D.; Stoll, S. Determining electron-nucleus distances and Fermi contact couplings from ENDOR spectra. *Phys. Chem. Chem. Phys.* **2021**, *23*, 8326–8335. <https://doi.org/10.1039/d1cp00229e>
 24. Kehl, A.; Sielaff, L.; Rimmel, L.; Ramisch, M. L.; Bennati, M.; Meyer, A. Frequency and time domain ^{19}F ENDOR spectroscopy: role of nuclear dipolar couplings to determine distance distributions. *Phys. Chem. Chem. Phys.* **2025**, *27*, 1415–1425. <https://doi.org/10.1039/d4cp04443f>
 25. Mukai, T.; Hoshi, H.; Ohtake, K.; Takahashi, M.; Yamaguchi, A.; Hayashi, A.; Yokoyama, S.; Sakamoto, K. Highly reproductive *Escherichia coli* cells with no specific assignment to the UAG codon. *Sci. Rep.* **2015**, *5*, 9699. <https://doi.org/10.1038/srep09699>
 26. Mims, W. B. Pulsed ENDOR experiments. *Proc. R. Soc. Lond. Ser. Math. Phys. Sci.* **1965**, *283*, 452–457. <https://doi.org/10.1098/rspa.1965.0034>

Chapter 4: Structural accommodation of 4-substituted tryptophans in maltose-binding protein

4.1 4-substituted tryptophans

Tryptophan (Trp) is the largest canonical amino acid, containing a bicyclic indole sidechain with a conjugated π -system, also making it the brightest fluorescent canonical amino acid. The hydrophobic bicyclic aromatic ring favours burial within protein cores, driving protein folding and stability. The indole nitrogen is essentially non-ionisable in physiological pH,¹ acting as a hydrogen bond donor for specific polar interactions. Unlike tyrosine (Tyr) and phenylalanine (Phe), Trp is asymmetrical when rotated 180° around the C $_{\beta}$ -C $_{\gamma}$ bond, doubling the distinguishable rotameric states relative to Tyr and Phe. The extensive π -electron surface participates in unique non-covalent interactions (cation- π and π - π stacking) making Trp commonly observed at protein-protein interfaces and in protein-ligand complexes.^{2,3} While many Trp positions are functionally restrained, some tolerate non-aromatic residues such as leucine (Leu) or methionine (Met) without loss of function.⁴ Trp's intrinsic fluorescence—sensitive to local environment polarity—can be used to monitor protein folding, conformational changes and ligand binding.^{5–7} However, its ultraviolet (UV) absorption (~280 nm) and emission (~350 nm), which are damaging to biological systems at high dosage,^{8,9} motivates the incorporation of isosteric analogues with better photophysical properties.

Fluorinated Trp analogues are powerful probes for Nuclear Magnetic Resonance (NMR) spectroscopy,^{10–13} and are among the most subtle variations to the Trp sidechain. The ¹⁹F nucleus has a spin $I = -\frac{1}{2}$, is 100% naturally abundant, and the ¹⁹F chemical shift is extremely sensitive to its local environment, with a chemical shift dispersion of over 200 ppm in proteins.¹⁴ This makes ¹⁹F a powerful reporter of protein conformation, dynamics and ligand binding, enabling studies in both solution and solid states. The van der Waals radius of fluorine is less than 0.3 Å larger than that of hydrogen (~1.47 Å for fluorine compared to ~1.20 Å for hydrogen), and the C–F bond length is only marginally longer than the C–H bond (~1.35 Å, compared to ~1.09 Å).^{15,16} Furthermore, the fluorine nucleus has a greater X-ray scattering coefficient compared to hydrogens, meaning that the position of the fluorine and the precise rotamer of the sidechain clearer in X-ray crystallography.^{17,18} The substitution position of the Trp has subtle differences in the

electrostatics and sterics of the sidechain conformer preferences,^{10,19,20} of particular interest is the 4-position, which points toward the backbone, and thermal unfolding experiments found that 4-fluoro-tryptophan (4FW) did not change the T_m of a tryptophan zipper peptide, but did alter the energy of its unfolding intermediate.²¹

Minor modifications to the Trp sidechain provide enhanced photophysical properties for fluorescence applications. Early analogues focused on analogues like 7-azatryptophan,^{22–24} but recent work has shown 4-cyanotryptophan (4CNW) is an exceptionally bright fluorescent amino acid.²⁵ The cyano substituent red-shifts excitation (up to 355 nm) and emission (up to 405 nm) into the near-visible spectrum where cellular autofluorescence and phototoxicity are greatly reduced.⁸ 4CNW has a remarkably high quantum yield ($QY > 0.8$ in water), a long fluorescence lifetime (~ 13.7 ns) and excellent photostability.^{25,26} These properties enable single molecule spectroscopy, protein localisation in microscopy and as a fluorescent reporter for ligand binding at nanomolar concentrations.^{27–29} However, compared to 4FW, the cyano substituent is a more substantial steric perturbation, extending approximately 2.60 Å (Figure 4.1C) away from the $C_{\epsilon 3}$ of the indole ring, compared to the extension of the fluorine atom of 1.35 Å (Figure 4.1B) Like fluorine, the cyano group extends perpendicular to the plane of the indole side chain, due to its linear geometry.

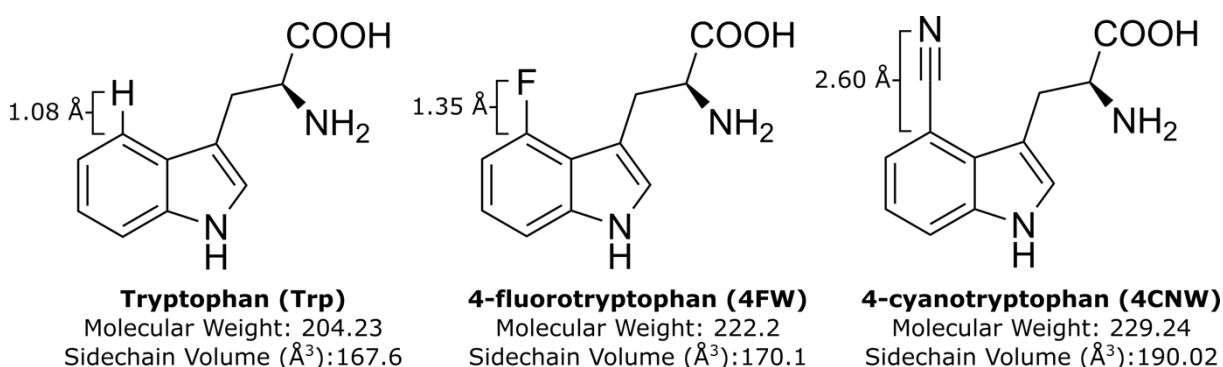


Figure 4.1 – Chemical structure of tryptophan, 4-fluorotryptophan and 4-cyanotryptophan, annotated with molecular weight and sidechain volume. The electron density of each sidechain was calculated with Orca³⁰ using the BL3YP basis set, the volumes were measured at a level of 0.002 a.u. to approximate the molecular envelope.³¹

A foundational assumption underpinning the use of ncAAs as biophysical probes is that these substitutions minimally perturb the structure in question, allowing the probe to report faithfully on native properties and dynamics. This assumption is typically supported by retained biological activity or low global backbone root-mean-square deviation (RMSD) between modified and wild-type (WT) structures. However, global metrics mask important local rearrangements. A sidechain can undergo a major conformational change, such as a rotameric flip, with negligible impact on the C_α trace and deceptively low RMSD. Crystallographic studies on fluorine-substituted hydrophobic sidechains shown that subtle accommodations are made by sidechain repacking,^{17,18} in order to accept this extra bulk, which extends even to trifluoromethylated Phe and Tyr.³² While these changes are invisible in the C_α trace, variations in sidechain dynamics can have significant implications to the properties of a protein.^{33–36} Despite the popularity of substituted Trp analogues as structural probes, a more detailed atomic analysis of the local accommodations required to incorporate them has been lacking. This chapter addresses that gap by leveraging high-resolution X-ray crystallography to dissect the structural consequences of substituting Trp with 4FW and 4CNW in three distinct local environments within a single, well characterised protein scaffold. By moving beyond global metrics, this work defines the precise local rules governing structural accommodation.

4.2 Maltose-binding protein

Maltose-binding protein (MBP) from *Escherichia coli* (*E. coli*) was selected to systematically investigate the structural impact of ncAA incorporation. MBP is a monomeric, ~40 kDa protein that is highly soluble, expresses to high levels, and is exceptionally well characterised, with hundreds of crystal structures available in the PDB.^{37–40} This provides a reliable baseline for its natural conformational variability, against which perturbations can be rigorously assessed. *E. coli* MBP has eight tryptophans that have been extensively used to monitor the dynamics of folding and ligand binding.⁴¹

Three native tryptophan residues showing different local environments were selected for substitution: positions 10, 62 and 340. Trp10 is located in a packed, buried region in the hydrophobic core. Trp62 is in a moderately buried environment atop a β -strand. Trp340 is positioned on the exposed surface of an α -helix, near the maltose-binding site. These sites represent distinct local environments in terms of solvent exposure, packing density, and secondary structure, providing a comprehensive testbed for assessing the rules of structural accommodation.

4.2.1 *Generating the MBP rotamer dataset*

A dataset of crystal structures was obtained from the PDB using EMBL-EBI's PDBeFold with the MBP WT crystal structure as query, using a Q-score threshold of 70%. Unique entries were downloaded using the PDB batch download module. A clean dataset of structures was generated by creating new PDB files for each chain with pairwise sequence identity $\geq 40\%$, yielding 678 unique structures. Each structure was annotated for the presence of molecule ID "GLC" (glucose, two of which are modelled to represent maltose). Only X-ray crystal structures were included for analysis.

Custom python scripts using the Biopython library computed sidechain dihedral angles (χ_1 and χ_2) for each residue of interest. Global pairwise sequence alignment between the WT reference sequence and each PDB chain facilitated mapping of the reference residue to its equivalent position in each structure and excluding mutations at that position. Due to this filter, the number of models analysed for each position varies. The final number of rotamers from the full rotamer dataset represents matching rotamers with matching residues at the designated position and is detailed in the figure captions. For most residues, both χ_1 and χ_2 angles were calculated and recorded. For proline (Pro), where its sidechain cyclised with the backbone, only the χ_1 was measured to describe ring pucker conformation. The resulting dihedral angle data was then used for comparative analysis. To refer to Trp rotamers, the following uses the naming convention established by Richardson and co-workers that is internally consistent, and the first initial refers to the χ_1 angle (**t** for trans, **m** for -60° , **p** for $+60^\circ$), and the χ_2 angle is the population mode for

that class of rotamer, the χ_2 range describes which values of χ_2 are accepted as that rotamer (Table 4.1).⁴²

Table 4.1 - Rotamer nomenclature for Tryptophan, adapted from Lovell *et. al.*⁴²

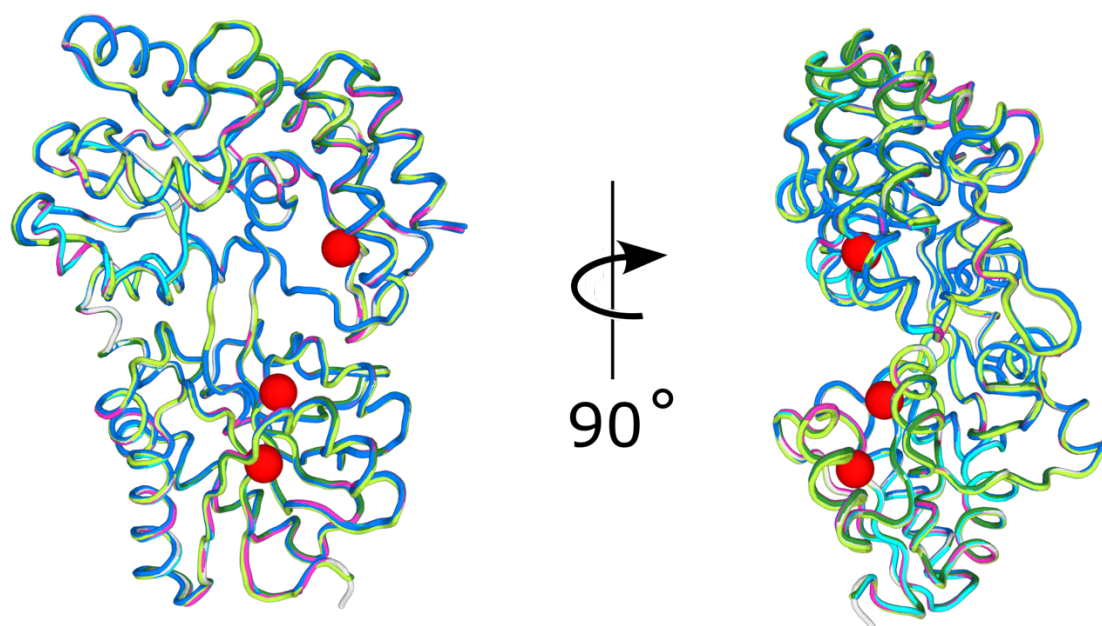
	χ_1	χ_2	χ_2 range
p -90°	+60°	-90°	-130° to -60°
p 90°	+60°	+90°	+60° to +130°
t -105°	180°	-105°	-130° to -60°
t 90°	180°	+90°	0° to +100°
m -90°	-65°	-90°	-130° to -60°
m 0°	-65°	-5°	-40° to 20°
m 95°	-65°	95°	60° to 130°

4.3 Homogeneity of global folds

Crystal structures were obtained for MBP variants containing single substitutions of 4FW or 4CNW at positions 10, 62 (4FW only), and 340, as well as the WT protein. All crystals belong to the $P 1 2_1 1$ space group, with similar unit cell dimensions, and all structures represent the apo, “open” conformation of MBP. The MBP WT structure has an alignment RMSD of 0.424 Å to the reference *E. coli* MBP crystal structure from Quijcho and co-workers.⁴³ An RMSD of ~0.4 Å is typical for the same apo-protein in different crystallisation conditions, at high resolution.^{44,45} All variant crystal structures showed pairwise backbone RMSD to WT of <0.3 Å, demonstrating that the proteins are unchanged at the global fold level and validating the use of these analogues as probes that do not fundamentally alter the protein’s overall architecture. All variant crystal structures showed sufficient resolution to analyse sidechain variability. Table 4.2 summarises resolution, B -factors and R_{free} values of the variants. Full crystallographic statistics are provided in Tables 4.3–4.5.

Table 4.2 - Summary of resolution, alignment RMSD to WT, R_{free} and B -factors of protein atoms for the crystal structures in this chapter.

	WT	W10 4FW	W10 4CNW	W62 4FW	W340 4FW	W340 4CNW
PDB ID	9CLD	9PQK	9CLC	9PQL	9PQN	9PQP
C_{α} RMSD (Å)	-	0.232	0.117	0.096	0.254	0.090
Resolution	1.58 Å	1.87 Å	1.48 Å	1.58 Å	1.07 Å	1.37 Å
R_{free}	0.1763	0.2668	0.1657	0.2016	0.1596	0.214
B-factors	25.0	30.0	23.0	17.2	17.9	27.3

**Figure 4.2** – Superimposition of all MBP variants with WT. Backbone geometry is illustrated in the ribbon representation. C_{β} atoms of Trp10, Trp62 and Trp340 in MBP WT shown as red spheres. Chains are coloured as follows: WT: white, W10 4FW: light blue, W10 4CNW: dark blue, W62 4FW: pink, W340 4FW: light green. W340 4CNW: dark green.

4.4 Trp10 substitution

The Trp10 residue is buried within the hydrophobic core of MBP, providing a stringent test of the protein's ability to accommodate the increased steric bulk. Examination of the crystal structures of WT, W10 4FW and W10 4CNW reveals that the indole ring, and the neighbouring Pro40 and Ile60 maintain essentially the same orientation (Figure 4.3). Electron density is well-defined for all three structures, with clear density visible for the 4-fluoro and 4-cyano substituents in the respective mutants (Figure 4.3B,C)

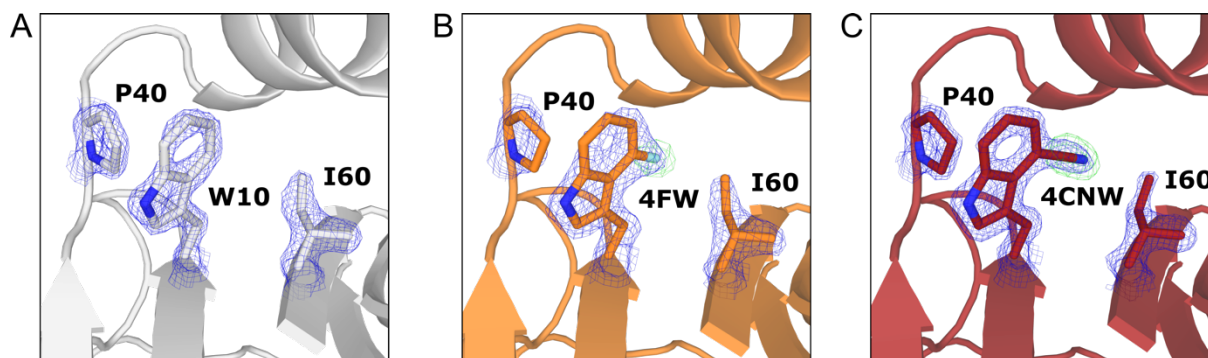


Figure 4.3 – Position W10, WT, 4FW and 4CNW crystal structures. $2mF_o-DF_c$ density shown as blue mesh contoured to 1.8σ . Residues 10, 47 and 60 annotated and shown as a stick illustration. (A) WT crystal structure 9CLD. (B) W10 4FW crystal structure 9PQK, mF_o-DF_c density shown as green mesh contoured to 4σ . Difference map generated from identical structure, without the 4-F atom. (C) W10 4CNW crystal structure 9CLC, mF_o-DF_c density shown as green mesh contoured to 4σ . Difference map generated from identical structure without the 4-CN substituent atoms.

Quantitative analysis of the sidechain dihedral angles (χ_1/χ_2) reveals a more nuanced story. In all three structures, the sidechain adopts the **m0°** ($\chi_1 \approx -90^\circ$, $\chi_2 \approx 5^\circ$) with its 4-position facing away from the backbone, the 4FW and 4CNW sidechains rotate clockwise less than 10° . This is not an outlier in either the Dunbrack⁴⁶ or Top8000⁴⁷ libraries for Trp. Figure 4.4A shows that this Trp rotamer is well represented in the MBP rotamer dataset. The ability of the protein to maintain the canonical rotamer even with the bulkier cyano group indicates a remarkable degree of local plasticity. This accommodation is achieved not by a change in the substituted sidechain's conformation, but through subtle, distributed shifts in the positions of neighbouring residues, primarily Ile60 and Pro40. Between WT and 4FW, the Ile60 sidechain remains static, however, in the 4CNW structure the χ_2 rotates clockwise by $\sim 20^\circ$, which directly moves it away from the cyano nitrogen. Inspecting the rotamer distribution in the MBP rotamer dataset, this rotation is within the magnitude of variation that occurs between homologous crystal structures (Figure 4.4B), suggesting this is an acceptable accommodation.

In both 4FW and 4CNW, Pro40 changes its pucker conformation from *endo* (WT) to *exo* (Figure 4.4C). This change is observed within the MBP rotamer dataset, suggesting that proline pucker inversion is one mechanism by which proteins can accommodate bulkier substituents. Pro40 is positioned facing the planar indole ring of Trp10, rather than directly toward the 4-position. Proline pucker is primarily influenced by backbone

dihedral angles, there is generally a low energy barrier between the two conformations.^{48,49} The slight counter-clockwise rotation of both the 4FW and 4CNW in this position, brings the indole towards the proline residue by $\sim 0.3\text{\AA}$. This shift may provide sufficient steric pressure to invert the Pro40 pucker, as the *exo* conformation moves the C_γ atom and its hydrogen substituents further from the indole.

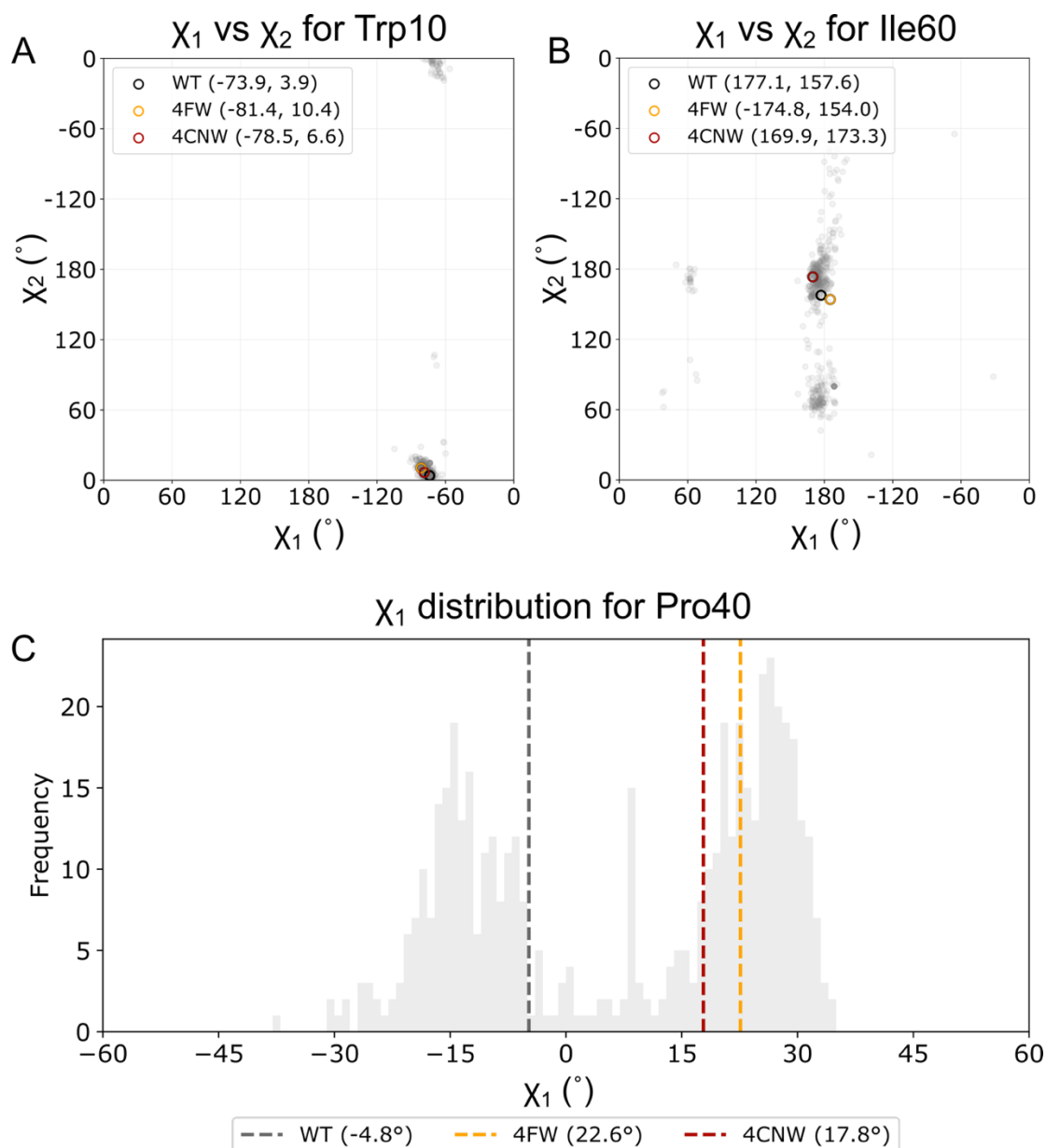


Figure 4.4 – χ_1/χ_2 plots of positions Trp10, Phe47 and Phe60 in MBP rotamer dataset. Each plot contains the χ_1/χ_2 distribution for the annotated residue in the rotamer dataset. Corresponding χ angles from the crystal structures in this study are marked on each plot and detailed in the legend. (A) χ_1/χ_2 distribution of Trp10, containing 525 datapoints from the full rotamer dataset. (B) χ_1/χ_2 distribution of Ile60, containing

510 datapoints from the full rotamer dataset. (C) χ_1/χ_2 distribution of Pro40, containing 504 datapoints from the full rotamer dataset. $\chi_1 < 0^\circ$ are representative of *endo* proline puckers, and $\chi_1 > 0^\circ$ are representative of *exo* proline puckers.

The mutations at position 10 in the buried core of MBP demonstrate that increased steric bulk in the 4-position of the indole sidechain can be accommodated through variations that are inherent to the protein. Both sidechains that exhibited perturbed dihedrals showed changes within the range of canonical variation in the MBP rotamer dataset.

4.5 Trp62 substitution

The Trp62 position, moderately buried atop a beta-strand, presents a different environmental challenge. The 4-position of the indole sidechain typically adopts the *t*-90° ($\chi_1 \approx 180^\circ$, $\chi_2 \approx -90^\circ$) rotamer and points between the backbone of the preceding residue (Phe61) and the backbone of the next β -strand in the β -sheet. The structure of W62 4FW was solved to 1.58 Å resolution and, like the W10 variants, shows little perturbation to the local or global structure (Figure 4.5B). The 4FW sidechain maintains the WT *t*-90° rotamer.

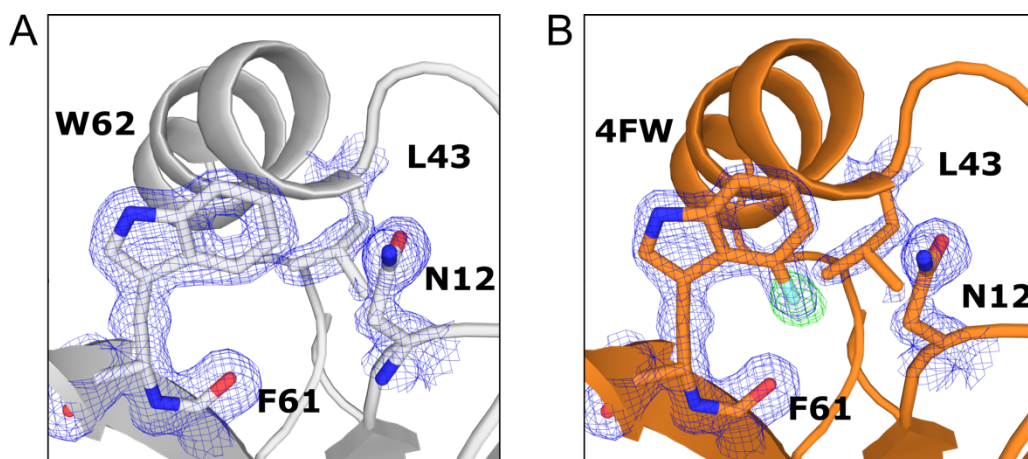


Figure 4.5 – Position W62, WT, 4FW crystal structures. 2mF_o-DF_c density shown as blue mesh contoured to 1.8 σ . Residues 10, 47 and 60 annotated and shown as a stick illustration. (A) WT crystal structure 9CLD. (B) W62 4FW crystal structure 9PQL, mF_o-DF_c density shown as green mesh contoured to 4 σ . Difference map generated from identical structure without the 4-F atom.

The fluorine makes van der Waals (vdW) contact with the backbone carbonyl oxygen of the preceding residue (Phe61), with an interatomic distance of 2.6 Å. The sum of vdW radii

is ~ 2.9 Å for fluorine and oxygen,¹⁵ making this a favourable contact that would become prohibitively repulsive if closer. This has been accommodated by a 12° rotation of the χ_1 in 4FW relative to the WT structure. Modelling the 4FW residue at position 62 in the WT model with χ_1 reverted to -157.5° reduces the F...O distance to 2.5 Å. However, this rotation is within the expected variation seen in the MBP rotamer dataset (Figure 4.6A). The reorientation of Leu43 correlates with the combined repositioning of the indole sidechain and the 4-fluoro substituent, as the rotation moves the fluorine toward the aliphatic sidechain. Figure 4.6B shows that the rotation in both χ_1 and χ_2 is slightly outside the main distribution peak for Leu43 rotamers. This rotation is accommodated by Asn12, which rotates clockwise to make space for the C_{δ2} methyl of Leu43 (Figure 4.6C).

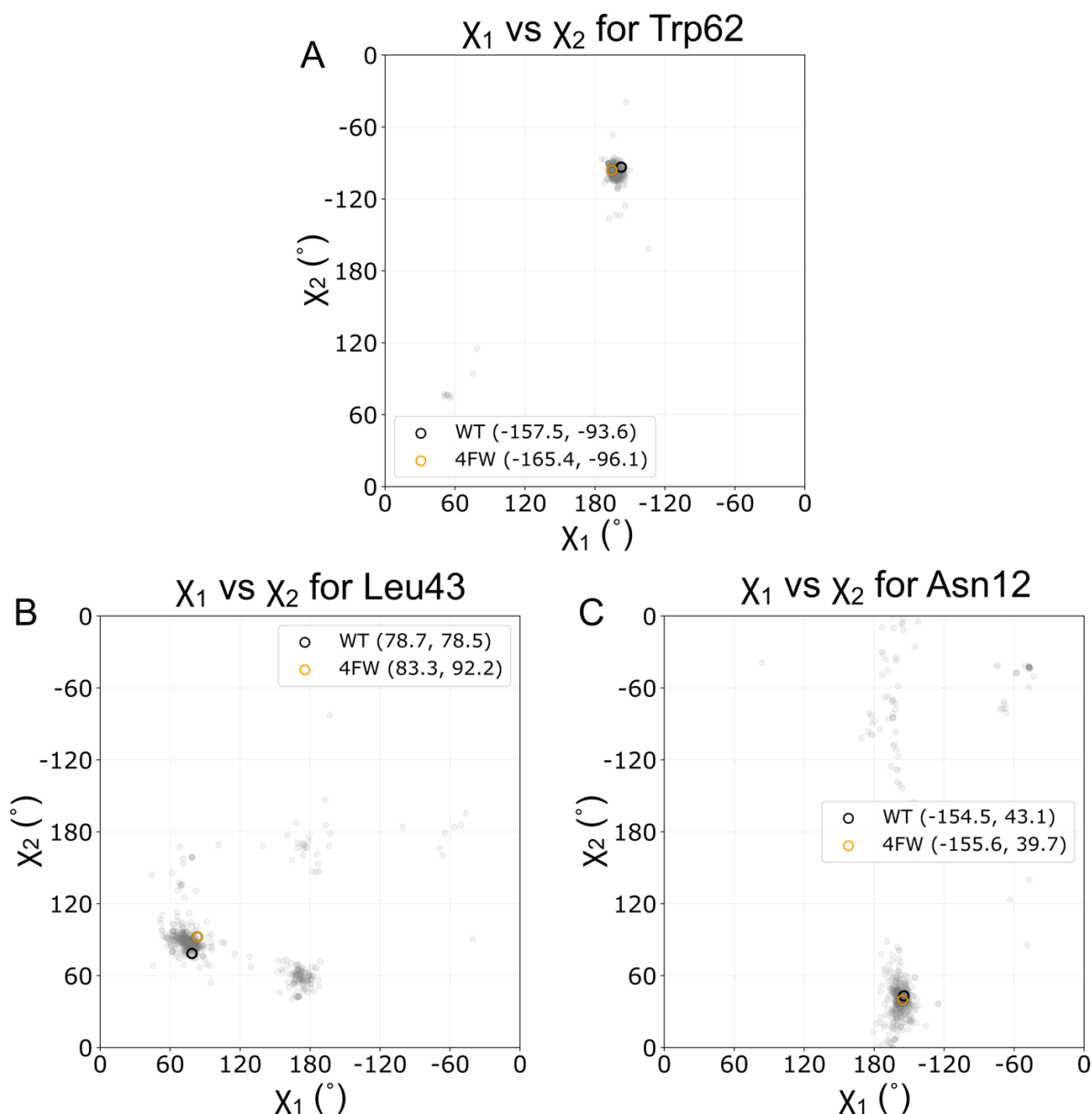


Figure 4.6 – χ_1/χ_2 plots of positions Trp62, Leu43 and Asn12 in MBP rotamer dataset. Each plot contains the χ_1/χ_2 distribution for the annotated residue in the rotamer dataset. Corresponding χ angles from the crystal structures in this study are marked on each plot and detailed in the legend. (A) χ_1/χ_2 distribution of Trp62, containing 500 datapoints from the full rotamer dataset. (B) χ_1/χ_2 distribution of Leu43, containing 528 datapoints from the full rotamer dataset. (C) χ_1/χ_2 distribution of Asn12, containing 541 datapoints from the full rotamer dataset.

As the 4-position of Trp62 faces rigid backbone atoms, the larger steric bulk of the cyano group at this position may not be accommodated without modifying nearby residues. Unlike in the W10 site, where the 4-position faces aliphatic sidechain residues, the protrusion at the W62 site faces backbone atoms. Therefore, there may not be sufficient malleable neighbouring residues to dissipate the steric strain through small, collective

movements. This hypothesis should be tested by biophysical measurements in future work.

4.6 Trp 340 substitution

The W340 position, located on an α -helix near the ligand-binding cleft, provides a third and uniquely informative environment. Here, the protein responds to the introduction of tryptophan analogues with distinct conformational changes. The WT Trp340 is in the **m90°** conformation ($\chi_1 \approx -60^\circ$, $\chi_2 \approx 90^\circ$), which is highly consistent in the MBP rotamer dataset. The structure of the W340 4FW variant, resolved at 1.07 Å resolution, reveals two distinct conformations of the sidechain (Figure 4.7B). The W340 4CNW variant shows a dramatic sidechain flip in both χ_1 and χ_2 (Figure 4.7C) suggesting that the increased steric bulk is not tolerated by the protein at this position.

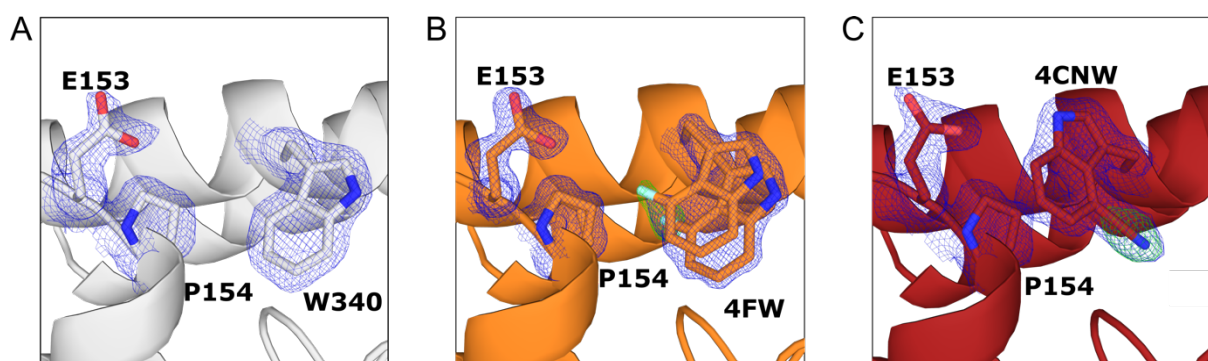


Figure 4.7 – Position W340, WT, 4FW, 4CNW crystal structures. $2mF_o-DF_c$ density shown as blue mesh contoured to 1.8 σ . Residues 340, 153 and 154 annotated and shown as a stick illustration. (A) WT crystal structure 9CLD. (B) W340 4FW crystal structure 9PQN, mF_o-DF_c density shown as green mesh contoured to 3.5 σ . Difference map generated from identical structure, without the 4-F atoms. (C) W340 4CNW crystal structure 9PQP, mF_o-DF_c density shown as green mesh contoured to 4 σ . Difference map generated from identical structure, without the 4-CN substituent atoms.

Comparing the observed rotamers to those in the MBP rotamer dataset, both 4FW conformers sample χ_1/χ_2 commonly represented within MBP crystal structures (Figure 4.8A). However, the **t-90°** rotamer of 4CNW ($\chi_1 \approx 180^\circ$, $\chi_2 \approx -90^\circ$) less frequently observed. The rotations for Glu153 of $\sim 20^\circ$ for χ_1 and $\sim 15^\circ$ for χ_2 are necessary adjustments, as the indole sidechain is now rotated towards Glu153 (Figure 4.8B). The pucker of Pro154 is largely conserved among the structures, weighted toward the *exo*

conformation in the MBP rotamer dataset. Conformer A of Pro154 would clash with conformer B of 4FW, thus conformer B of Pro154 adopts the *endo* pucker.

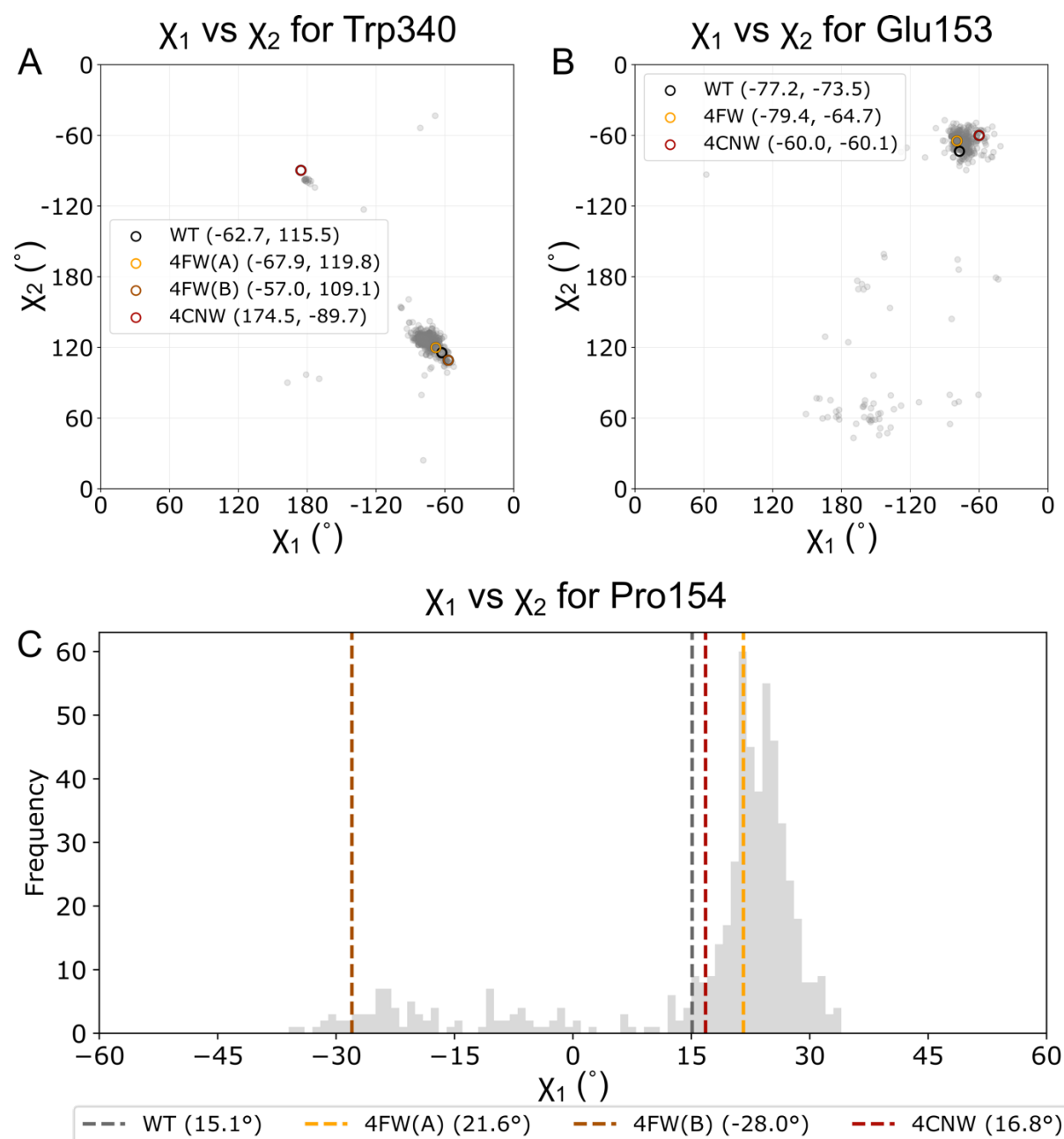


Figure 4.8 – χ_1/χ_2 plots of positions Trp340, Glu153 and Pro154 in MBP crystal structures. Each plot contains the χ_1/χ_2 distribution for the annotated residue in the rotamer dataset. Corresponding χ angles from the crystal structures in this study are marked on each plot and detailed in the legend. (A) χ_1/χ_2 distribution of Trp340, containing 550 datapoints from the full rotamer dataset. With corresponding χ angles from the crystal structures in this study. (B) χ_1/χ_2 distribution of Glu153, containing 524 datapoints from the full rotamer dataset. (C) χ_1/χ_2 distribution of Pro154, containing 543 datapoints from the full rotamer dataset. $\chi_1 < 0^\circ$ are representative of *endo* proline puckers, and $\chi_1 > 0^\circ$ are representative of *exo* proline puckers.

The two conformers in 4FW can be seen with clear electron density using omit maps (Figure 4.9) and were refined to 0.59 and 0.41 for conformer A and B, respectively. The maximum refined B-factors in the atoms of conformer A was 19.1, and 19.5 for conformer B. Omit maps were generated by omitting the atoms of one conformer while retaining the refined occupancy for the other. The $2mF_o-DF_c$ map contoured to 1.8σ in Figure 4.9A shows diffuse electron density over both conformers, with reduced density shown on the lower-population conformer at this contour level, particularly for the $C_{\epsilon 3}$ and 4-F atoms. Omit maps contoured to $3-4\sigma$ in Figure 4.9B and Figure 4.9C show density for conformer A and conformer B, respectively. The mF_o-F_c map for conformer B also shows lower density for atoms $C_{\epsilon 3}$ and 4-F at this contour level; however, the position of these atoms is well supported by strong density for the remaining sidechain atoms.

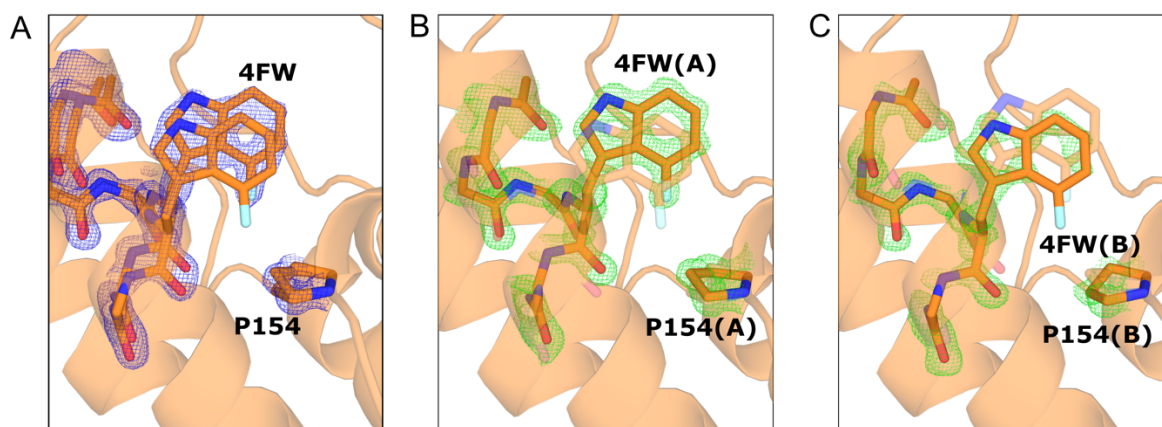


Figure 4.9 – WT and W340 4FW crystal structures, showing conformational variability around position 340. Cartoon representation shown at 50% opacity (A) MBP W340 4FW crystal structure with $2mF_o-DF_c$ density shown as a blue mesh contoured to 1.8σ . (B) W340 4FW crystal structure 9PQN. Conformer A is coloured opaque, conformer B is translucent. mF_o-DF_c density shown as a green mesh contoured to 4σ . Difference map generated from identical structure, with conformer B omitted. (C) W340 4FW crystal structure 9PQN. Conformer B is coloured opaque, conformer A is translucent. mF_o-DF_c density shown as a green mesh contoured to 3σ . Difference map generated from identical structure, with conformer A omitted.

The observation of dual conformations is coupled to the intrinsic conformation of the pyrrolidine ring of the adjacent Pro154. The occupancies of the surrounding residues were each independently refined for two conformers, yielding occupancies of 0.59 and 0.41, respectively, matching the values for 4FW. This indicates that this region samples at least two discrete states in the crystal lattice. The most notable differences between these conformers are illustrated in Figure 4.10. The $C=O\cdots H-N$ distance between the

carbonyl of Ser337 and the backbone amide of Tyr341 increases from 2.4 Å in conformer A to 4.0 Å in conformer B (Figure 4.10A,C), while the N–H···O=C bond angle contracts from 160.9° to 135.2°. The two conformers also show a translation of ~1 Å of the 4FW residue towards the C_γ atom of Pro154 (Figure 4.10B,D). In the most common proline conformation in the MBP rotamer dataset (*exo* pucker), which would cause a clash between the Pro154 H_{v2} and the 4-F atom in conformer B, as the modelled distance would be 1.5 Å. Instead, H_{v2} and 4-F in conformer B have a distance of 2.7 Å, the sum of their vdW radii. Both pucker states are sampled in the MBP rotamer dataset, confirming that this is a pre-existing mode of flexibility (Figure 4.8C). The modest steric perturbation introduced by the fluorine atom appears to be sufficient to shift the equilibrium, allowing the protein to populate an alternative state to relieve strain.

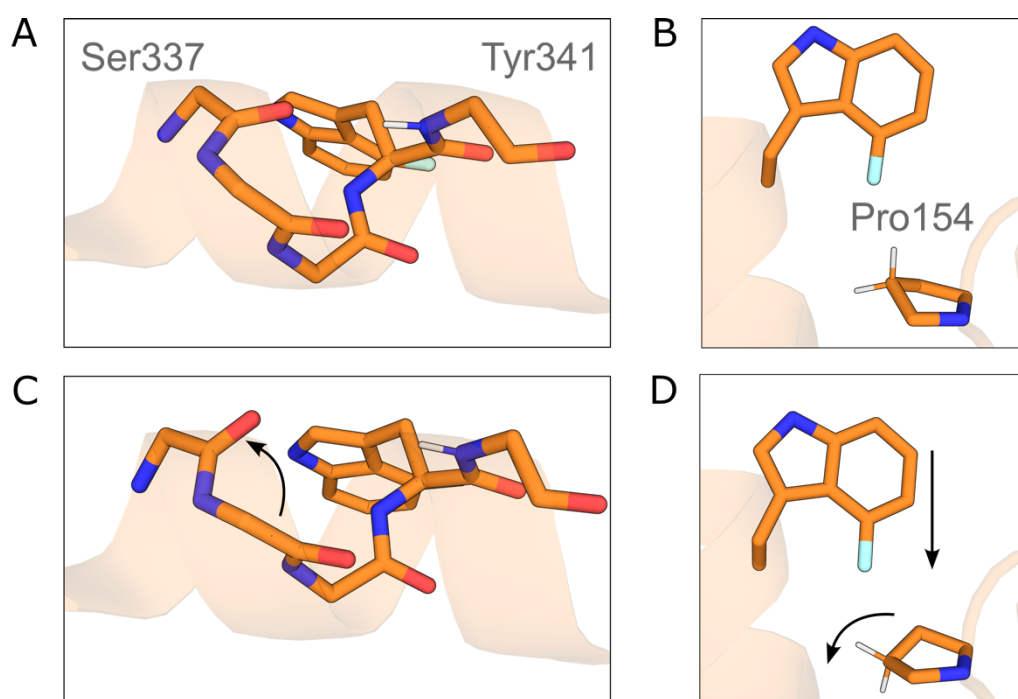


Figure 4.10 – Conformation differences between W340 4FW conformers A and B. (A, C) Backbone hydrogen bonding between Ser337 and Tyr341 differs between conformers. (B, D) Translation of 4FW toward Pro154 coupled with proline pucker change from *exo* (B) to *endo* (D). Cartoon representation at 70% transparency.

The bulkier cyano group elicits a far more dramatic response. The crystal structure of W340 4CNW shows a single, unambiguous conformation in which the sidechain has undergone a major rotameric flip from **m**90° to **t**–90°. Analysis of the MBP rotamer

dataset reveals that the $t90^\circ$ rotamer is never sampled by Trp340 in maltose-bound MBP structures (Figure 4.11), suggesting the $m90^\circ$ rotamer is important for binding site preorganisation. Superimposition of a holo MBP structure with similar χ_1/χ_2 to the MBP WT structure shows a favourable interaction between the WT Trp indole plane and bound maltose that would be lost in the 4CNW rotameric conformation (Figure 4.11B). Binding sites are preorganised to facilitate ligand binding, and this preorganisation is disrupted by the 4CNW substitution. This observed rotameric flip forces the sidechain into a conformation that is otherwise disfavoured for WT tryptophan at this position, indicating it is adopted by steric necessity rather than preference.

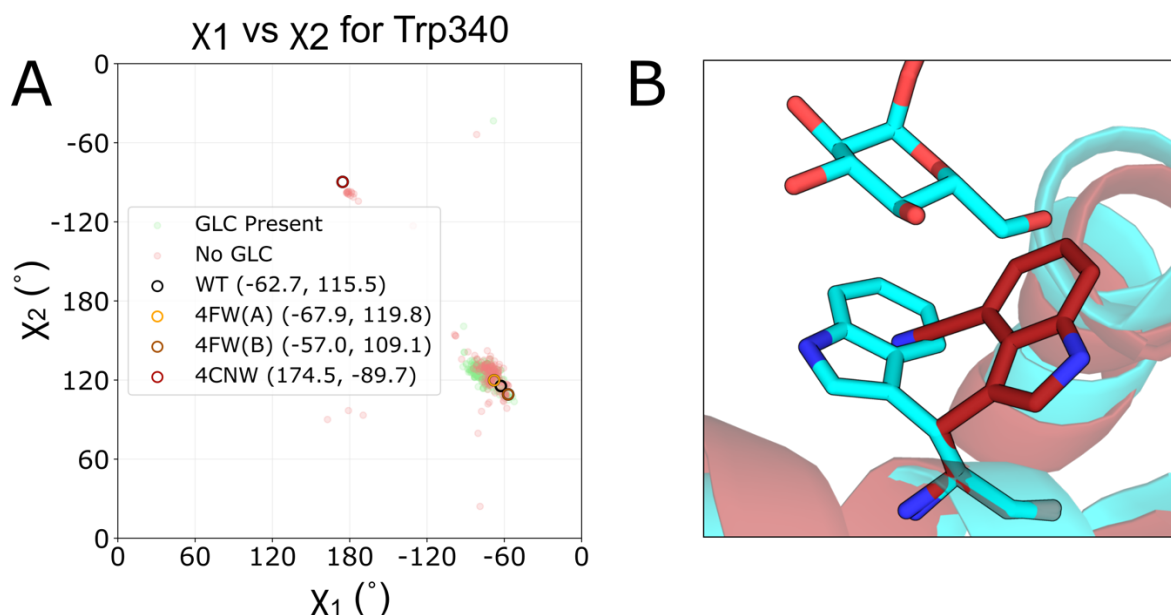


Figure 4.11 – Comparison of 340 4CNW with maltose bound MBP. (A) χ_1/χ_2 plots of positions W340 in MBP crystal structures, with (green) and without (red) GLC present. (B) Superimposition of maltose bound MBP (1DMB) coloured cyan with Trp340 and GLC in the stick representation. W340 4CNW coloured dark red with 4CNW in the stick representation. Cartoon representation is shown translucent.

To provide a mechanistic rationale for the observed rotameric flip in MBP 340 4CNW, a steric clash analysis was performed using Rosetta. The clash component of the standard Rosetta score function was used to compare crystal structure of MBP WT and MBP W340 4CNW. Custom residue parameters were generated for 4CNW, and four poses were created in PyRosetta: WT_Trp, WT_4CNW, 340_Trp and 340_4CNW, representing both crystal structures with both canonical tryptophan and 4CNW at position 340. The χ_1 and χ_2 angles were systematically rotated in 1° increments, and the clash component was

recorded at each position, yielding 360^2 values per pose. To generate the Δclash plot in Figure 4.12A, the clash component of WT_Trp was subtracted from WT_4CNW at matching χ_1/χ_2 pairs, revealing which conformations are forbidden in each model. Inspection of the WT model clash matrix shows that more conformations are forbidden with 4CNW than Trp in both models. More of the conformational space has significantly higher clash scores with 4CNW than with Trp, yet the canonical $m90^\circ$ rotamer is not the most disfavoured region. Nevertheless, experimental results demonstrate this clash difference is significant enough to cause the rotamer change. Analysis of the 340 model shows that placing a canonical tryptophan does not exclude the conformation observed in the 340 4CNW structure, indicating that the flipped rotamer is sterically permissible. However, this rotamer is rarely observed in the MBP rotamer dataset and not observed in the MBP WT structure, indicating that the $m90^\circ$ monomer occupies a local energy minimum. Taken together, these clash matrices suggest that 4CNW substitution introduces a steric barrier to the sidechain occupying the WT conformation, forcing adoption of an alternative, intrinsically less favourable but sterically viable conformation. The clash analysis doesn't prove this is the thermodynamically preferred state, only that it's sterically viable in the crystal structure.

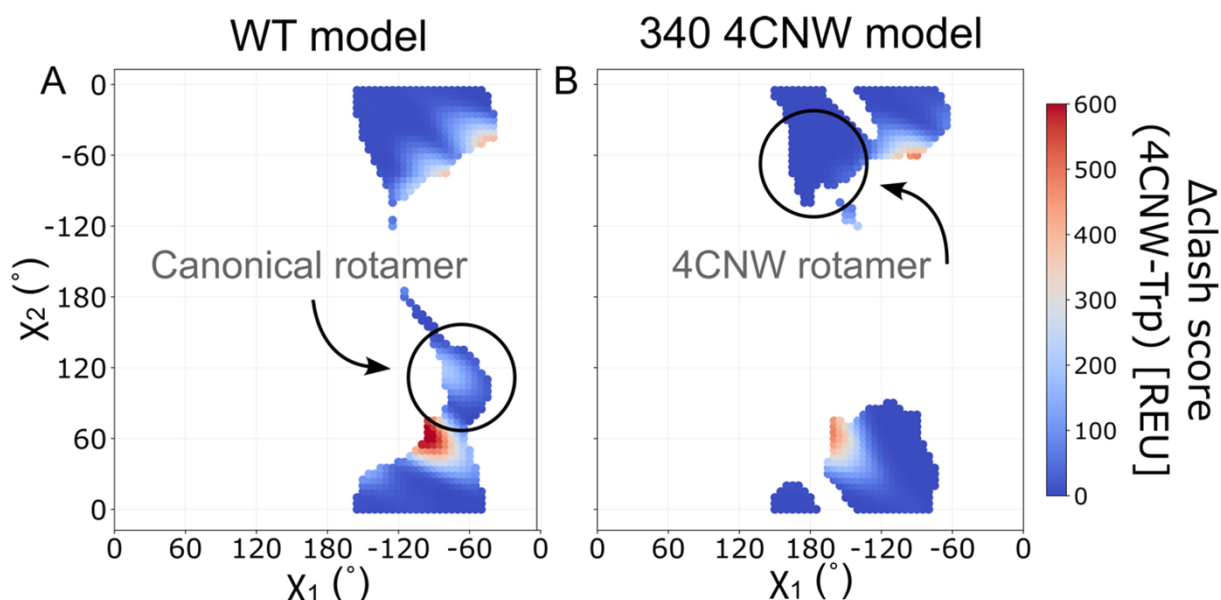


Figure 4.12 – Steric clash analysis of Trp and 4CNW at position 340. Δclash score heat maps showing the energetic penalty of 4CNW relative to Trp as a function of χ_1 and χ_2 angles. (A) WT model: canonical rotamer ($\chi_1 \approx -60^\circ$, $\chi_2 \approx 110^\circ$, circled) shows moderate clash penalty with 4CNW. (B) W340 4CNW model:

observed 4CNW rotamer ($\chi_1 \approx 180^\circ$, $\chi_2 \approx -90^\circ$, circled) is sterically permissible. Colour scale indicates Δclash score in Rosetta energy units (REU), with blue = low penalty, red = high penalty.

4.7 Conclusion

At position W10, collective sidechain movements accommodate both 4FW and 4CNW without rotamer change. W62 4FW is similarly tolerated through minor sidechain adjustments. At W340, 4FW inverts the Pro154 pucker while maintaining the canonical rotamer. W340 4CNW forces a rotameric flip from **m** 90° to **t** -90° , the nearest steric accommodation pathway available when local geometry cannot absorb the increased bulk of the cyano group.

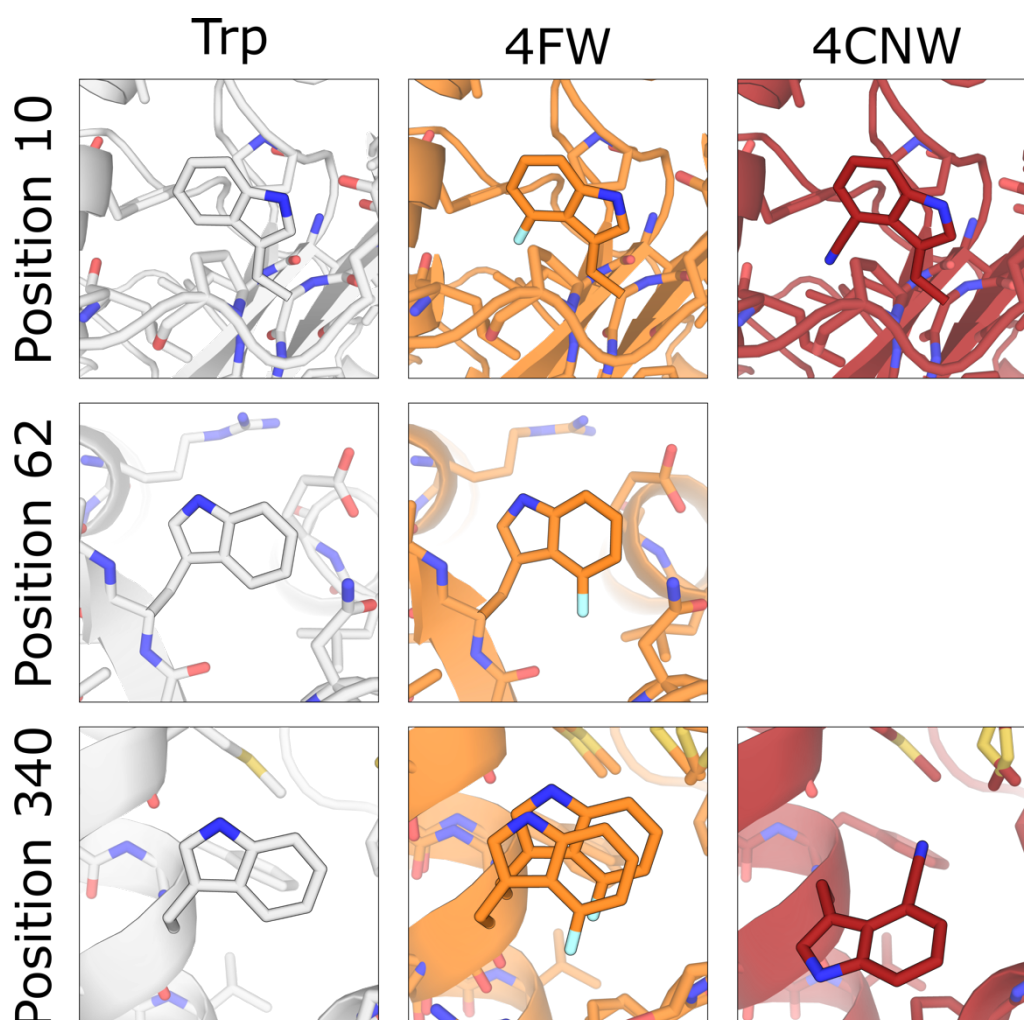


Figure 4.13 – Summary figure of tryptophan analogues in their local environments.

While the incorporation of 4-substituted tryptophans is indeed minimally perturbing at the global scale, the local structural consequences are highly context-dependent. Despite low global RMSD, this chapter illustrates distinct context-dependent accommodation mechanisms. Buried cores tolerate substitution through distributed sidechain movements: W10 maintains the canonical rotamer for both 4FW and 4CNW the subtle, correlated changes in Pro40 and Ile60. W340 4FW, with its 4-position facing towards Pro154, populates dual conformers coupled to Pro154 pucker conformations, and W340 4CNW is not accommodated as the native rotamer.

Before selecting a site for substitution, the local environment should be assessed for its capacity to tolerate the introduced steric bulk. Densely packed cores may accommodate bulk through distributed plasticity, as innocuous variations in the sidechains can comfortably accommodate the introduction. In position 62, where the 4-F extension makes vdW contact with the next residues carbonyl oxygen, crystals were unable to be obtained. Introduction of steric bulk that builds into the backbone of the protein is more likely to demand a greater conformational change. However, solvent-exposed sites may permit large rearrangements of the ncAA geometry relative to the substituted amino acid, if the protein cannot accommodate the substitution.

4.8 Methods

4.8.1 Cloning, expression and purification

MBP was cloned, expressed and purified by collaborators Elwy H. Abdelkader, Haocheng Qianzhu and Adarshi P. Welegedara. In short: *E. coli* B-95.ΔAΔ*fabR* cells were co-transformed with pRSF-4CNWRS or pRSF-4FWRS and pCDF plasmids containing amber codon-interrupted or wild-type MBP genes.⁵⁰ Cells were cultured in LB medium with 25 mg/L kanamycin and 25 mg/L spectinomycin, supplemented with 1 mM 4CNTrp or 4FW for mutant proteins. Expression was induced with 1 mM IPTG at OD₆₀₀ 0.6–1.0 and cells harvested after overnight growth at room temperature. Cell pellets were resuspended in buffer A (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 5% glycerol, 10 mM imidazole), lysed using an Avestin Emulsiflex C5 (10,000–15,000 psi, two passes), and clarified by

centrifugation (30,000 *g*, 60 min, 4°C). Lysate was loaded onto a 1 mL His GraviTrap column (Cytiva), washed with buffer B (buffer A with 20 mM imidazole), and eluted with buffer C (buffer A with 500 mM imidazole).

4.8.2 Size-exclusion chromatography

MBP WT was loaded onto a HiLoad 26/600 Superdex 75 pg column (Cytiva) equilibrated with buffer F (10 mM MES pH 6.2, 0.2% sodium azide). Fractions containing MBP WT were identified by SDS-PAGE, combined, and concentrated to 35 mg/mL using a 10 kDa MWCO Amicon ultrafiltration tube (Merck Millipore), then diluted to 15 mg/mL in buffer F. Following reverse IMAC, MBP-4CNW was dialyzed into buffer F and concentrated to 15 mg/mL using a 10 kDa MWCO Amicon ultrafiltration tube.

4.8.3 Crystallisation

Crystallisation screening was performed using Crystal Screen HT (Hampton Research) at 18°C with 1:1 protein:reservoir solution (0.5 µL total volume). MBP WT crystals grew within 36 hours in condition E12 (0.1 M cadmium chloride, 30% PEG 400, 0.1 M sodium acetate pH 4.6). Crystals were reproduced by hanging drop vapor diffusion (1 µL protein, 2 µL well solution, 500 µL reservoir). MBP-4CNW/4FW crystals were obtained by seeding: 3 µL MBP WT crystal drop was crushed in 50 µL condition E12 using a seed crusher, diluted to 550 µL, and serially diluted. Large cubic of all reported MBP-4CNW/4FW crystals grew within 36 hours using 10^{-4} – 10^{-6} dilution of seed stock. Crystals were cryoprotected with Parabar (Hampton Research) before flash-freezing in liquid nitrogen.

4.8.4 Structure determination and refinement

Diffraction data were collected at 100 K at the Australian Synchrotron MX2 beamline⁵¹ and processed using XDS⁵² in space group $P12_11$. Data were scaled in Aimless (CCP4).⁵³ Structures were solved by molecular replacement (Phaser MR, CCP4)⁵⁴ using PDB ID 1JW4 as the search model, followed by remodelling in Coot⁵⁵ and refinement in Phenix⁵⁶ and REFMAC.⁵⁷ Space group was validated using Zanuda.⁵⁸ For MBP-4FW structures, 4FW restraints from the CCP4 ligand library were used. For MBP-4CNW, custom CIF restraints

for 4-cyanotryptophan were generated by building the cyano-substituted tryptophan in PyMOL,⁵⁹ optimizing the ground state geometry in Orca using the B3LYP basis set,³⁰ and processing the resulting structure in eLBOW⁶⁰ to generate the CIF dictionary. Space group and metal geometry were validated using Zanuda⁵⁸ and the CheckMyMetal server,⁶¹ respectively. NcAAs were modelled using the replace residue function in Coot. Data collection and refinement statistics are in Table 4.3–5. Structures are deposited in the Protein Data Bank: 9CLD (MBP WT), 9CLC (MBP W10 4CNW), 9PQK (MBP W10 4FW), 9PQL (MBP W62 4FW), 9PQN (MBP W340 4CNW) and 9PQP (MBP W340 4FW).

4.9 Crystallographic statistics

Table 4.3 – Table of crystallographic statistics of MBP WT and MBP W10 4CNW.

	MBP WT	MBP W10 4CNW
PDB ID	9CLD	9CLC
Data collection		
Source	Australian Synchrotron MX2	Australian Synchrotron MX2
Space group	P 1 2 ₁ 1	P 1 2 ₁ 1
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	43.82 64.52 57.64	43.91 64.66 57.73
α , β , γ (°)	90.0 101.2 90.0	90.0 101.4 90.0
Resolution (Å)	42.99 – 1.58 (1.62 – 1.58)	42.56 – 1.48 (1.51 – 1.48)
<i>R</i> _{merge}	0.074 (0.933)	0.084 (0.851)
<i>R</i> _{pim}	0.037 (0.464)	0.048 (0.488)
<i>I</i> / σ <i>I</i>	10.6 (1.7)	10.0 (2.2)
CC _{1/2}	0.998 (0.635)	0.995 (0.464)
Completeness (%)	99.5 (99.3)	96.2 (94.4)
Redundancy	1.8 (1.8)	4.0 (4.0)
Refinement		
Resolution (Å)	42.99 – 1.58 (1.62 – 1.58)	42.56 – 1.48 (1.51 – 1.48)
No. reflections	42956 (2845)	50774 (2753)
<i>R</i> _{work} / <i>R</i> _{free}	0.1583 / 0.1763 (0.2373 / 0.2521)	0.1538 / 0.1657 (0.2639 / 0.2962)
Number of atoms		
Protein	2916	2928
Ligand/ion	88	107
Solvent	316	298
<i>B</i> -factors (overall)	25.0	23.0
Protein	22.3	21.0
Ligand/ion	48.7	48.9
Solvent	43.7	33.8
RMS deviations		
Bond lengths (Å)	0.007	0.006
Bond angles (°)	0.91	0.87

^aValues in parentheses are for the highest resolution shell.

Table 4.4 – Table of crystallographic statistics of MBP W10 4FW and MBP W62 4FW

	MBP W10 4FW	MBP W62 4FW
PDB ID	9PQK	9PQL
Data collection		
Source	Australian Synchrotron MX2	Australian Synchrotron MX2
Space group	P 1 2 ₁ 1	P 1 2 ₁ 1
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	43.96 64.97 57.63	43.84 64.62 57.77
α , β , γ (°)	90.0 101.0 90.0	90.0 101.4 90.0
Resolution (Å)	42.67 – 1.87 (1.94 – 1.87)	42.97 – 1.58 (1.62 – 1.58)
<i>R</i> _{merge}	0.226 (1.509)	0.107 (1.050)
<i>R</i> _{pim}	0.144 (0.919)	0.065 (0.658)
<i>I</i> / σ <i>I</i>	5.9 (1.5)	10.8 (1.8)
CC _{1/2}	0.994 (0.571)	0.999 (0.658)
Completeness (%)	99.6 (99.1)	99.6 (94.4)
Redundancy	6.7 (7.1)	7.2 (7)
Refinement		
Resolution (Å)	42.67 – 1.87 (1.94 – 1.87)	42.59 – 1.37 (1.39 – 1.37)
No. reflections	26253 (2868)	63758 (2808)
<i>R</i> _{work} / <i>R</i> _{free}	0.2099 / 0.2668 (0.3008 / 0.3433)	0.1611 / 0.2016 (0.3372 / 0.3441)
Number of atoms		
Protein	2886	2947
Ligand/ion	15	41
Solvent	127	406
<i>B</i> -factors (overall)	30.0	17.2
Protein	30.0	15.6
Ligand/ion	41.0	26.7
Solvent	30.8	27.9
RMS deviations		
Bond lengths (Å)	0.009	0.015
Bond angles (°)	0.97	1.80

^aValues in parentheses are for the highest resolution shell.

Table 4.5 – Table of crystallographic statistics of MBP W340 4CNW and MBP W340 4FW

	MBP W340 4CNW	MBP W340 4FW
PDB ID	9PQN	9PQP
Data collection		
Source	Australian Synchrotron MX2	Australian Synchrotron MX2
Space group	P 1 2 ₁ 1	P 1 2 ₁ 1
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	43.86 64.43 57.35	43.78 64.64 57.84
α , β , γ (°)	90.0 101.0 90.0	90.0 101.5 90.0
Resolution (Å)	43.05 – 1.37 (1.39 – 1.37)	42.62 – 1.07 (1.08 – 1.07)
<i>R</i> _{merge}	0.085 (2.528)	0.047 (0.804)
<i>R</i> _{pim}	0.053 (1.54)	0.029 (0.411)
<i>I</i> / σ <i>I</i>	10 (0.8)	16.9 (1.9)
CC _{1/2}	0.999 (0.556)	1.000 (0.556)
Completeness (%)	99.3 (98.2)	97.1 (88.5)
Redundancy	6.8 (5.5)	6.8 (5.5)
Refinement		
Resolution (Å)	43.05 – 1.37 (1.39 – 1.37)	42.62 – 1.07 (1.08 – 1.07)
No. reflections	65166 (2623)	63758 (2808)
<i>R</i> _{work} / <i>R</i> _{free}	0.1816 / 0.2140 (0.3545 / 0.3704)	0.1395 / 0.1596 (0.3016 / 0.3156)
Number of atoms		
Protein	2878	3221
Ligand/ion	60	37
Solvent	371	729
<i>B</i> -factors (overall)	27.3	17.9
Protein	25.6	14.0
Ligand/ion	44.0	40.3
Solvent	37.5	34.0
RMS deviations		
Bond lengths (Å)	0.007	0.005
Bond angles (°)	0.51	0.87

^aValues in parentheses are for the highest resolution shell.

4.10 References

1. Yagil, G. The proton dissociation constant of pyrrole, indole and related compounds. *Tetrahedron* **1967**, 23, 2855–2861. [https://doi.org/10.1016/0040-4020\(67\)85151-2](https://doi.org/10.1016/0040-4020(67)85151-2)
2. Gallivan, J. P.; Dougherty, D. A. Cation- π interactions in structural biology. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, 96, 9459–9464. <https://doi.org/10.1073/pnas.96.17.9459>
3. Crowley, P. B.; Golovin, A. Cation- π interactions in protein-protein interfaces. *Proteins* **2005**, 59, 231–239. <https://doi.org/10.1002/prot.20417>
4. Chapman, D.; Hochstrasser, R.; Millar, D.; Youderian, P. Engineering proteins without primary sequence tryptophan residues: Mutant trp repressors with aliphatic substitutions for tryptophan side chains. *Gene* **1995**, 163, 1–11. [https://doi.org/10.1016/0378-1119\(95\)00433-7](https://doi.org/10.1016/0378-1119(95)00433-7)
5. Gryczynski, I.; Wiczak, W.; Johnson, M. L.; Lakowicz, J. R. Lifetime distributions and anisotropy decays of indole fluorescence in cyclohexane/ethanol mixtures by frequency-domain fluorometry. *Biophys. Chem.* **1988**, 32, 173–185. [https://doi.org/10.1016/0301-4622\(88\)87005-4](https://doi.org/10.1016/0301-4622(88)87005-4)
6. Singh, P.; Chowdhury, P. K. Crowding-induced quenching of intrinsic tryptophans of serum albumins: A residue-level investigation of different conformations. *J. Phys. Chem. Lett.* **2013**, 4, 2610–2617. <https://doi.org/10.1021/jz401179z>
7. Fossum, C. J.; Johnson, B. O. V.; Golde, S. T.; Kielman, A. J.; Finke, B.; Smith, M. A.; Lowater, H. R.; Laatsch, B. F.; Bhattacharyya, S.; Hati, S. Insights into the mechanism of tryptophan fluorescence quenching due to synthetic crowding agents: A combined experimental and computational study. *ACS Omega* **2023**, 8, 44820–44830. <https://doi.org/10.1021/acsomega.3c06006>
8. Ludvikova, L.; Simon, E.; Deygas, M.; Panier, T.; Plamont, M. A.; Ollion, J.; Tebo, A.; Piel, M.; Jullien, L.; Robert, L.; Le Saux, T.; Espagne, A. Near-infrared co-illumination of fluorescent proteins reduces photobleaching and phototoxicity. *Nat. Biotechnol.* **2024**, 42, 872–876. <https://doi.org/10.1038/s41587-023-01893-7>
9. Debarre, D.; Olivier, N.; Supatto, W.; Beaupaire, E. Mitigating phototoxicity during multiphoton microscopy of live *Drosophila* embryos in the 1.0–1.2 μm wavelength range. *PLoS One* **2014**, 9, e104250. <https://doi.org/10.1371/journal.pone.0104250>
10. Qianzhu, H.; Abdelkader, E. H.; Otting, G.; Huber, T. Genetic encoding of fluoro-L-tryptophans for site-specific detection of conformational heterogeneity in proteins by NMR spectroscopy. *J. Am. Chem. Soc.* **2024**, 146, 13641–13650. <https://doi.org/10.1021/jacs.4c03743>
11. Qianzhu, H.; Abdelkader, E. H.; Herath, I. D.; Otting, G.; Huber, T. Site-specific incorporation of 7-fluoro-L-tryptophan into proteins by genetic encoding to monitor ligand binding by ^{19}F NMR spectroscopy. *ACS Sens.* **2022**, 7, 44–49. <https://doi.org/10.1021/acssensors.1c02467>
12. Toscano, G.; Rosati, M.; Barbieri, L.; Maier, K.; Banci, L.; Luchinat, E.; Konrat, R.; Lichtenecker, R. J. The synthesis of specifically isotope labelled fluorotryptophan and its use in mammalian cell-based protein expression for ^{19}F -NMR applications. *Chem. Commun.* **2024**, 60, 14188–14191. <https://doi.org/10.1039/d4cc04789c>
13. Pan, B.; Guo, C.; Liu, D.; Wüthrich, K. Fluorine-19 labeling of the tryptophan residues in the G protein-coupled receptor NK1R using the 5-fluoroindole precursor in *Pichia pastoris* expression. *J. Biomol. NMR* **2024**, 78, 133–138. <https://doi.org/10.1007/s10858-024-00439-6>
14. Chen, H.; Viel, S.; Ziarelli, F.; Peng, L. ^{19}F NMR: A valuable tool for studying biological events. *Chem. Soc. Rev.* **2013**, 42, 7971–7982. <https://doi.org/10.1039/c3cs60129c>
15. Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, 68, 441–451. <https://doi.org/10.1021/j100785a001>
16. Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1–S19. <https://doi.org/10.1039/P298700000S1>
17. Frkic, R. L.; Tan, Y. J.; Abdelkader, E. H.; Maleckis, A.; Tarcoveanu, E.; Nitsche, C.; Otting, G.; Jackson, C. J. Conformational preferences of the non-canonical amino acids (2S,4S)-5-fluoroleucine, (2S,4R)-5-fluoroleucine, and 5,5'-difluoroleucine in a protein. *Biochemistry* **2024**, 63, 1388–1394. <https://doi.org/10.1021/acs.biochem.4c00081>
18. Frkic, R. L.; Tan, Y. J.; Maleckis, A.; Chilton, N. F.; Otting, G.; Jackson, C. J. 1.3 Å crystal structure of *E. coli* peptidyl-prolyl isomerase B with uniform substitution of valine by (2S,3S)-4-fluorovaline reveals

- structure conservation and multiple staggered rotamers of CH₂F groups. *Biochemistry* **2024**, 63, 2602–2608. <https://doi.org/10.1021/acs.biochem.4c00345>
19. Ramirez-Cordero, B.; Traaseth, N. J. Quantification of membrane protein conformational free energy from mutations and a single atom. *J. Am. Chem. Soc.* **2025**, 147, 34316–34326. <https://doi.org/10.1021/jacs.5c04065>
20. Treiber-Kleinke, C.; Berger, A. A.; Adrian, L.; Budisa, N.; Koks, B. Escherichia coli adapts metabolically to 6- and 7-fluoroindole, enabling proteome-wide fluorotryptophan substitution. *Front. Synth. Biol.* **2023**, 1, 1345634. <https://doi.org/10.3389/fsybi.2023.1345634>
21. Reiter, D.; Mutlu, F.; Ebert, D.; Noutio, M.; Heberle, J.; Schubert, M.; Koks, B. Investigating the site-specific impact of fluorine substitution on aromatic interactions in a tryptophan zipper peptide. *Chem. Eur. J.* **2025**, 31, e202501263. <https://doi.org/10.1002/chem.202501263>
22. Abdelkader, E. H.; Qianzhu, H.; Huber, T.; Otting, G. Genetic encoding of 7-aza-L-tryptophan: Isoelectronic substitution of a single CH group in a protein for a nitrogen atom for site-selective isotope labeling. *ACS Sens.* **2023**, 8, 4402–4406. <https://doi.org/10.1021/acssensors.3c01904>
23. Rich, R. L.; Gai, F.; Lane, J. W.; Petrich, J. W.; Schwabacher, A. W. Using 7-azatryptophan to probe small molecule-protein interactions on the picosecond time scale: The complex of avidin and biotinylated 7-azatryptophan. *J. Am. Chem. Soc.* **1995**, 117, 733–739. <https://doi.org/10.1021/ja00107a016>
24. De Filippis, V.; De Boni, S.; De Dea, E.; Dalzoppo, D.; Grandi, C.; Fontana, A. Incorporation of the fluorescent amino acid 7-azatryptophan into the core domain 1-47 of hirudin as a probe of hirudin folding and thrombin recognition. *Protein Sci.* **2004**, 13, 1489–1502. <https://doi.org/10.1110/ps.03542104>
25. Yang, Y.; Feng, R.-r.; Gai, F. 4-cyanotryptophan as a sensitive fluorescence probe of local electric field of proteins. *J. Phys. Chem. B* **2023**, 127, 514–519. <https://doi.org/10.1021/acs.jpcc.2c07605>
26. Qianzhu, H.; Abdelkader, E. H.; Welegedara, A. P.; Habel, E.; Paul, N.; Frkic, R. L.; Jackson, C. J.; Huber, T.; Otting, G. Rendering proteins fluorescent inconspicuously: Genetically encoded 4-cyanotryptophan conserves their structure and enables the detection of ligand binding sites. *Angew. Chem., Int. Ed.* **2025**, 64, e202421000. <https://doi.org/10.1002/anie.202421000>
27. Acharyya, A.; Ahmed, I. A.; Gai, F. 4-cyanoindole-based fluorophores for biological spectroscopy and microscopy. *Methods Enzymol.* **2020**, 639, 191–215. <https://doi.org/10.1016/bs.mie.2020.04.014>
28. Wang, M.; Gai, F. Chemical and biological incorporation of the blue fluorescent[sic] amino acid 4-cyanotryptophan into proteins: Application to tune the absorption and emission wavelengths of GFP. *Phys. Chem. Chem. Phys.* **2025**, 27, 7199–7204. <https://doi.org/10.1039/d5cp00112a>
29. Zhang, K.; Ahmed, I. A.; Kratochvil, H. T.; DeGrado, W. F.; Gai, F.; Jo, H. Synthesis and application of the blue fluorescent amino acid L-4-cyanotryptophan to assess peptide-membrane interactions. *Chem. Commun.* **2019**, 55, 5095–5098. <https://doi.org/10.1039/c9cc01152h>
30. Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, 152, 224108. <https://doi.org/10.1063/5.0004608>
31. Bader, R. F. W. A quantum theory of molecular structure and its applications. *Chem. Rev.* **1991**, 91, 893–928. <https://doi.org/10.1021/cr00005a013>
32. Orton, H. W.; Qianzhu, H.; Abdelkader, E. H.; Habel, E. I.; Tan, Y. J.; Frkic, R. L.; Jackson, C. J.; Huber, T.; Otting, G. Through-space scalar ¹⁹F–¹⁹F couplings between fluorinated noncanonical amino acids for the detection of specific contacts in proteins. *J. Am. Chem. Soc.* **2021**, 143, 19587–19598. <https://doi.org/10.1021/jacs.1c10104>
33. Pritchard, R. B.; Hansen, D. F. Characterising side chains in large proteins by protonless (13)C-detected NMR spectroscopy. *Nat. Commun.* **2019**, 10, 1747. <https://doi.org/10.1038/s41467-019-09743-4>
34. Dahiyat, B. I.; Mayo, S. L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, 94, 10172–10177. <https://doi.org/10.1073/pnas.94.19.10172>
35. Maguire, J. B.; Boyken, S. E.; Baker, D.; Kuhlman, B. Rapid sampling of hydrogen bond networks for computational protein design. *J. Chem. Theory Comput.* **2018**, 14, 2751–2760. <https://doi.org/10.1021/acs.jctc.8b00033>
36. Fecker, T.; Galaz-Davison, P.; Engelberger, F.; Narui, Y.; Sotomayor, M.; Parra, L. P.; Ramirez-Sarmiento, C. A. Active site flexibility as a hallmark for efficient PET degradation by *I. sakaiensis* PETase. *Biophys. J.* **2018**, 114, 1302–1312. <https://doi.org/10.1016/j.bpj.2018.02.005>
37. Novokhatny, V.; Ingham, K. Thermodynamics of maltose-binding protein unfolding. *Protein Sci.* **1997**, 6, 141–146. <https://doi.org/10.1002/pro.5560060116>

38. Raran-Kurussi, S.; Waugh, D. S. The ability to enhance the solubility of its fusion partners is an intrinsic property of maltose-binding protein but their folding is either spontaneous or chaperone-mediated. *PLoS One* **2012**, *7*, e49589. <https://doi.org/10.1371/journal.pone.0049589>
39. Maina, C. V.; Riggs, P. D.; Grandea, A. G., III; Slatko, B. E.; Moran, L. S.; Tagliamonte, J. A.; McReynolds, L. A.; Guan, C. D. An *Escherichia coli* vector to express and purify foreign proteins by fusion to and separation from maltose-binding protein. *Gene* **1988**, *74*, 365–373. [https://doi.org/10.1016/0378-1119\(88\)90170-9](https://doi.org/10.1016/0378-1119(88)90170-9)
40. Boos, W.; Shuman, H. Maltose/maltodextrin system of *Escherichia coli*: transport, metabolism, and regulation. *Microbiol. Mol. Biol. Rev.* **1998**, *62*, 204–229. <https://doi.org/10.1128/MMBR.62.1.204-229.1998>
41. Martineau, P.; Szmecman, S.; Spurlino, J. C.; Quijcho, F. A.; Hofnung, M. Genetic approach to the role of tryptophan residues in the activities and fluorescence of a bacterial periplasmic maltose-binding protein. *J. Mol. Biol.* **1990**, *214*, 337–352. [https://doi.org/10.1016/0022-2836\(90\)90165-I](https://doi.org/10.1016/0022-2836(90)90165-I)
42. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The penultimate rotamer library. *Proteins* **2000**, *40*, 389–408.
43. Sharff, A. J.; Rodseth, L. E.; Spurlino, J. C.; Quijcho, F. A. Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry* **1992**, *31*, 10657–10663. <https://doi.org/10.1021/bi00159a003>
44. Carugo, O. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *J. Appl. Crystallogr.* **2003**, *36*, 125–128. <https://doi.org/10.1107/S0021889802020502>
45. Eyal, E.; Gerzon, S.; Potapov, V.; Edelman, M.; Sobolev, V. The limit of accuracy of protein modeling: Influence of crystal packing on protein structure. *J. Mol. Biol.* **2005**, *351*, 431–442. <https://doi.org/10.1016/j.jmb.2005.05.066>
46. Shapovalov, M. V.; Dunbrack, R. L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19*, 844–858. <https://doi.org/10.1016/j.str.2011.03.019>
47. Hintze, B. J.; Lewis, S. M.; Richardson, J. S.; Richardson, D. C. MolProbity's ultimate rotamer-library distributions for model validation. *Proteins* **2016**, *84*, 1177–1189. <https://doi.org/10.1002/prot.25039>
48. Kang, Y. K.; Choi, H. Y. Cis-trans isomerization and puckering of proline residue. *Biophys. Chem.* **2004**, *111*, 135–142. <https://doi.org/10.1016/j.bpc.2004.05.006>
49. Ganguly, H. K.; Basu, G. Conformational landscape of substituted prolines. *Biophys. Rev.* **2020**, *12*, 25–39. <https://doi.org/10.1007/s12551-020-00621-8>
50. Mukai, T.; Hoshi, H.; Ohtake, K.; Takahashi, M.; Yamaguchi, A.; Hayashi, A.; Yokoyama, S.; Sakamoto, K. Highly reproductive *Escherichia coli* cells with no specific assignment to the UAG codon. *Sci. Rep.* **2015**, *5*, 9699. <https://doi.org/10.1038/srep09699>
51. McPhillips, T. M.; McPhillips, S. E.; Chiu, H. J.; Cohen, A. E.; Deacon, A. M.; Ellis, P. J.; Garman, E.; Gonzalez, A.; Sauter, N. K.; Phizackerley, R. P.; Soltis, S. M.; Kuhn, P. Blu-Ice and the distributed control system: Software for data acquisition and instrument control at macromolecular crystallography beamlines. *J. Synchrotron Radiat.* **2002**, *9*, 401–406. <https://doi.org/10.1107/s0909049502015170>
52. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 125–132. <https://doi.org/10.1107/S0907444909047337>
53. Evans, P. R.; Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 1204–1214. <https://doi.org/10.1107/S0907444913000061>
54. McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser crystallographic software. *J. Appl. Crystallogr.* **2007**, *40*, 658–674. <https://doi.org/10.1107/S0021889807021206>
55. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 486–501. <https://doi.org/10.1107/S0907444910007493>
56. Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkoczi, G.; Chen, V. B.; Croll, T. I.; Hintze, B.; Hung, L. W.; Jain, S.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R. D.; Poon, B. K.; Prisant, M. G.; Read, R. J.; Richardson, J. S.; Richardson, D. C.; Sammito, M. D.; Sobolev, O. V.; Stockwell, D. H.; Terwilliger, T. C.; Urzhumtsev, A. G.; Videau, L. L.; Williams, C. J.; Adams, P. D. Macromolecular structure determination

- using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **2019**, 75, 861–877. <https://doi.org/10.1107/S2059798319011471>
57. Murshudov, G. N.; Skubak, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, 67, 355–367. <https://doi.org/10.1107/S0907444911001314>
58. Lebedev, A. A.; Isupov, M. N. Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, 70, 2430–2443. <https://doi.org/10.1107/S1399004714014795>
59. The PyMOL Molecular Graphics System, Version 3.0, Schrödinger, LLC.
60. Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. Electronic Ligand Builder and Optimization Workbench (eLBOW): A tool for ligand coordinate and restraint generation. *Acta Crystallogr. D Biol. Crystallogr.* **2009**, 65, 1074–1080. <https://doi.org/10.1107/S0907444909029436>
61. Gucwa, M.; Lenkiewicz, J.; Zheng, H.; Cymborowski, M.; Cooper, D. R.; Murzyn, K.; Minor, W. CMM- An enhanced platform for interactive validation of metal binding sites. *Protein Sci.* **2023**, 32, e4525. <https://doi.org/10.1002/pro.4525>

**Chapter 5: Minimalist fluorescent reporters and novel
autocatalytic chromophore chemistry with NBD amino
acids**

5.1 Introduction

Visible-light microscopy has become central to visualising biomolecular processes in their native cellular context, with marked advances in resolving power over the last two decades.¹ Yet, fluorescence readouts are often complicated by local physicochemical variables, such as solvent polarity,² viscosity,³ pH,⁴ and conformational dynamics,⁵ which modulate the fluorophore's quantum yield, extinction coefficient, spectra, and lifetime. While such sensitivity can be exploited when rigorously calibrated, many applications benefit from a compact reporter whose signal is stably correlated with the protein of interest and minimally perturbs native structure and function.^{6–8} Furthermore, dye-labelling typically targets native nucleophilic sidechains, most commonly cysteines, making them positionally non-specific when more than one of the target nucleophilic residue are present in the sequence.^{9–12} Genetically encoded fluorescent protein (FP) fusions remain a workhorse for adding fluorescence to specific proteins, but they are intrinsically large (~27 kDa), often restricted to termini, and can affect folding,¹³ trafficking,¹⁴ oligomerisation,¹⁵ or activity.¹⁶ By contrast, a single fluorescent amino acid embedded within a folded protein provides a minimalist label with far less steric burden, fixed stoichiometry, and a well-defined microenvironment. This chapter develops and validates such an approach using the nitrobenzodiazole (NBD) family of fluorophores.

5.1.1 Nitrobenzodiazole fluorophores

NBD scaffolds are classical push-pull systems in which an electron donor couples to a strong acceptor through a π -framework.¹⁷ Photoexcitation produces an intramolecular charge-transfer (ICT) state, whose dipole and radiative/non-radiative partitioning depend strongly on their local environment, yielding pronounced solvatochromism and fluorogenic “switch-on” behaviour.^{18,19} In polar media, NBD emission is weak and red-shifted; in non-polar or viscous environments it brightens and blue-shifts. These properties have been exploited in membrane and protein studies since the 1960s,^{18–21} and NBD excitation/emission (blue–green/green–yellow) is accessible to commonplace

lasers and filters, enabling routine live-cell imaging while avoiding deep-UV excitation that elevates phototoxicity^{22,23} and exacerbates cellular autofluorescence.²⁴

Figure 5.1 illustrates the chemical formulae of *N*-(7-nitro-2,1,3-benzoxadiazol-4-yl)-L-2,3-diaminopropionic acid (NBDO, $\lambda_{\text{ex}}/\lambda_{\text{em}} = 470/550$ nm) and its selenium analogue *N*-(7-nitro-2,1,3-benzoselenadiazol-4-yl)-L-2,3-diaminopropionic acid (NBDSe, $\lambda_{\text{ex}}/\lambda_{\text{em}} = 490/600$ nm).²⁵ Both maintain a compactness and neutrality that minimise steric and electrostatic disruption, yet deliver visible fluorescence and large Stokes shifts that facilitate clean separation of excitation and emission in multicolour experiments.

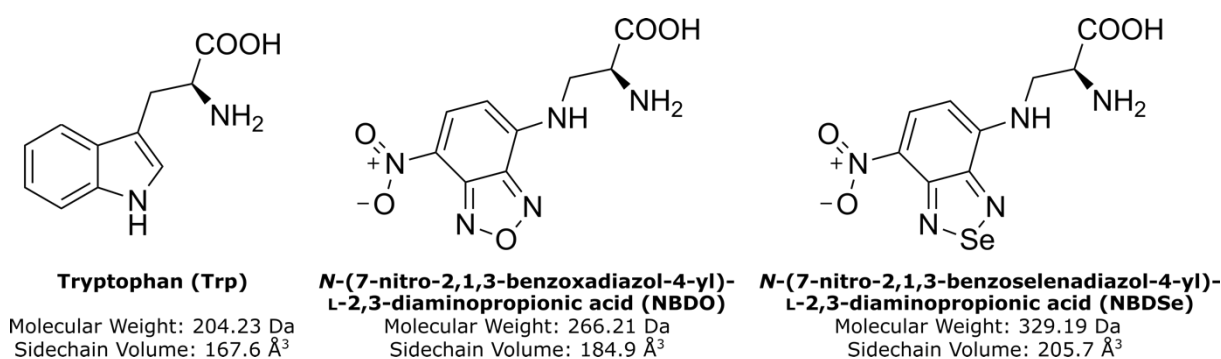


Figure 5.1 – Chemical structure of tryptophan, NBDO and NBDSe. The electron density of each sidechain was calculated with Orca²⁶ using the B3LYP basis set, the volumes were measured at a level of 0.002 a.u. to approximate the molecular envelope.²⁷

Despite their promise, NBD-based amino acids were not accessible as genetically encoded targets before this work. Their site-specific incorporation into proteins was limited to technically demanding *in vitro* strategies,^{28–30} which produce sufficient material for fluorescence readouts but insufficient quantities for structural investigation. This precluded systematic structural analysis of how proteins accommodate NBD fluorophores and how the local environment modulates their photophysical properties.

5.1.2 Fluorescent protein chromophore formation

Green fluorescent protein (GFP) from *Aequorea victoria* transformed cell biology with a genetically encoded tag that becomes fluorescent through autocatalytic chemistry.³¹ An 11-stranded β -barrel encloses the chromophore-forming tripeptide (positions 65–67) on

a central helix, creating a solvent-shielded microenvironment that both protects the nascent chromophore and catalyses its maturation.³² Maturation proceeds via a multi-step pathway. First, the backbone amide of residue 67 (usually glycine) attacks the carbonyl of residue 65 to cyclise into an imidazolinone. Subsequent dehydration and O₂-dependent oxidation of the C_α–C_β bond at residue 66 extends the π -system through the sidechain to yield the mature chromophore.³³ While the order of the final steps remains debated, it is agreed that an arginine (Arg96) stabilises the developing enolate/electrophilic centre, while a glutamate (Glu222) acts as a general base.^{34,36}

The modern palette of fluorescent proteins (FPs) spans blue³⁷ to near-IR³⁸ and varies in brightness,^{39,40} chromophore pK_a,⁴¹ photoswitching⁴² and oligomeric state^{43,44} (see FPbase).⁴⁵ Yet, this diversity is built on canonical amino acids (mostly tyrosine (Tyr), tryptophan (Trp) or phenylalanine (Phe)) alone. Naturally occurring anthozoan chromoproteins follow the same basic scaffold but often undergo an additional N–C_α oxidation to form an *N*-acylimine, extending conjugation and bathochromically shifting absorbance while frequently reducing fluorescence.^{46–49} Genetic code expansion (GCE) has been used to diversify this aromatic position and typically produces modest but diagnostic spectral shifts while retaining autocatalytic maturation.^{50–57} Collectively, these studies show that the β -barrel will accommodate a wide range of substituted aromatics. All cited examples rely on a C–C oxidation to form the methine bridge between the side-chain aromatic and the imidazolinone backbone. By contrast, the work presented here reveals π -conjugation extending across a methylene bridge between the side-chain and imidazolinone, mediated by an exocyclic secondary amine. This is a fundamentally distinct mode of chromophore formation that expands the chemical space accessible within the FP scaffold.

This chapter establishes NBD amino acids as compact, visible-light fluorescent reporters that preserve native protein architecture and function. A structure-guided analysis using a model nanobody to describes how NBD sidechains are accommodated into a protein core, and how the local environment modulates their photophysical properties. The

structure of the NBD amino acids incorporated into FPs is then presented. These amino acids are incorporated into the conserved aromatic residue site within the chromophore-forming tripeptide of a FP scaffold, where the fully mature chromophore is shown. This illustrates that the exocyclic secondary amine can participate in the autocatalytic maturation to furnish novel chromophores, highlighting how NBD chemistry can both report on and be integrated into protein function. Overall, by leveraging the small size, fluorogenicity and visible light compatibility of NBD amino acids, this work delivers a minimally perturbing route to protein-specific fluorescence that complements and improves upon FP fusions. The structural and spectroscopic data provided here are intended to guide rational placement of NBD reporters in other proteins and to enable reliable interpretation of their signals in biological settings.

5.2 Structural analysis of a nanobody functionalised with NBD-fluorophores

To assess whether NBD amino acids can be incorporated into buried sites without disrupting protein structure or function, the high-affinity H11 nanobody was selected as a model system.⁵⁸ H11 binds the N-terminal lobe of human activity-regulated cytoskeletal protein (ArcNL), which is ubiquitous in neuronal cells.⁵⁹ Tryptophan 37 (W37) lies buried within the nanobody core, and the model system was reported to crystallise well.⁵⁸

5.2.1 Tryptophan-NBD pseudoisosterism

NBDO and NBDSe can occupy similar 3-dimensional space to Trp, despite the L-2,3-diaminopropionic acid linker other extra heavy atoms. The secondary amine linker in NBDO introduces an additional dihedral angle (χ_3 , defined by $C_\beta-N_\gamma-C_\delta-C_{\epsilon 1}$). In Trp, the $\chi_{2,2}$ angle is defined by the $C_\alpha-C_\beta-C_\gamma-C_{\delta 2}$ dihedral is analogous to the χ_2 dihedral in NBDO (defined by $C_\alpha-C_\gamma-N_\gamma-C_\delta$). When χ_1 and $\chi_2/\chi_{2,2}$ dihedrals are matched and corresponding atoms are superimposed, overlap between NBDO and Trp has two maxima at 15° and 180° , corresponding to alignment of the ring planes (Figure 5.2). The bicyclic ring system of the NBD superimposes with the six-membered ring of Trp. In addition, there are fewer

H-atoms for NBDO/NBDSe relative to Trp, meaning the volume of NBDO is closer to that of Trp than would be inferred by looking at only the heavy atoms. While the overlap minima and maxima are identical for NBDSe, the increased steric bulk of NBDSe means the volume of the amino acid is ~20% larger than Trp, compared to ~10% for NBDO. The increased volume arises not only from the larger van der Waals (vdW) radius of selenium compared to oxygen, but also from the increased bond length of the associated N-Se covalent bonds, which projects the centre of the selenium atom 3.7 Å from the centre of the 6-membered ring, as opposed to 3.2 Å for oxygen in NBDO.

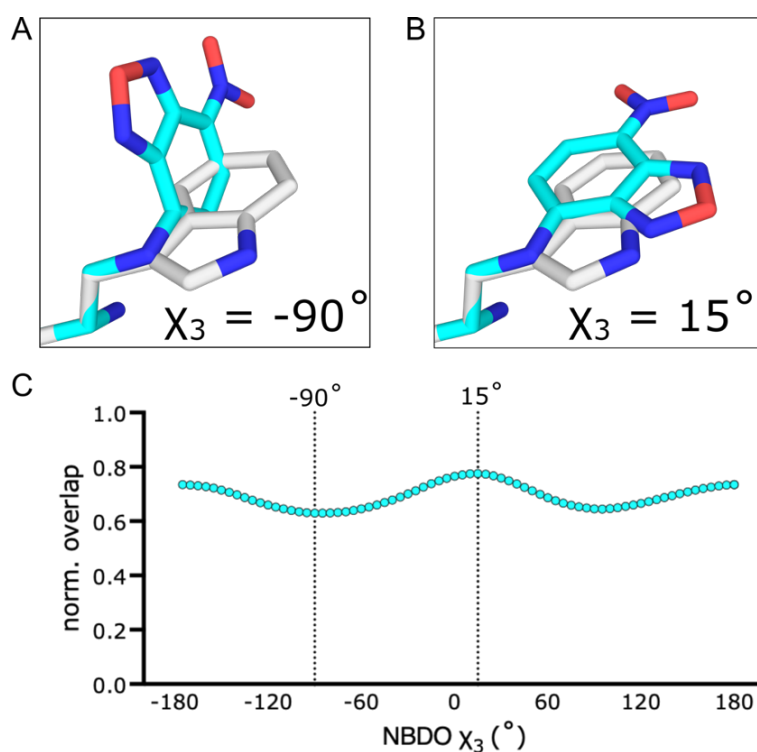


Figure 5.2 – Theoretical overlap between Trp (white) and NBDO (cyan). (A and B) Stick illustrations set to matching dihedral angles. (A) shows NBDO with $\chi_3 = -90^\circ$, (B) shows NBDO with $\chi_3 = 15^\circ$. (C) Normalised overlap calculated by using the overlap command in PyMOL for the superimposed amino acids, for every 5° χ_3 angle in NBDO, and dividing the result by the self-overlap of Trp.

5.2.2 Crystal structure of ArcNL-H11 W37 NBDO

The crystal structure of ArcNL-H11-W37 NBDO (H11 NBDO) was refined to 1.56 Å (PDB ID: 9PZZ). The asymmetric unit contains two independent copies of the ArcNL-H11 heterodimer. Superimposition of the two heterodimer pairs within the asymmetric unit produces a root-mean-square-deviation (RMSD) for all 190 C_α atoms of less than 0.2 Å.

The incorporation of NBDO at position 37 preserves the global protein architecture. Superimposition of the H11 NBDO nanobody onto the WT structure (PDB ID: 8QF4)⁵⁸ yields RMSD values of 0.248 Å (chain A) and 0.159 Å (chain B), confirming that the canonical immunoglobulin fold is maintained. Antigen binding is also preserved. The buried surface area (BSA) between H11 W37 NBDO and ArcNL is 1050 Å², which is nearly identical to the 1044 Å² reported for the wild-type (WT) complex (Figure 5.3).⁵⁸ The H11 native fold accommodates the steric bulk of NBDO relative to Trp through local sidechain repacking.

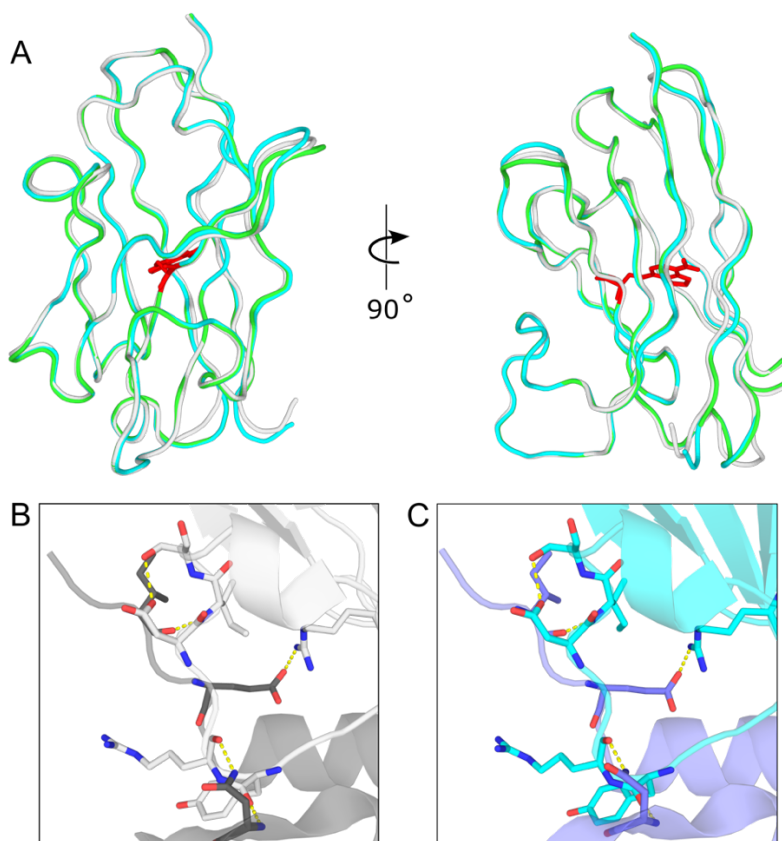


Figure 5.3 – Overall structure of H11 NBDO and protein binding interface with ArcNL. (A) Ribbon representation of WT ArcNL–H11 heterodimer (White, PDB ID: 8QF4), protomer 1 of ArcNL–H11 NBDO (cyan), protomer 2 of ArcNL–H11 NBDO (green). NBDO amino acid in stick representation in red. Key polar contacts are represented by dashed yellow lines (B, C) Protein binding interface between ArcNL and nanobody, (B) WT nanobody (white) interface with ArcNL (grey), buried surface area 1044 Å².⁵⁸ (C) H11 W37 (cyan) interface with ArcNL (dark blue), buried surface area 1050 Å².

The NBDO sidechain adopts a rotameric state that closely mimics the native tryptophan. The χ_1 dihedral is -83.6° ($\chi_1 = -73.2^\circ$ in WT), and the χ_2 is -73.3° ($\chi_{2,2} = -88.1^\circ$ in WT). The

χ_3 angle is 174.9° matching a predicted maximum overlap from Figure 5.2. Accommodation of the NBDO sidechain is achieved by repacking of neighbouring residues (Figure 5.4). Glu8 χ_1/χ_2 changes from $-54.6^\circ/93.1^\circ$ in WT to $-62.4^\circ/147.6^\circ$ in W37 NBDO, which is a $\sim 60^\circ$ rotation in its χ_2 angle. Val81 flips its χ_1 from -161.3° to 54.1° , which increases the distance between the aromatic sidechain and the Val81 C β . The high-resolution electron density map shows unambiguously that the NBD fluorophore is buried and excluded from bulk solvent, with no ordered water molecules observed within vdW contact. Relative to the free amino acid in PBS, buried NBDO sidechain in H11 has a blue-shifted emission maximum ($540\text{ nm} \rightarrow 500\text{ nm}$), and a $\sim 1.7\times$ increase in brightness (Figure 5.4C).

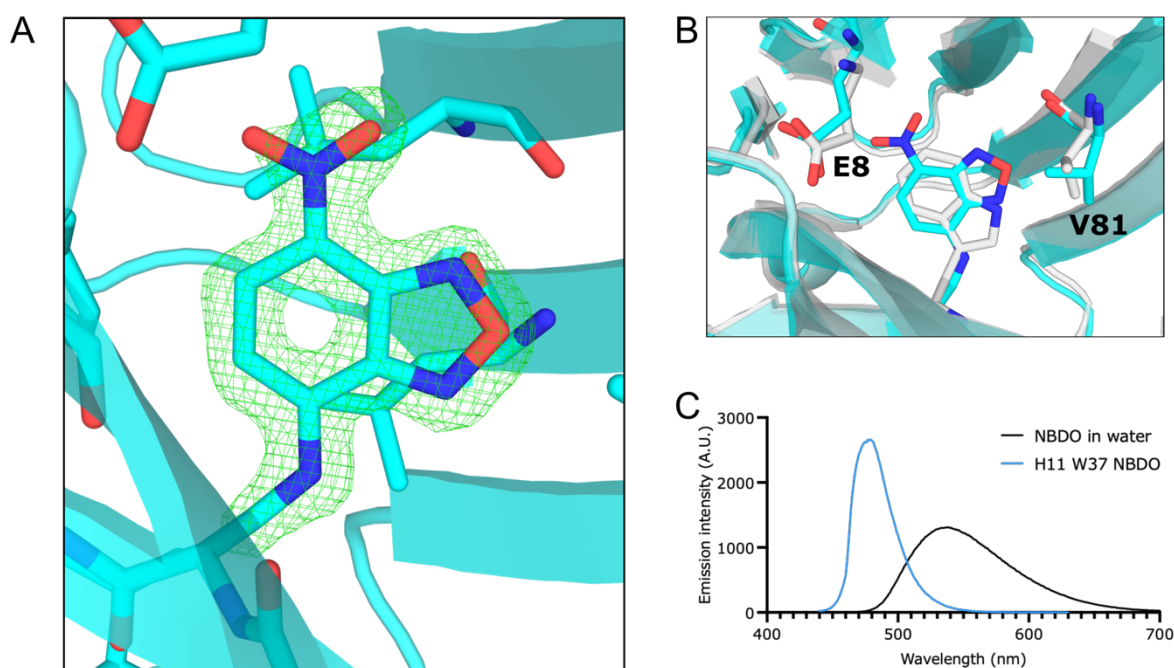


Figure 5.4 – Local environment of NBDO in the W37 position and fluorescence enhancement. (A) Cartoon and stick illustration of ArcNL–H11 W37 NBDO, green mesh represents mF_o-F_c contoured to 6σ , generated by omitting NBDO sidechain atoms. (B) Superimposition of ArcNL–H11 W37 NBDO and ArcNL–H11 WT. (C) Fluorescence emission spectra in PBS. Free NBDO: $\epsilon = 17,833 \pm 26\text{ M}^{-1}\text{cm}^{-1}$, QY = $7.2 \pm 0.1\%$, brightness = 1284, $\lambda_{\text{ex}}/\lambda_{\text{em}} = 470/540\text{ nm}$. H11 W37 NBDO: $\epsilon = 13,304 \pm 197\text{ M}^{-1}\text{cm}^{-1}$, QY = $16.3 \pm 0.5\%$, brightness = 2169, $\lambda_{\text{ex}}/\lambda_{\text{em}} = 435/500\text{ nm}$. Both samples were measured at $10\text{ }\mu\text{M}$. Brightness is calculated as $\epsilon \times \text{QY}$.

5.2.3 Crystal structure of ArcNL–H11 W37 NBDSe

The crystal structure of ArcNL–H11 W37 NBDSe was determined to a resolution of 2.8 Å (PDB ID: 9Q00). The position of the NBDSe residue was confirmed by a strong peak in the anomalous difference map corresponding to the selenium atom (Figure 5.6). Despite the increased steric bulk, the global architecture and its binding function remain intact. The C α RMSD relative to the WT structure is 0.462 Å (chain A, 124 C α pairs) and 0.464 Å (chain B, 124 C α pairs). The binding interface with ArcNL is conserved, with BSA values of 1048 Å² (chain A) and 1050 Å² (chain B).

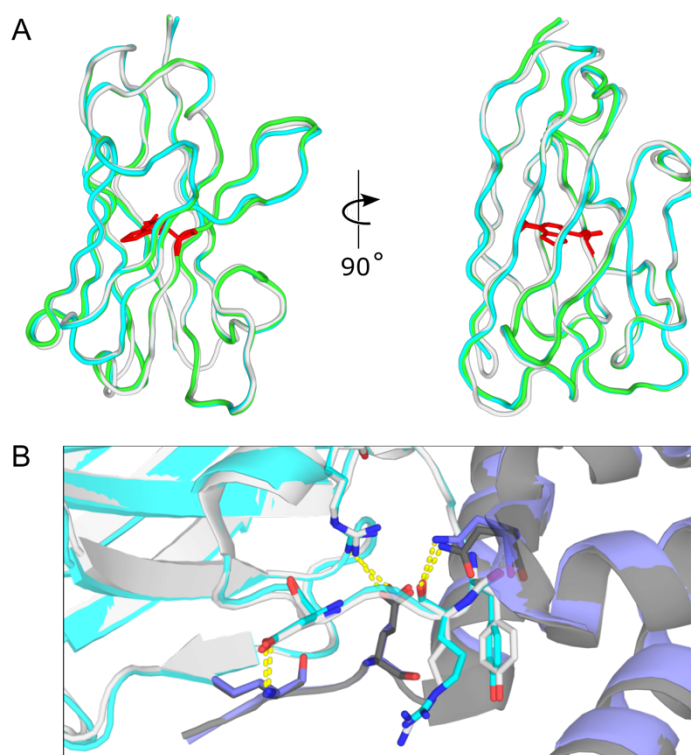


Figure 5.5 – Overall structure of H11 NBDSe and protein binding interface with ArcNL. (A) Superimposition of the ribbon representations of WT ArcNL–H11 (PDB ID: 8QF4, white), chain A of ArcNL–H11 NBDSe (PDB ID: 9Q00, cyan), and chain C of ArcNL–H11 NBDSe (green). The C α RMSD between the NBDSe variants and WT are 0.462 Å (chain A) and 0.464 Å (chain C), respectively. (B) The binding interface between H11 (chain A) and ArcNL (chain B) with a BSA of 1048 Å².

Like NBDSe, the NBDSe sidechain adopts a rotameric state similar to native tryptophan. The χ_1 dihedral is -77.7° ($\chi_1 = -73.2^\circ$ in WT) and χ_2 is -76.9° ($\chi_{2,2} = -88.1^\circ$ in WT). The χ_3 angle is 177.8° , maintaining the conformation observed in NBDSe. Accommodation of

the increased steric bulk requires repacking of neighbouring sidechains. Val81 flips its χ_1 from -161.3° in WT to 53.9° in NBDSe, moving the sidechain away from the selenium atom. Glu8 undergoes rotameric shifts: χ_1/χ_2 changes from $-54.6^\circ/93.1^\circ$ in WT to $-59.2^\circ/148.5^\circ$ in the NBDSe variant, which is the same magnitude of variation as in the H11–NBDO structure. The electron density map shows the NBDSe fluorophore remains buried and excluded from bulk solvent, with no ordered water molecules within vdW contact. The quantum yield increases 55-fold from 0.4% (free NBDSe in PBS) to 21.9% (H11 W37 NBDSe) (Figure 5.6C).

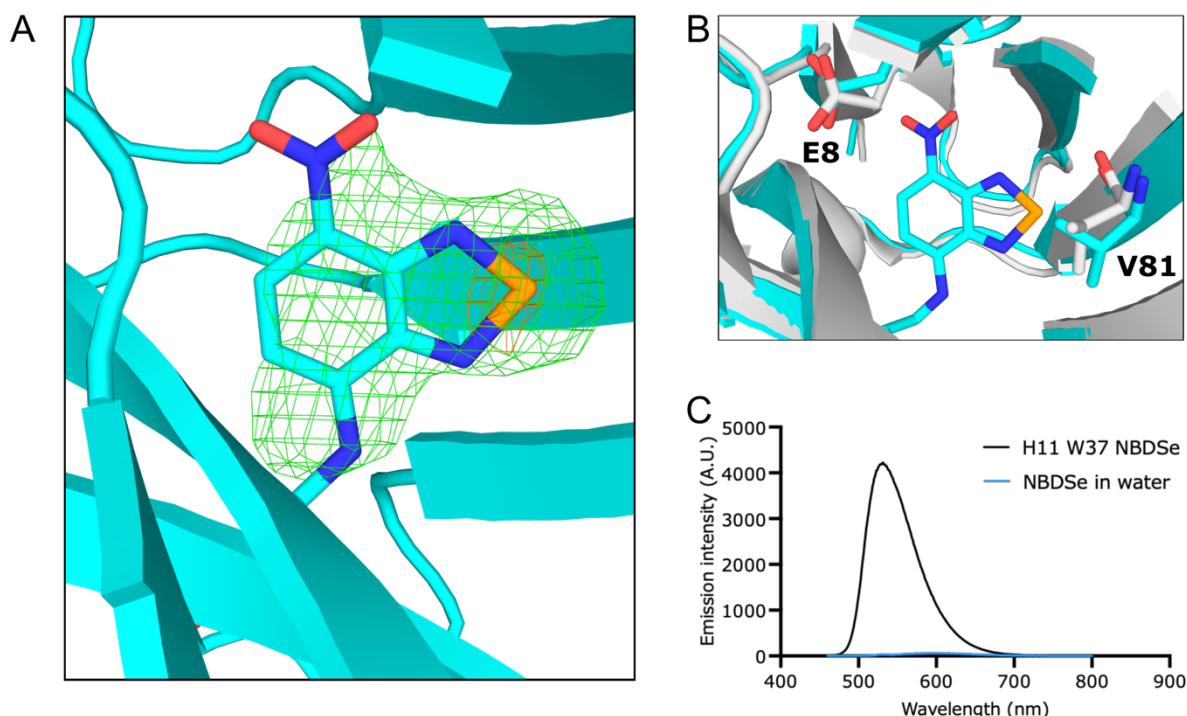


Figure 5.6– Local environment of NBDSe in the W37 position. (A) Cyan, cartoon, stick of NBDSe, $mF_o - mF_c$ density shown as green mesh contoured to 3σ , anomalous difference map shown as orange mesh contoured to 4σ (B) Superimposition of ArcNL–H11 NBDSe and ArcNL–H11 WT. (C) Fluorescence emission spectra in PBS. Free NBDSe: $\epsilon = 13,457 \pm 28 \text{ M}^{-1}\text{cm}^{-1}$, QY = $0.4 \pm 0.1\%$, brightness = 54, $\lambda_{\text{ex}}/\lambda_{\text{em}} = 490/595 \text{ nm}$. H11 NBDSe: $\epsilon = 8950 \pm 187 \text{ M}^{-1}\text{cm}^{-1}$, QY = $21.9 \pm 0.3\%$, brightness = 1960, $\lambda_{\text{ex}}/\lambda_{\text{em}} = 470/530 \text{ nm}$. Both samples were measured at $10 \mu\text{M}$. Brightness is calculated as $\epsilon \times \text{QY}$.

The nanobody structures demonstrate that NBD amino acids are pseudo-isosteric to tryptophan and can be accommodated at buried sites through local sidechain repacking without disrupting global fold or function. The dramatic fluorescence enhancement upon incorporation (brightness increases 2.3-fold for NBDO and 55-fold for NBDSe) likely

arises from two synergistic mechanisms which have been reported in studies on the free NBD sidechain moiety. First, the non-polar, solvent-excluded environment suppresses solvent-mediated quenching, to which NBD fluorophores are particularly susceptible.^{17,21} The greater enhancement for NBDSe reflects the higher sensitivity of its QY to solvent quenching. Second, viscosity has been established to reduce the non-radiative decay constant of NBD.¹⁸ The tight packing by surrounding residues restricts torsional motion of the amino acid, reducing dynamics-related non-radiative decay. These results establish that strategic placement of NBD amino acids at buried Trp sites can generate minimally perturbing, self-reporting proteins with significantly higher quantum yields in the protein relative to water.

5.3 NowGFP with NBD chromophores

To test whether the chromophore position of a fluorescent protein provides a buried environment for NBDO, NowGFP was selected as a scaffold. FPs are sequence optimised for expression and stability, and NowGFP natively has a tryptophan in the central helix tripeptide.⁶¹ The resultant proteins were unexpectedly dark blue with both NBDO and NBDSe (Figure 5.7A), contrasting with the yellow of the precursor amino acid and the green of NowGFP. This provided the first evidence that autocatalytic chromophore formation had occurred through the exocyclic secondary amine, which is not present in canonical amino acids. UV-Vis absorption spectroscopy confirmed the origin of the colour. The spectrum revealed a strong, broad absorption across most visible wavelengths. A distinct transmission window in the blue region produces the observed hue by selective absorption. This mechanism is fundamentally different from the colours of GFPs where colour arises mostly from the emission of photons at longer wavelengths following excitation. NowGFP NBDO and NowGFP NBDSe show similar UV-Vis spectra, with a bathochromic shift of the main absorption peak to 620–640 nm (Figure 5.7B). Fluorescence spectroscopy revealed weak emission peak at 690 nm from excitation at 620 nm for NowGFP NBDO, and an emission peak at 540 nm for excitation at 440 nm and 350 nm (Figure 5.7C). UV-Vis of the R96A variant, a mutation known to halt chromophore

formation,³³ lacks the ~470 nm absorbance peak, confirming that the NBD chromophore maturation requires the same autocatalytic machinery as canonical GFP.

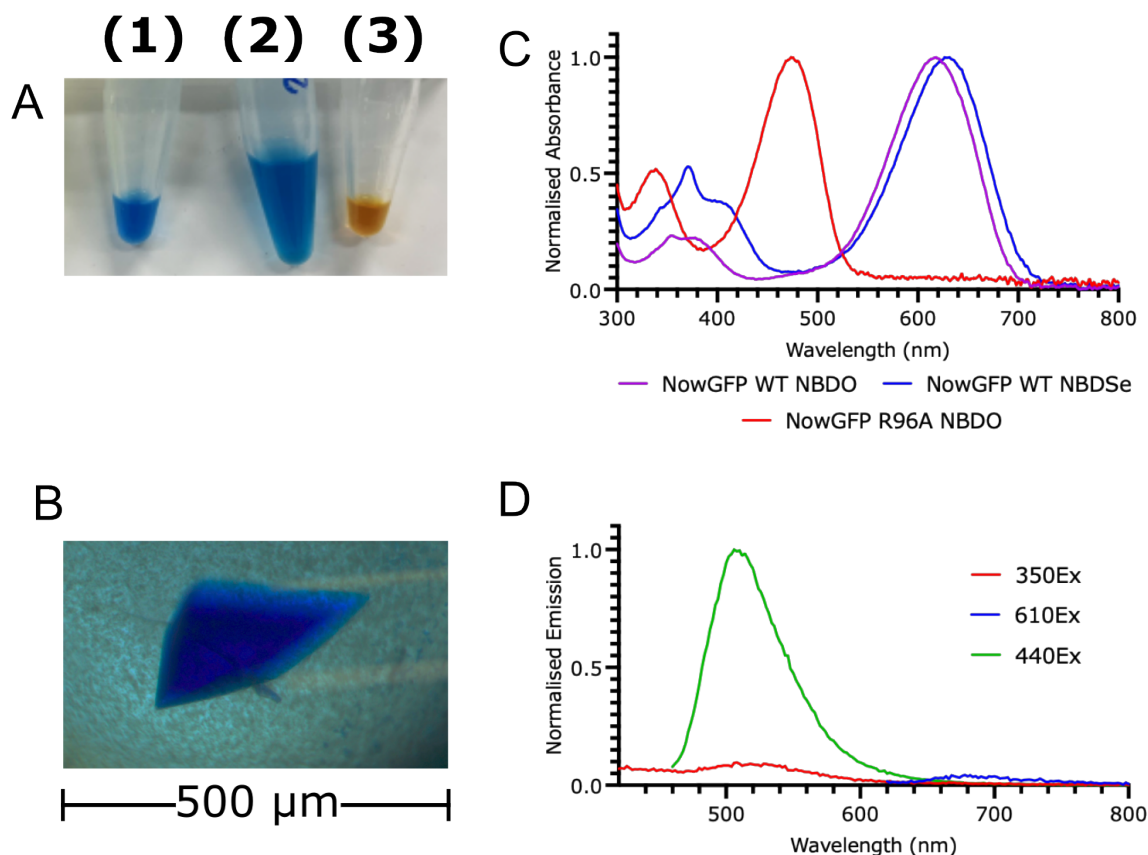


Figure 5.7 – Spectral comparison of NowGFP NBDO, NowGFP NBDSe and NowGFP R96A NBDO. (A) (1)NowGFP WT NBDO, (2)NowGFP NBDSe, (3) NowGFP R96A NBDO showing visible colour of purified proteins. (B) Crystal of NowGFP NBDO. (C) UV-Vis absorption spectra normalised to maximum signal. NowGFP WT NBDO (purple), NowGFP WT NBDSe (blue), NowGFP R96A NBDO (red). The R96A variant lacks the mature chromophore, retaining absorption at 470 nm characteristic of free NBDO. (D) Emission spectra of NowGFP NBDO at different excitation wavelengths, normalised to maximum peak height at $\lambda_{\text{ex}} = 440$ nm. $\lambda_{\text{ex}} = 350$ nm (red), $\lambda_{\text{ex}} = 610$ nm (blue), $\lambda_{\text{ex}} = 440$ nm (green).

5.3.1 Crystal structure of NowGFP NBDO

The crystal structure of NowGFP NBDO was determined to 1.36 Å. The asymmetric unit contains two protein molecules. The protein maintains the canonical 11-stranded β -barrel fold characteristic of GFP family proteins. Overlay of the C_{α} backbone with that of NowGFP WT yields RMSD values of 0.248 Å (chain A) and 0.157 Å (chain B) indicating that the incorporation of NBDO preserves the overall protein architecture. The omit map of

the chromophore model contoured at 8 σ confirms that a novel intramolecular condensation reaction has occurred.

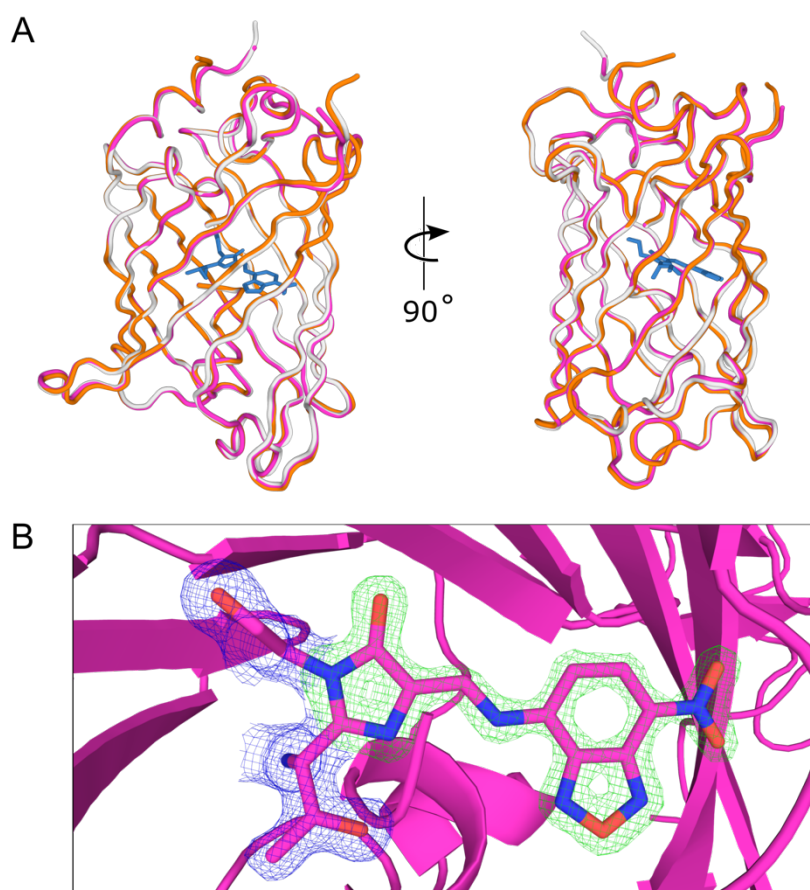


Figure 5.8 – Overall structure of NowGFP NBDO reveals preservation of the β -barrel fold and unambiguous density for the novel chromophore. (A) Superimposition of the C_{α} traces of NowGFP NBDO chain A (magenta) and chain B (orange) with NowGFP WT (white, PDB ID: 4RTC). The matured NBDO chromophore is shown as a teal stick model. (B) mF_o-DF_c omit electron density map (green mesh) contoured at 8.0 σ , calculated after omitting the entire matured chromophore from the model. The surrounding $2mF_o-DF_c$ density (blue mesh), contoured at 2.0 σ , highlights the well-resolved residues bound to the imidazolinone ring.

The NBDO sidechain adopts a rotameric state that closely mimics that of the native Trp. The χ_1 angle is 3.4° which is nearly identical to the 2.3° observed for Trp66 in NowGFP WT. The rotation of the sidechain around the bond between the β - and γ - atoms ($\chi_{2,2}$ in Trp and χ_2 in NBDO) is 168.6° for NBDO, and 172.4° in the WT structure. The χ_3 is in a trans-

conformation (178.2°). This planar arrangement allows the NBD ring system to occupy a similar three-dimensional volume to the indole ring of Trp, demonstrating the same pseudo-isosterism observed in H11. The nitro group of NBDO protrudes between two β -strands to face the solvent (Figure 5.9B) without disrupting any β -sheet hydrogen bonds that are present in the NowGFP WT structure. There are subtle shifts in the relative locations of the backbone atom coordinates relative to the WT, but these do not drastically change the distance between any paired carbonyls and amides (Figure 5.9B).

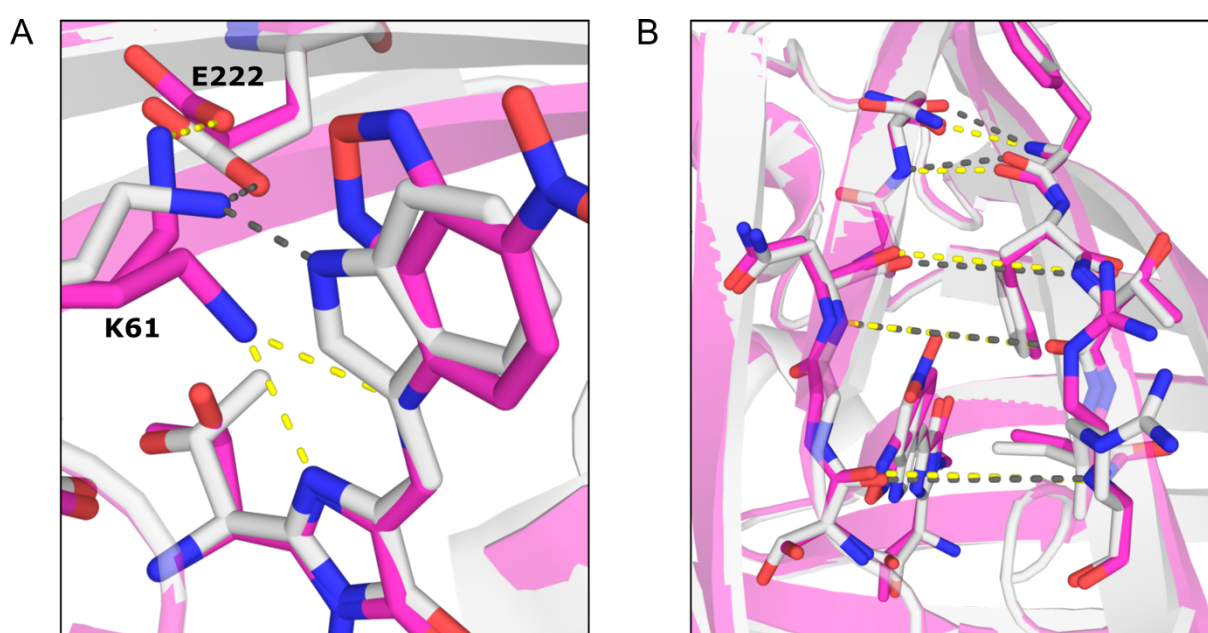


Figure 5.9 – Superimposition of NowGFP WT (white, PDB ID: 4RYW) and NowGFP NBDO (magenta). (A) Close up of chromophore sidechain, Lys61 and Glu222. Dashed lines represent polar contacts in WT (grey) and NBDO (yellow) protein. 2 dashes are equivalent to approximately 1 Å. (B) Protrusion of the NBDO nitro group through the β -sheet, with backbone distances marked with dashed lines.

The novel chromophore induces subtle changes in the local electrostatic environment. In NowGFP WT, a salt bridge between the sidechains of Lys61 and Glu222 stabilises the chromophore pocket, with Glu222 functioning as a general base that abstracts a proton during maturation.^{34,62} In crystal structures of NowGFP WT grown at pH 9.0, the Lys61 terminal amine is positioned 3.0 Å from the Trp $N_{\epsilon 1}$, which spectroscopic evidence indicates is deprotonated at this pH. In NowGFP NBDO, Lys61 is best modelled with a low occupancy alternate conformation (0.28 occupancy), indicating conformational flexibility. The minor conformer positions Lys61 within hydrogen-bonding distance of

both the imidazolinone nitrogen and the exocyclic secondary amine of the methine linker (Figure 5.9B). This configuration is consistent with partial negative charge on both nitrogen atoms.

5.4 Crystal structure of NowGFP NBDSe

The crystal structure of NowGFP NBDSe was determined to 1.87 Å resolution. The global 11-stranded β -barrel architecture is conserved. Alignment with NowGFP WT yields C_{α} RMSD values of 0.337 Å for chain A and 0.281 Å for chain B, marginally higher than those for NowGFP NBDO (0.248 Å and 0.157 Å). This indicates that NBDSe induces moderately greater structural perturbation. The position of NBDSe was confirmed by a strong peak in the anomalous difference map at 6 σ , corresponding to the selenium atom (Figure 5.10).

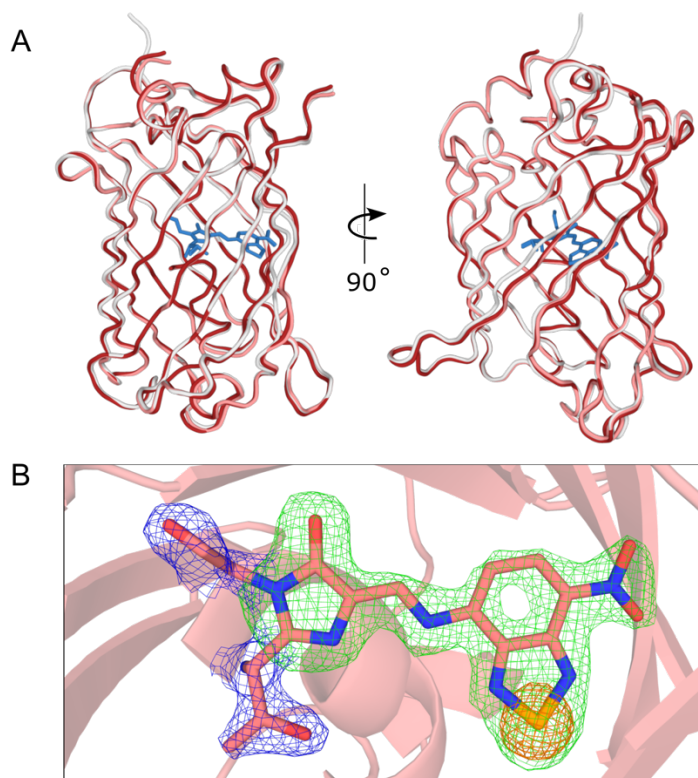


Figure 5.10 – Overall structure of NowGFP NBDSe and localisation of the selenium atom. (A) Superimposition of NowGFP NBDSe chain A (salmon) and chain B (deep red) with NowGFP WT (white, PDB ID: 4RTC). The matured NBDSe chromophore is shown as a teal stick model. (B) Electron density maps confirming the chromophore structure. The mF_o-DF_c is shown as a green mesh contoured to 4.5 σ . The anomalous difference map is shown as an orange mesh, contoured to 6.0 σ . The $2mF_o-DF_c$ density is shown as a blue mesh for surrounding residues contoured to 2.0 σ .

The NBDSe chromophore sidechain maintains a similar orientation to the WT tryptophan. The χ_1 and $\chi_{2,2}/\chi_2$ dihedrals differ by only 6.3° and 1.5° , respectively, and χ_3 is in the expected trans conformation (-3.8°). The β -strand distortion is more pronounced than in NBDSe, with changes in the hydrogen bond network. New backbone hydrogen bonds form between the β -strands of 201–209 and 146–149. Concurrently, the hydrogen bond between the carbonyl of Tyr164 and the backbone amide of Asn151 is lost. These rearrangements accommodate the larger selenium atom through local adjustment of the β -barrel wall.

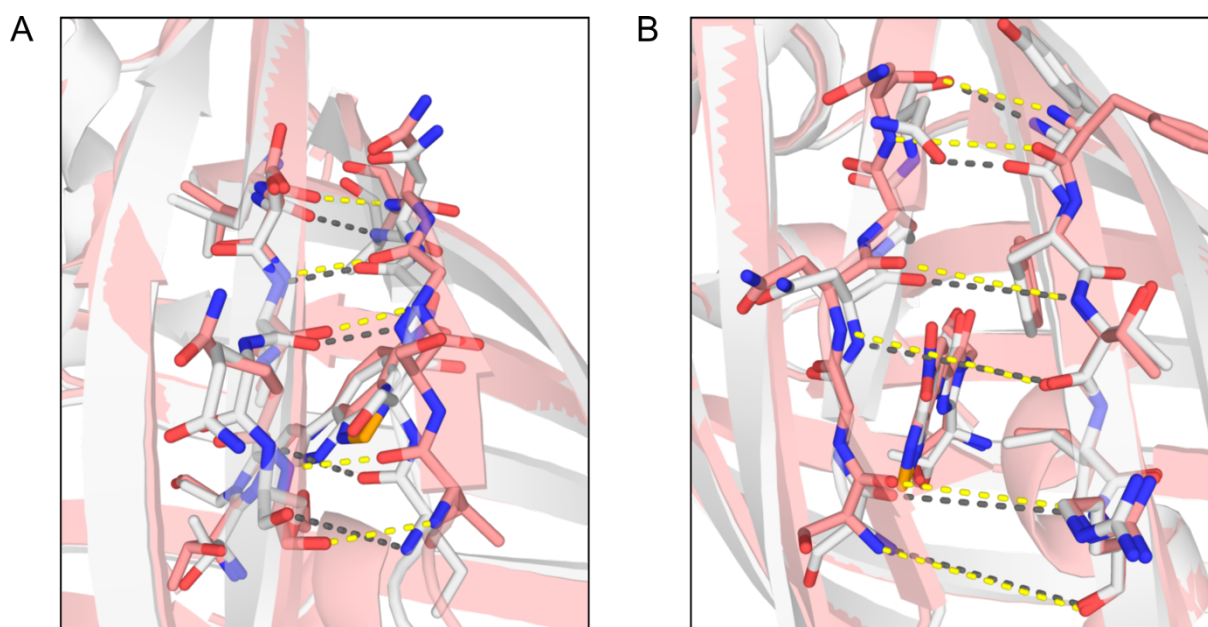


Figure 5.11 – Differences in the β -sheet of the NowGFP scaffold. NowGFP NBDSe (salmon) and NowGFP WT (white). (A) interactions between residues 201–209 and 146–149, with polar contacts illustrated as dashed lines. (B) interactions between residues 146–149 and 164–169, dashed lines represent polar contacts.

Increased steric pressure from NBDSe causes changes to nearby aliphatic residues and backbone positioning (Figure 5.12). In NowGFP NBDSe the Lys61 is locked into a single, 100% occupancy conformation. Ile146 shifts away from its WT position; its χ_1 has rotated from -159.3° in the WT to 54.8° , positioning both the methyl and ethyl groups away from the selenium atom. χ_2 changes from 100.7° to 179.1° , further rotating the ethyl arm away from the NBDSe. Concurrently, Ser205 rotates inward from -159.3° in the WT to -48.6° in NBDSe. Thr203 translates a short distance, but its χ_1 remains relatively unchanged

(-66.9° to -53.1°) This repacking positions Lys61 between Glu222 and the exocyclic secondary amine of the methine linker.

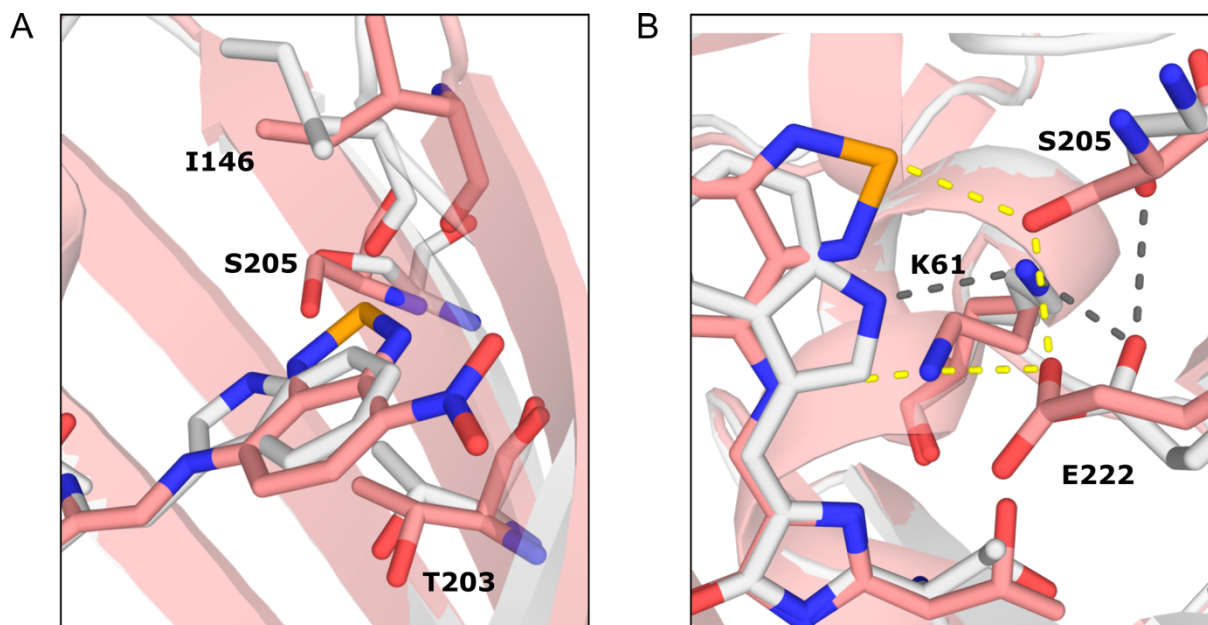


Figure 5.12 – Local environment changes in NowGFP NBDSe. (A) Superimposition of NowGFP WT (white) and NowGFP NBDSe (salmon) showing repacking of aliphatic residues Ile146, Ser205, and Thr203 around the chromophore. **(B)** Electrostatic network around the chromophore. Lys61 adopts a single conformation positioned between Glu222 and the exocyclic secondary amine. Yellow dashes: polar contacts in NBDSe; grey dashes: polar contacts in WT. Dashes are approximately 1 Å per 2 dashes.

5.5 Restoring NBD chromophore formation in a related scaffold

The NBDO forms an autocatalytic chromophore in NowGFP but not its ancestor mCerulean, despite a conservative set of 17 mutations differentiating the sequence (Figure 5.13). Both contain the autocatalytic residues Arg96 and Glu222, but mCerulean WT NBDO retains the colour of unmaturing NBDO. The NowGFP NBDO/NBDSe structures show that Lys61 contacts the mature chromophore, and the nitro group protrudes out between β -strands 146–149 and 164–169. mCerulean differs at several of these positions: Lys61 $\rightarrow$ Val61, Gly148 $\rightarrow$ Asp148, Tyr164 $\rightarrow$ Asn164, Thr166 $\rightarrow$ Lys166. Positions 164 and 166 point their sidechains toward the solvent, and so their mutation to different sidechains would minimally affect the chromophore. Reversion of Val61 and Asp148 to their NowGFP equivalents showed that chromophore maturation could be restored with the D148G mutation alone.

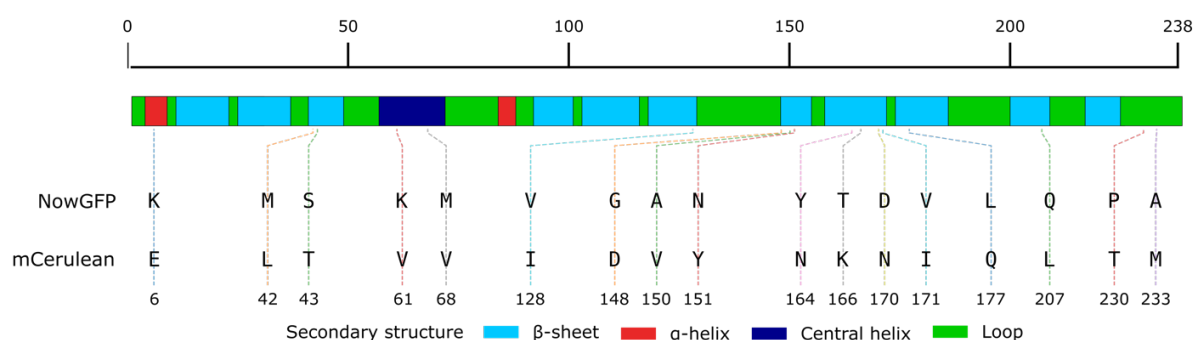


Figure 5.13 – Sequence alignment of NowGFP and mCerulean. Secondary structure elements mapped onto sequence: β-strands (cyan), α-helix (red), chromophore-forming central helix (dark blue), loops (green). Seventeen substitutions separate the scaffolds, with key differences at positions 61 (Lys → Val) and 148 (Gly → Asp) adjacent to the chromophore cavity. Residue numbering corresponds to protein sequence.

Superimposition of Cerulean and NowGFP structures shows that Asp148 directly clashes with the space occupied by the NBDO nitro group during chromophore maturation (Figure 5.14A). In NowGFP NBDO, this nitro substituent protrudes between β-strands 7 and 10, where Gly148 provides an open pocket. Asp148 in Cerulean closes this space with both steric bulk and electrostatic repulsion from the carboxylate. Position 148 also controls O₂ access to the chromophore cavity. Mechanistic studies on multiple GFP variants show that conformational flexibility at position 148 permits transient O₂ entry required for chromophore oxidation, despite crystal structures capturing predominantly "closed" conformations in the mature protein.^{35,63–65} The Cerulean crystal structure (PDB 2WSO, differing from mCerulean by A206K)⁶⁶ shows complete burial of the tryptophan chromophore with no solvent channels (Figure 5.14D), yet the smaller indole sidechain permits sufficient backbone mobility for O₂ diffusion during maturation. The bulkier NBDO substituent may restrict this flexibility when Asp148 is present, preventing both nitro group protrusion and O₂ access, both established as important for chromophore maturation.^{35,63–65} The mutation may also relieve steric clash between the NBD nitro group and the backbone of the central helix, which needs specific geometry for the initial cyclisation.⁶⁷

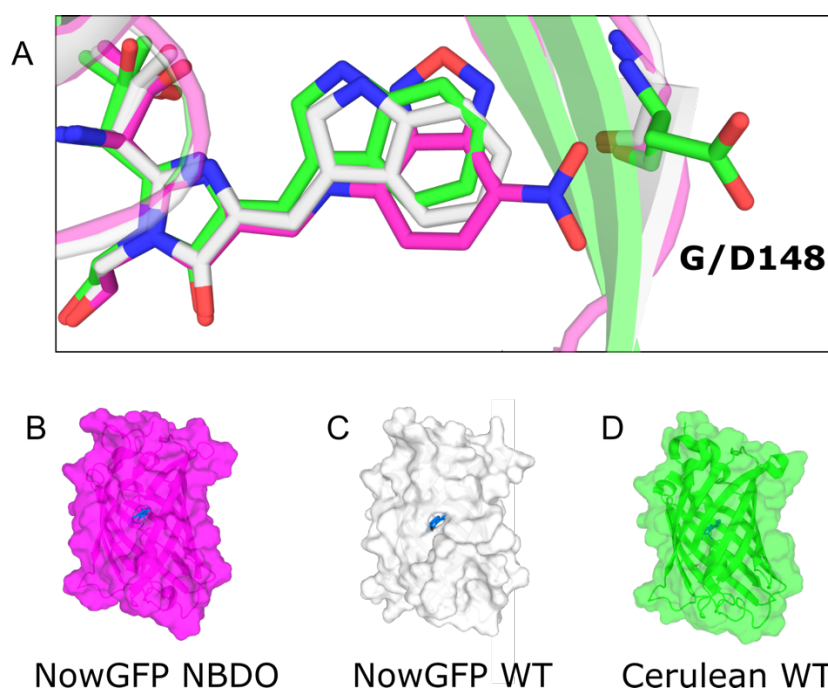


Figure 5.14 – Steric clash of Asp148 with the NBDO chromophore and solvent exclusion in the β -barrel cavity. (A) Superimposition of NowGFP NBDO (magenta), NowGFP WT (white, PDB 4RTC), and Cerulean WT (green, PDB 2WSO) showing chromophores and backbone positions of Gly148 (NowGFP) and Asp148 (Cerulean) in stick representation. The NBDO nitro group occupies the space near by the Asp148 sidechain in Cerulean. (B–D) Surface representations with chromophores shown in blue stick models. (B) NowGFP NBDO with Gly148 providing an open pocket for nitro group protrusion. (C) NowGFP WT with open pocket to the Trp chromophore. (D) Cerulean WT showing complete chromophore burial with no surface-accessible channels to the chromophore cavity.

5.5.1 Crystal structures of mCerulean D148G NBDO and mCerulean D148G/V61K NBDO

The D148G mutation restores chromophore formation. The UV-Vis spectrum of mCerulean D148G NBDO shows the characteristic mature chromophore absorption at 620–640 nm but retains residual unmaturation absorption at 440 nm (Figure 5.15C). Given that Lys61 in NowGFP NBDO adopts a conformation contacting the exocyclic secondary amine, the V61K mutation was introduced to test whether this interaction could eliminate the residual unmaturation absorption. The D148G/V61K double mutant loses this residual peak, yielding an absorption spectrum indistinguishable from NowGFP NBDO.

Crystal structures of mCerulean D148G NBDO (PDB 9Y97, 1.74 Å) and mCerulean D148G/V61K NBDO (PDB 9Y98, 1.56 Å) both reveal mature chromophores with

unambiguous electron density for the conjugated system (Figure 5.15A,B). The global architecture and chromophore geometry are essentially identical in both structures (C_{α} RMSD = 0.185 Å).

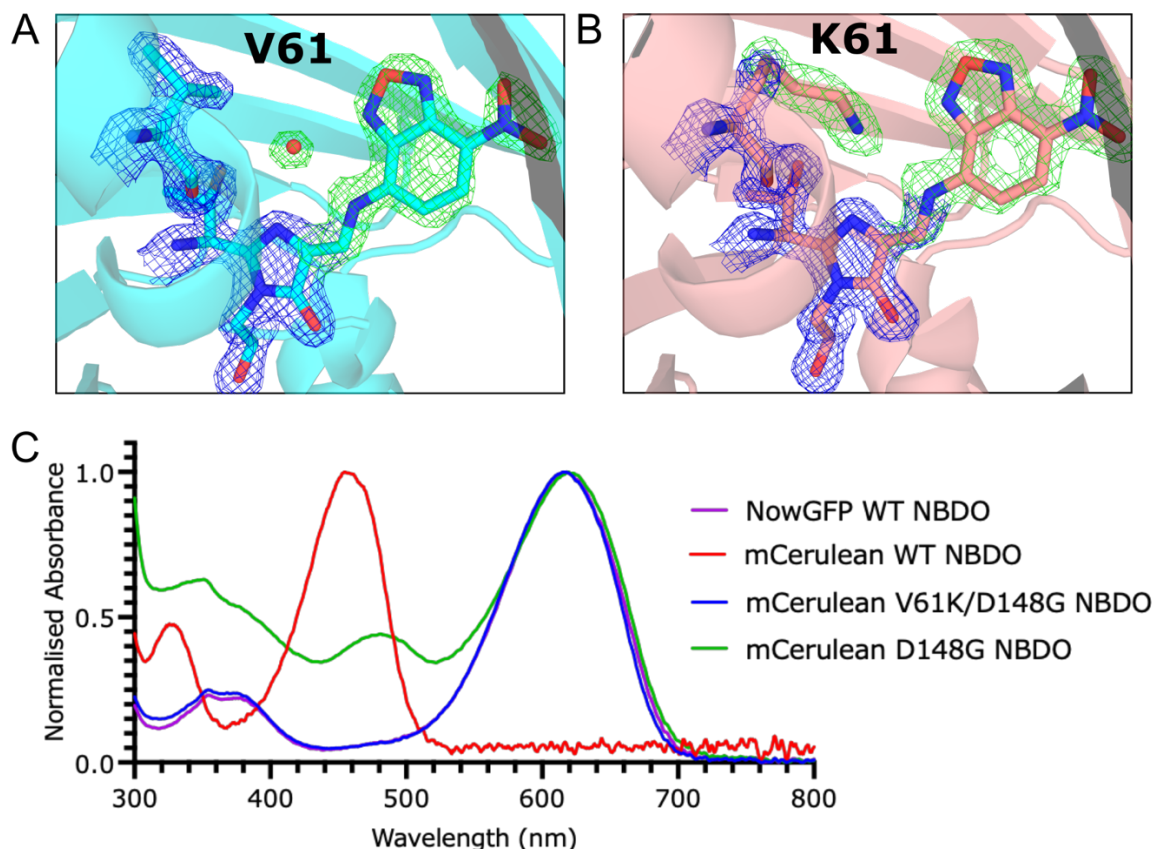


Figure 5.15 – Rescue of chromophore formation by D148G and complete maturation with D148G/V61K. (A) mCerulean D148G NBDO chromophore (cyan) with 2mF_o-DF_c density shown as blue mesh contoured to 2 σ and mF_o-DF_c omit density shown as green mesh contoured to 4 σ, calculated by omitting all chromophore atoms and pictured water. (B) mCerulean D148G/V61K-NBDO chromophore with 2mF_o-DF_c density shown as blue mesh contoured to 2 σ and mF_o-DF_c omit density shown as green mesh, contoured to 5.0 σ, calculating by omitting all chromophore atoms and Lys61 sidechain atoms. (C) UV-Vis absorption spectra normalized to maximum absorbance. mCerulean WT NBDO (red) shows only unmaturing NBDO absorption at 440 nm. mCerulean D148G NBDO (green) shows mature chromophore absorption at 620–640 nm with residual unmaturing NBDO at 440–490 nm. mCerulean D148G/V61K NBDO (blue) matches NowGFP WT NBDO (purple) with complete chromophore maturation.

In the mCerulean D148G NBDO structure, a water molecule is positioned within hydrogen-bonding distance of the exocyclic secondary amine of the mature chromophore (2.8 Å). In the mCerulean D148G/V61K NBDO structure, this water molecule is replaced by the terminal amino group of Lys61. Both crystal structures show fully mature chromophores with unambiguous electron density for the complete

conjugated system (Figure 5.15A,B), despite mCerulean D148G NBDO exhibiting residual unmaturred NBDO absorption at 440 nm in its UV-Vis spectrum.

In the mCerulean D148G/V61K NBDO structure, Lys61 adopts a rotamer outlier conformation ($\chi_1 = -27.8^\circ$, $\chi_2 = -114.7^\circ$), with the χ_1 angle deviating $\sim 30^\circ$ from the nearest preferred rotamer ($\chi_1 = 180^\circ$, $\chi_2 = -120^\circ$).⁶⁸ This conformation positions the Lys61 terminal amine 3.3 Å away from the exocyclic secondary amine. A similar minor occupancy conformation of Lys61 is observed in NowGFP NBDO. In the Trp-containing NowGFP WT, Lys61 stabilises an anionic Trp $N_{\epsilon 1}$.^{61,69–71} This suggests there may be a charge-charge interaction between the lysine and the NBDO chromophore. Lys can act as a neutral hydrogen bond acceptor in protein cores, but this has only been reported with direct experimental evidence twice.^{72,73} The shared chromophore structures, coupled with the complete disappearance of unmaturred NBDO absorption in mCerulean D148G/V61K NBDO suggests that chromophore maturation proceeds fully in the D148G single mutant, but one conformation sampled in solution does not have complete π -delocalization. The water molecule seen in the crystal structure of D148G could be H_3O^+ or H_2O playing the role of Lys61 when it is absent. In solution, the secondary amine equilibrium (protonated/deprotonated) would populate multiple states. Lys61 shifts this equilibrium toward the fully delocalised form, eliminating the 440 nm absorbance peak observed in the D148G single mutant.

5.5.2 Novel chromophore chemistry with NBD amino acids

A plausible chromophore formation pathway (Figure 5.16) can be constructed from the well-known FP chromophore maturation chemistry as well as the UV-vis spectra and structures of the above variants. The reaction begins with the canonical, autocatalytic cyclisation of the protein backbone involving the glycine at position 67, favoured by the canonical helical kink enforced by the β -barrel structure (1), which is not perturbed by the substitution of Trp with NBDO. Arg96 stabilises the dehydration forming the imidazolinone ring (2), maturation does not proceed with the R96A mutant. Maturation proceeds with oxidation by O_2 (3), the D148G mutation allows O_2 diffusion into the β -

barrel core. After oxidation, deprotonation must occur for the fully conjugated species (4), which is favoured by Lys61.

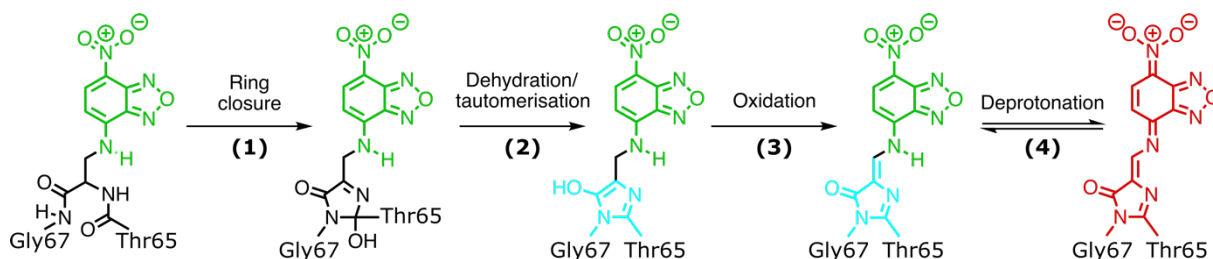


Figure 5.16 – Proposed autocatalytic maturation pathway for the NBDO chromophore in NowGFP.

NBDO chromophore formation proceeds through four steps: (1) nucleophilic attack by Gly67 backbone nitrogen on Thr65 carbonyl forms the imidazolone ring; (2) dehydration eliminates H₂O, tautomerisation forms the imidazole (3) oxidation of the C α -C β bond forms amino-substituted methine bridge; (4) The mature chromophore exists in protonation equilibria (+H⁺/–H⁺) involving the imidazolinone nitrogen and the bridging secondary amine, extending π -conjugation across the NBD and imidazolinone rings. This pathway diverges from canonical GFP chromophores by forming a C=N–CH= bridge instead of a C–CH= bridge.

Canonical GFP chromophores, built from Tyr, Trp, or Phe, mature through C α -C β oxidation that forms a methine bridge directly between the aromatic ring and the imidazolinone (Figure 5.16A). Previous ncAA incorporations at position 66 have followed this same mechanism, producing spectroscopically different proteins through modified aromatic systems but retaining the canonical bridging chemistry. The NBD chromophore is unique: maturation proceeds canonically via oxidation of the C α -C β bond, but a further deprotonation must occur for π -conjugation across the L-2,3-diaminopropionic acid linker (Figure 5.16B). This extended conjugation across the methylene spacer illustrates that chromophore formation can occur from sidechains with an additional nitrogen between the sidechain and the backbone.

For context, Anthozoan chromoproteins achieve extended π -systems compared to standard GFP chromophores (Figure 5.17A) and bathochromic shifts through NH–CH $\rightarrow$ N=C oxidation (Figure 5.17C), but this occurs on the Asn65 backbone rather than through the bridging linker. The rational rescue of chromophore formation in mCerulean through the D148G mutation demonstrates that this new pathway is predictable and amenable to engineering. These results expand the chemical space accessible to autocatalytic

chromophore formation in FPs, establishing that autocatalytic chromophore formation occurs with non-canonical L-2,3-diaminopropionic acid linkers.

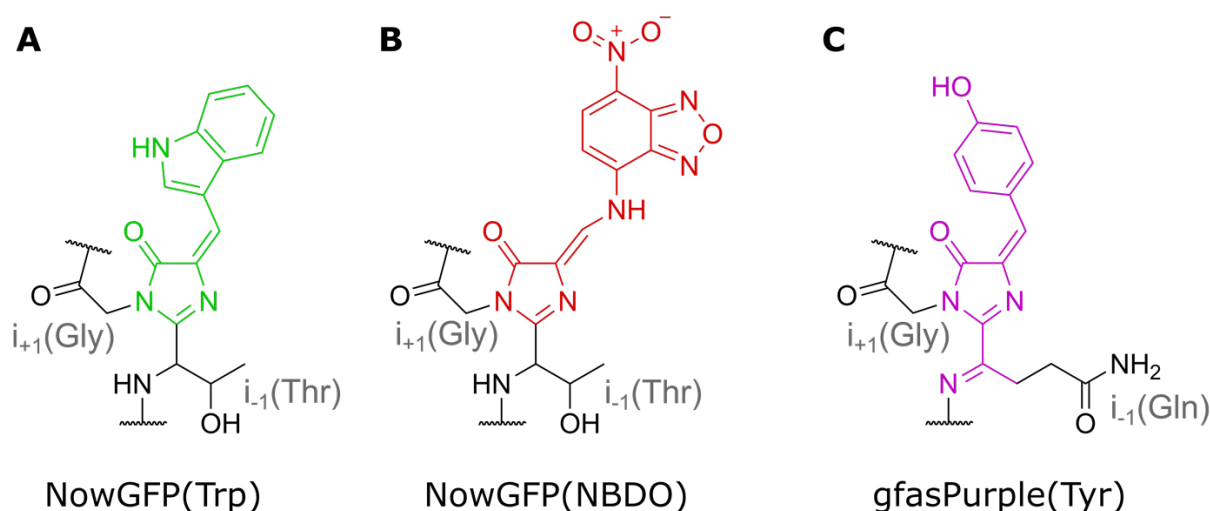


Figure 5.17 – Distinct mature chromophores in fluorescent protein scaffolds. (A) Canonical GFP chromophore derived from Trp65 in NowGFP WT (PDB 4RTC). C_{α} - C_{β} oxidation forms a methine bridge ($C-CH=C$) directly connecting the indole ring to the imidazolinone. (B) NBDO-derived chromophore in NowGFP NBDO showing amino-substituted methine bridge ($C=N-CH=C$) formed via the exocyclic secondary amine. π -Conjugation extends across the extended amino substituted methylene spacer. (C) Chromoprotein chromophore from gfasPurple (Tyr65) showing N-acylimine extension formed by additional N- C_{α} oxidation. All three chromophores share the imidazolinone core but differ in bridging chemistry and aromatic components.

5.6 Conclusion

This work investigates the biophysical principles for engineering minimally perturbing fluorescent reporters through genetic code expansion. The H11 nanobody structures demonstrate that NBD amino acids are pseudo-isosteric to Trp and can replace buried Trp residues while preserving protein fold and function (C_{α} RMSD ≤ 0.45 Å). Solvent exclusion and restricted conformational dynamics in the hydrophobic core produce dramatic fluorescence enhancement: brightness increases 2.3-fold for NBDO and 55-fold for NBDSe relative to unincorporated amino acids in aqueous solution. This validates a labelling strategy that avoids both the burden of fluorescent protein fusions and the non-specificity of chemical labelling.

The NowGFP and mCerulean structures reveal that autocatalytic chromophore formation can accommodate novel conjugation pathways. Structure-guided restoration

of chromophore formation allowed the partial rescue of NBDO chromophore maturation with the mCerulean D148G mutant, and full chromophore maturation was restored with the mCerulean D148G/V61K mutant. The proposed maturation pathway proceeds via the canonical C_{α} – C_{β} bond oxidation step, and following an additional deprotonation step, π -conjugation extends across the amino-substituted methine bridge ($C=N-CH=C$) rather than the direct linker ($C-CH=C$) found in all previous GFP chromophores. This extended conjugation establishes that autocatalytic maturation can proceed in the presence of an L-2,3-diaminopropionic acid linker. These results provide structural evidence for a novel chromophore framework that proceeds autocatalytically, and a structural rationale for placing fluorescent amino acids in a defined microenvironment to enable self-reporting fluorescent nanobodies.

5.7 Methods

5.7.1 Development of translation system to incorporate NBD amino acids

The development of the orthogonal translation system to incorporate the NBD amino acids in Figure 5.1 was developed by Haocheng Qianzhu. Selection was carried out in *E. coli* DH10B cells co-transformed with pBK-G1RS, encoding the G1PylRS mutant library, and pBAD-H6RFP, a reporter plasmid expressing a red fluorescent protein (RFP) with an amber codon inserted after an N-terminal His₆ tag (His₆-TAG-RFP). A series of alternating positive and negative selection rounds were performed using fluorescence-activated cell sorting (FACS) to enrich for active and orthogonal synthetase variants NBDO and NBDSe. Following multiple rounds of selection, colonies were isolated and RFP expression levels were evaluated. The top-performing clones were sequenced and used in the following experiments.

5.7.2 Cloning, expression and protein purification

Protein constructs were cloned, expressed and isolated by collaborators Elwy H. Abdelkader, Haocheng Qianzhu, and Nathan Paul. In short: *E. coli* B-95.ΔAΔfabR cells were co-transformed with pRSF-NBDORS or pRSF-NBDSeRS and pCDF plasmids

containing amber codon-interrupted or wild-type genes.⁷⁴ Cells were cultured in LB medium with 25 mg/L kanamycin and 25 mg/L spectinomycin, supplemented with 0.5 mM NBDO/NBDSe. Expression was induced with 1 mM IPTG at OD₆₀₀ 0.6–1.0 and cells harvested after overnight growth at room temperature. Cell pellets were resuspended in buffer A (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 5% glycerol, 10 mM imidazole), lysed using an Avestin Emulsiflex C5 (10,000–15,000 psi, two passes), and clarified by centrifugation (30,000 g, 60 min, 4°C). Lysate was loaded onto a 1 mL His GraviTrap column (Cytiva), washed with buffer B (buffer A with 20 mM imidazole), and eluted with buffer C (buffer A with 500 mM imidazole).

5.7.3 Size-exclusion chromatography

NowGFP variants were loaded onto a HiLoad 26/600 Superdex 75 pg column (Cytiva, USA) equilibrated with buffer D (20 mM Tris pH 8.0, 200 mM NaCl). Fractions containing NowGFP-WT were identified by SDS-PAGE, combined, and concentrated to 33.6 mg mL⁻¹ using a 10 kDa MWCO Amicon centrifugal filter (Merck Millipore, USA), then diluted to 25 mg mL⁻¹ in buffer D.

ArcNL was run on a HiLoad 26/600 Superdex 75 pg column equilibrated in phosphate-buffered saline (PBS); peak fractions were pooled and concentrated (3 kDa MWCO). H11-NBDO and H11-NBDSe were prepared identically. Equimolar ArcNL and H11 variants were mixed and re-run on the same SEC column; the monodisperse heterodimer peak was verified by SDS-PAGE and concentrated to 10 mg mL⁻¹. Heterodimer stocks were diluted to 5, 2.5 and 1.25 mg mL⁻¹ for crystallisation trials.

5.7.4 Protein crystallisation

For NowGFP, samples of 25 mg mL⁻¹ were subjected to sparse-matrix screening (SG1 Screen HT-96, MD1-89; Molecular Dimensions) at 18 °C used 0.5 µL total drops (1:1 protein:reservoir). Multiple crystals grew within 18 h in 0.2 M ammonium sulfate, 25% PEG 4 K, 0.1 M sodium acetate pH 4.6 (condition C1). After fine-screening, crystals were reproduced by hanging-drop vapour diffusion (1 µL protein + 2 µL well) over 500 µL of

buffer E (35% PEG 3350, 0.2 M ammonium sulfate, 0.1 M sodium acetate pH 4.6). A 5 μ L NowGFP-WT crystal drop was transferred into 50 μ L buffer E and crushed with a melted-tip seed crusher. The slurry was brought to 550 μ L and serially diluted; large hexagonal crystals of NowGFP NBDO grew within 36 h using 10^{-6} seed dilutions in 1 μ L + 1 μ L drops against buffer E. Following the crystallisation of NowGFP NBDO, crystals of NowGFP NBDOSe, NowGFP A150V NBDO, mCerulean D148G NBDO and mCerulean D148G V61K NBDO could be reproduced by using seeds of NowGFP NBDO and fine screening of protein concentration (± 5 mg mL $^{-1}$) and PEG 3350 concentration ($\pm 10\%$).

For ArcNL-H11 (NBDO/NBDOSe), fine-screening by hanging-drop vapour-diffusion (2 μ L total drops) was guided by Hampton pre-crystallisation test conditions and protein concentrations of 5, 2.5 and 1.25 mg mL $^{-1}$. Initial clusters formed in buffer F 0.1 M Tris-HCl pH 8.5, 1 M ammonium sulfate (condition B1) at all protein concentrations. After iterative seeding and ammonium-sulfate gradients, single crystals suitable for data collection were obtained for both ArcNL-H11-NBDO and ArcNL-H11-NBDOSe at 1.25 mg mL $^{-1}$ in 2 M ammonium sulfate, 0.1 M Tris-HCl pH 8.5.

All crystals were cryoprotected with Parabar (Hampton Research) and flash-cooled in liquid nitrogen.

5.7.5 Structure determination and refinement

Diffraction data were collected at 100 K on the Australian Synchrotron MX2 beamline⁷⁵. Data were indexed and integrated with XDS⁷⁶ and scaled with Aimless (CCP4)⁷⁷. Molecular replacement used Phaser (CCP4);⁷⁸ for ArcNL-H11 datasets a suitable homolog (PDB ID: 8QF4) was used as the search model, for both NowGFP and mCerulean, the NowGFP crystal structure was used as the search model (PDB ID: 4RTC). Iterative rebuilding was performed in Coot⁷⁹ and refinement in Phenix⁸⁰. For ArcNL-H11, extensive space group validation⁶⁰ was used to confirm that the non-crystallographic symmetry (NCS) operator was indeed a property of the unit cell in the C 1 2 $_1$ 1 space group, and not higher order crystallographic symmetry.

For ncAA-derived chromophores, CIF restraints were produced by building the molecule in PyMOL,⁸¹ the ground state geometry in Orca using the BL3YP basis set,²⁶ and CIF dictionaries were generated with eLBOW.⁸² Where applicable, anomalous difference Fourier maps were inspected to confirm selenium sites. Data-collection and refinement statistics are reported in Table 5.1–1.5, and coordinates have been deposited under: 9PZZ (ArcNL–H11 W37 NBDO), 9Q00 (ArcNL–H11 W37 NBDSe), 9Q01 (NowGFP W66 NBDO), 9Q02 (NowGFP W66 NBDSe), 9Y97 (mCerulean D148G NBDO) and 9Y98 (mCerulean D148G V61K NBDO).

5.7.6 UV-Vis and fluorescence

UV-Vis and fluorescence spectra were collected by collaborators Elwy H. Abdelkader and Haocheng Qianzhu. UV-Vis spectra and extinction coefficient measurements were performed on a Cary 50 UV-Vis spectrophotometer (Agilent, USA) in PBS, with a cuvette path length of 1 cm. Fluorescence measurements were performed on a Cary Eclipse fluorescence spectrophotometer (Agilent, USA), using samples diluted to 10 μ M in PBS with a cuvette path length of 1 cm. Extinction coefficients were measured in PBS in triplicate. Quantum yields (QYs) were determined using the protocol of Würth *et. al.*⁸³ using either tryptophan (280 nm) or quinine sulfate (305 nm) as reference standard. Brightness is calculated as the product of extinction coefficient and QY.

5.8 Crystallographic statistics

Table 5.1 - Table of crystallographic statistics of ArcNL-H11 W37 NBDO and ArcNL-H11 W37 NBDS_e.

	ArcNL-H11 W37 NBDO	ArcNL-H11 W37 NBDS _e
PDB ID	9PZZ	9Q00
Data collection		
Source	Australian Synchrotron MX2	Australian Synchrotron MX2
Space group	C 1 2 ₁ 1	C 1 2 ₁ 1
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	113.0 44.1 88.8	113.8 44.3 89.0
α , β , γ (°)	90.0 92.7 90.0	90.0 98.2 90.0
Resolution (Å)	48.68 – 1.56 (1.58 – 1.56)	44.65 – 2.84 (3.13 – 2.84)
<i>R</i> _{merge}	0.105 (1.107)	0.433 (1.107)
<i>R</i> _{pim}	0.066 (0.690)	0.286 (0.755)
<i>I</i> / σ <i>I</i>	11.5 (1.6)	6.1 (1.7)
CC _{1/2}	0.998 (0.442)	0.951 (0.434)
Completeness (%)	100.0 (100.0)	99.9 (99.8)
Redundancy	6.8 (6.9)	6.1 (5.7)
Refinement		
Resolution (Å)	48.68 – 1.56 (1.58 – 1.56)	44.65 – 2.84 (3.13 – 2.84)
No. reflections	62461 (2704)	10612 (2634)
<i>R</i> _{work} / <i>R</i> _{free}	0.1856 / 0.2267 (0.2962 / 0.3437)	0.2156 / 0.2807 (0.2156 / 0.3933)
Number of atoms		
Protein	3138	3109
Ligand/ion	36	36
Solvent	683	115
<i>B</i> -factors (overall)	23.8	34.7
Protein	21.8	35.2
Ligand/ion	14.1	27.3
Solvent	33.7	21.7
RMS deviations		
Bond lengths (Å)	0.008	0.003
Bond angles (°)	0.83	0.66

^aValues in parentheses are for the highest resolution shell.

Table 5.2 - Table of crystallographic statistics of NowGFP NBDO and NowGFP NBDS_e

	NowGFP W66 NBDO	NowGFP W66 NBDS _e
PDB ID	9Q01	9Q02
Data collection		
Source	Australian Synchrotron MX2	Australian Synchrotron MX2
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	51.80 52.00 196.70	50.76 50.83 192.95
α , β , γ (°)	90.0 90.0 90.0	90.0 90.0 90.0
Resolution (Å)	49.18 – 1.36 (1.38 – 1.36)	49.15– 1.87 (1.91 – 1.87)
<i>R</i> _{merge}	0.108 (2.255)	0.150 (2.058)
<i>R</i> _{pim}	0.032 (0.911)	0.062 (0.843)
<i>I</i> / σ <i>I</i>	13.7 (1.3)	12.4 (0.9)
CC _{1/2}	0.999 (0.338)	0.999 (0.450)
Completeness (%)	100.0 (100.0)	100.0 (100.0)
Redundancy	13.3 (13.7)	12.8 (13.3)
Refinement		
Resolution (Å)	49.18 – 1.36 (1.38 – 1.36)	49.15– 1.87 (1.91 – 1.87)
No. reflections	114914 (3759)	42304 (2581)
<i>R</i> _{work} / <i>R</i> _{free}	0.1646 / 0.1895 (0.3372 / 0.3580)	0.1688 / 0.2105 (0.3002 / 0.3325)
Number of atoms		
Protein	3774	3720
Ligand/ion	60	63
Solvent	794	617
<i>B</i> -factors (overall)	22.6	31.6
Protein	20.5	30.3
Ligand/ion	14.5	25.7
Solvent	33.14	39.9
RMS deviations		
Bond lengths (Å)	0.007	0.007
Bond angles (°)	0.90	0.92

^aValues in parentheses are for the highest resolution shell.

Table 5.3 - Table of crystallographic statistics of mCerulean D148G NBDO and mCerulean D148G/V61K NBDO.

	mCerulean D148G NBDO	mCerulean D148G/V61K NBDO
PDB ID	9Y97	9Y98
Data collection		
Source	Australian Synchrotron MX2	Australian Synchrotron MX2
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁
Cell dimensions		
a, b, c (Å)	51.5 62.3 65.9	51.2 62.7 69.4
α, β, γ (°)	90.1 89.9 90.1	90.0 90.0 90.0
Resolution (Å)	45.43 – 1.74 (1.77 – 1.74)	46.52 – 1.56 (1.61 – 1.56)
R_{merge}	0.232 (3.528)	0.257 (1.773)
R_{pim}	0.093 (1.427)	0.116 (0.784)
$I/\sigma I$	13.4 (1.1)	9.6 (1.7)
$CC_{1/2}$	0.998 (0.500)	0.996 (0.303)
Completeness (%)	99.8 (94.8)	99.8 (100.0)
Redundancy	13.4 (13)	11.4 (11.8)
Refinement		
Resolution (Å)	45.43 – 1.74 (1.77 – 1.74)	46.52 – 1.56 (1.61 – 1.56)
No. reflections	22530 (2718)	32415 (2655)
$R_{\text{work}}/R_{\text{free}}$	0.1580 / 0.1938 (0.2618 / 0.2976)	0.1849 / 0.2209 (0.3133 / 0.3392)
Number of atoms		
Protein	1842	1857
Ligand/ion	30	28
Solvent	289	315
B -factors (overall)	26.9	20.9
Protein	25.5	19.5
Ligand/ion	23.2	12.9
Solvent	36.12	29.2
RMS deviations		
Bond lengths (Å)	0.007	0.007
Bond angles (°)	0.92	0.91

^aValues in parentheses are for the highest resolution shell.

5.9 References

1. Schermelleh, L.; Ferrand, A.; Huser, T.; Eggeling, C.; Sauer, M.; Biehlmaier, O.; Drummen, G. P. C. Super-resolution microscopy demystified. *Nat. Cell Biol.* **2019**, *21*, 72–84. <https://doi.org/10.1038/s41556-018-0251-8>
2. Duuren, B. L. v. Solvent effects in the fluorescence of indole and substituted indoles. *J. Org. Chem.* **1961**, *26*, 2954–2960. <https://doi.org/10.1021/jo01066a079>
3. Haidekker, M. A.; Theodorakis, E. A. Environment-sensitive behavior of fluorescent molecular rotors. *J. Biol. Eng.* **2010**, *4*, 11. <https://doi.org/10.1186/1754-1611-4-11>
4. Han, J.; Burgess, K. Fluorescent indicators for intracellular pH. *Chem. Rev.* **2010**, *110*, 2709–2728. <https://doi.org/10.1021/cr900249z>
5. Doose, S.; Neuweiler, H.; Sauer, M. Fluorescence quenching by photoinduced electron transfer: A reporter for conformational dynamics of macromolecules. *ChemPhysChem* **2009**, *10*, 1389–1398. <https://doi.org/10.1002/cphc.200900238>
6. Giepmans, B. N.; Adams, S. R.; Ellisman, M. H.; Tsien, R. Y. The fluorescent toolbox for assessing protein location and function. *Science* **2006**, *312*, 217–224. <https://doi.org/10.1126/science.1124618>
7. Sankaranarayanan, S.; De Angelis, D.; Rothman, J. E.; Ryan, T. A. The use of pHluorins for optical measurements of presynaptic activity. *Biophys. J.* **2000**, *79*, 2199–2208. [https://doi.org/10.1016/S0006-3495\(00\)76468-X](https://doi.org/10.1016/S0006-3495(00)76468-X)
8. Adams, S. R.; Campbell, R. E.; Gross, L. A.; Martin, B. R.; Walkup, G. K.; Yao, Y.; Llopis, J.; Tsien, R. Y. New biarsenical ligands and tetracysteine motifs for protein labeling *in vitro* and *in vivo*: Synthesis and biological applications. *J. Am. Chem. Soc.* **2002**, *124*, 6063–6076. <https://doi.org/10.1021/ja017687n>
9. Grimm, J. B.; Lavis, L. D. Caveat fluorophore: An insiders' guide to small-molecule fluorescent labels. *Nat. Methods* **2022**, *19*, 149–158. <https://doi.org/10.1038/s41592-021-01338-6>
10. Jing, C.; Cornish, V. W. Chemical tags for labeling proteins inside living cells. *Acc. Chem. Res.* **2011**, *44*, 784–792. <https://doi.org/10.1021/ar200099f>
11. Ren, H.; Xiao, F.; Zhan, K.; Kim, Y. P.; Xie, H.; Xia, Z.; Rao, J. A biocompatible condensation reaction for the labeling of terminal cysteine residues on proteins. *Angew. Chem., Int. Ed.* **2009**, *48*, 9658–9662. <https://doi.org/10.1002/anie.200903627>
12. Gunnoo, S. B.; Madder, A. Chemical protein modification through cysteine. *ChemBioChem* **2016**, *17*, 529–553. <https://doi.org/10.1002/cbic.201500667>
13. Pedelacq, J. D.; Cabantous, S.; Tran, T.; Terwilliger, T. C.; Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **2006**, *24*, 79–88. <https://doi.org/10.1038/nbt1172>
14. Huang, L.; Pike, D.; Sleat, D. E.; Nanda, V.; Lobel, P. Potential pitfalls and solutions for use of fluorescent fusion proteins to study the lysosome. *PLoS One* **2014**, *9*, e88893. <https://doi.org/10.1371/journal.pone.0088893>
15. Zacharias, D. A.; Violin, J. D.; Newton, A. C.; Tsien, R. Y. Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells. *Science* **2002**, *296*, 913–916. <https://doi.org/10.1126/science.1068539>
16. Agbulut, O.; Huet, A.; Niederlander, N.; Puceat, M.; Menasche, P.; Coirault, C. Green fluorescent protein impairs actin-myosin interactions by binding to the actin-binding site of myosin. *J. Biol. Chem.* **2007**, *282*, 10465–10471. <https://doi.org/10.1074/jbc.M610418200>
17. Jiang, C.; Huang, H.; Kang, X.; Yang, L.; Xi, Z.; Sun, H.; Pluth, M. D.; Yi, L. NBD-based synthetic probes for sensing small molecules and proteins: Design, sensing mechanisms and biological applications. *Chem. Soc. Rev.* **2021**, *50*, 7436–7495. <https://doi.org/10.1039/d0cs01096k>
18. Mukherjee, S.; Chattopadhyay, A.; Samanta, A.; Soujanya, T. Dipole moment change of NBD group upon excitation studied using solvatochromic and quantum chemical approaches: Implications in membrane research. *J. Phys. Chem.* **1994**, *98*, 2809–2812. <https://doi.org/10.1021/j100062a014>
19. Amaro, M.; Filipe, H. A.; Prates Ramalho, J. P.; Hof, M.; Loura, L. M. Fluorescence of nitrobenzoxadiazole (NBD)-labeled lipids in model membranes is connected not to lipid mobility but to probe location. *Phys. Chem. Chem. Phys.* **2016**, *18*, 7042–7054. <https://doi.org/10.1039/c5cp05238f>

20. Ghosh, P. B.; Whitehouse, M. W. 7-chloro-4-nitrobenzo-2-oxa-1,3-diazole: A new fluorogenic reagent for amino acids and other amines. *Biochem. J.* **1968**, *108*, 155–156. <https://doi.org/10.1042/bj1080155>
21. Fery-Forgues, S.; Fayet, J.-P.; Lopez, A. Drastic changes in the fluorescence properties of NBD probes with the polarity of the medium: Involvement of a TICT state? *J. Photochem. Photobiol., A* **1993**, *70*, 229–243. [https://doi.org/10.1016/1010-6030\(93\)85048-D](https://doi.org/10.1016/1010-6030(93)85048-D)
22. Ludvikova, L.; Simon, E.; Deygas, M.; Panier, T.; Plamont, M. A.; Ollion, J.; Tebo, A.; Piel, M.; Jullien, L.; Robert, L.; Le Saux, T.; Espagne, A. Near-infrared co-illumination of fluorescent proteins reduces photobleaching and phototoxicity. *Nat. Biotechnol.* **2024**, *42*, 872–876. <https://doi.org/10.1038/s41587-023-01893-7>
23. Debarre, D.; Olivier, N.; Supatto, W.; Beaurepaire, E. Mitigating phototoxicity during multiphoton microscopy of live *Drosophila* embryos in the 1.0–1.2 μm wavelength range. *PLoS One* **2014**, *9*, e104250. <https://doi.org/10.1371/journal.pone.0104250>
24. Jun, Y. W.; Kim, H. R.; Reo, Y. J.; Dai, M.; Ahn, K. H. Addressing the autofluorescence issue in deep tissue imaging by two-photon microscopy: The significance of far-red emitting dyes. *Chem. Sci.* **2017**, *8*, 7696–7704. <https://doi.org/10.1039/c7sc03362a>
25. de Moliner, F.; Konieczna, Z.; Mendive-Tapia, L.; Saleeb, R. S.; Morris, K.; Gonzalez-Vera, J. A.; Kaizuka, T.; Grant, S. G. N.; Horrocks, M. H.; Vendrell, M. Small fluorogenic amino acids for peptide-guided background-free imaging. *Angew. Chem., Int. Ed.* **2023**, *62*, e202216231. <https://doi.org/10.1002/anie.202216231>
26. Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, *152*, 224108. <https://doi.org/10.1063/5.0004608>
27. Bader, R. F. W. A quantum theory of molecular structure and its applications. *Chem. Rev.* **1991**, *91*, 893–928. <https://doi.org/10.1021/cr00005a013>
28. Bragg, P. D.; Hou, C. Effect of NBD chloride (4-chloro-7-nitrobenzo-2-oxa-1,3-diazole) on the pyridine nucleotide transhydrogenase of *Escherichia coli*. *Biochim. Biophys. Acta* **1999**, *1413*, 159–171. [https://doi.org/10.1016/S0005-2728\(99\)00090-0](https://doi.org/10.1016/S0005-2728(99)00090-0)
29. Turcatti, G.; Nemeth, K.; Edgerton, M. D.; Meseth, U.; Talabot, F.; Peitsch, M.; Knowles, J.; Vogel, H.; Chollet, A. Probing the structure and function of the tachykinin neurokinin-2 receptor through biosynthetic incorporation of fluorescent amino acids at specific sites. *J. Biol. Chem.* **1996**, *271*, 19991–19998. <https://doi.org/10.1074/jbc.271.33.19991>
30. Bernal-Perez, L. F.; Prokai, L.; Ryu, Y. Selective N-terminal fluorescent labeling of proteins using 4-chloro-7-nitrobenzofurazan: A method to distinguish protein N-terminal acetylation. *Anal. Biochem.* **2012**, *428*, 13–15. <https://doi.org/10.1016/j.ab.2012.05.026>
31. Örmö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science* **1996**, *273*, 1392–1395. <https://doi.org/10.1126/science.273.5280.1392>
32. Tsien, R. Y. The green fluorescent protein. *Annu. Rev. Biochem.* **1998**, *67*, 509–544. <https://doi.org/10.1146/annurev.biochem.67.1.509>
33. Barondeau, D. P.; Putnam, C. D.; Kassmann, C. J.; Tainer, J. A.; Getzoff, E. D. Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12111–12116. <https://doi.org/10.1073/pnas.2133463100>
34. Sniegowski, J. A.; Lappe, J. W.; Patel, H. N.; Huffman, H. A.; Wachter, R. M. Base catalysis of chromophore formation in Arg96 and Glu222 variants of green fluorescent protein. *J. Biol. Chem.* **2005**, *280*, 26248–26255. <https://doi.org/10.1074/jbc.M412327200>
35. Grigorenko, B. L.; Krylov, A. I.; Nemukhin, A. V. Molecular modeling clarifies the mechanism of chromophore maturation in the green fluorescent protein. *J. Am. Chem. Soc.* **2017**, *139*, 10239–10249. <https://doi.org/10.1021/jacs.7b00676>
36. Bartkiewicz, M.; Kazazic, S.; Krasowska, J.; Clark, P. L.; Wielgus-Kutrowska, B.; Bzowska, A. Non-fluorescent mutant of green fluorescent protein sheds light on the mechanism of chromophore formation. *FEBS Lett.* **2018**, *592*, 1516–1523. <https://doi.org/10.1002/1873-3468.13051>

37. Subach, O. M.; Gundorov, I. S.; Yoshimura, M.; Subach, F. V.; Zhang, J.; Gruenwald, D.; Souslova, E. A.; Chudakov, D. M.; Verkhusha, V. V. Conversion of red fluorescent protein into a bright blue probe. *Chem. Biol.* **2008**, *15*, 1116–1124. <https://doi.org/10.1016/j.chembiol.2008.08.006>
38. Shcherbakova, D. M.; Verkhusha, V. V. Near-infrared fluorescent proteins for multicolor in vivo imaging. *Nat. Methods* **2013**, *10*, 751–754. <https://doi.org/10.1038/nmeth.2521>
39. Cranfill, P. J.; Sell, B. R.; Baird, M. A.; Allen, J. R.; Lavagnino, Z.; de Gruiter, H. M.; Kremers, G. J.; Davidson, M. W.; Ustione, A.; Piston, D. W. Quantitative assessment of fluorescent proteins. *Nat. Methods* **2016**, *13*, 557–562. <https://doi.org/10.1038/nmeth.3891>
40. Henderson, J. N.; Remington, S. J. Crystal structures and mutational analysis of amFP486, a cyan fluorescent protein from *Anemonia majano*. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12712–12717. <https://doi.org/10.1073/pnas.0502250102>
41. Costantini, L. M.; Balaban, M.; Markwardt, M. L.; Rizzo, M. A.; Guo, F.; Verkhusha, V. V.; Snapp, E. L. A palette of fluorescent proteins optimized for diverse cellular environments. *Nat. Commun.* **2015**, *6*, 7670. <https://doi.org/10.1038/ncomms8670>
42. Patterson, G. H.; Lippincott-Schwartz, J. A photoactivatable GFP for selective photolabeling of proteins and cells. *Science* **2002**, *297*, 1873–1877. <https://doi.org/10.1126/science.1074952>
43. Shaner, N. C.; Campbell, R. E.; Steinbach, P. A.; Giepmans, B. N.; Palmer, A. E.; Tsien, R. Y. Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nat. Biotechnol.* **2004**, *22*, 1567–1572. <https://doi.org/10.1038/nbt1037>
44. Campbell, R. E.; Tour, O.; Palmer, A. E.; Steinbach, P. A.; Baird, G. S.; Zacharias, D. A.; Tsien, R. Y. A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7877–7882. <https://doi.org/10.1073/pnas.082243699>
45. Lambert, T. J. FPbase: A community-editable fluorescent protein database. *Nat. Methods* **2019**, *16*, 277–278. <https://doi.org/10.1038/s41592-019-0352-8>
46. Matz, M. V.; Fradkov, A. F.; Labas, Y. A.; Savitsky, A. P.; Zaraisky, A. G.; Markelov, M. L.; Lukyanov, S. A. Fluorescent proteins from nonbioluminescent Anthozoa species. *Nat. Biotechnol.* **1999**, *17*, 969–973. <https://doi.org/10.1038/13657>
47. Verkhusha, V. V.; Lukyanov, K. A. The molecular properties and applications of Anthozoa fluorescent proteins and chromoproteins. *Nat. Biotechnol.* **2004**, *22*, 289–296. <https://doi.org/10.1038/nbt943>
48. Ahmed, F. H.; Caputo, A. T.; French, N. G.; Peat, T. S.; Whitfield, J.; Warden, A. C.; Newman, J.; Scott, C. Over the rainbow: structural characterization of the chromoproteins gfasPurple, amilCP, spisPink and eforRed. *Acta Crystallogr. D Struct. Biol.* **2022**, *78*, 599–612. <https://doi.org/10.1107/S2059798322002625>
49. Yang, X.; Yang, N.; Wu, S.; He, S.; Jiang, W.; Zhong, W.; Du, J.; Zhang, G.; Wang, X.; Yang, S. Establishment of a visible reporter system in *Zymomonas mobilis* through random mutagenesis and rational design of a chromoprotein eforRed. *ACS Synth. Biol.* **2025**, *14*, 2548–2558. <https://doi.org/10.1021/acssynbio.4c00872>
50. Olenginski, G. M.; Piacentini, J.; Harris, D. R.; Runko, N. A.; Papoutsis, B. M.; Alter, J. R.; Hess, K. R.; Brewer, S. H.; Phillips-Piro, C. M. Structural and spectrophotometric investigation of two unnatural amino-acid altered chromophores in the superfolder green fluorescent protein. *Acta Crystallogr. D Struct. Biol.* **2021**, *77*, 1010–1018. <https://doi.org/10.1107/S2059798321006525>
51. Pal, P. P.; Bae, J. H.; Azim, M. K.; Hess, P.; Friedrich, R.; Huber, R.; Moroder, L.; Budisa, N. Structural and spectral response of *Aequorea victoria* green fluorescent proteins to chromophore fluorination. *Biochemistry* **2005**, *44*, 3663–3672. <https://doi.org/10.1021/bi0484825>
52. Reddington, S. C.; Rizkallah, P. J.; Watson, P. D.; Pearson, R.; Tippmann, E. M.; Jones, D. D. Different photochemical events of a genetically encoded phenyl azide define and modulate GFP fluorescence. *Angew. Chem., Int. Ed.* **2013**, *52*, 5974–5977. <https://doi.org/10.1002/anie.201301490>
53. Wang, F.; Niu, W.; Guo, J.; Schultz, P. G. Unnatural amino acid mutagenesis of fluorescent proteins. *Angew. Chem., Int. Ed.* **2012**, *51*, 10132–10135. <https://doi.org/10.1002/anie.201204668>
54. Bae, J. H.; Paramita Pal, P.; Moroder, L.; Huber, R.; Budisa, N. Crystallographic evidence for isomeric chromophores in 3-fluorotyrosyl-green fluorescent protein. *ChemBioChem* **2004**, *5*, 720–722. <https://doi.org/10.1002/cbic.200300818>

55. Zhang, H.; Tian, X.; Zhang, J.; Ai, H. W. Engineering and characterization of 3-aminotyrosine-derived red fluorescent variants of circularly permuted green fluorescent protein. *Biosensors (Basel)* **2024**, *14*, 54. <https://doi.org/10.3390/bios14010054>
56. Oscar, B. G.; Zhu, L.; Wolfenden, H.; Rozanov, N. D.; Chang, A.; Stout, K. T.; Sandwisch, J. W.; Porter, J. J.; Mehl, R. A.; Fang, C. Dissecting optical response and molecular structure of fluorescent proteins with non-canonical chromophores. *Front. Mol. Biosci.* **2020**, *7*, 131. <https://doi.org/10.3389/fmolb.2020.00131>
57. Wang, L.; Xie, J.; Deniz, A. A.; Schultz, P. G. Unnatural amino acid mutagenesis of green fluorescent protein. *J. Org. Chem.* **2003**, *68*, 174–176. <https://doi.org/10.1021/jo026570u>
58. Godoy Munoz, J. M.; Neset, L.; Markusson, S.; Weber, S.; Krokengen, O. C.; Sutinen, A.; Christakou, E.; Lopez, A. J.; Bramham, C. R.; Kursula, P. Structural characterization of two nanobodies targeting the ligand-binding pocket of human Arc. *PLoS One* **2024**, *19*, e0300453. <https://doi.org/10.1371/journal.pone.0300453>
59. Vazdarjanova, A.; Ramirez-Amaya, V.; Insel, N.; Plummer, T. K.; Rosi, S.; Chowdhury, S.; Mikhael, D.; Worley, P. F.; Guzowski, J. F.; Barnes, C. A. Spatial exploration induces ARC, a plasticity-related immediate-early gene, only in calcium/calmodulin-dependent protein kinase II-positive principal excitatory and inhibitory neurons of the rat forebrain. *J. Comp. Neurol.* **2006**, *498*, 317–329. <https://doi.org/10.1002/cne.21003>
60. Lebedev, A. A.; Isupov, M. N. Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, *70*, 2430–2443. <https://doi.org/10.1107/S1399004714014795>
61. Pletnev, V. Z.; Pletneva, N. V.; Sarkisyan, K. S.; Mishin, A. S.; Lukyanov, K. A.; Goryacheva, E. A.; Ziganshin, R. H.; Dauter, Z.; Pletnev, S. Structure of the green fluorescent protein NowGFP with an anionic tryptophan-based chromophore. *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71*, 1699–1707. <https://doi.org/10.1107/S1399004715010159>
62. Di Donato, M.; van Wilderen, L. J.; Van Stokkum, I. H.; Stuart, T. C.; Kennis, J. T.; Hellingwerf, K. J.; van Grondelle, R.; Groot, M. L. Proton transfer events in GFP. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16295–16305. <https://doi.org/10.1039/c1cp20387h>
63. Rekas, A.; Alattia, J. R.; Nagai, T.; Miyawaki, A.; Ikura, M. Crystal structure of Venus, a yellow fluorescent protein with improved maturation and reduced environmental sensitivity. *J. Biol. Chem.* **2002**, *277*, 50573–50578. <https://doi.org/10.1074/jbc.M209524200>
64. Seifert, M. H.; Georgescu, J.; Ksiazek, D.; Smialowski, P.; Rehm, T.; Steipe, B.; Holak, T. A. Backbone dynamics of green fluorescent protein and the effect of histidine 148 substitution. *Biochemistry* **2003**, *42*, 2500–2512. <https://doi.org/10.1021/bi026481b>
65. Auhim, H. S.; Grigorenko, B. L.; Harris, T. K.; Aksakal, O. E.; Polyakov, I. V.; Berry, C.; Gomes, G. D. P.; Alabugin, I. V.; Rizkallah, P. J.; Nemukhin, A. V.; Jones, D. D. Stalling chromophore synthesis of the fluorescent protein Venus reveals the molecular basis of the final oxidation step. *Chem. Sci.* **2021**, *12*, 7735–7745. <https://doi.org/10.1039/d0sc06693a>
66. Rizzo, M. A.; Springer, G. H.; Granada, B.; Piston, D. W. An improved cyan fluorescent protein variant useful for FRET. *Nat. Biotechnol.* **2004**, *22*, 445–449. <https://doi.org/10.1038/nbt945>
67. Craggs, T. D. Green fluorescent protein: Structure, folding and chromophore maturation. *Chem. Soc. Rev.* **2009**, *38*, 2865–2875. <https://doi.org/10.1039/b903641p>
68. Hintze, B. J.; Lewis, S. M.; Richardson, J. S.; Richardson, D. C. MolProbity's ultimate rotamer-library distributions for model validation. *Proteins* **2016**, *84*, 1177–1189. <https://doi.org/10.1002/prot.25039>
69. Sarkisyan, K. S.; Goryashchenko, A. S.; Lidsky, P. V.; Gorbachev, D. A.; Bozhanova, N. G.; Gorokhovatsky, A. Y.; Pereverzeva, A. R.; Ryumina, A. P.; Zherdeva, V. V.; Savitsky, A. P.; Solntsev, K. M.; Bommarius, A. S.; Sharonov, G. V.; Lindquist, J. R.; Drobizhev, M.; Hughes, T. E.; Rebane, A.; Lukyanov, K. A.; Mishin, A. S. Green fluorescent protein with anionic tryptophan-based chromophore and long fluorescence lifetime. *Biophys. J.* **2015**, *109*, 380–389. <https://doi.org/10.1016/j.bpj.2015.06.018>
70. Sarkisyan, K. S.; Yampolsky, I. V.; Solntsev, K. M.; Lukyanov, S. A.; Lukyanov, K. A.; Mishin, A. S. Tryptophan-based chromophore in fluorescent proteins can be anionic. *Sci. Rep.* **2012**, *2*, 608. <https://doi.org/10.1038/srep00608>

71. Laricheva, E. N.; Goh, G. B.; Dickson, A.; Brooks, C. L., III. pH-dependent transient conformational states control optical properties in cyan fluorescent protein. *J. Am. Chem. Soc.* **2015**, *137*, 2892–2900. <https://doi.org/10.1021/ja509233r>
72. Nagae, T.; Takeda, M.; Noji, T.; Saito, K.; Aoyama, H.; Miyanoiri, Y.; Ito, Y.; Kainosho, M.; Hirose, Y.; Ishikita, H.; Mishima, M. Direct evidence for a deprotonated lysine serving as a H-bond "acceptor" in a photoreceptor protein. *Proc. Natl. Acad. Sci. U.S.A.* **2024**, *121*, e2404472121. <https://doi.org/10.1073/pnas.2404472121>
73. Takayama, Y.; Castaneda, C. A.; Chimenti, M.; Garcia-Moreno, B.; Iwahara, J. Direct evidence for deprotonation of a lysine side chain buried in the hydrophobic core of a protein. *J. Am. Chem. Soc.* **2008**, *130*, 6714–6715. <https://doi.org/10.1021/ja801731g>
74. Mukai, T.; Hoshi, H.; Ohtake, K.; Takahashi, M.; Yamaguchi, A.; Hayashi, A.; Yokoyama, S.; Sakamoto, K. Highly reproductive *Escherichia coli* cells with no specific assignment to the UAG codon. *Sci. Rep.* **2015**, *5*, 9699. <https://doi.org/10.1038/srep09699>
75. McPhillips, T. M.; McPhillips, S. E.; Chiu, H. J.; Cohen, A. E.; Deacon, A. M.; Ellis, P. J.; Garman, E.; Gonzalez, A.; Sauter, N. K.; Phizackerley, R. P.; Soltis, S. M.; Kuhn, P. Blu-Ice and the distributed control system: Software for data acquisition and instrument control at macromolecular crystallography beamlines. *J. Synchrotron Radiat.* **2002**, *9*, 401–406. <https://doi.org/10.1107/s0909049502015170>
76. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 125–132. <https://doi.org/10.1107/S0907444909047337>
77. Evans, P. R.; Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 1204–1214. <https://doi.org/10.1107/S0907444913000061>
78. McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser crystallographic software. *J. Appl. Crystallogr.* **2007**, *40*, 658–674. <https://doi.org/10.1107/S0021889807021206>
79. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 486–501. <https://doi.org/10.1107/S0907444910007493>
80. Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkoczi, G.; Chen, V. B.; Croll, T. I.; Hintze, B.; Hung, L. W.; Jain, S.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R. D.; Poon, B. K.; Prisant, M. G.; Read, R. J.; Richardson, J. S.; Richardson, D. C.; Sammito, M. D.; Sobolev, O. V.; Stockwell, D. H.; Terwilliger, T. C.; Urzhumtsev, A. G.; Videau, L. L.; Williams, C. J.; Adams, P. D. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **2019**, *75*, 861–877. <https://doi.org/10.1107/S2059798319011471>
81. The PyMOL Molecular Graphics System, Version 3.0, Schrödinger, LLC.
82. Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. Electronic Ligand Builder and Optimization Workbench (eLBOW): A tool for ligand coordinate and restraint generation. *Acta Crystallogr. D Biol. Crystallogr.* **2009**, *65*, 1074–1080. <https://doi.org/10.1107/S0907444909029436>
83. Würth, C.; Grabolle, M.; Pauli, J.; Spieles, M.; Resch-Genger, U. Relative and absolute determination of fluorescence quantum yields of transparent samples. *Nat. Protoc.* **2013**, *8*, 1535–1550. <https://doi.org/10.1038/nprot.2013.087>

Chapter 6: Conclusion and future directions

6.1 Conclusion and future directions

This thesis elucidates biophysical principles for the design of photoluminescent proteins. Genetic code expansion enabled systematic characterisation of fluorescent non-canonical amino acids as lanthanide-sensitising antennae. Then, *de novo* design yielded a compact lanthanide-binding mini-protein featuring a non-canonical antenna with desired photoluminescent properties. A new methodology for inferring indole sidechain orientations at a dynamic site was developed. The structural consequences of incorporating tryptophan analogues were determined through X-ray crystallography. Visible-light fluorescent amino acids were introduced into proteins, demonstrating enhanced photophysical properties in a nanobody and a novel autocatalytic chromophore formation in a fluorescent protein.

The first part of Chapter 2 identified 6-azatryptophan as an exceptional sensitiser and quantified how antenna structure trades off energy-transfer efficiency against lanthanide-binding affinity. Future work should decompose the S_1 and T_1 contributions to energy transfer, alongside determining the quantum yield of each complex. This would resolve the photophysical pathways that ultimately govern antenna choice. As the spectral properties of S_1 and T_1 may correlate with local environment, such analyses could underpin rational design of lanthanide-luminescent proteins as environment-responsive probes.

The second part of Chapter 2 developed RF2, a *de novo* mini-protein with strong, lanthanide-selective binding affinity that maintains robust binding despite variation of the antenna chromophore. RF2 provides a compact, genetically encodable binding scaffold for lanthanide ions, producing robust lanthanide luminescence in a small protein. Such a small, bright, lanthanide-luminescent protein is well suited for future applications in time-resolved microscopy, visualisation of slow biomolecular processes, and multiplexed labelling where long-lived emission is advantageous. Future iterations of this design should decrease the q -value to ~ 0 , enhancing luminescence and likely increasing binding affinity. Determining the precise lanthanide coordination environment in RF2 at

will clarify the structural basis of its modest ion discrimination and guide design of proteins with stronger lanthanide selectivity.

Chapter 3 applied ^{19}F - Gd^{3+} ENDOR spectroscopy to resolve rotamer populations of the tryptophan sidechain in a highly dynamic, solvent-exposed position. By integrating genetic code expansion, clash-filtered rotamer libraries and ENDOR-derived distance distributions, this work showed that low-population χ_1/χ_2 conformations of a tryptophan side chain can be distinguished and quantitatively analysed in a protein environment. This establishes a general framework for using ^{19}F - Gd^{3+} ENDOR to read out sidechain conformations in systems where crystal structures are difficult to obtain, or multiple conformational states coexist. Beyond protein design, understanding how fluorinated groups are oriented and accommodated in protein contexts is directly relevant to drug design, where many small-molecule therapeutics contain fluorine substituents; the conceptual and analytical tools developed in Chapter 3 can therefore inform both the design of fluorinated biophysical probes and the interpretation of fluorine-containing ligands at protein targets.

Chapter 4 determined high-resolution crystal structures of maltose-binding protein containing 4-fluorotryptophan and 4-cyanotryptophan, which represent two degrees of steric extension at the 4-position of tryptophan. These structures revealed that aliphatic sidechains can readily shift to accommodate extra steric bulk, and retain the native tryptophan rotamer, but backbones are less accommodating. Because current structure prediction models remain limited to canonical amino acids, expanding this type of structural data will refine principles for choosing non-canonical amino acid incorporation sites that preserve protein activity while acting as faithful fluorescent reporters.

Chapter 5 applied the understanding of fluorescent sidechain accommodation to nanobody engineering. Two visible-light fluorescent amino acids were substituted at position W37, sidechain repacking accommodated their incorporation and significantly enhanced the fluorescence of the introduced amino acids. Whether W37 represents a generalisable substitution position for immunoglobulin folds warrants further

investigation, but this work identifies it as a promising site. Nanobodies have applications as targeting modules in imaging, diagnostics, and therapeutics. Therefore, these bright, site-specifically incorporated amino acids provide a route to genetically encoded, specialised fluorescent probes with well-defined photoluminescence. Canonical chromophore maturation chemistry was extended across an introduced secondary amine, yielding an unprecedented chromophore with near-IR emission within a genetically encoded scaffold; this spectral region is particularly attractive for deep-tissue imaging and multiplexed readouts. Future work should stabilise its fluorescent geometry and evaluate its performance in cellular and *in vivo* imaging contexts.

Collectively, this work establishes a biophysically grounded framework for designing photoluminescent proteins that is directly translatable to practice. While contemporary protein design tools provide unprecedented structural predictive power, photoluminescent protein engineering requires principles that extend beyond what can be inferred from canonical amino acid sequences alone. This thesis establishes such principles by systematically characterising how critical elements—lanthanides, antennae, fluorescent amino acids, chromophore maturation, and conformational dynamics—determine photophysical properties in designed and natural protein scaffolds.

Appendix A: Thesis supporting information

Appendix A : Thesis supporting information

A.1 Supporting information for Chapter 2

- A.1.1 Intact protein mass spectrometry
- A.1.2 Lanthanide titration experiments
- A.1.3 Integrated lifetime measurements
- A.1.4 Spectral overlap calculations
- A.1.5 Sequence design of EF2-derived mini-proteins
- A.1.6 NMR measurements of $^{13}\text{C}/^{15}\text{N}$ -labeled RF2
- A.1.7 Photoluminescence properties of RF2 and RF2 6AW with Tb^{3+} , Eu^{3+} , Dy^{3+} and Sm^{3+}
- A.1.8 q-value determination
- A.1.9 Supplementary tables

A.2 Supporting information for Chapter 3

- A.2.1 Protein sequence and expression yield
- A.2.2 Intact protein mass spectrometry
- A.2.3 Rotamer distributions
- A.2.4 Model generation and clash score calculation

A.3 References

A.1 Supporting information for Chapter 2

A.1.1 Intact protein mass spectrometry

Intact protein analysis was performed on an Orbitrap Fusion™ Tribrid™ mass spectrometer (Thermo Fisher Scientific, USA) connected to a Thermo Fisher Scientific UltiMate 3000 HPLC system equipped with ZORBAX 300SB-C3, 3.5 μ m, 4.6 x 50 mm HPLC column (Agilent Technologies, USA). Approximately 50 pmol of sample was injected using a 500 μ L/min linear gradient of solvent A (0.1% (v/v) formic acid in water) and solvent B (0.1% (v/v) formic acid in acetonitrile), ramping solvent B from 5% solvent B at the start to 80% after 12 min. Data were collected using an electrospray ionization (ESI) source in positive ion mode. Protein intact mass was determined by deconvolution using the program Xcalibur 3.0.63 (Thermo Fisher Scientific, USA).

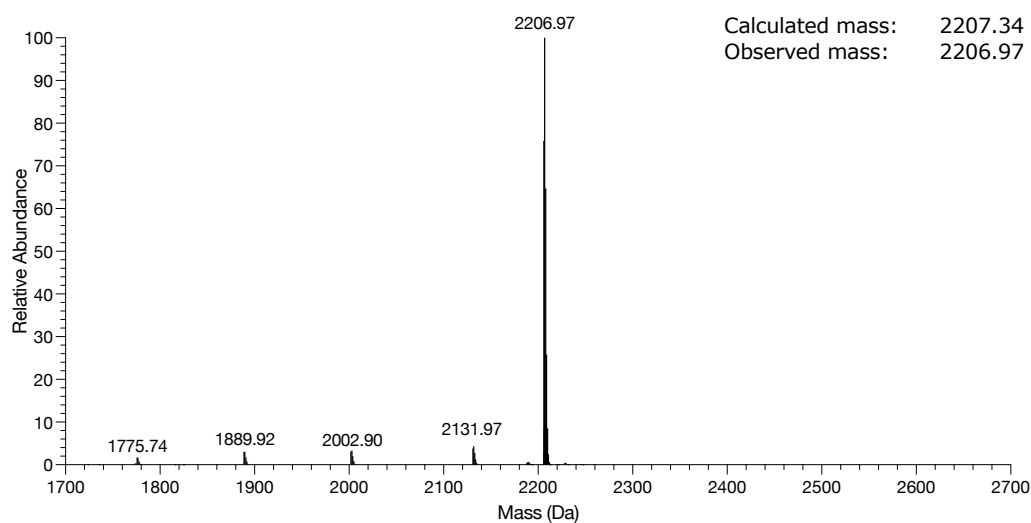


Figure A.18 - Intact protein mass spectrometry of LBT.

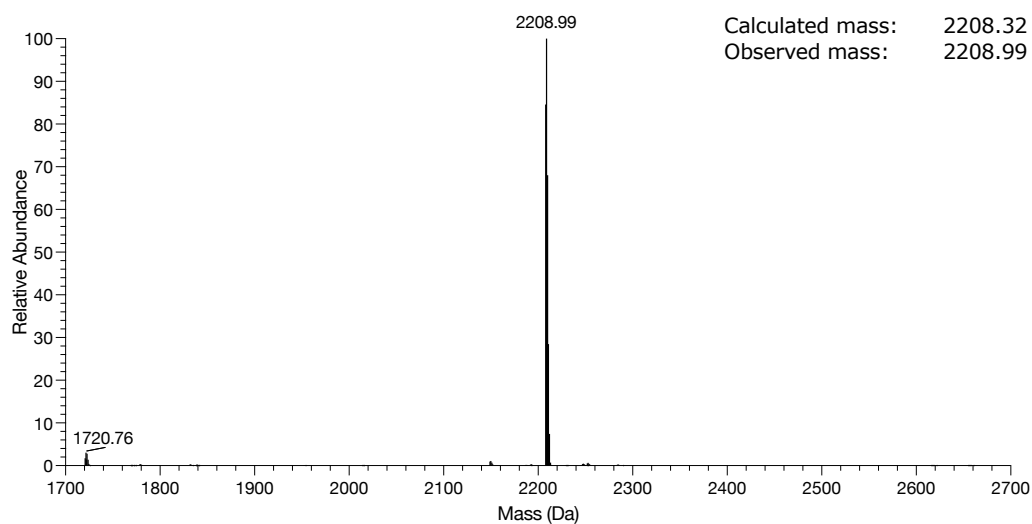


Figure A.19 - Intact protein mass spectrometry of LBT 4AW.

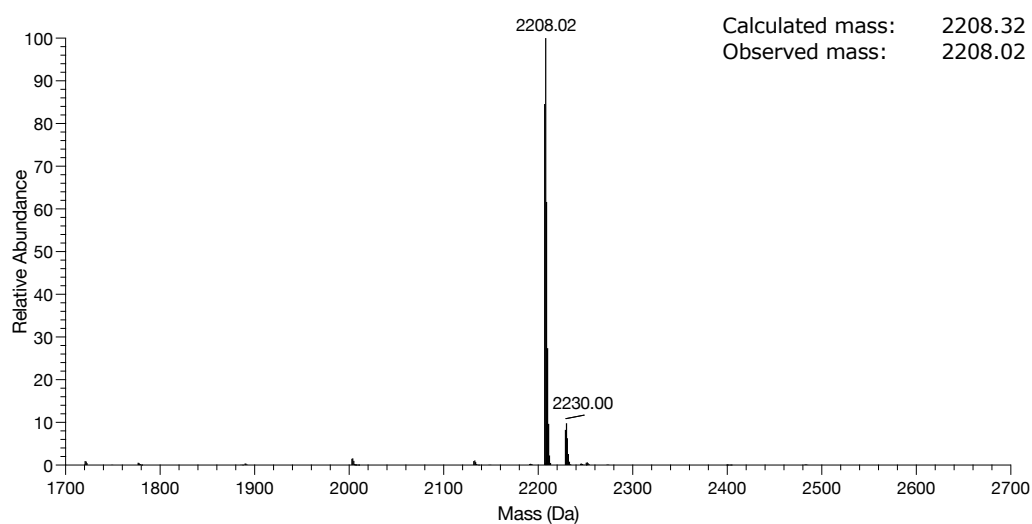


Figure A.20 - Intact protein mass spectrometry of LBT 6AW.

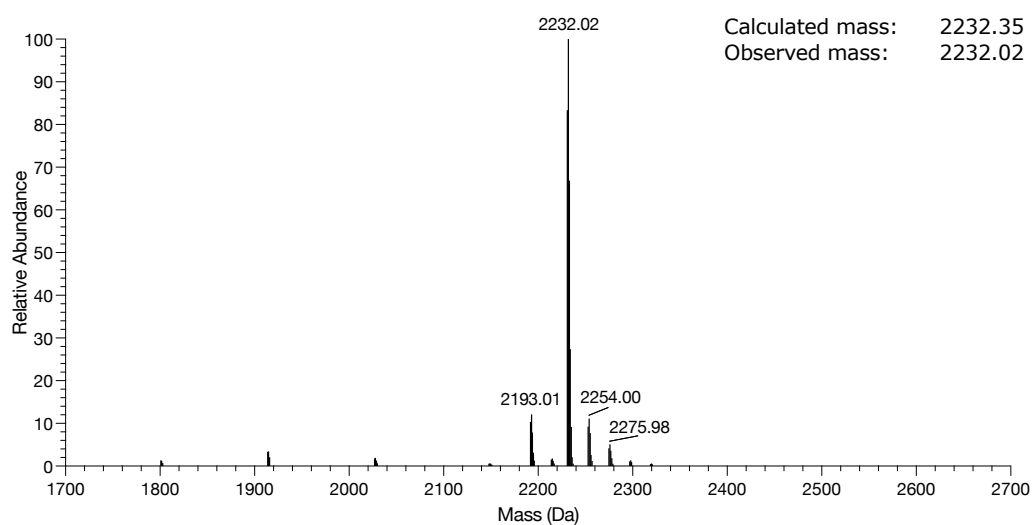


Figure A.21 - Intact protein mass spectrometry of LBT 4CNW.

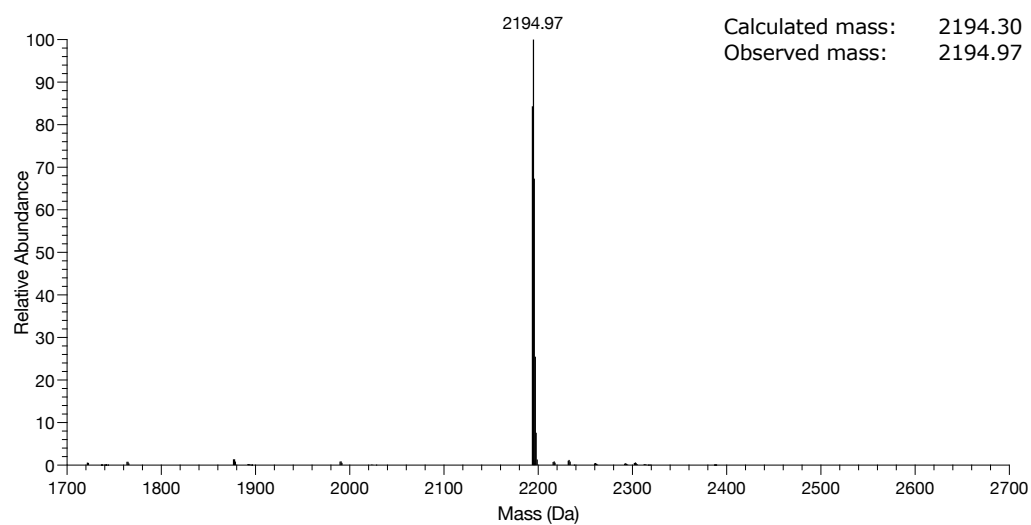


Figure A.22 - Intact protein mass spectrometry of LBT mCNP.

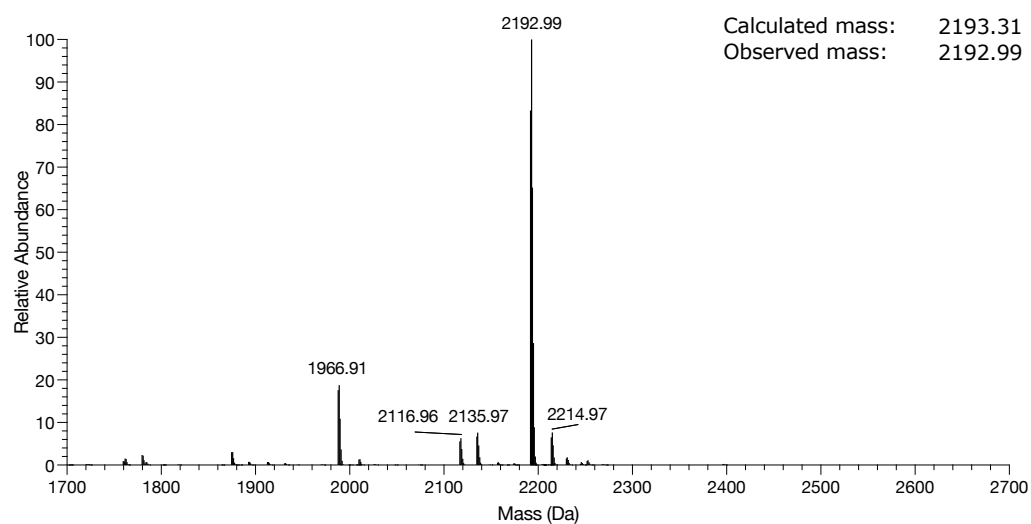


Figure A.23 - Intact protein mass spectrometry of LBT mCNF.

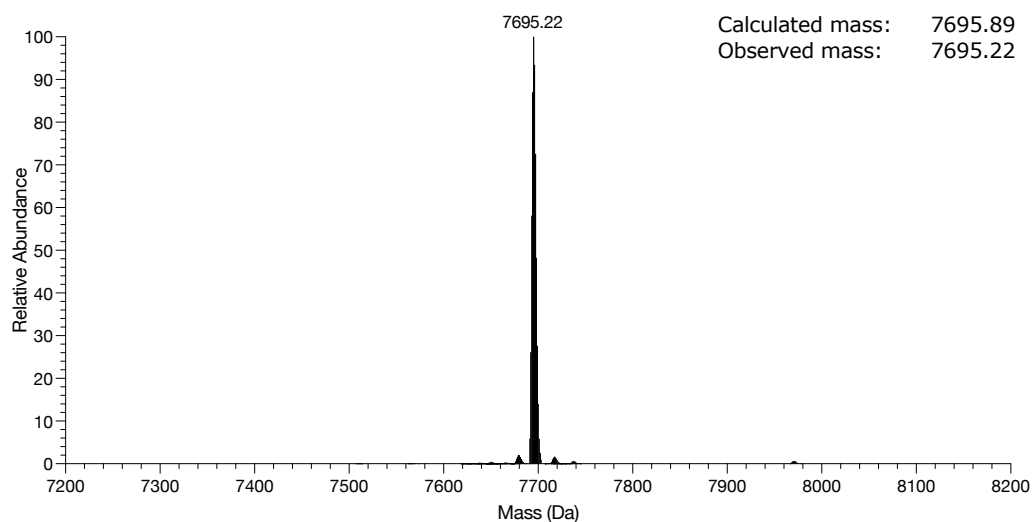


Figure A.24 - Intact protein mass spectrometry of the EF2 variant RF1.

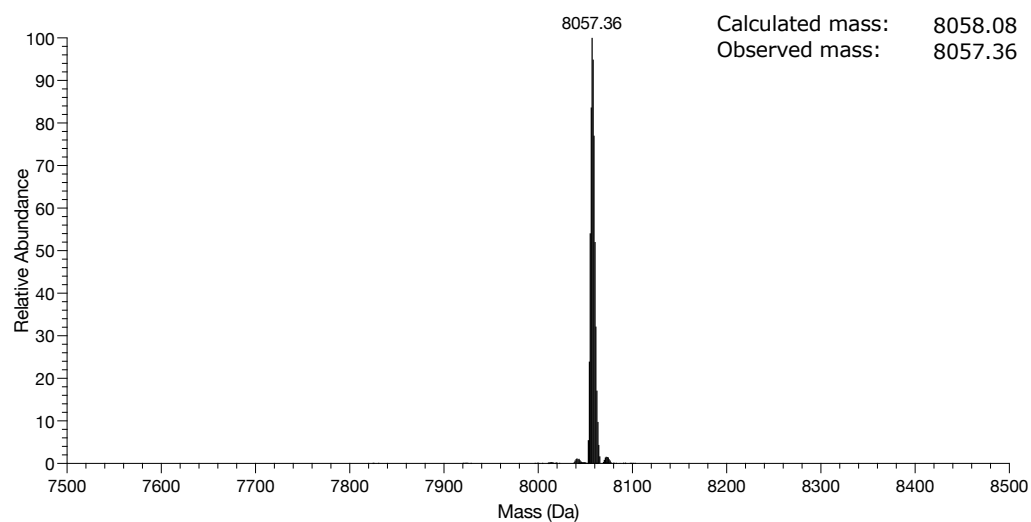


Figure A.25 - Intact protein mass spectrometry of the EF2 variant RF2.

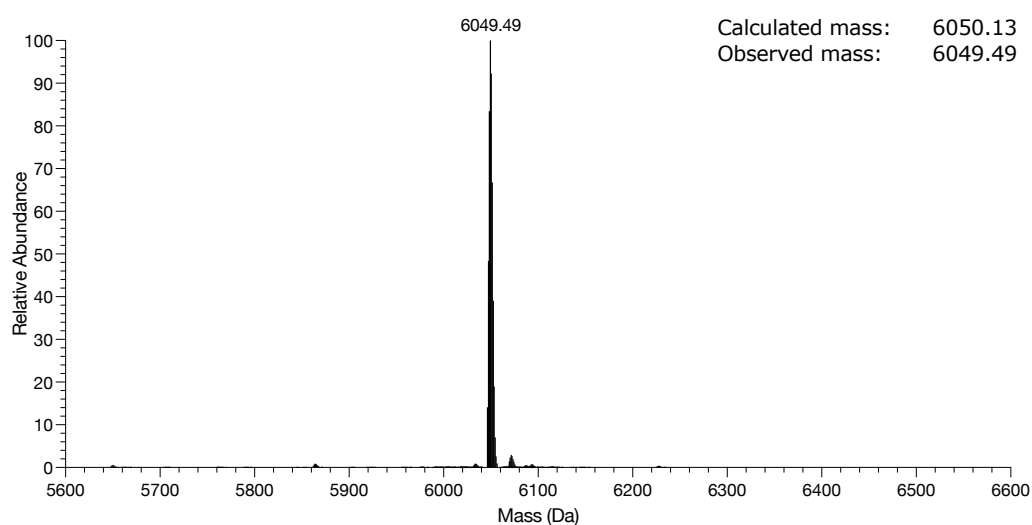


Figure A.26 - Intact protein mass spectrometry of the EF2 variant RF3.

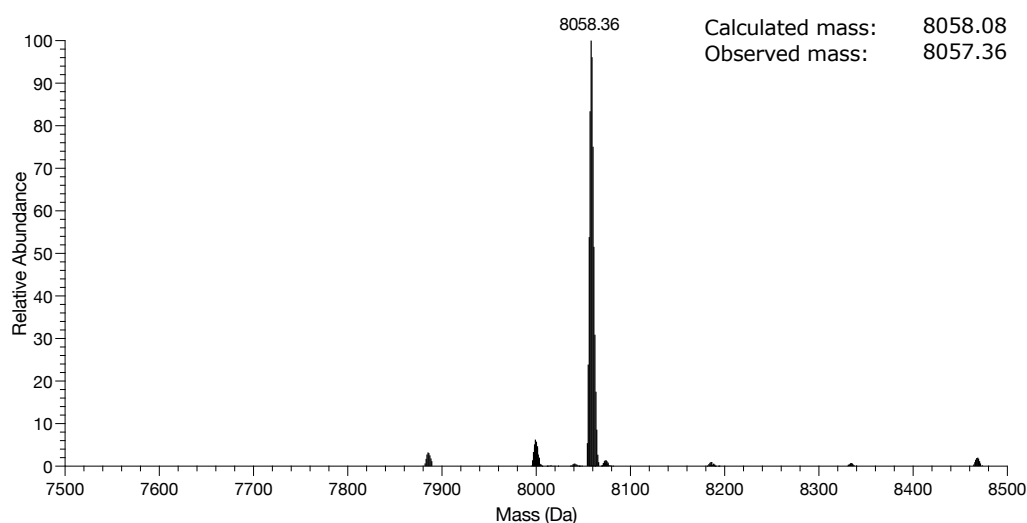


Figure A.27 - Intact protein mass spectrometry of the RF2 42 6AW.

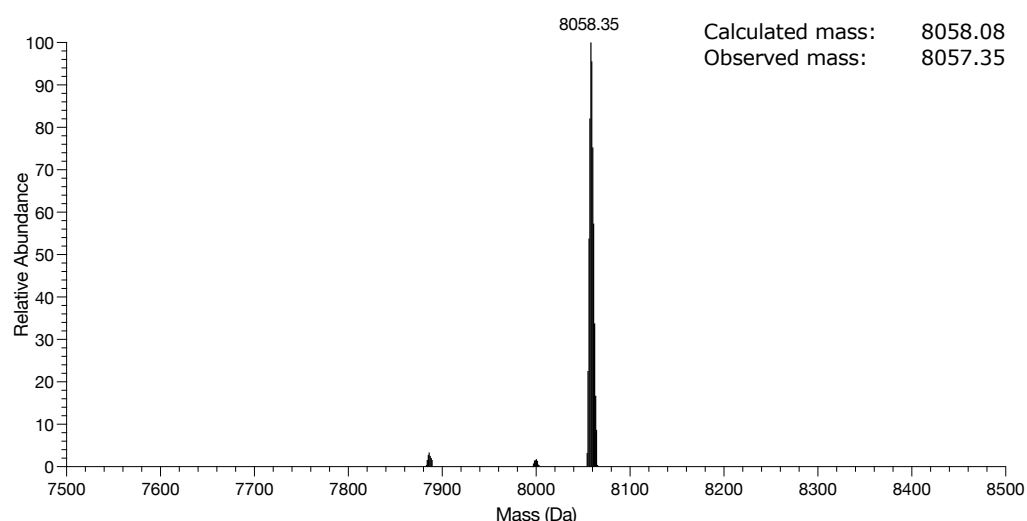


Figure A.28 - Intact protein mass spectrometry of the RF2 50 6AW.

A.1.2 Lanthanide titration experiments

Lanthanide binding affinity titrations were performed on a TECAN Infinite 200 Pro M Plex plate reader (Tecan, Switzerland) using black 384-well flat bottom plates (Corning, USA). Protein samples were prepared in Chelex-treated buffer A (20 mM MOPS pH 7.0, 50 mM KCl). Each well was a series of 1:2 serial dilutions of the corresponding lanthanide chloride stock, whose concentration was determined colourimetrically using the serial dilution described above. For each well, a background blank containing the same concentration of lanthanide ion was used for blank subtraction. Time-resolved luminescence measurements were performed for Eu^{3+} and Tb^{3+} with a 100 μs delay

followed by a 1 ms integration window. For Dy^{3+} , time-resolved settings were not used, instead collecting the emission directly under steady-state conditions. Excitation wavelengths were matched to the known absorption maxima of each fluorinated amino acid residue, with emission measured at 545 nm for terbium, 615 nm for europium and 570 nm for dysprosium. Samples were equilibrated for 15 minutes at room temperature prior to measurement. Each point reflects the average of three replicate wells from replicate titrations. The averaged dataset was fit to the Hill binding equation, with the Hill coefficient constrained to 1, using GraphPad Prism:

$$I = \frac{B_{\text{max}} \cdot [\text{Ln}^{3+}]}{K_D + [\text{Ln}^{3+}]}$$

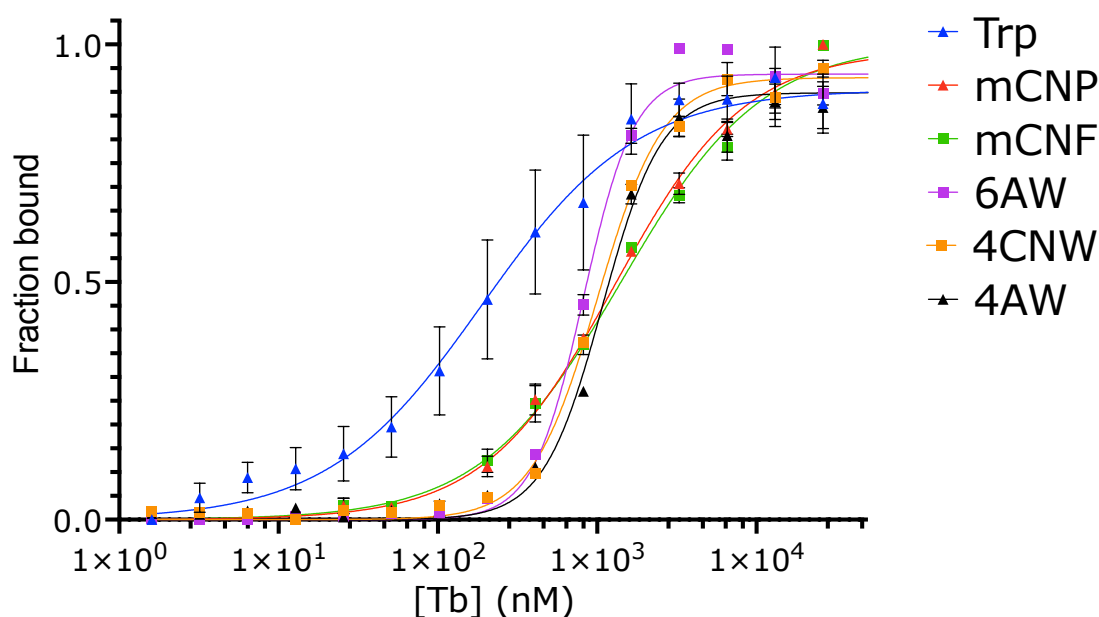


Figure A.29 – Binding affinity determination by titration of LBT variants with Tb^{3+} .

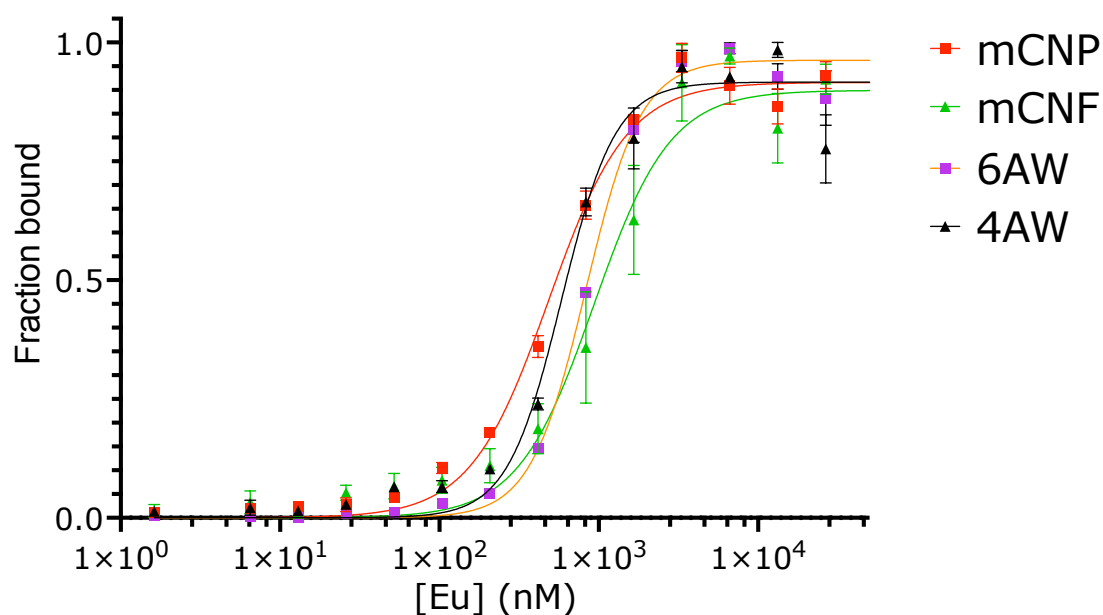


Figure A.30 – Binding affinity determination by titration of LBT variants with Eu^{3+} .

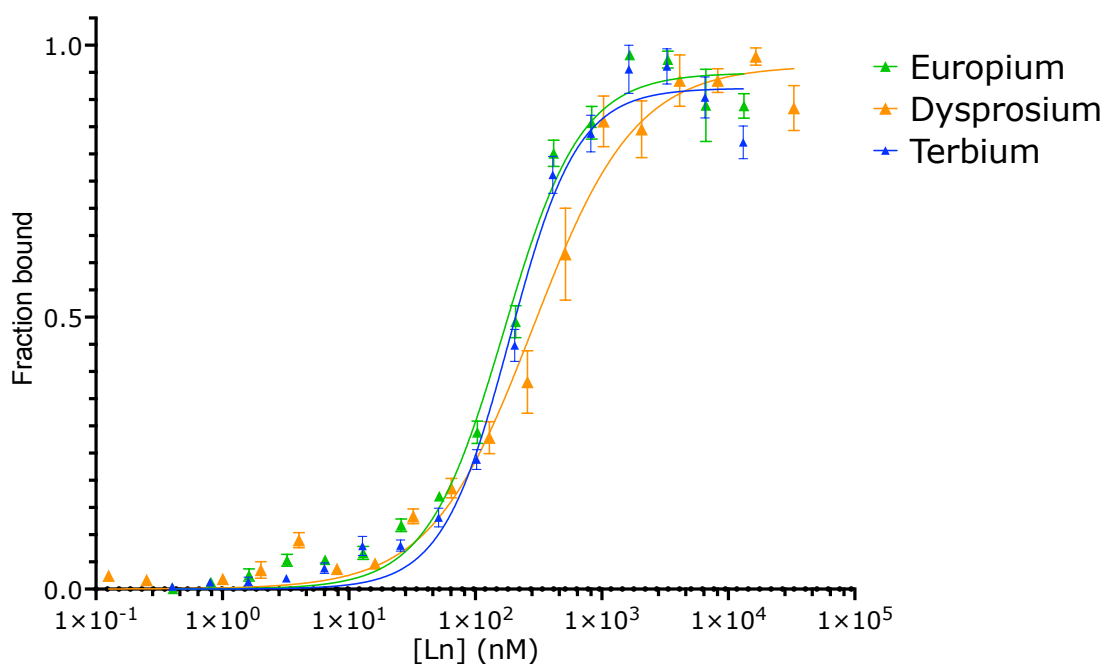


Figure A.31 - Binding affinity determination by titration of RF2 6AW complexes with Tb^{3+} , Eu^{3+} , and Dy^{3+} .

A.1.3 Integrated lifetime measurements

Luminescence lifetimes were measured on a Cary Eclipse fluorescence spectrophotometer (Agilent, USA) equipped with a lifetime measurement module. Each

measurement employed a 100 μ s delay followed by a 50 μ s gate time, with both excitation and emission slit widths set to 20 nm. A 360–1100 nm filter was used to avoid signal from harmonic doubling. A total of 25 excitation flashes was used for a 15 ms decay window, and each dataset reflects the average of 50 repeated measurement cycles. Samples had a total volume of 600 μ L in a 1 cm quartz cuvette. Samples were excited at their respective excitation maxima with emission recorded at 545 nm for Tb³⁺ complexes and 615 nm for Eu³⁺ complexes with a 20 nm slit width. Each sample was excited at the protein's excitation maximum with a 20 nm slit width and a recording of 500 nM Ln³⁺ with no protein was subtracted. The instrument gain was set to 900. To account for variability in binding affinity, protein was iteratively added to the cuvette at small volumes (2 μ L) until the initial intensity value did not change after addition of more protein. The lifetime measurement was then adjusted by subtracting the blank measurement, and then fit to the following exponential equation using GraphPad Prism, with the constraint on the fit that $K > 0$.

$$Y = Plateau + (Y_0 - Plateau) \cdot e^{-KX} \quad 2$$

The values of Y_0 , Plateau, and K were then tabulated with a standard error of the fit. Residuals were inspected to assess the quality of fit. The integrated emission intensity was then calculated as:

$$I = \frac{Y_0 - Plateau}{K} \quad 3$$

The standard error of the best-fit parameters was then propagated as follows:

$$\delta I = \sqrt{\left(\frac{1}{K}\right)^2 \cdot ((\delta Y_0)^2 + (\delta Plateau)^2) + \left(\frac{Y_0 - Plateau}{K^2}\right)^2 \cdot (\delta K)^2} \quad 4$$

Except for three very weakly luminescent samples, (RF2 6AW/Dy³⁺, RF2/Eu³⁺, LBT 4CNW/Eu³⁺) propagated error did not exceed 1% of the integrated value. Exact values are tabulated in Table A.1.

Table A.4 – Variables used in integrated lifetime calculations^a

Protein ^b	Ln ³⁺ Y ₀	Plateau	K	Integral	Error (%)
LBT	Tb 259.17 ± 0.20	0.00 ± 0.07	0.38 ± 0.00	682.88 ± 1.14	0.17
LBT	Eu 3.32 ± 0.03	0.00 ± 0.00	0.71 ± 0.00	4.65 ± 0.04	0.79
LBT mCNP	Tb 233.43 ± 0.20	-0.15 ± 0.07	0.37 ± 0.00	630.27 ± 1.17	0.19
LBT mCNP	Eu 253.05 ± 0.18	-0.01 ± 0.03	0.70 ± 0.00	361.21 ± 0.48	0.13
LBT mCNF	Tb 156.29 ± 0.15	-0.14 ± 0.05	0.37 ± 0.00	423.25 ± 0.85	0.20
LBT mCNF	Eu 35.33 ± 0.06	0.00 ± 0.01	0.70 ± 0.00	50.39 ± 0.16	0.32
LBT 4AW	Tb 468.61 ± 0.92	2.16 ± 0.18	0.64 ± 0.00	728.16 ± 2.67	0.37
LBT 4AW	Eu 61.71 ± 0.09	0.01 ± 0.02	0.72 ± 0.00	85.77 ± 0.24	0.28
LBT 6AW	Tb 402.22 ± 0.30	-0.14 ± 0.10	0.38 ± 0.00	1072.23 ± 1.71	0.16
LBT 6AW	Eu 268.10 ± 0.24	0.01 ± 0.04	0.71 ± 0.00	377.86 ± 0.62	0.16
LBT 4CNW	Tb 209.92 ± 0.31	0.16 ± 0.06	0.65 ± 0.00	323.19 ± 0.88	0.27
LBT 4CNW	Eu 3.77 ± 0.02	-0.00 ± 0.00	0.70 ± 0.01	5.38 ± 0.06	1.20
RF1	Tb 867.20 ± 0.48	0.17 ± 0.09	0.72 ± 0.00	1207.33 ± 1.24	0.10
RF2	Tb 1236.34 ± 0.51	0.02 ± 0.07	0.69 ± 0.00	1783.83 ± 1.20	0.07
RF2	Eu 0.67 ± 0.02	-0.01 ± 0.00	1.11 ± 0.04	0.61 ± 0.03	4.77
RF3	Tb 357.29 ± 0.59	0.28 ± 0.10	0.77 ± 0.00	462.69 ± 1.38	0.30
RF2 42 6AW ^b	Tb 1027.07 ± 0.54	0.07 ± 0.10	0.68 ± 0.00	1507.04 ± 1.46	0.10
RF2 42 6AW ^b	Eu 920.91 ± 0.47	0.03 ± 0.04	1.82 ± 0.00	506.04 ± 0.43	0.09
RF2 42 6AW ^b	Dy 1631.29 ± 119.71	0.02 ± 0.00	56.77 ± 0.98	28.73 ± 2.17	7.54
RF2 50 6AW ^b	Tb 1200.19 ± 0.67	0.03 ± 0.10	0.71 ± 0.00	1700.41 ± 1.60	0.09
RF2 50 6AW ^b	Eu 24.18 ± 0.07	-0.02 ± 0.01	1.88 ± 0.01	12.84 ± 0.06	0.48

^aUncertainties were derived from the deviation of the best-fit equation to the data. Final values are rounded to whole numbers in the main text. ^bThe notations used here are different from the main text to facilitate discrimination of the different 6AW positions.

Table A.5 – Extinction coefficients of the amino acids used in Chapter 2

Amino acid	λ_{ex} (nm)	Extension coefficient ^a (M ⁻¹ cm ⁻¹)
4-CN-Trp ^{a,b}	290	4708 ± 58 (at 290 nm) 2304 ± 48 (at 280 nm)
6-Aza-Trp	265	3087 ± 8 (at 265 nm) 1013 ± 4 (at 280 nm)
4-Aza-Trp	290	7335 ± 4 (at 290 nm) 6272 ± 4 (at 280 nm)
<i>m</i> -cyano-pyridylalanine	265	1386 ± 9 (at 265 nm) 255 ± 9 (at 280 nm)
<i>m</i> -cyano-phenylalanine	275	1059 ± 30 (at 275 nm) 843 ± 19 (at 280 nm)

^aExtinction coefficients at 280 nm are useful for determining protein concentrations based on their amino acid sequence. For comparison, the ϵ_{280} value of tryptophan is 5422 M⁻¹ cm⁻¹. ^bValue from Abdelkader *et al.*¹

A.1.4 Spectral overlap calculations

The spectral overlap calculations were performed using pre-processed emission and absorption spectra. Spectra were truncated to the range of 280–600 nm. The area under each spectrum was normalised to 1 using trapezoidal integration. Spectral overlap was calculated as the integral of the scalar product between the normalised Ln^{3+} absorption spectrum, and the normalised amino acid emission spectrum, resulting in a dimensionless overlap value between 0 and 1. These calculations were performed in MATLAB.

The absorption spectra of Eu^{3+} and Tb^{3+} were recorded on a Cary 50 UV-vis spectrophotometer using a 400 μL quartz cuvette. Spectra were acquired in dual beam mode and baseline corrected by blanking with water. Both lanthanide ions were provided as 1 M chloride salt solutions.

Fluorescence emission spectra of the amino acids were recorded in a 400 μL quartz cuvette using a Cary Eclipse fluorescence spectrophotometer. The excitation slit width was set to 10 nm and the emission slit width to 1.5 nm. For all amino acids except mCNP and mCNF, a 295–1100 nm long-pass emission filter was used to suppress scattered excitation light and harmonic doubling. For mCNP and mCNF, the emission filter was left open.

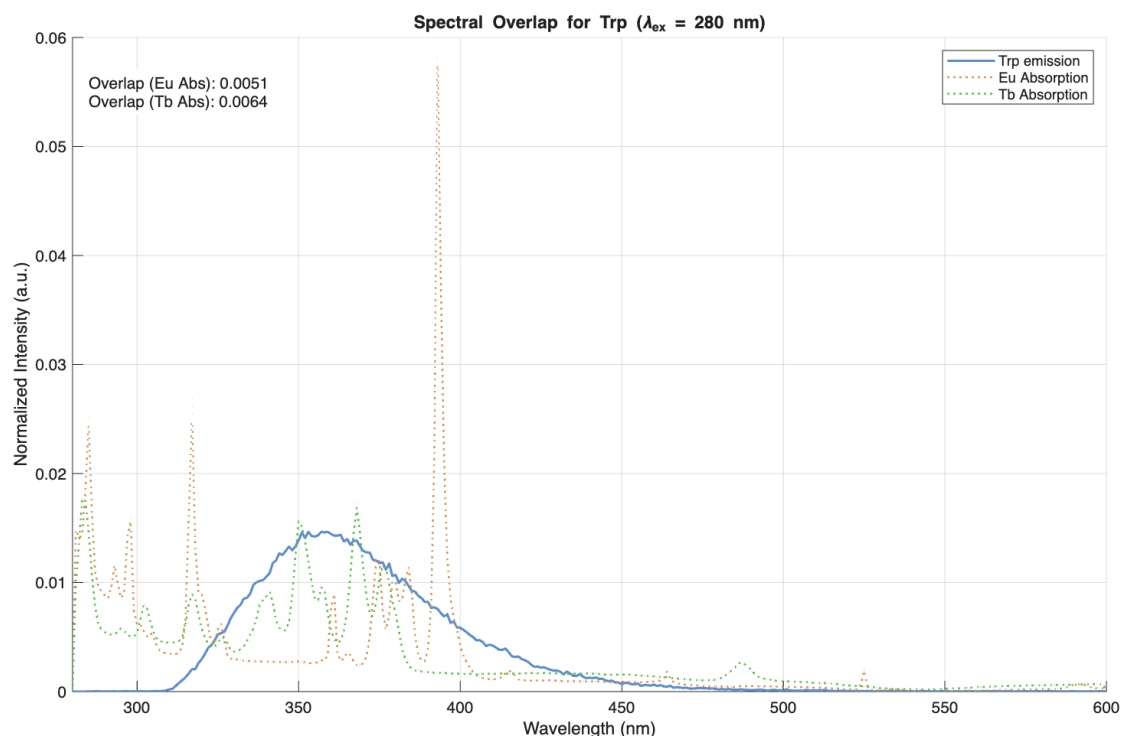


Figure A.32 - Spectral overlap of the Trp emission (free amino acid in 20 mM MOPS pH 7.0, 50 mM KCl) with the absorption of 1M Eu^{3+} or Tb^{3+} (LnCl_3 in 20 mM MOPS pH 7.0, 50 mM KCl). Areas of each spectrum are normalized to 1 before plotting.

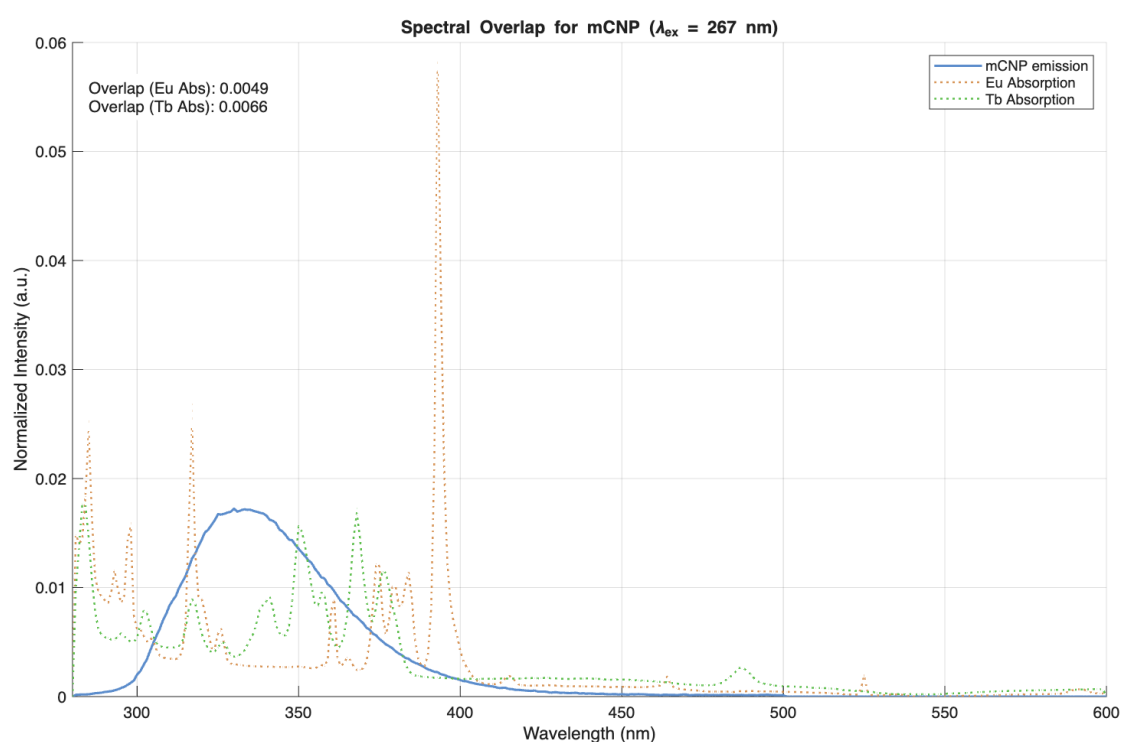


Figure A.33 - Spectral overlap of the mCNP emission (free amino acid in 20 mM MOPS pH 7.0, 50 mM KCl) with the absorption of 1M Eu^{3+} or Tb^{3+} (LnCl_3 in 20 mM MOPS pH 7.0, 50 mM KCl). Areas of each spectrum are normalized to 1 before plotting.

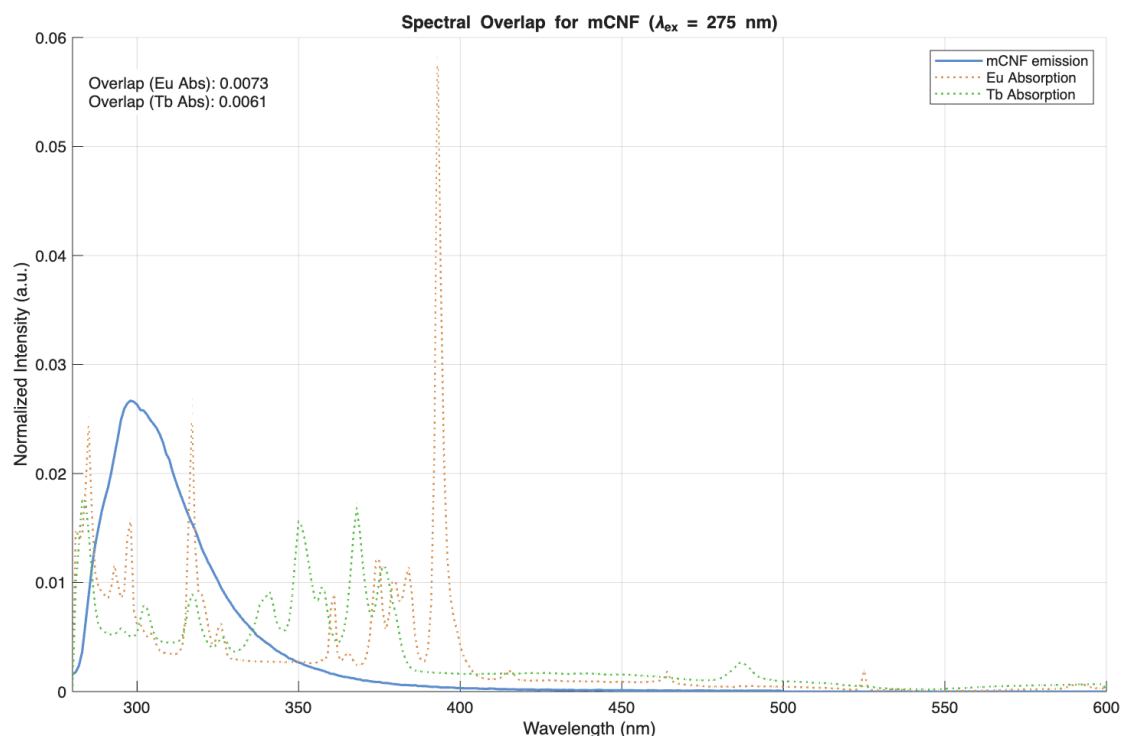


Figure A.34 - Spectral overlap of the mCNF emission (free amino acid in 20 mM MOPS pH 7.0, 50 mM KCl) with the absorption of 1M Eu^{3+} or Tb^{3+} (LnCl_3 in 20 mM MOPS pH 7.0, 50 mM KCl). Areas of each spectrum are normalized to 1 before plotting.

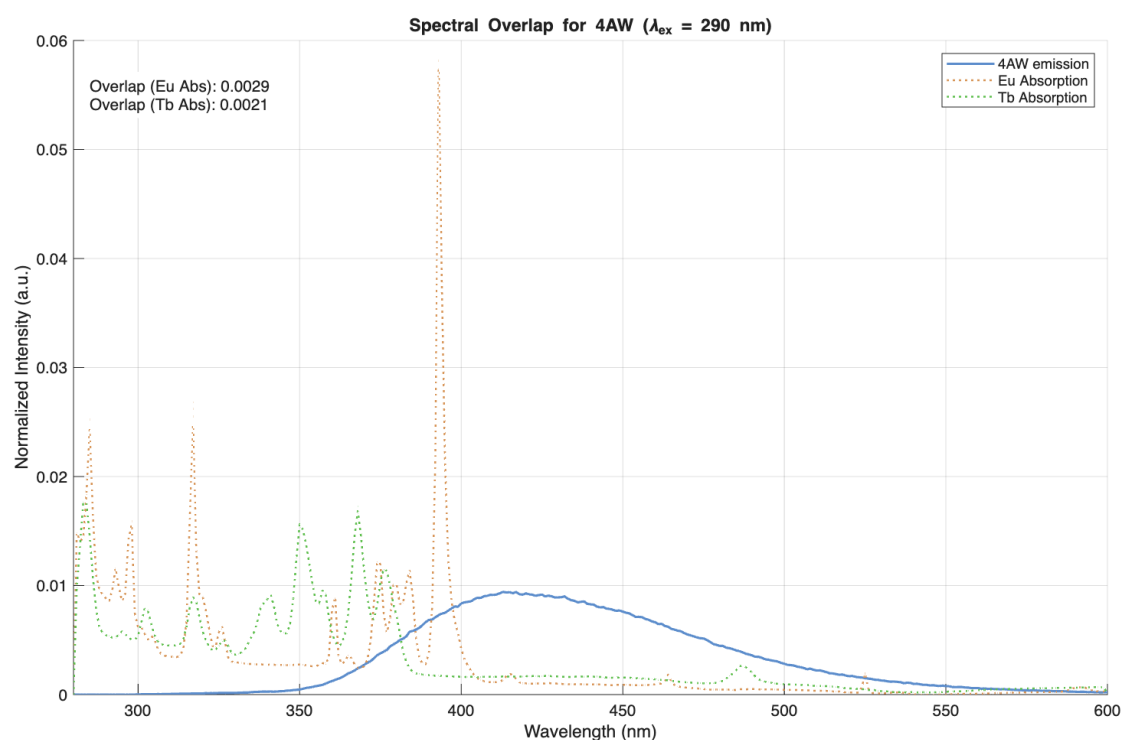


Figure A.35 - Spectral overlap of the 4AW emission (free amino acid in 20 mM MOPS pH 7.0, 50 mM KCl) with the absorption of 1M Eu^{3+} or Tb^{3+} (LnCl_3 in 20 mM MOPS pH 7.0, 50 mM KCl). Areas of each spectrum are normalized to 1 before plotting.

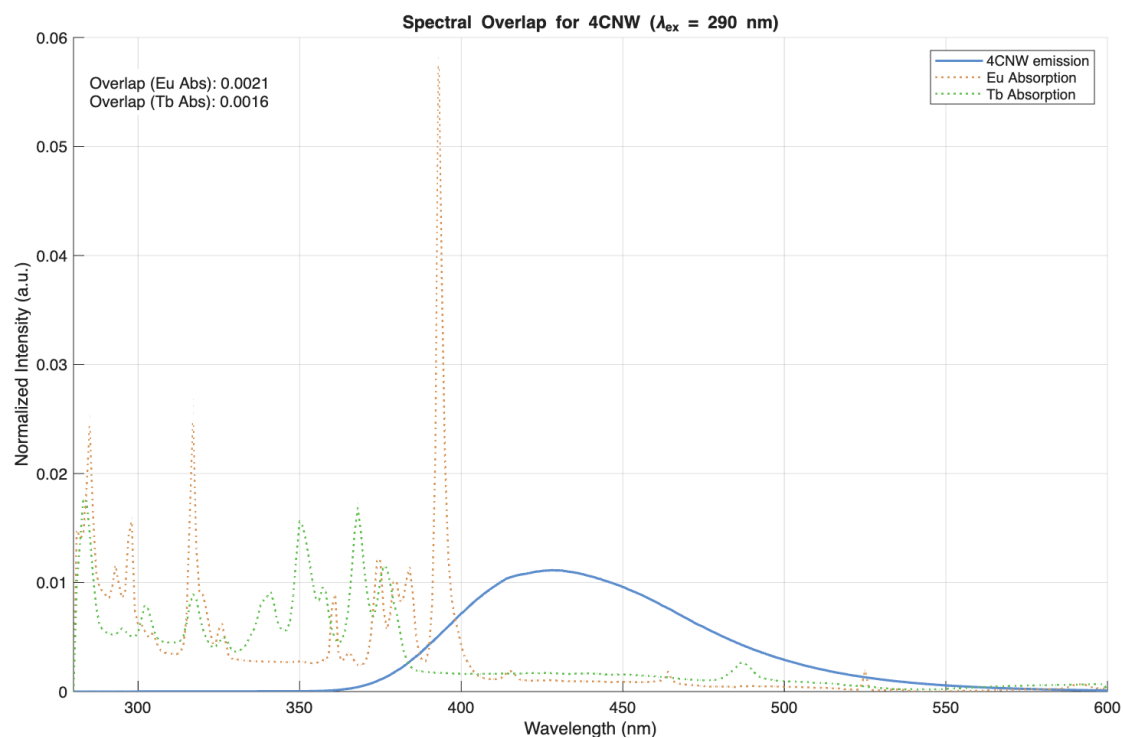


Figure A.36 - Spectral overlap of the 4CNW emission (free amino acid in 20 mM MOPS pH 7.0, 50 mM KCl) with the absorption of 1M Eu^{3+} or Tb^{3+} (LnCl_3 in 20 mM MOPS pH 7.0, 50 mM KCl). Areas of each spectrum are normalized to 1 before plotting

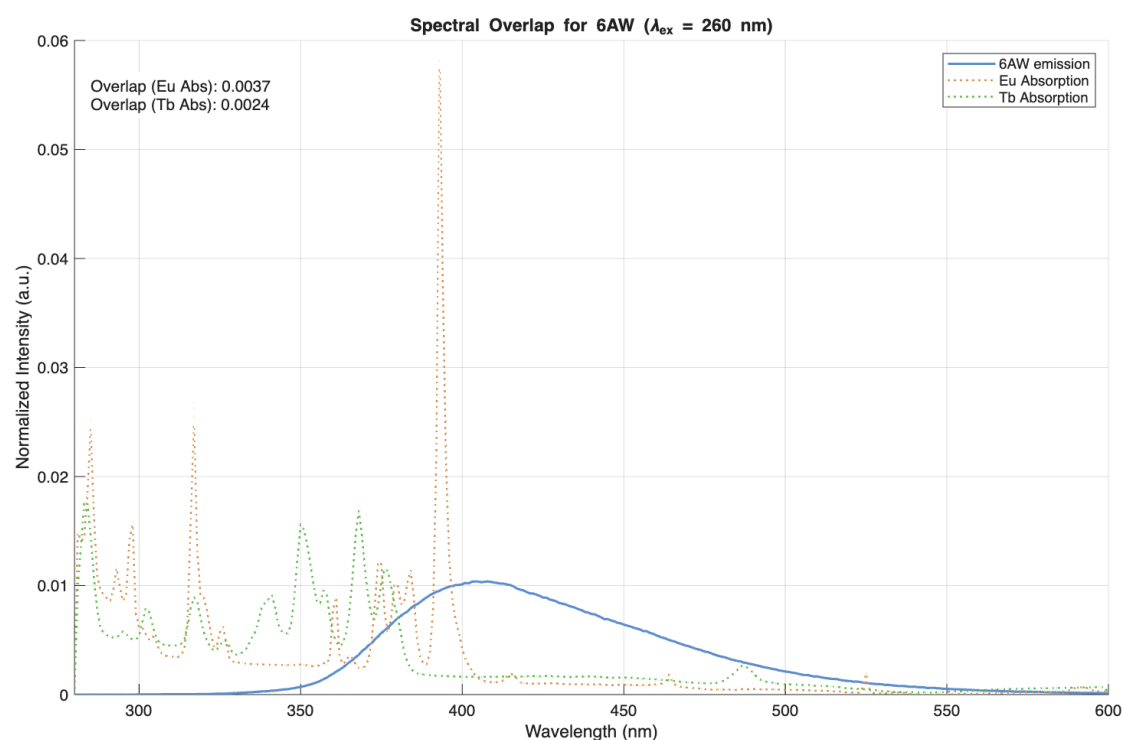


Figure A.37 - Spectral overlap of the 6AW emission (free amino acid in 20 mM MOPS pH 7.0, 50 mM KCl) with the absorption of 1M Eu^{3+} or Tb^{3+} (LnCl_3 in 20 mM MOPS pH 7.0, 50 mM KCl). Areas of each spectrum are normalized to 1 before plotting.

A.1.5 Sequence design of EF2-derived mini-proteins

Using PyMOL,² the coordinates of EF2 (residues 59–70) were extracted from the first frame of the lanmodulin NMR ensemble (PDB 6MI5).³ Thr65 was mutated to tryptophan and the isolated segment was reindexed to begin at residue 1. The structure was energy minimised using the PyRosetta MinMover and the *ref2015* score function⁴⁻⁵ The minimised coordinates were uploaded to the RFdiffusion Google Colab (v1.1.1).⁶ Two sets of RFdiffusion backbones and sequences were generated using the built-in ProteinMPNN module. RF1 and RF2 were taken from set A with contigs set to “35:A1-15:24,” and RF3 was taken from set B with contigs set to “20:A1-15:20.” For both sets, iterations=50, num_designs=32, and symmetry=“none” ProteinMPNN parameters were num_sequences=8, num_recycles=3, and mpnn_sampling_temp=0.1, with cysteine excluded from suggestions. Set A and B were downloaded and combined for processing. The final set included 64 sets of backbone coordinates, each with 8 associated sequences.

For each backbone, PyRosetta was used to mutate the coordinates to match each of the eight ProteinMPNN sequences. A Tb³⁺ ion was placed 2.5 Å from the tryptophan backbone carbonyl along the C-O bond vector. Each model was minimized with *ref2015* and the final score was recorded as “rosettascore.” The C_α alignment RMSD between lanmodulin EF2 residues 59–70 and the corresponding segment in each design was computed and recorded as “bb_RMSD.” Distances between side-chain oxygen atoms and the antenna tryptophan carbonyl were measured, the count of O-Tb³⁺ distances below 3.0 Å was recorded as “nO,” and an RMSD to a 2.5 Å target for these O-Tb³⁺ distances was calculated and recorded as “O_RMSD.” The 2.5 Å target approximates an Ln³⁺-O bond length in protein crystal structures containing lanthanides. The combination of rosettascore, bb_RMSD, nO, and O_RMSD was used to select one sequence per backbone. The resulting models were inspected in PyMOL and three designs were selected to represent different tertiary structures. For expression, the first residue was changed to glycine to introduce an N-terminal TEV site.

A.1.6 NMR measurements of $^{13}\text{C}/^{15}\text{N}$ -labeled RF2

All NMR measurements were carried out at 30 °C using a Bruker 800 MHz NMR spectrometer. The backbone resonance assignments were obtained from the analysis of 3D HNCO, HN(CA)CO, HNCA, HNCACB, together with NOESY- ^{15}N -HSQC and TOCSY- ^{15}N -HSQC. Chemical shift assignments were deposited in the BMRB under the accession number 53251 and the title EF2-RF2. Spectra are available on Zenodo (DOI: 10.5281/zenodo.1578834).

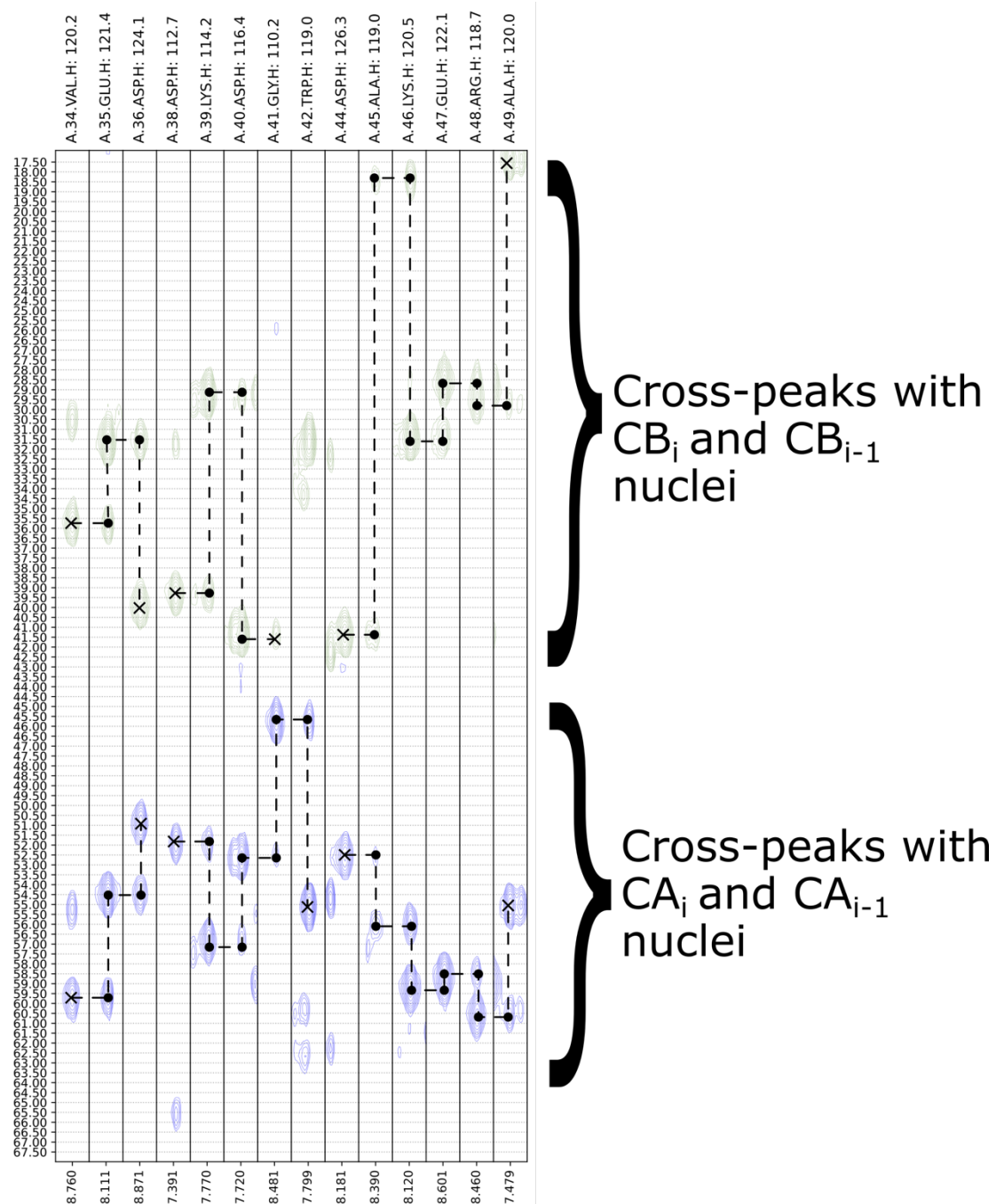


Figure A.38 - HNCACB data of $^{15}\text{N}/^{13}\text{C}$ -labeled RF2. The strip plots are of the 3D HNCACB spectrum of the lanthanide-binding EF-Hand of RF2 including 4 flanking residues (residues 34–49): VEDPDKDGWLDAKERA, where underlined residues do not have assigned NH peaks. The strips illustrate the sequential CB_i - CB_{i-1} peaks along the sequence.

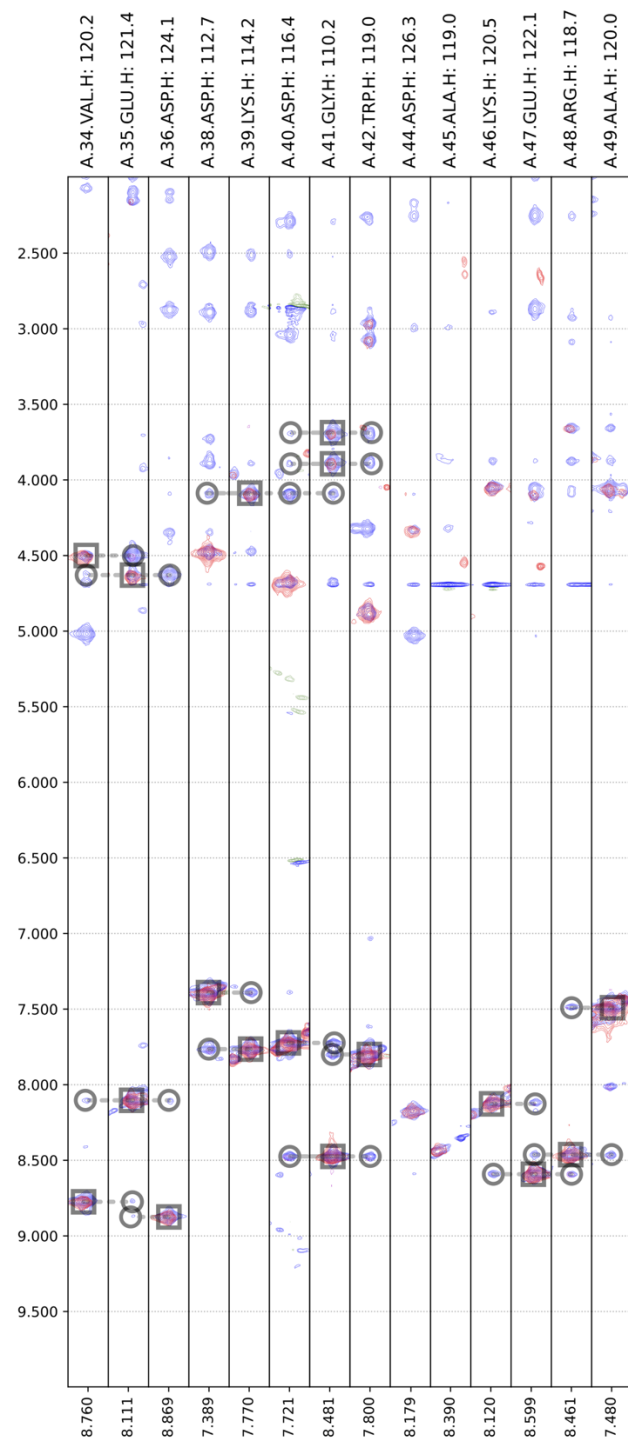


Figure A.39 Overlay of ^{13}C -decoupled 3D NOESY- ^{15}N -HSQC (blue) and TOCSY- ^{15}N -HSQC (red) spectra of $^{15}\text{N}/^{13}\text{C}$ -labeled RF2. The strips belong to the lanthanide-binding EF-Hand of RF2 and 4 flanking residues (res 34–49) VEDPKDGWLDAKERA, where underlined residues do not have assigned NH peaks. Intra- and inter-residual cross-peaks are identified by squares and circles, respectively.

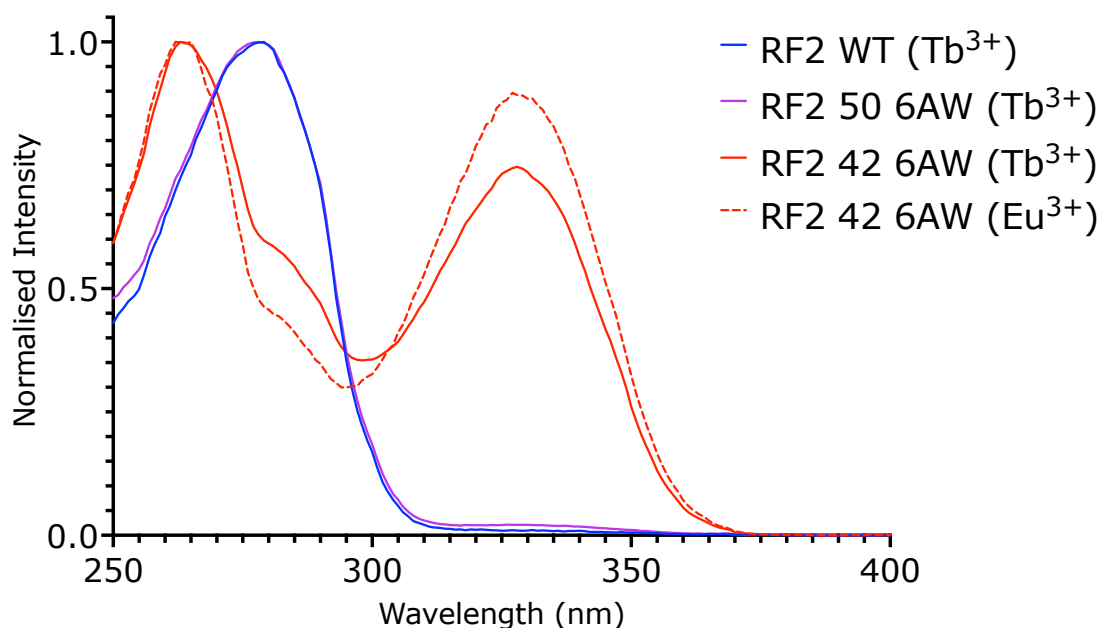
A.1.7 Photoluminescence properties of RF2 and RF2 6AW with Tb^{3+} , Eu^{3+} , Dy^{3+} and Sm^{3+} 

Figure A.40 - Excitation scans of the RF2 complexes with Tb^{3+} and Eu^{3+} . Emission at 545 nm or 615 nm was integrated for 5000 μs after a delay of 100 μs , using an excitation slit width of 5 nm and an emission slit width of 10 nm. Protein and lanthanide chloride were at 1 μM and 2 μM , respectively. The samples were prepared in buffer A. Each spectrum is normalized to the maximum intensity.

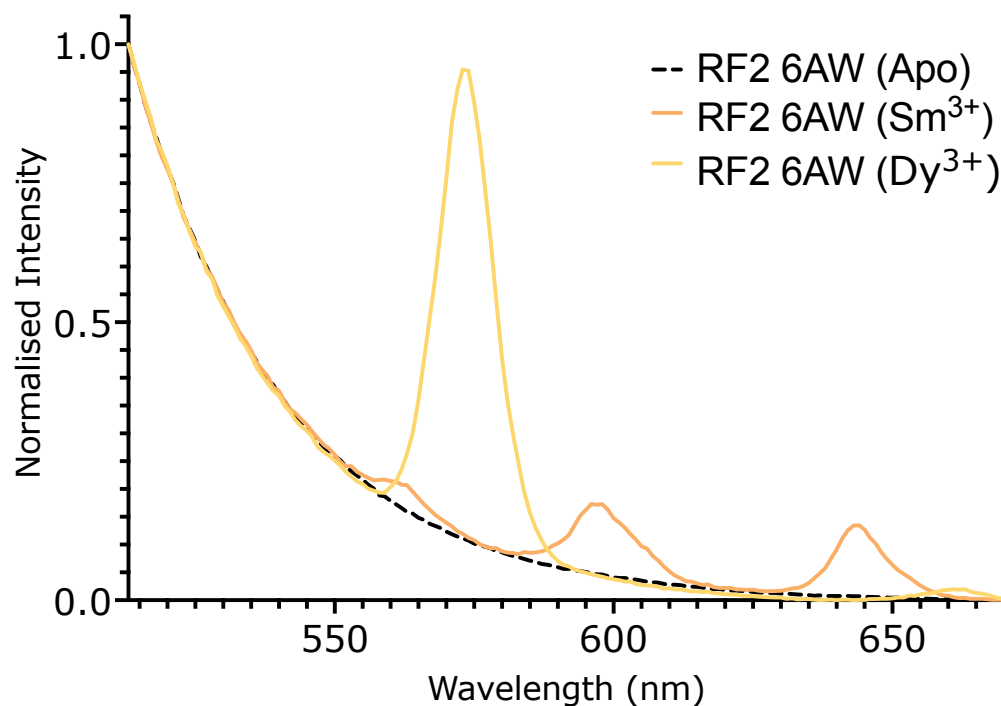


Figure A.41 - Fluorescence emission spectra of RF2 6AW with Sm^{3+} and Dy^{3+} . Each spectrum is normalised to the intensity at 520 nm.

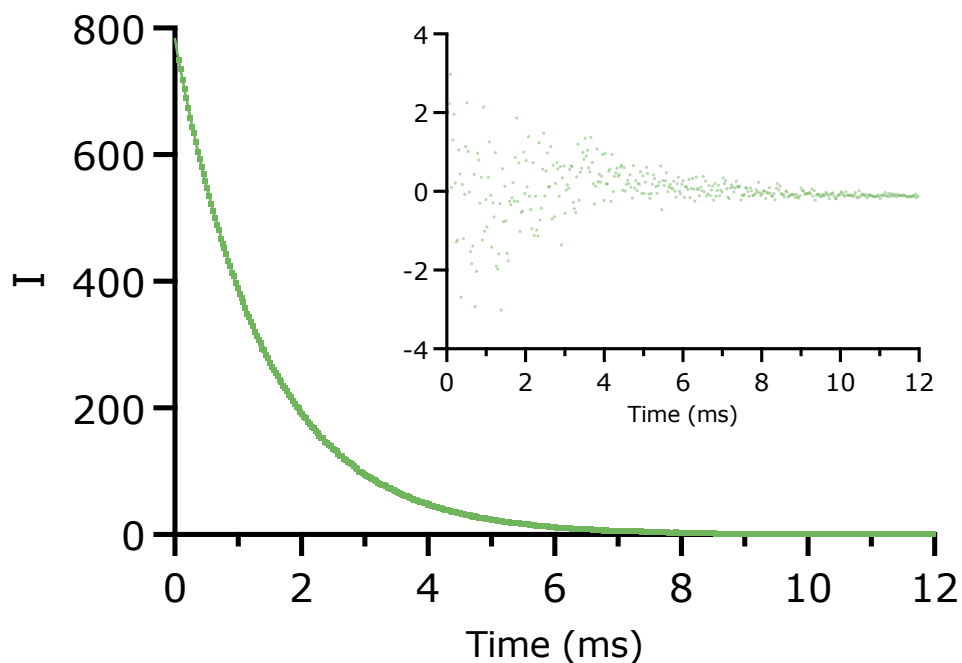


Figure A.42 - Luminescence lifetime measurement of RF2 with Tb^{3+} . The excitation wavelength was 280 nm and emission was recorded at 545 nm. Inset: residuals of the experimental values from the curve fit. The best-fit value for τ_{Tb} is 1.4 ms.

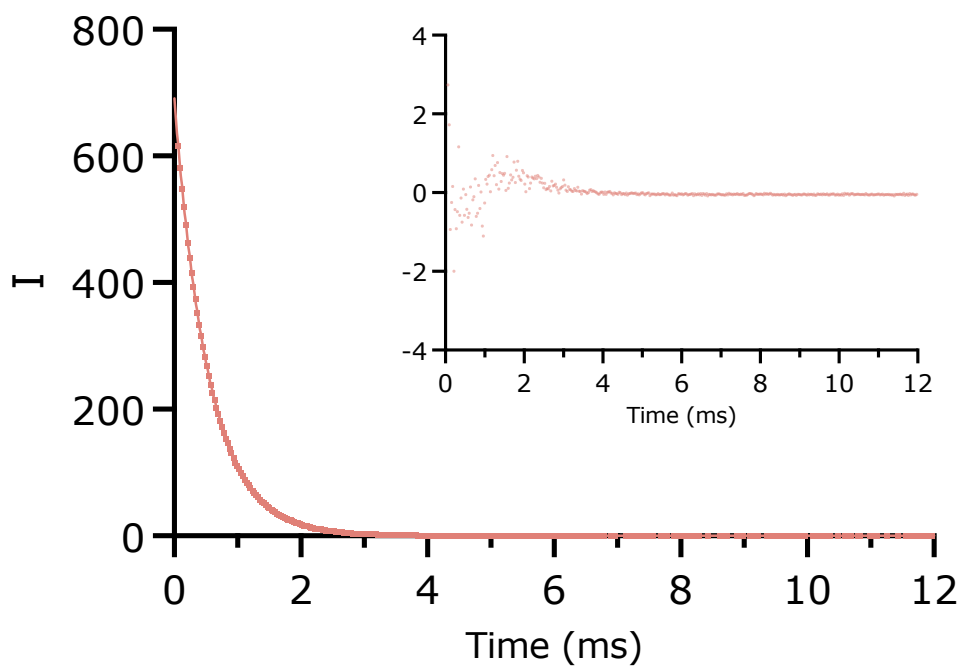


Figure A.43 - Luminescence lifetime measurement of RF2 6AW with Eu^{3+} . The excitation wavelength was 330 nm and emission was recorded at 615 nm. Inset: residuals of the experimental values from the curve fit. The best-fit value for τ_{Eu} is 0.54 ms.

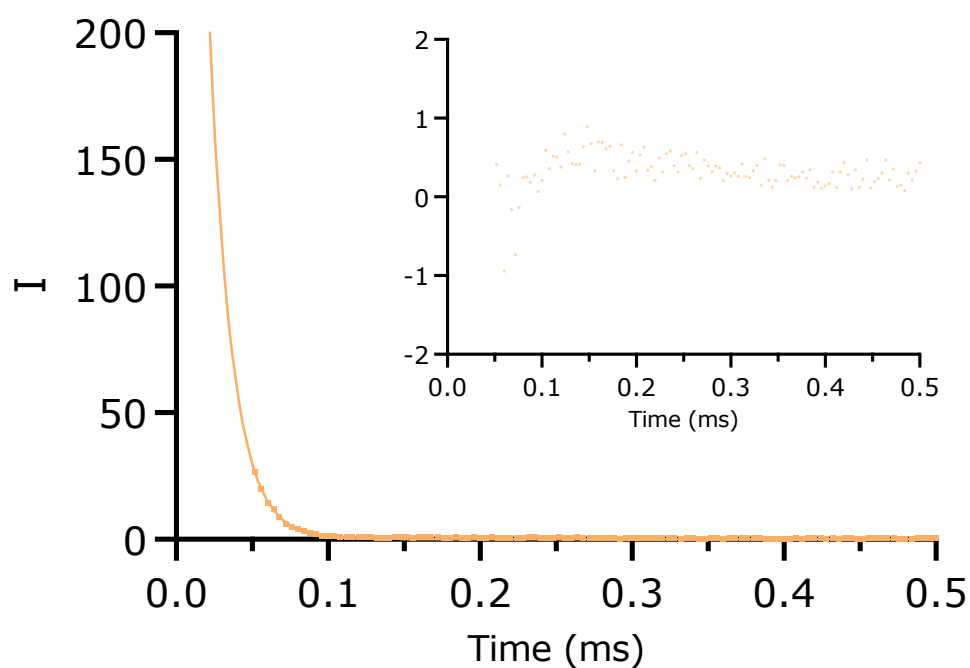


Figure A.44 - Luminescence lifetime measurement of RF2 6AW with Sm^{3+} . The excitation wavelength was 265 nm and emission was recorded at 650 nm. The instrument gain was set to 1000. Inset: residuals of the experimental values from the curve fit. The best-fit value for τ_{Sm} is 12 μs .

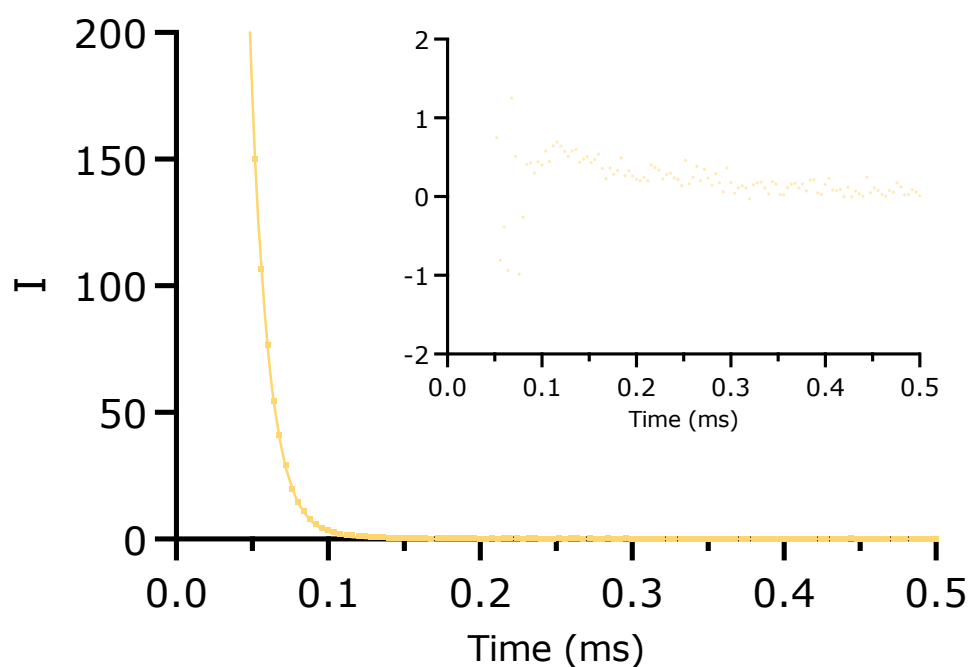


Figure A.45 - Luminescence lifetime measurement of RF2 6AW with Dy^{3+} . The excitation wavelength was 265 nm and emission was recorded at 570 nm. Inset: residuals of the experimental values from the curve fit. The best-fit value for τ_{Dy} is 14 μs .

A.1.8 q-value determination

The q-value was determined using a version of the Horrock's equation, originally:

$$q = A((k_{H_2O} - k_{D_2O}) - B) - C \quad 5$$

where A is a proportionality constant for the quenching of the lanthanide ion by O-H oscillators, k is the rate constant for the radiative decay of the lanthanide emission, parameter B is the corrective factor, accounting for other deactivating vibrations surrounding the ion, and parameter C is the outer-sphere contribution of closely diffusing solvent molecules.⁷

As the outer-sphere contributions are expected to be equivalent between the complex in H₂O and D₂O, the difference between k_{H_2O} and k_{D_2O} is attributed primarily to the inner-sphere solvent. C was therefore omitted for all ions giving the equation:

$$q = A((k_{H_2O} - k_{D_2O}) - B) \quad 6$$

For samarium, this equation is:⁸

$$q_{Sm} = 0.026(1/\tau_{H_2O} - 1/\tau_{D_2O}) \quad 7$$

For europium, this equation is:⁹

$$q_{Eu} = 1.2((1/\tau_{H_2O} - 1/\tau_{D_2O}) - 0.25) \quad 8$$

For terbium, this equation is:⁹

$$q_{Tb} = 5((1/\tau_{H_2O} - 1/\tau_{D_2O}) - 0.06) \quad 9$$

For dysprosium, this equation is:⁸

$$q_{Dy} = 0.024(1/\tau_{H_2O} - 1/\tau_{D_2O}) \quad 10$$

The lifetime in D₂O was determined by plotting the lifetime of the Ln³⁺-complex at mole fractions of D₂O (0.2, 0.4, 0.6 and 0.8) followed by a linear fit to this data. The τ_{D_2O} value was then determined by extrapolating to a D₂O mole fraction of 1.0. These parameters are empirically determined based on the assumption that O-D oscillators effectively do not contribute to non-radiative deactivation, and that all other deactivation paths are the same between D₂O and H₂O. The estimated uncertainties with using this technique are listed with the calculations of the q-values below.

A.1.9 Supplementary tables

Table A.6 – Amino acid sequences of the proteins used in this study

Protein	Amino acid sequence ^a
NT-LBT-WT	MASMTGHHHHHSHTTPWTNPGLAENFMNSFMQGLSSMPGFTASQLDK MSTIAQSMVQSIQSLAAQGRTPNDLQALNMAFASSMAEIAASEEGGGSLS TKTSSIASAMSNAFLQTTGVVNQPFINEITQLVSMFAQAGMNDVSAGNSEN LYFQGFIDTNNDGWIEGDELLLEG
NT-LBT-Amb	MASMTGHHHHHSHTTPWTNPGLAENFMNSFMQGLSSMPGFTASQLDK MSTIAQSMVQSIQSLAAQGRTPNDLQALNMAFASSMAEIAASEEGGGSLS TKTSSIASAMSNAFLQTTGVVNQPFINEITQLVSMFAQAGMNDVSAGNSEN LYFQGFIDTNNDGZIEGDELLLEG
EF2 RF1-WT	MASMTGHHHHHHGGSENLYFQGKKVITISEDTPESFEVKMGKGLLKVTVP GKTFIVEDPDKDGLDAKEIKALLAAKAATGAKTIEIEDIPK
EF2 RF2-WT	MASMTGHHHHHHGGSENLYFQGRRAYLLRVDPDTLVEISPEEAERLAKTR PVLEVEDPDKDGLDAKERAWILEHLRETHPDASGAVIVVVD
EF2 RF3-WT	MASMTGHHHHHHGGSENLYFQGKTITTYKFATKEDEKKAKELDPDKDG WLDKELKKLKKAGLKITKVTKTIK
EF2 RF2-42Z	MASMTGHHHHHHGGSENLYFQGRRAYLLRVDPDTLVEISPEEAERLAKTR PVLEVEDPDKDGZLDAKERAWILEHLRETHPDASGAVIVVVD
EF2 RF2-50Z	MASMTGHHHHHHGGSENLYFQGRRAYLLRVDPDTLVEISPEEAERLAKTR PVLEVEDPDKDGLDAKERAZILEHLRETHPDASGAVIVVVD

^a Z indicates the positions of the amber sites.

Table A.7 – Yields of different proteins co-expressed with G1(ncAA)RS in LB media with 1 mM ncAA^a

Amber site	G1(ncAA)RS	Yield (mg per 1 L cell culture)
NT-LBT	-	31 ^b
NT-LBT 4AW	4AWRS ¹⁰	16 ^c
NT-LBT 6AW	6AWRS ¹⁰	26 ^c
NT-LBT 4CNW	4CNWRS ¹	20 ^c
NT-LBT mCNP	mCNPRS ¹¹	10 ^c
NT-LBT mCNF	mCNPRS ¹¹	15 ^c
EF2 RF1	-	50 ^b
EF2 RF2	-	>400 ^b
EF2 RF3	-	10 ^b
EF2 RF2 42 6AW	6AWRS	51 ^d
EF2 RF2 50 6AW	6AWRS	19 ^e

^a All proteins were expressed and purified as described in the methods section. Yields were calculated by measuring the protein concentration at a known volume and adjusting for expression volume. ^b From 1 L expression volume. ^c From 100 mL expression volume. ^d From 50 mL expression volume. ^e From 30 mL expression volume.

Table A.8 – EF2 RF variants with Tb³⁺

Protein	Intensity Tb	Apparent K_D (Tb ³⁺) (nM)	Lifetime (τ_{Tb})
EF2 RF1	1207	320 ± 50	1.4
EF2 RF2	1734	20 ± 3	1.4
EF2 RF3	462	3300 ± 130	1.3

Table A.9 – Determination of q-values from D₂O titrations^a

Complex	τ_{H_2O}	τ_{D_2O}	q	Uncertainty on q
RF2 WT Tb	1.42 ms	2.34 ms	1.07	±0.3
RF2 42 6AW Eu	0.54 ms	1.42 ms	1.06	±0.3
RF2 42 6AW Sm	12.12 μ s	48.64 μ s	1.30	±0.5
RF2 42 6AW Dy	14.18 μ s	35.63 μ s	1.31	±0.5

^a The q values were determined using the modified Horrocks equation as described above. Uncertainties are from errors in the empirical parameters used to determine the q value.

A.2 Supporting information for Chapter 3

A.2.1 Protein sequence and expression yield

Table A.10 – Amino acid sequence of the protein used in this study

Protein	Amino acid sequence ^a
RF2-W42Z	MASMTGHHHHHHGGSE NYFQ RRAYLLRVDPDTLVEISPEEAERLAKT RPVLEVEDPDKDG Z LDAKERAWILEHLRETHPDASGAVIVVVD

Bold text indicates the TEV protease cleavage site, where the protein is cleaved between Q/G.^a**Z** indicates the positions of Amber site

Table A.11 – Plasmid combinations and yields for variants used in this study

Variant	G1(ncAA)RS	Supplemented fluoroindole / concentration	Yield (mg per 1 L cell culture)
RF2-W42-4FW	G1F4W27	4-fluoroindole / 2 mM	26 ^a
RF2-W42-5FW	G1F4W27	5-fluoroindole / 0.5 mM	34 ^a
RF2-W42-6FW	G1F7W36	6-fluoroindole / 0.5 mM	26 ^a
RF2-W42-7FW	G1F7W36	7-fluoroindole / 0.5 mM	18 ^a

^aYield is determined from the volume and the protein concentration as measured after primary IMAC, using the extinction coefficient calculated by ExPASy ProtParam,¹² with Z substituted for F and a molecular weight of 8075 Da. The final mass from 50 mL cultures was then extrapolated to yield per 1 L cell culture.

A.2.2 Intact protein mass spectrometry

Intact protein analysis was performed on an Orbitrap Fusion™ Tribrid™ mass spectrometer (Thermo Fisher Scientific, USA) connected to a Thermo Fisher Scientific UltiMate 3000 HPLC system equipped with ZORBAX 300SB-C3, 3.5 μm, 4.6 x 50 mm HPLC column (Agilent Technologies, USA). Approximately 50 pmol of sample was injected using a 500 μL/min linear gradient of solvent A (0.1% (v/v) formic acid in water) and solvent B (0.1% (v/v) formic acid in acetonitrile), ramping solvent B from 5% solvent B at the start to 80% after 12 min. Data were collected using an ESI source in positive ion mode. Protein intact mass was determined by deconvolution using the program Xcalibur 3.0.63 (Thermo Fisher Scientific, USA).

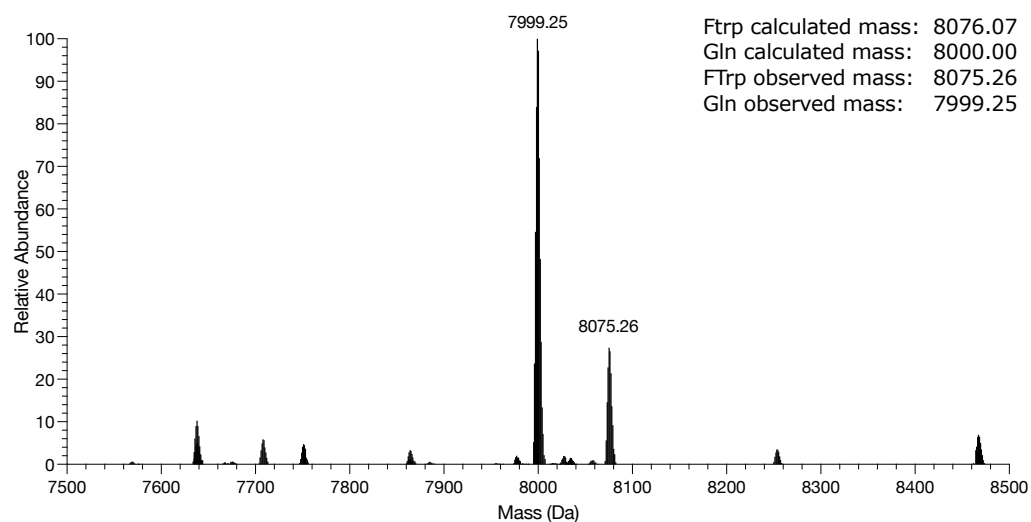


Figure A.46 - Intact protein mass spectrometry of RF2 W42 4FW

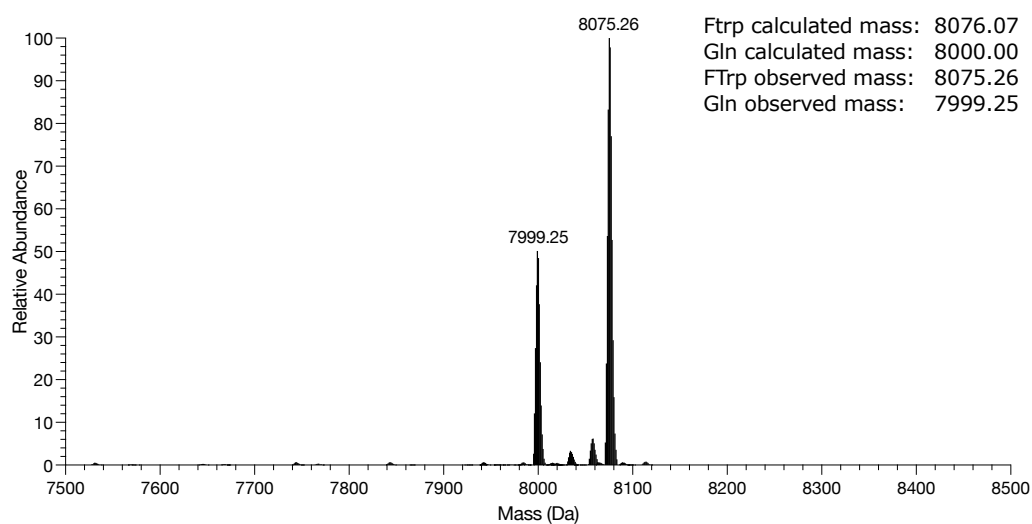


Figure A.47 - Intact protein mass spectrometry of RF2 W42 5FW

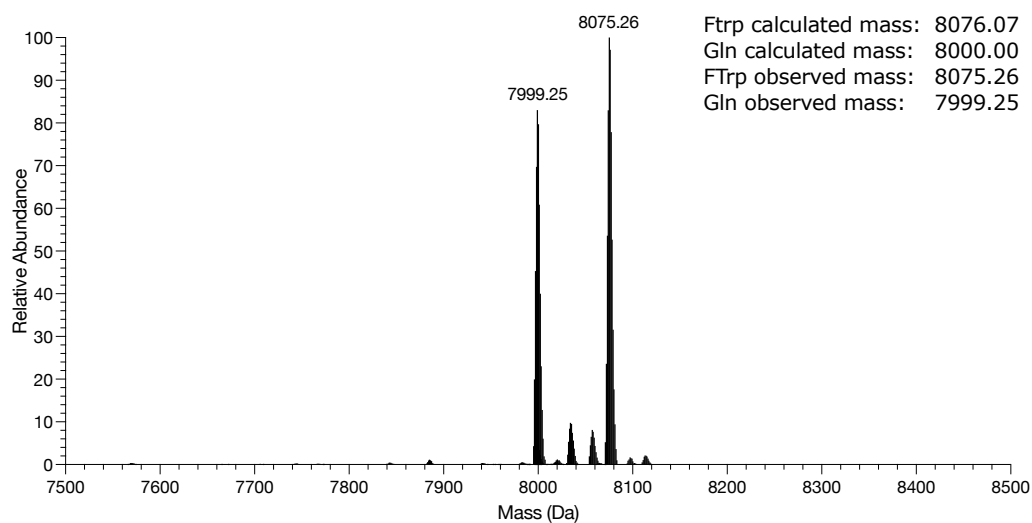


Figure A.48 - Intact protein mass spectrometry of RF2 W42 6FW

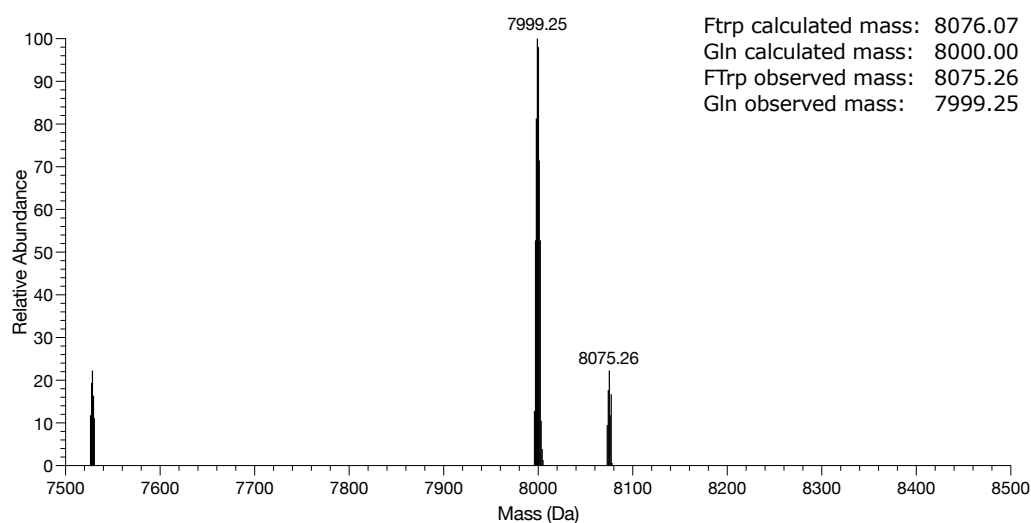


Figure A.49 - Intact protein mass spectrometry of RF2 W42 7FW

A.2.3 Rotamer distributions

Using the TEV-cleaved RF2 sequence from Table A.10, a request was made to the Alphafold3 (AF3) webserver¹³ using 1 copy of the protein sequence, 1 copy of Ca^{2+} , and a seed of 42. The output pseudo-ensemble was then used for subsequent modelling. C_α -alignment root mean square deviations (RMSDs) of each model to the highest confidence model were less than 0.5 Å.

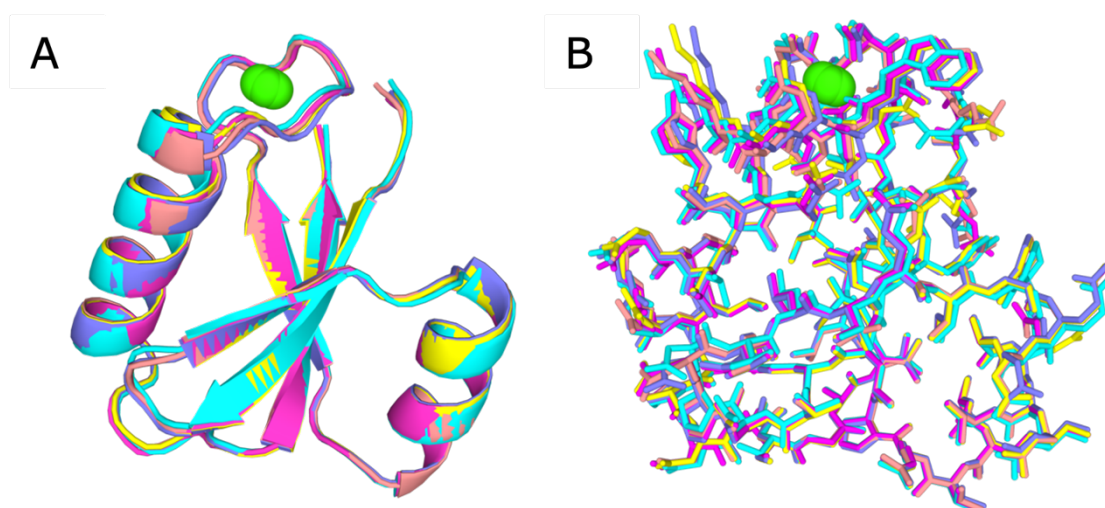


Figure A.50 - Superimposed AF3 models of RF2 with Ca^{2+} , C_α -alignment RMSD to model_0 (dark blue) was 0.484 Å (model_1, cyan), 0.238 Å (model_2, magenta), 0.177 Å (model_3, yellow) and 0.242 Å (model_4, salmon). Ca^{2+} is represented as green spheres. A) models shown in the cartoon illustration. B) models shown in the stick illustration.

A.2.4 Model generation and clash score calculation

Each model was loaded into PyMOL,² atoms ' $H_{\epsilon 3}$ ', ' $H_{\zeta 3}$ ', ' $H_{\eta 2}$ ' and ' $H_{\zeta 2}$ ' were rebuilt with fluorine atoms correspondingly named 'F4', 'F5', 'F6' and 'F7'. In PyMOL, the fluorine coordinates are automatically adjusted to match a C-F bond length. The C_{α} was replaced with a Gd atom. A custom script for running in PyMOL was written to sample χ_1 and χ_2 angles at regular intervals and perform operations on the sampled structures (calculation of distances, angles, and clashes). Sampling intervals of 5° were used.

The Gd-F distance for all fluorines was measured for each interval, for each model. For each (χ_1, χ_2) the mean distance and standard deviation (SD) for each Gd-F was computed. Figure A.51 shows the plotted distances, and Figure A.52 below shows the SD for each distance. For any given (χ_1/χ_2) , SD did not exceed $0.2 \text{ \AA}$.

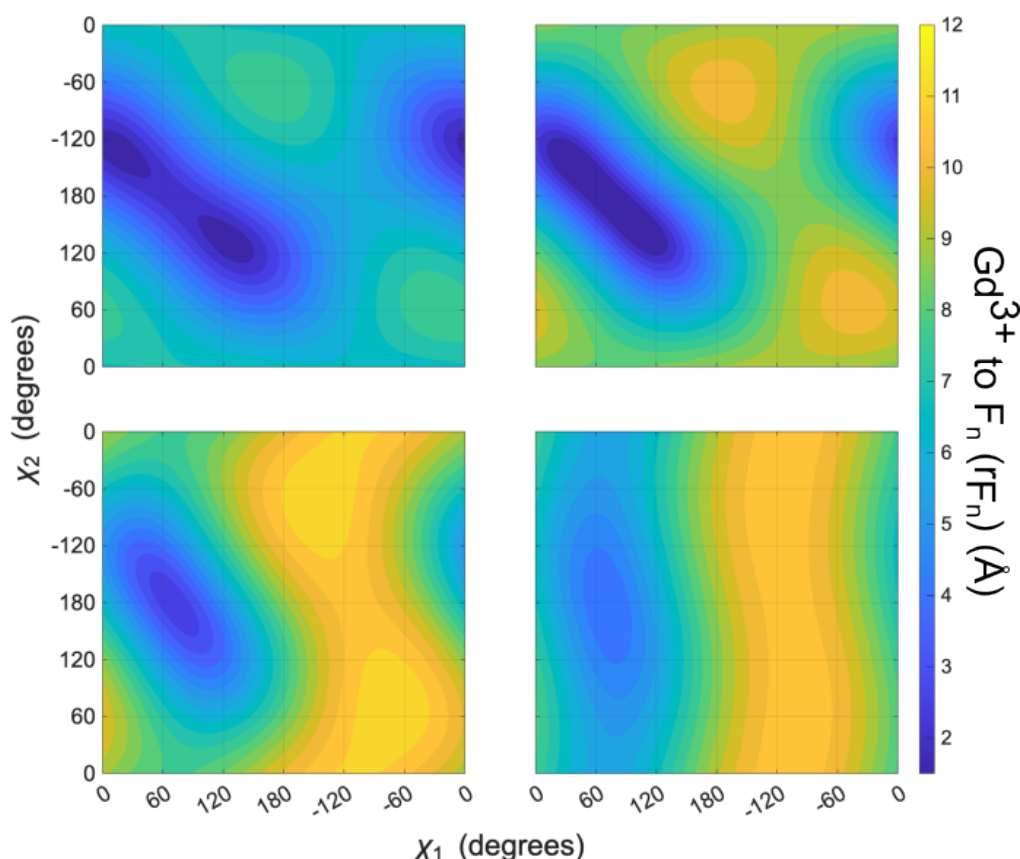


Figure A.51 – Contour plot of parametrisation of mean ^{19}F - Gd^{3+} distances ($\text{\AA}$) in generated models using 5° binning.

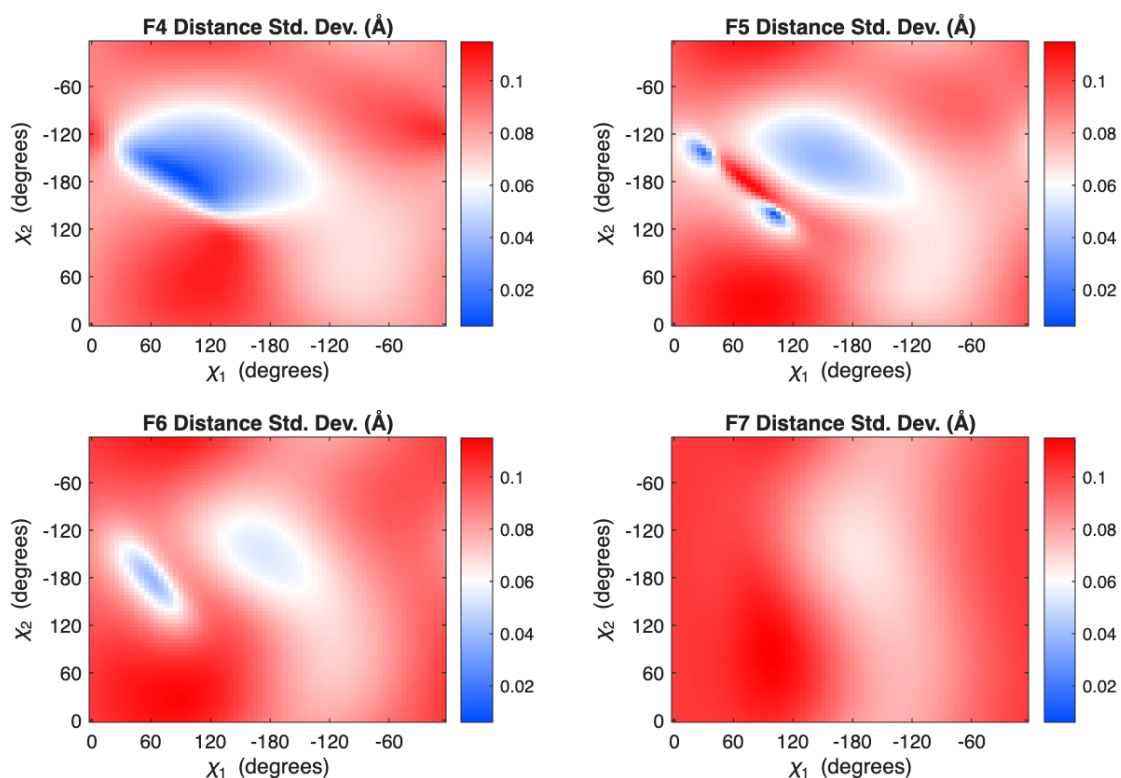


Figure A.52 - SD plots of F-Gd distances (Å) in generated models using 5° binning.

For each (χ_1/χ_2) two sets of selected coordinates were temporarily generated. Set A contains sidechain atoms of the sampled residue at that (χ_1/χ_2) . Set B contains all atoms whose minimum distance to any atom in A was ≤ 4.0 Å; atoms in A and the sampled residue's C_α were excluded. For every pair (i,j) , where $i \in A$ and $j \in B$, the centre-centre distance between each atom (d_{ij}) was calculated and then subtracted from the sum of van der Waals radii R for pair i,j , giving the atomic overlap Δ_{ij} :

$$\Delta_{ij} = (R_i + R_j) - d_{i,j} \quad 11$$

Pairs with $\Delta_{ij} > 0$ were counted as clashes. We recorded the count $N = |\{(i,j) \in A \times B : \Delta_{ij} > 0\}|$ and the sum $S = \sum_{(i,j) \in A \times B} \Delta_{ij} \mathbf{1}[\Delta_{ij} > 0]$. Clash severity for that (χ_1/χ_2) could then be calculated as:

$$\mu = \frac{S}{N}$$

12

Mean clash metrics for (χ_1/χ_2) across 5 models are plotted in Figure A.53.

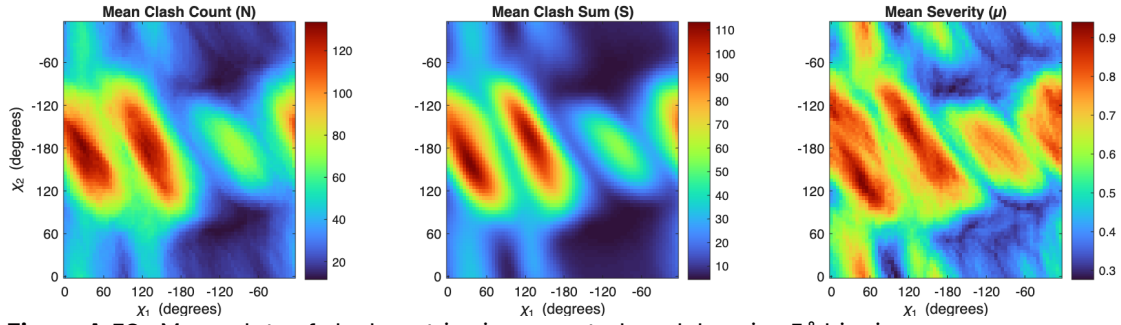


Figure A.53 - Mean plots of clash metrics in generated models using 5° binning.

A.3 References

1. Qianzhu, H.; Abdelkader, E. H.; Welegedara, A. P.; Habel, E.; Paul, N.; Frkic, R. L.; Jackson, C. J.; Huber, T.; Otting, G. Rendering proteins fluorescent inconspicuously: Genetically encoded 4-cyanotryptophan conserves their structure and enables the detection of ligand binding sites. *Angew. Chem., Int. Ed.* **2025**, 64, e202421000. <https://doi.org/10.1002/anie.202421000>
2. The PyMOL Molecular Graphics System, Version 3.0, Schrödinger, LLC.
3. Cook, E. C.; Featherston, E. R.; Showalter, S. A.; Cotruvo, J. A. Structural basis for rare earth element recognition by lanmodulin. *Biochemistry* **2019**, 58, 120–125. <https://doi.org/10.1021/acs.biochem.8b01019>
4. Chaudhury, S.; Lyskov, S.; Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **2010**, 26, 689–691. <https://doi.org/10.1093/bioinformatics/btq007>
5. Park, H.; Bradley, P.; Greisen, P., Jr.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **2016**, 12, 6201–6212. <https://doi.org/10.1021/acs.jctc.6b00819>
6. Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, 620, 1089–1100. <https://doi.org/10.1038/s41586-023-06415-8>
7. Bünzli, J.-C. G. Lanthanide luminescence for biomedical analyses and imaging. *Chem. Rev.* **2010**, 110, 2729–2755. <https://doi.org/10.1021/cr900362e>
8. Kimura, T.; Kato, Y. Luminescence study on determination of the hydration number of Sm(III) and Dy(III). *J. Alloys Compd.* **1995**, 225, 284–287. [https://doi.org/10.1016/0925-8388\(94\)07084-9](https://doi.org/10.1016/0925-8388(94)07084-9)
9. Beeby, A.; Clarkson, I. M.; Dickins, R. S.; Faulkner, S.; Parker, D.; Royle, L.; de Sousa, A. S.; Williams, J. A. G.; Woods, M. Non-radiative deactivation of the excited states of europium, terbium and ytterbium complexes by proximate energy-matched OH, NH and CH oscillators: An improved luminescence method for establishing solution hydration states. *J. Chem. Soc., Perkin Trans. 2* **1999**, 493–504. <https://doi.org/10.1039/A808692C>
10. Abdelkader, E. H.; Qianzhu, H.; Huber, T.; Otting, G. Genetic incorporation of azatryptophans enhances PETase activity. *Unpublished work*.
11. Abdelkader, E. H.; Qianzhu, H.; George, J.; Frkic, R. L.; Jackson, C. J.; Nitsche, C.; Otting, G.; Huber, T. Genetic encoding of cyanopyridylalanine for in-cell protein macrocyclization by the nitrile–aminothiol click reaction. *Angew. Chem., Int. Ed.* **2022**, 61, e202114154. <https://doi.org/10.1002/anie.202114154>
12. Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R. D.; Bairoch, A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **2003**, 31, 3784–8. <https://doi.org/10.1093/nar/gkg563>
13. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C. C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Zengulyte, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlín, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Zidek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, 630, 493–500. <https://doi.org/10.1038/s41586-024-07487-w>