CREATION OF NOVEL PROTEINS FOR PEPTIDE RECOGNITION AND INTERACTION

BITASADAT HOSSEINI

Doctor of Philosophy

ASTON UNIVERSITY

December 2024

Creation of novel proteins for peptide recognition and interaction

Bitasadat Hosseini

Doctor of Philosophy

Health and Life Sciences, Aston University

2024

Thesis Abstract

The SpyTag-SpyCatcher system, developed by the Howarth lab at the University of Oxford, is based on splitting the immunoglobulin-like domain CnaB2 from the fibronectin-binding protein FbaB of *Streptococcus pyogenes* into two functional components: the 13-amino-acid SpyTag peptide and the 116-amino-acid SpyCatcher protein. Upon incubation, SpyTag and SpyCatcher spontaneously form a covalent isopeptide bond between Asp7 of SpyTag and Lys31 of SpyCatcher.

This study investigates whether the specificity of the SpyTag-SpyCatcher interaction can be modulated through targeted substitutions within the hydrophobic binding pocket of SpyCatcher and corresponding SpyTag residues, by exploring alternative hydrophobic residues, with the aim to develop orthogonal SpyTag-SpyCatcher pairs. Molecular modelling guided the design of positionally-fixed SpyCatcher and SpyTag libraries, constructed using overlap PCR and MAX randomisation. To assess the specificity, a novel screening strategy was developed, combining SDS-PAGE for semi-quantitative binding analysis with mass photometry for precise interaction detection within complex mixtures.

Screening the SpyCatcher variants against SpyTag libraries demonstrated that substitution of larger, hydrophobic, aliphatic residues is well-tolerated, while aromatic bulky residues can eliminate the interaction. Comparative analyses between SDS-PAGE and mass photometry demonstrated strong consistency, highlighting the reliability of both methods in assessing binding specificity. The findings underscore the structural importance of specific residues in the SpyTag-SpyCatcher interaction and validate mass photometry as a transformative tool for screening combinatorial protein libraries. This represents the first application of mass photometry for screening protein libraries, showcasing its potential to evaluate protein-ligand specificity with minimal sample requirements. This study opens new avenues for the application of mass photometry by validating both the quality and specificity of the constructed libraries and opening of new avenues for its use in protein engineering.Key words: SpyTag, SpyCatcher, combinatorial libraries, MAX randomisation, overlap PCR, orthogonal pair, mass photometry.

Dedicated to the pure soul of my Dad…..

Acknowledgment

I would like to express my heartfelt gratitude to my supervisor, Prof. Anna V. Hine, for her exceptional guidance, unwavering support, and constant encouragement throughout my project. Your invaluable insights and the lessons you have imparted will remain with me always. I am deeply thankful for your mentorship. I also extend my sincere thanks to my co-supervisor, Prof. Corrinne Spickett, for her guidence and endless support, and to Dr. Andrew Sutherland for his invaluable guidance and contributions to my work.

My deepest appreciation goes to Dr. Mohammed Ashraf, whose endless support and technical assistance have been pivotal throughout this journey. Your countless favors, generous help, and delicious meals have truly made a huge impact on me. I am especially thankful to Dr. Anupama Chembath for her valuable insights, training, and constant presence. Your encouragement has been a beacon of strength for me, and I could not have reached this milestone without your support.

I would also like to acknowledge Dr. Philip Kitchen for his guidance with mass photometry, Professor Alan Goddard for his support, Professor Ahmadi,as my best mentor since I started my master. My thanks also go to Dr. Russel Collighan, Dr. John Reynolds for the great discussions, Dr. Caroline Kardeby, Dr. Nourbakhsh, and Dr. Fadaie and many others who have contributed to my journey.

A special thanks to my dear lab mates: Idoia, Maral, Ophelie, Yanis, Jacob, Fatima, and Marcella, for their camaraderie and collaboration. I am also grateful to my friends: Annelise, Carolina, Laura, Zahra, Sarah, and especially my best friends, Yalda and Polina, for their increcible support and encouragement throughout this journey.

To my husband, Navid, I owe my deepest gratitude for your unwavering support, patience, and love. You believed in me even during the most challenging moments of this PhD, when I lost my Dad. You have been my rock, my cheerleader, and my inspiration. I could not have achieved this without you by my side.

Lastly, I am deeply thankful to my family for providing me with a strong foundation and encouragement throughout this journey. A special thanks to my incredible mom—your support has been invaluable, and having you by my side has been a blessing ! To my wonderful sisters; Neda and Vida, and my dear in law's; Mitra and Neda,I am endlessly grateful for your love and encouragement. Thank you all for being my greatest source of strength.

List of Contents

Chapter 1: Introduction.....16

1.1. Spytag-Spycatcher technology 16

1.1.1. History of Spytag-Spycatcher technology 16

1.1.2. SpyTag-SpyCatcher structure 17

1.1.3. The isopeptide-bond.....17

1.1.4. Mechanism of action of SpyTag-SpyCatcher.....18

1.1.5. Engineering SpyCatcher for enhanced reaction speed.....20

1.1.6. Alternatives to SpyTag-SpyCatcher systems.....21

1.1.7. Key Differences between SnoopTag/SnoopCatcher and SpyTag-SpyCatcher.....22

1.2. Application of SpyTag-SpyCatcher technology..... 24

1.2.1. Protein engineering and modification 24

1.2.2. Protein characterisation and analysis 24

1.2.3. Cellular and in vivo applications 25

1.2.4. Vaccine development.....25

1.2.5. Other applications.....25

1.2.6. Advantages of SpyTag-SpyCatcher 26

1.2.7. Limitations of SpyTag-SpyCatcher 27

1.3. Engineering novel proteins via saturation mutagenesis 28

1.3.1. Reducing degeneracy in saturated libraries.....29

1.3.2. Nondegenerate saturation mutagenesis 31

1.4. Combinatorial libraries and screening..... 37

1.4.1. Historical context of combinatorial chemistry.....37

1.4.2. Methods for creating combinatorial libraries 38

1.4.3. Applications of combinatorial libraries 40

1.4.4. Directed evolution in combinatorial protein libraries 41

1.4.5. Screening biological protein libraries.....41

1.4.6. Identifying zinc finger proteins using positional fixing 42

1.4.7. Future direction 43

1.5. Mass photometry 44

1.5.1. Mass photometry as an analytical tool.....44

1.5.2. The principle of mass photometry 45
1.5.3. Applications of mass photometry 47
1.5.4. Advantages of MP over other methods 49
1.5.5. Limitations of mass photometry 51
1.5.6. Future directions of mass photometry 52
1.6. The aims and objectives of the project 52
Chapter 2: Materials and Methods 54
2.1. Materials 54
2.1.1. Risk assessment 54
2.1.2. Vectors 54
2.1.3. Cell lines 54
2.1.4. Media recipes 54
2.1.5. Buffer recipes 55
2.1.6. Solutions by Application 57
2.2. Methods 58
2.2.1. Methodology for utilising PyMOL in molecular modelling 58
2.2.2. Molecular procedures 65
2.2.3. Electrophoresis 68
2.2.4. Competent cells preparation 69
2.2.5. Transformation into E. coli 70
2.2.6. Sequencing 70
2.2.7. Gene Expression 72
2.2.8. Lysate production using BugBuster 73
2.2.9. Protein purification 73
2.2.10. Optimisation of peptide-protein binding conditions 74
2.2.11. Mass photometry 74
Chapter 3: Molecular visualisation of SpyTag and SpyCatcher 79
3.1. Introduction to PyMOL for molecular modeling 79
3.2. Visualisation of the native SpyTag-SpyCatcher interaction 80
3.3. Visualisation of the SpyTag and SpyCatcher residues proposed for mutagenesis 82
3.4. In silico mutagenesis 83

3.5. Discussion 85
Chapter4: SpyCatcher library construction 86
4.1. Creation of a SpyCatcher expression vector 86
4.2. Library design 87
4.3. Creating fixed-position library fragments via overlap PCR 88
4.4. Library synthesis strategy 89
4.4.1. To fix position 27, while randomising position 44 and position 90 90
4.4.2. To fix position 44, while randomising position 27 and position 90 92
4.4.3. To fix position 90, while randomising position 27 and position 44 94
4.4.4. Preparing the library cassettes for cloning 96
4.5. Cloning the libraries 98
4.6. Sanger sequencing of the SpyCatcher libraries 98
4.6.1. Observed verses expected distribution of bases of first-fixed libraries 100
4.6.2. Observed verses expected distribution of bases of second-fixed libraries 105
4.6.3. Observed verses expected distribution of bases of third-fixed libraries 112
4.7. Next Generation Sequencing of the SpyCatcher libraries 118
4.7.1. Amino acid distribution of the first-fixed position libraries 118
4.7.1. Amino acid distribution of the second-fixed position libraries 122
4.7.3. Amino acid distribution of the third-fixed position libraries 126
4.8. Comparative analysis of Sanger sequencing and next-generation sequencing results 130
4.9. Gene expression 135
4.10. Expression of 18 SpyCatcher libraries 136
4.10.1. IPTG Concentration 136
4.10.2. Temperature 137
4.11. Affinity purification of 18 SpyCatcher libraries 137
4.12. Discussion 139
Chapter 5: Peptide library construction 140
5.1. Generation of native SpyTag-mCherry plasmid 141
5.2. SpyTag library design 143

5.3. Construction of SpyTag libraries 145
5.4. Cloning of SpyTag libraries 148
5.5. Sanger sequencing of the 12 SpyTag libraries 148
5.6. SpyTag library expression 151
5.7. Discussion 153
Chapter 6: Interactions of native and mutant SpyTag-SpyCatcher proteins 154
6.1. Interactions of native and mutant SpyTag-SpyCatcher proteins 154
6.1.1. Interaction of native-SpyCatcher with native-SpyTag 154
6.1.2. Specificity of native SpyCatcher for SpyTag 155
6.1.3. Specificity of SpyCatcher for position 3 of SpyTag as assessed by SDS-PAGE 156
6.1.4. Specificity of SpyCatcher for position 5 of SpyTag as assessed by SDS-PAGE 156
6.1.5. Mass photometry analysis of the native SpyCatcher - native SpyTag interaction 156
6.1.6. Specificity of native SpyCatcher for position 3 of SpyTag as assessed by mass photometry 158
6.1.7. Specificity of native SpyCatcher for position 5 of SpyTag as assessed by mass photometry 160
6.2. Interactions of mutated SpyCatcher proteins (libraries) with native-SpyTag 162
6.2.1. Fixed position 27 libraries (native SpyCatcher = isoleucine) 164
6.2.2. Fixed position 44 libraries (native SpyCatcher = methionine) 164
6.2.3. Fixed Position 90 libraries (native SpyCatcher = isoleucine) 164
6.2.4. Mass photometry of the fixed-position SpyCatcher libraries with native SpyTag 165
6.3. Discussion 171
Chapter 7: Interactions of novel SpyCatcher variants 175
7.1. Creation of newly discovered SpyCatcher proteins 175
7.1.1. Creation of genes encoding novel SpyCatcher “binders” 176
7.1.2. Gene cloning 179
7.1.3. Sanger sequencing of the newly constructed SpyCatcher genes 179
7.1.4. Expression and purification of novel SpyCatcher proteins 181

7.2. Examination of novel SpyCatcher interactions with SpyTag libraries via SDS-PAGE analysis 181

7.3. Examination of novel SpyCatcher interactions with SpyTag libraries via mass photometry 186

7.4. Discussion 192

Chapter 8: Conclusion 194

References 199

Annex 1 Oligonucletide sequences……..…205

Annex 2 SpyCatcher-mNeongreen plasmid sequence (6335bp)…….207

Annex 3 SpyTag-mCherry plasmid sequence (602bp) ……...208

List of Figures

Figure 1.1. SpyTag-SpyCatcher structure (PDB ID:4MLI) obtained from ……..……17
Figure 1.2. Isopeptide bond formation……..……18
Figure 1.3. How the iso-peptide bond is formed between the residues ……..……19
Figure 1.4. SpyCatcher002 to 003 mutations ……..……..……21
Figure 1.5. Applications of the SpyCatcher-SpyTag system ……25
Figure 1.6. Comparison of the effectiveness of common saturation mutagenesis ……..…..29
Figure 1.7. Schematic representation of MAX randomisation technique ……..32
Figure 1.8. Schematic representation of overlap PCR ……..…..35
Figure 1.9. Schematic representation of positionally fixed for peptide library screening …...38
Figure 1.10. Working principles of mass photometry ……...44
Figure 1.11. An example of capturing the contrast of numerous molecules ……..44
Figure 1.12. Example screenshots from the MP camera ……..…… ..45
Figure 1.13. Single-molecule mass photometry (SMMP)……..……..46
Figure 2.1. The PDB ID for SpyTag-SpyCatcher complex ……..59
Figure 2.2. PyMOL interface …… ..59
Figure 2.3. Representation of protein structure ……..…… 60
Figure 2.4. Representation of various options for viewing the protein ……..……..61
Figure 2.5. Renaming selected residues in PyMOL …… 61
Figure 2.6. Displaying polar contacts using the Action menu …… 62
Figure 2.7. Visualisation of residue interactions ……..…… 63
Figure 2.8. Displaying water molecules and interactions …….. 63
Figure 2.9. Measuring the distance between two selected residues …… 64
Figure 2.10. Mutagenesis menu for residue mutation ……..…….. 65
Figure 2.11. Mutation of reside using the mutagenesis Wizard …….. 65
Figure 2.12. Detecting bubbles in the oil. …….. 75
Figure 2.13. Representative examples of sample concentrations molecule landing events . 76
Figure 2.14. A screenshot of the data analysis software for mass photometry ……..….. 77
Figure 3.1. Schematic representation of SpyTag-SpyCatcher …… 81
Figure 3.2. Main-Chain hydrogen bonds between SpyCatcher and SpyTag ……... 82
Figure 3.3. Visualisation of key SpyCatcher ……..…… 83
Figure 3.4. Mutagenesis in PyMOL …… 84
Figure 3.5. Schematic representation of the residues in the hydrophobic pocket …….. 85
Figure 4.1. Amplification and assembly of the SpyCatcher gene …….. 86
Figure 4.2. 18 Fixed positional SpyCatcher libraries …… 88
Figure 4.3. SpyCatcher library mechanism of construction ……..…… 89

Figure 4.4. Library synthesis strategy ……..…..90
Figure 4.5. Agarose gel electrophoresis of SpyCatcher fragments …….. 91
Figure 4.6. Overlap PCR of fixed position 27 libraries …… 92
Figure 4.7. Agarose gel electrophoresis of SpyCatcher fragments ……. 93
Figure 4.8. Agarose gel electrophoresis of SpyCatcher fragments ……..…… 95
Figure 4.9. PCR Amplification for Generating Overlap Products …….. 96
Figure 4.10. Full length PCR products of 18 SpyCatcher libraries ……97
Figure 4.11. PCR Amplification of plasmid backbone for cloning …….. 98
Figure 4.12. The expected distribution of bases ……..… 99
Figure 4.13. Observed vs. expected result of Sanger sequencing of the IXX library ..……...100
Figure 4.14. Observed vs. expected result of Sanger sequencing of the VXX library .....….101
Figure 4.15. Observed vs. expected result of Sanger sequencing of the LXX library …….....102
Figure 4.16. Observed vs. expected result of Sanger sequencing of the FXX library ……...103
Figure 4.17. Observed vs. expected result of Sanger sequencing of the MXX library ……...104
Figure 4.18. Observed vs. expected result of Sanger sequencing of the YXX library ……...105
Figure 4.19. Observed vs. expected result of Sanger sequencing of the XIX library …….106
Figure 4.20. Observed vs. expected result of Sanger sequencing of the XVX library ……...107
Figure 4.21. Observed vs. expected result of Sanger sequencing of the XLX library ……...108
Figure 4.22. Observed vs. expected result of Sanger sequencing of the XFX library ……...109
Figure 4.23. Observed vs. expected result of Sanger sequencing of the XMX library ……..110
Figure 4.24. Observed vs. expected result of Sanger sequencing of the XYX library ……...111
Figure 4.25. Observed vs. expected result of Sanger sequencing of the XXI library ……..…112
Figure 4.26. Observed vs. expected result of Sanger sequencing of the XXV library ……...113
Figure 4.27. Observed vs. expected result of Sanger sequencing of the XXL library ……...114
Figure 4.28. Observed vs. expected result of Sanger sequencing of the XXF library ……115
Figure 4.29. Observed vs. expected result of Sanger sequencing of the XXM library 116
Figure 4.30. Observed vs. expected result of Sanger sequencing of the XXY library ……...117
Figure 4.31. Observed and expected distribution of encoded amino acids ……....119
Figure 4.32. Observed and expected distribution of encoded amino acids …….. 123
Figure 4.33. Observed and expected distribution of codons ……..…. 127
Figure 4.34. Sanger sequencing versus next generation sequencing of the IXX library ….. 131
Figure 4.35. Sanger sequencing versus next generation sequencing of the XMX library … 132
Figure 4.36. Sanger sequencing versus next generation sequencing of the XXY library … 133
Figure 4.37. Fluorescent colonies of native SpyCatcher-mNeongreen …….. 135
Figure 4.38. Analysis of wild-type SpyCatcher protein expression …….. 136
Figure 4.39. SDS-PAGE analysis of wild-type SpyCatcher protein expression ……. 137

Figure 4.40. Purification of (A) wt. SpyCatcher, and (B) mutated SpyCatcher …….. 138
Figure 4.41. Purification of the SpyCatcher Libraries ……. 138-139
Figure 5.1. Amplification of the genes to make native SpyTag-mCherry plasmid ……..... 141
Figure 5.2. Amplification of the SpyTag-mCherry plasmid for cloning ……..…. 142
Figure 5.3. The amino acid sequence of SpyTag peptide ……..…. 143
Figure 5.4. Schematic representation of SpyTag library design …….. 143
Figure 5.5. Schematic representation of the DNA oligonucleotide design …… 144
Figure 5.6. Denaturation temperature gradient using neat ligation product …….. 145
Figure 5.7. Agarose gel electrophoresis of 12 SpyTag library fragments ……..…….. 147
Figure 5.8. Sanger sequencing of SpyTag-libraries ……..……..148-151
Figure 5.9. Fluorescent colonies of SpyTag-mCherry ……..… 152
Figure 5.10. Purification of native SpyTag-mCherry protein ……..…….. 152
Figure 6.1. Covalent binding reconstitution between native-SpyTag ……..…….. 154
Figure 6.2. Covalent binding between native-SpyCatcher and SpyTag libraries ……..….. 155
Figure 6.3. Mass photometry analysis …….. 157
Figure 6.4. Mass Photometry analysis of native-SpyCatcher with fixed position 3 ……..… 159
Figure 6.5. Mass photometry analysis of native SpyCatcher with fixed position 5 ……..…161
Figure 6.6. Covalent bonding analysis …….. 163
Figure 6.7. Mass Photometry analysis of fixed-position 27 SpyCatcher libraries (NXX) …. 166
Figure 6.8. Mass Photometry analysis of fixed-position 44 SpyCatcher libraries (XNX). …. 168
Figure 6.9. Mass Photometry analysis of fixed-position 90 SpyCatcher libraries (XXN) …. 170
Figure 7.1. Agarose gel electrophoresis of novel SpyCatcher fragments ……..…… 177
Figure 7.2. PCR Amplification for generating overlap and full-length products ……..….. 178
Figure 7.3. Sanger sequencing of the newly synthesised SpyCatcher genes ……..… 180
Figure 7.4. Purification of mutated SpyCatcher by affinity chromatography …….. 181
Figure 7.5. Covalent binding reconstitution …….. 183
Figure 7.6. Covalent binding reconstitution ……..…….. 184
Figure 7.7. Mass photometry analysis of SpyCatcher ILI with selected-fixed position ……. 187
Figure 7.8. Mass photometry analysis of SpyCatcher LLI with selected-fixed position ……. 188
Figure 7.9. Mass photometry analysis of SpyCatcher LLV with selected-fixed position ..… 189
Figure 7.10. Mass photometry analysis of non-binders SpyCatcher ……..……..…. 190
Figure 8.1. Schematic representation of targeted residues …….. 196
Figure 8.2. Interaction between tyrosine (Y9) of SpyTag and aspartic acid (D35) ……..… 197
Figure 8.3. Interaction between lysine (K10) of SpyTag and glutamic acid (E85) ……. 197

List of Tables

Table 2.1. Overview of vectors utilised in the study ……54
Table 2.2. Overview of Cell lines utilised in the study …… 54
Table 2.3. Components of SDS-PAGE gels …… 69
Table 5.1. Sequences of MAX selection oligonucleotide pools …… 144
Table 7.1. SpyCatcher residue substitution compared to native-SpyCatcher …… 176
Table 8.1. Comparative analysis of SpyCatcher variants interacting with SpyTag libraries .. 195
Table 9.1. Oligos utilised to make SpyCatcher-mNeongreen plasmid……..205
Table 9.2. Oligos utilized to make Native SpyTag-mCherry plasmid……..205
Table 9.3. Oligos utilised to construct 18 distinct SpyCatcher libraries ……206

List of Abbreviations

APS	Ammonium Persulfate
ATG	Start Codon in DNA (Adenine, Thymine, Guanine)
ATP	Adenosine Triphosphate
BCA	Bicinchoninic Acid (commonly used in protein quantification)
BLI	Biolayer Interferometry
BSA	Bovine Serum Albumin
CLIP	Cross-Linking and Immunoprecipitation
COUNTIF	Excel function for counting cells that meet a criterion
CTG	Codon encoding Leucine in DNA (Cytosine, Thymine, Guanine)
DC	Double Combinatorial
DNA	Deoxyribonucleic Acid
DTT	Dithiothreitol (reducing agent)
GTG	Codon encoding Valine in DNA
DTA	Ethylenediaminetetraacetic Acid
HC	Heavy Chain (in immunology)
HEPES	4-(2-Hydroxyethyl)-1-Piperazineethanesulfonic Acid (buffer)
HF	High-Frequency
HIV	Human Immunodeficiency Virus
HPLC	High-Performance Liquid Chromatography
IIS	Type II Secretion System
IPTG	Isopropyl β-D-1-thiogalactopyranoside (inducer in protein expression)
ITC	Isothermal Titration Calorimetry
LB	Luria-Bertani (nutrient-rich medium for bacteria)
MAX	Non-degenerate saturation mutagenesis method
MP	Mass Photometry
MW	Molecular Weight
NC	Negative Control
NDT	Degenerate codon subset for saturation mutagenesis
NEB	New England Biolabs (biotech company)

NGS	Next-Generation Sequencing
NNK	Degenerate Codon (used in mutagenesis)
NNN	Fully Degenerate Codon (encodes all amino acids)
NNS	Degenerate Codon (used in mutagenesis)
NTA	Nitrilotriacetic Acid (used in affinity chromatography)
OD	Optical Density
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Phosphate-Buffered Saline
PC	Positive Control
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PNG	Portable Network Graphics
PYMOL	Molecular visualization software
RMS	Root Mean Square
RNA	Ribonucleic Acid
RSB	Reaction Sample Buffer
SAM	S-Adenosyl Methionine
SC	SpyCatcher
SDS	Sodium Dodecyl Sulfate
SMMP	Single-Molecule Mass Photometry
SNR	Signal-to-Noise Ratio
SOC	Super Optimal Broth with Catabolite Repression
ST	SpyTag
TAE	Tris-Acetate-EDTA Buffer
TEMED	Tetramethylethylenediamine (used in gel electrophoresis)
TGG	Codon encoding Tryptophan in DNA
TRIM	Technique for reducing degeneracy in mutagenesis
TTA	Codon encoding Leucine in DNA
TTT	Codon encoding Phenylalanine in DNA
UV	Ultraviolet
ZFH	Zinc Finger Homolog

Chapter 1 Introduction

1.1. Spytag-Spycatcher technology

The SpyTag-SpyCatcher system is a powerful and versatile tool used in protein engineering. The technology enables protein ligation, labelling, and enhanced protein stability, harnessing proteins' natural ability to form covalent bonds, to create covalent bonding between protein components under physiological conditions. It has emerged as a powerful tool for fields requiring stable protein-protein linkages, such as protein assembly, bioconjugation, and synthetic biology, where robust, precise, and predictable covalent interactions are often essential (Zakeri et al., 2012).

1.1.1. History of Spytag-Spycatcher technology

The SpyTag-SpyCatcher system was first developed in 2010 by the Howarth lab in Oxford. The concept came from a called a protein called fibronectin-binding protein (FbaB) within the gram-positive bacterium Streptococcus pyogenes. Within this protein there is an immunoglobulin-like domain termed CnaB2, which plays a crucial role in enabling the bacterium to invade human cells (Amelung et al., 2011). The scientists split this domain into a larger incomplete immunoglobulin-like domain, of 116 residues (15 kDa), which they named SpyCatcher and a shorter peptide of 13 residues (1.6 kDa) named SpyTag. SpyCatcher and SpyTag retain high affinity for each other (Kd = 0.2 μM) and following interaction, form a stable covalent complex (Zakeri et al., 2017). When these two components bind to each other, they form a covalent bond between a lysine residue on SpyCatcher and an aspartic acid residue on SpyTag, without requiring additional reagents or extreme conditions (Li et al., 2014). Under typical experimental conditions (such as room temperature), this bond forms efficiently within minutes. SpyTag is comparable in size to other epitope tags and can be fused to proteins at either the N- or C-terminus or even internally, enabling it to bind with SpyCatcher when fused to target proteins (Wu et al., 2020; Howarth, 2016). This technology offers a highly specific and stable bond with rapid reaction kinetics, enabling flexibility in protein engineering applications, where maintaining structural integrity under a wide range of environmental conditions (e.g., temperature, pH) is crucial (Zakeri et al., 2012).

1.1.2. SpyTag-SpyCatcher structure

SpyCatcher adopts an immunoglobulin-like β-sandwich fold, resembling the structure of its parent domain, CnaB2.This structure comprises eight β-strands arranged in two sheets, forming a compact and stable domain (Li et al., 2014).

The crystal structure of the SpyTag-SpyCatcher complex provides detailed insights into the interaction. SpyTag-SpyCatcher structure (PDB ID:4MLI) obtained from the Protein Data Bank (PDB). Splitting CnaB2 (shown as cartoon in Figure 1.1) generates a protein of 116 residues called SpyCatcher and the remaining, a 13 amino acid tag as SpyTag. The N- and C-terminal segments of SpyCatcher are not essential for SpyTag binding and can be removed without significantly affecting the reaction rate. This has led to the development of a minimised SpyCatcher, which is 32 residues shorter than the original SpyCatcher protein but retains its ability to rapidly react with SpyTag (Li et al., 2014).

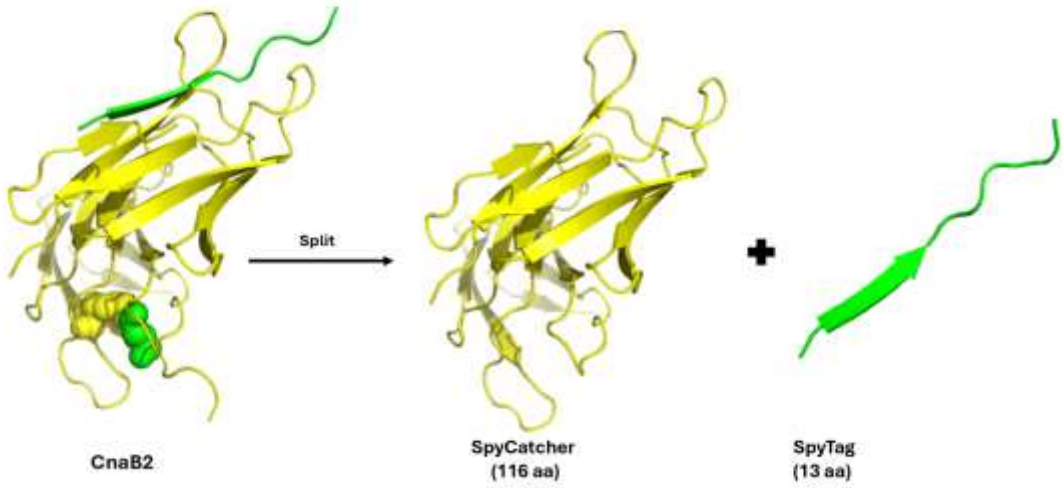


Figure 1.1. SpyTag-SpyCatcher structure (PDB ID:4MLI) obtained from the Protein Data Bank (PDB). *Splitting CnaB2 (shown in cartoon) generates a protein of 116 residues called SpyCatcher and the remaining, a 13 amino acid tag considered as SpyTag. The bond between the two partner has been shown in space-filling at the bottom of the CnaB2.*

1.1.3. The isopeptide-bond

Isopeptide bonds in bacteria are predominantly found in surface proteins of Gram-positive species. The two main classes of proteins to feature these bonds are the microbial surface components that recognise adhesive matrix molecules (MSCRAMMs) and pili proteins (Kang & Baker, 2011a). In pili, pilin subunits are covalently linked by an intermolecular isopeptide bond, connecting the C-terminal carboxyl of a threonine in one subunit to the ε-amine of a

lysine in the adjacent subunit. This bond is facilitated by a pilin-specific enzyme called sortase (Hendrickx et al., 2011). In MSCRAMMs, isopeptide bonds form spontaneously within the same domain, which typically belongs to the immunoglobulin fold superfamily (Sridharan & Ponnuraj, 2016).

An isopeptide bond is an amide bond formed between the side (or "R") groups of residues in a protein or peptide – i.e. outside of the main chain carbon and nitrogen atoms. As above, isopeptide bonds can form spontaneously, which can be considered as an indicator of gram-positive bacteria (Kang & Baker, 2011). However, spontaneous isopeptide bond formation was first observed in the bacteriophage HK97, where the bond forms between lysine and asparagine residues (Duda, 1998). Spontaneous isopeptide bond formation is an approach to cross-link polypeptides (Hae et al., 2007). Three residues which are typically involved in the formation of isopeptide bonds are Lys, Asn, and Glu (Zakeri & Howarth, 2010) (Figure 1.2).

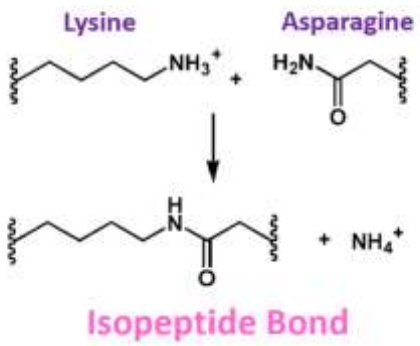


Figure 1.2. Isopeptide bond formation. *In this reaction, the amino group of the lysine side chain reacts with the carbonyl group of the asparagine side chain, resulting in the formation of a covalent isopeptide bond. This reaction releases an ammonium ion (NH_4^+) as a byproduct (Zakeri and Howarth, 2010).*

1.1.4. Mechanism of action of SpyTag-SpyCatcher

The interaction between the SpyTag and SpyCatcher involves two main steps:

Initially, SpyTag binds to SpyCatcher through non-covalent interactions, forming a complex which is driven by the insertion of SpyTag's Ile3 and Met5 into a hydrophobic pocket of SpyCatcher. Numerous parallel hydrogen bonds then form between SpyTag and residues 25–32 of the β-strand 1 of SpyCatcher (Li et al., 2014).

Specifically, following non-covalent association, the reactive lysine residue (Lys31) in SpyCatcher forms an isopeptide bond with the aspartate residue (Asp7) of SpyTag. This reaction is autocatalytic, meaning that it is facilitated by the catalytic glutamate (Glu77) within

SpyCatcher, and proceeds rapidly under a wide range of conditions (Reddington & Howarth, 2015) (Figure 1.3).The lysine residue (Lys31) is the site of covalent attachment to SpyTag, the aspartate (Asp7) is involved in the reaction mechanism, and the glutamate (Glu77) acts as a catalytic proton shuttl, facilitating the reaction (Reddington & Howarth, 2015; Hatlem et al., 2019). The reaction proceeds through a neutral tetrahedral intermediate and culminates in the release of a water molecule and the formation of the stable isopeptide bond (Hatlem et al., 2019). Mutation of Glu77 can disturb the formation of the bond (Hagan et al., 2010).

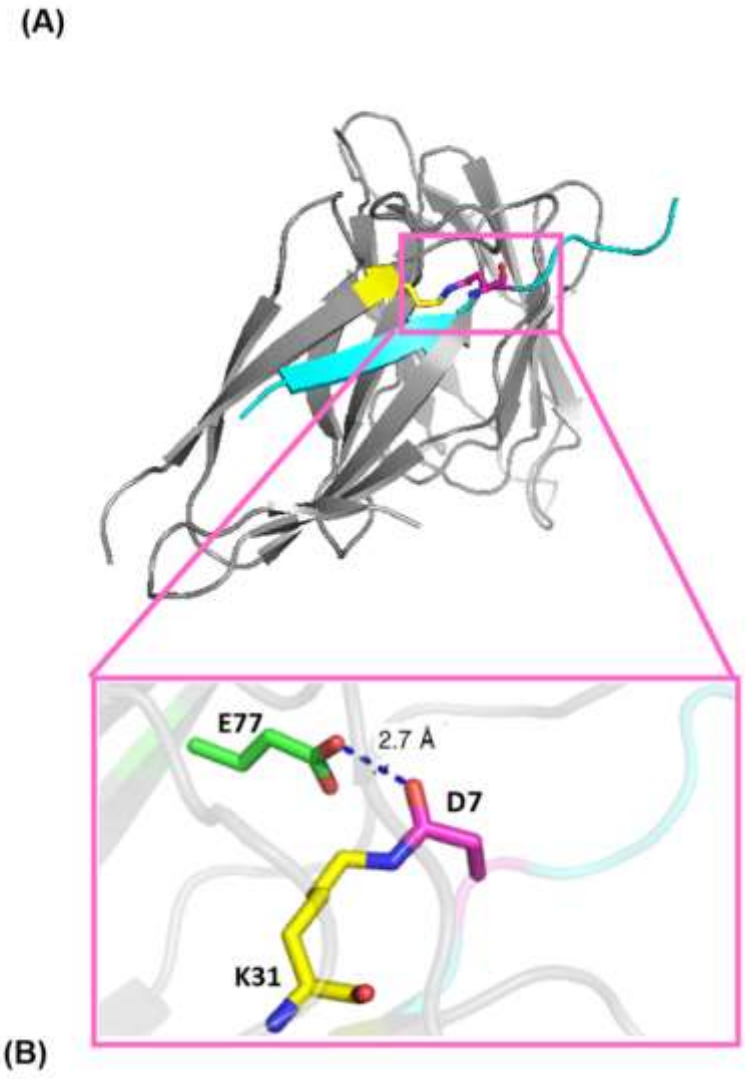


Figure 1.3. How the iso-peptide bond is formed between the residues. (A) *Cartoon representation of SpyTag-SpyCatcher. SpyCatcher is shown in grey and the SpyTag in cyan.* ***(B)****The bond forms between the Lysine 31 (K31) on SpyCatcher (shown in yellow as sticks) and the Asp 7 (D7) on SpyTag (shown in magenta as sticks) with the neighbouring residues Glu77 (E77) (shown in green as sticks) which acts as proton shuttle to help facilitate the formation of isopeptide bond (Reddington & Howarth, 2015a).*

The crystal structure of the SpyTag-SpyCatcher complex reveals that SpyTag forms a β-strand that complements one of the β-sheets in SpyCatcher (Hatlem et al., 2019; Li et al., 2014). This integration of SpyTag into SpyCatcher's structure positions the reactive residues in close proximity, enabling efficient isopeptide bond formation.

The formation of this bond has been tested under a number of conditions: the bond can form at pH values between 5 to 8, and in a range of buffers including PBS, phosphate citrate, HEPES and Tris. Since there is no cysteine residues in either SpyTag or SpyCatcher, the use of reducing agents does not affect the reaction . Isopeptide bond formation is relatively insensitive to temperature and pH changes and occurs in the presence of non-ionic detergents. However, the reaction rate is higher at lower pH (5-6) compared with pH 7.9 (Zakeri et al., 2012).

1.1.5. Engineering SpyCatcher for enhanced reaction speed

Interestingly, only 10 out of the 13 amino acids of SpyTag are directly involved in the interaction with SpyCatcher (Li et al., 2014). This suggests that specific residues within SpyTag are crucial for recognition and binding, while others may be less critical. Understanding the structure of SpyCatcher and its interaction with SpyTag provides valuable information for optimising and expanding the applications of this system. This knowledge has facilitated the development of faster reacting variants through rational design and researchers have engineered SpyTag-SpyCatcher (Hartzell et al., 2021; Keeble et al., 2017a). The initial SpyTag-SpyCatcher pair reacted with a half-life of just over 1 minute at 10 μM concentration (Reddington & Howarth, 2015a). Rational design based on structural insights combined with identified beneficial mutations, led to the development of SpyTag002-SpyCatcher002 (Keeble et al., 2017; Khairil Anuar et al., 2019), which reacts 12 times faster than the original SpyTag-SpyCatcher pair (Keeble et al., 2017a). For further optimisation of SpyTag002-SpyCatcher002, scientists introduced mutations to SpyCatcher002 to stabilise a key loop involved in SpyTag binding (Keeble et al., 2017a). In SpyTag002, they added positive charges at the N-terminus to enhance interactions with a negatively charged patch on SpyCatcher002 (Keeble et al., 2017a).

SpyTag002 has the sequence, PTIVMVDAYKRYK, which differs from the original SpyTag by the substitution of the first two residues (AH to PT) and the replacement of the residues 11 & 12 from PT to RY. These modifications resulted in the SpyTag003-SpyCatcher003 pair (Figure 1.4) with a reaction rate that is 400-fold faster than the original SpyTag-SpyCatcher pair (Keeble et al., 2019). Even at low nanomolar concentrations, SpyTag003-SpyCatcher003 demonstrated a rapid reaction - within minutes (Ahmadi et al., 2020). The outcome of this new

engineering revealed that the efforts have significantly enhanced the SpyTag-SpyCatcher system, pushing its reaction rate closer to the diffusion limit for protein-protein interactions (Keeble et al., 2019). This near "infinite affinity" unlocks possibilities for various applications, including rapid protein labelling in vivo, intracellular protein engineering, and sensitive protein detection techniques (Keeble et al., 2019)

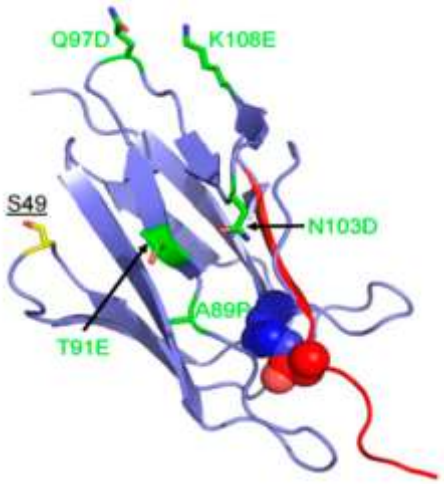


Figure 1.4. SpyCatcher002 to 003 mutations. Locations of mutations in SpyCatcher003 are highlighted. *Mutations introduced in SpyCatcher002 to create SpyCatcher003 are displayed in green in stick format, while SpyTag is highlighted in red. The isopeptide-bond is shown in space-filling format. The structure is based on PDB entries 2X5P and 4MLI (Keeble et al., 2019).*

1.1.6. Alternatives to SpyTag-SpyCatcher systems

A parallel system called SnoopTag-SnoopCatcher, derived from a pilin protein of Streptococcus pneumoniae, has been developed to enable the assembly of multi-component protein fusions. Tripartite applications incorporating these systems have also been created, using proteins SpyLigase or SnoopLigase to join two peptides via an additional catalytically active protein unit (Hatlem et al., 2019). Additionally, SpyLigase and SnoopLigase have been engineered to catalyse the formation of isopeptide bonds between two peptide tags, allowing for the creation of protein fusions with minimal molecular scars (Hatlem et al., 2019; Fierer et al., 2014). The SnoopTag-SnoopCatcher system is a protein ligation technology orthogonal to the SpyTag-SpyCatcher system, meaning that SnoopTag will only react with SnoopCatcher and not with SpyCatcher. The two systems can be used simultaneously to create more complex protein architectures (Hatlem et al., 2019). This system is based on the pilus adhesin RrgA of Streptococcus pneumoniae (Hartzell et al., 2021 ;Lin et al., 2020a). Similar to the SpyTag-SpyCatcher system, SnoopTag and SnoopCatcher spontaneously form an isopeptide bond upon mixing (Hatlem et al., 2019). SnoopCatcher is the protein component derived from the D4 domain of RrgA. It contains the reactive asparagine (N854) involved in isopeptide bond formation, as well as the catalytic glutamate (E803) that facilitates the reaction. SnoopTag is

a 12-residue peptide representing the segment containing the reactive lysine (K742) that forms the isopeptide bond with SnoopCatcher (Hatlem et al., 2019; Lin et al., 2020b).

1.1.7. Key Differences between SnoopTag/SnoopCatcher and SpyTag-SpyCatcher

In SpyTag-SpyCatcher, the reactive lysine is on SpyCatcher, and the reactive aspartate is on SpyTag. In SnoopTag-SnoopCatcher, the reactive lysine is on SnoopTag, and the reactive asparagine is on SnoopCatcher (Hatlem et al., 2019). SpyTag-SpyCatcher is derived from the fibronectin-binding protein FbaB of S. pyogenes, while SnoopTag-SnoopCatcher comes from the pilus adhesin RrgA of S. pneumoniae (Lin et al., 2020b).

A key distinction between the SpyTag-SpyCatcher and SnoopTag-SnoopCatcher systems is their compatibility and orthogonality. The SnoopTag-SnoopCatcher system is designed to operate alongside SpyTag-SpyCatcher without cross-reactivity, enabling simultaneous and independent interactions within a single experimental setup (Hatlem et al., 2019). This orthogonality offers researchers greater flexibility to design complex molecular assemblies and multifunctional protein constructs. Moreover, SnoopTag-SnoopCatcher has been employed to expand and integrate functions within the Spy toolkit. Its introduction allows for more sophisticated protein engineering by enabling the creation of multi-component systems where distinct tags and catchers can attach specific protein domains or functional groups in a highly selective manner (Keeble & Howarth, 2020a). This capability is especially valuable for constructing intricate protein networks and for applications such as vaccine development, where precise control over protein assembly is essential (Hatlem et al., 2019). The potential, in conjunction with SpyTag-SpyCatcher, for creating "polyproteams" refers to modular polyproteins created by using both systems simultaneously to attach multiple proteins together. One example is the construction of a modular vaccine using Hbp, a bacterial protein, as a scaffold. The researchers created a tripartite binding complex by fusing SpyTag and SnoopTag to different antigens (PspAα and SP1690) and then attaching them to Hbp-SpyCatcher and Hbp-SnoopCatcher, respectively. This demonstrates the possibility of building complex protein assemblies using both systems for targeted applications (Hatlem et al., 2019).

While the SnoopTag-SnoopCatcher system has broadened the scope of the original SpyTag-SpyCatcher technology, the addition of SnoopLigase - a catalyst enabling covalent bonding between SnoopTag and DogTag peptides - further overcomes certain limitations of SpyTag-SpyCatcher, particularly regarding the size of the Catcher protein partner. This advancement supports more efficient and flexible bioconjugation techniques, enhancing existing methods and unlocking new possibilities for precise peptide-peptide conjugation and purification in various biological applications (Hatlem et al., 2019). SnoopLigase catalyses isopeptide bond formation between a modified SnoopTag called SnoopTagJr and another peptide called

DogTag. This allows for the creation of protein fusions with a smaller molecular scar than using a complete SnoopCatcher protein (Hatlem et al., 2019; Karimi Baba Ahmadi et al., 2020; Andersson et al., 2019).

The development of SnoopTag-SnoopCatcher and SnoopLigase demonstrates the expanding toolbox of peptide-protein interaction systems available for protein engineering and bioconjugation. They offer a modular approach for building complex protein architectures with high specificity and efficiency.

1.2. Application of SpyTag-SpyCatcher technology

1.2.1. Protein engineering and modification

Protein cyclisation and stabilisation:

By fusing SpyTag and SpyCatcher to the N- and C-termini of a protein, respectively, researchers can induce the formation of a cyclic structure called a "SpyRing" (Reddington & Howarth, 2015; Song et al., 2022). This cyclisation can significantly enhance the protein's stability and resilience to denaturation. Studies have shown that SpyRing cyclisation can confer resistance to boiling on mesophilic enzymes (Schoene et al., 2014; Song et al., 2022) (Figure 1.5A).

Construction of protein conjugates:

The SpyTag-SpyCatcher system enables the site-specific and irreversible conjugation of proteins to various molecules and surfaces (Li et al., 2014).This has been used to create protein-protein conjugates (Khairil Anuar et al., 2019), protein-dye conjugates (Dovala et al., 2016; Reddington & Howarth, 2015b), protein-nanoparticle conjugates and bacterial outer membrane vesicles (Hatlem et al., 2019) (Figure 1.5B&C).

Generation of bioactive hydrogels:

SpyCatcher can be incorporated into hydrogels, allowing for post-hydrogelation decoration with SpyTag-fused proteins (Hatlem et al., 2019; Li et al., 2014). This approach enables the creation of hydrogels with specific functionalities, such as mimicking the extracellular matrix or presenting bioactive molecules (Li et al., 2014).

1.2.2. Protein characterisation and analysis

Protein expression and solubility analysis:

SpyTag-SpyCatcher can be used for rapid analysis of protein expression levels and solubility (Dovala et al., 2016). By attaching a fluorescent label to SpyCatcher, researchers can easily follow tagged proteins via SDS-PAGE. This technique allows for rapid screening of protein constructs and optimisation of expression conditions (Dovala et al., 2016).

Protein purification:

SpyTag can also be used as an affinity tag for protein purification. A technique called Spy&Go utilises a modified SpyCatcher, termed SpyDock, to create a reversible interaction with SpyTag-fused proteins (Hatlem et al., 2019; Khairil Anuar et al., 2019). This method allows for the efficient capture and release of SpyTag-proteins from cell lysates, offering an alternative to traditional His-tag purification (Khairil Anuar et al., 2019).

Western Blotting:

Fluorescently-labeled SpyCatcher can be used as a sensitive and specific detection reagent in western blot analysis. SpyCatcher's ability to react with denatured protein on nitrocellulose membranes and form a covalent bond with SpyTag makes it a highly specific alternative to antibodies (Dovala et al., 2016).

Analysis of protein-RNA interactions:

The SpyTag-SpyCatcher system has also been employed in a technique called SpyCLIP to characterise protein-RNA interactions (Tian et al., 2021). In this method, a SpyTagged RNA-binding protein is allowed to interact with RNA and is then covalently captured on beads functionalised with SpyCatcher. This approach offers advantages over traditional CLIP methods by reducing non-specific interactions and eliminating the need for gel purification steps (Hatlem et al., 2019; Tian et al., 2021) (Figure 1.5D).

1.2.3. Cellular and in vivo applications

Live cell imaging:

SpyTag-SpyCatcher has been employed to label and visualise proteins in living cells (Pessino et al., 2017). By fusing SpyTag to a target protein and introducing fluorescently labelled SpyCatcher, researchers can track protein localisation and dynamics in real-time (Keeble & Howarth, 2020a) (Figure 1.5E).

Investigating bacterial virulence factors:

The SpyTag-SpyCatcher system has been used to study the surface topology and localisation of bacterial virulence factors, particularly autotransporter proteins (Keeble et al., 2019). By fusing SpyTag to these proteins, researchers can use SpyCatcher-fused reporters to assess surface exposure and probe the mechanisms of protein secretion(Hatlem et al., 2019).

1.2.4. Vaccine development

Antigen delivery and display: The SpyTag-SpyCatcher system has emerged as a powerful tool for vaccine development, particularly in the context of virus-like particles (VLPs) and outer membrane vesicles (OMVs) (Schoene et al., 2014b; Tian et al., 2021). By fusing SpyTag or SpyCatcher to VLPs or OMVs, researchers can efficiently decorate these particles with SpyTag- or SpyCatcher-fused antigens (Hatlem et al., 2019; Schoene et al., 2014c). This modular approach allows for the creation of multivalent vaccines that can elicit strong immune responses (Hatlem et al., 2019; Tian et al., 2021) (Figure 1.5F).

1.2.5. Other applications

Protein immobilisation:

SpyTag-SpyCatcher can immobilise enzymes on various surfaces (Khairil Anuar et al., 2019; Williams et al., 2018).

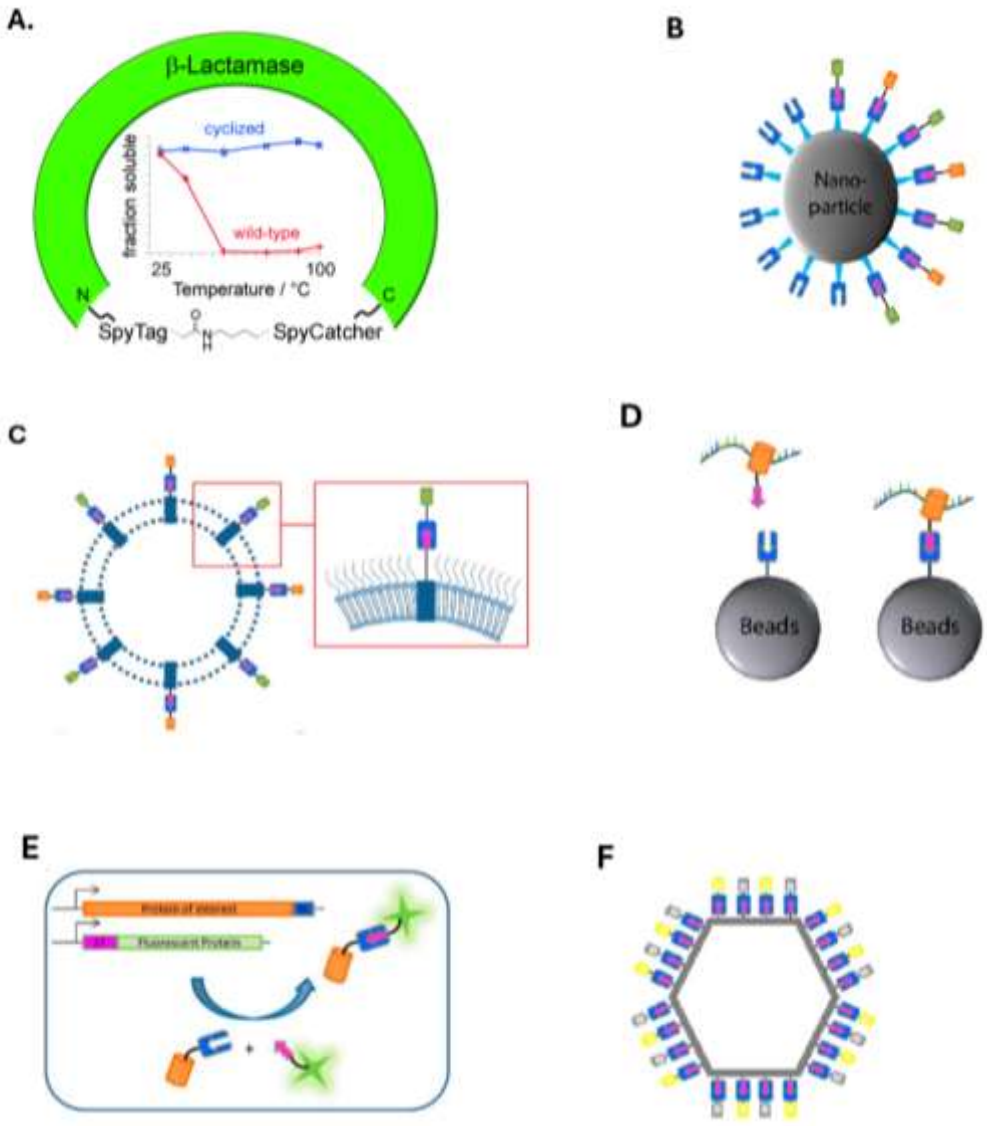


Figure 1.5. Applications of the SpyCatcher-SpyTag system. ***(A)*** *SpyRing: Fusion of SpyCatcher and SpyTag to the terminal ends of a protein of interest induces protein cyclisation, enhancing its resilience to denaturation.* ***(B)*** *Conjugating target proteins to nanoparticles.* ***(C)*** *Binding to outer membrane vesicles,* ***(D)*** *SpyCLIP: A method where a SpyTagged RNA-binding protein binds RNA and is covalently linked to SpyCatcher-coated beads, useful for pull-down assays.* ***(E)*** *Fluorescent protein labelling: SpyTag can be fused to fluorescent proteins for applications such as microscopy (SpyTag represented as ST).* ***(F)*** *Integration with virus-like particles (Hatlem et al., 2019).*

1.2.6. Advantages of SpyTag-SpyCatcher

There are distinct advantages of the SpyTag-SpyCatcher system over other protein tagging techniques. As described previously, the reaction between SpyTag and SpyCatcher is rapid and efficient, with a half-life of just over one minute at 10 mM (Hatlem et al., 2019). Unlike

most other peptide tags that bind to their protein partners reversibly, the SpyTag-SpyCatcher interaction results in an isopeptide bond between the peptide tag and its protein partner (Reddington & Howarth, 2015 ;Li et al., 2014). This covalent linkage makes the complex extremely stable, enabling its use under harsh conditions, such as boiling in SDS (Zakeri et al., 2012;Karimi Baba Ahmadi et al., 2020). This stability is crucial for applications that require resistance to force, such as single-molecule force measurements, or harsh conditions, like protein purification (Karimi Baba Ahmadi et al., 2020). SpyTag is small and can be placed at various locations within a protein without disturbing its function. The small size of the SpyTag (13 amino acids) allows it to be inserted at different positions in a protein, including the N-terminus, C-terminus, and even internal loops, without significantly affecting protein folding or function (Cooley et al., 2014; Hatlem et al., 2019; Karimi Baba Ahmadi et al., 2020). This versatility in tag placement makes the SpyTag-SpyCatcher system more adaptable when compared with other covalent protein tagging methods, such as split inteins or sortases, that require specific terminal placement and can disrupt protein function (Hatlem et al., 2019; Zakeri et al., 2012). Overall, the unique properties of SpyTag and SpyCatcher, along with the development of related tools and improvements to the original system, make it a powerful and versatile platform for protein engineering, cellular studies, and various other biotechnological applications and its applications continue to expand (Keeble & Howarth, 2020;Reddington & Howarth, 2015).

1.2.7. Limitations of SpyTag-SpyCatcher

Designing proteins for compatibility with the SpyTag-SpyCatcher system poses several critical challenges that researchers must address to ensure optimal functionality and efficiency. A primary challenge is preserving the stability and function of the target protein when fused with SpyTag or SpyCatcher sequences, as these additions may disrupt the protein's native structure and performance. This requires careful design strategies to minimise any potential negative effects (Keeble & Howarth, 2020). The insertion site for SpyTag or SpyCatcher is also crucial; improper placement can lead to misfolding or limited accessibility of the tag, which in turn impacts the binding efficiency between SpyTag and SpyCatcher (Hatlem et al., 2019). To counter this, researchers perform detailed structural analyses and experimental tests to find ideal insertion points that maintain the protein's activity (Keeble & Howarth, 2020). Another consideration would be the potential immunogenicity of SpyTag and SpyCatcher sequences when used in vivo. To address this, protein engineers often use in silico analyses to predict and adjust immunogenic epitopes, employing bioinformatics tools to identify and redesign regions likely to trigger immune responses without affecting the system's function (Hatlem et

al., 2019). Overcoming these challenges requires a multidisciplinary approach that combines protein engineering, structural biology, and computational modelling, allowing successful integration of SpyTag-SpyCatcher technology into diverse applications (Keeble & Howarth, 2020a).

1.3. Engineering novel proteins via saturation mutagenesis

Engineering novel proteins through saturation mutagenesis is a pivotal approach in protein engineering, allowing for the exploration of sequence space to enhance or alter protein functions. Thus the approach involves modifying specific residues which have key role(s) in protein function. When structural information about the protein is available, these key residues can be identified with reasonable confidence. At this point, the focus shifts to introducing as many variations as possible at these critical positions. This process generates a DNA library that encodes a diverse collection of proteins (protein library), ideally encompassing every possible amino acid combination at the targeted sites. The library is then screened, with the aim that one or more variants will demonstrate specificity for the desired new substrate or ligand. Saturation mutagenesis is a widely used technique in protein engineering and directed evolution for exploring protein sequence space and creating libraries with diverse protein variants (Kille et al., 2013a). However, the creation of diversity at the codon level introduces redundancy, where multiple codons can encode the same amino acid, leading to an increase in library size and screening effort (Kille et al., 2013a). This redundancy, referred to as degeneracy, presents a significant challenge as it can severely impact the efficiency of identifying the desired protein variants (Ferreira Amaral et al., 2017).

In the past, mutations were introduced one at a time using various methods, making it impractical to create individual genes with each specific combination of changes. To address this, the concept of saturation mutagenesis was introduced years ago. In this approach, each key codon is replaced with a degenerate codon (usually NNN, NNK, or NNS) to allow encoding of all 20 amino acids at each target position. As a result, each gene within the mixture is unique, but together, the library encompasses all possible combinations of codons, and thus, all potential amino acid variations at the critical sites (Kille et al., 2013a). Moreover, degenerate codons can randomly introduce termination codons, resulting in truncated, nonfunctional proteins that are prone to aggregation and precipitation (Ferreira Amaral et al., 2017). One challenge in subsequent screening arises from the degeneracy of the genetic code. With NNN codons, each nucleotide triplet is a unique combination of the four bases—adenine (A), guanine (G), cytosine (C), and thymine (T)—resulting in 64 possible codon combinations. Of these, 61 codons code for 20 amino acids, while 3 serve as termination codons. This means

that some amino acids are encoded by multiple codons: two codons (Cys, Asp, Glu, Phe, His, Lys, Asn, Gln, Tyr), three codons (Ile), four codons (Ala, Gly, Pro, Thr, Val), or even six codons (Leu, Ser, Arg). Only tryptophan (Trp) and methionine (Met) are specified by a single codon. This codon redundancy can reduce the diversity of a gene library, introducing bias and impacting library size. Higher codon degeneracy increases the number of variants, which affects the choice of screening methods in later steps, as each method has its own capacity for handling variant numbers. Furthermore, codon bias influences screening outcomes, especially in ligand-based methods that require equal protein concentrations for accuracy. The use of degenerate codons can also result in a vast number of DNA sequences, making it impractical to screen the entire library (Hughes et al., 2003a).

1.3.1. Reducing degeneracy in saturated libraries

Various strategies have been developed to address the limitations associated with using degenerate codons, including the use of NNK and NNS saturation codons (where K = T/G and S = G/C). These approaches minimise redundancy by restricting the third codon position, reducing the possible combinations from 64 to 32 codons for encoding 20 amino acids and a single termination codon. This reduction not only limits redundancy but also simplifies library construction and screening by decreasing the number of necessary variants. Another key innovation was the 22c-Trick, which employs just 22 codons for 20 amino acids, significantly reducing redundancy and the overall library size (Kille et al., 2013a). The 22c-Trick utilises subsets of oligonucleotides with controlled degeneracy at specific codons. It requires three primers for generating saturated codons during PCR: the first contains NDT (A/C/G/T; A/G/T; T), the second VHG (A/C/G; A/C/T; G), and the third TGG. When these primers are used together, valine and leucine each appear twice, giving this method its name (22c refers to 22 codons). For optimal codon saturation, it is essential to adjust the primers' annealing temperature carefully, as insufficient optimisation can compromise the diversity of the resulting library (Kille et al., 2013a). Figure 1.6 illustrates the effects of these methods compared with complete nondegenerate saturation, which uses exactly 20 codons for the 20 amino acids. In both the construction and screening phases of a DNA library, diversity — defined as the percentage of unique variants — is crucial. However, when mutating a single codon using degenerate methods (like NNN or NNK/NNS), diversity immediately drops to around 60-70%. With each additional targeted codon, diversity declines even more sharply, reaching only about 10% when 12 codons are mutated. The 22c-Trick, despite being nearly nondegenerate, follows a similar trend, with diversity decreasing as the number of targeted codons increases, resulting in about 40% diversity for 12 targeted codons. In contrast, with truly nondegenerate

mutagenesis, diversity theoretically remains constant at 100%, regardless of the number of targeted codons (Figure 1.6A). Diversity directly impacts a second key library assessment parameter known as encoded bias. Protein libraries often exhibit bias due to the genetic code's structure, where some amino acids are encoded by multiple codons while others are represented by only one or two. When constructing gene libraries, equal representation of all amino acids is typically desired, which can only be achieved through nondegenerate saturation methods. Bias increases with the number of amino acids introduced, as shown in the theoretical ratios provided in the bias table (Figure 1.6B), where codons for more frequently encoded amino acids (like Ser with six codons) contrast with those for less frequent ones (like Trp with one codon). Figure 1.6B illustrates that this bias at its most extreme when using NNN/NNK/NNS methods. Nondegenerate methods, however, eliminate this bias by assigning a single codon to each amino acid during saturation, as seen in approaches like MAX, ProxiMAX, or Slonomics® approaches. Another key issue with degenerate methods is the risk of introducing termination codons, which can lead to premature truncation and the production of nonfunctional proteins that may aggregate within cells (Ferreira Amaral et al., 2017) (Figure 1.6C). In conclusion, nondegenerate saturation mutagenesis enables the inclusion of all 20 amino acids at theoretically equal ratios. These methods also avoid termination codons, preventing the generation of truncated proteins. By eliminating degenerate codons and using only 20 codons instead of 64, library size is minimised, and diversity—represented by unique DNA sequences—is maximised to 100%. These features of nondegenerate methods support the creation of high-quality, diverse libraries, which enhances the efficiency of screening processes (Tang et al., 2012a).

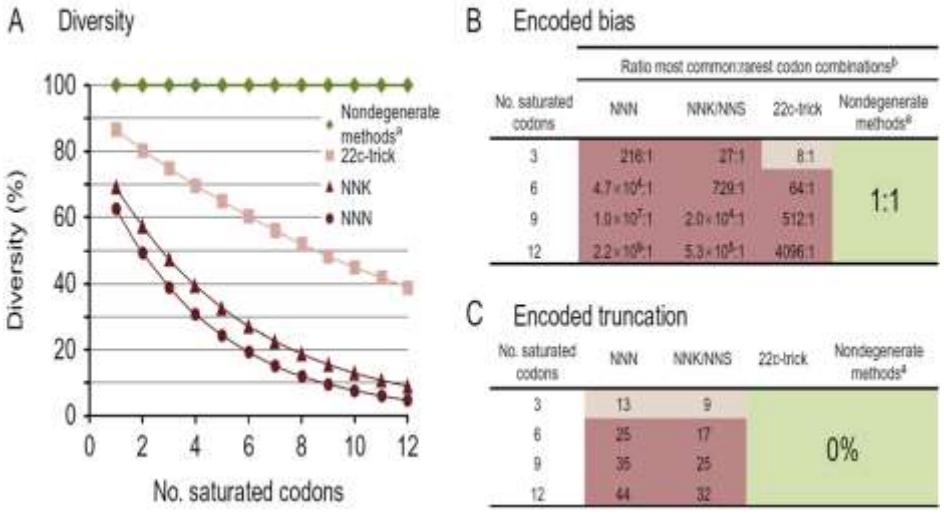


Figure 1.6. Comparison of the effectiveness of common saturation mutagenesis techniques. *Green shading indicates ideal performance, light pink indicates acceptable performance, and dark pink shows unacceptable performance. Nondegenerate methods can be created using various techniques. (A) Diversity was calculated using the formula d=1/(N∑kpk2)d = 1/(N\sum_{k} p_{k}^2)d=1/(N∑kpk2)*

(Makowski & Soares, 2003), aligning with findings for a 12-mer peptide saturated with NNN codons (Krumpe, Schumacher, McMahon, Makowski, & Mori, 2007). (B) Ratios represent the theoretical relative concentrations of genes containing any of the most common codons (e.g., Leu/Arg/Ser for NNN/NNK; or Leu/Val for the 22c-Trick) versus those containing combinations of the rarest codons (e.g., Met/Trp for NNN; Cys/Asp/Glu/Phe/His/Ile/Lys/Met/Asn/Gln/Trp/Tyr for NNK; or 18 codons excluding Leu/Val for the 22c-Trick). (C) Truncation is calculated as the percentage of sequences that include one or more stop codons within the saturated region (Ashraf et al., 2013).

1.3.2. Nondegenerate saturation mutagenesis

Several approaches have been developed to mitigate degeneracy and improve the quality of saturated libraries. Nondegenerate saturation mutagenesis offers the advantage of including all twenty codons in theoretically equal proportions. Importantly, this method excludes termination codons, mitigating the risk of generating truncated proteins. Additionally, the library size is minimised by replacing degenerate codons with a precise set of 20 codons instead of 64. This ensures maximum diversity, achieving 100% unique DNA sequences. These features make nondegenerate approaches ideal for creating diverse, high-quality libraries with a reduced yet entirely unique variant pool, ultimately enhancing the efficiency of screening methods (Tang et al., 2012b).

Regardless of whether saturation mutagenesis is degenerate, near-nondegenerate or fully nondegenerate, all approaches are based on the same fundamental techniques (Kille et al., 2013). The fundamental techniques encompass cassette mutagenesis, simple primer extension mutagenesis, and overlap extension mutagenesis. Nondegenerate techniques include TRIM technology, MAX randomisation, ProxiMAX randomisation, Slonomics®/SlonoMax™, DC Analyzer, MDC Analyzer, and the 22c-Trick (classified as near-nondegenerate saturation) (Ferreira Amaral et al., 2017).

Of the techniques mentioned, MAX randomisation and overlap PCR will be discussed in detail, as these two methods were employed in this study.

1.3.2.1. MAX Randomisation

MAX randomisation is a technique used in protein engineering to create highly diverse gene libraries with minimal redundancy and bias and was one of the first nondegenerate saturation mutagenesis methods that required no specialised equipment or reagents and could be easily implemented in any lab (Hughes et al., 2003a). In other words, MAX randomisation was created to remove the degeneracy of the genetic code by employing a one-to-one codon-to-amino acid ratio, thereby eliminating representational bias in engineered libraries without requiring specialised chemistry. MAX randomisation utilises a technique called "selectional hybridisation" to generate the desired randomisation cassettes, the selection oligonucleotides hybridise to a complementary DNA template while introducing saturation mutagenesis at specific codons (Figure 1.7). Once the selection oligonucleotides have hybridised, they are

ligated to close the gaps between them, followed by asymmetric PCR amplification to create a randomised cassette.

The selection oligos are short, typically nine nucleotides long, with six nucleotides conserved to match the template and three corresponding to a single MAX codon, allowing a mix of twenty oligos to saturate a specific position. Additionally, two terminal oligos are used as primer sites for the final amplification step in the MAX randomisation process. (Hughes et al., 2003a). However, the utility of MAX randomisation extends beyond simply encoding all 20 amino acids. It also offers the flexibility to restrict the range of amino acids encoded in a gene library (Ferreira Amaral et al., 2017). Instead of using all 20 selection oligonucleotides, researchers can choose a subset of selection oligonucleotides that correspond to the desired amino acids. This allows for the creation of gene libraries that are focused on specific types of amino acids. This targeted approach can be particularly useful when there is prior knowledge about the functional role of specific amino acids or when exploring specific biochemical properties. This would allow for a more focused exploration of sequence space and could lead to the identification of novel protein variants with enhanced binding affinity. Hughes et al. (2003a) provide a specific example of this concept. They used MAX randomisation to create a library restricted to 8 amino acids; Asp, Glu, His, Lys, Asn, Gln, Arg, and Trp. This subset was chosen to represent amino acids capable of participating in hydrogen bonding interactions, along with the unique properties of tryptophan, which can participate in hydrophobic, hydrogen-bonding, and π-π interactions. The results demonstrated that MAX randomisation successfully allowed for the near-exclusive cloning of the desired subset of amino acids. This ability to precisely control the amino acid composition of a gene library offers a powerful tool for protein engineering and expands the possibilities for exploring and optimising protein function (Hughes et al., 2003a).

Figure 1.7 demonstrates the approach, where three selected positions are saturated with three distinct MAX oligonucleotide pools. These MAX oligonucleotides bind to complementary regions on the constant template sequence, which is flanked by conserved regions essential for PCR priming and amplification. Through specific annealing, the selected oligonucleotides and flanking regions (End 1 and End 2) are ligated in a precise configuration, creating a DNA strand with the desired combination of MAX codons at each saturated position. Following ligation, asymmetric PCR amplification of the strand with MAX codons is performed using primers that bind to the conserved flanking regions. The overlap length between the flanking oligonucleotides (End 1 and End 2) and the traditionally randomised template strand is crucial, needing to be short enough to avoid amplifying the template strand itself. Experimentally, a 6

bp overlap is effective, whereas an overlap of 9 bp can still result in amplification of the NNN-codon strand.

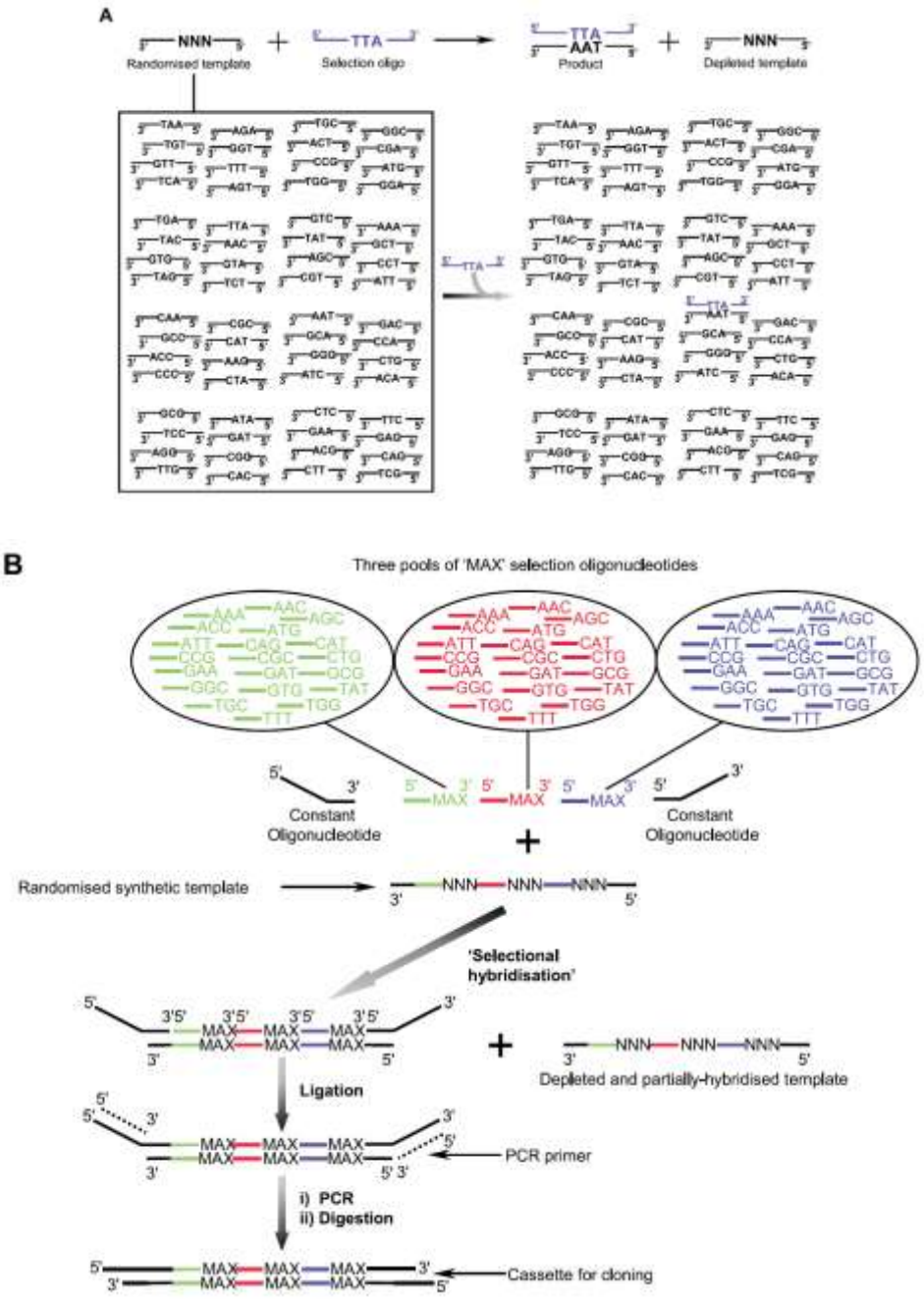


Figure 1.7. Schematic representation of MAX randomisation technique. *A synthetic template oligonucleotide is prepared with specific codons randomised as NNN and invariant bases (shown as continuous lines) that match the parent gene sequence. For each randomised codon, 20 synthetic selection oligonucleotides are used, each containing the necessary complementary invariant region and a codon (the MAX codon) optimised for expressing a specific amino acid. These selection oligonucleotides hybridise with the template oligonucleotide, with base pairing ensuring that each selection oligonucleotide binds to its complementary sequence in the template.* ***(A)*** *An example of hybridisation: a selection oligonucleotide with the MAX codon 5'-TTA-3' pairs with a template oligonucleotide containing 5'-TAA-3' at the randomised codon position.* ***(B)*** *Selection-based hybridisation to create a synthetic cassette for gene randomisation, where the template oligonucleotide contains three randomised codons. The invariant regions on the template are color-coded to align with*

their corresponding selection oligonucleotides. Two additional unique constant oligonucleotides are also included to provide primer-binding and restriction sites at the cassette ends. Broken lines indicate the primer-binding sites, positioned to ensure that only the selection strand is amplified during PCR. The resulting DNA cassette is then digested with restriction enzymes, dephosphorylated (to prevent concatemer formation), and cloned (Hughes et al., 2003a).

The randomisation cassette produced in this process can subsequently be used to create gene libraries via subcloning or overlap PCR or serve as a double-stranded primer for site-directed mutagenesis. Although MAX randomisation enables saturation mutagenesis across multiple codons, it is limited to targeting only two contiguous codons simultaneously, which can be a drawback when applying this technique, for codons far from each other, in the generation of libraries. It cannot randomise more than two contiguous codons due to the conserved sections in the selection oligonucleotides that are necessary for addressing each MAX codon to the appropriate site (Hughes et al., 2003a). ProxiMAX randomisation builds upon the principles of MAX randomisation but allows for the saturation of multiple contiguous codons, overcoming this limitation (Ashraf et al., 2013).

MAX randomisation has been successfully applied in various protein engineering studies, including engineering Zinc Finger Proteins, which used MAX randomisation to create a series of 60 overlapping gene libraries encoding zinc finger proteins (Hughes et al., 2005). These libraries were randomised at three key DNA-contacting residues and screened against five different target DNA sequences, leading to the identification of novel zinc finger proteins with high affinity for their targets (Hughes et al., 2005). A high-quality library is characterised by high diversity and completeness, encompassing a wide range of protein variants. MAX randomisation contributes to improved library quality by ensuring that all 20 amino acids are present and evenly represented (Hughes et al., 2005; Tang et al., 2012). This maximises the chances of identifying rare and valuable variants that might be missed in a library with significant amino acid bias. MAX randomisation plays a crucial role in enabling the use of positional fixing for library deconvolution. This technique, where one amino acid position is fixed while others are randomised, relies on the accurate representation of all amino acids at the randomised positions (Hughes et al., 2005).

MAX randomisation represents a significant advancement in gene library construction, addressing the limitations of codon redundancy and amino acid bias inherent in traditional methods. By ensuring uniform representation of encoded proteins, this technique leads to higher-quality libraries, reduced screening effort, and more accurate identification of desired variants. While complexity considerations exist, the benefits of MAX randomisation make it a valuable tool for protein engineering and directed evolution, particularly in applications requiring high-quality libraries and efficient screening strategies.

1.3.2.2. Overlap PCR

1.3.2.2.1. Historical context

Overlap extension PCR for mutagenesis was first introduced in 1988 by Higuchi, Krummell, and Saiki. Essentially, PCR fragments with complementary 3' regions—created from primers that incorporate mutations—can prime each other, facilitating the joining of these fragments. This technique is widely used in both degenerate and nondegenerate saturation mutagenesis. It is applied as originally described by Higuchi et al. (1988), where the overlaps include the mutations, or for linking cassettes where mutations are located internally. Overlap extension PCR is a versatile technique for manipulating DNA sequences in vitro. This method enables the creation of specific mutations, insertions, or deletions at any location within a DNA fragment, eliminating the need for traditional cloning methods and facilitating the study of protein-DNA interactions (Higuchi et al., 1988a).

1.3.2.2.2. Basic principles of overlap PCR

This technique involves amplifying the DNA sequences to be combined in such a way that the amplicons have overlapping sequences that can prime each other during a subsequent PCR reaction, leading to the fusion of the fragments (Hilgarth & Lanigan, 2020). Overlap extension PCR relies on the ability of DNA polymerase to extend overlapping DNA strands. The process starts with two primary PCR reactions, each using a different set of primers. One primer in each reaction contains the desired mutation and overlaps with the other primary PCR product. After the primary PCR, the two products are mixed, denatured, and allowed to reanneal. During reannealing, the overlapping regions of the two fragments hybridise, creating a template for DNA polymerase to extend and generate a full-length, double-stranded DNA fragment containing the mutation. This fragment can be further amplified using only the outermost primers from the initial PCR reactions (Higuchi et al., 1988a). The overlap extension PCR technique thus involves a three-step process in which the final step amplifies the full-length fusion product using primers designed to bind to the ends of the fused sequence (Figure 1.8). This step generates a sufficient amount of the desired product for downstream applications such as subcloning and sequencing (Hilgarth & Lanigan, 2020).

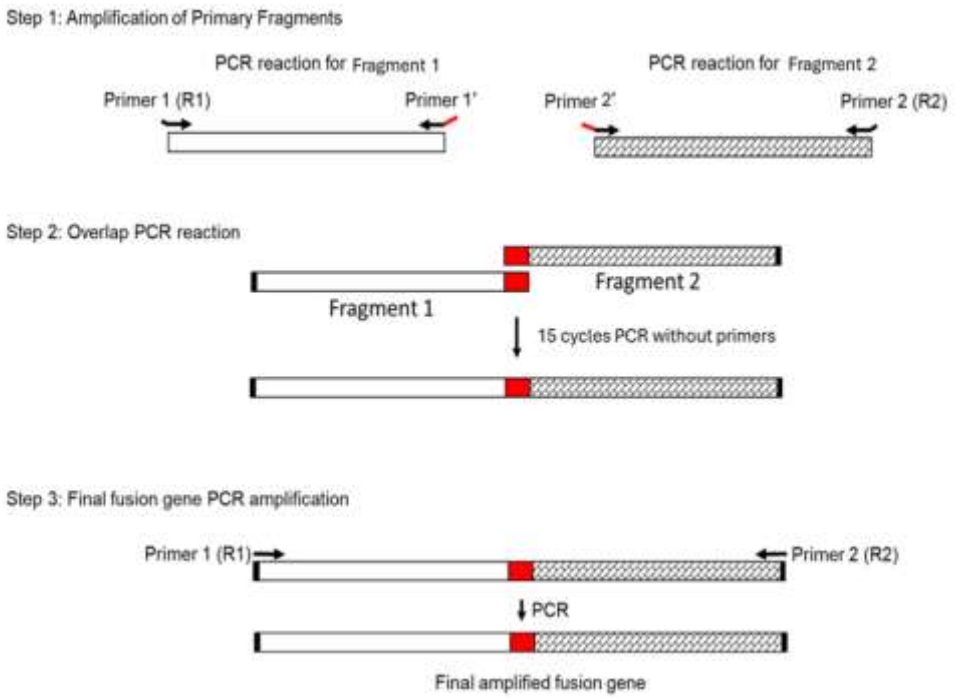


Figure 1.8. Schematic representation of overlap PCR. *The figure illustrates the three main steps in the overlap PCR process to generate a full-length product. Step 1: Amplification of Primary Fragments. PCR is performed separately for two fragments (Fragment 1 and Fragment 2) using specific primers. For Fragment 1, Primer 1 (R1) and Primer 1' are used, while for Fragment 2, Primer 2' and Primer 2 (R2) are used. The overlapping regions (shown in red) are introduced through these primers to facilitate subsequent fusion. Step 2: Overlap PCR Reaction. The two fragments with complementary overlapping regions are mixed and subjected to 15 cycles of PCR without additional primers. This step allows the overlapping regions to anneal, resulting in the fusion of Fragment 1 and Fragment 2. Step 3: Final Fusion Gene PCR Amplification. The fused fragments are amplified using Primer 1 (R1) and Primer 2 (R2) to generate the final amplified fusion gene. The product includes the complete sequence of both fragments joined together at the overlapping region (Hilgarth & Lanigan, 2020).*

One of the key advantages of overlap PCR is its efficiency, requiring minimal optimisation in all amplification steps. The protocol is straightforward, making it accessible to researchers of all skill levels (Hilgarth & Lanigan, 2020). However, overlap extension PCR has some limitations. For example, the accuracy of the DNA polymerase used in the reaction can influence the fidelity of the final product. Therefore, it is recommended to use high-fidelity polymerases with lower error rates, to minimise this issue (Higuchi et al., 1988a). In addition, proper primer design is essential for successful overlap extension PCR. The overlapping sequences should have sufficient length and melting temperature to ensure efficient annealing and extension.

1.4. Combinatorial libraries and screening

1.4.1. Historical context of combinatorial chemistry

Combinatorial chemistry encompasses synthetic methods that enable the preparation of extensive numbers of compounds, ranging from tens to millions, in a single process. These compound libraries can be created as mixtures, or sets of individual compounds. The field was conceived almost 40 years ago, initially focusing on peptide and oligonucleotide libraries (Lowe, 1995). The development of combinatorial libraries marked a paradigm shift in chemical synthesis and screening, enabling the rapid generation and evaluation of vast numbers of compounds. This revolution was rooted in the limitations of existing methods to synthesise and screen large numbers of peptides. Prior to the advent of combinatorial libraries, synthesising large numbers of peptides was a laborious and time-consuming process, where existing methods could not produce the millions of individual peptides required for comprehensive analyses. Moreover, these methods often struggled to generate sufficient quantities of unmodified free peptides necessary for solution-based assays (Houghten,1991). The foundation for combinatorial libraries was laid by Merrifield's groundbreaking work on solid-phase peptide synthesis in the 1960s (Houghten 1991; Ostresh et al., 1994).

This technique allowed peptides to be synthesised on solid supports, simplifying purification and paving the way for automation. Initially, solid-phase synthesis was a linear process, producing one peptide at a time. However, in the mid-1980s, parallel synthesis techniques like the "pin", "tea bag", and "spot" methods emerged, enabling the simultaneous synthesis of hundreds of individual peptides (Houghten et al., 1999; Ostresh et al., 1994). This advancement significantly reduced the time and cost of peptide synthesis and served as a stepping stone to the development of combinatorial libraries. The late 1980s and early 1990s witnessed the emergence of true combinatorial libraries, primarily peptide-based. A paper published in 1991 by Houghten et al. described the creation and use of synthetic peptide combinatorial libraries (SPCLs) (Houghten,1991; Houghten et al., 1999). These libraries were distinguished by their composition of mixtures of free peptides, making them directly compatible with various assay systems (Houghten et al., 1999). This approach circumvented the limitations of traditional methods and allowed researchers to explore a much larger chemical space. Notably, Houghten's work demonstrated the successful identification of an antigenic determinant recognised by a monoclonal antibody and the development of potent antimicrobial peptides using these libraries (Houghten ,1991). Merrifield's pioneering method established the foundation for future advancements in combinatorial chemistry techniques. Combinatorial chemistry evolved into a complex suite of methods aimed at synthesising,

purifying, analysing, and screening vast libraries of chemical compounds more efficiently and affordably than conventional approaches.

Initially advancing within the pharmaceutical sector to accelerate drug discovery, the field has since broadened its reach to numerous other areas in chemistry. Terms like parallel array synthesis and high-throughput chemistry were also frequently associated with this discipline (Terrett, 1998).

1.4.2. Methods for creating combinatorial libraries

1.4.2.1. Design and screening of libraries

Some factors need to be considered when designing a library, such as size, diversity, and screening method. A thorough understanding of assay parameters is crucial for successful screening of combinatorial libraries. Key parameters include signal-to-background ratio, variability, and sensitivity (Houghten et al., 1999).

The choice between screening on beads or in solution depends on the nature of the target and the library. If the biological target is soluble, screening libraries attached to beads is advantageous as it allows direct selection of beads carrying the compound with the highest affinity (Lowe, 1995), though care must be taken to choose a solid support that is suitable for the following assay.

Conversely, there are advantages to screening in solution, especially if the solid support might affect biological activity. However, solution-phase screening might require iterative deconvolution or an orthogonal combinatorial approach, which uses separate sublibraries that represent the same set of compounds in different arrangements (Houghten, 1991; Lowe, 1995). Whatever the means of screening, it is essential to have a means of identifying the active compound within a library. This process is called deconvolution. Common deconvolution strategies include iterative deconvolution and positional scanning.

1.4.2.2. Positional scanning versus iterative synthesis

Positional scanning involves complete “up front” synthesis. Deconvolution then involves screening a set of sub-libraries, with each sub-library having a different position defined by a single building block, while the remaining positions are occupied by a mixture of all building blocks used in the library (Houghten et al., 1999). Each sub-library represents the same collection of compounds but differs in the location of the defined position.

All sub-libraries are screened, and the most active mixture from each sub-library is selected. By combining the building blocks that define the most active mixtures in each position,

individual active compounds are identified and subsequently synthesised for testing (Houghten et al., 1999).

As an alternative, iterative deconvolution involves synthesising and testing increasingly smaller mixtures until a single active compound is identified. Iterative deconvolution begins with an initial set of mixtures, typically with one or two positions in the molecule defined by specific building blocks, while the remaining positions are occupied by a mixture of all building blocks (Houghten et al., 1999). These initial mixtures are screened in the relevant assay, and the most active mixture is selected. The next step involves synthesising a new set of mixtures, iterating on the identified active mixture. This new set of mixtures will have the previously defined position(s) retained, and additional position(s) will be defined with individual building blocks, creating a new set of mixtures for screening. This iterative process of synthesis and screening is repeated until all positions are defined, resulting in the identification of individual active compounds (Houghten et al., 1999). For example, a hexapeptide library could start with mixtures defined by the first two amino acids, and with each iteration, one additional amino acid position would be defined until a single hexapeptide sequence is identified (Lowe, 1995). The iterative method is relatively straightforward and conceptually easy to understand. It can be useful when limited knowledge is available about the target and a broader exploration of chemical space is desired (Houghten et al., 1999), while positional scanning is generally much faster than iterative deconvolution, as it requires fewer rounds of synthesis and allows the identification of key building blocks at each position simultaneously (Houghten et al., 1999).

An example of positionally fixed peptide library for a tetrapeptide with five amino acids building blocks (labelled A – E) is shown in Figure 1.9.

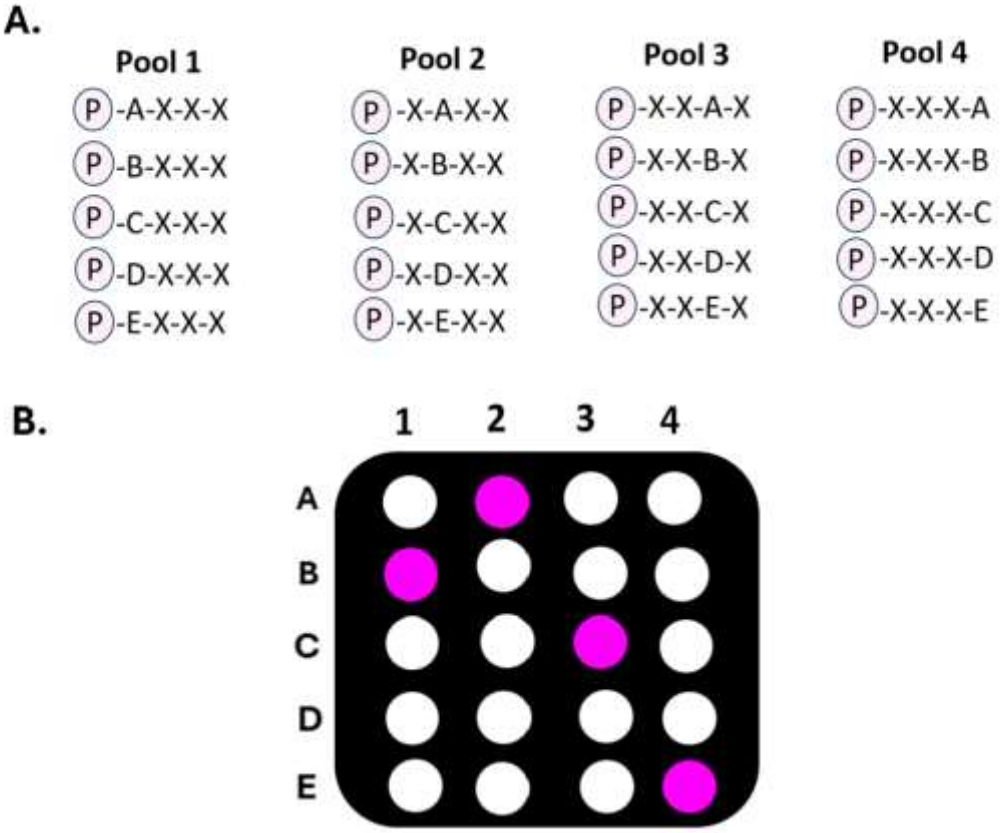


Figure 1.9. Schematic representation of positionally fixed for peptide library screening. (A)*In the positionally fixed libraries method, each pool has one position fixed for a specific building block (A, B, C, etc.), while other positions remain randomised (X).* ***(B)****This approach enables parallel screening of all combinations, allowing direct identification of active compounds based on screening results. Herein, the active compound would be B-A-C-E.*

1.4.3. Applications of combinatorial libraries

Combinatorial libraries have been used successfully to discover enzyme inhibitors, receptor agonists and antagonists, antimicrobial and antiviral agents, and ligands for antibodies and T-cell receptors (Houghten et al., 1999). Combinatorial chemistry plays a significant role in drug development, especially in identifying lead compounds that can be further optimised into drug candidates. Researchers have employed combinatorial libraries to study the interactions between antibodies and antigens, leading to a better understanding of immunological recognition and the development of potential vaccines (Houghten,1991). Combinatorial libraries have been instrumental in identifying novel enzyme inhibitors, particularly for targets involved in critical biological processes such as HIV protease and have been screened to discover potent and selective ligands for various receptors, including opioid receptors and endothelin receptors (Houghten et al., 1999).

1.4.4. Directed evolution in combinatorial protein libraries

The process of selecting proteins with enhanced or novel characteristics from combinatorial libraries is known as directed evolution (Mössner and Plückthun, 2001). This method mirrors natural evolution, where selection pressures are applied to promote advantageous traits in combinatorial libraries. In natural selection, favourable genetic mutations lead to beneficial phenotypic traits that are passed down to future generations through the genetic material of the selected parent organisms. Similarly, in directed evolution, mutations introduced in selected proteins are retained in subsequent iterations, driving the development of optimised protein variants. In combinatorial methods, genetic mutations within the coding sequence of a target protein are introduced experimentally. This approach results in the creation of a mutant gene library. The selection pressure applied to these mutant proteins is also determined by the user, allowing the screening conditions to favour proteins that express desired phenotypic traits. After selecting a protein from the library, it is necessary to identify the mutations responsible for the desired characteristics. Effective screening of peptides from combinatorial libraries requires establishing a link between the selected protein and its encoding nucleic acid. This genotype-phenotype linkage has been achieved through methods such as restricting each library's contents to specific proteins (Jamieson et al., 1996; Choo and Klug, 1994) and encoding the identity of DNA-binding protein target sites by varying the lengths of DNA target sites (Desjarlais and Berg, 1994). These strategies support the selection of proteins from conventional libraries. More commonly, however, physical linkages between genotype and phenotype are created using display technologies, which streamline the selection of proteins from combinatorial libraries.

Randomised gene libraries are generated by purposefully introducing mutations into the coding sequence of a target gene. These mutations, or the randomisation of the target gene, can be introduced using various techniques. Similar to display technologies, selecting the appropriate randomisation strategy is essential for the success of combinatorial experiments. One of the most promising uses of targeted randomisation has been in creating novel, sequence-specific DNA-binding proteins based on Cys2His2 zinc finger frameworks This approach suggests potential for designing artificial transcription factors and gene-silencing elements on demand.

1.4.5. Screening biological protein libraries

Biological protein libraries are typically screened using display technologies that establish a direct link between phenotype (the activity or function of the encoded protein) and genotype

(the DNA sequence encoding the protein). This linkage allows researchers to identify and isolate proteins with desired characteristics along with their corresponding genetic information. In contrast, chemists often lack this inherent genotype-phenotype connection. Biological protein libraries are generally screened by a process called biopanning, which is often used with display libraries. Examples of display libraries include phage display, bacterial display, ribosome display and RNA display (Li et al., 2019). Biopanning involves multiple rounds of screening that gradually reduce the number of proteins in the library. In each round, the complex of protein and its encoding sequence is incubated with an immobilised target molecule. Weakly binding proteins are removed by washing. The remaining strongly bound proteins are then eluted and used for the next round of biopanning. Eventually, only proteins with strong binding to the target are isolated. The identity of the active protein is then determined by sequencing the linked DNA or RNA sequence (Hughes et al., 2003a). Chemists typically do not have the advantage of directly linking phenotype to genotype like in biological systems. However, DNA-encoded libraries can be used. In this case, the DNA does not encode the product itself, but instead acts as a covalently linked identification tag. This provides a reference to identify the product. Alternative methods to biopanning, such as positional fixing, is more commonly used for deconvolution in combinatorial chemistry, but can also be applied in biological scenarios through a series of screens of overlapping sets of protein mixtures.

1.4.6. Identifying zinc finger proteins using positional fixing

As described above, positional fixing has been routinely used in combinatorial chemistry, but is rarely applied to biological protein libraries. In 2005, Hughes et al. applied positional fixing to create self-deconvoluting zinc finger libraries. This involved generating and screening a series of overlapping protein libraries in a way that allows the identity of active proteins to be determined directly from the initial library screens, eliminating the need for purification or sequence tags. In this study the researchers focused on saturating three key DNA-contacting residues in the second finger of a three-zinc finger protein called ZFH, which was fused to green fluorescent protein (GFP) to create His6-ZFH-GFP. They created 60 randomised gene libraries, fixing one of the amino acids at each of the three positions while randomising the other two, each library encoded 400 individual proteins. This setup allowed them to test 40,000 potential interactions in just 300 assays (60 libraries × 5 DNA targets).

These libraries were designed to ensure high clonal representation and minimal protein bias. Over-representation of the parental protein was avoided by introducing a frameshift mutation and a termination codon, ensuring that only the correct insertion of a randomised DNA

cassette would produce intact zinc finger/GFP fusions. Finally, to eliminate bias from the genetic code, MAX randomisation was used, allowing for non-degenerate encoding of amino acids at each randomised position. The libraries were screened against five different target DNA sequences using a plate-based protein-DNA interaction assay. Screening interactions were performed directly from crude bacterial lysates. The interaction data was then scaled according to the total GFP fluorescence, normalised, and plotted to identify the putative interacting proteins. The results of their encoded deconvolution method were validated using both in vivo and in vitro assays, by yeast one-hybrid assays to confirm the biological activity of the identified proteins in a cellular environment. They then constructed, expressed, and purified individual zinc finger genes and measured their dissociation constants to confirm DNA-binding activity in vitro.

The results demonstrated that encoded deconvolution, based on the principle of positional fixing, is a feasible and effective method for identifying novel zinc finger proteins with high affinity for their target DNA sequences. This approach has several advantages over conventional methods, including the ability to screen large number of protein-DNA interactions rapidly and efficiently, eliminating the need for multiple rounds of biopanning, protein purification, and sequence tagging. The success of this approach highlighted the appropriateness of positional fixing for identifying novel proteins as it offers a more direct and efficient alternative to conventional methods. This method could significantly impact future research by enabling the rapid discovery and characterisation of zinc finger proteins for various applications, including gene regulation and protein engineering.

1.4.7. Future direction

Combinatorial chemistry has dramatically transformed pharmaceutical research and development compared to traditional drug discovery approaches. While conventional drug discovery typically involves the step-by-step synthesis and testing of individual compounds - a process that is both time-consuming and resource-intensive - combinatorial chemistry enables the simultaneous synthesis and screening of extensive compound libraries. This approach speeds up the discovery process and enhances the likelihood of finding effective drug candidates. The use of combinatorial libraries has shifted pharmaceutical research from focusing on single compounds to exploring vast compound collections.

This shift has paved the way for dynamic combinatorial libraries, fragment-based drug discovery, and virtual library screening. These strategies allow researchers to rapidly generate and assess millions of compounds, greatly improving drug discovery efficiency and creating new possibilities for developing innovative therapeutics (Furka, 2022).

The future of combinatorial libraries lies in expanding chemical diversity, improving screening methodologies, and integrating computational approaches. As technology advances, the integration of artificial intelligence and machine learning is poised to revolutionise library design and screening (Monti et al., 2023). Future advancements in screening methods are likely to focus on miniaturisation, automation, and increased throughput, making it possible to screen even larger and more diverse libraries. New analytical techniques, such as mass spectrometry and high-content imaging, will enhance the identification and characterisation of active compounds. Combinatorial libraries, once limited to peptides, are evolving to encompass a vast chemical space, promising an exciting future for drug discovery and research. The integration of high-throughput screening and programmable robotic instruments has significantly transformed the efficiency and scale at which combinatorial libraries can be synthesised and analysed. Programmable robotic instruments, on the other hand, enable the precise and automated synthesis of combinatorial libraries, reducing manual labour and human error (Potyrailo et al., 2011). Additionally, the development of DNA-encoded combinatorial libraries, combined with next-generation sequencing, has brought even more revolutionary changes to the field. These innovations enable the amplification and highly precise screening of billions of compounds within a single experiment, greatly enhancing both the throughput and accuracy of drug discovery (Suay-García et al., 2022).

1.5. Mass photometry

Mass photometry, though less than five years old, has rapidly gained widespread use in laboratories, is commercially available, and has been cited in hundreds of publications. This non-destructive, label-free technique provides unparalleled sensitivity for measuring molecular mass, making it highly useful for studying nucleic acids, single membrane proteins, and protein–protein interactions. Despite its recent development, mass photometry is rooted in the well-established principle of light scattering, a concept dating back to the 17th century. Through innovative methods to overcome the challenge of detecting weak scattering signals from single molecules, mass photometry enables precise mass measurements of molecules that would otherwise be nearly invisible (Graciani & Yoon, 2023).

1.5.1. Mass photometry as an analytical tool

Mass photometry (MP) is a label-free technique that estimates the molecular mass of single biomolecules in solution (Becker et al., 2023, Soltermann et al., 2020). MP relies on the linear relationship between the intensity of light scattered by a molecule and its mass. The method detects single biomolecules by measuring the light they scatter as they land on a microscope

coverslip (Soltermann et al., 2020). The resulting change in refractive index at the glass-water interface alters the local reflectivity. By optimising the interference between the scattered and reflected light, each binding event can be detected with high accuracy (Li et al., 2020, Soltermann et al., 2020). The change in reflectivity is directly proportional to the mass of the molecule, so the technique is calibrated using molecules of known mass, allowing for mass determination of unknown samples with high accuracy and precision (Soltermann et al., 2020, Becker et al., 2023).

1.5.2. The principle of mass photometry

MP works by imaging the interference of light scattered by a molecule with light reflected from a glass-water interface (Wu & Piszczek, 2020a). A dilute solution of biomolecules is placed on a microscope coverslip and a laser beam illuminates the interface between the coverslip and the solution. When a molecule binds to the coverslip, it scatters light. This scattered light interferes with the light reflected from the glass-water interface, creating a change in the local reflectivity that can be detected by a camera (Wu & Piszczek, 2020).This change in reflectivity is proportional to the mass of the molecule (Soltermann et al., 2020). Figure 1.10 illustrates how individual binding events appear as diffraction-limited spots in the ratiometric images. The mass of an unknown molecule can be determined by comparing its scattering signal to the scattering signals of molecules of known mass.

The light scattered by the biomolecule and the reflected light from the glass-water interface are captured by a camera. The resulting interference contrast increases linearly with the molecule's polarisability, which is proportional to its refractive index and volume as shown in Figure 1.11 (Asor & Kukura, 2022).

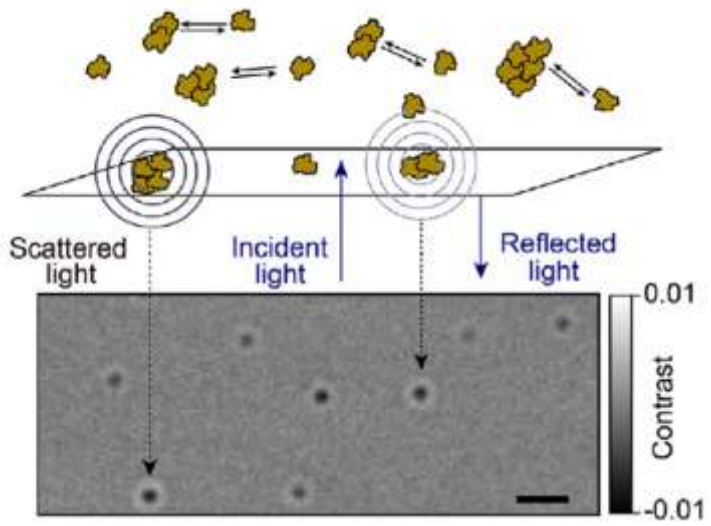


Figure 1.10. Working principles of mass photometry. *Binding events show up as distinct diffraction-limited spots (scale bar 1 mm) in the rolling averaged ratiometric images, highlighting that intensity changes due to reflectivity variations result from biomolecules binding to the glass surface. The contrast of each spot corresponds to the mass of the bound macromolecule (Asor & Kukura, 2022).*

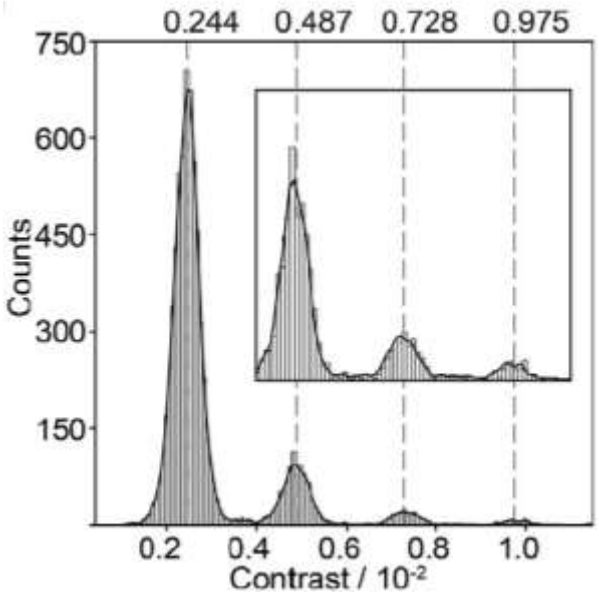


Figure 1.11. An example of capturing the contrast of numerous molecules. *A histogram reflects the distribution of solution (Asor & Kukura, 2022).*

When using MP, the cleanliness of the coverslip matters and checking for a proper MP image is essential (Figure 1.12A&B). The glass surface should be free from significant imperfections, and the signal value (RMS deviation of the MP image) should be 0.05% or lower. If this value exceeds 0.05%, the coverslip should be discarded and the opposite side of the coverslip should be tested. If neither side meets the acceptable signal threshold, the cleaning process was insufficient, thus the experiment should start again with a new coverslip and more rigorous cleaning (Wu & Piszczek, 2021). Protein sample concentration range for MP measurements is also important. Ideally, there should be a high frequency of landing events while maintaining adequate spatial separation. To meet these conditions, the suggested 20 nM sample concentration can be modified as needed (Figure 1.12C; Wu & Piszczek, 2021).

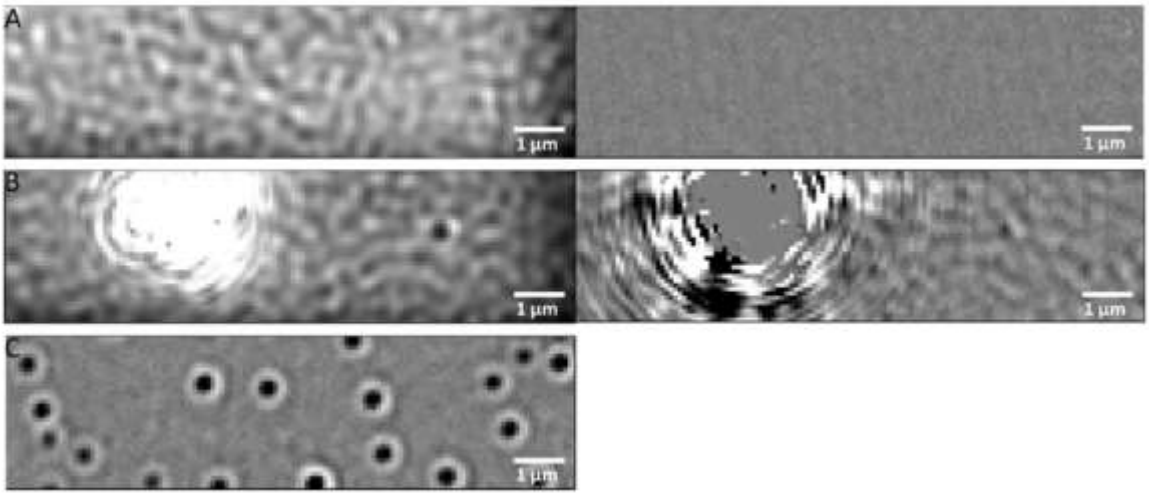


Figure 1.12. Example screenshots from the MP camera. (A) *Native (left) and ratiometric (right) buffer images on a clean coverslip.* ***(B)*** *Native (left) and ratiometric (right) images showing an inadequately cleaned coverslip surface.* ***(C)*** *Ratiometric view of protein landing events at the optimal sample concentration (Wu & Piszczek, 2021).*

1.5.3. Applications of mass photometry

1.5.3.1. Characterising biomolecular interactions

MP can be used to study protein-protein interactions ,enables real-time monitoring of protein-protein interactions without requiring labels or complex sample preparation, measures binding affinities by detecting changes in the mass distribution of protein complexes upon interaction. (Soltermann et al., 2020), protein-DNA interactions, provides a direct method for observing binding between proteins and DNA, which is critical for studying regulatory and repair mechanisms, and allows detailed analysis of protein-DNA complex stoichiometry, revealing the precise assembly states (Li et al., 2020), and antibody-antigen interactions, MP can assess the specificity and binding affinity of antibodies to antigens, facilitating drug development and diagnostic applications or Kinetics analysis**,** by tracking mass distributions over time, MP provides insight into the kinetics of binding and dissociation (Wu & Piszczek, 2021). MP can determine binding affinities and kinetics of interactions, as well as stoichiometries of complexes (Soltermann et al., 2020, Wu & Piszczek, 2021).

1.5.3.2. Analysing protein oligomerisation and aggregation

MP can be used to study the formation of oligomers and aggregates, which are implicated in various diseases (Asor & Kukura, 2022). The technique called single-molecule mass photometry (SMMP) can be applied to track the evolution of different oligomeric species over time, during aggregation (Figure 1.13)(Paul et al., 2022). SMMP allows for the quantification of different oligomer populations, including transient states, as they evolve over time during

aggregation. This capability is particularly useful as small oligomers are thought to be key neurotoxic species in diseases like Alzheimer's. By monitoring changes in oligomer populations over time, SMMP enables the development and testing of kinetic models of aggregation. The microscopic rates for each step in the aggregation mechanism can be measured.This helps in identifying the most likely mechanisms and quantifying the rates of each step of the aggregation.

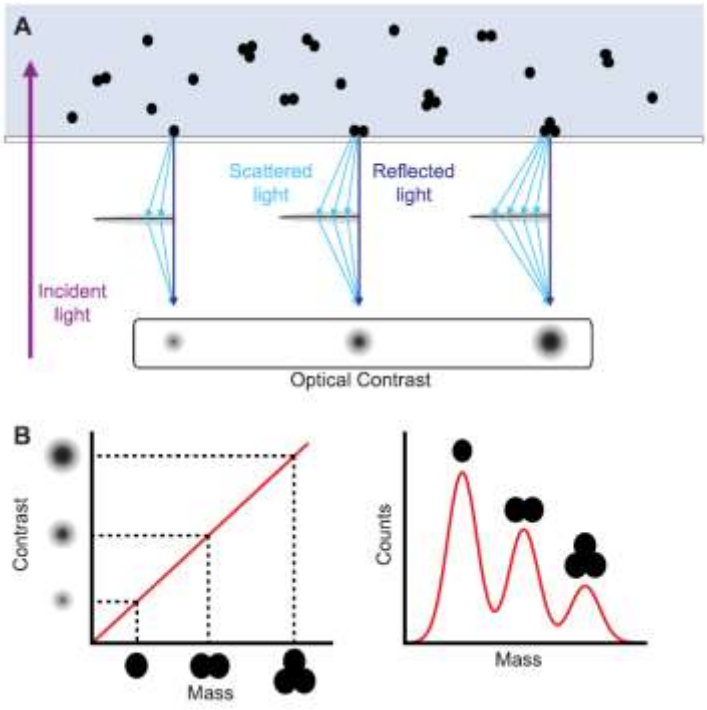


Figure 1.13. Single-molecule mass photometry (SMMP). (A) *When protein molecules attach to a coverslip surface, they scatter light according to their size. The interference of this scattered light with the incident light reflected off the coverslip generates an optical contrast, which varies based on the protein's size.* ***(B)*** *This optical contrast is linearly related to protein mass, allowing the determination of each protein's mass during binding events (left) and enabling the measurement of mass distributions (right) (Paul et al., 2022).*

1.5.3.3. Characterising antibody-antigen binding affinities

MP offers advantages over traditional methods for measuring antigen-antibody affinities, such as isothermal titration calorimetry (ITC) and biolayer interferometry (BLI) (Wu & Piszczek, 2021) demonstrated the potential of MP for rapid and accurate measurement of antigen-antibody binding affinities using minimal sample volumes and concentrations. MP requires much smaller sample volumes and lower concentrations compared to ITC, which typically requires high sample concentrations and large volumes to detect heat changes during binding. This is particularly important for antigen-antibody studies where the availability of reagents (e.g., antigens or engineered antibodies) may be limited. Apart from that, ITC measures the

heat changes associated with binding, which may not always be significant for weak or low-affinity interactions, while MP directly measures binding events by detecting mass changes, making it more versatile for a wider range of interactions, including weak antigen-antibody affinities. MP enables rapid detection and quantification of binding interactions, while ITC requires multiple injections and equilibration steps, making it time-intensive (Wu & Piszczek, 2021).

1.5.3.4. Studying nucleic acids

MP can be used to measure the mass of nucleic acids, provides precise mass measurements of nucleic acids, enabling direct quantification of their molecular weight. This is critical for analyzing nucleic acid purity, integrity, and modifications, and to study their interactions with proteins, MP enables real-time monitoring of interactions between nucleic acids and proteins without the need for complex setups or secondary labeling, for example, it can directly measure the formation of nucleoprotein complexes, such as transcription factors binding to DNA (Y. Li et al., 2020, Asor & Kukura, 2022). The relationship between the number of bases in dsDNA and the corresponding MP contrast has been found to be linear up to 1200 bp, enabling length quantification of dsDNA with up to 2 bp accuracy (Li et al., 2020).

1.5.3.5. Membrane protein characterisation

Dynamic MP has been developed to study membrane proteins on supported lipid bilayers (Foley et al., 2021). By implementing a new background processing methodology, dynamic MP can image, track, and measure the mass of diffusing protein complexes on the membrane (Foley et al., 2021). Mass photometry has also been used to simplify the characterisation of membrane proteins using various membrane mimetics, to assess homogeneity (Marty, 2021). Dynamic MP has also been used to track single membrane-associated proteins diffusing on supported lipid bilayers. Specifically, Foley et al., (2021) applied this technique to study dynamin-1, a membrane-remodelling GTPase, and revealed heterogeneous mixtures of dimer-based oligomers. This study expanded the capabilities of MP beyond analysing stationary molecules, allowing for the investigation of complex membrane-associated processes.

1.5.4. Advantages of MP over other methods

1.5.4.1. Label-free detection

MP does not require the use of fluorescent labels or tags, which can be time-consuming and expensive to implement, and may interfere with the biological activity of the molecule being studied (Wu & Piszczek, 2021, Graciani & Yoon, 2023).

1.5.4.2. Solution-based measurements

MP measurements are performed in solution, providing a more native environment for studying biomolecules than surface-based techniques (Wu & Piszczek, 2020,Graciani & Yoon, 2023) and provide direct information on the mass and relative abundance of molecules in solution (Wu & Piszczek, 2021).

1.5.4.3. Sensitivity

The ability to detect weak scattering signals from small biomolecules is crucial for achieving a low detection limit and expanding the applicability of MP to a wider range of molecules. The detection limit of commercially available MP instruments was initially reported as being around 50 kDa (Soltermann et al., 2020, Claasen et al., 2024), but a lower detection limit of 40 kDa has been reported in research settings (Wu & Piszczek, 2021) and recent personal communication from the manufacturers in now suggests a lower limit of 30 kDa (Hine, pers. comm.) This limitation arises from the difficulty in distinguishing weak scattering signals from background noise (Asor & Kukura, 2022). Several factors contribute to the sensitivity of MP, including the incident light intensity, the collection efficiency of the detection optics, the quantum efficiency of the detector, and the efficiency of the background removal algorithms (Becker et al., 2023, Claasen et al., 2024).

1.5.4.4. Mass resolution

The ability to accurately differentiate between molecules with similar masses is essential for studying heterogeneous samples and for characterising biomolecular interactions with high precision (Soltermann et al., 2020). The mass resolution of MP is determined by the width of the peaks in the contrast distribution histograms (Becker et al., 2023). This resolution is typically reported as the full width at half maximum (FWHM) of the peaks and is influenced by factors such as the signal-to-noise ratio (SNR), the precision of the mass calibration, and the heterogeneity of the sample itself (Asor & Kukura, 2022; Becker et al., 2023). A higher mass resolution enables researchers to study subtle changes in mass, such as the binding of small ligands or post-translational modifications (Asor & Kukura, 2022).

1.5.4.5. Concentration range

The optimal concentration range for MP is determined by the density of landing events on the coverslip surface (Wu & Piszczek, 2020a). At low concentrations, the number of binding events may be too low to provide statistically significant data (Claasen et al., 2024). Conversely, high concentrations can lead to overlapping binding events that are difficult to resolve. In standard

MP, the ideal concentration range is typically between 100 pM and 100 nM (Claasen et al., 2024). Techniques like microfluidics can be used to extend this concentration range, allowing the study of weaker interactions at higher concentrations (Claasen et al., 2024).

1.5.4.6. General utility

The time required to perform an MP experiment is another factor that contributes to its utility (Wu & Piszczek, 2020). Fast measurements allow for high-throughput screening and enable the study of dynamic processes (Wu & Piszczek, 2020). A typical MP measurement may take less than 5 minutes for data acquisition and analysis (Wu & Piszczek, 2020).

MP measurements also typically require very small sample volumes, which is advantageous when working with precious or limited samples. Typical MP experiments use sample volumes around 10-20 µL at nanomolar concentrations. The low sample consumption allows for multiple measurements and reduces the cost of experiments (Wu & Piszczek, 2020a).

1.5.5. Limitations of mass photometry

The lower detection limit of MP is currently around 30-50 kDa, limiting its application to larger biomolecules (Wu & Piszczek, 2021). The technique struggles to measure smaller molecules due to their weak scattering signals (Graciani & Yoon, 2023). MP measurements are typically performed at low nanomolar concentrations. Measuring samples outside this range (either too dilute or too concentrated) can lead to inaccuracies (Claasen et al., 2024). Concentrated samples can result in overlapping binding events that cannot be accurately resolved, while dilute samples may not generate sufficient data for analysis (Claasen et al., 2024). In terms of nucleic acid analysis, factors such as GC content, DNA nicks, supercoiling, circularity, and variations in secondary structure could influence the accuracy of mass determination for nucleic acids (Li et al., 2020). The requirement for molecules to bind to the coverslip surface can limit the study of certain biomolecules and interactions (Soltermann et al., 2020). The binding process might also influence the measured properties of the molecules, especially the kinetics (Wu & Piszczek, 2020a).

One of the limitations regarding the sample concentration was improved by a recent study, in which mass photometry was combined with rapid-dilution microfluidics. This method allowed researchers to analyse samples at higher initial concentrations, expanding the range of protein-protein interactions that can be studied. The microfluidic system rapidly dilutes the sample to the appropriate concentration range immediately before it reaches the mass photometry observation area. This rapid dilution allows for the detection of transient, low-affinity complexes that may have dissociated during slower, manual dilution because the measurement is completed before the equilibrium of the interaction is significantly altered. The

diluted sample is continuously delivered to the detector for analysis. By combining these two powerful techniques, researchers can now investigate a wider range of biomolecular interactions, including those with weaker affinities and faster dissociation rates, leading to a better understanding of complex biological processes (Claasen et al., 2024).

1.5.6. Future directions of mass photometry

MP is a rapidly evolving technique with continuous improvements and expansions. Reducing the lower mass limit through improved sensitivity and signal-to-noise ratios will open up new avenues for studying smaller biomolecules (Foley et al., 2021), which might be achieved by developing more efficient background removal algorithms and by improving the sensitivity of the detectors (Foley et al., 2021). Higher mass resolution will enable more precise differentiation between different species in a sample, facilitating the study of subtle changes in mass (Becker et al., 2023). Advancements in instrumentation and data analysis techniques are expected to contribute to this goal. Meanwhile, developing new applications by combining MP with other techniques such as microfluidics, electrophoresis, and other imaging modalities holds potential for expanding the range of applications and addressing current limitations (Claasen et al., 2024, Paul et al., 2022). Continued advancements in the technique and the development of new applications are expected to further expand its utility and impact in the field of biomolecular research.

1.6. The aims and objectives of the project

While previous studies have successfully increased the interaction rate between SpyTag and SpyCatcher (e.g. Keeble et al., 2019), few have focused on generating orthogonal specificity. Rather, where controlled immobilisation of different proteins is required, combinations of SpyCatcher-SpyTag / SnoopCatcher-SnoopTag have been used, such as for the patterning of a surface monolayer, as described by Regan and co-workers (Williams et al., 2018). Accordingly, the current study aimed to examine whether it would be possible to alter the specificity of the Spy-Tag-SpyCatcher interaction by substituting residues both in the hydrophobic binding pocket of SpyCatcher and the corresponding positions of SpyTag with alternative hydrophobic residues and further, by making such mutations, whether it might be possible to generate orthogonal Spy-Tag-SpyCatcher pairs. The project was based on a preliminary discussion with Prof Lynne Regan (University of Edinburgh, pers comm.) who, following molecular modelling, had proposed that substitution of residues 27, 44 and 90 within SpyCatcher and corresponding residues 3 and 5 within Spy-Tag with hydrophobic residues

phenylalanine, isoleucine, leucine, methionine, valine and lastly, tyrosine might achieve that goal.

The aim of the project was therefore divided into the following objectives:

1. Verification of residues selected for mutagenesis via molecular imaging
2. Application of the principles of combinatorial chemistry to make combinatorial libraries both of SpyCatcher and SpyTag
3. Identification of a suitable screening protocol
4. Analyses of the combinatorial SpyCatcher libraries with native SpyTag (to identify novel SpyCatcher proteins with either modified or lost specificity for native SpyTag)
5. Analyses of individual SpyCatcher proteins identified within objective 4 with SpyTag libraries to determine specificity and/or orthogonality.

Chapter 2 Materials and Methods

Media and buffers were prepared using distilled water (dH_2O). For molecular procedures, such as Polymerase Chain Reaction (PCR), DNA elution during extraction, and other related processes, ultrapure 18.2 MΩ water (also referred to as double distilled water, ddH_2O) was used.

2.1. Materials

2.1.1. Risk assessment

The necessary risk assessments were completed and filed online and in the laboratory.

2.1.2. Vectors

Vector	Comments
pET24(+)His-WT spytag-linker-mNeonGreen-His6	A gift from University of Edinburgh, School of Biological Sciences
pET24mCherry-WtSpyCatcher	A gift from University of Edinburgh, School of Biological Sciences

Table 2.1. Overview of vectors utilised in the study

2.1.3. Cell lines

Organism	Strain	Genotype
E. coli	DH5α	F– φ80*lacZ*ΔM15 Δ(*lacZYA-argF*)U169 *recA1 endA1 hsd*R17(rK–, mK +) *phoA supE44 λ– thi-1 gyrA96 relA1*
E. coli	Tuner™(DE3)	F– ompT hsdSB (rB– mB–) gal dcm lacY1(DE3)

Table 2.2. Overview of Cell lines utilised in the study

2.1.4. Media recipes

2.1.4.1. Agar

The medium was prepared using Bacto-Agar (Difco™, 0140-01) according to manufacturer's instructions. The solution was autoclaved at 121 °C for 20 minutes and allowed to cool to approximately 55 °C. For preparing solid medium plates, appropriate antibiotic was added to the medium as required and 25-30 mL of the agar-media solution was aseptically poured into each petri dish. The plates were left to solidify at room temperature overnight and then stored at 4 °C the next day.

2.1.4.2. LB broth medium

To prepare a 2% (w/v) medium solution, LB broth powder (Sigma LB Broth, L3022) was dissolved in distilled H2O. The solution was autoclaved at 121 °C for 20 minutes. After cooling, 50µg/mL of sterile kanamycin solution was added for liquid media.

2.1.4.3. SOC medium (Super Optimal broth with Catabolite repression) (Invitrogen, 15544-034)

- Tryptone 2 % (w/v)
- Yeast extract (YE) 0.5 % (w/v)
- Sodium chloride (NaCl) 10 mM
- Potassium chloride (KCl) 2.5 mM
- Magnesium chloride (MgCl2) 10 mM
- Magnesium sulfate (MgSO4) 10 mM
- Glucose 20 mM

2.1.5. Buffer recipes

2.1.5.1. 1X Tris-Acetate-EDTA (TAE) Buffer (50X) (Thermo Scientific™, B49)

- Tris base 40 mM
- Acetic acid 20 mM
- EDTA 1 mM

2.1.5.2. Blue/orange DNA Gel Loading Dye (6X) (Promega, G1881)

- Ficoll®-400 15%
- Xylene cyanol FF 0.03%
- Tris-HCl (pH:7.5) 10mM
- EDTA (pH:8.0) 50mM
- Orange G 0.4%
- Bromophenol blue 0.03%

2.1.5.3. Gel Loading Dye Purple (6X) (NEB, B7024S)

2.1.5.4. Sample Buffer Laemmli 2x Concentrate (Sigma-Aldrich, S3401)

2.1.5.5. T4 DNA Ligase Buffer (10X) (NEB, B0202S)

- Tris-HCl 40 mM
- Magnesium chloride (MgCl2) 10 mM
- ATP 0.5 mM
- Dithiothreitol (DTT) 10 mM

2.1.5.6. CutSmart® Restriction Digest Buffer (10X) (NEB, B6004S)

- Tris-acetate 20 mM
- Magnesium acetate 10 mM
- Potassium acetate 50 mM
- BSA 100 µg/mL

2.1.5.7. Phusion HF Buffer (5X) (supplied with Phusion Thermo Scientific™, F530S)

Contains 7.5 mM MgCl2. Other components not disclosed by manufacturers.

2.1.5.8. Pfu DNA polymerase buffer (10x) (Promega, M7741)

- Tris-HCl (pH 8.8) 200 mM
- KCl 100 mM
- (NH4)2SO4 100 mM
- MgSO4 420 mM
- Triton® X-100 1% v/v
- Nuclease-free BSA 1 mg/mL.

2.1.5.9. SDS-PAGE and protein purification buffers

2.1.5.9.1. Binding buffer

- Tris-HCl pH 8 50 mM
- NaCl 30 mM
- Glycerol 10% (v/v)
- Imidazole 10 mM

2.1.5.9.2. Wash buffer

- Tris-HCl, pH 8.0 50 mM
- NaCl 30 mM
- Glycerol 10% (v/v)
- Imidazole 20 mM

2.1.5.9.3. Elution buffer

- Tris-HCl, pH 8.0 50 mM
- NaCl 30 mM
- Glycerol 10% (v/v)
- Imidazole 250 mM

2.1.5.9.4. SDS-PAGE Running Buffer (10X)

- Tris base 250 mM

- Glycine 1.92 M
- Sodium Dodecyl Sulphate (SDS) 1% w/v

2.1.5.9.5. 1X SDS-PAGE sample buffer

- Tris-HCl (pH 6.8) 50 mM
- Dithiothreitol (DTT) 100 mM
- SDS 2 % (w/v)
- Glycerol 8 % (v/v)
- Bromophenol blue 0.1 % (w/v)

2.1.6. Solutions by Application

2.1.6.1. Antibiotics

- Kanamycin Sulphate (Sigma-Aldrich, B5264)
- Stock solution (50 mg/mL) prepared and stored at -20°C.

2.1.6.2. Electrophoresis and protein analysis

- Ethidium bromide solution (Sigma-Aldrich, E7637)
- Concentration of 1 µg/mL for gel staining.
- 10% Ammonium persulfate (APS) (Sigma-Aldrich, 215589)
- Sodium Dodecyl Sulfate (SDS) (Thermo Scientific™, 28365)
- 10% (w/v) solution.
- N,N,N′,N′-Tetramethylethylenediamine (TEMED) (Sigma-Aldrich, 1.10732)
- SureCast 40% (w/v) Acrylamide (Thermo Scientific™, HC2040)
- Acrylamide: Bis-acrylamide ratio of 29:1.
- InstantBlue™ Protein Stain (Sigma-Aldrich, RSB-1L)

2.1.6.3. Protein extraction and purification

- BugBuster® Protein Extraction Reagent (Merck Millipore, 70584-3)
- HisPur TM Ni^{2+}-NTA Resin (Thermo Scientific™, 88222)
- Imidazole (Sigma-Aldrich, I3386)
- Isopropyl β-D-1-thiogalactopyranoside (IPTG) (Sigma-Aldrich, 16758)
- Prepared as a 1 M solution and stored at -20°C.

2.1.6.4. PCR and DNA Amplification

- Phusion™ Plus PCR Master Mix (Thermo Scientific™, F631S)

- BamHI restriction endonuclease (NEB, R0136S)
- dNTP Mix (Promega, U1511)

2.1.6.5. Molecular Cloning

- NEBridge Golden Gate Assembly Kit (BsaI-HF v2) (NEB, E1601S)
- NEBuilder HiFi DNA Assembly Master Mix (NEB, E2621G)

2.1.6.6. General Laboratory Solvents and solutions

- Ethanol
- Prepared at 70% (v/v).
- Isopropanol
- Phosphate-buffered saline tablet (PBS) (Thermo Scientific™, 003002)
- Prepared using PBS tablets and sterilised.

2.1.6.7. Oligonucleotides

Purchased from Eurofins Genomics (HPLC purified, 100 nmol scale) or Sigma (Desalted, 20 nmol scale).

2.1.6.8. Ladders for DNA and Protein

- DNA Step Ladders
- 1 kb and 50 bp ladders (Promega, G694A and G452A,respectively).
- Protein Ladders
- PageRuler™ Prestained and Unstained Protein Ladder (Thermo Scientific™, (26617 and 26614,respectively).

2.2. Methods

2.2.1. Methodology for utilising PyMOL in molecular modelling

2.2.1.1. Download protein code

In the first step, a file must be downloaded from the Protein Data Bank (PDB) at www.rcsb.org for use in PyMOL. PDB files are identified by a unique alphanumeric code. After obtaining the PDB ID, the file can be downloaded in PDB format by selecting the "Download Files" option (Figure 2.1).

Figure 2.1. The PDB ID for SpyTag-SpyCatcher complex. *The PDB is displayed, and the structure can be downloaded in PDB format from the right-hand side of the page.*

2.2.1.2. Loading files into PYMOL

Upon opening PyMOL, two main windows are available: the main viewing window and the upper control menu. On the right side of the main window, where the protein structure is displayed, there is an object panel containing several buttons labelled A, S, H, L, and C. Each button provides specific functions: A stands for "Action," allowing various actions, analyses, and calculations; S stands for "Show," which applies graphical effects to selected residues or molecules; H stands for "Hide," which removes graphical effects; L stands for "Labelling," and C for "Colouring" (Figure 2.2). A file can be loaded into PyMOL either by using the pull-down menu and navigating to File > Open to select the desired PDB ID, or by typing the Fetch command, such as "Fetch 4MLI."

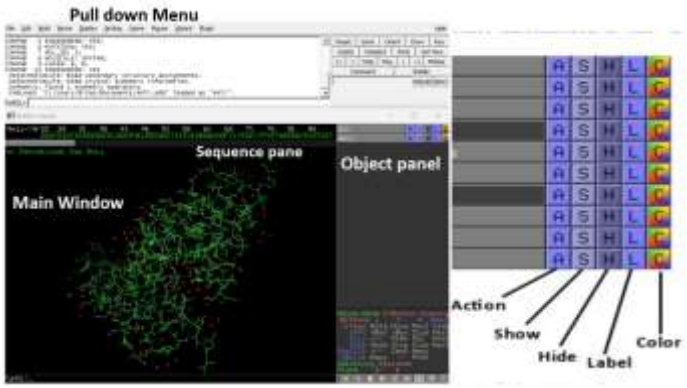


Figure 2.2. PyMOL interface showing the main window and the upper "pull-down" menu. *The object panel is located on the right side, and the sequence pane displays the protein sequence. The buttons labelled A (Action), S (Show), H (Hide), L (Label), and C (Colour) provide various options for manipulating and visualising molecular structures.*

2.2.1.3. Using object menu panel

To begin, the structure should be displayed in cartoon format by selecting 'S' > 'Cartoon,' which will show the structure as a green cartoon. Next, the lines can be removed by selecting 'H' > 'Lines,' and the water molecules (represented as red crosses) should be hidden by choosing 'H' > 'Waters.' The sequence can be displayed in the sequence pane by pressing the 'S' button located at the bottom right of the screen (Figure 2.3). The residue numbers above the amino acid sequence can be viewed by navigating to Display > Sequence Mode > All residue numbers. Since PyMOL lacks an undo button, it is important for the session to be saved regularly by going to the upper menu and selecting File > Save Session.

Figure 2.3 Representation of protein structure. *The protein structure is displayed as a cartoon in the main window after removing water molecules and hiding the lines for a clearer representation. The 'S' button located at the bottom of the menu to shows all the residues.*

One way to display structures using various presets is by navigating to A → Action → Preset, as shown in Figure 2.4. This menu provides different options for visualising a protein structure.

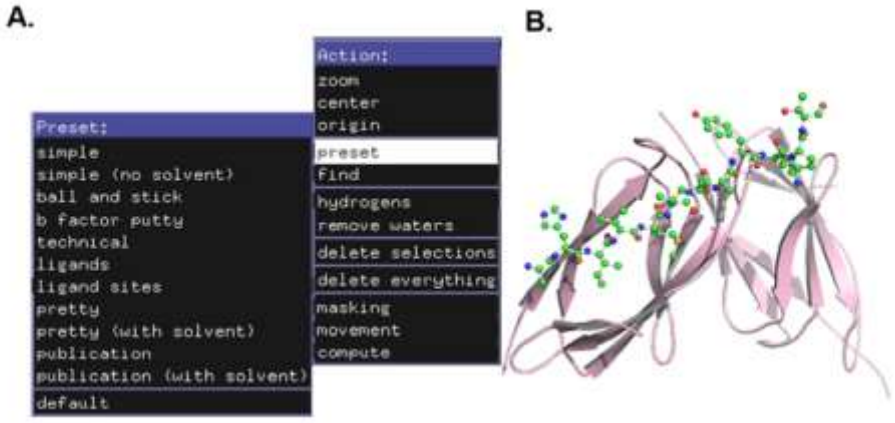


Figure 2.4.Representation of various options for viewing the protein in different formats. (A) *By navigating to Action > Preset, users can choose from several visualisation presets such as simple, ball and stick, technical, ligands, ligand sites, and more, allowing for tailored representation of the protein*

structure. ***(B)*** *The peptide is displayed using the ball-and-stick representation, with atoms coloured by element.*

2.2.1.4. Selecting and displaying

To select and display specific residues to understand the structure-function relationship, the relevant residues must be highlighted. Selection should begin by choosing one or more residues from the sequence pane, after which 'S' > 'Sticks' should be selected to display them. The color of the selected residues can be altered by using 'C' > 'By SS' or by coloring them according to their element.

For clarity, the selected residues should be renamed by changing "sele" to the name of the residue, using 'A' > 'Rename Selection.' For example, the selection could be renamed to "SpyTag," and the entire tag displayed as sticks with a distinct color (Figure 2.5). To change the background color, the upper menu should be used to select Display > Background > White.

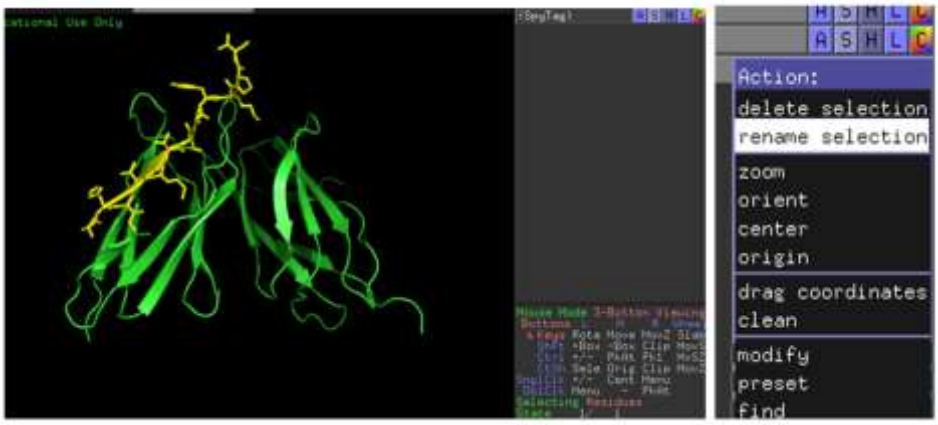


Figure 2.5. Renaming selected residues in PyMOL via the Action menu. *When selecting residues, the selection can be renamed by navigating to 'A' > 'Rename Selection'. This allows for easy identification and management of specific residues or groups within the structure.*

2.2.1.5. Display hydrogen bonds

The option to display polar contacts is provided by PyMOL and can be accessed by navigating to 'A' > 'Find' > 'Polar Contacts' and selecting from various submenus. The option "to other atoms in object" can be chosen to reveal additional hydrogen bonds between the helix side-chains and other regions of the protein. As shown in Figure 2.6, a specific residue can be selected, and its polar contacts can be displayed, allowing the hydrogen bonds and other polar interactions involving that residue to be visualized.

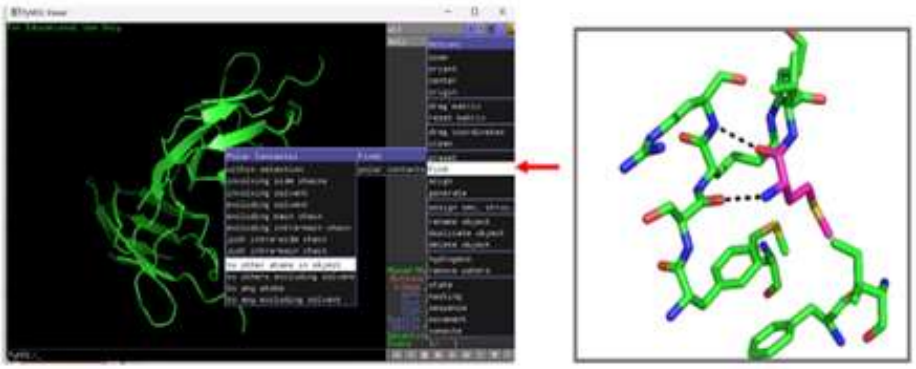


Figure 2.6. Displaying polar contacts using the Action menu. *A residue highlighted in magenta forms two polar contacts with surrounding residues, as indicated by the dotted lines. This interaction occurs within the selected residue, demonstrating the molecular interactions contributing to the protein's structure.*

2.2.1.6. Display molecular distances

Understanding how specific residues interact with other residues or ligands is crucial for determining protein function. To identify residues likely to interact with others, the structure (e.g., 4MLI) should be selected, and 'A' > 'Modify' > 'Expand' > 'by 4 Å' should be navigated to expand the selection to nearby residues (Figure 2.7A). It is recommended for the cartoon representation to be made semi-transparent by selecting Setting > Transparency > Cartoon > 50% in the upper menu (Figure 2.7B).

When interactions are displayed, some may involve water molecules. To show water molecules, the command "show nonbonded" should be typed in the command line and executed by pressing enter, causing water molecules to appear as stars. A water molecule can also be selected and displayed as a surface to make it more visible. To remove extra water molecules, the command "hide nonbonded" should be typed (Figure 2.8). For a clearer view of polar contacts and interactions, it is recommended that the cartoon be hidden by selecting 4MLI > Hide > Cartoon.

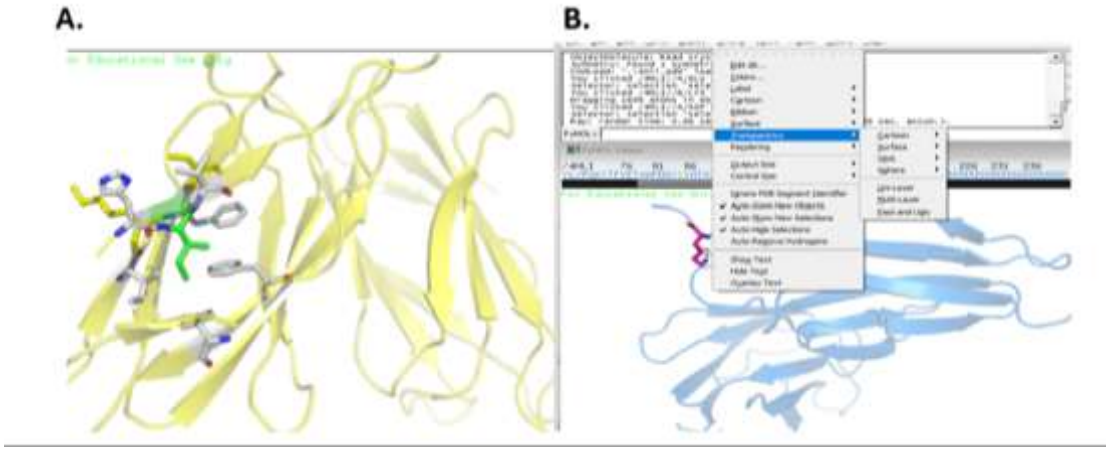


Figure 2.7. Visualisation of residue interactions within: (A) *The distances between a selected residue (shown in green) and surrounding residues within a 4 Å radius are displayed. The interacting residues are represented as grey sticks, highlighting the close contacts and potential interactions between these residues.* ***(B)*** *Adjusting cartoon transparency in PyMOL. To make the cartoon representation transparent.*

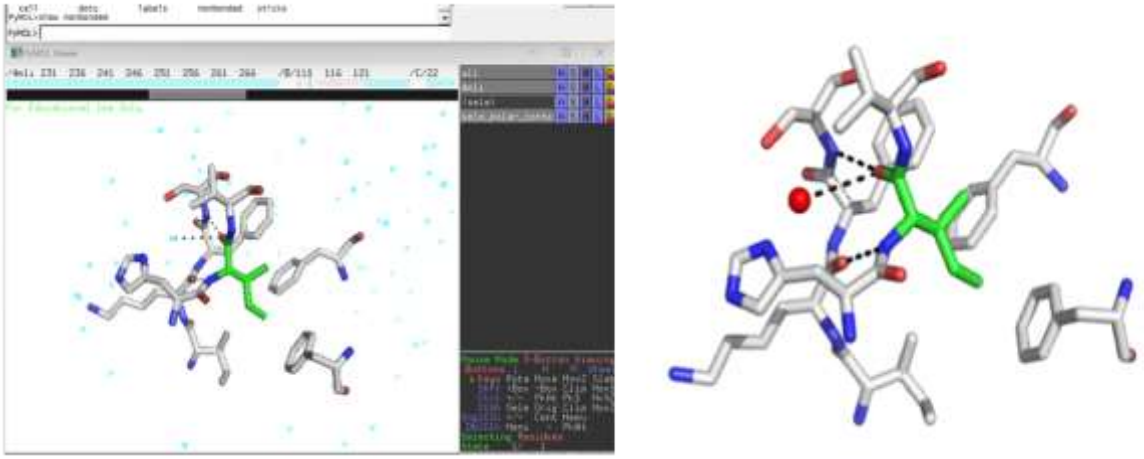


Figure 2.8. Displaying water molecules and interactions. *The interactions are displayed while showing water molecules by typing "show nonbonded" in the command line. All nonbonded waters are represented as cyan stars. The interacting water molecule is shown as a surface, while all other water molecules have been hidden for clarity.*

2.2.1.7. Taking measurement

To measure the distance between two residues, the upper menu should be used to select Wizard > Measurement. The first atom should then be clicked, followed by clicking the second atom, and the process should be completed by pressing Done**,** as shown in Figure 2.9.

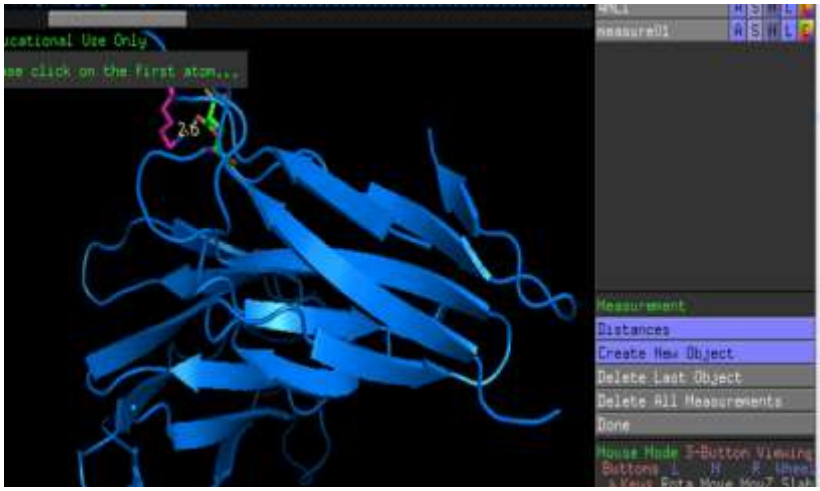


Figure 2.9. Measuring the distance between two selected residues. *The distance between two selected residues is displayed as 2.6 Å, using the measurement tool in PyMOL. This allows for precise analysis of spatial relationships between atoms in the protein structure.*

2.2.1.8. In silico mutagenesis

The ability to perform mutagenesis is offered by PyMOL, allowing one amino acid to be changed to another to assess its impact on the protein structure. The mutagenesis tool can be accessed via the "Wizard" menu at the top of the screen by selecting Wizard > Mutagenesis. Various options, including Mutation, Apply, Clear, and Done**,** are provided in this menu. The "Apply" option is used to apply the selected mutation, "Clear" is used to remove the current mutation, and "Done" is used to exit the mutagenesis tool (Figure 2.10).

Through the Mutation menu, any of the 20 standard amino acids, as well as specific ionization states for certain residues, can be selected. As an example, the residue I3 has been selected for mutation and is displayed in yellow stick representation. The residue I3 has been mutated to tyrosine using the Wizard Mutagenesis tool (Figure 2.11). The targeted residue is highlighted, making it easier to visualize and manipulate during the mutation process.

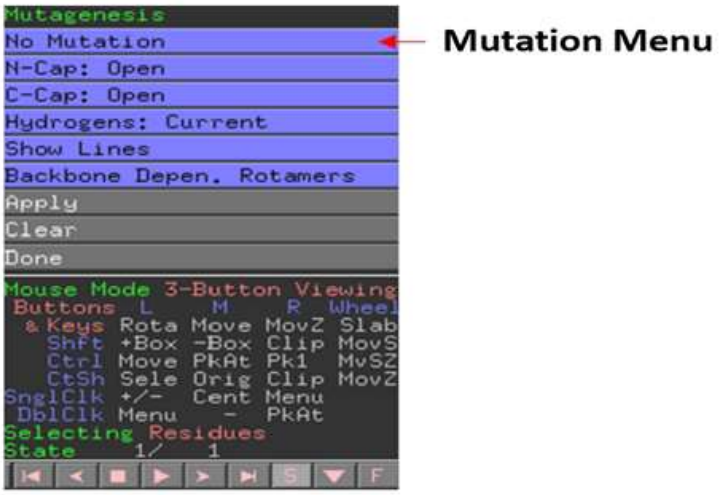


Figure 2.10. Mutagenesis menu for residue mutation. *The mutation menu displays "No Mutation" in the first row. By clicking on it, a list of twenty possible residues for mutation is revealed. After selecting the desired mutation, click "Apply" and then press "Done" to complete the process.*

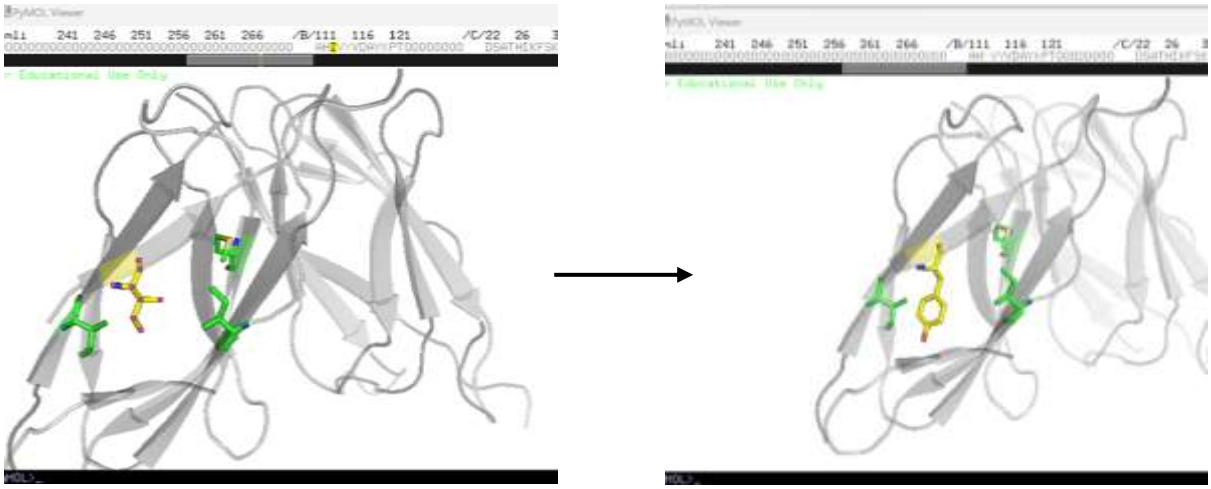


Figure 2.11. Mutation of reside using the mutagenesis Wizard. *The residue I3 has been mutated to Tyrosine (Tyr) using the Wizard Mutagenesis tool. The newly applied Tyr residue is shown in yellow stick representation, replacing the original Isoleucine (I3).*

2.2.2. Molecular procedures

2.2.2.1. Polymerase chain reaction (PCR)

For routine PCR amplification, a commercially available PCR Master Mix (Phusion™ Plus PCR Master Mix) was utilised, which contains DNA polymerase, dNTPs, Mg^{2+}, and reaction buffer at optimised concentrations. The final reaction volume for each PCR was set to 50 µL, the forward and reverse primers were used at a final concentration of 0.5 µM each (both primers were reconstituted in nuclease-free water to a stock concentration of 100 µM, a working solution was prepared by diluting the stock concentration to 10 µM, and 1 µL of each

primer was added to a 50 µL PCR reaction, resulting in a final primer concentration of 0.5 µM). DNA template concentrations were optimised for each reaction and varied based on the initial concentration of the extracted DNA. PCR cycling parameters, including temperatures and durations, were adapted based on the manufacturer's instructions. For high-fidelity polymerases, typical conditions included an initial denaturation step at 98°C for 30 seconds, followed by 30 cycles of denaturation at 98°C for 10 seconds, annealing at a primer-specific temperature for 30 seconds, and extension at 72°C for 20-30 seconds per kilobase (kb), concluding with a final extension at 72°C for 5 minutes. Conditions for Phusion Plus were the same except that a constant annealing temperature of 60 °C for 30 seconds was used in all experiments. All reactions were run alongside negative controls without DNA to check for potential contamination.

2.2.2.1.1. Gradient PCR

Gradient PCR reactions were established as outlined in section 2.2.2.1. Cycling conditions were optimised by testing a gradient of annealing temperatures to determine the optimal conditions for primer binding. The reaction volumes typically ranged from 10 to 25 µL.

2.2.2.1.2. Overlap PCR

Overlap PCR was used to join two DNA fragments that have overlapping regions, enabling the seamless assembly of constructs without the need for restriction enzymes or ligation. Equal volumes of each construct, or equimolar amounts of each oligonucleotide, were combined with 0.025 U/µL Pfu DNA polymerase, and milliQ water to 50 µL reaction. The PCR cycling conditions were set as follows: an initial denaturation at 95°C for 2 minutes, followed by 15 cycles of 95°C for 30 seconds, 60°C for 30 seconds, and 72°C for 1 minute, with a final extension at 72°C for 5 minutes and a hold at 4°C.

2.2.2.1.3. Inverse PCR

Inverse PCR was employed to amplify and modify circular plasmids, to introduce specific deletions or insertions. Reactions were prepared using Phusion Plus master mix, 50 pmol each of forward and reverse primers and water. The PCR cycling conditions were as follows: an initial denaturation at 95°C for 2 minutes, followed by 30 cycles of 95°C for 30 seconds, 60°C for 30 seconds, and 72°C for 5 minutes, with a final extension at 72°C for 5 minutes and a hold at 4°C.

2.2.2.2. Ligation (Golden Gate assembly)

Golden Gate Assembly is a specialised ligation method that simultaneously uses both digestion (via a type IIS restriction enzyme, such as BsaI) and ligation (via T4 DNA ligase) in a single reaction. This method allows for the seamless and directional assembly of multiple DNA fragments by creating specific overhangs that facilitate correct fragment order during ligation. Reactions were performed using the NEBridge Golden Gate Assembly Kit (2.1.6.5) according to the manufacturer's protocol, with slight modifications to optimise for multiple fragment assemblies. Ligation was carried out at a molar ratio of 1:3 vector to insert. The Golden Gate assembly reaction was set up by combining the following components in a single tube: 100 ng of vector, insert concentrations were calculated based on their size and molar ratios to the vector, T4 DNA ligase buffer (10X), NEBrige Golden Gate enzyme mix containing type IIS restriction enzyme BsaI-HFv2, with nuclease-free water to a final volume of 20 µL. The reaction was incubated in a thermocycler with the following program: an initial incubation at 37°C for 1 minute, followed by 30 cycles of 16°C for 1 minute, 37°C for 1 minute and then heat-inactivation at 60°C for 5 minutes.

2.2.2.3. Restriction digest

Restriction digestions were carried out using up to 1µg of DNA in a 1X digestion buffer with 10 units of the appropriate restriction enzyme (2.1.6.4) in a total reaction volume of 20 µL. Incubation and inactivation conditions were adjusted according to the specific requirements of the enzyme used.

2.2.2.4. Saturation mutagenesis

2.2.2.4.1. MAX oligonucleotide selection pools

Each oligonucleotide used in the MAX randomisation process was synthesised externally by Eurofins Genomics and provided at a concentration of 100 µM in water.

2.2.2.4.2. MAX randomisation

Master mixes prepared for the MAX randomization process were composed of 1X T4 DNA ligase buffer, combined with the NNN template oligonucleotide at a final concentration of 10 µM, MAX oligonucleotide selection pools (with the total pool concentration maintained at 100 µM and each pool at a final concentration of ≈16.67 µM), the upstream oligonucleotide of the saturated region at a final concentration of 10 µM, and the downstream oligonucleotide of the saturated region at a final concentration of 10 µM. The remaining volume, excluding T4 DNA ligase, was adjusted with milliQ water to achieve a total reaction volume of 20 µL. The reaction was incubated starting at 95°C, with the temperature reduced at a rate of 1°C per minute until

4°C was reached. T4 DNA ligase, at a concentration of 200 Weiss units/µL, was then added, and the reaction was incubated overnight at 4°C.

2.2.2.5. Nucleic acid purification

2.2.2.5.1. DNA purification including PCR products

Samples were purified using spin column purification kits Wizard® SV Gel and PCR Clean-Up System (Promega), or Zymoresearch kit (D4001T), following the manufacturers' protocols. Purified DNA was quantified using a NanoDrop™ 2000 spectrophotometer (2.2.2.5.4).

2.2.2.5.2. DNA plasmids

Plasmid DNA was isolated from bacterial cultures grown in LB broth using PureYield™Plasmid Miniprep System (Promega, A1221), following the manufacturer's instructions.

2.2.2.5.3. Gel extraction

DNA was extracted from agarose gels using the Gel Extraction Kit (QIAquick®, 28704), following the manufacturer's instructions. The concentration of the gel-extracted DNA was quantified using a NanoDrop spectrophotometer (2.2.2.5.4).

2.2.2.5.4. Nucleic acid quantification

DNA concentrations and A260/A280 ratios were measured using a Thermo Scientific NanoDropTM 2000 spectrophotometer, which initially was calibrated using a blank solution, typically 2 µL of ddH_2O. For each measurement, 2 µL of the sample was pipetted onto the pedestal of the NanoDrop. After each calibration or measurement, the pedestal was thoroughly cleaned with a tissue.

2.2.3. Electrophoresis

2.2.3.1. Agarose gel electrophoresis of DNA

Gels of varying concentrations were prepared depending on the size of the DNA to be analysed. Standard Agarose powder, LE Multi-Purpose Agarose (Geneflow) was dissolved in 1X TAE buffer (2.1.5.1). The agarose mixture was heated in a microwave until fully dissolved. Once cooled, ethidium bromide (2.1.6.2) was added to a final concentration of 1 µg/mL for nucleic acid visualisation. The solution was then poured into a casting tray with a comb to create wells and allowed to set.

Samples for analysis were pre-mixed with sample loading buffer before being loaded into the wells, alongside DNA markers/ladders (2.1.6.8) for reference. Electrophoresis was conducted in 1X TAE at 10 V/cm. Visualisation of the DNA bands was achieved through ethidium bromide staining and it was visualised using a Syngene G:BOX.

2.2.3.2. Polyacrylamide gel electrophoresis (SDS-PAGE)

Electrophoresis was conducted using hand-cast gels with the Mini-PROTEAN® Tetra Vertical Electrophoresis Cell (BioRad) and the PowerPac™ HC High-Current Power Supply (BioRad). A 12% resolving gel was first prepared according to the recipe outlined in Table 2.3. The components were added sequentially, thoroughly mixed, and poured between 0.7 mm glass plates. To ensure uniform polymerisation, the gel surface was covered with isopropanol. Once polymerised, the isopropanol layer was removed by rinsing with deionised water three times. Next, a 4% stacking gel was prepared following the recipe in Table 2.3, mixed well, and poured on top of the set resolving gel. A comb was inserted to form wells, and the gel was allowed to polymerise fully for about an hour.

For sample preparation, lysed and purified samples were prepared by mixing with 5 x Laemmli Buffer. The samples mixed with buffer were incubated at 95°C for 7 minutes to denature the proteins and then allowed to cool before being loaded onto the gel. Samples were loaded into the wells alongside the appropriate protein standard (2.1.6.8). Electrophoresis was carried out in 1X SDS-PAGE running buffer (2.1.5.8.4) at a constant voltage of 180 V until the dye front reached the bottom of the gel. The gel was then removed from the plates and stained with InstantBlue™ Protein Stain (2.1.6.2) on a rocking platform for up to 1 hour. Images of the gel were captured using a Bio-Rad imager or UV transilluminator.

Component	Resolving gel	Stacking gel
Polyacrylamide	12%	4%
Tris-HCL	0.375 M (PH:8.8)	0.126 M (PH:6.8)
SDS	0.1%	0.1%
Ammonium persulfate	0.1%	0.1%
TEMED	0.1%	0.1%

Table 2.3. Components of SDS-PAGE gels

2.2.4. Competent cells preparation

A single colony of the required bacterial strain was inoculated into 5 mL of sterile liquid growth medium. The culture was incubated overnight at 37°C with constant shaking. A suitable volume of the overnight culture was transferred to a larger flask containing fresh growth

medium to achieve an optical density (A600) of approximately 0.3–0.5. The culture was further incubated until it reached the mid-logarithmic phase.

The bacterial culture was then harvested by centrifugation at 5,000 x g for 10 minutes at 4°C. The supernatant was carefully decanted, and the cell pellet was gently resuspended in ice-cold CaCl2 solution. The resuspended cells were incubated on ice for 15 min and centrifuged again at 5,000 x g for 10 minutes at 4°C. The supernatant was discarded, and the cell pellet was gently resuspended in an equal volume of ice-cold 15% glycerol. The competent cells were distributed into small aliquots, flash-frozen on dried ice, and stored at -80°C until further use. To test transformation efficiency, a small aliquot of competent cells was thawed on ice. A control transformation with a known plasmid was performed to determine the transformation efficiency. The remaining competent cells were stored in labelled, sterile tubes at -80°C for long-term use.

2.2.5. Transformation into *E. coli*

Aliquots 50 µL of DH5α or Tuner (DE3) cells were thawed on ice, and DNA was added to the competent cells, followed by incubation on ice for 30 minutes. The tubes were then heat-shocked at 42°C for 30 seconds in a water bath and immediately placed back on ice for 5 minutes. Subsequently, the cells were suspended in pre-heated S.O.C medium (section 2.1.4.3) and incubated on a shaker at 250 rpm for 1 hour at 37°C. Following incubation, the mixture was directly added to LB broth containing the appropriate antibiotic and incubated at 37°C with shaking at 180 rpm, overnight.

2.2.6. Sequencing

2.2.6.1. Sanger sequencing

Purified plasmids at a concentration of 30-100 ng/µL were sequenced using the Pre-Mix Sanger sequencing service at Genewiz (capillary electrophoresis-based Sanger sequencing) which simplifies the preparation of sequencing samples by allowing the user to combine the template DNA and primer (5 pmol/ µl (µM)) in a single tube before submission.

2.2.6.2. Next generation sequencing (NGS)

DNA libraries were sent to Genewiz for sequencing using their Amplicon EZ service, adhering to Genewiz's sample submission guidelines (Amplicon Sequencing | GENEWIZ from Azenta). DNA samples were normalised to a concentration of 20 ng/µL with water prior to submission. Genewiz perform NGS using an IIlumina MiSeq with 2 × 250 bp paired-end sequencing.

2.2.6.2.1. NGS data analysis

Raw NGS files were processed using the Galaxy platform (available at https://usegalaxy.org/), along with the appropriate reference library sequence (in fasta format) as described (Chembath et al., 2022) In essence:

1) Paired-end reads were merged using the “fastq-join” function (default settings), which joins the reads at overlapping ends.
2) The merged sequences underwent quality control using the “filter fastq” function, filtering reads by length.
3) The quality-controlled reads were then aligned and mapped against the reference sequence using the Bowtie2 function, selecting the reference genome from “history to build index” and using default settings for the analysis mode and other options.
4) The Bowtie2 output file was then converted to SAM format using the “BAM to SAM” conversion option, selecting “exclude header” under header options.
5) The SAM file was then converted to an interval file using the “convert SAM to interval” option, selecting “yes” for the print all option.
6) The interval file was then saved as a text file by right-clicking the download button, choosing “All files” in the “save as type” dropdown, and renaming the file with a .txt extension.
7) The saved aligned data was opened in Excel using the text import wizard with “delimited” as the file type and “general” as the column data format, then saved as an .xls file.

For analysis, all columns except the one containing the aligned sequences were deleted, and sequence lengths measured using the “LEN” function in a neighbouring column. The reads were sorted by length (large to small) using the sort and filter option. The data was then ready for codon occurrence counting at individual positions by analysing in Excel file as follows:

Analysis of Excel data using delimiting

This section outlines how to use Excel's delimit function to separate randomised regions in library sequences into distinct columns. The process is guided by using the eight bases preceding each randomised region as an anchor to differentiate and organise the sequence data. The steps are as follows:

First, the "Find and Replace" feature was used to replace the eight conserved bases immediately preceding the first randomised codon with a hyphen (-). This step was repeated for each set of codons if multiple sets are present within the sequence.

Navigating to the "Data" tab, "Text to Columns was selected," and "Delimited" chosen as the data type. Under delimiters, "Other" was chosen and a hyphen (-) inserted in the box, followed

by clicking "Finish." Excel then splits the sequence into separate columns wherever a hyphen appears, placing randomised codons at the start of each new column.

The first column containing the sequences preceding the randomised region was moved to another spreadsheet. The column with the randomised sequences was highlighted. In the data preview box, the break line was selected and dragged to capture the first three nucleotides (corresponding to the first randomised codon) before clicking "Next" and "Finish." This split the data so that the first three nucleotides appeared in one column, while the remaining part of the sequence was shifted to a new column. This process was repeated for each randomised codon.

To analyse codon occurrences, a reference column listing all 64 possible codons was created. The COUNTIF function was used to tally the codons in each column of the sequence data. For example, if the codon data spans columns A-F and the reference codons are listed in column H (cells 1-64), the following formula was entered in cell J1: =COUNTIF(A:A, $H1). Enter was pressed, and the autofill handle was dragged to count occurrences of each codon across the sequence data columns.

For further data manipulation, such as organising codons by their encoded amino acids, the counted data was copied and pasted alongside the reference codons on a new sheet. Then the codons were rearranged by amino acid and a stacked histogram was generated to compare observed versus expected codon distributions. For equimolar encoding, the expected values will be the total number of sequences divided by the number of specified codons.

2.2.7. Gene Expression

DNA libraries (or individual expression vectors) were transformed into *E. coli* Tuner™(DE3) cells in LB broth containing 50 µg/mL Kanamycin. The starting culture was inoculated with transformed cells added directly to the medium and incubated overnight at 37°C (180 rpm). The overnight culture was added in 1:200 dilution to flasks containing LB and the flasks were incubated at 37°C with shaking at 220 rpm. After approximately 3 hours the OD A600 was checked using a spectrophotometer to achieve the appropriate density of 0.5-0.6. Once the cultures reached the required OD, the recombinant proteins were induced by adding IPTG (1.0 mM) and the cultures returned to incubator to allow protein expression at 20°C for 24 h. The cells then were pelleted by centrifugation at 5,000 x g and the pellets were stored at -20°C.

2.2.8. Lysate production using BugBuster

Bacterial cell pellets were resuspended in BugBuster® MasterMix (Merck Millipore) at a ratio of 5 mL per gram of cell pellet. The suspension was then incubated on a rocking platform at room temperature for 20 minutes. Following incubation, the insoluble fraction was separated by centrifugation at 16,000 x g for 20 minutes at 4°C. The resulting pellet was discarded, and the supernatant containing the soluble fraction was carefully transferred to a new tube for further experiment or else stored at -20°C until required.

2.2.9. Protein purification

2.2.9.1. Using Ni-NTA resins by immobilised metal affinity chromatography

Proteins were purified from other solubilised proteins using Ni^{2+}-NTA resins. For small-scale purification, 1 mL of the cell lysate (2.2.8) was incubated at 4°C on an end-over-end rotator with 100 µL of His-select nickel resin affinity gel (2.1.6.3) for 2 hours. The mixture was then loaded onto a 1 mL Pierce™ Spin Column (ThermoFisher Scientific, 89897), and the cap was removed to allow flow-through. The column was washed three times with wash buffer (2.1.5.8.2). The bound protein was eluted with 1 column volume at a time with elution buffer (2.1.5.8.4) four times, containing 250mM imidazole. Each fraction was collected in separate tubes, and 20 µL from each was taken for analysis by SDS-PAGE. For large-scale purification, 10 mL of the solubilised fraction was incubated overnight at 4°C on an end-over-end rotator with 1 mL of Ni-NTA Resin (which had been pre-equilibrated with equilibration buffer without imidazole).

2.2.9.2. Protein quantification

Protein samples were quantified using the Pierce™ BCA Protein Assay Kit (Thermo Scientific™, 23227) according to the manufacturer's instructions. Albumin Standard (BSA, 2 mg/mL) was diluted with phosphate buffer (100 mM, pH 7.4) to prepare assay standards within a working range of 20-2,000 µg/mL. The working reagent was prepared by mixing 50 parts of BCA Reagent A with 1 part of BCA Reagent B. For each standard or unknown sample, 25 µL were pipetted in triplicate into a microplate well (Thermo Scientific™ Pierce 96-well plate), followed by the addition of 200 µL of the working reagent to each well. The microplate was shaken for 30 seconds and incubated at 37°C for 30 minutes. Absorbance was then measured at 562 nm using the Multiskan™ GO Microplate Spectrophotometer (N10588, Thermo Scientific™).

2.2.10. Optimisation of peptide-protein binding conditions

The native SpyCatcher protein was incubated with the native SpyTag peptide under various experimental conditions to evaluate binding efficiency. A range of incubation times, temperatures, and concentrations were tested to optimise the interaction between SpyCatcher and SpyTag. SDS-PAGE analysis was subsequently performed to identify the conditions yielding the most prominent bands. The optimised conditions were as follows: 10 μM each of the peptide and protein in PBS (pH 7.4), using a 1:1 ratio (native SpyTag-native SpyCatcher), a 1:18 ratio (native SpyTag-SpyCatcher libraries), and 20 μM at 1:6 ratio (native SpyCatcher-SpyTag libraries). Incubations were carried out for 3 hours at 25°C.

2.2.11. Mass photometry

Some parts of this protocol and the included figures were adapted from a recent paper regarding mass photometry (Claasen et al., 2024).

2.2.11.1. Setting up the instruments and starting the software

First, the mass photometer was powered on and left to warm up for at least one hour before taking any measurements. Allowing the instrument time to reach a stable temperature is crucial, as skipping this step can result in mass shifts and inaccurate readings.

2.2.11.2. Protein samples and buffer preparation

Fresh phosphate-buffered saline (PBS) was prepared and filtered through 0.22 μm filters to ensure removal of particulate matter that could interfere with measurements. Protein samples were diluted to a concentration range of 50-200 nM, optimised to minimise aggregation and support high-resolution single-molecule detection in the mass photometer. All protein samples were kept on ice during preparation.

2.2.11.3. Mass photometry calibration

To ensure precision, the mass photometer (Refeyn) was calibrated at the start of each session using protein standards with known molecular weights. Calibration allowed consistent mass determination across experiments. To prepare β-Amylase as a calibrant, 20 mg of β-Amylase powder was dissolved in 3.57 mL of PBS (5% glycerol) to achieve a concentration of 5.6 mg/mL, equivalent to 100 μM (given the molecular weight of 56 kDa). To dilute the calibrant stock to 20 μM, 200 μL of the 100 μM stock was mixed with 800 μL filtered PBS (pH 7.4, at

room temperature) in a 1 mL centrifuge tube. A 20 µL sample of the freshly-diluted 20 µM β-Amylase was then used for calibration.

2.2.11.4. Experimental setup

The optical surfaces, including the plastic gasket within the 6-well sample cassette, were cleaned according to the manufacturer's guidelines to prevent sample contamination and enhance signal reliability. Initially, a small drop of immersion oil was placed on the mass photometer's objective lens and then a coverslip was placed on the Mass Photometer. The lid of the mass photometer was closed. To achieve focus, the Droplet-Dilution Find Focus option was selected in the data acquisition software. The white focus ring was checked in the bottom left of the software interface. In the Focus Control tab of the data collection software, the Up and Down buttons were used for coarse stage movement to make initial adjustments. If gaps appeared in the ring, indicating an air bubble in the immersion oil, the stage was gently moved laterally at maximum speed to remove the bubble. Note: impurities may come from the glass surface or the buffer (Figure 2.12).

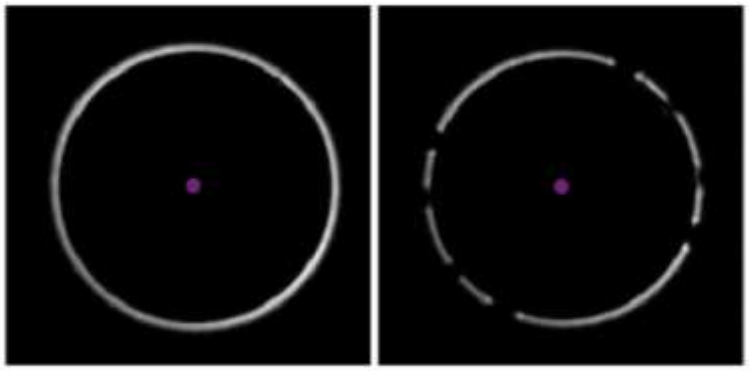


Figure 2.12. Detecting bubbles in the oil. *The examples from the data acquisition software display an intact, continuous focus ring (left) and a "broken" ring, which indicates air bubbles in the immersion oil (right).*

2.2.11.5. Sample measurement

For each measurement, an initial 18 µL of buffer was pipetted onto the coverslip to enable optimal instrument focus. Subsequently, 2 µL of each diluted protein sample was added, followed by gentle mixing to ensure homogeneity without introducing air bubbles. Once focus had been achieved, the first recording was started. To capture a 1-minute measurement, ensuring no visible impurities, the Record button was pressed, and measurements were conducted immediately at room temperature, with data acquisition lasting approximately 60 seconds per sample, to capture a comprehensive dataset. To optimise the number of

molecules landing on the measurement surface, the landing event density was checked as shown in Figure 2.13. If the density was not ideal, the sample concentration was adjusted accordingly.

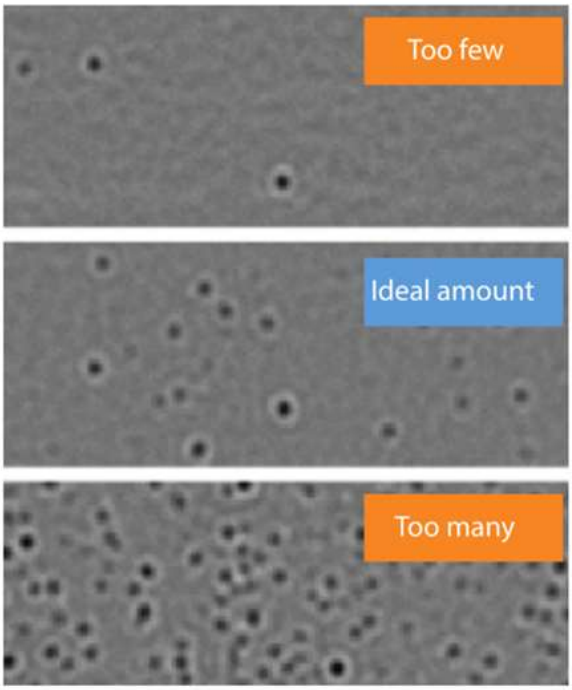


Figure 2.13. Representative examples of sample concentrations molecule landing events. *In the ratiometric view of the mass photometry image, landing molecules appear as dark spots. The optimal concentration should produce an ideal density of landing events (centre) for data acquisition. If the event density is too low (top, "Too few"), statistical analysis of the mass photometry data cannot be accurately performed. If it is too high (bottom, "Too many"), molecules overlap spatially, degrading data quality. These images were captured using the mass photometer's data acquisition software.*

2.2.11.6. Data analysis

After acquisition, the scattering signals were processed to generate histograms representing the mass distribution of the detected particles, and data analysis was subsequently initiated.In these histograms, the x-axis was configured to display molecular mass in kilodaltons (kDa), while the y-axis represented the number of molecules or detection frequency, allowing visualisation of particle mass across populations. Once the data collection was completed, the data analysis software was opened. The plus (+) icon in the top left was selected, then the .mp calibrant file was chosen to begin the analysis. Software processing could take a few minutes depending on the data size and quantity. The Create Mass Calibration button at the bottom right was then clicked. A dialog box appeared with a table showing contrast values for the fitted peaks. Note: Analysing .mp files in the data analysis software should be avoided while data acquisition is ongoing, as this may reduce data quality.

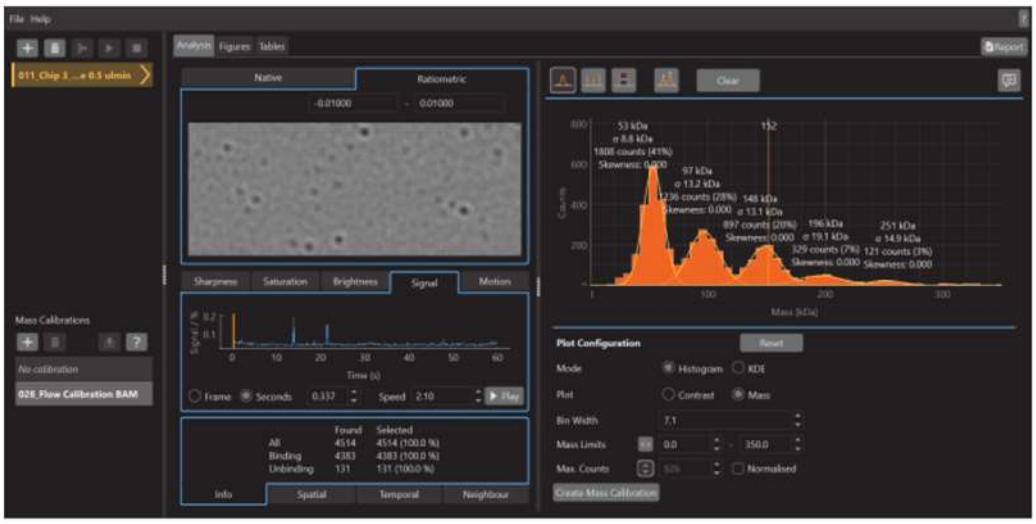


Figure 2.14. A screenshot of the data analysis software for mass photometry. *Where .mp files exported from the data acquisition software can be loaded for data analysis. Processed data can then be used to generate figures.*

To generate a mass histogram (as in Figure 2.14), the analysis tab was navigated, by selecting Histogram mode, and choosing the Mass plot option. Bin width, mass limits, and other parameters were adjusted as necessary. For additional customisation, the Figures tab was used before exporting the figure or saving the workspace as a .dmp file.

2.2.11.7. Representative results

In mass photometry experiments, mass histograms are plotted with the scattering signal strength (or "contrast") of each landing event on the x-axis, which, after calibration, is represented as molecular mass. The y-axis is used to display the count of molecules detected at each mass (or contrast). A peak is indicated as a group of molecules within the molecular weights represented on the x-axis, with the peak reflecting the size of that population. A mass histogram consistent with the expected molecular weights was produced by the diluted sample.

2.2.11.8. Essential steps for accurate graph interpretation

Distinct peaks on the histogram were observed to correspond to specific oligomeric states of the protein, such as monomers, dimers, or trimers. Each peak was carefully identified based on its mass and expected position, which was aligned with theoretical molecular weights.

The height and area of each peak were quantified by the software, and these values were correlated with the relative abundance of each oligomeric state. This quantification allowed the population ratios of different oligomers in solution to be estimated.

To convert scattering intensity to molecular mass, a calibration curve was applied. This curve was constructed using known standards and adjusted based on the expected mass range (typically 10 kDa to 500 kDa). The accuracy of the mass assignments for each peak was verified by comparing them with theoretical masses.

Background correction was repeatedly performed to ensure minimal interference from noise, particularly in low-intensity regions of the histogram. Normalisation was conducted to enable cross-experimental comparisons, ensuring that results could be consistently replicated.

The bin size (the width of each interval on the x-axis) was adjusted to affect the resolution and smoothness of the histogram. A smaller bin size provided more detailed resolution but introduced noise. Adjustments to the bin size were made to balance detail with readability and were fine-tuned based on the distribution of molecular weights in the sample.

To improve the signal-to-noise ratio, an intensity threshold was set. Low-intensity (background) signals, which may have represented noise rather than actual protein particles, were excluded. The threshold was adjusted to isolate relevant peaks by removing smaller artifacts or background noise.

Normalisation of scattering intensities was carried out to ensure consistency across multiple measurements. Scaling adjustments were applied to account for differences in laser intensity or sample concentration, enabling data from different samples to be compared on a single graph.

Background signal subtraction was performed to improve the accuracy of mass determination. This correction was refined to remove baseline noise and highlight actual peaks. The level of correction was adjusted based on the baseline noise detected in control or buffer-only measurements.

Smoothing options provided by the software were applied to reduce the appearance of jagged peaks in the histogram. Smoothing parameters were adjusted to help visualise distinct peaks for different oligomeric states, though care was taken to avoid excessive smoothing that might obscure smaller peaks or fine details.

The calibration curve, which relates scattering intensity to molecular mass, was modified as needed based on the specific mass range or calibration standards used. This curve was fine-

tuned for accurate mass estimation, particularly when the sample masses fell outside the standard range.

Sensitivity for peak detection was adjusted to identify smaller peaks corresponding to less abundant oligomers. Increasing sensitivity enabled the detection of minor populations but also increased the likelihood of detecting noise as false peaks.

The final histogram, with annotated peaks for different oligomeric states, was exported in high-resolution formats (PDF, PNG) to ensure publication-ready quality.

Chapter 3 Molecular visualisation of SpyTag and SpyCatcher

3.1. Introduction to PyMOL for molecular modeling

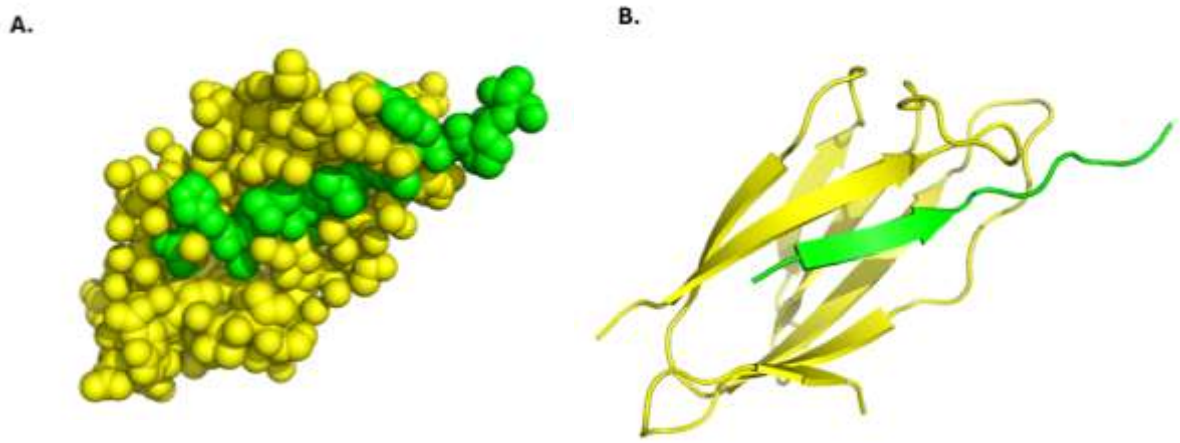
PyMOL is a widely used molecular visualisation tool that plays an essential role in computational biology and chemistry. It allows for the creation, analysis, and presentation of 3D molecular structures, providing insights into the spatial arrangement and interactions of atoms and molecules at the atomic level. Initially developed as an open-source software, PyMOL has evolved into a powerful platform for researchers to explore macromolecular structures such as proteins, nucleic acids, and ligands. PyMOL's capabilities extend beyond simple visualisation; it offers advanced features for molecular modelling, including protein-

ligand docking, mutagenesis, and simulation of molecular dynamics. Its user-friendly graphical interface and robust script-based functionality make it a versatile tool for both experimentalists and computational researchers. Importantly, PyMOL integrates with various databases such as the Protein Data Bank (PDB), allowing for the rapid import of experimental data for analysis. In the context of structural biology, PyMOL enables the exploration of protein conformations, ligand binding sites, and the effect of mutations at a high level of detail. By allowing the manipulation of molecular geometries and the real-time adjustment of parameters, PyMOL supports the design of novel compounds in drug discovery and the prediction of molecular interactions. The ability to generate publication-quality images also made PyMOL an invaluable tool for presenting structural data and communicating findings clearly and effectively.

For this study, PyMOL was used to visualise the structure and key residues of SpyCatcher and SpyTag, to examine the proposed residues for mutagenesis in terms of their positions and distances from each other.

3.2. Visualisation of the native SpyTag-SpyCatcher interaction

The structural co-ordinates of SpyCatcher-SpyTag were downloaded from the Protein Data Bank (PDB, ref 4MLI) and loaded into PyMOL (2.2.1.2). Initially, the whole complex was visualised, in space filling, ribbon and stick formats (Figure 3.1).



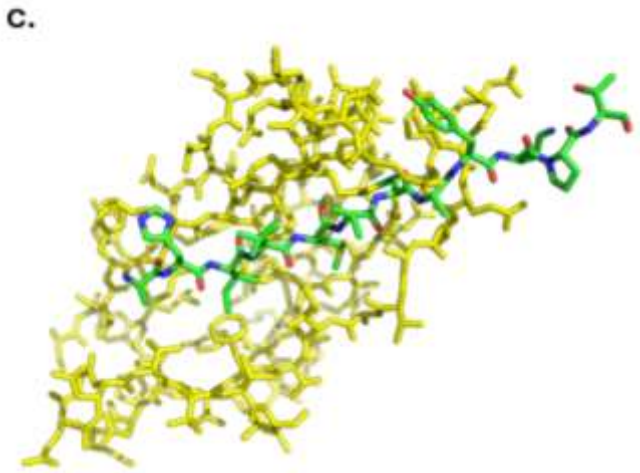


Figure 3.1. Schematic representation of SpyTag-SpyCatcher. *SpyTag is depicted in green and SpyCatcher in yellow. Representations include* ***(A)*** *space-filling****, (B)*** *ribbon-ribbon, and* ***(C)*** *stick-stick, with SpyTag colouredjf by elements for enhanced visualisation.*

SpyCatcher consists of 116 residues, only residues 21-104 are visible in structure 4MLI. Meanwhile, SpyTag peptide consists of 13 amino acids (AHIVMVDAYKPT), as shown in Figure 3.1C. The first eight residues of SpyTag engage in hydrophobic interactions with SpyCatcher (Li et al., 2014), but SpyTag forms backbone hydrogen bonds with residues 25-32 of SpyCatcher (Figure 3.2).

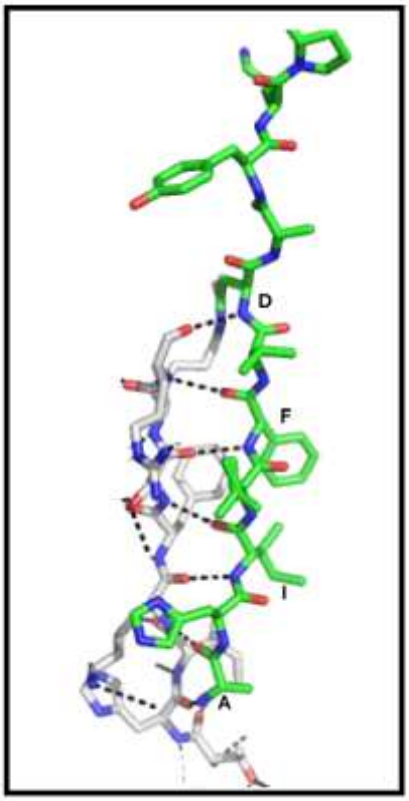


Figure 3.2. Main-Chain hydrogen bonds between SpyCatcher and SpyTag. *The main-chain hydrogen bonds between residues 25-32 of SpyCatcher and SpyTag are illustrated. SpyTag residues are shown in green, while SpyCatcher residues are represented in grey. The hydrogen bonds are depicted as black dashed lines, highlighting the interactions between the two proteins.*

3.3. Visualisation of the SpyTag and SpyCatcher residues proposed for mutagenesis

The residues proposed for mutagenesis in both SpyTag: I3 & M5 and SpyCatcher: I27, M44 & I90 were illustrated as sticks (2.2.1.4) while SpyTag-SpyCatcher were shown in transparent ribbon format (Figure 3.3).

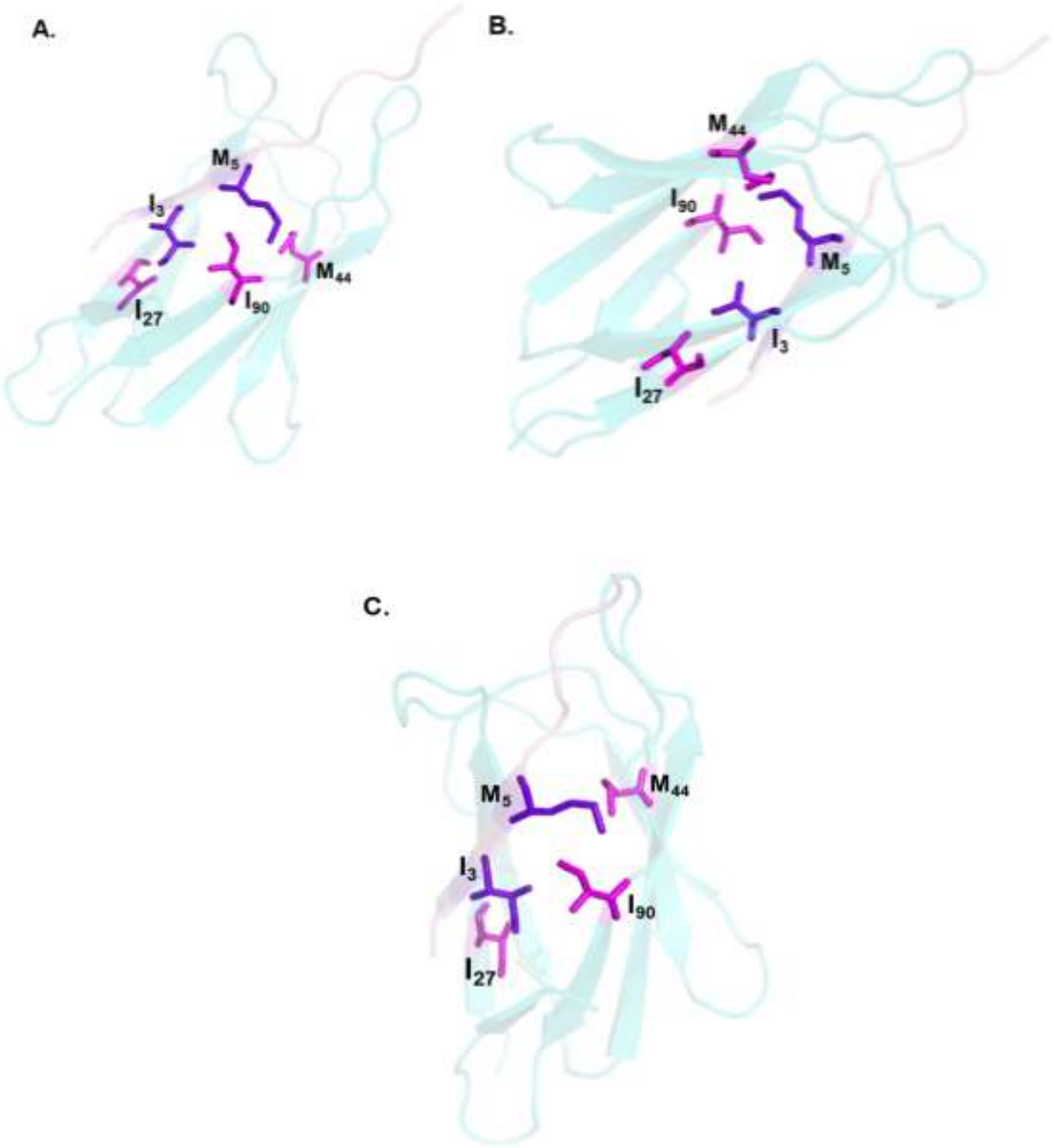


Figure 3.3. Visualisation of key SpyCatcher and SpyTag residues targeted for mutagenesis in various orientations. *SpyTag is represented as a pink cartoon, while SpyCatcher appears in cyan. The selected residues on SpyCatcher (I27, M44, and I90) are shown as pink sticks, and those on SpyTag (I3 and M5) as purple sticks, emphasising their positions within the structure.*

3.4. In silico mutagenesis

PyMOL provides the functionality to perform mutagenesis, enabling the substitution of one amino acid for another (2.2.1.8). Here, a couple of proposed residues have been visualised to demonstrate their replacement with bulky aromatic residues, as depicted in Figure 3.4.

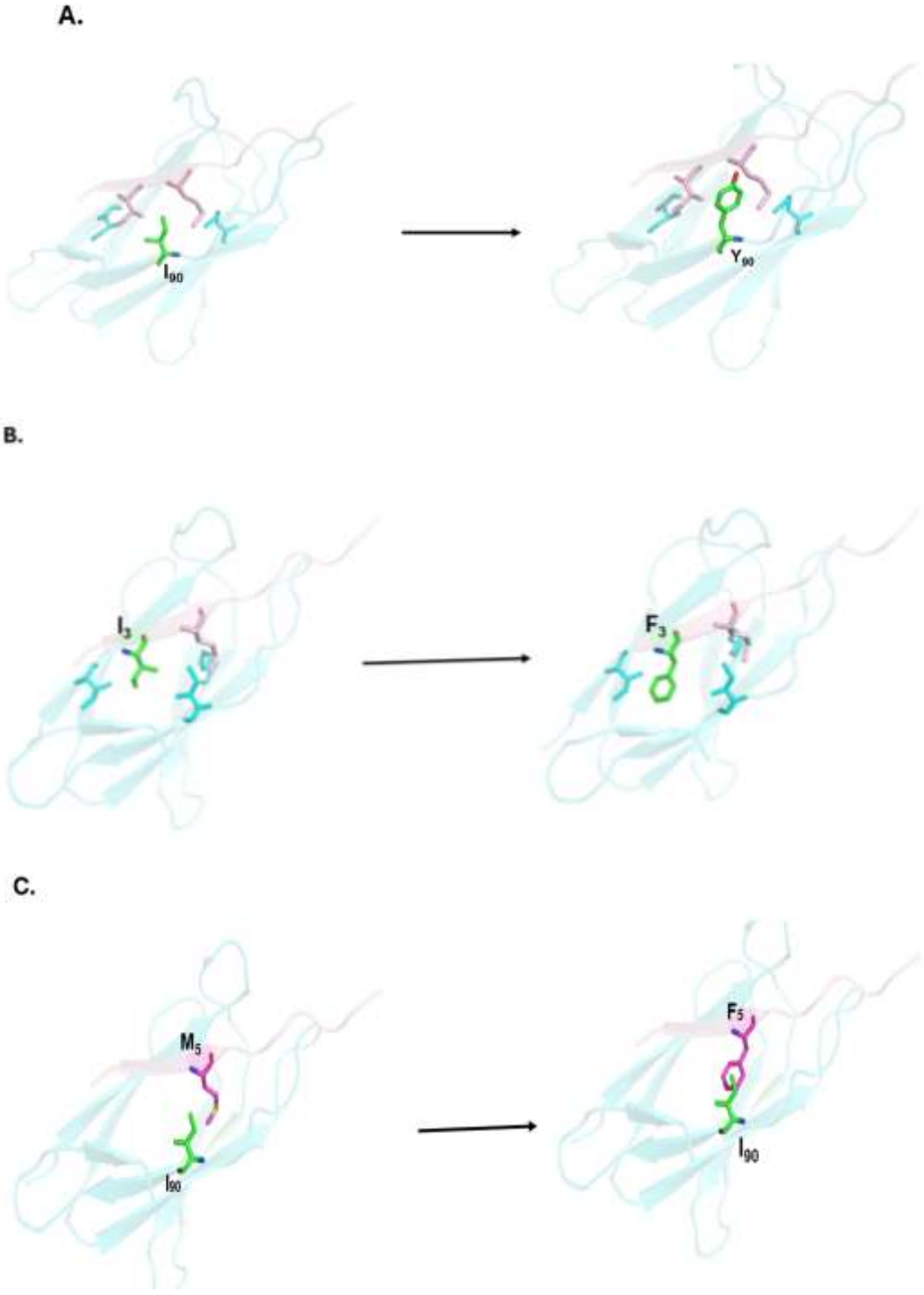


Figure 3.4. Mutagenesis in PyMOL. *SpyTag residues are shown in pink, and SpyCatcher residues are displayed in green, with the entire complex represented as a transparent cartoon.* ***(A)*** *Isoleucine on SpyCatcher mutated to tyrosine,* ***(B)*** *isoleucine on SpyTag mutated to phenylalanine, and* ***(C)*** *methionine on SpyTag mutated to phenylalanine.*

3.5. Discussion

The utilisation of PyMOL in this study provided a platform for the exploration of SpyTag-SpyCatcher interactions for the proposed mutagenesis of the key residues. By integrating data from the Protein Data Bank (PDB), specifically the 4MLI structure, the visualisation of the SpyTag-SpyCatcher complex highlighted critical regions and their contributions to molecular interactions.

The study verified proposed residues (I3 and M5) in SpyTag and (I27, M44, and I90) in SpyCatcher as potential targets for mutagenesis. The residues were visualised in PyMOL using the mutagenesis functionality to simulate amino acid substitutions, enabling the replacement of selected residues with six proposed hydrophobic residues for mutagenesis. This approach facilitated the visualisation of structural changes caused by the mutations. All substitutions were verified, and a few examples are illustrated in Figure 3.4.

As prior studies indicated that a hydrophobic environment in the binding pocket can enhance reaction rates (Li et al., 2014), the proposed residues within this pocket are highlighted in Figure 3.5. The side chains of I3 and M5 are positioned within the residues in SpyCatcher, which includes I27, M44, and I90, along with three phenylalanine residues (F27, F29, and F92) also on SpyCatcher.

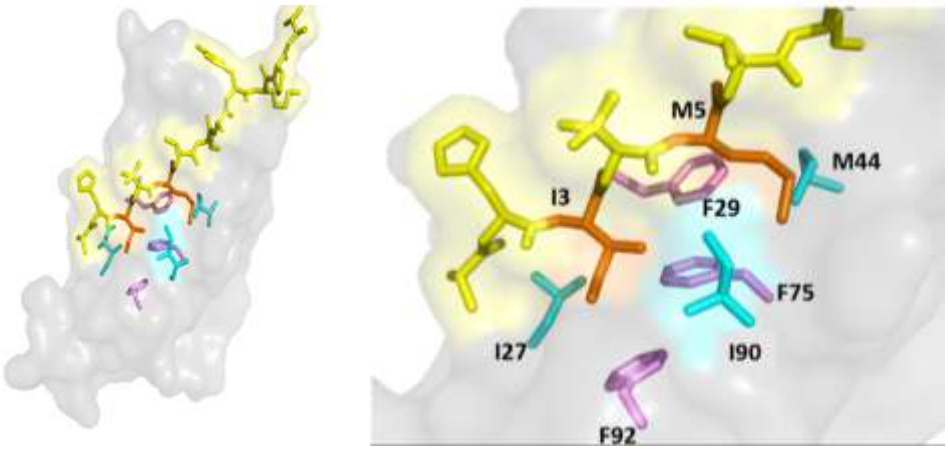


Figure 3.5. Schematic representation of the residues in the hydrophobic pocket . *SpyTag residues I3 and M5 are shown as sticks in orange, with other SpyTag residues coloured yellow. Residues of SpyCatcher that form a putative hydrophobic pocket are shown in pink (F27, F29, and F92) and cyan (I27, M44, and I90).*

In conclusion, the use of PyMOL visualisation to simulate residue mutations confirmed that the chosen residues were suitable for mutagenesis as originally proposed.

Chapter 4 SpyCatcher library construction

4.1. Creation of a SpyCatcher expression vector

To create SpyCatcher libraries, it was necessary to first clone the SpyCatcher gene in a suitable expression vector. To maximise options for future screening, it was decided to express SpyCatcher as a fusion to mNeonGreen. A SpyTag mNeonGreen plasmid had already been provided as a gift from the Regan lab (University of Edinburgh). Meanwhile, a plasmid containing the original SpyCatcher gene was provided as gift, too (University of Edinburgh). The SpyCatcher gene was amplified from the plasmid SpyCatcher-BirA using primers incorporating suitable BsaI restriction sites.

Meanwhile, the SpyTag-mNeonGreen expression vector was amplified by inverse PCR again using primers incorporating suitable BsaI restriction sites, to copy all parts of the plasmid except for the SpyTag gene (Table 9.1, Annex 1). The two PCR products were then joined by seamless cloning, via Golden Gate assembly (2.2.2.2; Figure 4.1).

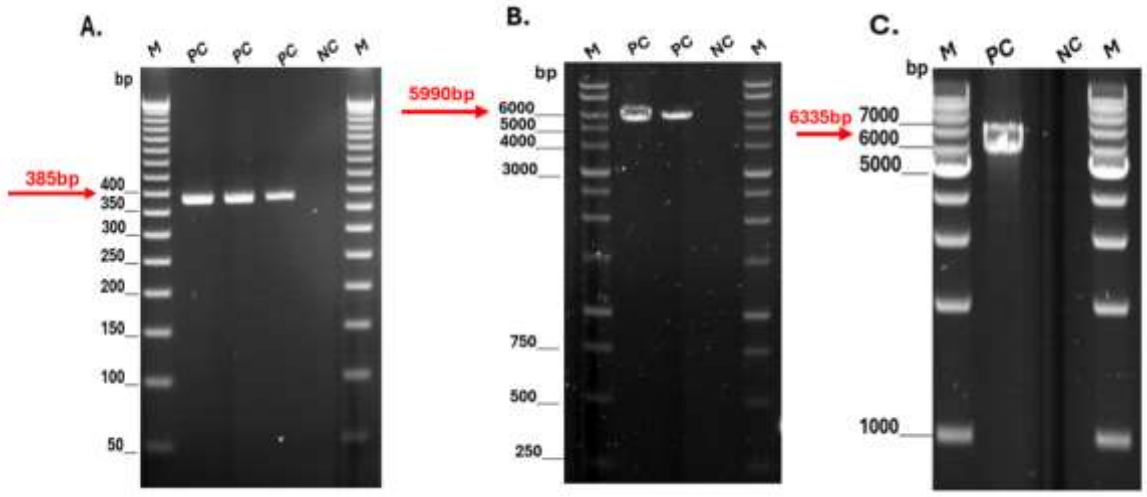


Figure 4.1. Amplification and assembly of the SpyCatcher gene. (A) *PCR amplification of SpyCatcher gene: The PCR product, amplified with Phusion enzyme, (385 bp), was electrophoresed on a 3% agarose gel and stained with ethidium bromide.* ***(B)*** *Inverse PCR Product: The resulting inverse PCR product, amplified with Phusion enzyme with an expected size of 5990 bp, was electrophoresed on a 1% agarose gel and stained with ethidium bromide. (C) Assembly of the SpyCatcher fragment: The final assembled product, expected to be 6335 bp, was analysed on a 1% agarose gel and stained with ethidium bromide. Lanes: PC positive control, NC, negative (no template) control, M, 50bp and 1Kbp ladder MW marker.*

4.2. Library design

As discussed in Chapter 3, three residues in the hydrophobic cleft of SpyCatcher (positions 27, 44 and 90) were selected for mutagenesis, with the intention of maintaining hydrophobic interactions while investigating whether alternative, orthogonal SpyCatcher-SpyTag pairs might be created. Accordingly, the hydrophobic amino acids; valine, isoleucine, leucine, methionine, phenylalanine were selected for substitution at each of the three identified positions. Tyrosine was additionally selected to evaluate whether a little polarity might also be tolerated within the interaction. Thus, in total, 216 (6 x 6 x 6) SpyCatcher mutants were envisaged.

Chapter 3 also confirms two positions for mutagenesis in SpyTag. Again, the same six substitutions were envisaged. Collectively, if both sets of mutants were made individually, this would mean creating, cloning, expressing and purifying 216 SpyCatcher mutants plus an additional 36 (6 x 6) SpyTag mutants, or 252 different mutants in total. In itself, that is a large, but possibly feasible task. However, subsequent screening of these mutants would make the endeavour impossible in the absence of robotics, since 7,776 (216 x 36) independent interaction analyses of the individual mutants would be required. Houghton's positional fixing methodology (Houghten;1991), which has been applied previously to screen positionally-fixed zinc finger protein libraries (Hughes et al., 2005), was therefore employed to design SpyCatcher libraries that could be screened initially against individual SpyTag targets.

To enable screening via positional fixing, the SpyCatcher mutants were divided into 18 sub-libraries, each encoding 36 protein variants. Specifically, the libraries were organised into three distinct sets, each comprising six sub-libraries that collectively encode all 216 SpyCatcher variants, as illustrated in Figure 4.2. For example, when fixing position 27 with a single identity, the other two positions 44 and 90, are randomised (Figure 4.2, "X") to encode all six hydrophobic amino acids. Similarly, the second set fixes position 44 while randomising positions 27 and 90 and the third set fixes position 90, while randomising positions 27 and 44 (Figure 4.2).

First-Fixed Positions			
Library	27	44	90
1	I	X	X
2	V	X	X
3	L	X	X
4	F	X	X
5	M	X	X
6	Y	X	X

Second-Fixed Positions			
Library	27	44	90
7	X	I	X
8	X	V	X
9	X	L	X
10	X	F	X
11	X	M	X
12	X	Y	X

Third-Fixed Positions			
Library	27	44	90
13	X	X	I
14	X	X	V
15	X	X	L
16	X	X	F
17	X	X	M
18	X	X	Y

Figure 4.2. 18 Fixed positional SpyCatcher libraries. *Schematic representation of the 18 fixed positional SpyCatcher libraries highlighting fixed positions at 27, 44, and 90, respectively in pink. Each row corresponds to a different library, where the fixed amino acid at position 27, 44, or 90 is indicated by a specific letter (I, V, L, F, M, Y), while 'X' represents variable positions where mixture of hydrophobic amino acids are allowed. This design enables systematic exploration of amino acid substitutions at the indicated positions to study their effects on SpyCatcher function.*

4.3. Creating fixed-position library fragments via overlap PCR

Owing to the distance between the selected residues in SpyCatcher, overlap PCR was chosen as the technique to introduce targeted mutations. Overlap extension PCR for introducing mutations into DNA sequences was initially reported in 1988. This method revolutionised site-directed mutagenesis by allowing researchers to create specific changes in DNA molecules using a polymerase chain reaction-based approach. PCR fragments (produced from primers that include mutations) with complementary 3' ends can act as primers for each other, allowing these fragments to be joined together. This method is widely employed in both degenerate and non-degenerate saturation mutagenesis, either using fragments with overlapping regions that contain the mutations, or for assembling cassettes where the mutations are located internally (Higuchi et al., 1988b).

The 18 sub-libraries illustrated in Figure 4.2 were designed to be synthesised via two mutated PCR fragments, to be joined subsequently by overlap PCR, as illustrated in Figure 4.3.

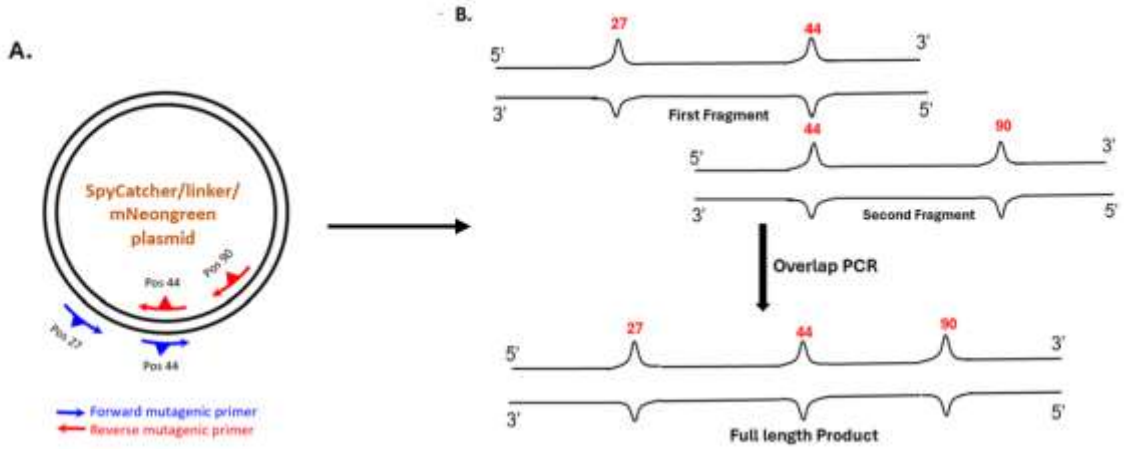


Figure 4.3. SpyCatcher library mechanism of construction. (A) *SpyCatcher plasmid, with positions 27(Pos 27), position 44(Pos 44), and position 90 (Pos 90) highlighted for mutagenesis and the arrows indicate the positions of the forward and the reverse mutagenic primers targeting these position* ***(B).****Schematic representation of overlap PCR, to produce first fragment from primer 27 forward and primer 44 reverse, second fragment from primer 44 forward and primer 90 reverse, then overlap to make the full length product from primer 27 as forward and primer 90 as reverse****.***

4.4. Library synthesis strategy

To create all 18 libraries, a mutagenic PCR grid was designed as illustrated in Figure 4.4. For example, to create the IXX library, in which position 27 was fixed as isoleucine, Fragment 1 (Figure 4.3B) was synthesised using forward primer position 27I with a mixture of all six position 44 reverse primers, in equimolar quantities (Table 9.3, Annex 1). Subsequently, Fragment 2 was synthesised using an equimolar mixture of position 44 forward primers and an equimolar mixture of position 90 reverse primers. Following verification by gel electrophoresis, the two fragments would be joined by using overlap PCR.

A.

Mutagenic PCR grids

Pos 27 forward primers \ Pos 44 reverse primers	I	V	L	F	M	Y
I	II	IV	IL	IF	IM	IY
V	VI	VV	VL	VF	VM	VY
L	LI	LV	LL	LF	LM	LY
F	FI	FV	FL	FF	FM	FY
M	MI	MV	ML	MF	MM	MY
Y	YI	YV	YL	YF	YM	YY

Pos 44 forward primers \ Pos 90 reverse primers	I	V	L	F	M	Y
I	II	IV	IL	IF	IM	IY
V	VI	VV	VL	VF	VM	VY
L	LI	LV	LL	LF	LM	LY
F	FI	FV	FL	FF	FM	FY
M	MI	MV	ML	MF	MM	MY
Y	YI	YV	YL	YF	YM	YY

B.

Fixed position libraries

Pos 27				
IXX	Overlap PCR between	Row I, 27/44	and	all 44/90
VXX	Overlap PCR between	Row V, 27/44	and	all 44/90
LXX	Overlap PCR between	Row L, 27/44	and	all 44/90
FXX	Overlap PCR between	Row F, 27/44	and	all 44/90
MXX	Overlap PCR between	Row M, 27/44	and	all 44/90
YXX	Overlap PCR between	Row Y, 27/44	and	all 44/90
Pos 44				
XIX	Overlap PCR between	Column I, 27/44	and	Row I, 44/90
XVX	Overlap PCR between	Column V, 27/44	and	Row V, 44/90
XLX	Overlap PCR between	Column L, 27/44	and	Row L, 44/90
XFX	Overlap PCR between	Column F, 27/44	and	Row F, 44/90
XMX	Overlap PCR between	Column M, 27/44	and	Row M, 44/90
XYX	Overlap PCR between	Column Y, 27/44	and	Row Y, 44/90
Pos 90				
XXI	Overlap PCR between	All 27/44	and	Column I, 44/90
XXV	Overlap PCR between	All 27/44	and	Column V, 44/90
XXL	Overlap PCR between	All 27/44	and	Column L, 44/90
XXF	Overlap PCR between	All 27/44	and	Column F, 44/90
XXM	Overlap PCR between	All 27/44	and	Column M, 44/90
XXY	Overlap PCR between	All 27/44	and	Column Y, 44/90
Where X = all 6 codons				

Figure 4.4. Library synthesis strategy. (A) *Mutagenic PCR Grids: Two grids display the combinations of forward and reverse primers used to introduce mutations at specific positions in the SpyCatcher sequence.* ***(B)*** *Fixed position libraries: The combinations generated from the mutagenic PCR grids would be joined through overlap PCR, resulting in 18 distinct fixed position libraries. Each library is named according to the specific fixed amino acids at positions 27, 44, and 90 (represented by I, V, L, F, M, Y, or X, where X indicates all six possible codons). The overlap between the rows and columns of the grids ensures coverage of all desired combinations.*

4.4.1. To fix position 27, while randomising position 44 and position 90

The process described above was iterated six times to fix position 27 alternatively as all six residues, resulting in the libraries as follows, IXX, VXX, LXX, FXX, MXX, and YXX, respectively (Figure 4.5).

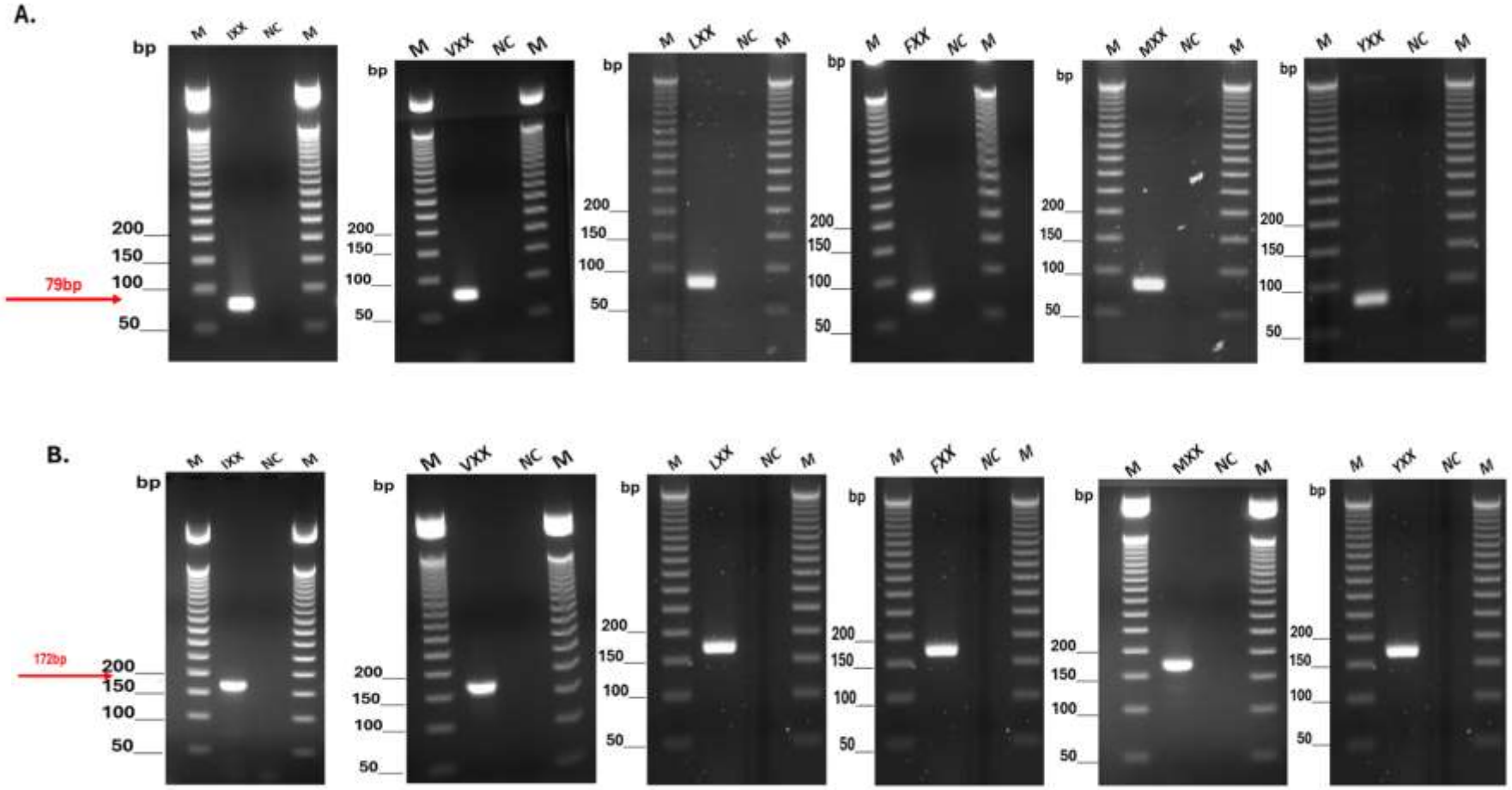


Figure 4.5. Agarose gel electrophoresis of SpyCatcher fragments of libraries for fixed position 27. *Fragments were amplified from plasmid SpyCatcher-mNeongreen with primers listed in Table 9.3. Resulting PCR reactions were electrophoresed in a 3% agarose gel and stained with ethidium bromide, where each sample represents 20% of a 50µl PCR reaction.* ***(A)*** *Fragment 1.* ***(B)*** *Fragment 2. Lanes: NXX, fixed position 27 as indicated, NC, negative (no template) control, M, 50bp ladder MW marker.*

To generate the desired overlap PCR products for constructing the fixed position libraries, Fragment 1 and Fragment 2 were each adjusted to a concentration of 15 µM and an overlap PCR reaction performed (2.2.2.1.2) and the resulting products examined by electrophoresis (Figure 4.6).

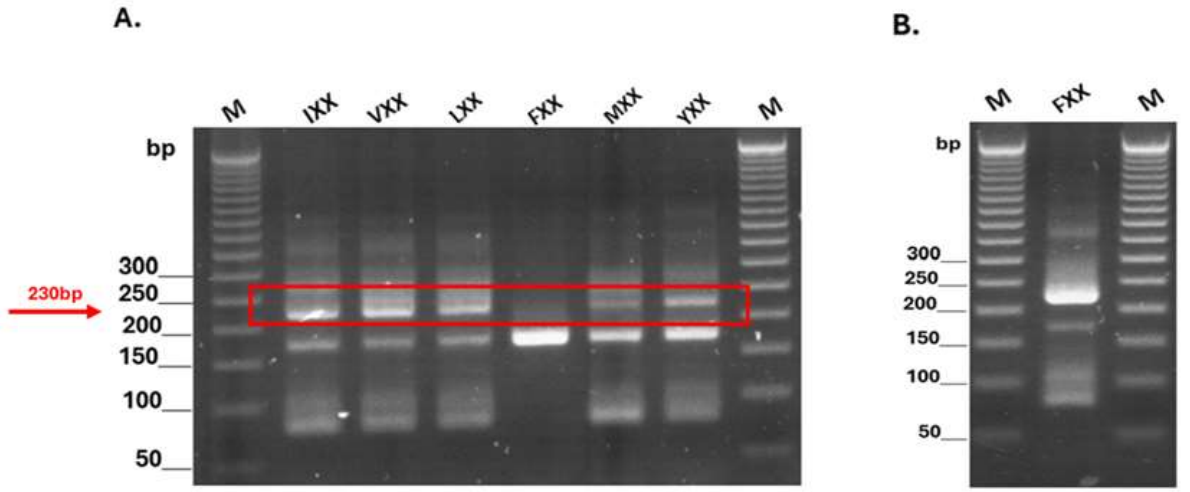


Figure 4.6. Overlap PCR of fixed position 27 libraries. (A) *Overlap PCR products were electrophoresed on a 3% agarose gel and stained with ethidium bromide. The expected size of the overlap PCR product is 230 bp, as indicated by the red box.* ***(B)*** *Repeated amplification for library FXX: A repeat of the overlap PCR for the FXX library, Lanes: first-fixed position libraries, M, 50bp ladder MW marker.*

4.4.2. To fix position 44, while randomising position 27 and position 90

To generate the second-fixed positional libraries (e.g. XIX), the Fragment 1 (Figure 4.3B) was created by performing a PCR with an equimolar blend of all six position 27 forward primers and one position 44 reverse primer (e.g. position 44I reverse). Meanwhile Fragment 2 was created using position 44I as the forward primer and a blend of all six position 90 reverse primers. The two fragments were then again assembled via overlap PCR. This process was iterated six times, with each iteration fixing one position at a time, resulting in libraries XIX, XVX, XLX, XFX, XMX, and XYX (Figure 4.7).

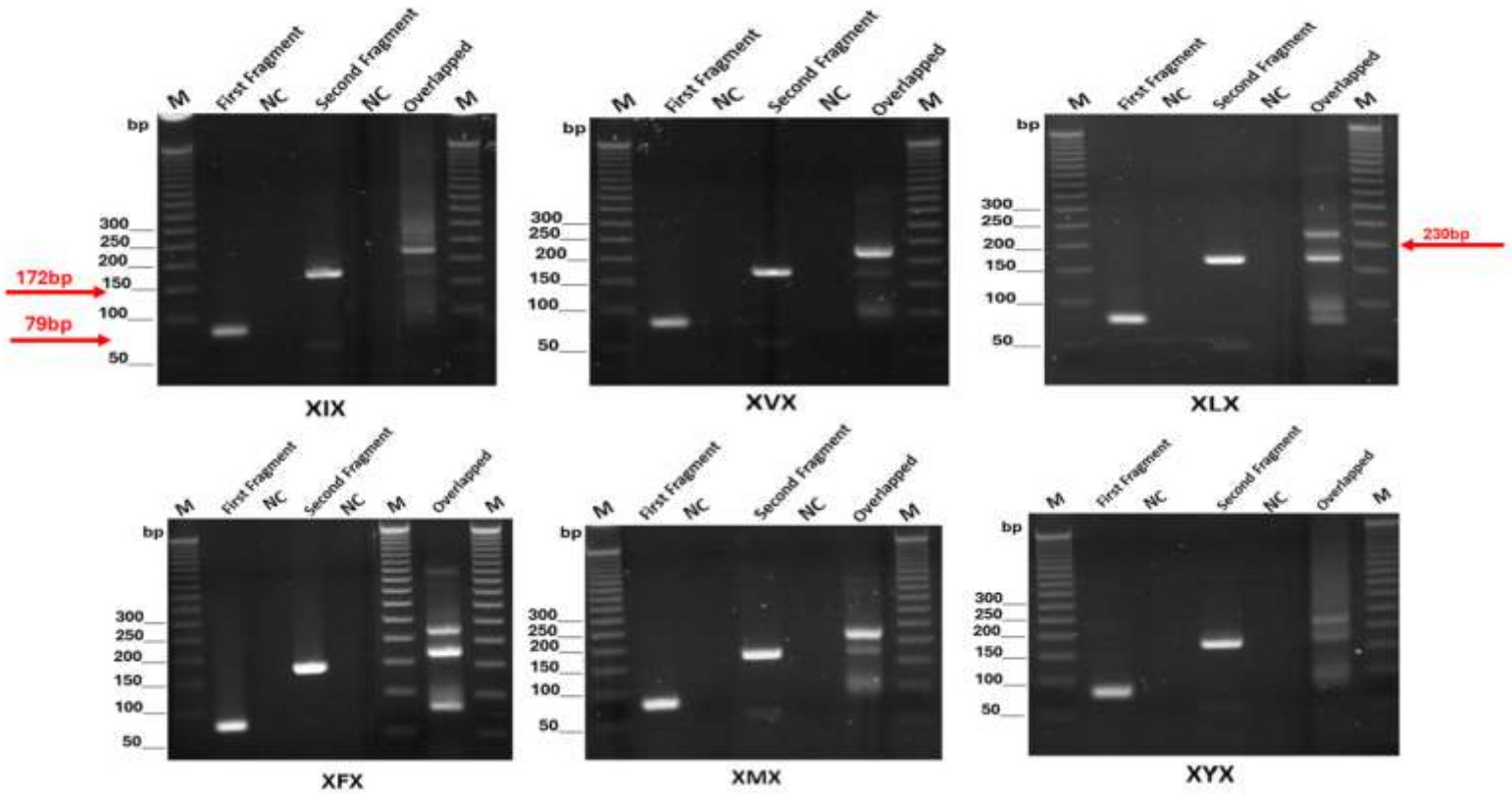


Figure 4.7. Agarose gel electrophoresis of SpyCatcher fragments of libraries for fixed position 44. *Fragments were amplified from plasmid SpyCatcher-mNeongreen with primers listed in Table 9.3. Resulting PCR reactions were electrophoresed in a 3% agarose gel and stained with ethidium bromide, where each sample represents 20% of a 50µl PCR reaction. Each gel displays Fragment 1, Fragment 2, and the corresponding overlapped products for each library. Lanes: XNX, fixed position 44 as indicated, NC, negative (no template) control, M, 50bp ladder MW marker.*

4.4.3. To fix position 90, while randomising position 27 and position 44

Finally, to generate the third-fixed positional libraries (e.g., XXI), Fragment 1 was produced using a mixture of all six position 27 forward primers and all six position 44 reverse primers, while Fragment 2 was produced using all six position 44 forward primers and a single position 90 reverse primer (Figure 4.3B). The products were examined by electrophoresis(2.2.3.1) and then assembled by overlap PCR (Figure 4.9).

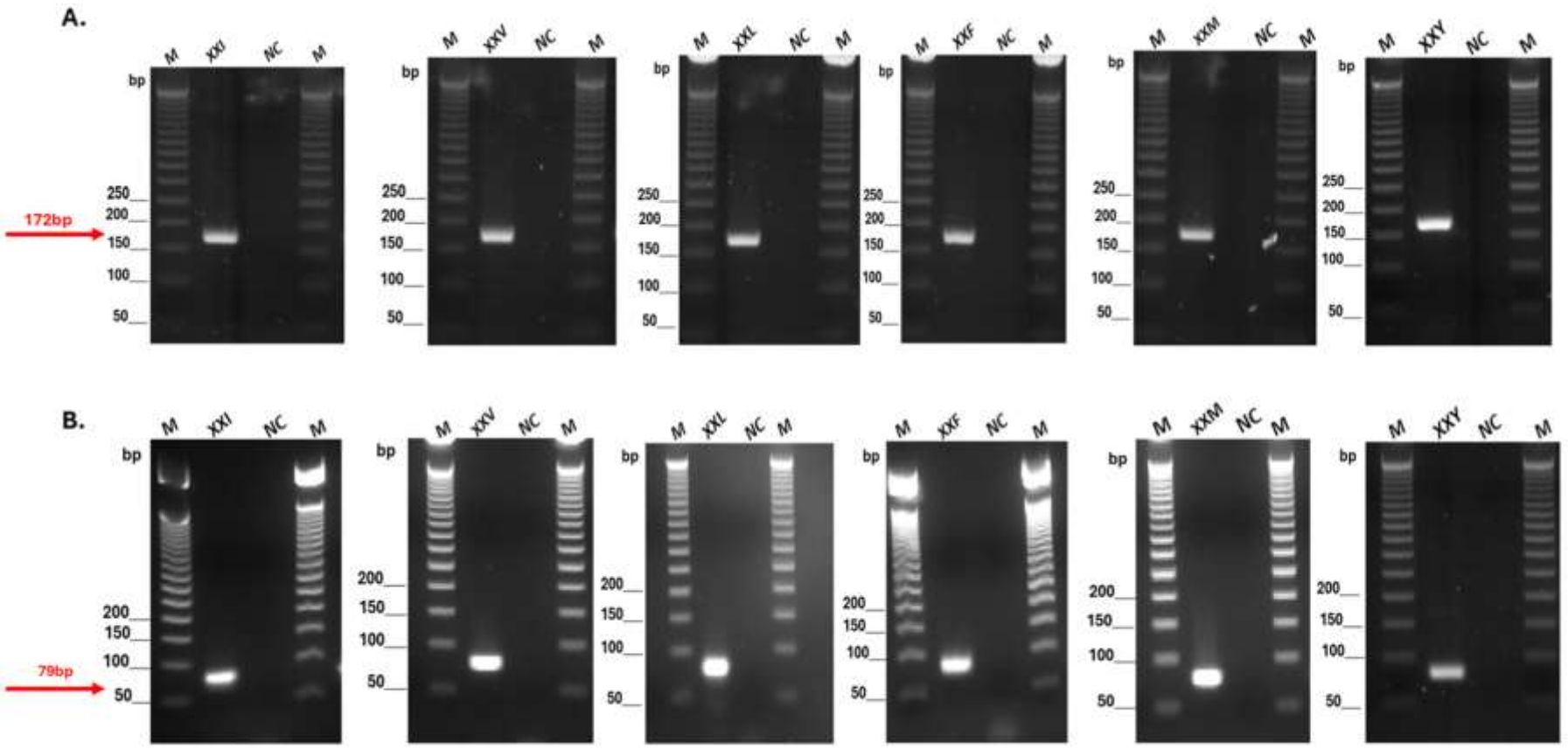


Figure 4.8. Agarose gel electrophoresis of SpyCatcher fragments of libraries for fixed position 90. *Fragments were amplified from plasmid SpyCatcher-mNeongreen with primers listed in Table 9.3. Resulting PCR reactions were electrophoresed in a 3% agarose gel and stained with ethidium bromide, where each sample represents 20% of a 50µl PCR reaction.* ***(A)*** *Fragment 1.* ***(B)*** *Fragment 2. Lanes: XXN, fixed position 90 as indicated, NC, negative (no template) control, M, 50bp ladder MW marker.*

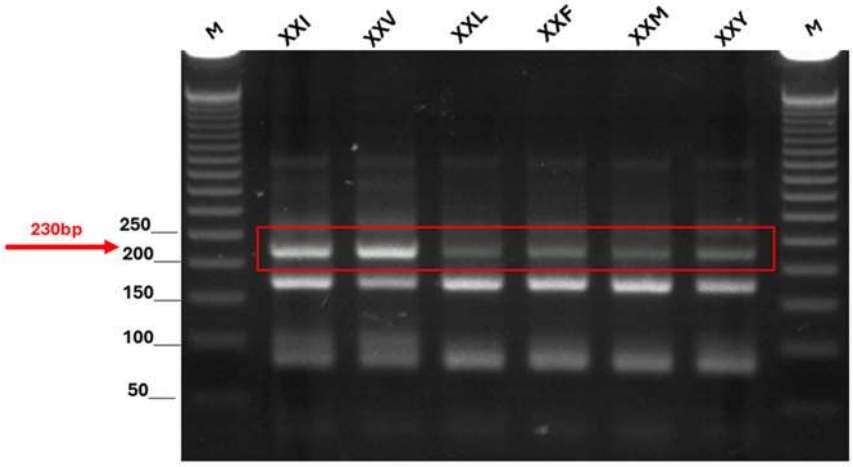


Figure 4.9. PCR Amplification for Generating Overlap Products. *The results of overlap PCR for third-fixed position libraries are shown on a 3% agarose gel stained with ethidium bromide. The expected size of the PCR product is 230 bp, as indicated by the red box. Lanes: third-fixed position libraries, M, 50bp ladder MW marker.*

4.4.4. Preparing the library cassettes for cloning

The second stage of any overlap PCR reaction is to amplify a dilution of the overlapped product with terminal primers, in order to amplify only the required full-length product. Accordingly, a 1/100 dilution of each of the 18 overlapped library cassettes (Figures 4.6, 4.7 & 4.9) was amplified with extended 27 forward and 90 reverse primers (single primers or mixtures, according to the fixed position of each library) to incorporate BsaI sites on each end of the fragment. The resulting library fragments each of expected length 252 bp were examined by 3% agarose gel electrophoresis (Figure 4.10).

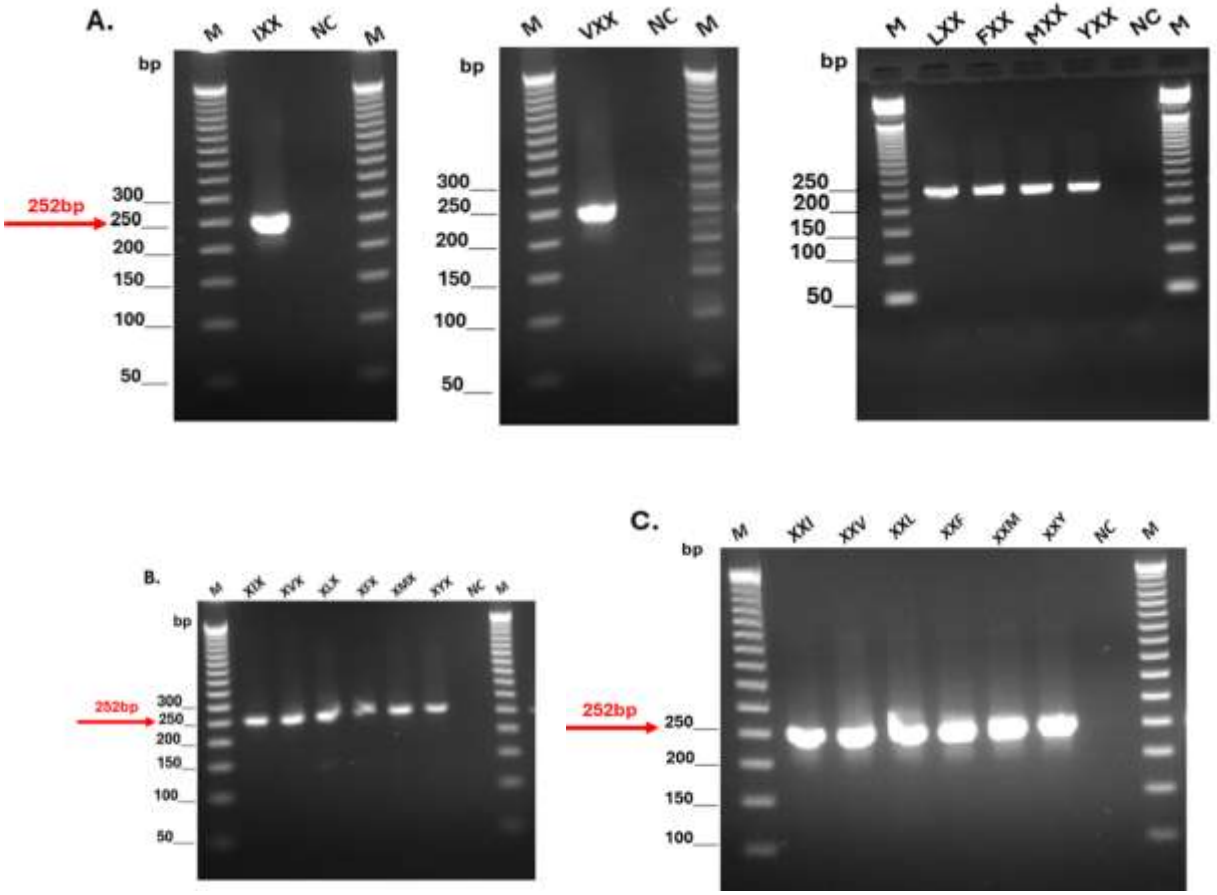


Figure 4.10. Full length PCR products of 18 SpyCatcher libraries. *PCR amplification results of 18 SpyCatcher libraries were visualised on 3% agarose gels stained with ethidium bromide. The expected full-length product is 252 bp, including overhangs generated by BsaI digestion.* ***(A)*** *Libraries with the first position fixed.* ***(B)*** *Amplifications of the second-fixed position libraries* ***(C)*** *Amplifications of the third-fixed position libraries. Lanes: PC, positive control; NC, negative control (no template); M, 50 bp molecular weight marker.*

4.5. Cloning the libraries

An inverse PCR of the SpyCatcher-mNeongreen plasmid described in section 4.1 was performed with primers containing BsaI sites (Table 9.1, Annex 1) to omit the native 252bp section of SpyCatcher. The resulting product was examined by electrophoresis (Figure 4.11).

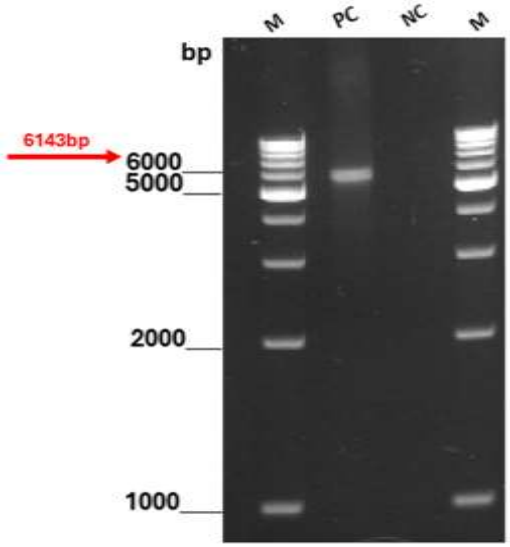


Figure 4.11. PCR Amplification of plasmid backbone for cloning. *A 1% agarose gel stained with ethidium bromide displays the results of inverse PCR, producing an overhang-containing product with the expected size of 6,143 bp. Lanes: PC, positive control; NC, negative (no template) control; M, 1 kb ladder molecular weight marker.*

Following purification of the PCR product (2.2.2.5.1) each of the 18 library cassettes was inserted into the purified plasmid backbone at a 3:1 molar ratio of inserts (libraries) to backbone using Golden Gate assembly (2.2.2.2). The resulting products were transformed into chemically competent *Escherichia coli* DH5α cells (2.2.5) and post-transformation, the cells were inoculated directly into a 50 mL starter culture containing 50 µg/mL kanamycin to selectively propagate the transformed cells. After sufficient growth in the selective medium, the bacterial cultures were harvested for plasmid DNA extraction using a mini-prep protocol (2.2.2.5.2).

4.6. Sanger sequencing of the SpyCatcher libraries

To check the insertion of the randomised DNA cassettes into the SpyCatcher-mNeongreen plasmid, each plasmid library was purified using a Zymoresearch kit (2.2.2.5.1) and sent for Sanger sequencing (2.2.6.1). It was known that Sanger sequencing would show whether or

not individual codons were randomised, but unknown as to whether peak height in a chromatogram would correspond with proportions of each base within a randomised codon. During library construction, codons ATT, GTG, CTG, TTT, ATG and TAT were selected to encode isoleucine, valine, leucine, phenylalanine, methionine, and tyrosine respectively. Accordingly, if peak height in a chromatogram were to be a reliable indication of base representation within a randomised codon, the first base should contain A:C:G:T at a ratio of 2:1:1:2, whilst the second and third positions of the codons should contain these bases at ratios 1:0:0:5 and 0:0:3:3 respectively (Figure 4.12).

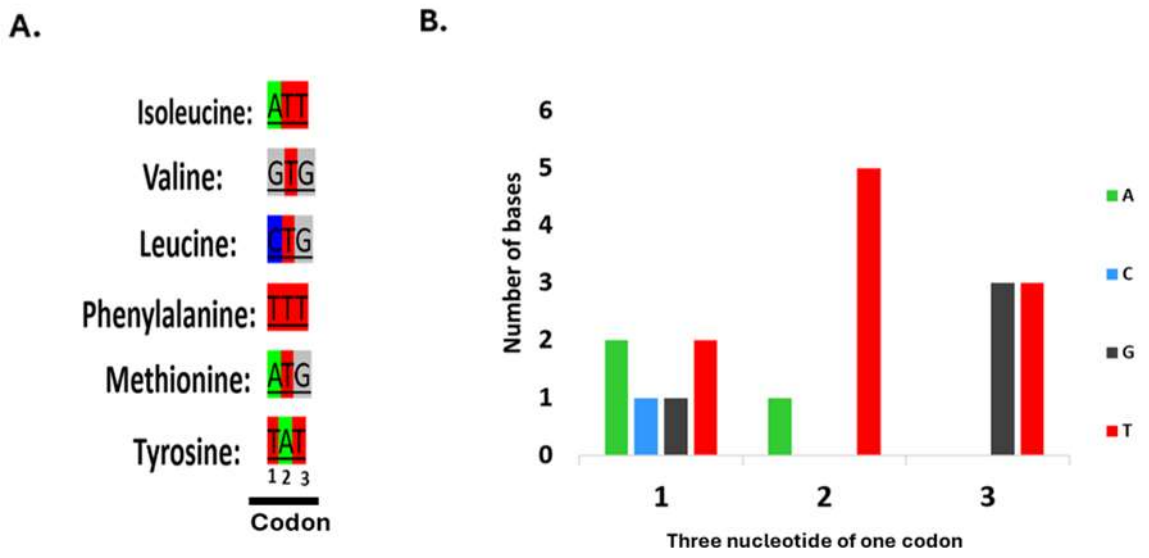


Figure 4.12. The expected distribution of bases in the randomised codons of the DNA library. (A) *According to the codons chosen to encode each of the six hydrophobic residues, two adenine, two thymine, one cytosine and one guanine in the first base of the codon were expected. In the second base, five thymine and one adenine with no cytosine and guanine were expected, while in the third base an equal number of thymine and guanine, with no adenine and cytosine were expected.* ***(B)*** *The base distribution is depicted using the Sanger sequencing colour code, where adenine (A) is green, cytosine (C) is blue, guanine (G) is black, and thymine (T) is red. This color-coding aids in the visualisation and interpretation of the sequencing data.*

The results of Sanger sequencing of all 18 SpyCatcher libraries are illustrated in Figures 4.13-4.30. Sanger sequencing clearly demonstrated codon randomisation at the required positions and in some cases demonstrated the predicted distribution of bases at each position of a randomised codon (e.g. Figure 4.16), but in others, suggested omission of base A (in particular) at the second position of the codon (e.g. Figure 4.25, position 44).

4.6.1. Observed verses expected distribution of bases of first-fixed libraries

Sanger sequencing of the first-fixed position libraries was performed to verify the expected distribution of fixed and randomised positions. Sequencing results indicated high-quality data. At position 27, the codon was fixed as anticipated. Additionally, positions 44 and 90 showed a randomised nucleotide distribution, with all possible bases represented at approximately equal frequencies, consistent with the theoretical randomisation design. All the Sanger sequencing results of the first-fixed libraries are depicted below (Figure 4.13, Figure 4.14, Figure 4.15, Figure 4.16, Figure 4.17 and Figure 4.18).

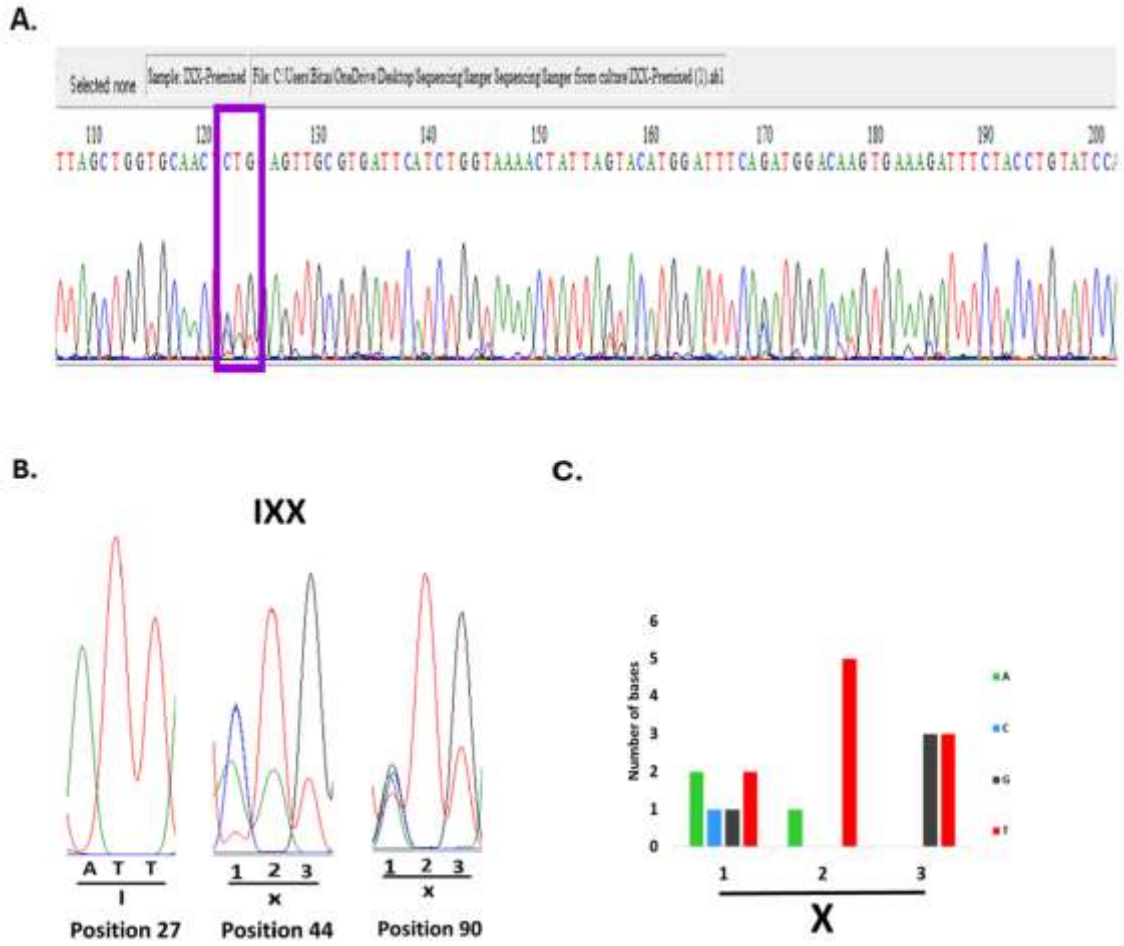


Figure 4.13. Observed vs. expected result of Sanger sequencing of the IXX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 44 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the IXX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 27, the sequencing data indicated that codon ATT was fixed to encode isoleucine. In contrast, positions 44 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presenting in the right chart* ***(C).*** *At position 90, the green peak corresponding to A in the second base of the codon is absent or not observed.*

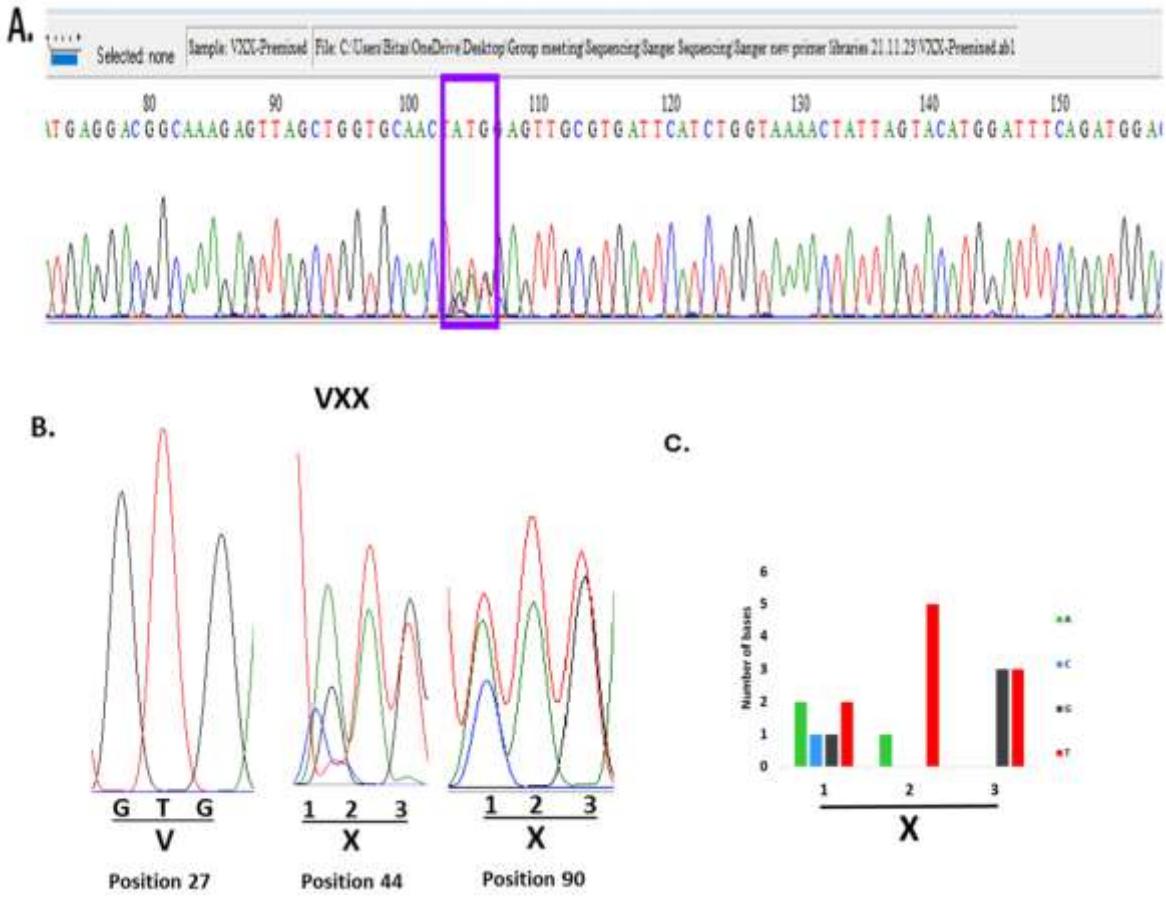


Figure 4.14. Observed vs. expected result of Sanger sequencing of the VXX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised positions 44 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the VXX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 27, the sequencing data indicated that valine was fixed to encode the codon GTG. In contrast, positions 44 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

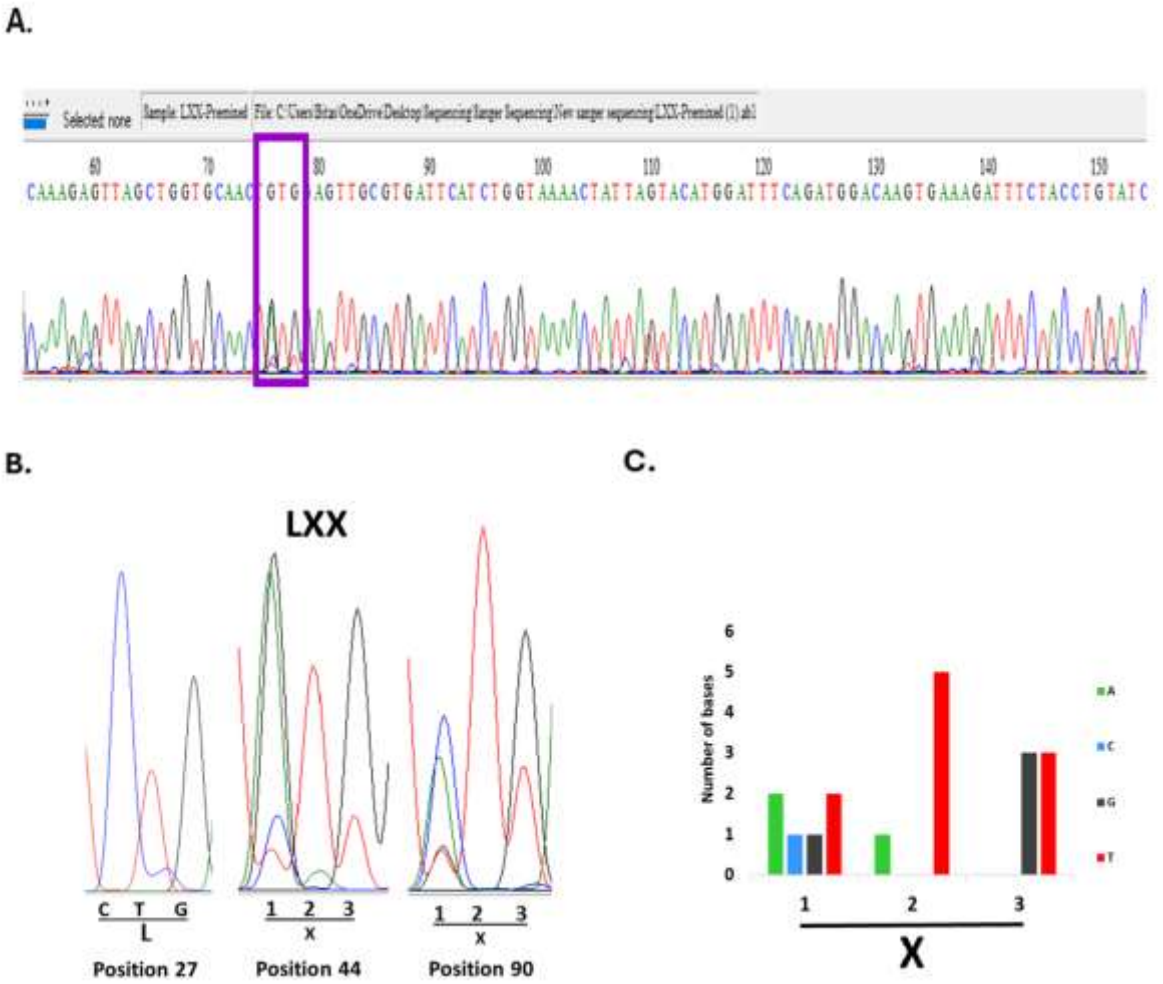


Figure 4.15. Observed vs. expected result of Sanger sequencing of the LXX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 44 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the LXX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 27, the sequencing data indicated that codon CTG was fixed to encode leucine. In contrast, positions 44 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presenting in panel* ***(C).*** *At position 90, the green peak corresponding to A in the second base of the codon is absent or not observed.*

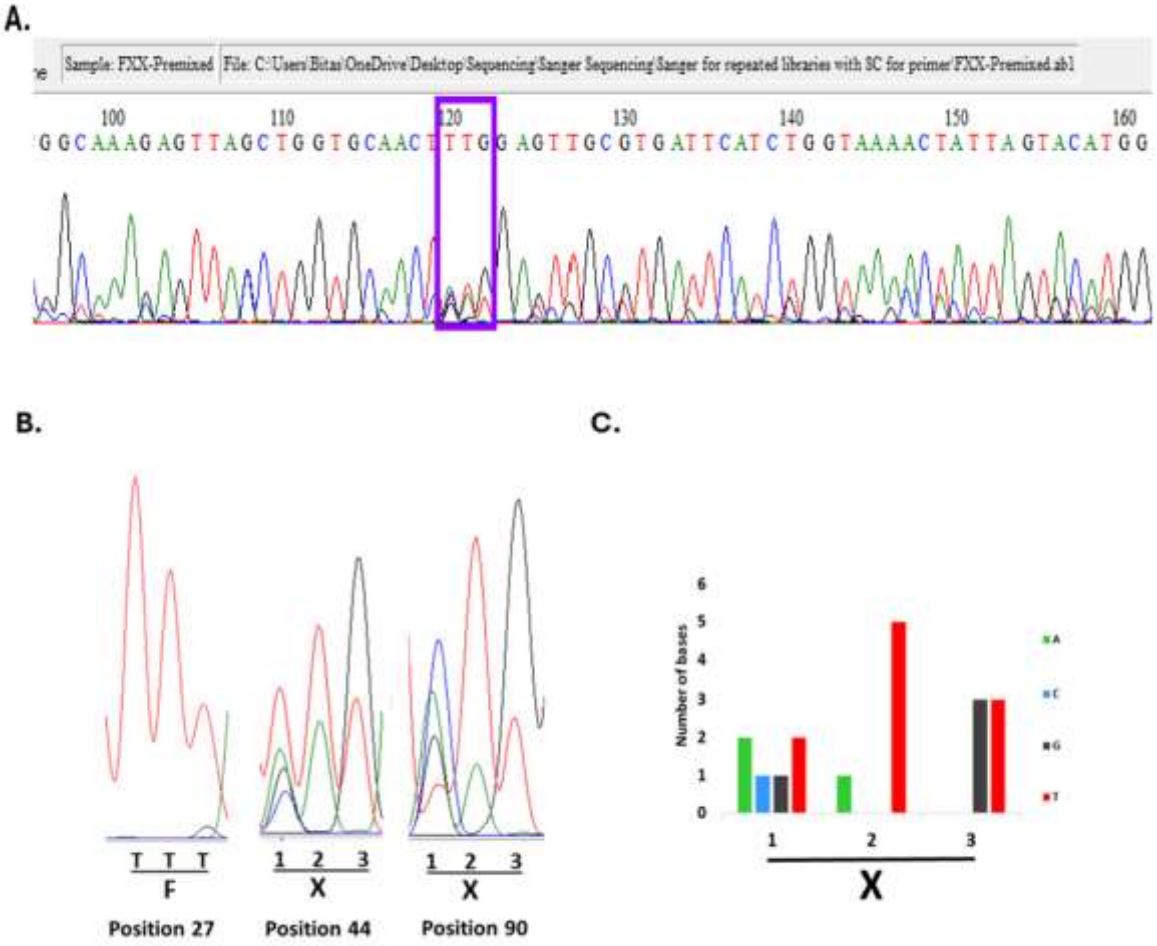


Figure 4.16. Observed vs. expected result of Sanger sequencing of the FXX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised positions 44 is highlighted with a purple box.* ***(B****) Sanger sequencing result (observed) for the FXX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 27, the sequencing data indicated that codon TTT was fixed to encode phenylalanine. In contrast, positions 44 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

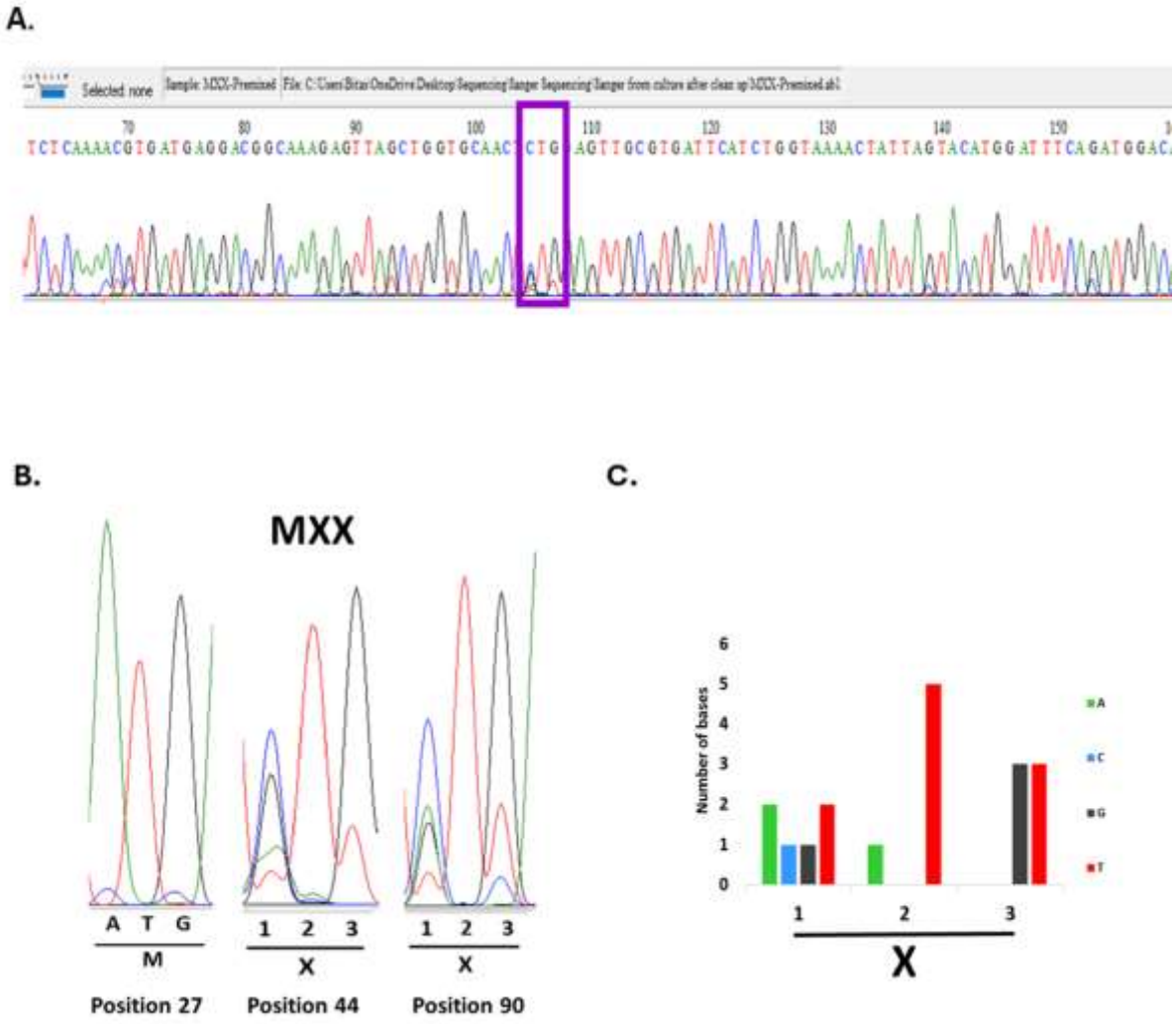


Figure 4.17. Observed vs. expected result of Sanger sequencing of the MXX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 44 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the MXX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 27, the sequencing data indicated that codon ATG was fixed to encode methionine. In contrast, positions 44 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).*** *At position 44 and 90, the green peaks corresponding to A in the second base of the codon are absent or not observed.*

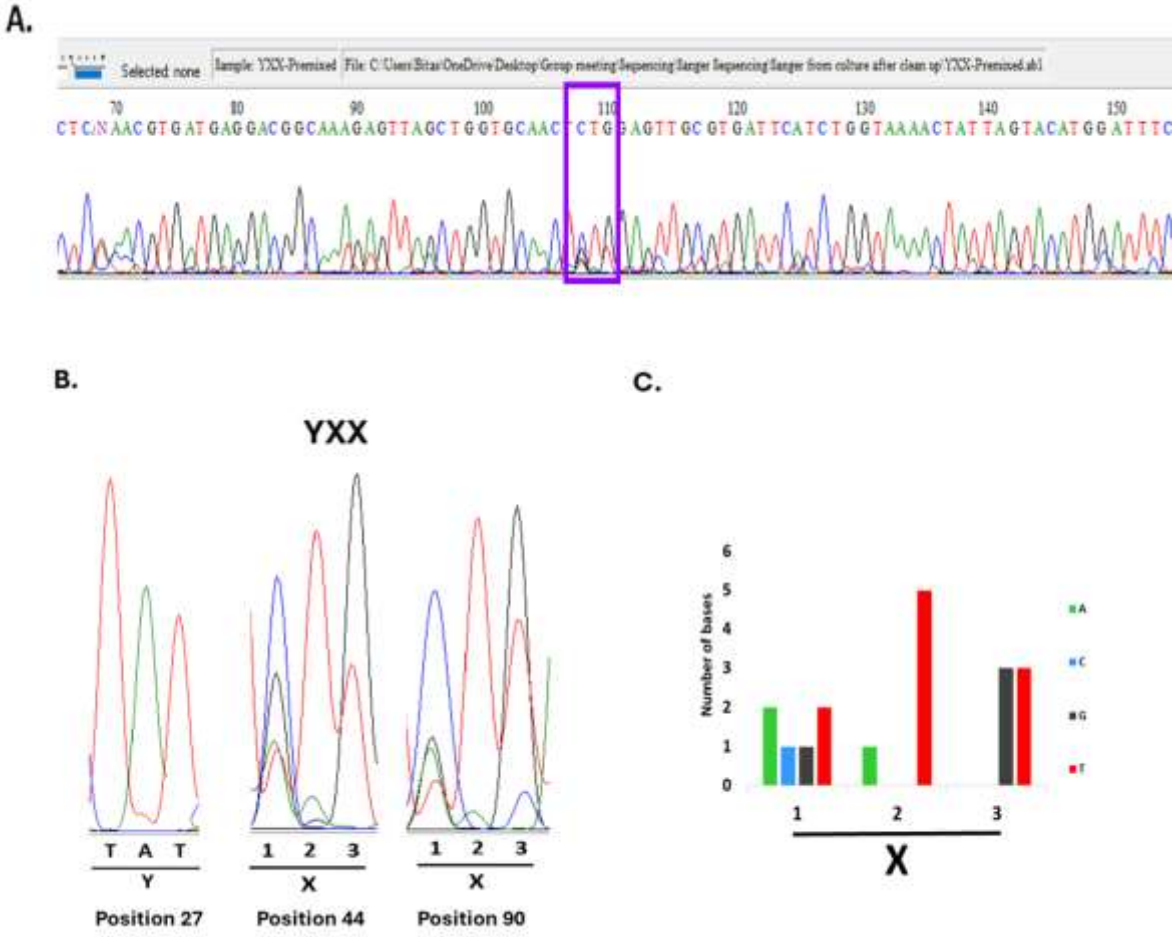


Figure 4.18. Observed vs. expected result of Sanger sequencing of the YXX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 44 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the YXX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 27, the sequencing data indicated that codon TAT was fixed to encode tyrosine. In contrast, positions 44 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

4.6.2. Observed verses expected distribution of bases of second-fixed libraries

Sanger sequencing of the second-fixed position libraries was performed to verify the expected distribution of fixed and randomised positions. Sequencing results indicated high-quality data. At position 44, the codon was fixed as anticipated. Additionally, positions 27 and 90 showed a randomised nucleotide distribution, with all possible bases represented at approximately equal frequencies, consistent with the theoretical randomisation design. All the Sanger sequencing results of the second-fixed libraries are depicted below (Figure 4.19, Figure 4.20. Figure 4.21, Figure 4.22, Figure 4.23 and Figure 4.24).

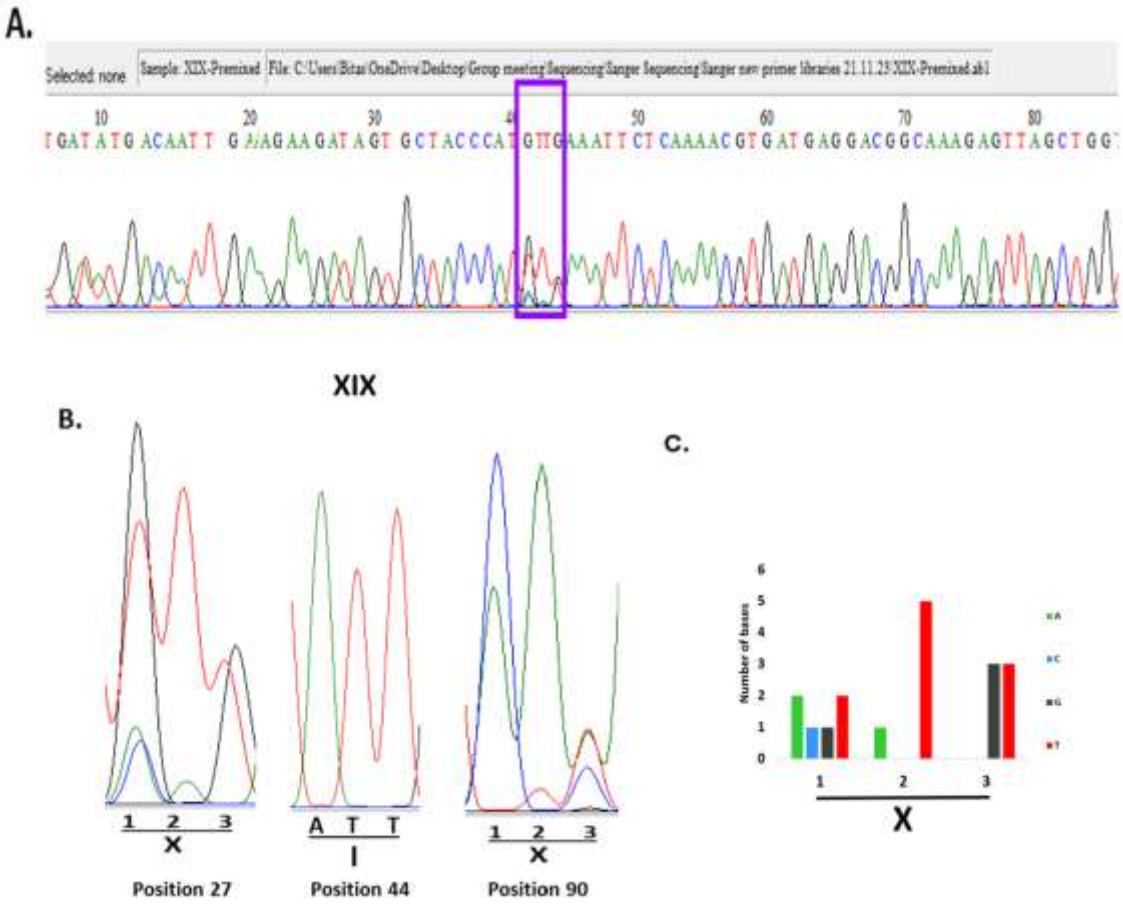


Figure 4.19. Observed vs. expected result of Sanger sequencing of the XIX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised positions 27 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the XIX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 44, the sequencing data indicated that codon ATT was fixed to encode isoleucine. In contrast, positions 27 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

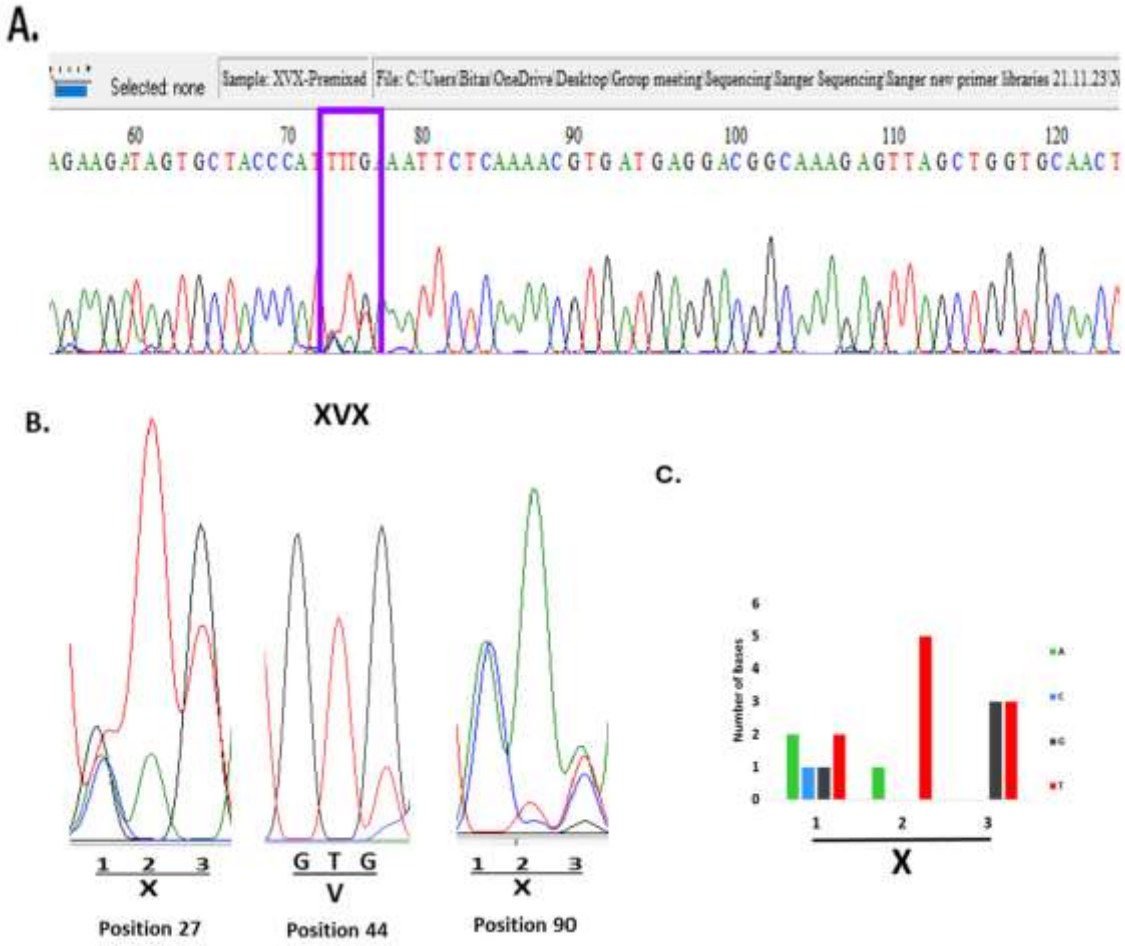


Figure 4.20. Observed vs. expected result of Sanger sequencing of the XVX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 27 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the XVX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 44, the sequencing data indicated that codon GTG was fixed to encode valine. In contrast, positions 27 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).*** *At position 27, the green peak corresponding to A in the second base of the codon is absent or not observed.*

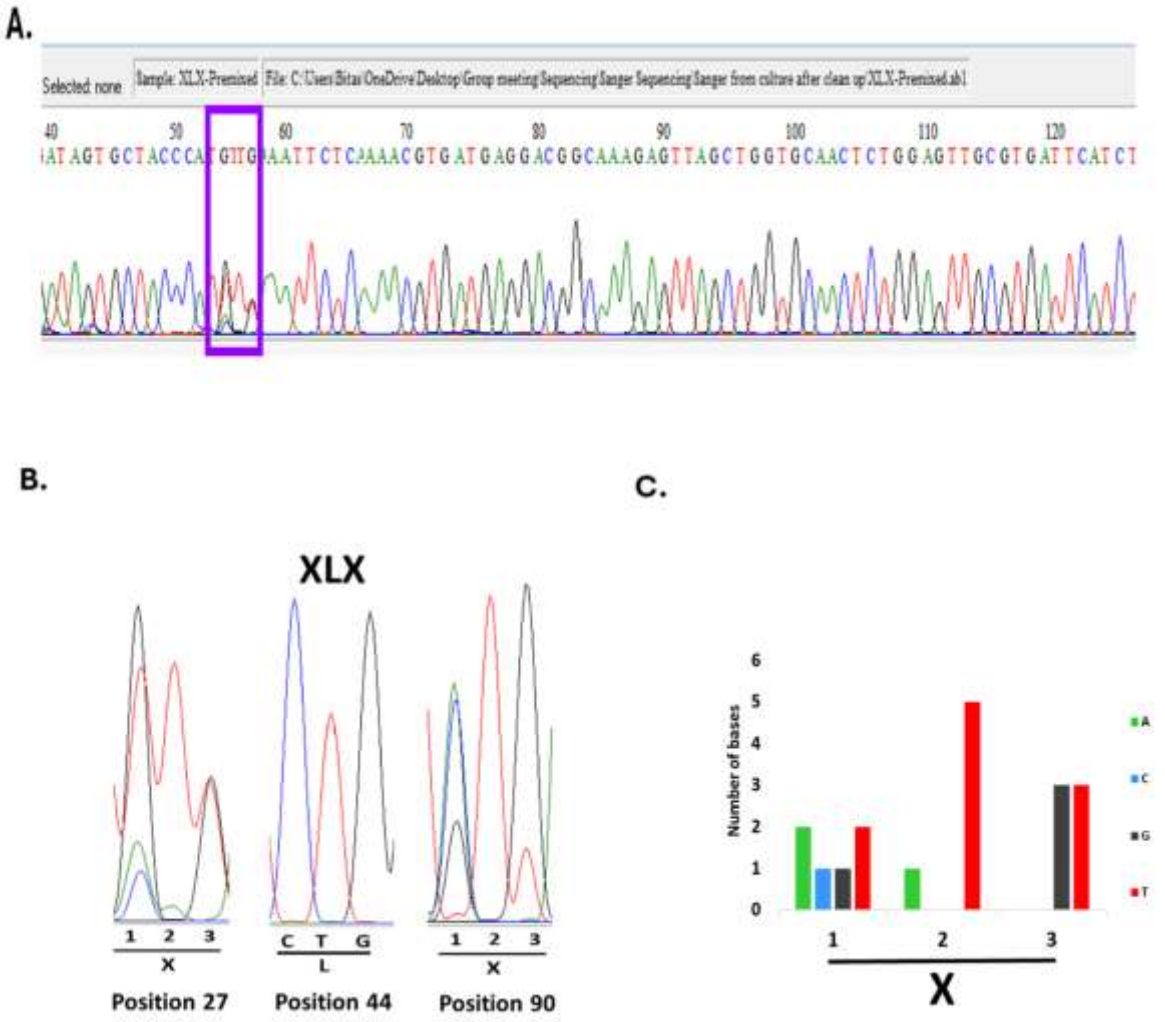


Figure 4.21. Observed vs. expected result of Sanger sequencing of the XLX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 27 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the XLX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 44, the sequencing data indicated that codon CTG was fixed to encode leucine. In contrast, positions 27 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).*** *At position 90, the green peak corresponding to A in the second base of the codon is absent or not observed.*

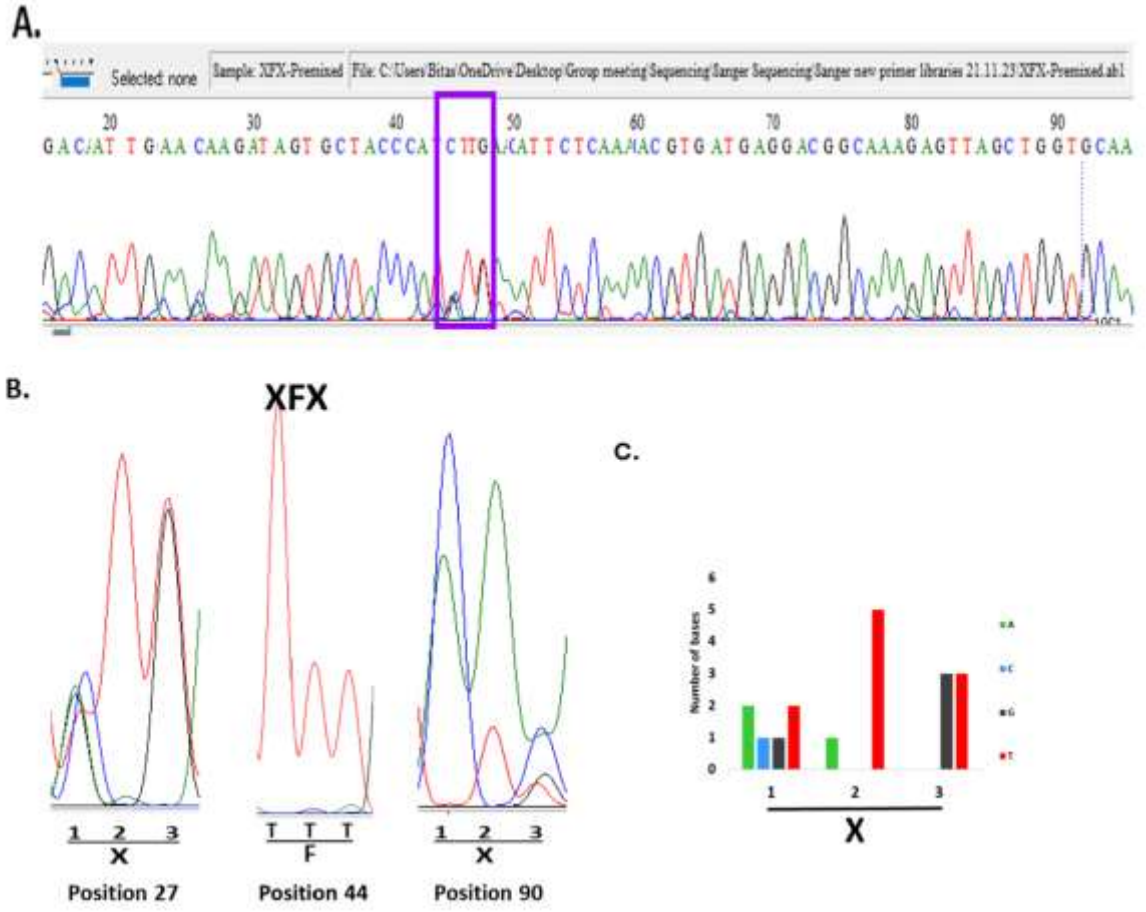


Figure 4.22. Observed vs. expected result of Sanger sequencing of the XFX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 27 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the XFX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 44, the sequencing data indicated that codon TTT was fixed to encode phenylalanine . In contrast, positions 27 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

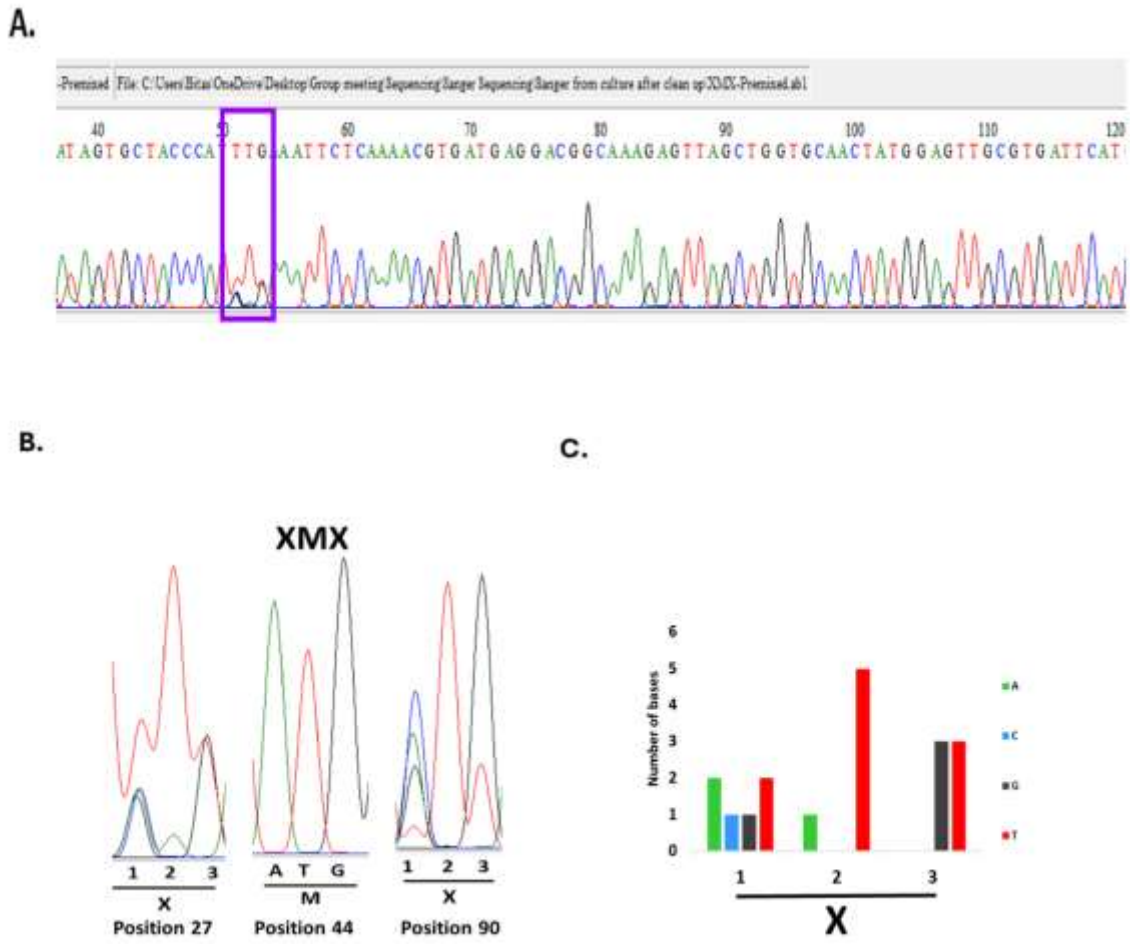


Figure 4.23. Observed vs. expected result of Sanger sequencing of the XMX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 27 is highlighted with a purple box.* ***(B)*** *Sanger sequencing result (observed) for the XMX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 44, the sequencing data indicated that codon ATG was fixed to encode methionine. In contrast, positions 27 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).*** *At position 90, the green peak corresponding to A in the second base of the codon is absent or not observed.*

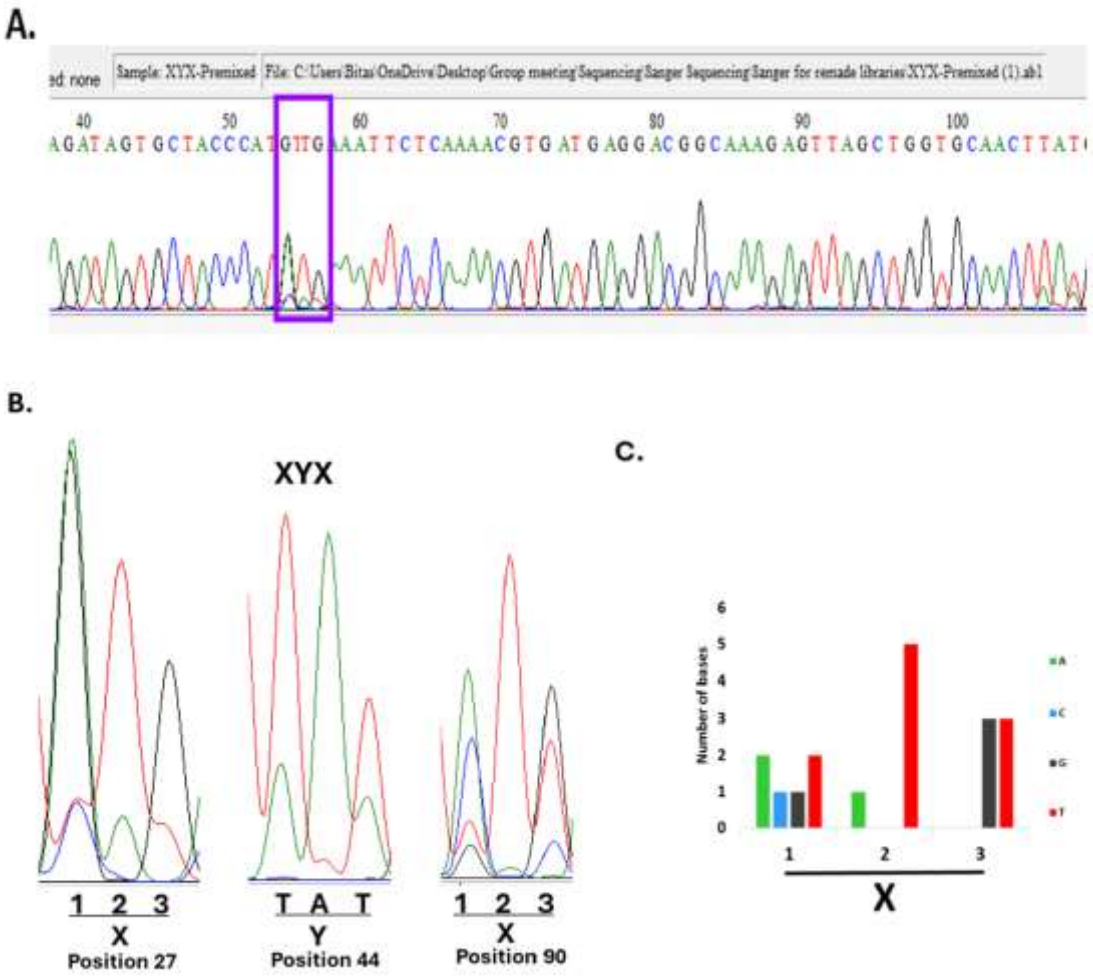


Figure 4.24.Observed vs. expected result of Sanger sequencing of the XYX library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. The randomised position 27 is highlighted with a purple box.* ***(B****) Sanger sequencing result (observed) for the XYX library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 44, the sequencing data indicated that codon TAT was fixed to encode tyrosine. In contrast, positions 27 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

4.6.3. Observed verses expected distribution of bases of third-fixed libraries

Sanger sequencing of the third-fixed position libraries was performed to verify the expected distribution of fixed and randomised positions. Sequencing results indicated high-quality data (Figure 4.25, Figure 4.26, Figure 4.27, Figure 4.28, Figure 4.29 and Figure 4.30). At position 90, the codon was fixed as anticipated. Additionally, positions 27 and 90 showed a randomised nucleotide distribution, with all possible bases represented at approximately equal frequencies, consistent with the theoretical randomisation design.

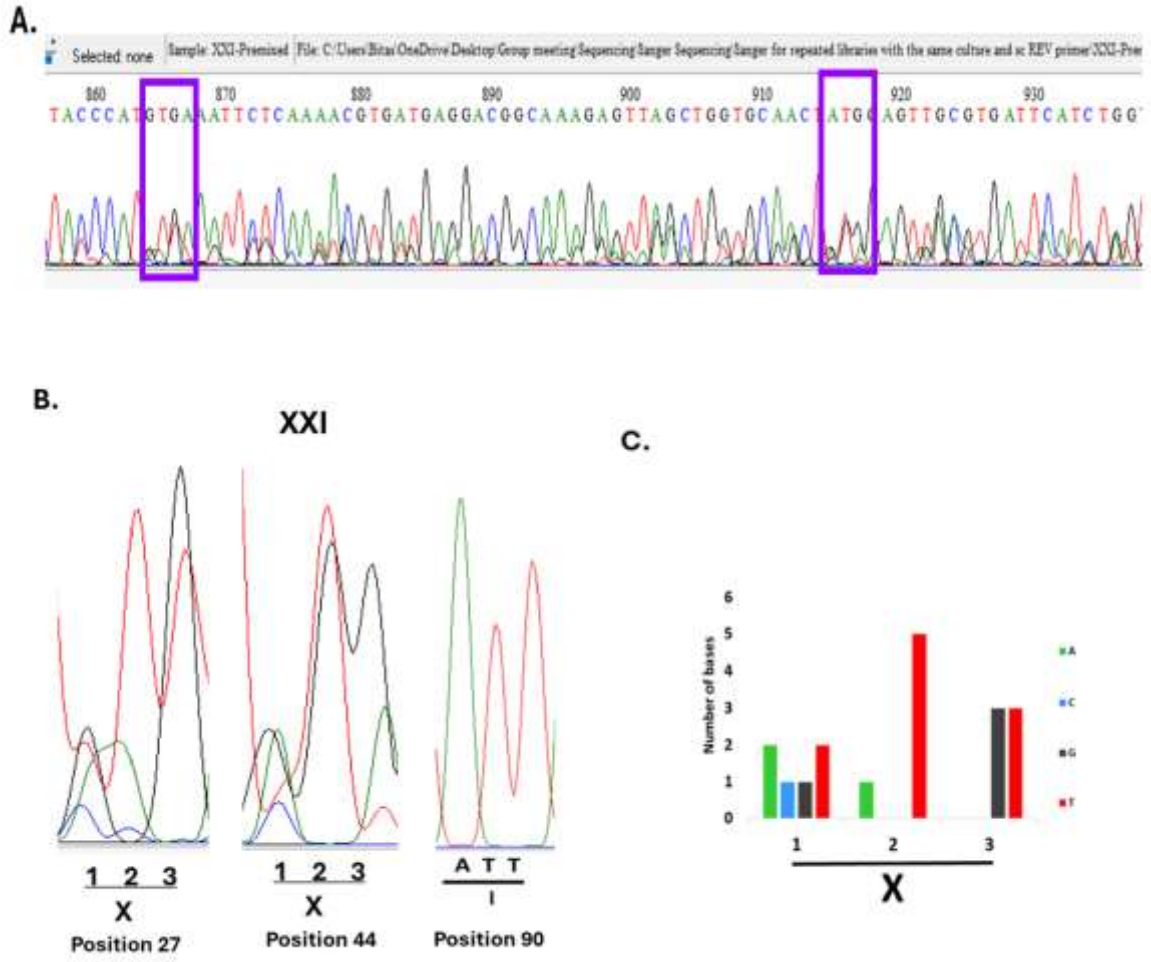


Figure 4.25. Observed vs. expected result of Sanger sequencing of the XXI library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. Both randomised positions are highlighted with purple boxes.* ***(B)*** *Sanger sequencing result (observed) for the XXI library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 90, the sequencing data indicated that codon ATT was fixed to encode isoleucine. In contrast, positions 27 and 44 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C****). At position 44, the green peak corresponding to A in the second base of the codon is absent or not observed.*

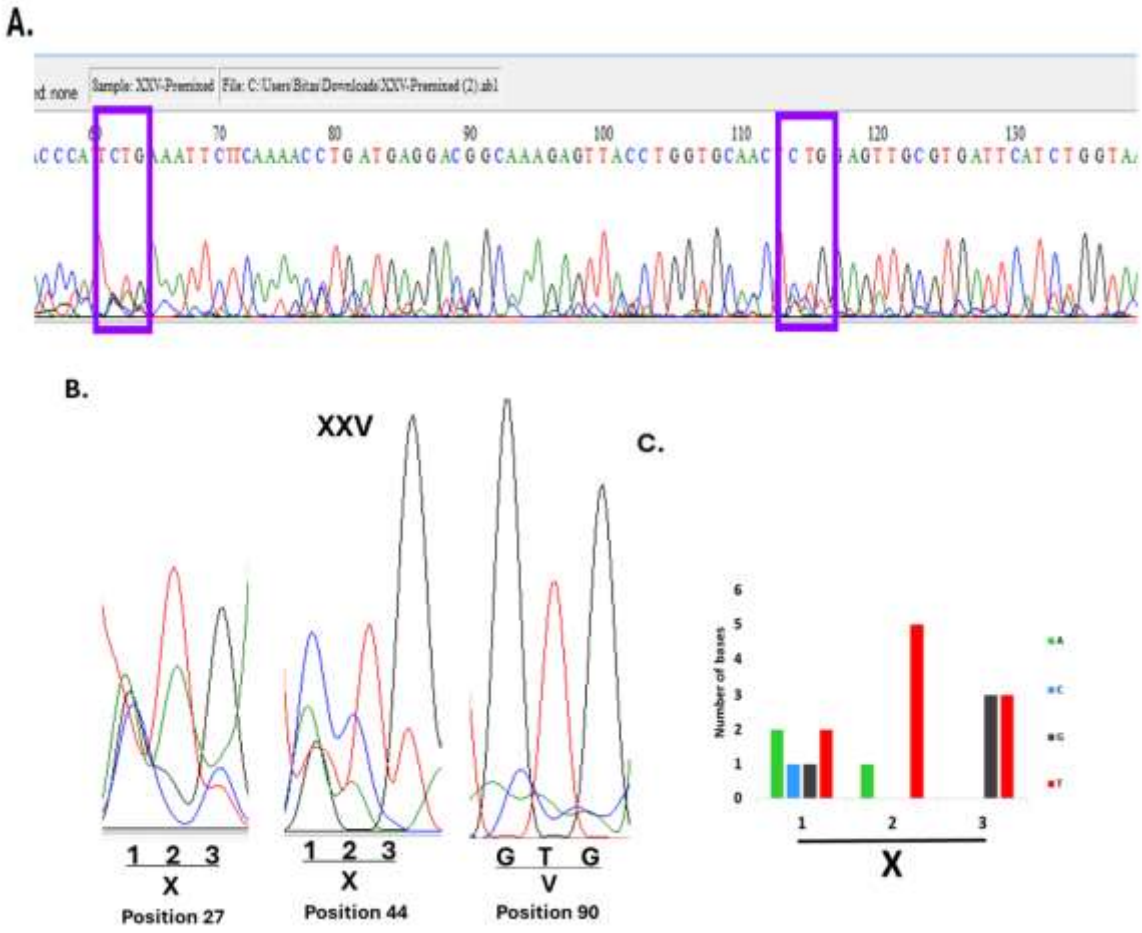


Figure 4.26.Observed vs. expected result of Sanger sequencing of the XXV library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the sequence, although with some background noise. Both randomised positions are highlighted with purple boxes.* ***(B)*** *Sanger sequencing result (observed) for the XXV library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 90, the sequencing data indicated that codon GTG was fixed to encode valine. In contrast, positions 27 and 44 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

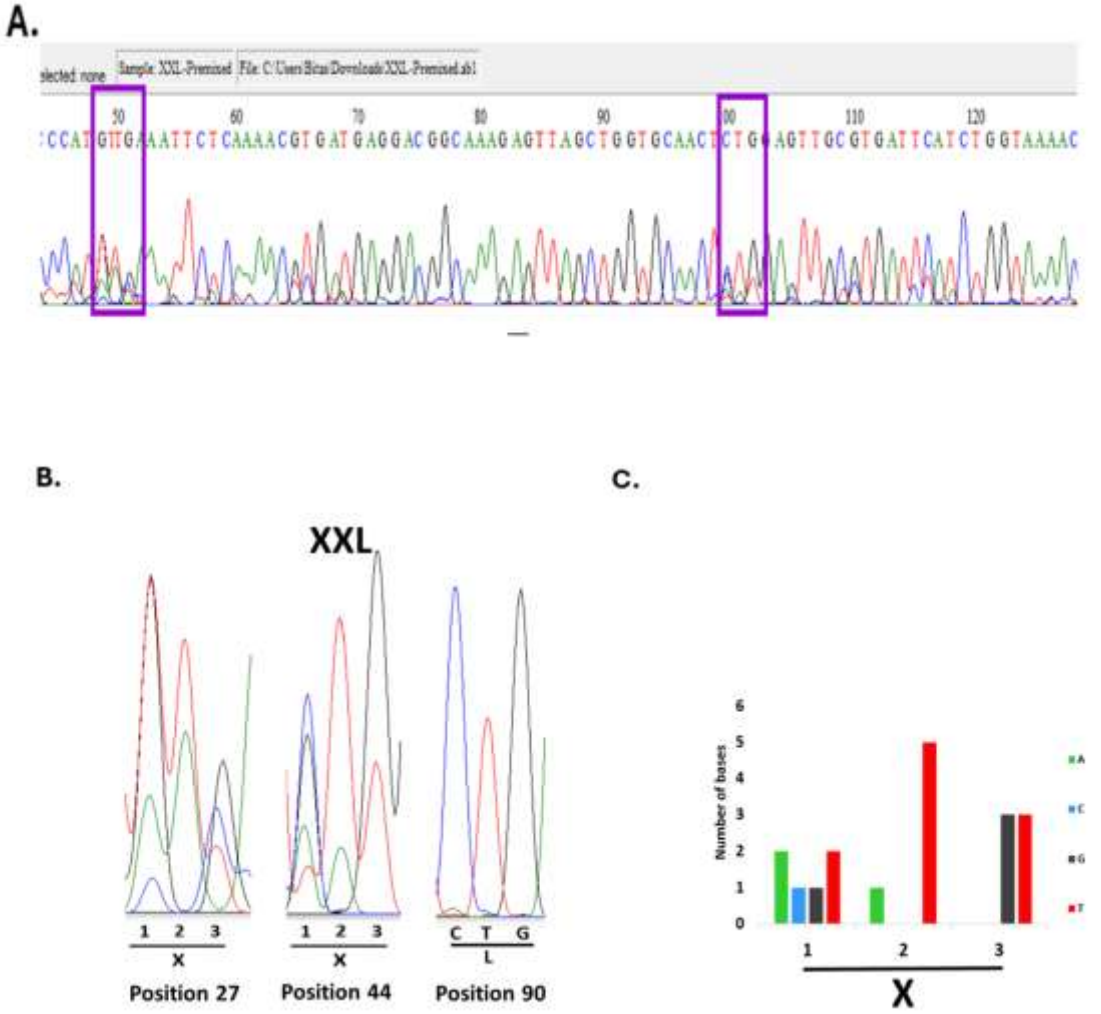


Figure 4.27. Observed vs. expected result of Sanger sequencing of the XXL library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. Both randomised positions are highlighted with purple boxes.* ***(B)*** *Sanger sequencing result (observed) for the XXL library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 90, the sequencing data indicated that codon CTG was fixed to encode leucine. In contrast, positions 27 and 44 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).***

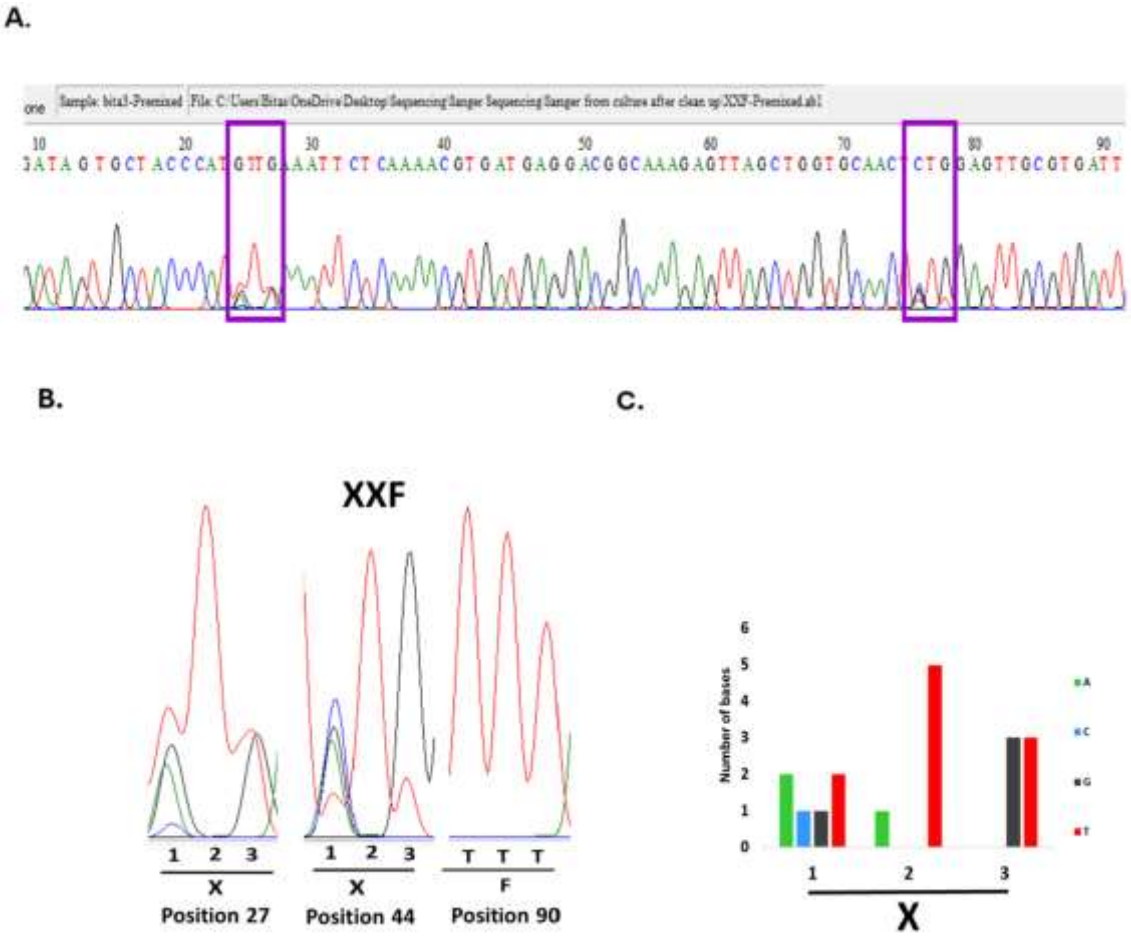


Figure 4.28.Observed vs. expected result of Sanger sequencing of the XXF library. ***(A****) A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. Both randomised positions are highlighted with purple boxes.* ***(B)*** *Sanger sequencing result (observed) for the XXF library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 90, the sequencing data indicated that codon TTT was fixed to encode phenylalanine. In contrast, positions 27 and 44 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).*** *At both randomised position 27 and 44, the green peaks corresponding to A in the second base of the codons are absent or not observed.*

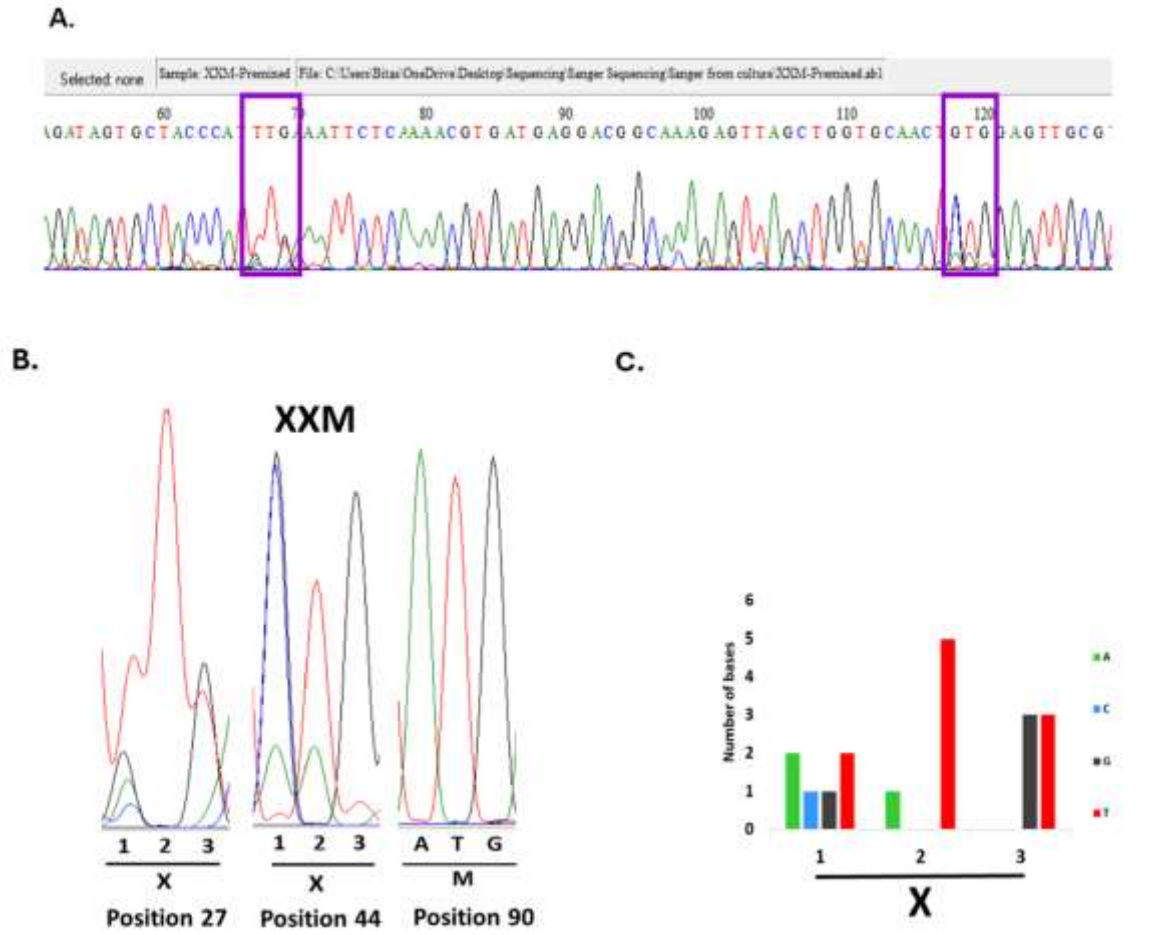


Figure 4.29. Observed vs. expected result of Sanger sequencing of the XXM library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. Both randomised positions are highlighted with purple boxes.* ***(B)*** *Sanger sequencing result (observed) for the XXM library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 90, the sequencing data indicated that codon ATG was fixed to encode methionine. In contrast, positions 27 and 44 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).*** *At position 27, the green peak corresponding to A in the second base of the codon is absent or not observed.*

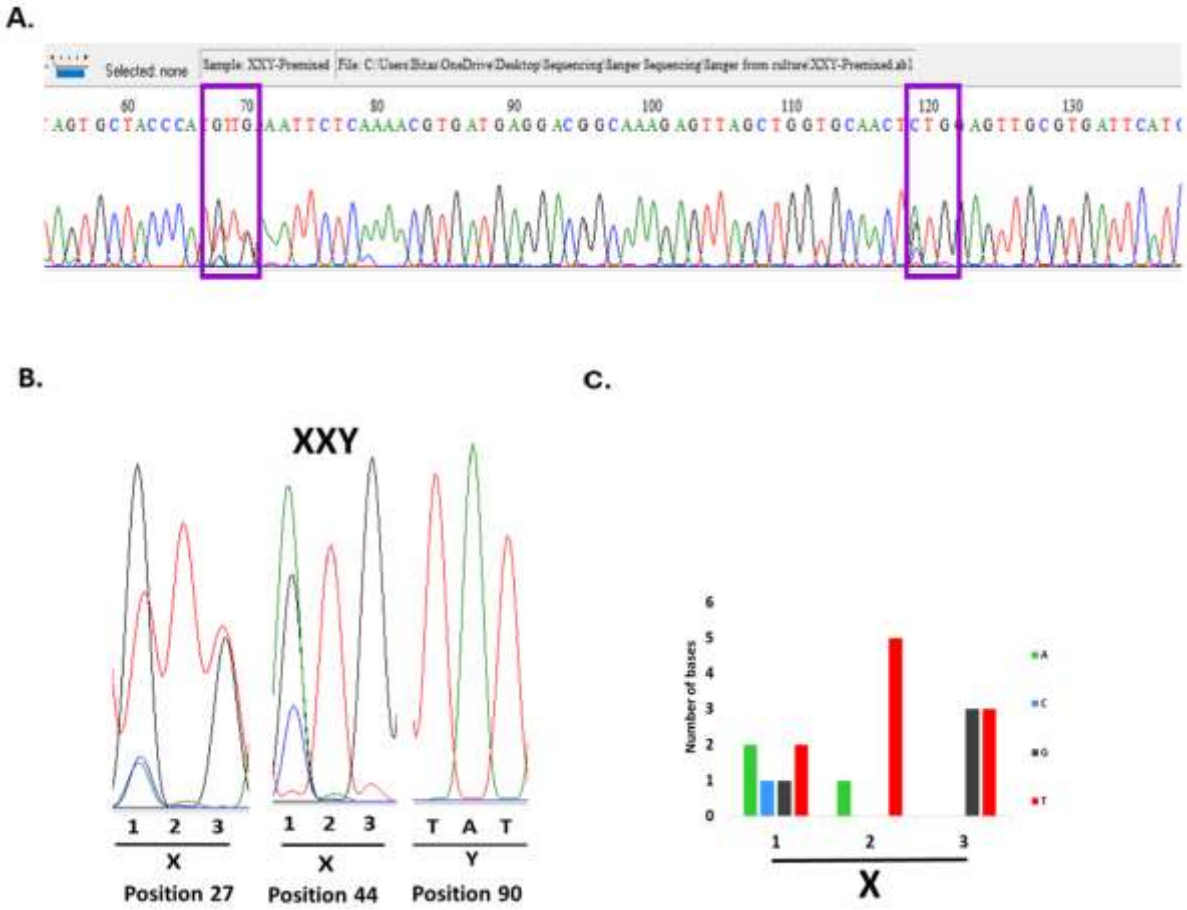


Figure 4.30. Observed vs. expected result of Sanger sequencing of the XXY library. ***(A)*** *A segment of the sequencing chromatogram is presented to demonstrate the high-quality peaks with minimal background noise. Both randomised positions are highlighted with purple boxes.* ***(B)*** *Sanger sequencing result (observed) for the XXY library was extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 90, the sequencing data indicated that codon TAT was fixed to encode tyrosine. In contrast, positions 27 and 44 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in panel* ***(C).*** *At both randomised position 27 and 44, the green peaks corresponding to A in the second base of the codons had weak signals.*

4.7. Next Generation Sequencing of the SpyCatcher libraries

Since the results of Sanger sequencing were inconclusive (with respect to codon distribution), to properly verify the codon distribution at the randomised positions, the DNA cassettes were analysed by Next Generation Sequencing using the Illumina MiSeq Sequencing System. The randomisation had been designed to distribute the amino acids evenly, with each substituted amino acid expected to comprise 16.67% of the total sequences (100% / 6 = 16.67%).

The SpyCatcher library cassette PCR products were sequenced at Genewiz, using the Amplicon EZ service, that employs paired reads in both 5'-3' and 3'-5' directions to ensure comprehensive library coverage. The sequencing data was subsequently processed and analysed (2.2.6.2.1), with the results presented graphically.

4.7.1. Amino acid distribution of the first-fixed position libraries

The observed and expected distribution of amino acids at randomised positions and fixed position for fragments of the first-fixed libraries are depicted below (Figure 4.31).

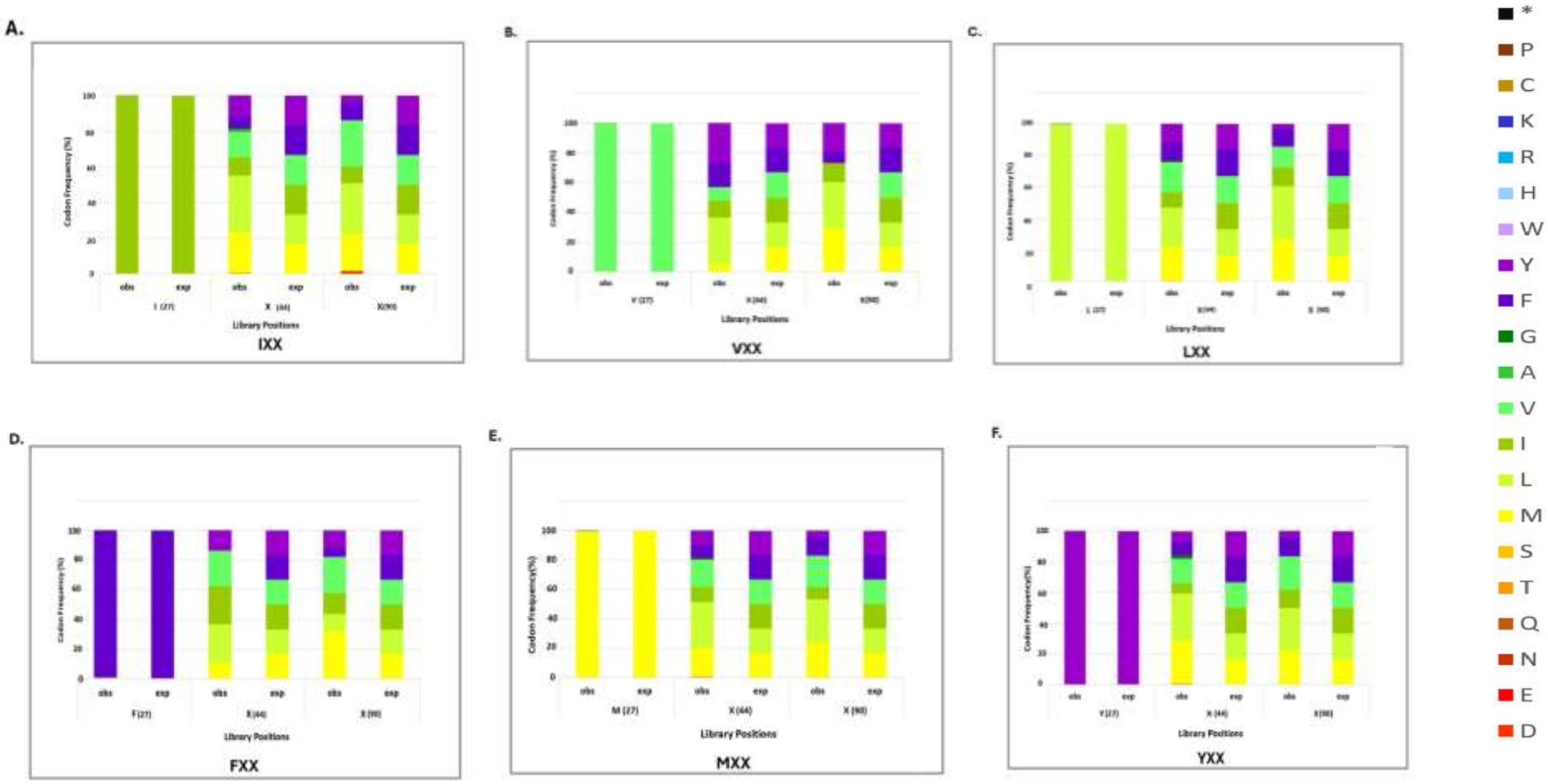


Figure 4.31. Observed and expected distribution of encoded amino acids at randomised and fixed position for fragments of the first fixed position libraries. *The colour coding indicates encoded amino acids, with their corresponding percentages contributing to the overall distribution at each position. The data suggests successful randomisation at positions 44 and 90, with the required range of amino acids represented, while position 27 remains fixed.*

4.7.1.1. IXX Library

Figure 4.31A shows remarkable consistency between observed (obs) and expected (exp) frequencies. Position 27 shows a fixed isoleucine codon, with the observed frequency (obs) aligning with the expected (exp) frequency of 100%. The distribution at positions 44 and 90 as randomised positions, reflects the variability introduced during library construction. Codons for six hydrophobic amino acids are represented, with their respective frequencies shown in the corresponding bars. The observed frequencies at these positions are compared with the theoretical expected distribution.

At both positions 44 and 90, the observed distribution of codons largely reflects the expected distribution. However, there are slight variations in the frequency of encoding certain residues, as follows: at position 44, the codon for leucine showed higher frequencies than anticipated. Conversely, codons for Phenylalanine and isoleucine appear less frequently than expected, whereas in position 90, over-representation of codons for leucine and valine was observed verses a lower frequency than expected for isoleucine and tyrosine.

Overall, both positions 44 and 90 showed an over-representation of leucine (Figure 4.31A).

4.7.1.2. VXX library

Figure 4.31B demonstrates that position 27 is fixed to encode valine, with the observed frequency at almost 100%. At position 44, codons for tyrosine and leucine were observed at higher frequencies than expected, suggesting a potential over-representation of these hydrophobic amino acids, possibly due to biases in the library construction or sequencing processes. Phenylalanine was represented at nearly expected levels, while methionine, valine and isoleucine were observed at lower frequencies.

At position 90, a notable observation is the absence of any valine codons, while codons for leucine and methionine, were over-represented at this position. The codon for isoleucine appeared at frequencies close to the expected levels, while phenylalanine was slightly under-represented.

Overall, leucine again exhibited over-representation in both positions. Nevertheless, the VXX library demonstrates successful randomisation for six amino acids at position 44, although the absence of valine residue at position 90 suggests potential biases in the randomisation process, which could be due to poor synthesis of the fragments during PCR (Figure 4.31B).

4.7.1.3. LXX library

Figure 4.31C demonstrates that position 27 is fixed to encode leucine, with the observed frequency matching the expected frequency at almost 100%. At position 44, codons for

leucine, valine and methionine are over-represented, appearing at higher frequencies than expected. This suggests a potential bias during the library construction or sequencing process, favouring the codons for these amino acids. Codons for phenylalanine and tyrosine both showed frequencies that closely match the expected values, indicating successful randomisation for these amino acids. Isoleucine is encoded at slightly lower than expected levels.

At position 90, leucine and methionine, similar to position 44, are over-represented, indicating a consistent trend of bias towards these codons across multiple positions within the library. Codons for valine, isoleucine and phenylalanine appear at frequencies close to the expected values, suggesting effective randomisation at this position, with underrepresentation for tyrosine.

In conclusion, while the LXX library achieves effective randomisation for both randomised positions, the consistent overrepresentation of leucine could be based on the randomisation approach of overlap PCR (Figure 4.31C).

4.7.1.4. FXX library

Position 27 consistently exhibited a fixed phenylalanine codon, with an observed frequency of 100%. At position 44, the absence of phenylalanine codons is noteworthy. In contrast, codons for leucine, isoleucine, and valine were observed at slightly higher frequencies than anticipated, suggesting a possible over-encoding of these hydrophobic amino acids. Tyrosine was observed at frequencies nearly matching the expected values, whereas methionine was underrepresented.

At position 90, codons for both methionine and valine were over-represented. Isoleucine and tyrosine codons appeared at frequencies close to expected levels, while phenylalanine and leucine were slightly under-represented (Figure 4.31D).

The analysis of the FXX library indicates successful randomisation of six amino acids at position 90. However, the absence of one residue at position 44 suggests failure in PCR or else an error in combining and mixing primers.

4.7.1.5. MXX library

Position 27 consistently exhibited a fixed methionine codon, with an observed frequency of almost 100%. At position 44, the leucine codon was predominantly observed, and the valine codon appeared at slightly higher frequencies than anticipated, suggesting a potential over-representation of these hydrophobic amino acids. Methionine was observed at frequencies

nearly matching the expected values, whereas codon for isoleucine, phenylalanine, and tyrosine were under-represented.

At position 90, codons for leucine, methionine and valine were over-represented. Phenylalanine codons appeared at a frequency close to the expected levels, while isoleucine and tyrosine were under-represented.

The analysis of the MXX library indicates successful randomisation of six amino acids at both randomised positions. However, leucine codons were predominantly over-represented at both positions, suggesting potential biases in the randomisation process (Figure 4.31E).

4.7.1.6. YXX library

Position 27 consistently exhibited a fixed tyrosine codon, with an observed frequency of 100%. At position 44, leucine and methionine codons were observed at higher frequencies than anticipated. The codon for valine was observed at frequencies nearly matching the expected values, whereas codons for isoleucine, phenylalanine and tyrosine were under-represented. Glycine was also present but at very low frequencies, which might be due to the sequence similarity between glycine codons and valine codons and could result either from mutation or sequence mis-calls. At position 90, similar to position 44, codons for leucine and methionine, as well as valine, were over-represented. Phenylalanine and isoleucine appeared at frequencies close to the expected levels, while tyrosine was under-represented.

The analysis of the YXX library indicates successful randomisation of six amino acids at both randomised positions. The distribution of these six residues at the randomised positions was closely aligned, with only slight differences observed between the two positions (Figure 4.31F).

4.7.1. Amino acid distribution of the second-fixed position libraries

The observed and expected distribution of amino acids at fixed and randomised positions for fragments of the second-fixed libraries are depicted below (Figure 4.32).

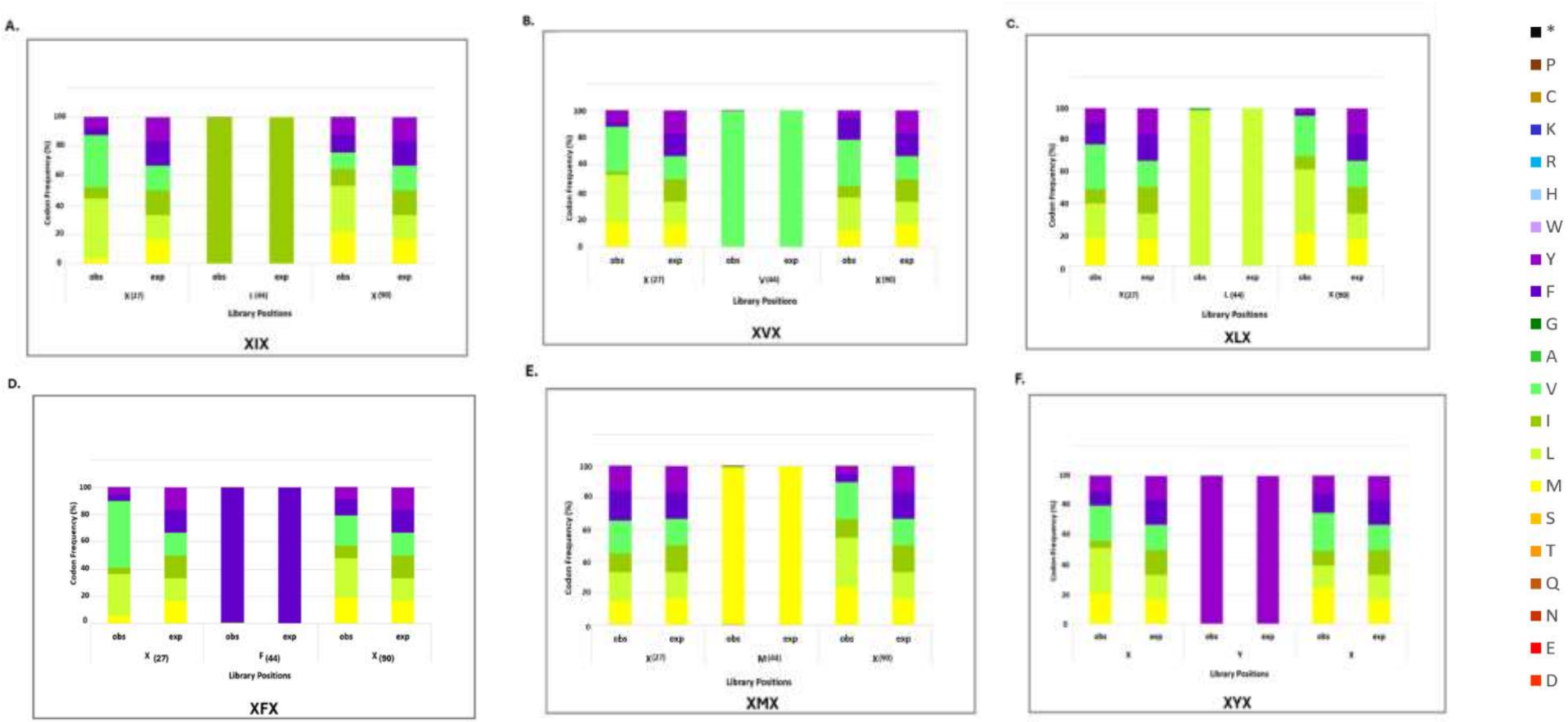


Figure 4.32. Observed and expected distribution of encoded amino acids at randomised and fixed position for fragments of the second fixed position libraries. *The colour coding indicates encoded amino acids, with their corresponding percentages contributing to the overall distribution at each position. The data suggests successful randomisation at positions 27 and 90, with the required range of amino acids represented, while position 44 remains fixed.*

4.7.2.1. XIX library

In this library, position 44 consistently exhibited a fixed isoleucine codon, with an observed frequency of almost 100%. At position 27, there was a significant over-representation of codons for leucine and valine, while codons for the remaining residues were notably under-represented.

At position 90, the codon for leucine was predominantly over-represented, with codons for the other residues appearing at frequencies relatively close to the expected levels.

The analysis of the XIX library indicates successful randomisation of six amino acids at both randomised positions, although the randomisation at position 90 exhibited a more consistent distribution of the amino acids (Figure 4.32A).

4.7.2.2. XVX library

Position 44 consistently exhibited a fixed valine codon, with an observed frequency of almost 100%. At position 27, there was a significant overrepresentation of codons for leucine and valine, while codons for phenylalanine and isoleucine were notably under-represented. Codons for methionine and tyrosine showed frequencies that closely matched the expected values.

At position 90, valine was the predominantly over-represented codon. The distribution of amino acids at this position was similar to that observed at position 27, with the primary difference being in an improved representation of the codon for phenylalanine.

The analysis of the XVX library indicates successful randomisation of six amino acids at both randomised positions, with the randomisation at both positions showing a similar distribution of codons (Figure 4.32B).

4.7.2.3. XLX library

Position 44 consistently exhibited a fixed leucine codon, with an observed frequencies of approximately 98%. At position 27, there was an over-representation primarily of the valine codon, followed by leucine, while the remaining codons displayed distributions relatively close to the expected values, with the exception of the codon for isoleucine, which was under-represented.

At position 90, leucine was the most predominantly over-represented codon, followed by valine. Codons for tyrosine and phenylalanine were observed at lower frequencies compared with the other four codons.

The analysis of the XLX library indicates successful randomisation of six amino acids at both randomised positions, with a more balanced distribution at position 27 compared with position 90 (Figure 4.32C).

4.7.2.4. XFX library

Position 44 consistently exhibited a fixed phenylalanine codon, with an observed frequency of 100%. At position 27, there was significant over-representation of the codon for valine, followed by leucine, while the remaining codons displayed distributions with frequencies lower than the expected values.

At position 90, leucine was the most over-represented codon, followed by valine. Codons for methionine and phenylalanine were relatively similar to expected frequencies and there was slight under-representation of codons for tyrosine and isoleucine.

The analysis of the XFX library indicates successful randomisation of six amino acids at both randomised positions, with a more balanced distribution at position 90 compared with position 27 (Figure 4.32D).

4.7.2.5. XMX library

Position 44 consistently exhibited a fixed methionine codon, with an observed frequency of 99%. At position 27, there was a slight overrepresentation of the codon for valine, while the other codons were distributed evenly in line with the expected values.

At position 90, leucine was the most over-represented codon, followed by valine and, to a lesser extent, methionine. In contrast, codons for tyrosine, phenylalanine, and isoleucine were underrepresented.

The analysis of the XMX library suggests successful randomisation of six amino acids at both randomised positions, with a more balanced distribution observed at position 27 compared with position 90 (Figure 4.32E).

4.7.2.6. XYX library

Position 44 consistently exhibited a fixed tyrosine codon, with an observed frequency of 100%. At position 27, there was an overrepresentation of the codons for leucine, followed by valine, while the remaining codons showed distributions closely aligned with the expected values, except for isoleucine, which was under-represented.

At position 90, codons for valine and methionine were over-represented, while the remaining residues displayed distributions relatively consistent with the expected values.

The analysis of the XYX library suggests successful randomisation of six amino acids at both randomised positions, with a more balanced distribution observed at position 90 compared with position 27 (Figure 4.32F).

4.7.3. Amino acid distribution of the third-fixed position libraries

The observed and expected distribution of amino acids at fixed and randomised positions for fragments of the third-fixed libraries are depicted below (Figure 4.33).

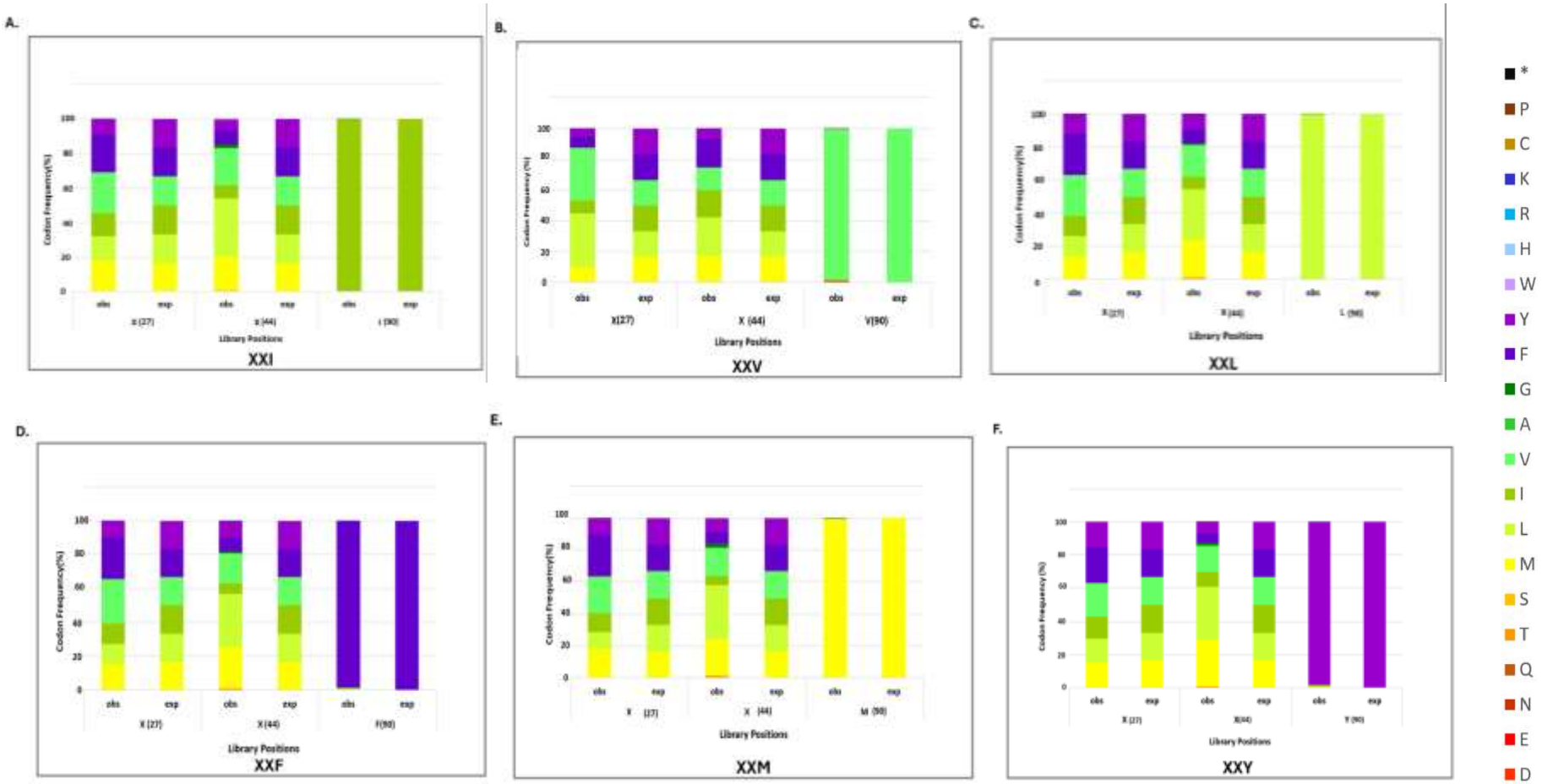


Figure 4.33. Observed and expected distribution of codons at randomised and fixed position for fragments of the third fixed position libraries. *The colour coding indicates different required codons, with their corresponding percentages contributing to the overall distribution at each position. The data suggests successful randomisation at positions 27 and 44, with a range of amino acids encoded, while position 90 remains fixed.*

4.7.3.1. XXI Library

Position 90 consistently exhibited a fixed isoleucine codon, with an observed frequency of almost 100%. At position 27, there was a slightly over-representation of the codon for valine, followed by phenylalanine, while the remaining codons showed distributions closely aligned with the expected values, except for tyrosine, which was under-represented.

At position 44, codons for leucine and valine were over-represented, while the remaining codons displayed relative underrepresentation compared to the expected values and slightly presence of glycine.

The analysis of the XXI library suggests successful randomisation of six amino acids at both randomised positions, with a more balanced distribution observed at position 27 compared with position 44 (Figure 4.33A).

4.7.3.2. XXV Library

Position 90 consistently exhibited a fixed valine codon, with an observed frequency of 99%. At position 27, the codon for valine was over-represented, followed by leucine, while the other codons appeared at frequencies lower than expected.

At position 44, an even distribution was observed among all six codons.

The analysis of the XXV library indicates successful randomisation of six amino acids at both randomised positions, with a nice balanced distribution at position 44 compared with position 27 (Figure 4.33B).

4.7.3.3. XXL Library

Position 90 consistently exhibited a fixed leucine codon, with an observed frequency of almost 100%. At position 27, there was a slight over-representation of the codon for valine, followed by phenylalanine, while the remaining codons showed distributions closely aligned with the expected values.

At position 44, the codon for leucine was over-represented and there was also a slight over-representation of valine, followed by methionine, while codons for the remaining residues displayed relative under-representation compared with the expected values.

The analysis of the XXL library suggests successful randomisation of six amino acids at both randomised positions, with a more balanced distribution observed at position 27 compared with position 44 (Figure 4.33C).

4.7.3.4. XXF Library

Position 90 consistently exhibited a fixed phenylalanine codon, with an observed frequency of 100%. At position 27, the codon for valine was slightly over-represented, followed by phenylalanine, while the remaining codons showed distributions closely matching the expected values.

At position 44, the codon for leucine was predominantly over-represented, followed by a slight over-representation of methionine, while the other codons were relatively under-represented compared with the expected values, with the valine codon maintaining the expected distribution and a very slight presence of a codon for glycine.

The analysis of the XXF library suggests successful randomisation of six amino acids at both randomised positions, with a balanced distribution observed across both positions (Figure 4.33D).

4.7.3.5. XXM Library

Position 90 consistently exhibited a fixed methionine codon, with an observed frequency of almost 100%. At position 27, the codon for valine was slightly over-represented, followed by phenylalanine, while the remaining codons showed distributions closely matching the expected values.

At position 44, the codon leucine was predominantly over-represented, followed by a slight over-representation of methionine, while the other codons were relatively under-represented compared with the expected values. The codon for valine was matched the expected distribution and there was an additional tiny distribution of codons for glycine and threonine.

The analysis of the XXM library suggests successful randomisation of six amino acids at both randomised positions, with a more balanced distribution observed at position 27 compared with position 44 (Figure 4.33E)

4.7.3.6. XXY Library

Position 90 consistently exhibited a fixed tyrosine codon, with an observed frequency of almost 100%. At position 27, a relatively even distribution of all codons was observed.

At position 44, the codon for leucine was predominantly over-represented, followed by methionine, while the other codons were relatively under-represented compared with the expected values. The codon for valine maintained the expected distribution, with glycine codons present at very low levels.

The analysis of the XXY library indicates successful randomisation of six amino acids at both randomised positions, with a good balanced distribution observed at position 27 compared with position 44 (Figure 4.33F).

4.8. Comparative analysis of Sanger sequencing and next-generation sequencing results

Among the 18 SpyCatcher libraries, three libraries - one representing the first, second, and third fixed positions - were selected for comparative analysis of their sequencing data obtained via Sanger sequencing and Next-Generation Sequencing (NGS).

The objective was to compare the expected Sanger chromatogram with the base frequencies derived from NGS reads. Specifically, these values comprised a 1:1 ratio for A and T (A:T), and separately a 1:1 ratio for C and G (C:G) at the first base of the randomised codon, a 5:1 ratio of T to A (T:A) at the second base, with the absence of C and G, and a 1:1 ratio of G with T (G:T) with the absence of A and C at the third base, as described previously (Figure 4.12).

In the Sanger chromatogram, these patterns were observed; however, it was not possible to quantitatively confirm the expected ratios. Notably, adenine (A) was frequently absent in the second base of the randomised codon in the Sanger sequencing results, whereas NGS data confirmed the presence of adenine with an almost accurate ratio of 5:1 (T) in the randomised positions (Figure 4.34, Figure 4.35 and Figure 4.36).

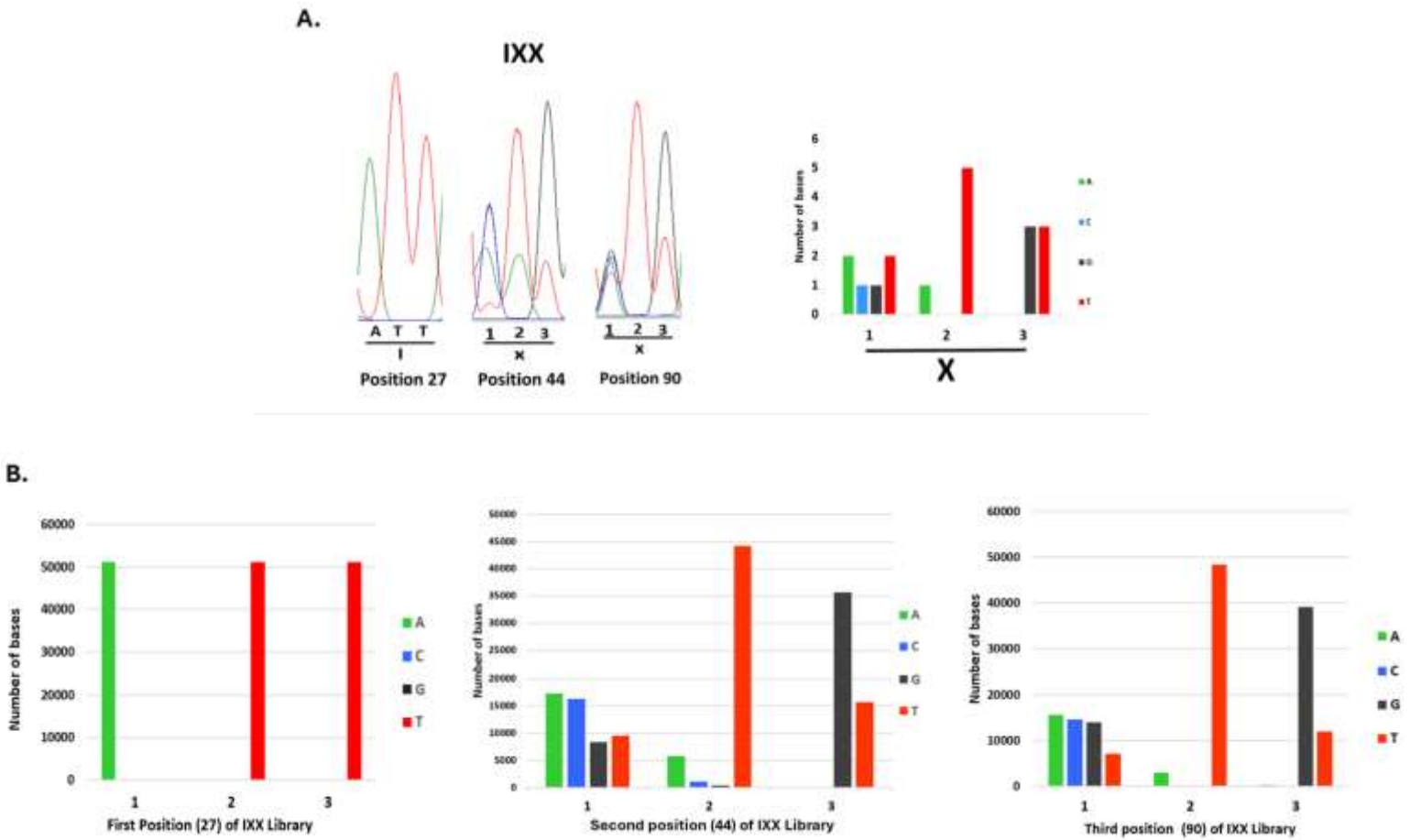


Figure 4.34. Sanger sequencing versus next generation sequencing of the IXX library. (A) *Sanger sequencing results for the IXX library were extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 27, the sequencing data indicated that codon ATT was fixed to encode isoleucine. In contrast, positions 44 and 90 exhibited randomisation, showing a mixture of bases corresponding to all six possible amino acid residues, aligning with theoretical expectation presented in the right hand histogram.* ***(B)*** *Base frequencies derived from Next-Generation Sequencing (NGS) reads, each representing the number of bases calculated from the NGS results at each position.*

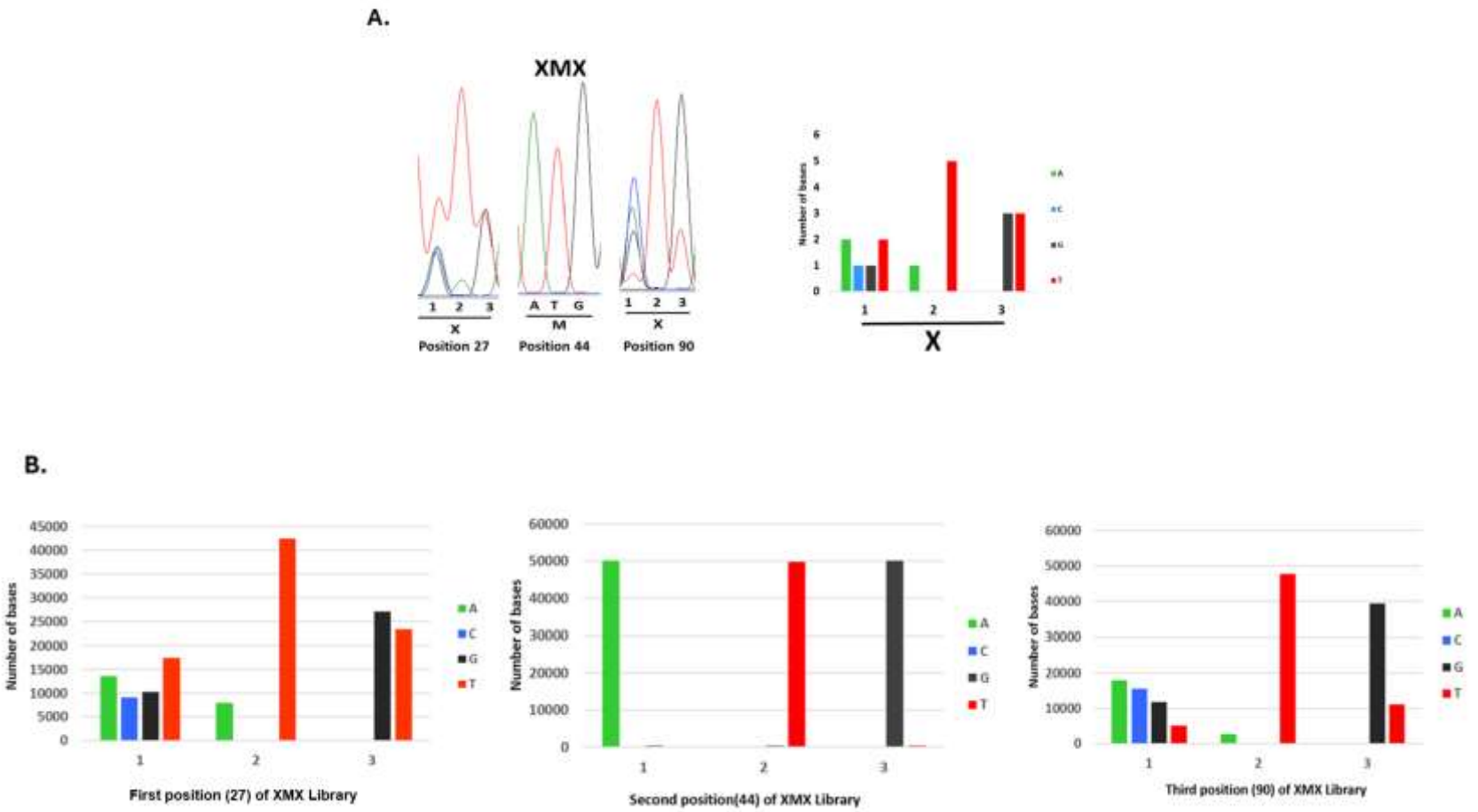


Figure 4.35. Sanger sequencing versus next generation sequencing of the XMX library. (A) *Sanger sequencing results for the XMX library were extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 44, the sequencing data indicated that the codon ATG was fixed to encode methionine. In contrast, positions 27 and 90 exhibited randomisation, showing a mixture of bases encoding all six possible amino acid residues, aligning with theoretical expectation presenting in the right hand histogram.* ***(B)*** *Base frequencies derived from Next-Generation Sequencing (NGS) reads, each representing the number of bases calculated from the NGS results at each position.*

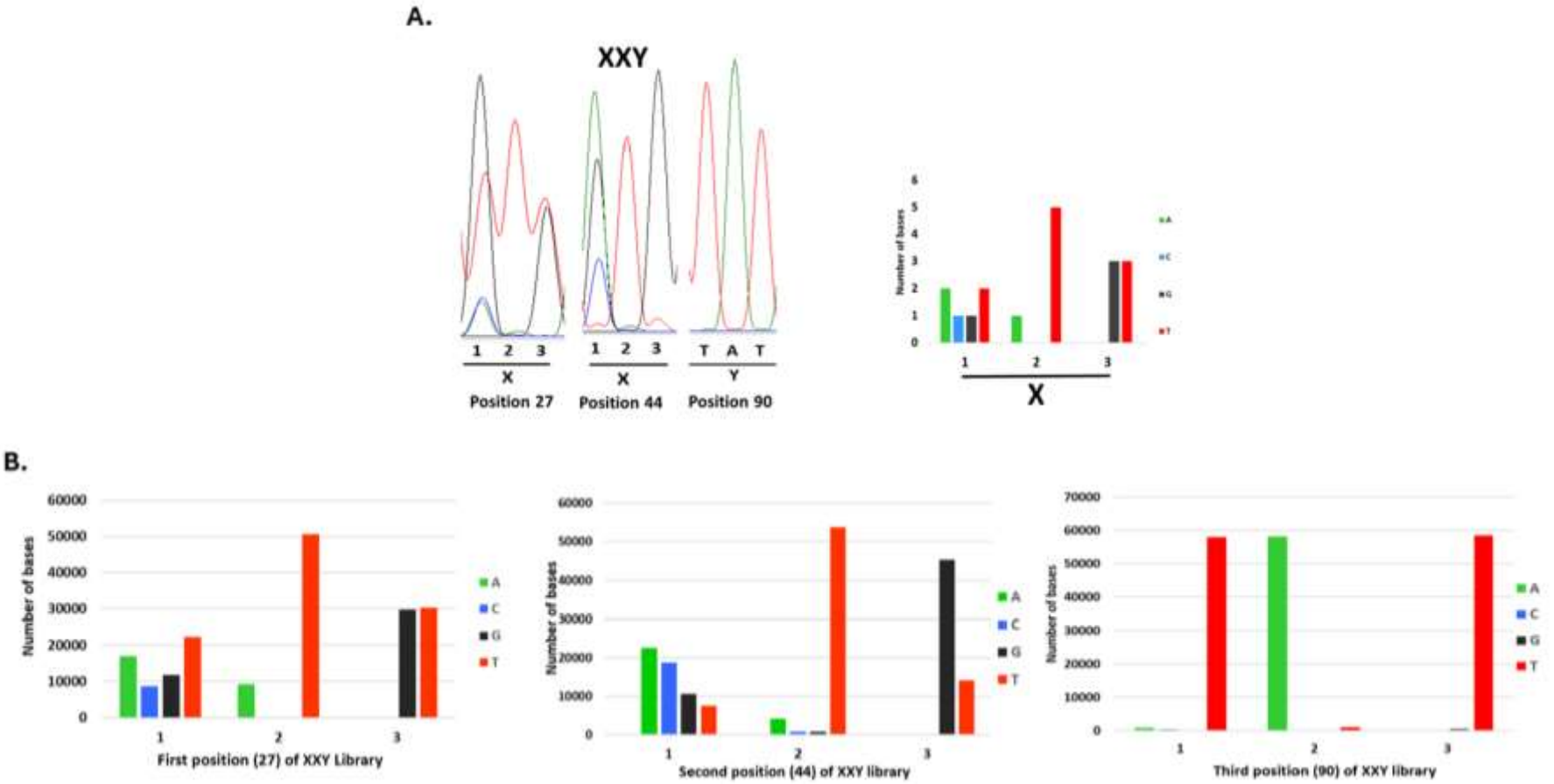


Figure 4.36. Sanger sequencing versus next generation sequencing of the XXY library. (A) *Sanger sequencing results for the XXY library were extracted from the chromatogram at each of the three targeted positions and combined for analysis. At position 90, the sequencing data indicated that codon TAT was fixed to encode tyrosine. In contrast, positions 27 and 44 exhibited randomisation, showing a mixture of bases encoding all six possible amino acid residues, aligning with theoretical expectation presented in the right-hand histogram.* ***(B)*** *Base frequencies derived from Next-Generation Sequencing (NGS) reads, each representing the number of bases calculated from the NGS results at each position.*

In Figure 4.34, in position 27, the chromatogram in panel A shows a clear peak for the fixed codon at this position, indicating consistent incorporation of the expected bases (ATT) which is consistent with NGS (panel B), where the histogram similarly confirms the fixed nature of the bases at position 27. In position 44, the chromatogram indicates a mixed representation of nucleotides at this codon, while NGS shows near-expected distribution for the first base, and a 5:1 ratio of T to A in the second base, but a higher representation of G compared with A in the third base, rather than the expected 1:1 ratio. In position 90, similar to Position 44, the chromatogram showed a mixed nucleotide presence, confirming randomisation, even though the signal for A is absent at the second base. However with NGS (panel B), results for position 90 display a more refined distribution in the first base as expected, highlighting an even distribution of C and G and the anticipated presence of A in the second base.

In Figure 4.35, for position 27, the chromatogram indicates a mixed representation of nucleotides, confirming the randomised position, while NGS (panel B), shows a more detailed distribution of nucleotides, with an almost expected even distribution for the first base, a 5:1 ratio of T to A in the second base, and an even distribution for the third base. In position 44, the chromatogram in Panel A shows a clear peak for the fixed nucleotide at this position, indicating consistent incorporation of the expected bases (A, T, G) and with NGS (panel B) the histogram similarly confirms the fixed nature of the bases at position 44. In position 90, similar to position 44, the chromatogram shows a mixed nucleotide presence, confirming randomisation, though the signal for A is absent in the second base. While NGS (panel B) of position 90 displays a more refined distribution in the first base as expected, the anticipated presence of A in the second base, and an excess of G compared with T at third base.

In Figure 4.36, for position 27, the chromatogram indicates a mixed representation of nucleotides confirming the randomisation, but with a missing signal for A in the second base, while with NGS (panel B), the histogram shows a more detailed distribution of nucleotides, with an almost expected even distribution for the first base, the presence of a 5:1 ratio of T to A in the second base, and an even distribution for the third base. In position 44, similar to Position 27, the chromatogram shows a mixed nucleotide presence, confirming the presence of the randomised codon, even though the signal for T is weak at the third base. While with NGS (panel B), results for position 44 display a more refined distribution in the first base as expected, the anticipated presence of A in the second base, and an excess of G compared with T at the third base. At position 90, the chromatogram in Panel A shows a clear peak for the fixed nucleotide at this position, indicating consistent incorporation of the expected bases (TAT) and with NGS (panel B) the histogram similarly confirms the fixed nature of the bases at position 90.

Overall Comparison: As might be expected, NGS offers superior resolution and quantification of nucleotide distributions across all positions when compared with Sanger sequencing, with respect to randomised codons. While Sanger sequencing provides a general overview and can confirm fixed or randomised nucleotides, NGS allows for a more detailed understanding of the nucleotide composition, especially in cases of mixed or over-represented sequences. NGS provides precise counts of each nucleotide, making it easier to identify over- or under-representations in the libraries, whereas Sanger sequencing offers more qualitative insights with less precision in quantification. In terms of sensitivity, NGS offers higher sensitivity, capable of detecting low-frequency variants as seen in the detailed graphs in panels 4.32B, 4.33B and 4.34B. Thus, in summary, NGS offers more comprehensive and precise data, particularly for positions with mixed nucleotide distributions.

4.9. Gene expression

After confirming the sequences of the SpyCatcher libraries using both Sanger sequencing and NGS, the DNA libraries were determined to be well-suited for subsequent protein expression. The confirmed constructs were subsequently transformed into the chemically competent *E. coli* Tuner™ (DE3) cells (2.2.5) in preparation for gene expression. These cells are specifically designed to work with inducible expression systems, making them ideal for producing a wide range of protein variants from the libraries (Baumgarten et al., 2018). Simultaneously, to evaluate transformation efficiency, a 50 µL aliquot of the transformed cells was plated onto Luria-Broth (LB). If the various stages of cloning had worked properly, resulting colonies should be fluorescent since the SpyCatcher libraries contained a C-terminal fusion of the mNeonGreen gene. As shown in Figure 4.37, fluorescent colonies were indeed apparent, indicating successful cloning.

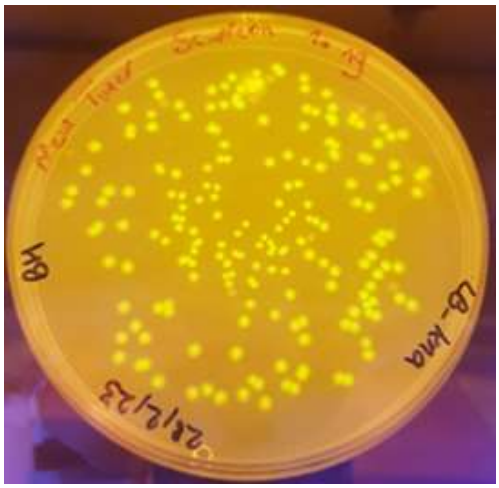


Figure 4.37. Fluorescent colonies of native SpyCatcher-mNeongreen. *Transformed cells were plated on LB agar containing kanamycin and incubated overnight to allow for the selection of successfully transformed colonies which were then observed under UV transillumination.*

4.10. Expression of 18 SpyCatcher libraries

Initially, wild-type SpyCatcher mNeongreen was expressed in *Escherichia coli* Tuner™(DE3) cells under a range of conditions to optimise protein yield. Variables such as IPTG concentration, induction time and growth temperature were adjusted to identify the conditions that maximise protein expression (2.2.7). Protein expression was subsequently analysed by SDS-PAGE electrophoresis (2.2.3.2). Cell lysates were prepared using the BugBuster reagent (2.2.8) to ensure efficient protein extraction.

4.10.1. IPTG concentration

Various IPTG concentrations were evaluated to determine the optimal level for inducing protein expression (Figure 4.38).

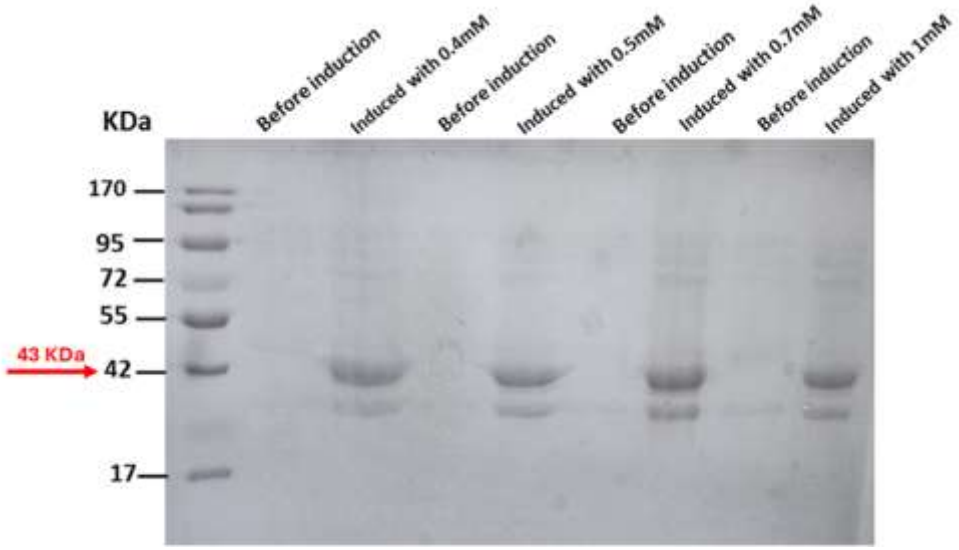


Figure 4.38. Analysis of wild-type SpyCatcher protein expression at various IPTG concentrations*. Polyacrylamide gel electrophoresis using a 12% gel was performed from lysate after induction, induced with 0.4, 0.5, 0.7 and 1.0 mM IPTG for 3 hours at 30°C with shaking at 200 rpm. The gel was stained with InstantBlue for visualisation.*

The predicted molecular weight of the wild-type SpyCatcher, including the 1 kDa His tag, is approximately 43 kDa. This estimation was confirmed by the predominant bands observed in Fig 4.38 (indicated by red arrow), which correspond to the expected protein size. As the IPTG concentration showed similar SpyCatcher production at all concentrations tested, 1.0 mM IPTG was selected for inducing SpyCatcher protein expression in further experiments.

4.10.2. Temperature

Post-induction growth temperature can sometimes greatly influence both the solubility and yield of recombinant proteins. To assess this, the effects of different temperatures were tested during the overnight expression of the wild-type SpyCatcher protein. After protein extraction with BugBuster, the supernatant was analysed on SDS-PAGE as shown in Figure 4.39.

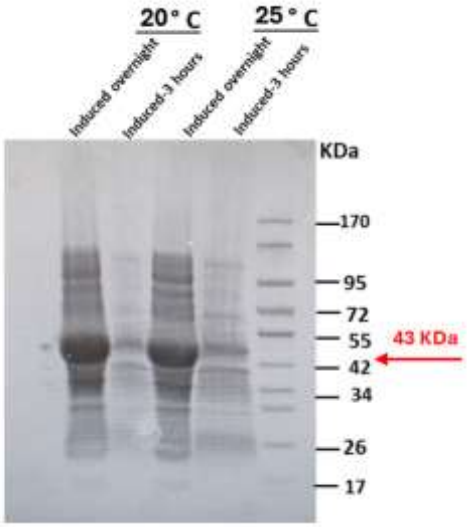


Figure 4.39. SDS-PAGE analysis of wild-type SpyCatcher protein expression at different temperatures. *Polyacrylamide gel electrophoresis using a 12% gel was performed to analyse the expression of SpyCatcher, induced with 1.0 mM IPTG for 3 hours and overnight, at 20°C and 25°C with shaking at 200 rpm. The gel displays the whole cell lysate, showing overexpression of the full-length protein (MW = 43 kDa), as indicated by red arrow. The gel was stained with InstantBlue for visualisation.*

The SpyCatcher protein was successfully expressed at both 20°C and 25°C temperatures; however, the differences between these temperatures were minimal, leading to the selection of 20°C with overnight expression as the chosen conditions for protein production.

4.11. Affinity purification of 18 SpyCatcher libraries

After determining the optimal expression conditions including 20°C as optimal temperature and 1.0 mM IPTG concentration for induction, cell lysates were prepared using BugBuster for efficient lysis. Each SpyCatcher library was expressed in a 200 mL culture and purified using affinity chromatography (2.2.9). The SpyCatcher protein was separated from other proteins using affinity chromatography with nickel-NTA resins, taking advantage of the His tag included in the construct. The SpyCatcher proteins were eluted from the Ni-NTA resin with 250 mM imidazole.

PAGE analysis of the starting material, flow-through, wash, and eluted fractions showed a strong expression of the protein as demonstrated in Figure 4.40.

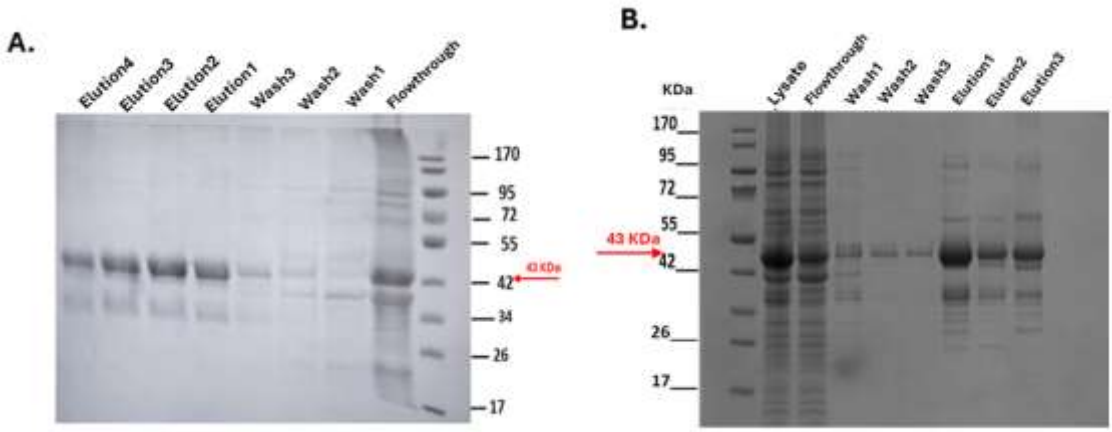
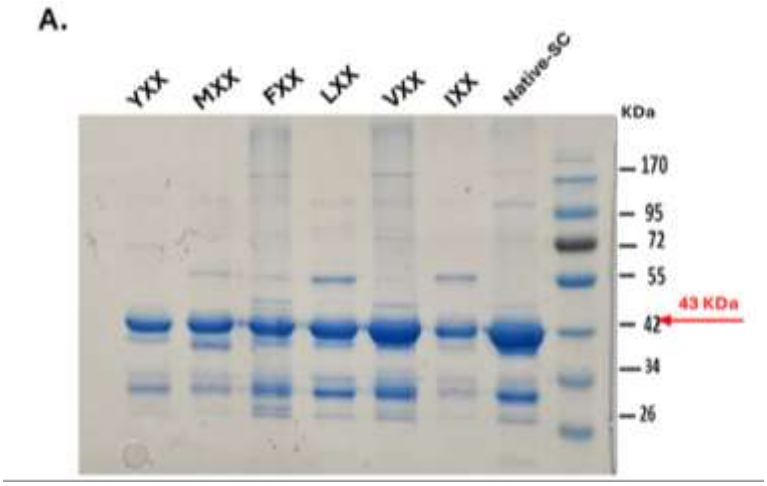


Figure 4.40. Purification of (A) wt. SpyCatcher, and (B) mutated SpyCatcher by affinity chromatography using Ni-NTA resin. *Polyacrylamide gel electrophoresis using a 12% gel shows samples obtained during purification using affinity column chromatography. The expected size of the protein is 43 kDa (with His tag). The gels were stained with InstantBlue.*

The highest yield of soluble, recombinant protein was detected in the first elution, while the final fraction (elution 4) indicates that some protein was still eluting. The purified samples display a prominent band near the 43 kDa marker, corresponding to the SpyCatcher-mNeonGreen protein, along with a some other proteins that could represent either co-purification (i.e. *E. coli* proteins that have natural affinity for the His6 Tag) or else degradation products of the SpyCatcher-NeonGreen fusion protein. SDS-PAGE followed by staining was conducted to assess the 18 SpyCatcher libraries shown in Figure 4.41.



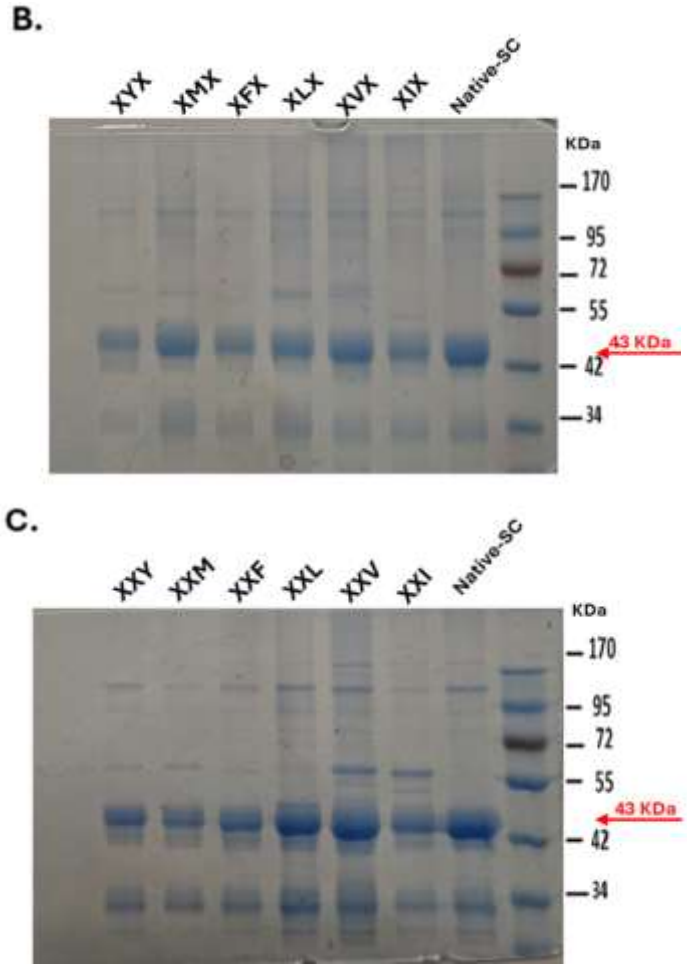


Figure 4.41. Purification of the SpyCatcher Libraries: (A) First-fixed, (B) Second-fixed, and (C) Third-fixed*. Polyacrylamide gel electrophoresis (SDS-PAGE, 4-12% gradient) was performed to analyse samples from the first elution of each purification. Purification was conducted using affinity column chromatography, eluted with 250 mM imidazole with the expected protein size being 43 kDa, including the His tag. Gels were stained with InstantBlue.*

4.12. Discussion

The construction of SpyCatcher libraries is pivotal to this study, aimed at exploring the potential for engineering proteins with novel properties. The creation of fixed-position libraries via overlap PCR allowed for targeted mutagenesis at specific amino acid residues within the binding pocket of the SpyCatcher protein. This strategic approach was essential for generating a broad diversity of protein variants, to identify (in future experiments) mutations that might modify protein functionality.

The library design was focused on three positions (27, 44, and 90) within the SpyCatcher protein. By targeting these positions, the study aimed to explore their role in protein-peptide interactions, particularly in modifying the binding specificity of the SpyCatcher protein. By fixing one position at a time and randomising the others, the study created a comprehensive set of

variants. This approach ensured a manageable library size while maximising the diversity of potential protein variants. Such a design is consistent with strategies used in previous studies, where targeted mutagenesis in critical functional regions has led to significant improvements in protein characteristics (Tang et al., 2012a). The choice of overlap PCR as the primary method for generating these libraries was based on the location of chosen codons. Overlap PCR has been demonstrated as a robust technique for introducing specific mutations, allowing for the seamless assembly of mutated DNA fragments (Higuchi et al., 1988b).

The comparative analysis of Sanger and NGS results highlighted the strengths and limitations of each method. While Sanger sequencing is limited by its lower throughput, it provided an overview of randomised positions. In contrast, NGS offered a broader view of the library, revealing subtleties in codon representation that were not evident from Sanger sequencing. This dual approach ensured robust validation of the libraries.

The effectiveness of overlap PCR in terms of codon inclusion / exclusion is evident from the content of all but one position in one library (VXX position 90, Figure 4.31B), where the valine codon at position 90 was absent, despite a repeat synthesis (data not shown). However, at least in this study, results demonstrate that overlap PCR is less effective in terms of representation, which rarely met the balanced 16.67% representation of each chosen codon and discrepancies were observed with respect to consistent over- and under-representation of specific codons. These anomalies might be attributed to various factors.

Inaccurate pipetting of oligonucleotides while preparing the PCR reactions could have contributed to the uneven representation. Another plausible factor could be the differential annealing efficiency of the oligonucleotides, where one oligonucleotide may hybridise more effectively than others, leading to imbalanced incorporation. Lastly, errors in adjusting oligonucleotide concentrations to the desired levels during preparation may have resulted in incorrect stoichiometric ratios, skewing the final distribution. Despite these discrepancies, the overall consistency of residue representation across most libraries suggests that the issue is not due to a systematic error in the protocol. Overall, the libraries generated in this study provide a solid foundation for subsequent screening and selection processes aimed at identifying SpyCatcher variants with modified binding properties.

Chapter 5 Peptide library construction

5.1. Generation of native SpyTag-mCherry plasmid

To create SpyTag libraries, it was necessary to create a vector suitable for expression of the SpyTag peptide. Since SpyCatcher had already been expressed as a fusion to mNeonGreen it was decided to express SpyTag as a fusion to mCherry. Accordingly, primers were designed using NEBuilder (NEB Assembly Tool) to amplify the SpyTag-mCherry vector by inverse PCR, omitting the mNeonGreen gene and including BsaI sites to enable subsequent cloning. Complementary primers were similarly designed to amplify mCherry (Table 9.2, Annex 1). Both PCR reactions were performed (2.2.2.1 and 2.2.2.1.3) and the products were examined by electrophoresis (Figure 5.1).

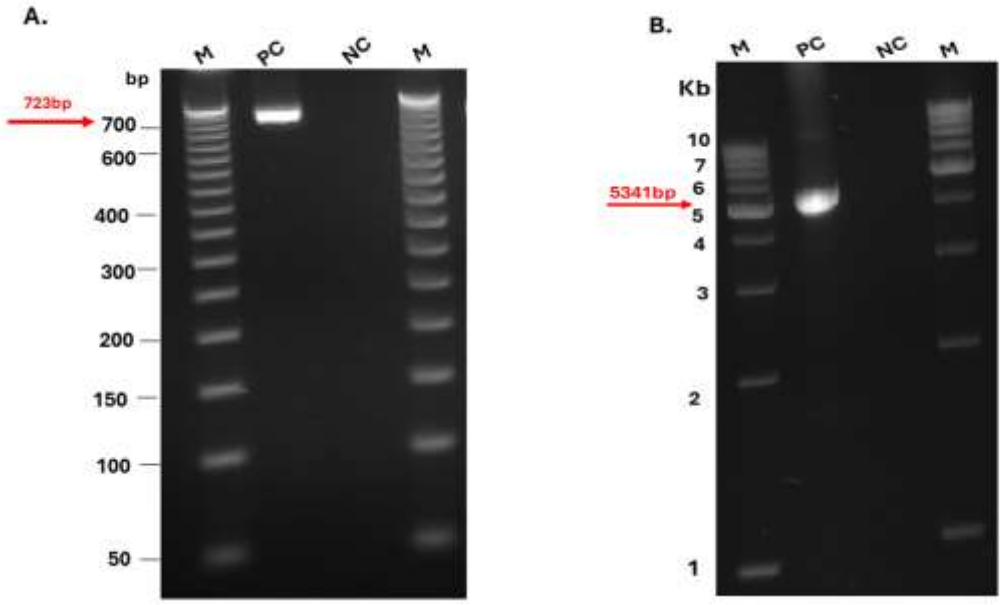


Figure 5.1. Amplification of the genes to make native SpyTag-mCherry plasmid. ***(A)*** *PCR amplification of mCherry gene: The PCR product amplified with Phusion enzyme, expected to be 723 bp, was electrophoresed on a 3% agarose gel and stained with ethidium bromide.* ***(B)*** *Inverse PCR product, amplified with Phusion enzyme with an expected size of 5341 bp, to make the backbone for cloning, was electrophoresed on a 1% agarose gel and stained with ethidium bromide. Lanes: PC positive control, NC, negative (no template) control, M, 50bp and 1Kbp ladder MW marker, respectively.*

The assembled plasmid was subsequently sent for sequencing using the Plasmid-EZ service at Genewiz, employing a plasmid concentration of 50 ng/µL to sequence the entire construct. Sequencing results verified the accurate assembly of the plasmid, confirming the expected size of 6026 bp and the correct incorporation of the mCherry gene in the newly assembled plasmid. The sequencing data, as illustrated in Annex 3, so validating the use of the plasmid for further experiments.

The plasmid was then prepared for cloning SpyTag libraries. Primers were again designed for inverse PCR (Table 9.2, Annex 1) to amplify the whole plasmid, omitting SpyTag and an unwanted ribosome binding site and incorporating BsaI restriction sites. However, upon

analysis of the PCR products, additional smaller fragments (~700 bp) were observed along with the expected 6002 bp band (Figure 5.2A). These smaller fragments were likely non-specific amplification products, which required further optimisation of the PCR conditions. To improve the specificity of the amplification and enhance the intensity of the desired 6002 bp band, PCR conditions were adjusted by changing factors such as annealing temperature, number of cycles and the primer concentrations. This optimisation included modifying reaction components and cycling conditions to reduce non-specific amplification and increase the yield of the target fragment. As a result, in subsequent experiments, the intensity of the expected 6002 bp band was significantly improved, but it proved impossible to eliminate the smaller amplicon entirely (Figure 5.2B). Accordingly, the target fragments were excised from the agarose gel and purified using a gel extraction kit (2.2.2.5.3), which effectively concentrated the desired plasmid product and removed any contaminating sequences. This purified product, now containing BsaI sites, was ready for the ligation of the amplified SpyTag library cassettes, facilitating the assembly of the SpyTag-mCherry plasmid with the mutated SpyTag libraries. This preparation step was crucial for the subsequent Golden Gate cloning, ensuring that only the desired fragments were used in the ligation reactions.

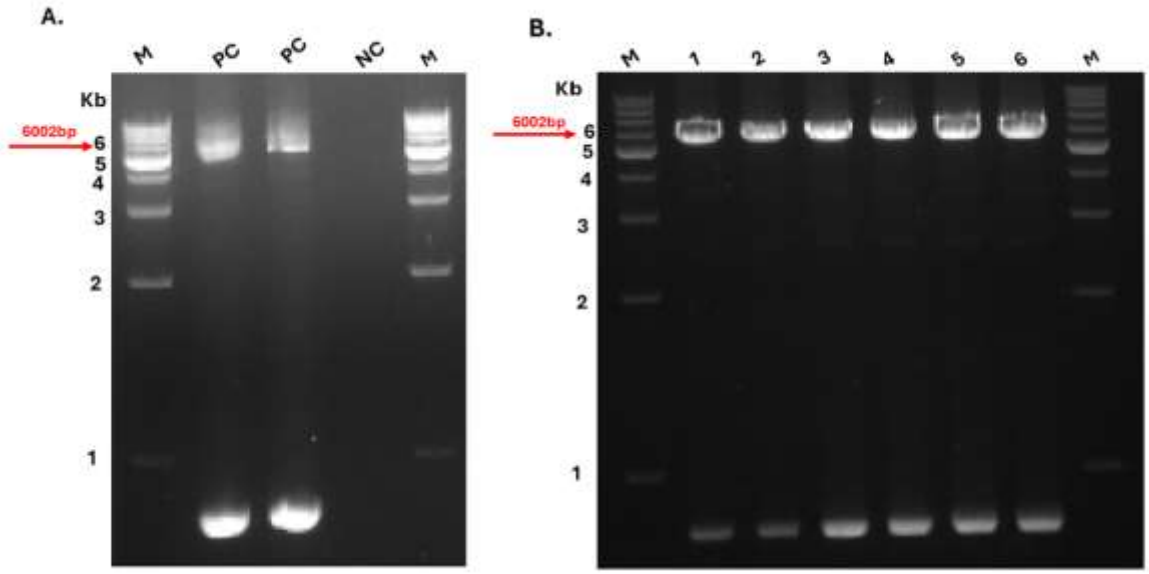


Figure 5.2. Amplification of the SpyTag-mCherry plasmid for cloning. ***(A)*** *Inverse PCR amplification of ST-mCherry plasmid, to produce an overhang-containing product with BsaI, was electrophoresed in a 1% gel, stained with ethidium bromide.* ***(B)*** *Optimised inverse PCR amplification, intended for gel purification was electrophoresed in a 1% gel, stained with ethidium bromide. Lanes: PC, positive control; NC, negative control (no template); M; 1 kb molecular weight marker.*

5.2. SpyTag library design

As detailed in Chapter 3, two critical residues of the SpyTag peptide had been selected for saturation mutagenesis. These residues were chosen based on their key roles in determining the specificity of the interaction between the SpyTag and SpyCatcher (Figure 5.3).

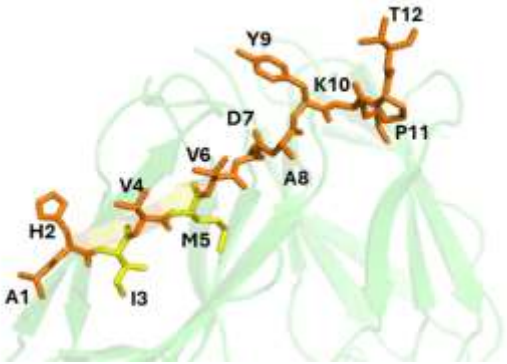


Figure 5.3. The amino acid sequence of SpyTag peptide. *The SpyTag peptide is illustrated with residues from 1 to 12 represented in stick format, except for the final residue that is not visible in the crystal structure (PDB:4MLI). The two key residues targeted for saturation mutagenesis; isoleucine at position 3 (I3) and methionine at position 5 (M5), are highlighted in yellow. The SpyCatcher protein (to which SpyTag binds) is rendered as a green ribbon in the background.*

Owing to the proximity of these residues, MAX randomisation was selected as the method of saturation for these positions and again, the same six amino acids (phenylalanine, isoleucine, leucine, methionine, tyrosine and valine) were selected for the saturation. Twelve positionally-fixed SpyTag sub-libraries were designed, fixing one position at a time and randomising the other, such that each library contained 6 unique gene variants (Figure 5.4).

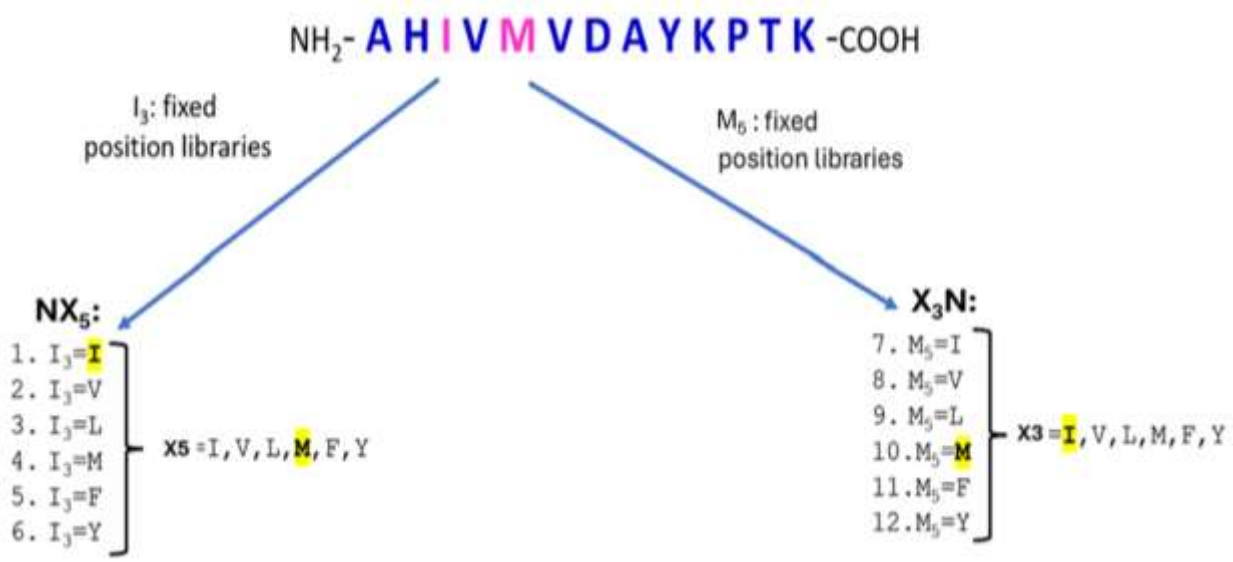


Figure 5.4. Schematic representation of SpyTag library design. *The amino acid sequence of SpyTag peptide is shown in blue. Key residues are highlighted in pink as isoleucine(I) and methionine(M). A total of 12 SpyTag libraries were designed, with position 3 fixed and position 5 partially randomised (NX5) or alternatively, with position 3 partially randomised and position 5 fixed (X3N). Only two of these libraries contain the native SpyTag sequence, highlighted in yellow (library number 1 and 10) (where X = all six possible residues and N = one of six hydrophobic residues).*

MAX selection oligonucleotides targeting positions 3 and 5 of the SpyTag peptide sequence were designed and synthesised for each saturated position. As depicted in Figure 5.5 a template containing NNN, along with the cassette start and end regions, was also created. The MAX randomisation process utilises a unique 6-base pair sequence to enable the specific annealing of the MAX selection oligonucleotides. This region ensures precise binding and accurate incorporation of the desired mutations during the peptide library construction.

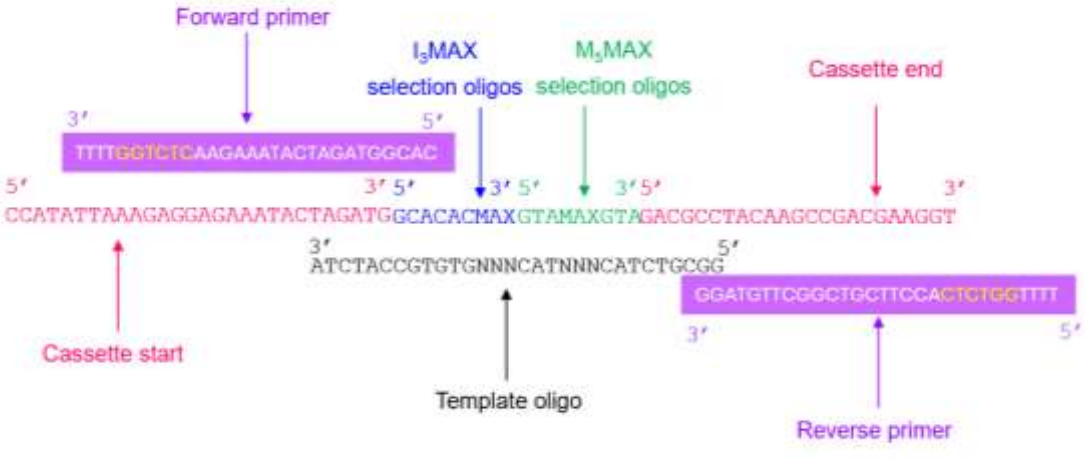


Figure 5.5. Schematic representation of the DNA oligonucleotide design with MAX selection oligonucleotides for constructing the SpyTag library. *The sequence illustrates the design strategy for the SpyTag MAX library. The cassette sequence is marked in pink, indicating the start and end regions that create overhangs. Positions I3 and M5, targeted for saturation mutagenesis, are labelled as I3MAX (blue) and M5MAX (green), with each oligo complementary to a specific region of the template. The synthetic template, highlighted in black, includes randomised codons (NNN) at the targeted positions.*

The sequences of the individual selection oligonucleotides, illustrated collectively as “MAX” in Figure 5.5 are listed in Table 5.1.

	Position	I3	M5
Identity of MAX codon	**Sequence showing MAX position**	**GCACACMAX**	**GTAMAXGTA**
	Isolucine(I)	GCACACATT	GTAATTGTA
	Valine(V)	GCACACGTG	GTAGTGGTA
	Leucine(L)	GCACACCTG	GTACTGGTA
	Phenylalanine(F)	GCACACTTT	GTATTTGTA
	Methionine(M)	GCACACATG	GTAATGGTA
	Tyrosine(Y)	GCACACTAT	GTATATGTA

Table 5.1. Sequences of MAX selection oligonucleotide pools created for saturation mutagenesis at each target position in the SpyTag peptide. *The 9mer MAX selection oligonucleotides comprise 3 bases of one MAX codon and a 6 base conserved region to ensure accurate annealing to the template (Figure 5.3). The codons chosen to encode each amino acid were*

selected based on the codon usage preferences of Escherichia coli, utilising the GenScript Codon Frequency Table for Escherichia coli, since E. coli was the chosen organism for gene expression.

5.3. Construction of SpyTag libraries

Before beginning the construction of the libraries, pools of MAX selection oligonucleotides, each corresponding to a specific target position, were prepared (2.2.2.4.1). Each MAX selection oligonucleotide was received at a concentration of 100 µM in aqueous solution. To create the MAX selection oligonucleotide pool, equal volumes of either NX5 or X3N oligonucleotides (Figure 5.4) were combined and then diluted by 50%. The final concentration of each selection pool was 50 µM, with the concentration of individual MAX selection oligonucleotides at 8.33 µM. A test MAX randomisation reaction (2.2.2.4.2) was then performed by hybridising together the template, cassette start and cassette end oligonucleotides (Figure 5.5) along with the I3 MAX pool and M5 isoleucine oligonucleotides (Table 5.1). Previous studies (Huang et al., 2022) have shown that the denaturing temperature is most important when creating a MAX randomisation cassette and accordingly, following ligation, the resulting cassette was amplified with forward and reverse primers (Figure 5.5) using a denaturation temperature gradient. The resulting products were examined by electrophoresis (Figure 5.6).

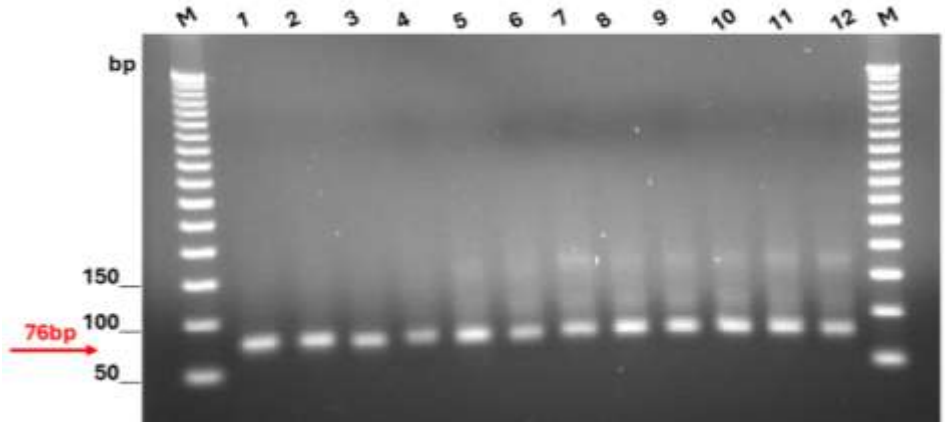


Figure 5.6. Denaturation temperature gradient using neat ligation product as template to determine the optimal denaturation temperature for amplification of library fragments. *A single PCR reaction was divided into 12 equal parts. Denaturation was performed at the temperatures listed below, with annealing at 60 °C and an in-cycle extension time of 10 seconds. Lanes: M; 50 bp MW ladder. Denaturation temperatures: 1. 80.0°C, 2. 80.3°C, 3. 81.2°C, 4. 82.6°C, 5. 84.6°C, 6. 86.5°C, 7. 88.4°C, 8. 90.3°C, 9. 92.2°C, 10. 93.7°C, 11. 94.6°C, 12. 95.0°C.*

As can be seen in Figure 5.6, a 76 bp product was observed across all denaturation temperatures, although higher denaturation temperatures led to some concatemer formation. Accordingly 80.0°C was selected as the preferred denaturation temperature for MAX randomisation (2.2.2.4.2) was then performed 12 times under the optimised conditions to

create all twelve SpyTag library fragments, which were examined by electrophoresis (Figure 5.7).

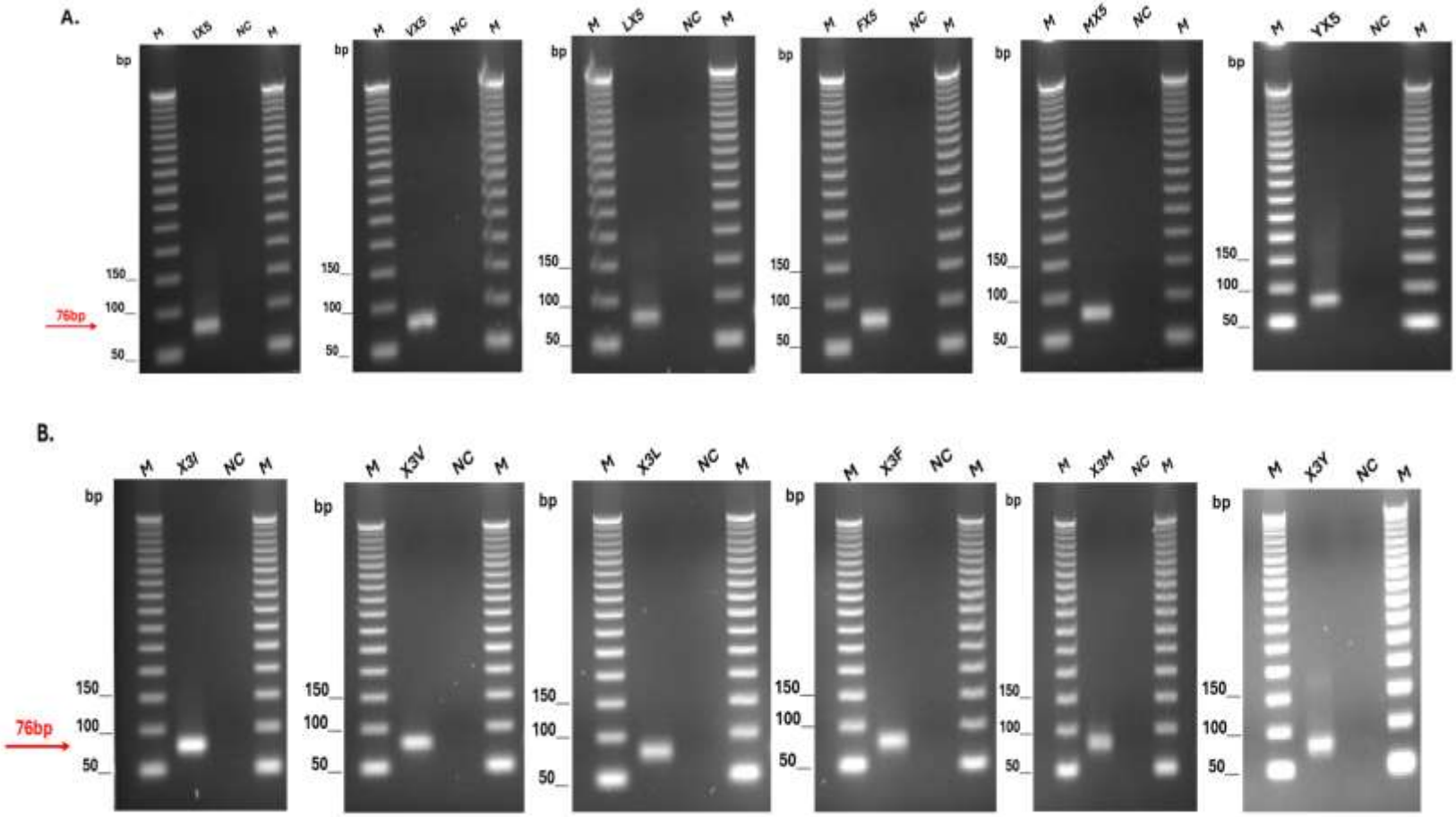


***Figure 5.7. Agarose gel electrophoresis of 12 SpyTag library fragments (A**). Position 3-fixed and position 5 partially randomised (NX5) libraries; **(B)** Position-5 fixed and position 3 partially randomised (X3N) libraries. Fragments were amplified using the primers illustrated in Figure 5.5. The resulting PCR reactions were electrophoresed in a 3% agarose gel and stained with ethidium bromide, where each sample represents 20% of a 50µl PCR reaction. Lanes: NC (negative control, no template), M; 50bp ladder MW marker.*

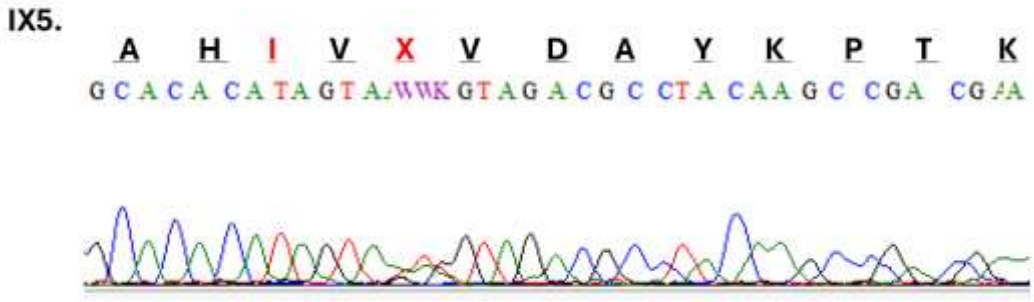
As seen in Figure 5.7, all 12 SpyTag library cassettes were of the expected size of 76 bp. Although two libraries, IX5 and LX5, exhibited slightly smeared bands, suggesting possible minor inconsistencies in amplification, they were deemed acceptable for downstream applications. The SpyTag libraries were named as follows: IX5, VX5, LX5, FX5, MX5, YX5/ X3I, X3V, X3L, X3F, X3M, and X3Y, where "X" represents all of six possible codon combinations.

5.4. Cloning of SpyTag libraries

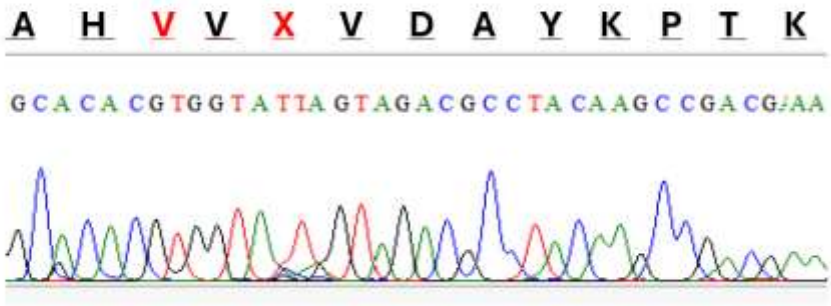
The library fragments created in section 5.3 were purified using a PCR purification kit (2.2.2.5.1) and were then joined with the plasmid backbone via Golden Gate assembly (2.2.2.2) using a 3:1 molar ratio of library fragments to backbone. The resulting ligation products were transformed into chemically competent *E. coli* DH5α cells (2.2.5) and the cells were immediately inoculated into a 50 mL starter culture containing 50 µg/mL kanamycin. Once adequate growth was achieved in the selective medium, the bacterial cultures were harvested for plasmid DNA extraction using a mini-prep protocol (2.2.2.5.2).

5.5. Sanger sequencing of the 12 SpyTag libraries

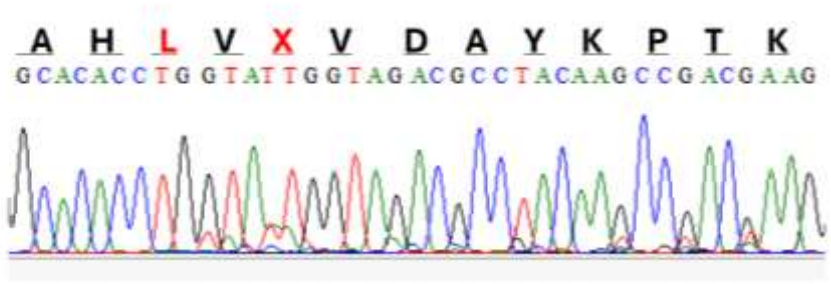
Since the number of variant clones in each library was so small (six components per library only), rather than sequencing by NGS, each plasmid library was purified using the Zymoresearch kit, the concentrations were normalised and the libraries sent for Sanger sequencing (2.2.6.1). Each library's chromatogram was analysed using the BioEdit sequence alignment tool, with the quality of the chromatograms assessed based on well-defined sharp peaks and a clear baseline (Figure 5.8).



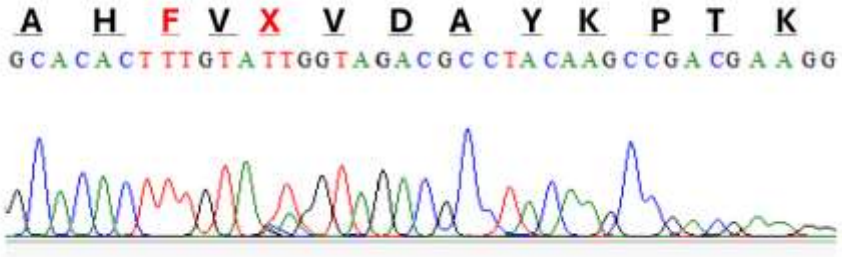
VX5.



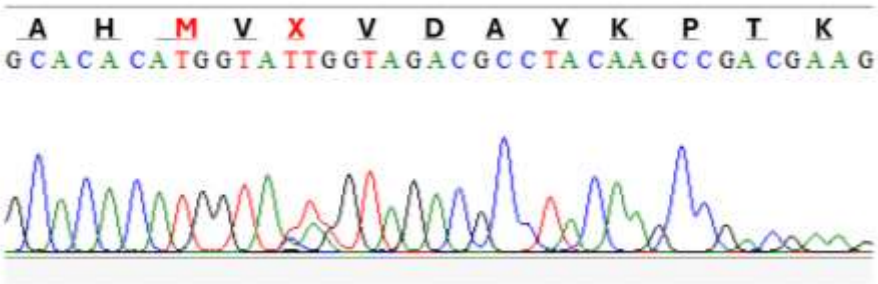
LX5.



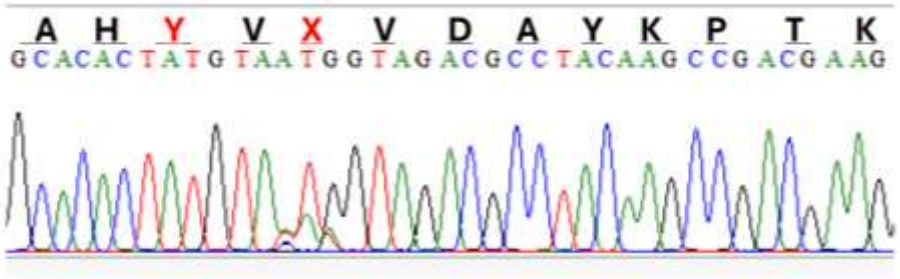
FX5.



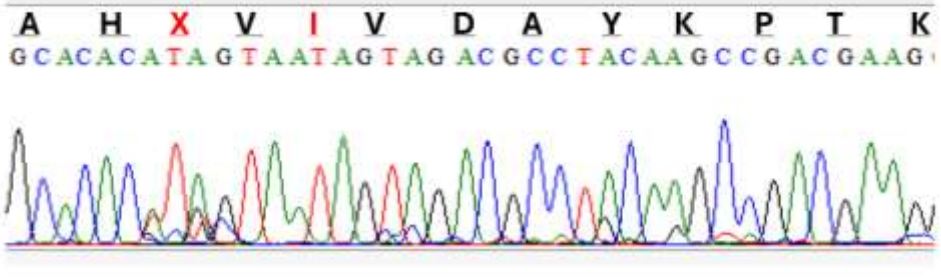
MX5.



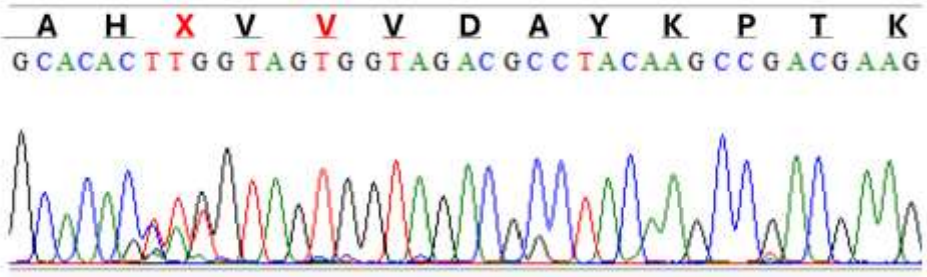
YX5.



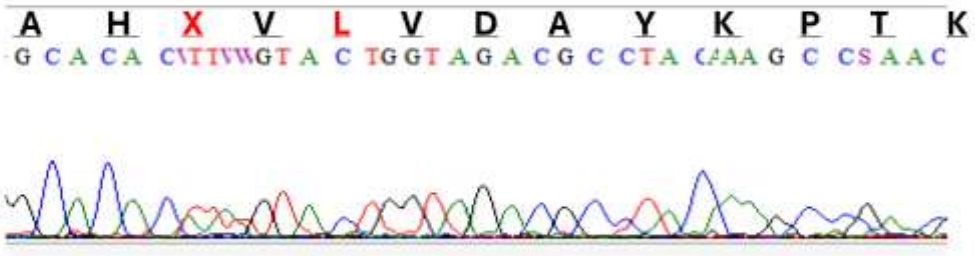
X3I.



X3V



X3L.



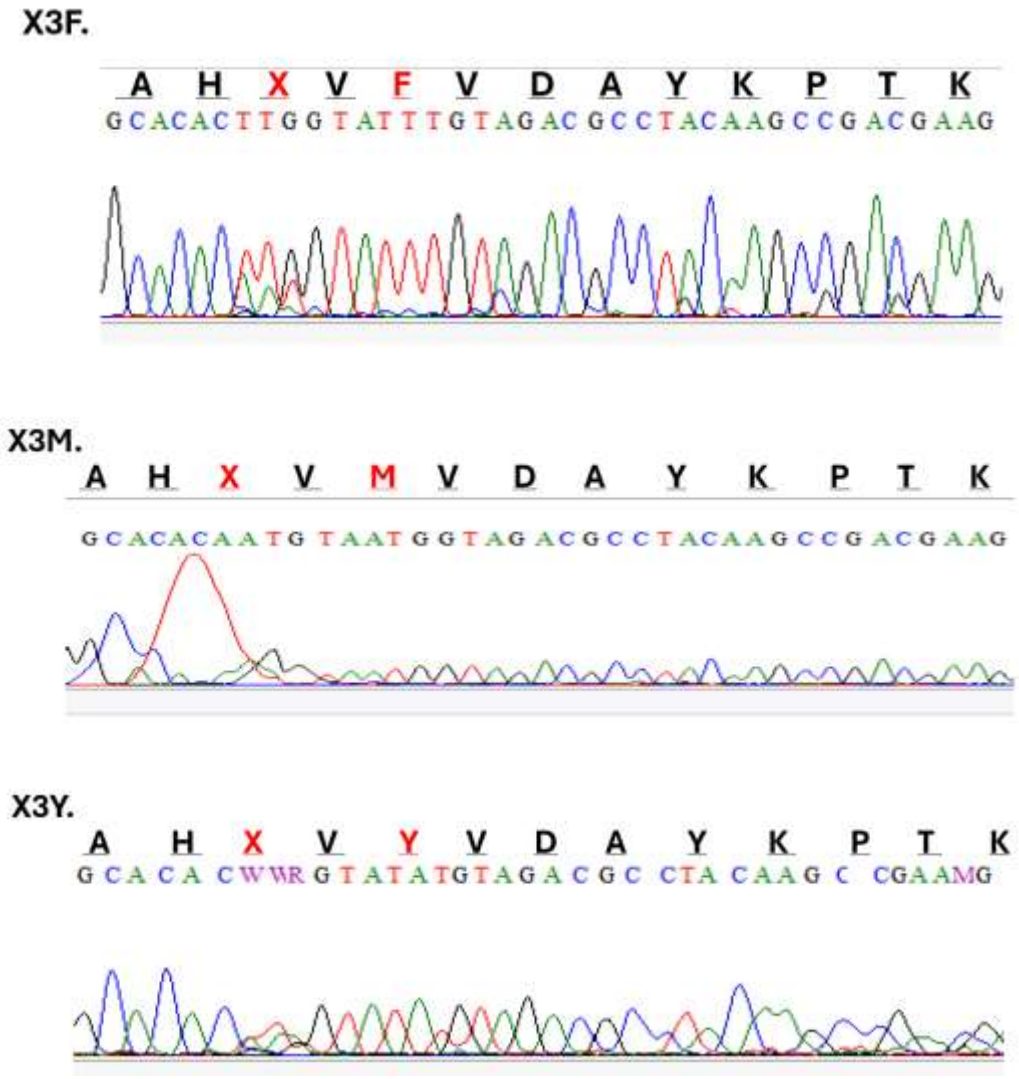


Figure 5.8. Sanger sequencing of SpyTag-libraries. *A segment of the chromatogram corresponding to the SpyTag peptide is presented, with the residue sequence displayed above the codons. This sequence represents 13 residues of the SpyTag. Fixed and randomised positions are highlighted in red, where "X" denotes the randomised codon.*

As illustrated in Figure 5.8, Sanger sequencing confirmed the presence of both the fixed and randomised codons .

5.6. SpyTag library expression

After confirming the sequences of the SpyTag libraries by Sanger sequencing, the plasmid libraries were transformed into chemically competent *E. coli* Tuner (DE3) cells (2.2.5). A transformation of the native SpyTag-mCherry plasmid was also performed. All libraries plus SpyTag-mCherry were cultured in 200 ml volumes and expression induced with 1.0 mM IPTG, with subsequent growth at 20°C overnight (2.2.7). As a preliminary test of expression, a 50 µL

aliquot of native SpyTag-mCherry was plated onto Luria-Broth (LB) agar. Fluorescent colonies, visible due to the inclusion of the mCherry fluorescent gene, appeared (Figure 5.9).

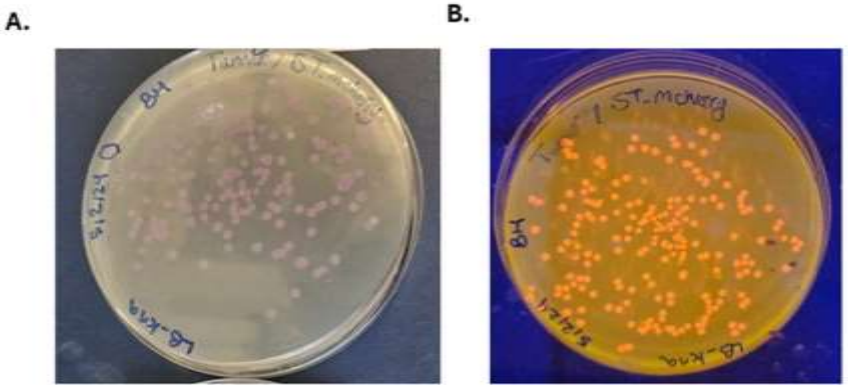


Figure 5.9. Fluorescent colonies of SpyTag-mCherry. *Transformed cells were plated on LB agar containing kanamycin and incubated overnight to select for successfully transformed colonies. The colonies were observed (A) without UV light and (B) under a UV transilluminator.*

Cells were harvested by centrifugation and cell lysates were prepared using BugBuster reagent (2.2.8) to ensure efficient protein extraction. The resulting proteins were purified using nickel-NTA resins affinity chromatography (2.2.9). The various stages of protein purification were examined by PAGE analysis. An exemplar purification of native SpyTag-mCherry protein is illustrated in Figure 5.10.

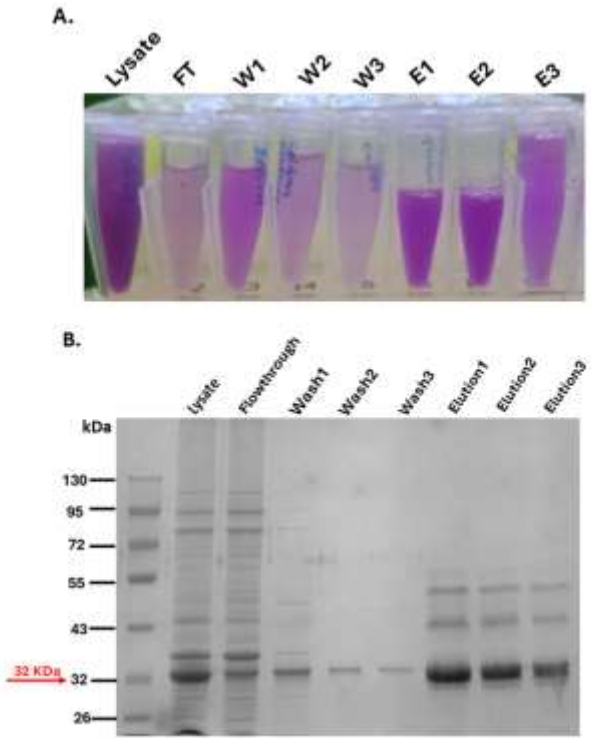


Figure 5.10. Purification of native SpyTag-mCherry protein by affinity chromatography using Ni-NTA resin. (A) *Protein fractions collected during the purification process.* ***(B)*** *Polyacrylamide gel*

electrophoresis (12% SDS-PAGE) of samples from the purification process. The purification was performed using affinity column chromatography, with the expected protein size being 32 kDa (including the His tag). The gel was stained with InstantBlue.

Figure 5.10B illustrates a prominent band near the 32 kDa marker, corresponding to the SpyTag-mCherry protein, along with a small amount of co-purified protein.

5.7. Discussion

The approach used in designing and constructing the SpyTag peptide libraries via MAX randomisation, has yielded promising results, as evidenced by the successful sequencing transformation, expression, and subsequent purification of the randomised libraries. The MAX randomisation method was employed since MAX randomisation specifically optimises the codon usage to minimise redundancy, reducing the number of variants while retaining a significant level of genetic diversity. One of the key advantages of the MAX randomisation approach lies in its ability to focus on the generation of a comprehensive set of mutants while using a smaller pool of variants, reducing the overall complexity of the library screening process, while traditional saturation mutagenesis methods often suffer from excessive library sizes, which can make high-throughput screening laborious and inefficient. The MAX strategy effectively addresses this issue, allowing the creation of libraries with a manageable number of mutants that still capture a wide range of functional diversity (Hughes et al., 2003).

The success of the library construction was validated through Sanger sequencing, after cloning the libraries into the host organism, which confirmed the presence of the expected randomised and fixed codons within the libraries.

Following the cloning of the SpyTag libraries into the mCherry-plasmid, transformation into *E. coli* cells resulted in the expression of functional SpyTag variants, as evidenced by fusion to the mCherry fluorescent reporter gene. The expression of the SpyTag libraries in Tuner (DE3) cells under optimised conditions (20°C, 1.0 mM IPTG) yielded high levels of soluble recombinant protein.

In conclusion, through the use of MAX randomisation, the study achieved its goal of creating distinct SpyTag libraries for subsequent screening experiments.

Chapter 6 Interactions of native and mutant SpyTag-SpyCatcher proteins

6.1. Interactions of native and mutant SpyTag-SpyCatcher proteins

In this chapter, both the specificity of native SpyCatcher (as a fusion to mNeonGreen) and native SpyTag (as a fusion to mCherry) have been evaluated by examining interactions of the native fusion proteins with respective libraries generated in chapters 4 and 5.

6.1.1. Interaction of native-SpyCatcher with native-SpyTag

As an initial assessment of the expected interaction, native SpyCatcher-mNeonGreen fusion protein (hereafter termed “native SpyCatcher”) was incubated with native-SpyTag-mCherry fusion protein (hereafter termed “native SpyTag”) under published conditions (2.2.10) to determine the baseline results and characteristics of the resulting SpyTag-SpyCatcher product. As discussed in Chapter 1, a highly-specific covalent bond is formed spontaneously following binding between SpyCatcher and its cognate SpyTag. Clearly the resulting product will remain intact under the denaturing conditions of SDS-PAGE, permitting visualisation of the combined product of molecular weight 72 kDa on an SDS-PAGE gel.

Accordingly, purified native SpyTag-mCherry protein (30 kDa) and native SpyCatcher-mNeongreen protein (42 kDa) were allowed to interact (2.2.10) and the product examined by SDS-PAGE (Figure 6.1). This result would serve as a comparison of binding efficiency and bond formation during analysis of the SpyTag and SpyCatcher libraries created in this project.

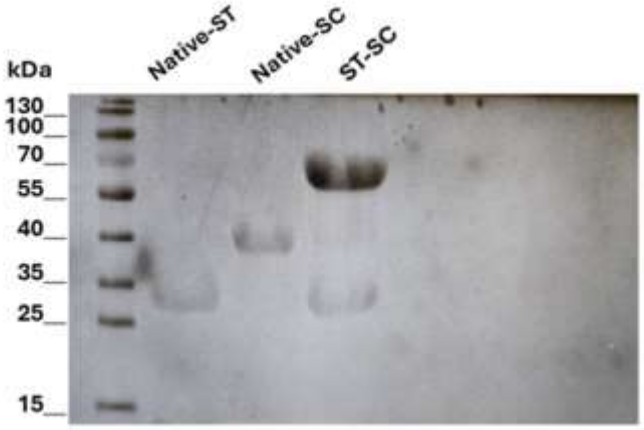


Figure 6.1. Covalent binding reconstitution between native-SpyTag (ST) and native-SpyCatcher (SC). *Polyacrylamide gel electrophoresis (SDS-PAGE, 4-12% gradient) of purified SpyTag-mCherry protein, SpyCatcher-mNeonGreen protein, and their covalent complex following incubation to assess their interaction. The gel was stained with InstantBlue. 10 µg of protein per sample was loaded into each well.*

From Figure 6.1, it can be seen that the purified SpyTag-mCherry and SpyCatcher proteins migrate in SDS-PAGE approximately as would be expected (i.e. 30 kDa and 42 kDa respectively), whilst the major product of their covalent bonding migrates at ~70 kDa , which corresponds well with the expected covalent product (ST-SC) of 72 kDa.

6.1.2. Specificity of native SpyCatcher for SpyTag

To evaluate the specificity of native SpyCatcher for its SpyTag partner, incubations of native SpyCatcher with each of the 12 SpyTag libraries (5.2) were performed using a 1:6 molar ratio of SpyCatcher:SpyTag libraries as described (2.2.10). This ratio was chosen since each SpyTag library contains six variants.

All incubated samples were analysed initially by SDS-PAGE to evaluate isopeptide bond formation (Figure 6.2). Prior to staining, the resulting gels were examined on a transilluminator. Interestingly, while the mNeonGreen-SpyCatcher fusion was not visible on the gel (as would be expected), the mCherry fluor (fused to SpyTag and SpyTag variants) survived SDS-PAGE conditions and remained visible).

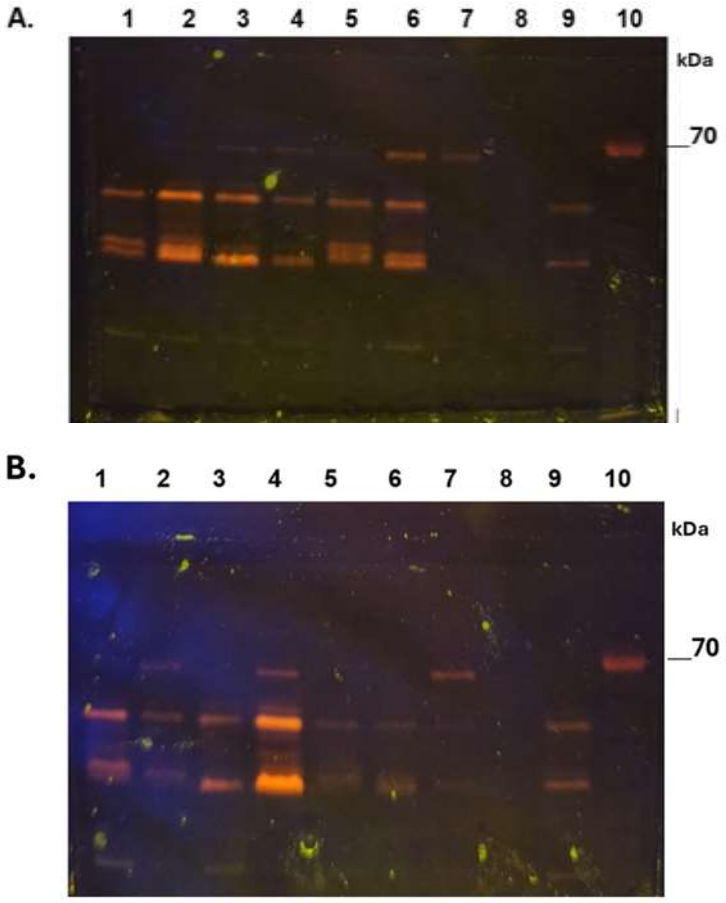


Figure 6.2. Covalent binding between native-SpyCatcher and SpyTag libraries. Polyacrylamide gel electrophoresis (SDS-PAGE, 4-12% gradient). (A) *fixed-position 3 SpyTag libraries**, (B)** fixed position 5 SpyTag libraries, visualised by UV transillumination. Library fixed identities: lanes 1, tyrosine;*

2, methionine; 3, phenylalanine; 4, leucine; 5, valine; 6, isoleucine. Lanes 7, native SpyCatcher-native SpyTag complex; 8, native SpyCatcher-mNeonGreen; 9, native SpyTag-mCherry; 10, MW marker.

Clearly in these experiments, there has been some degradation of the SpyTag-mCherry fusion as evidenced by the lower band in lanes 1-7 and 9, but nevertheless, the 72 kDa complex is clearly visible, both in the native SpyCatcher-native SpyTag control (lane 7) and in certain of the SpyTag library analyses, as described below. Note that the excess of protein in lanes 1-6 results from the molar excess of SpyTag fusion proteins in these lanes.

6.1.3. Specificity of SpyCatcher for position 3 of SpyTag as assessed by SDS-PAGE

Native SpyTag contains isoleucine at position 3 and it is clear that this amino acid is favoured by SpyCatcher at position 3 of its cognate SpyTag (compare Figure 6.2A, lanes 6&7). Nevertheless, very low levels of the 72 kDa SpyCatcher-SpyTag complex are also visible in Figure 6.2A lanes 2-5, where position 3 is fixed as methionine, phenylalanine, leucine and valine respectively, noting that these libraries contain no isoleucine at position 3 of SpyTag. Conversely, substitution of tyrosine for isoleucine at position 3 of SpyTag eliminates the interaction between SpyCatcher and SpyTag (Figure 6.2A, lane 1).

6.1.4. Specificity of SpyCatcher for position 5 of SpyTag as assessed by SDS-PAGE

Native SpyTag contains methionine at position 5. Here, Figure 6.2B demonstrates that SpyCatcher is more discriminating at that position of SpyTag. As would be expected, clear complexation is seen between the native SpyCatcher and SpyTag (Figure 6.2B lane 7) and the one fixed-position library that contains native SpyTag (Figure 6.2B lane 2), but interestingly, Figure 6.2B lane 4 demonstrates that leucine is also tolerated at position 5 of the SpyTag. Conversely, neither tyrosine, phenylalanine, valine or isoleucine (Figure 6.2B lanes 1, 3, 5 & 6) can substitute for methionine at position 5 of SpyTag, as assessed by SDS-PAGE.

6.1.5. Mass photometry analysis of the native SpyCatcher - native SpyTag interaction

As described in chapter 1, mass photometry enables detection of label-free molecular interactions within a minimum MW cut-off of approximately 30 kDa. Initially, as with the PAGE analysis, native SpyCatcher-mNeonGreen, native SpyTag-mCherry and a complex of the two proteins in a 1:1 ratio were examined by mass photometry (2.2.11) to determine viability of this method of analysis and if successful, to establish a baseline for subsequent assessment of the fixed position SpyCatcher libraries.

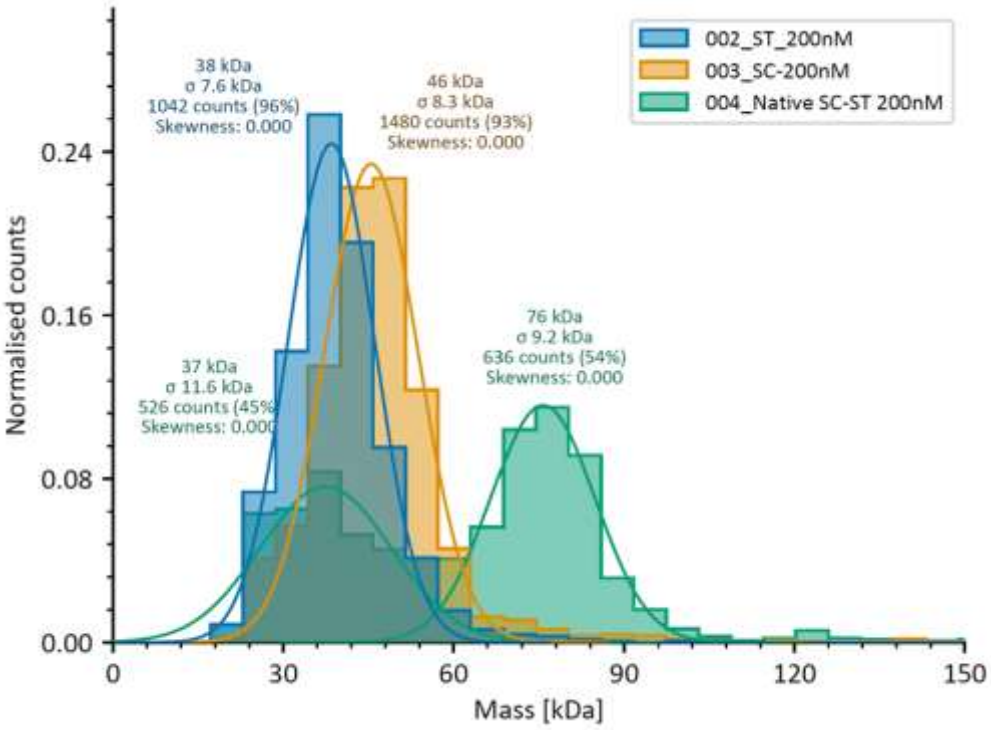


Figure 6.3. Mass photometry analysis of native peptide (ST), native protein (SC) and their interaction (SC-ST). *The graph illustrates the molecular mass assessment of individual SpyTag (ST_200nM, blue) and SpyCatcher protein (SC_200nM, orange) components and the results of their interaction (Native SC-ST_200nM, green). Each dataset represents normalised counts within a specific colour (results of each measurement have been overlayed to permit comparison).*

As illustrated in Figure 6.3 which represents an overlay of three independent spectra, mass photometry was indeed able to detect both the component Spy Catcher and SpyTag proteins and the covalently bonded product that results from their interaction. SpyTag alone (ST_200nM, blue) gave a peak at 38 kDa with a standard deviation (σ) of 7.6 kDa and 1,042 counts (actual MW = 30 kDa) and representing 96% of the sample. SpyCatcher alone (SC_200nM, orange), exhibited a peak at 46 kDa (σ = 8.3 kDa) with 1,480 counts, accounting for 93% of the sample, consistent with the expected molecular weight of SpyCatcher protein at 42 kDa and suggesting a slight over-estimation of molecular weight by the technique. Following the incubation of SpyTag with SpyCatcher, a new peak (Native SC-ST_200nM, green) was evident at 76 kDa (σ = 9.2 kDa), with 636 counts representing 54% of the sample and corresponding to the anticipated 72 kDa covalent SpyTag-SpyCatcher product, whilst the remaining material in this sample (37 kDa, 526 counts, 45%) represents a combination of unreacted SpyTag and SpyCatcher fusion proteins. Collectively, the covalently bonded and individual components of the reaction comprise 99% (54% + 45%) of detected molecules. In summary, the mass shift from the individual components to the new peak indicated successful

bonding between the two molecules, leading to the formation of a stable covalent complex with a mass approximately equivalent to the sum of SpyTag and SpyCatcher.

6.1.6. Specificity of native SpyCatcher for position 3 of SpyTag as assessed by mass photometry

Native SpyCatcher was next incubated with each of the six SpyTag libraries fixed at position 3 in a 1:6 molar ratio) and freshly diluted to the required concentration before being analysed using a mass photometer (2.2.11). Each experiment was conducted alongside native-SpyTag-native-SpyCatcher (Native SC-ST) interaction as a control for comparison (Figure 6.4).

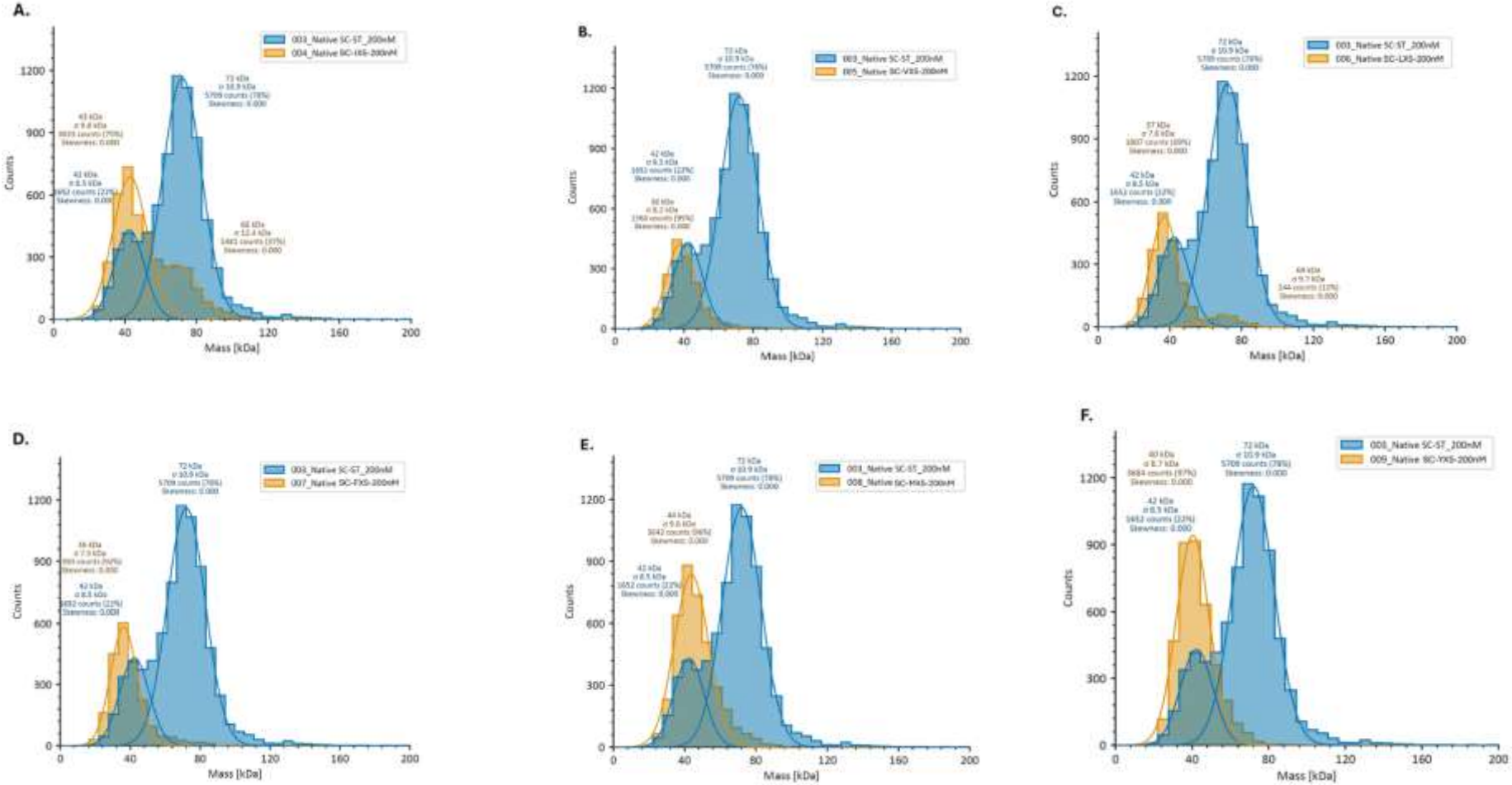


Figure 6.4. Mass Photometry analysis of native-SpyCatcher with fixed position 3 SpyTag libraries. *The spectra show the molecular mass distribution for resulting complex after the interaction between native SpyCatcher and native-SpyTag (Native SC-ST, blue), and native SpyCatcher with SpyTag libraries (SC_NX5, orange), Each dataset represents normalised counts plotted against molecular mass (kDa). Panels depict data for:* ***(A)*** *IX5,* ***(B)*** *VX5,* ***(C)*** *LX5,* ***(D)*** *FX5,* ***(E)*** *MX5 and* ***(F)*** *YX5 libraries.*

Each SpyTag library contains six variants, with position 3 of SpyTag fixed and position 5 containing a mixture of isoleucine, valine, leucine, phenylalanine, methionine, and tyrosine. As expected, the best interaction is seen between the library in which position 3 of SpyTag is fixed as the native isoleucine (Figure 6.4A), but leucine is also accepted at this position (Figure 6.4C), to a lesser extent. However, the weaker interactions with valine, methionine and phenylalanine observed by SDS-PAGE analysis (Figure 6.2A) have not been recognised by the mass photometry software, although manual inspection of Figure 6.4B, D and E arguably reveal small shoulders in the orange histogram trace around the 70-80 kDa range, when compared with Figure 6.4F, where position 3 was fixed as tyrosine which showed no interaction at all by SDS-PAGE.

This interpretation might suggest that SDS-PAGE analysis is more sensitive than mass photometry. However, the normalisation within the mass photometry plots allows more relevant comparison between libraries, since it overcomes differences in loading that can be seen in Figure 6.2A – particularly with respect to lane 2 in that figure.

6.1.7. Specificity of native SpyCatcher for position 5 of SpyTag as assessed by mass photometry

Native SpyCatcher was again incubated with each of the six SpyTag libraries fixed at position 5 in a 1:6 molar ratio) and freshly diluted to the required concentration before being analysed using a mass photometer (2.2.11). Each experiment was conducted alongside the native-SpyTag-native-SpyCatcher (Native SC-ST) interaction as a control for comparison (Figure 6.5).

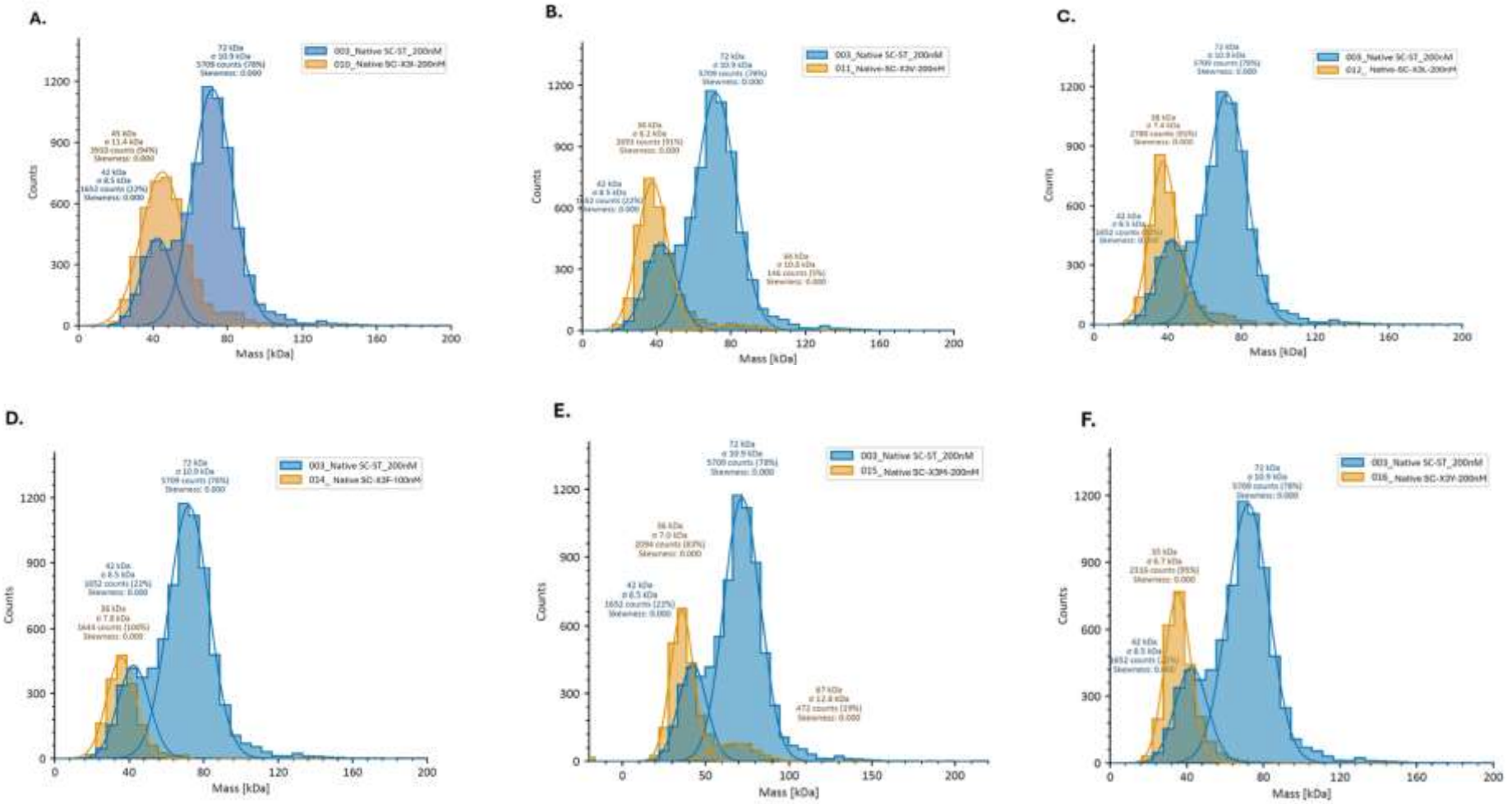


Figure 6.5. Mass photometry analysis of native SpyCatcher with fixed position 5 SpyTag libraries. *The spectra show the molecular mass distribution for resulting complex after the interaction between native SpyCatcher and native SpyTag (Native SC-ST, blue), as well as native-SpyCatcher with SpyTag-libraries (SC_X3N, orange), Each dataset represents normalised counts plotted against molecular mass (kDa). Panels depict data for:* ***(A)*** *X3I,* ***(B)*** *X3V,* ***(C)*** *X3L,* ***(D)*** *X3F,* ***(E)*** *X3M and* ***(F)*** *X3Y libraries.*

Here, the mass photometry spectra again show the native methionine as the favoured residue at position 5 of SpyTag. Again, there are visible shoulders in the histogram traces at 70-80 kDa in spectra 6.5A, B and C (isoleucine, valine and leucine at position 5 of SpyTag respectively), although the mass photometry software has only detected the interaction with respect to valine at position 5 (Figure 6.5B). This result is not entirely consistent with the SDS-PAGE analysis, where no covalently-linked product was visible with respect to isoleucine and valine at position 5, though again, differential loading versus normalisation may be responsible for the apparent disparity between the SDS-PAGE and mass photometry results at this position. Accordingly, it was decided to analyse interactions both by SDS-PAGE and mass photometry in future experiments.

6.2. Interactions of mutated SpyCatcher proteins (libraries) with native-SpyTag

Section 6.1. demonstrated, unsurprisingly, that native SpyCatcher is relatively specific for residues isoleucine and methionine at positions 3 and 5 respectively of SpyTag. It was next investigated whether mutations at the modelled positions (27, 44 and 90) within SpyCatcher would be tolerated and if so, whether the mutated proteins would have any effect on SpyTag specificity – either in terms of retained binding to native SpyTag, recognition of an alternative SpyTag sequence (i.e. an orthogonal pair) or elimination of the SpyTag interaction altogether. Chapter 4 described the synthesis of positionally-fixed SpyCatcher libraries in which one of the modelled positions 27, 44 and 90 was fixed alternately as valine, leucine, isoleucine, methionine, phenylalanine and tyrosine while the other two positions were randomised to all six residues. Each resulting library therefore contains 36 variant SpyCatcher proteins which would suggest a molar ratio of 36:1 SpyCatcher:SpyTag for subsequent assays. However, preliminary experiments suggested that an 18:1 molar ratio of protein:peptide worked well and avoided overloading the PAGE gels. Accordingly, each of the 18 positionally fixed SpyCatcher libraries was incubated with native SpyTag fusion protein at an 18:1 molar ratio as described (2.2.10).

The 18 SpyCatcher-SpyTag analyses were all examined by SDS-PAGE, organised by fixed position. Gels were visualised both before and after staining for comparison (Figure 6.6).

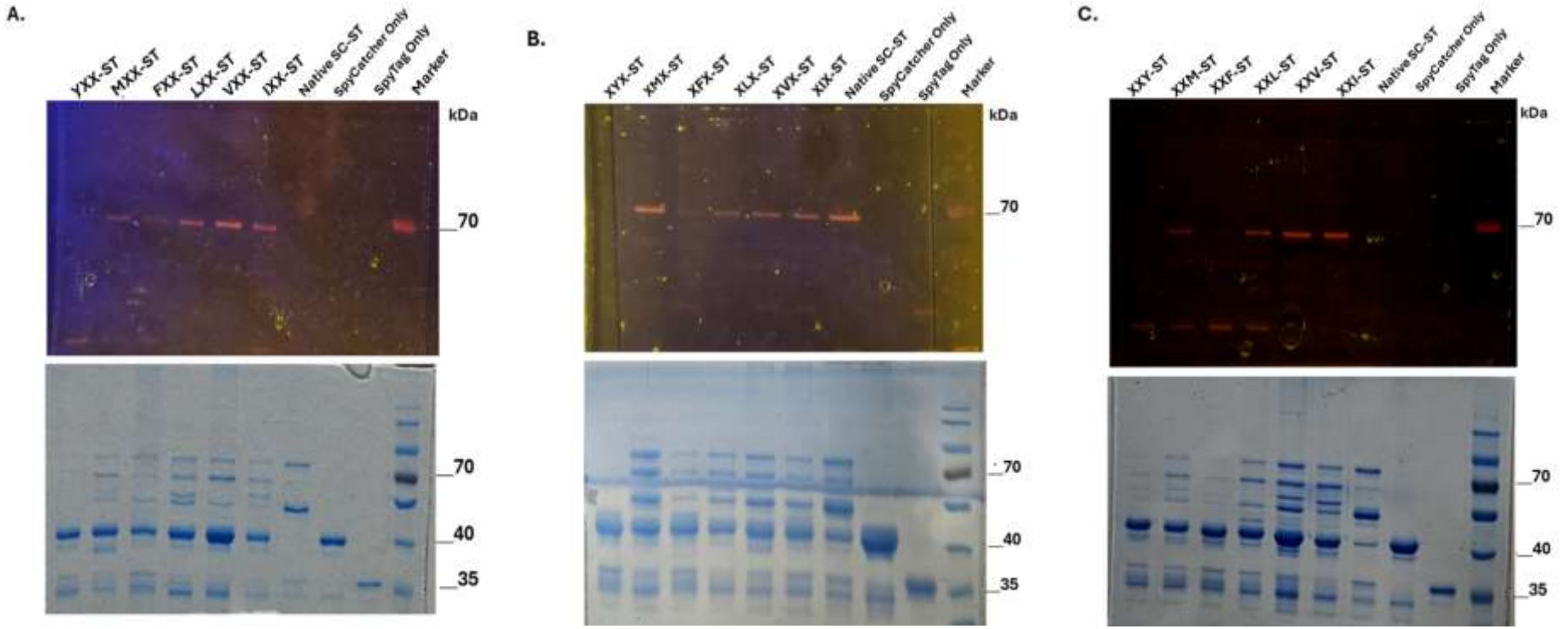


Figure 6.6. Covalent bonding analysis between fixed position SpyCatcher libraries and native SpyTag peptide. *Polyacrylamide gel electrophoresis (SDS-PAGE, 4-12% gradient) of* ***(A)*** *fixed-position 27 SpyCatcher libraries,* ***(B)*** *fixed position 44 SpyCatcher libraries****, (C)*** *fixed-position 90 SpyCatcher libraries. Fixed positions are as indicated above lanes. Gels were first visualised by transillumination and then stained with InstantBlue and visualised via white light and conventional photography.*

6.2.1. Fixed position 27 libraries (native SpyCatcher = isoleucine)

Among the fixed position 27 libraries, IXX, VXX, LXX libraries all exhibited good complexation with native SpyTag, as evidenced by intense bands at the expected molecular weight (~72 kDa) in Figure 6.6A. Of these, only the IXX library contains native SpyCatcher, meaning that any of the large, aliphatic hydrophobic amino acids can be tolerated at position 27 of SpyCatcher when forming a covalent bond with native SpyTag. Conversely, YXX library showed no complex formation, while libraries FXX and MXX displayed weak 72 kDa bands (Figure 6.6A), suggesting minimal isopeptide bond formation and that accordingly, of the residues tested, efficient interaction and subsequent isopeptide bond formation between SpyTag and SpyCatcher, a large, aliphatic, hydrophobic residue is required at position 27 of SpyCatcher.

6.2.2. Fixed position 44 libraries (native SpyCatcher = methionine)

In this set of SpyCatcher libraries, four of the six SpyCatcher libraries: XIX, XVX, XLX, and XMX demonstrated good isopeptide bond formation with native SpyTag, with the strongest interaction being visible in library XMX, which is the only one of this set of libraries to contain the native SpyCatcher. XFX displayed a minimal complex formation, while XYX libraries showed no observable interaction, again suggesting that a larger, aliphatic and hydrophobic residue is required at position 44 of SpyCatcher for isopeptide bond formation with native SpyTag (Figure 6.6B).

6.2.3. Fixed Position 90 libraries (native SpyCatcher = isoleucine)

The binding patterns in the third-fixed position were largely consistent with those observed in the second-fixed position, though with some slight variations; XXI, XXV, XXL libraries again showed strong binding, with prominent bands at the expected molecular weight. The XXM library showed a slightly weaker band, suggesting that methionine (M) in this position still allows binding and isopeptide bond formation but not as effectively as the other amino acids tested. No isopeptide bond formation was evident in libraries XXF and XXY (Figure 6.6C).

In summary, the libraries exhibited distinct behaviours, with weak or no binding observed for phenylalanine (F) and tyrosine (Y) depending on the selected positions, while leucine (L) and valine (V) generally exhibited stronger binding responses. Libraries containing the native variants, including IXX, XMX, and XXI displayed bright bands on the gels, as expected.

6.2.4. Mass photometry of the fixed-position SpyCatcher libraries with native SpyTag

As with the SpyTag libraries, analysis using mass photometry was repeated with all 18 SpyCatcher libraries, with analysis of each library interacting with native SpyTag.

6.2.4.1. Fixed-position 27 SpyCatcher libraries interacting with native SpyTag

Libraries NXX (fixed at position 27) were incubated with native SpyTag at a molar ratio of 18:1 (2.2.10) and prepared for mass photometry analysis (2.2.11). The results are presented in Figure 6.7 A-F.

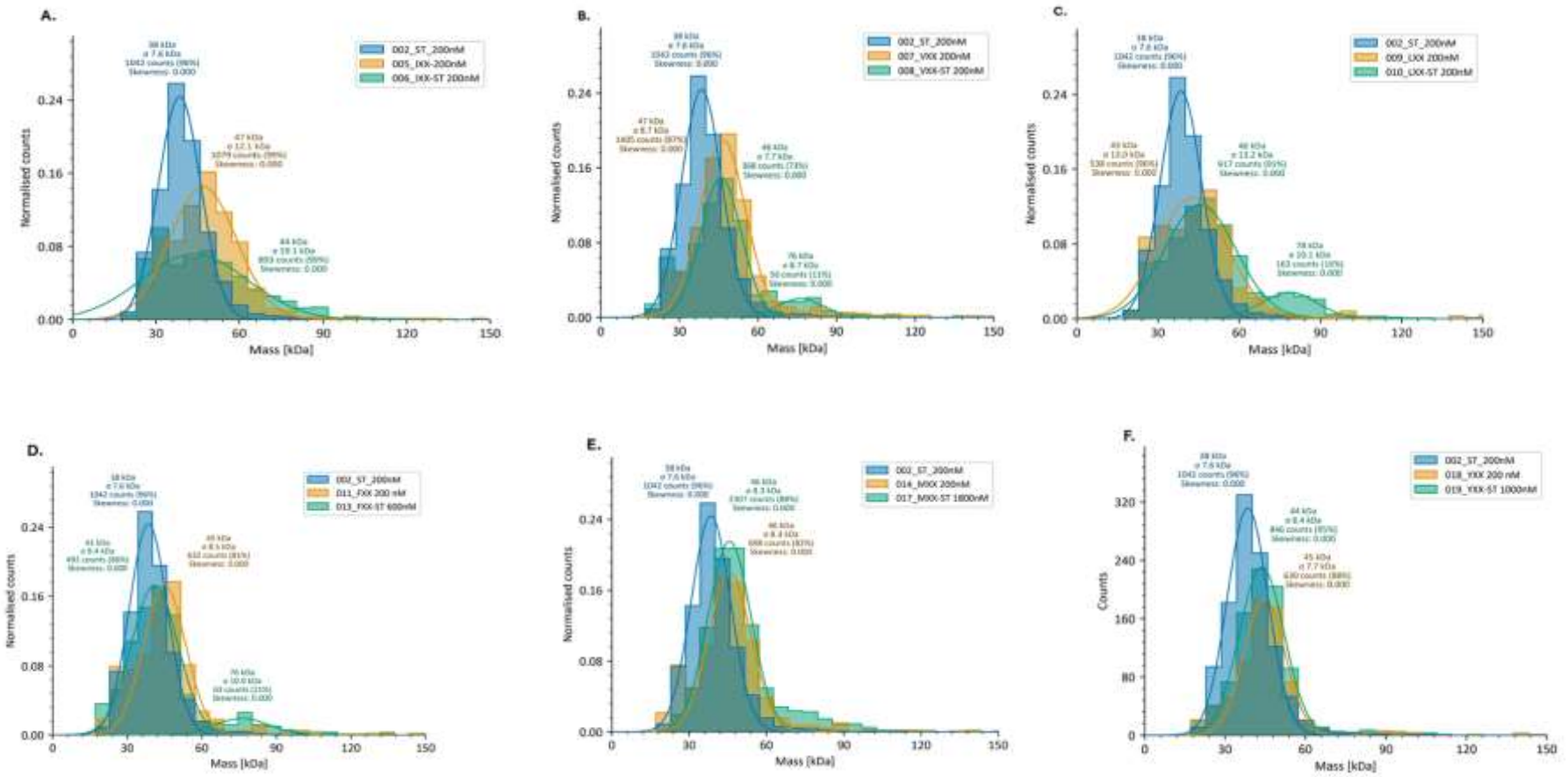


Figure 6.7. Mass Photometry analysis of fixed-position 27 SpyCatcher libraries (NXX). *Molecular mass distribution graphs show individual components and complexes of SpyCatcher-SpyTag interactions. Each panel presents normalised counts of particles as a function of molecular mass (kDa), highlighting the binding behaviour between native-SpyTag peptide (ST, 200 nM; blue), SpyCatcher libraries (NXX, 200 nM; orange), and the complex post-interaction (NXX-ST, 200 nM; green). Panels display data for the following libraries:* ***(A)*** *IXX,* ***(B)*** *VXX,* ***(C)*** *LXX,* ***(D)*** *FXX,* ***(E)*** *MXX and (****F****) YXX.*

In common with the PAGE analyses (Figure 6.6A) all fixed position libraries except for YXX showed some degree of covalent bond formation with native SpyTag as evidenced by the shoulder in the 70-90kDa regions of the green histogram traces in Figures 6.7A-E. In some cases this shoulder was detected by the mass photometry software (Figure 6.7B-D; position 27 = valine, leucine and phenylalanine respectively) while in others (Figures 6.7A and E; position 27 = isoleucine and methionine respectively) manual inspection of the spectra is required in order to see the shoulder that represents the interaction. Interestingly, this includes the only fixed position library to contain the native isoleucine residue at position 27 (Figure 6.7A). However, there is clearly a complete absence of any such shoulder in Figure 6.7F, where position 27 is fixed as tyrosine. Thus for the position 27 libraries, both SDS-PAGE analysis and mass photometry analyses are in complete agreement: of the residues tested, all large hydrophobic residues, whether aliphatic or aromatic can be tolerated at position 27 of SpyCatcher; interaction with native SpyCatcher still occurs. Conversely, introduction of a single hydroxyl moiety at position 27 eliminates native SpyTag binding beyond detectable levels.

6.2.4.2. Fixed-position 44 SpyCatcher libraries interacting with native SpyTag

The experiment in section 6.3.1 was repeated with the fixed position 44 SpyCatcher libraries and results are illustrated in Figure 6.8A-F.

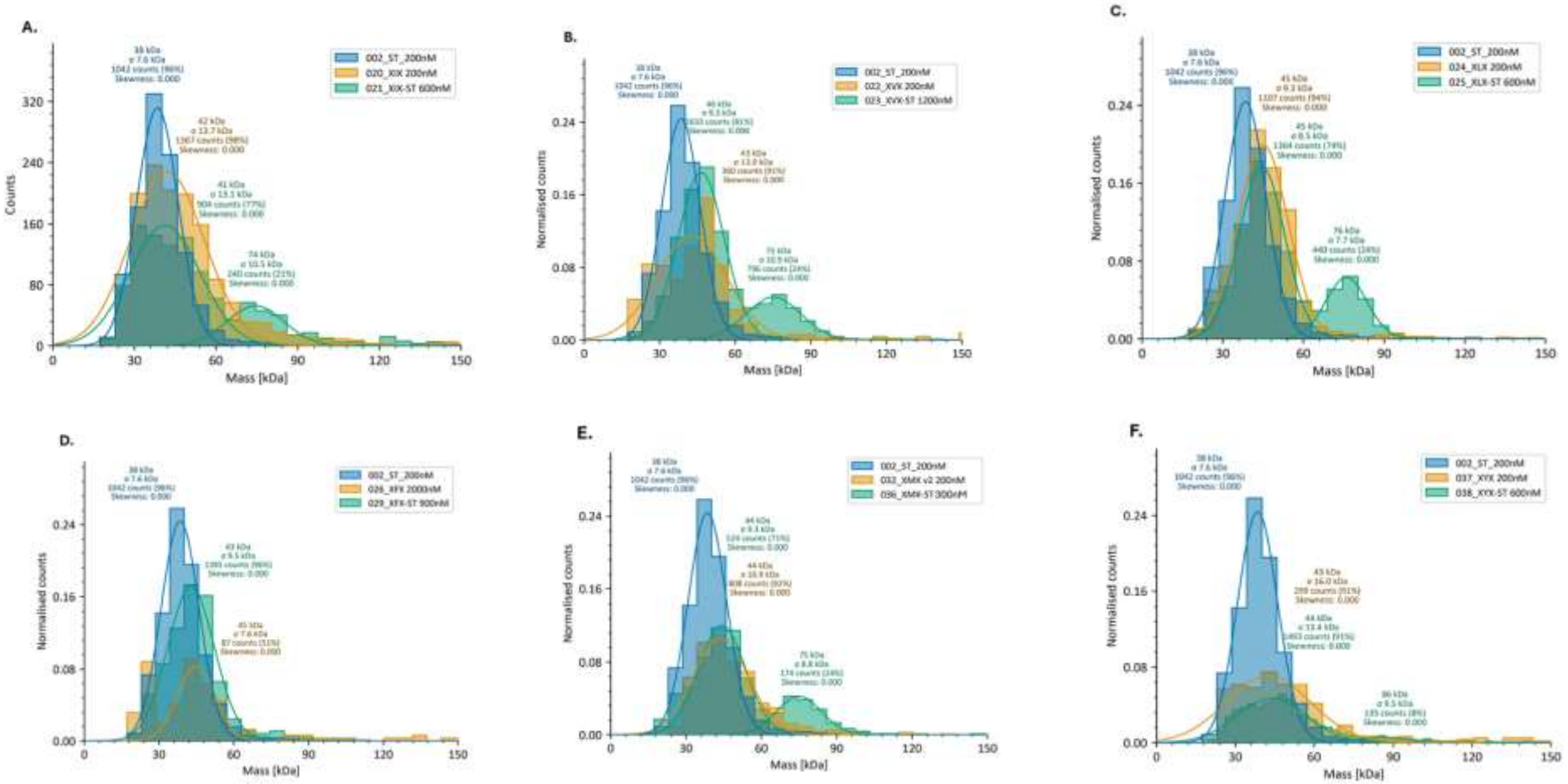


Figure 6.8. Mass Photometry analysis of fixed-position 44 SpyCatcher libraries (XNX). *Molecular mass distribution spectra show individual components and complexes of SpyCatcher-SpyTag interactions. Each panel presents normalised counts of particles as a function of molecular mass (kDa), highlighting the binding behaviour between native-SpyTag peptide (ST, 200 nM; blue), SpyCatcher libraries (XNX, 200 nM; orange), and the complex post-interaction (XNX-ST, 200 nM; green). Panels display data for the following libraries:* ***(A)*** *XIX,* ***(B)*** *XVX,* ***(C)*** *XLX,* ***(D)*** *XFX,* ***(E)*** *XMX and* ***(F)*** *XYX.*

The results from mass photometry were again mostly consistent with results obtained by SDS-PAGE analysis. Covalent 72 kDa SpyCatcher-SpyTag products can be seen clearly in Figures 6.8A-C and 6.8E, whereas no shoulder is apparent in Figure 6.8F where position 44 is fixed as tyrosine. The only inconsistency between the two methods of analysis concern library XFX where position 44 is fixed as phenylalanine. Here, SDS-PAGE analysis showed a faint 72 kDa band (Figure 6.8B) whereas no convincing shoulder is seen in Figure 6.8D.
In summary, residues that promote hydrophobic interactions (isoleucine, valine, leucine, and methionine) exhibited strong binding to SpyTag in position 44 of SpyCatcher, while the bulky aromatic residue phenylalanine is less effective and in common with position 27, introduction of a single hydroxyl group abolishes the interaction.

6.2.4.3. Fixed-position 90 SpyCatcher libraries interacting with native SpyTag

The experiment in section 6.3.1 was again repeated, this time with fixed-position 90 SpyCatcher libraries. The results are illustrated in Figure 6.9A-F.

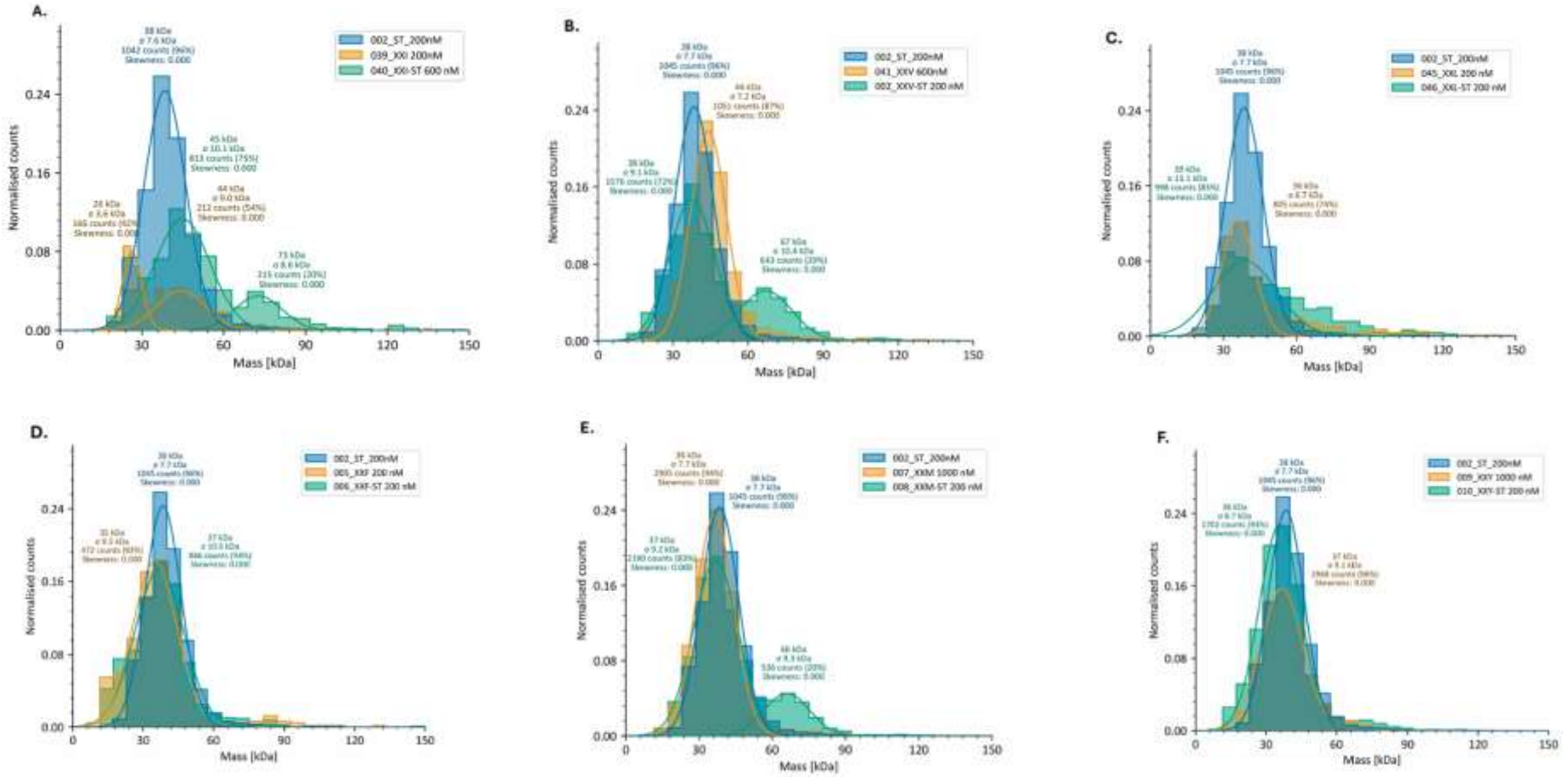


Figure 6.9. Mass Photometry analysis of fixed-position 90 SpyCatcher libraries (XXN). *Molecular mass distribution graphs show individual components and complexes of SpyCatcher-SpyTag interactions. Each panel presents normalised counts of particles as a function of molecular mass (kDa), highlighting the binding behaviour between native-SpyTag peptide (ST, 200 nM; blue), SpyCatcher libraries (XXN, 200 nM; orange), and the complex post-interaction (XXN-ST, 200 nM; green). The distribution analysis provides insights into the molecular masses and binding affinities for each fixed positional variant. Panels depict data for:* ***(A)*** *XXI,* ***(B)*** *XXV****, (C)*** *XXL****, (D)*** *XXF****, (E)*** *XXM and* ***(F)*** *XXY libraries.*

The results from mass photometry were again mostly consistent with results obtained by SDS-PAGE analysis. Covalent 72 kDa SpyCatcher-SpyTag products can be seen clearly in Figures 6.9A-B & 6.9E, whereas no shoulder is apparent Figure 6.9D & F where position 90 is fixed as phenylalanine and tyrosine. The only inconsistency between the two methods of analysis concern library XXL where position 90 is fixed as Leucine. Here, SDS-PAGE analysis showed a bright 72 kDa band (Figure 6.6C) whereas a small unconvincing shoulder is seen in Figure 6.9C. In addition, a faint band was seen on the gel for the methionine (Figure 6.6C), while covalent product is observed clearly in Figure 6.9E.

In summary, residues that promote hydrophobic interactions (isoleucine, valine, and methionine) exhibited strong binding to SpyTag in position 90 of SpyCatcher, while surprisingly, leucine at this position did not exhibit binding as determined by mass photometry. In common with positions 27 and 44, aromatic residues phenylalanine and tyrosine are less well tolerated at position 90 of SpyCatcher.

6.3. Discussion

The accurate characterisation of biomolecules is fundamental in biochemistry and molecular biology. Two techniques employed for this study are Polyacrylamide Gel Electrophoresis (PAGE) and mass photometry. While PAGE has been a cornerstone in laboratories for decades, mass photometry represents an innovative approach that has gained traction very recently. This chapter delves into the relative merits of PAGE and mass photometry analyses, critically evaluating their capabilities and limitations. Additionally, it discusses the reliability of the percentage values derived from normalised mass photometry counts.

In terms of sensitivity and detection limits, PAGE usually seems to have moderate sensitivity, often requiring amplification steps or enhanced staining methods for low-abundance proteins. Mass photometry excels in sensitivity, detecting single molecules and providing mass measurements with high precision. Regarding quantitation, PAGE offers semi-quantitative results, with quantitation relying on densitometry of bands, which can be variable, while mass photometry provides accurate quantitative data on molecular masses and population distributions. For sample requirements, PAGE requires microgram quantities of sample, which can be limiting, while mass photometry requires significantly less sample, preserving valuable or limited materials. In terms of throughput, PAGE is time-consuming, with each gel run taking several hours, while mass photometry allows rapid data acquisition, analysing multiple samples in a shorter time frame. Regarding data interpretation, PAGE results are straightforward to interpret visually even without staining while analysing fluorescent proteins

but provide limited quantitative data (Ladner-Keay et al., 2018), while mass photometry data is rich but requires specialised analysis tools and expertise.

According to the significance of normalised mass photometry counts, the percentage values obtained from normalised mass photometry counts represent the relative abundance of different molecular species within a sample. Interpreting these values requires careful consideration of several factors and as such, accurate mass determination depends on proper calibration using standards. Any deviation can affect the mass assignments and, consequently, the percentage counts. Interactions between molecules and the detection surface can lead to preferential adsorption or desorption, skewing the population distributions. Since mass photometry counts individual molecules, statistical fluctuations can impact the percentage values, especially with low-count species. There are factors affecting percentage values such as complex mixtures which may contain overlapping species, making it challenging to resolve individual populations accurately. Variations in temperature, buffer composition, and other environmental factors can influence molecule behaviour during measurement. The algorithms used for peak assignment and integration can introduce biases if not properly optimised. While mass photometry provides quantitative data, the actual percentage values should be interpreted with caution, percentage values are most reliable when comparing samples measured under identical conditions, allowing for relative comparisons rather than absolute quantitation. Small differences in percentage values may not translate to significant functional differences (Asor & Kukura, 2022).

The mass photometry results presented in this chapter demonstrate strong alignment with those obtained through SDS-PAGE analysis. While minor variations between the two methods were observed, these discrepancies provide complementary insights. Notably, PAGE appeared to exhibit greater sensitivity compared to mass photometry, contrary to the expected superior sensitivity of the latter. The analysis revealed that the SpyCatcher-SpyTag interaction is strongly influenced by specific amino acid properties, with a marked preference for hydrophobic residues of sufficient size at key positions. These findings offer valuable insights for understanding and optimising the SpyCatcher-SpyTag system for novel SpyCatcher proteins.

The analysis provided understanding of interaction frequencies between specific amino acids and their positions within both SpyCatcher and SpyTag libraries. This data highlights the preferences and specificities of these interactions, shedding light on the amino acids and positions most conducive to robust interactions. Comparative analysis of interactions between native SpyCatcher with SpyTag libraries and native SpyTag with SpyCatcher libraries yielded several insights into interaction specificity.

In the SpyTag library analysis, native isoleucine (position 3) and methionine (position 5) emerged as particularly favourable residues, as expected. Other hydrophobic amino acids, such as leucine and valine, were also tolerated by SpyCatcher at positions 3 and 5 of the SpyTag, while aromatic residues like phenylalanine and tyrosine completely disrupted binding between native SpyCatcher and variant SpyTags. The presence of leucine at position 3 suggests that certain hydrophobic residues can serve as effective substitutes, although native residues remain preferred. Discrepancies observed between SDS-PAGE results and mass photometry in these libraries may stem from differences in sample loading procedures or variations in the normalisation of count data.

In the SpyCatcher library analysis, substitution of leucine or valine at position 27 of SpyCatcher generated variant SpyCatcher proteins still capable of binding native SpyTag. Variant SpyCatcher proteins containing either phenylalanine or methionine at position 27 also showed moderate interaction with native SpyTag. Conversely, substitution of tyrosine at position 27 of SpyCatcher disrupted the interaction with native SpyTag, reaffirming the requirement for large, aliphatic, hydrophobic residues at this position. Similar preferences were observed at position 44, where variant SpyCatcher proteins containing substitution of residues isoleucine, leucine, valine, and methionine all bound native SpyTag, while bulky aromatic residues (phenylalanine and tyrosine) disrupted interactions. These findings align closely with SDS-PAGE results. At position 90, preferred residues largely mirrored those at position 44, except for leucine, which showed no binding in mass photometry despite forming a sharp band in SDS-PAGE. This discrepancy warrants further investigation. Across positions 27, 44, and 90, aromatic residues such as phenylalanine and tyrosine generally had a negative impact on interactions, although phenylalanine displayed moderate tolerance at position 44. Tyrosine, in particular, completely prevented covalent complex formation. Collectively, the data indicate a strong preference for larger, aliphatic hydrophobic residues.

In conclusion, the choice between PAGE and mass photometry—or their combined use—should be guided by specific research objectives, available resources, and the required level of analytical detail. Understanding the relative advantages and limitations of each method enables researchers to select the most appropriate tool for their studies. The consistent findings between mass photometry and SDS-PAGE in this study confirm that both methods are reliable for evaluating specificity of interaction. In addition, this analysis emphasises the role of specific amino acids and their positions in determining the specificity of SpyCatcher-SpyTag interactions. These insights are instrumental in designing and optimising these systems to ensure that the most effective amino acid combinations are employed to achieve desired interaction properties. The observed preference for hydrophobic, aliphatic residues

over polar or aromatic ones underscores the critical role of hydrophobic interactions in the formation of the SpyCatcher-SpyTag complex. This finding provides a valuable basis for further experiments aimed at identifying novel SpyCatcher variants with either altered or completely diminished specificity for SpyTag, facilitating the development of orthogonal pairs.

Chapter 7 Interactions of novel SpyCatcher variants

7.1. Creation of newly discovered SpyCatcher proteins

The simplest interpretation of the results in Chapter 6 is to hypothesise that by taking the most effective substitution from each of the three mutated positions in SpyCatcher and combining those substitutions together into a single SpyCatcher protein, it should be possible to create a novel SpyCatcher that still forms a covalent bond with the native SpyTag peptide. The purpose of such an experiment would not be to generate a new SpyCatcher protein per se, but rather to validate the principle of the positionally fixed screening protocol with respect to the SpyCatcher-SpyTag interaction.

Comparing both SDS-PAGE and mass photometry data gathered from the fixed-position SpyCatcher libraries, the “best” substitutions at positions 27, 44 and 90 were judged to be leucine, leucine and valine respectively. To test the hypothesis, three proteins were selected for individual synthesis in order of increasing risk: one that substituted just one of the selected mutations into native SpyCatcher, one that made two substitutions and a third that substituted all three residues (Table 7.1).

Conversely, again based on chapter 6 results, three further combinations of substitutions were chosen that were deemed unlikely to bind native SpyTag. Here, the aim was to determine whether it might be possible to create one or more orthogonal pairs with novel SpyTag sequences already contained within the SpyTag libraries. Substitution of tyrosine in any one of positions 27, 44 and 90 of SpyCatcher abolished interaction with native SpyTag so far as the combinatorial assays were able to determine. Substitution of phenylalanine in either position 44 or 90 was similarly deemed disruptive to the SpyCatcher-SpyTag interaction. With regards to position 27, both phenylalanine and methionine showed similar but weaker interaction with native SpyTag (Figures 6.6A and 6.7D&E), so to give a second alternative at that position (and to avoid testing only aromatic residues in all three positions) methionine was selected as an additional option for position 27. The putative “non-binders” chosen for individual synthesis, with all three positions of native SpyCatcher substituted, are listed in Table 7.1.

	SpyCatcher position		
	27	**44**	**90**
Native	I	M	I
Novel binders	I	L	I
	L	L	I
	L	L	V
Novel non-binders	Y	F	Y
	Y	F	F
	M	Y	F

Table 7.1. SpyCatcher residue substitution compared to native-SpyCatcher for novel binders and non-binders. *Substitutions are compared to the native SpyCatcher residues (I27, M44, I90), with green cells representing native identities ("Novel Binders"), yellow cells representing amino acids that retain some degree of binding to native SpyTag and red cells indicating loss of binding ("Novel Non-Binders").*

7.1.1. Creation of genes encoding novel SpyCatcher “binders”

Genes encoding individual SpyCatcher proteins ILI, LLI, LLV, YFY, YFF and MYF (Table 7.1) were constructed. Two mutated fragments were first amplified by PCR (2.2.2.1) for each gene. For example, for protein LLI, the first fragment was amplified using primers 27(L) & 44(L) reverse , while the second fragment was synthesised using primer 44(L) as the forward primer and primer 90(I) as the reverse primer (Table 9.3, Annex 1). This process was iterated for the other five genes. Gel electrophoresis was then used to examine the resulting fragments (Figure 7.1).

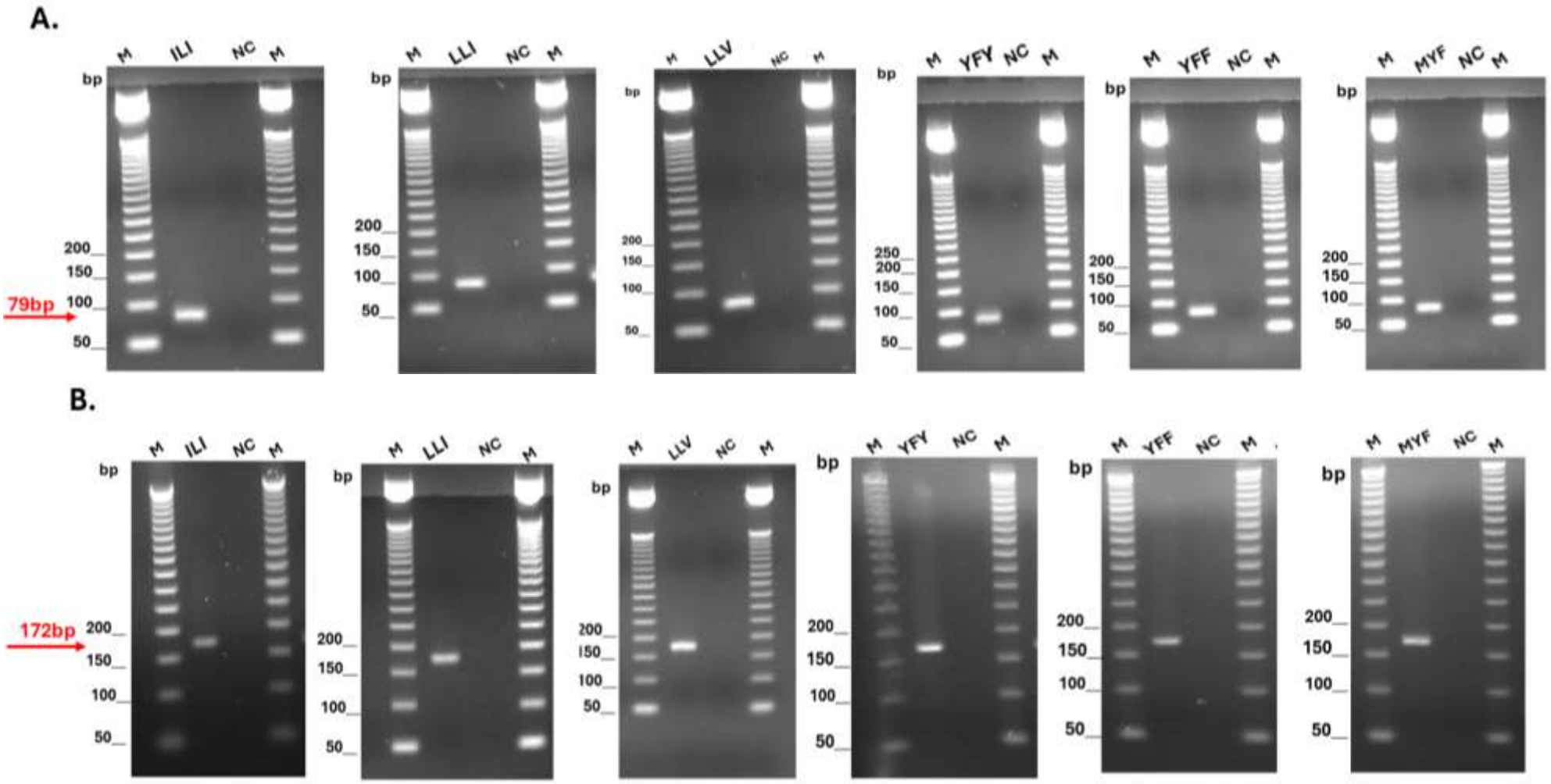


Figure 7.1. Agarose gel electrophoresis of novel SpyCatcher fragments. *Fragments were amplified from plasmid SpyCatcher-mNeongreen with primers listed in Table 9.3.Resulting PCR reactions were electrophoresed in a 3% agarose gel and stained with ethidium bromide, where each sample represents 20% of a 50µl PCR reaction.* ***(A)*** *Fragment 1****. (B)*** *Fragment 2. Lanes: ILI, LLI, LLV, YFY, YFF, MYF, NC; negative (no template) control, M, 50bp ladder MW marker.*

To generate the desired overlap PCR products, fragments 1 and 2 appropriate to each gene were each adjusted to a concentration of 15 µM. These fragments were then mixed together in an equimolar ratio in a PCR reaction mixture containing high-fidelity Pfu DNA polymerase and subjected to overlap PCR, without any additional primers (2.2.2.1.2). The products were then subjected to gel electrophoresis (3% agarose gel), to confirm that the overlap had occurred (Figure 7.2A). Resulting fragments were then diluted 100-fold and full-length products amplified by PCR using terminal primers 27 F and 90 R (2.2.2.1, Annex 1) and the resulting products examined by agarose gel electrophoresis (Figure 7.2.B).

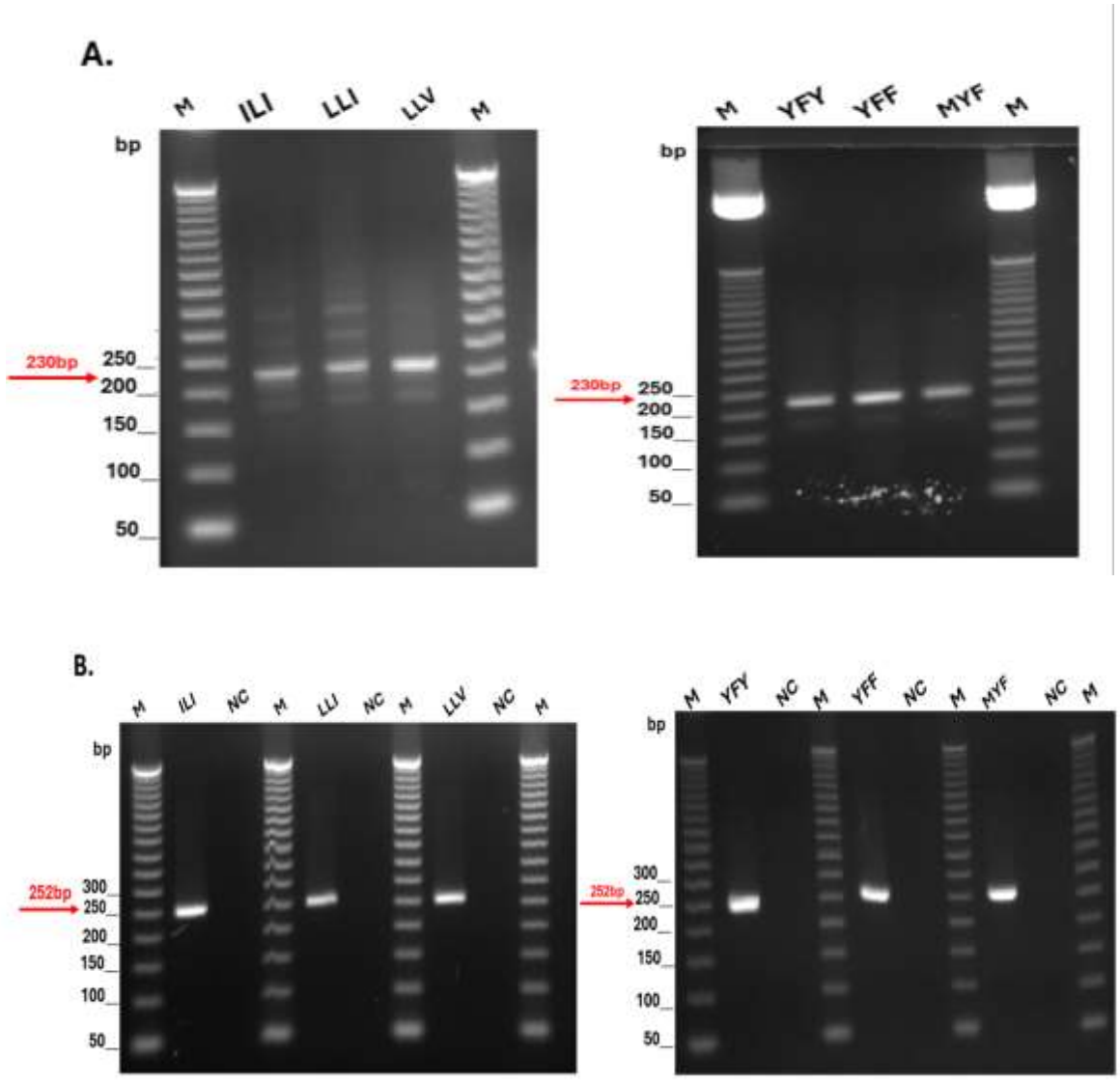


Figure 7.2. PCR Amplification for generating overlap and full-length products. (A) *Overlap PCR products: The results of overlap PCR for new SpyCatcher proteins are shown on a 3% agarose gel stained with ethidium bromide.* ***(B)*** *The expected full-length product, including overhangs generated by BsaI digestion. Lanes: ILI, LLI, LLV, YFY, YFF, MYF, NC, negative (no template) control, M, 50bp ladder MW marker.*

7.1.2. Gene cloning

Having confirmed the correct size of the overlapped products (Figure 7.2B), each fragment was purified using the Wizard® PCR Clean-Up System (Promega) PCR purification kit to ensure their suitability for Golden Gate cloning. The full length fragments were cloned into the mNeonGreen plasmid using Golden Gate cloning (2.2.2.2) to allow for the directional assembly of DNA fragments into the plasmid vector in a 3:1 molar ratio of inserts to backbone. The ligation products generated from the molecular cloning process were introduced into chemically competent *Escherichia coli* DH5α cells via transformation (2.2.5). Post-transformation, a 50 µL aliquot of the transformed cells was plated onto Luria-Broth (LB) agar, following inoculation of a single colony into a 5 mL starter culture containing 50 µg/mL kanamycin to selectively propagate the transformed cells. After sufficient growth in the selective medium, the bacterial cultures were harvested for plasmid DNA extraction using a mini-prep protocol (2.2.2.5.2).

7.1.3. Sanger sequencing of the newly constructed SpyCatcher genes

Plasmids were further purified using a Zymoresearch kit, and their concentrations were normalised to 50 ng/µl. The plasmids were then sent for Sanger sequencing (2.2.6.1) at Genewiz. On receipt of the chromatograms, each was analysed using the BioEdit sequence alignment tool, and the quality of the chromatograms was verified by checking for sharp, well-defined peaks and a clear baseline. The sequencing results confirmed the presence of the inserts in the plasmids, validating the cloning process (Figure 7.3).

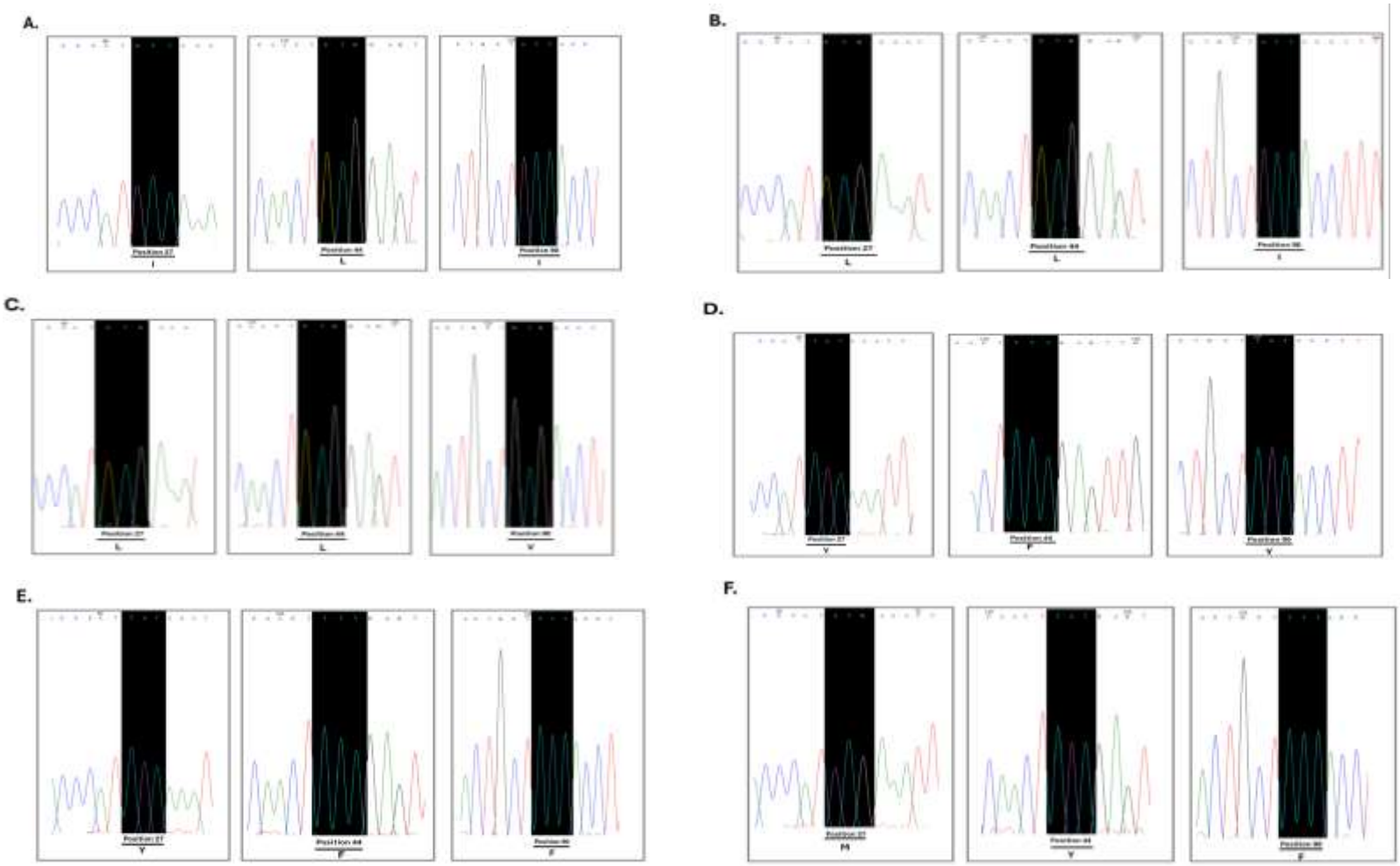


Figure 7.3. Sanger sequencing of the newly synthesised SpyCatcher genes. *Sanger sequencing data of newly synthesised SpyCatcher genes were extracted from the chromatogram at each of the three targeted positions and combined for analysis. The mutated positions are highlighted in black. At position 27, the sequencing data indicated that isoleucine was replaced with the expected codon. At positions 44, methionine was replaced with the expected codon and at position 90, isoleucine was replaced, aligning with theoretical expectation to encode each selected amino acid* ***(A)*** *ILI****, (B)*** *LLI,* ***(C)*** *LLV****, (D)*** *YFY,* ***(E)*** *YFF and* ***(F)*** *MYF.*

7.1.4. Expression and purification of novel SpyCatcher proteins

The verified expression vectors were transformed into chemically competent *E. coli* Tuner (DE3) cells (2.2.5) to enable expression of the new SpyCatcher variants.

Each of the six SpyCatcher genes was expressed in 200 ml cultures (2.2.7) and purified using affinity chromatography (2.2.9) usings nickel-NTA resins, taking advantage of the His tag included in the construct (2.2.9.1). The various stages of protein purification were examined by SDS-PAGE. An exemplar purification is shown in Figure 7.4.

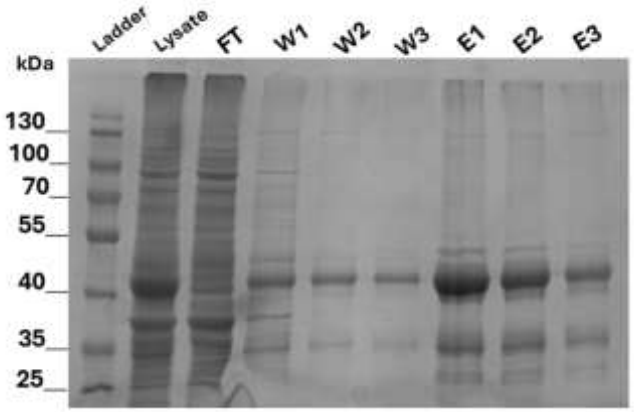


Figure 7.4. Purification of mutated SpyCatcher by affinity chromatography using Ni-NTA resin. *Polyacrylamide gel electrophoresis using a 12% gel shows fractions taken at stages of the purification process. The purification was via affinity column chromatography. The expected size of the protein is 43 kDa. FT= flow through, W= wash and E= elution. The gels were stained with InstantBlue.*

7.2. Examination of novel SpyCatcher interactions with SpyTag libraries via SDS-PAGE analysis

A total of six newly designed SpyCatcher proteins were synthesised, and each of these was incubated with all the twelve SpyTag libraries, leading to a total of 72 unique sample combinations (6 SpyCatchers × 12 SpyTags). The primary objective was to first analyse these 72 samples using SDS-PAGE to evaluate whether successful SpyTag-SpyCatcher binding occurred in each combination, which would be indicated by the formation of bands corresponding to the expected SpyTag-SpyCatcher complex.

The presence of a strong, clear band at the appropriate molecular weight would indicate successful binding between the SpyTag libraries and SpyCatcher proteins. Since the SpyTag libraries were fused with a fluorescent mCherry tag, binding interactions could be observed on the gel without the need for additional staining, allowing for a quick and efficient assessment of binding. Given the large number of samples (72 combinations), the SDS-PAGE results served as a screening tool to prioritise samples for more detailed mass photometry analysis.

Only those samples that showed clear evidence of binding (e.g. bands at the expected molecular weight) were selected for subsequent mass photometry.

Each newly synthesised SpyCatcher protein was analysed by SDS-PAGE using two sets of gels; one with the SpyTag libraries that had a fixed residue at position 3 (NX5 libraries), and the other with libraries that had a fixed residue at position 5 (X3N libraries). The results are illustrated in Figure 7.5 (predicted binders) and Figure 7.6 (predicted non-binders).

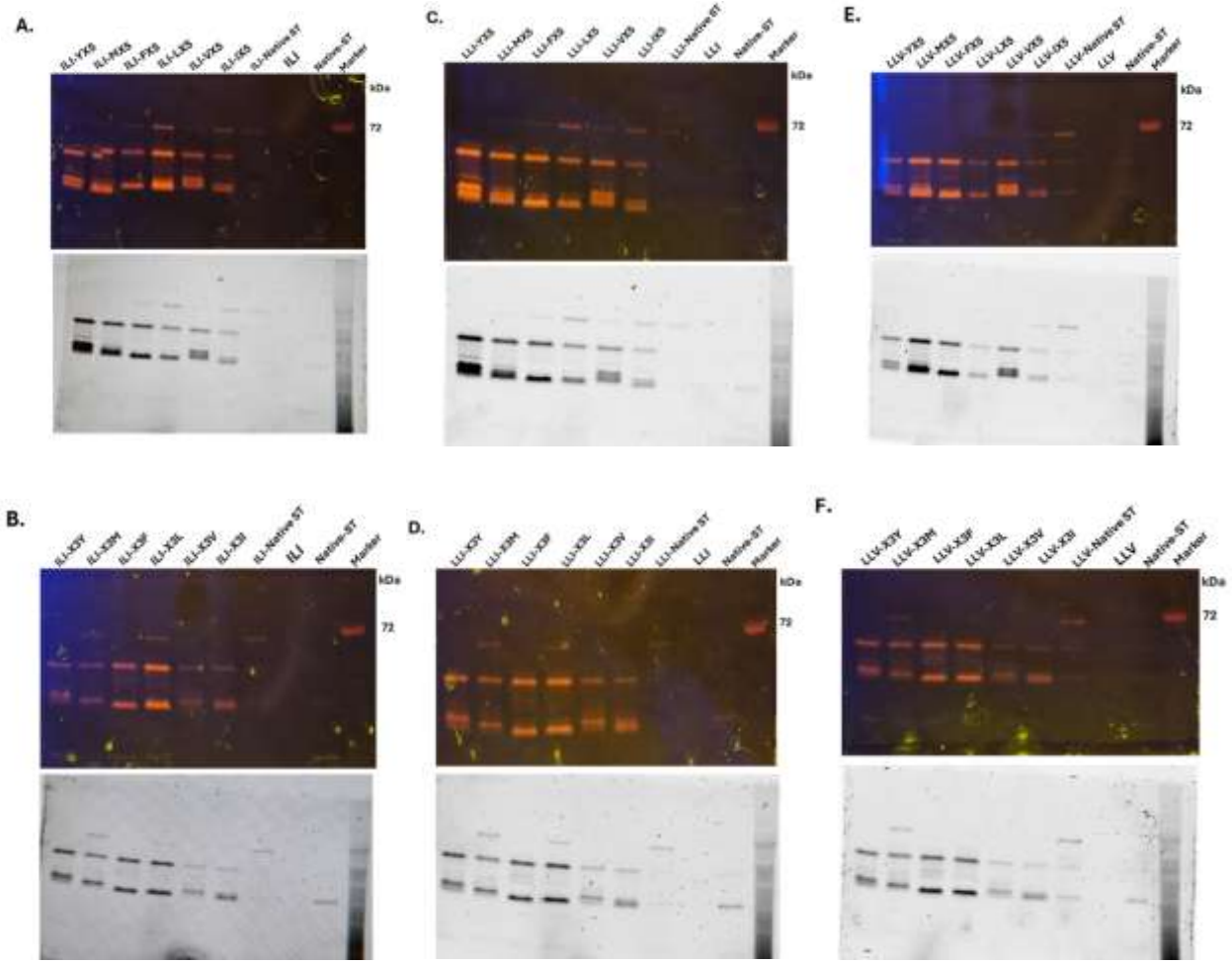


Figure 7.5. Covalent binding reconstitution between newly synthesised SpyCatcher proteins and SpyTag libraries*. Polyacrylamide gel electrophoresis (SDS-PAGE, 4-12% gradient) of* ***(A)*** *fixed positions libraries (NX5),* ***(B)*** *fixed positions libraries (X3N). Covalent complexes were assessed following 3 hours incubation. The native SC-ST was run alongside for comparison. Gels were visualised without staining under UV transillumination.*

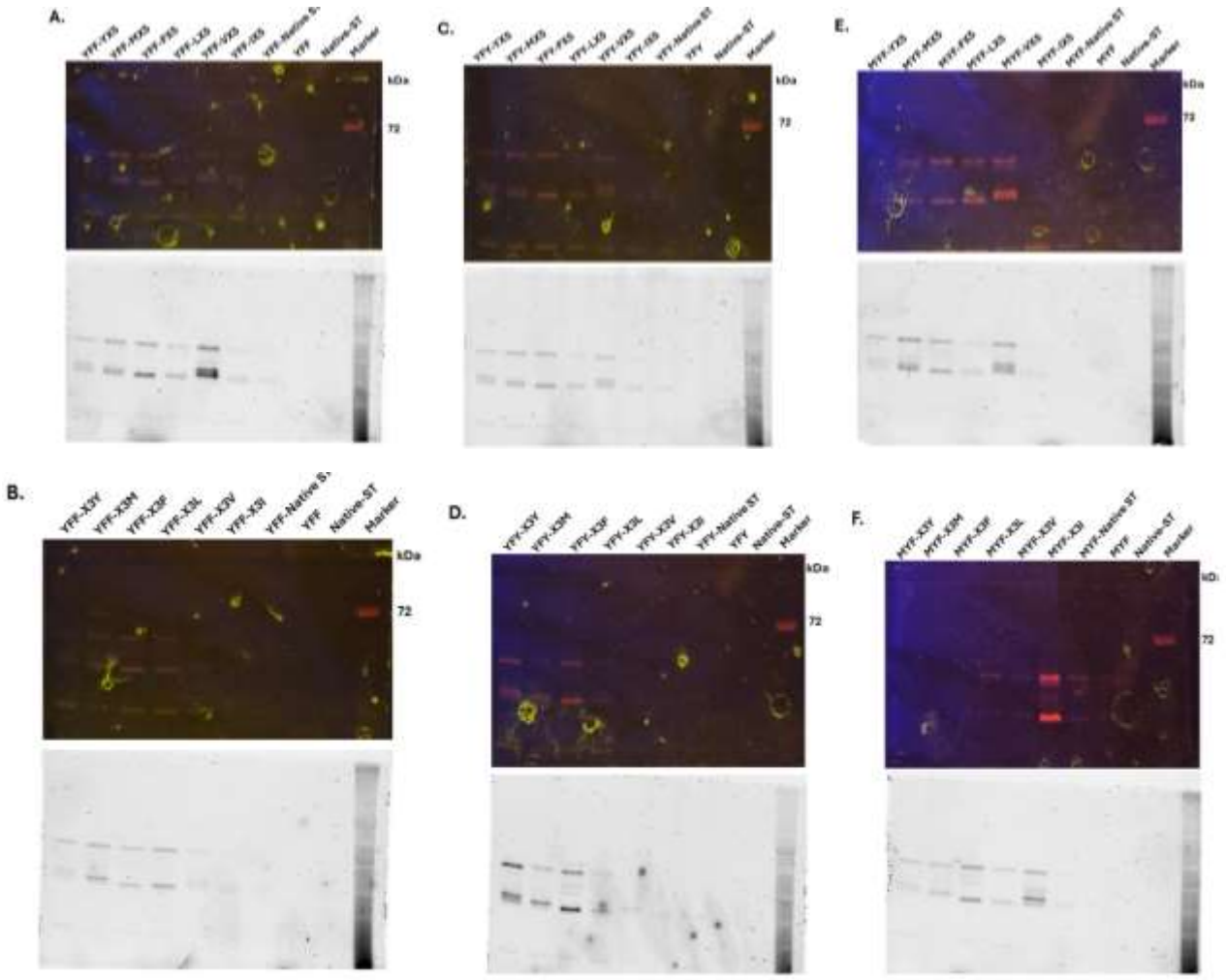


Figure 7.6. Covalent binding reconstitution between newly synthesised SpyCatcher proteins (predicted non-binders) and SpyTag libraries. *Polyacrylamide gel electrophoresis (SDS-PAGE, 4-12% gradient) of* ***(A)*** *fixed positions libraries (NX5),* ***(B)*** *fixed positions libraries (X3N). Covalent complexes were assessed following 3 hours incubation. The native SC-ST was run alongside for comparison. Gels were visualised without staining under UV transillumination.*

The SDS-PAGE analysis of SpyCatcher ILI (in which leucine is substituted for native methionine at position 44, and positions 27 and 90 are conserved) with SpyTag libraries NX5 showed complexation with nearly all NX5 libraries and good complexation with IX5 and LX5 libraries, as evidenced by intense bands at the expected molecular weight (~72 kDa), while libraries VX5, FX5 and MX5 showed faint 72 kDa bands on the gel (Figure 7.5A). There was no binding for the YX5 library. Of these, only the IX5 library contains native SpyTag, meaning that any of the large, aliphatic hydrophobic amino acids can be tolerated at position 3 of SpyTag when forming a covalent bond with variant SpyCatcher ILI. These results are similar to those obtained for native SpyCatcher with the NX5 libraries (Section 6.1.3). Conversely, among the SpyTag fixed position 5 libraries (X3N), only X3L and X3M exhibited complexation with SpyCatcher ILI (Figure 7.5B). Of these, only library X3M contains native SpyTag. No complex formation was observed for the rest of the X3N libraries with ILI protein, meaning that ILI variant protein, similar to native SpyCatcher (see section 6.1.4) is more selective at position 5, and of the residues tested by SDS PAGE, requires either leucine or methionine at position 5 in order to form an isopeptide bond with SpyTag.

The SDS-PAGE analysis of SpyCatcher LLI (with leucine substituted for native isoleucine and methionine both at positions 27 and 44 respectively) behaved similarly to SpyCatcher ILI and native SpyCatcher with respect to position 3 of SpyTag. There was isopeptide bond formation with components of all NX5 libraries, except for library YX5. In particular, LX5 library showed good complexation as demonstrated by a bright band at the expected molecular weight when compared with the rest of the libraries (Figure 7.5C), suggesting a possible, if minor alteration of specificity of SpyCatcher LLI for leucine over the native isoleucine at position 3 of SpyTag. Among the SpyTag fixed position 5 libraries (X3N), only X3L and X3M exhibited complex formation at the expected size (~72 kDa Figure 7.5D). Thus there is no change in preference for position 3 of SpyTag between native SpyCatcher and variant SpyCatchers ILI and LLI. Of the residues tested, all three proteins require either native isoleucine or leucine at position 3 of SpyTag.

Interestingly, PAGE analysis of variant SpyCatcher LLV (with all three residues substituted) suggested a possible minor enhancement of specificity for native isoleucine at position 3 of SpyTag. Whereas the native SpyCatcher and variants ILI and LLI all showed broad specificity for large hydrophobic residues at position 3 of SpyTag, LLV showed a clear preference for isoleucine at this position 3; faint bands for VX5 and LX5 libraries, where position 3 was fixed as valine and leucine respectively are also visible in Figure 7.5E, while there are no complexes with the other NX5 libraries (Figure 7.5E), suggesting that these mutations may confer enhanced specificity for the native target residue, isoleucine. Again, enhanced specificity is seen between SpyCatcher LLV and the X3N set of libraries, where a band of the expected

molecular weight (~72 kDa) is observed only in the lanes which contain native SpyTag – i.e. SpyTag itself and the IX5 library which contains native SpyTag (Figure 7.5F). Clearly when other hydrophobic variants are in positions 3 and 5 of SpyTag, SpyCatcher fails to form isopeptide bonds in detectable quantities, suggesting that the triple mutant $L_{27}L_{44}V_{90}$ has enhanced specificity for native SpyTag when compared with native SpyCatcher ($I_{27}M_{44}I_{90}$).

The analysis of predicted non-binding variants of SpyCatcher (Table 7.1) failed to show any substantial levels of isopeptide bond formation as assessed by SDS-PAGE. Specifically no clear complexion was observed with the YFY SpyCatcher protein with any of the SpyTag libraries or with native SpyCatcher (Figure 7.6A & B), while a very faint band was seen with SpyCatcher YFF with native SpyTag and SpyTag libraries IX5 and X3M, which each contain native SpyTag (Figure 7.6C & D). Analysis of SpyCatcher MYF revealed no binding with any SpyTag library except for a very faint band with the LX5 library (Figure 7.6E), suggesting minimal complexation when leucine is present at position 3 of SpyTag. However, there was no corresponding complexation seen in any of the XN5 libraries and no complexation visible between SpyCatcher MYF and native SpyTag (Figure 7.6F).

7.3. Examination of novel SpyCatcher interactions with SpyTag libraries via mass photometry

Following the SDS-PAGE results, samples were chosen for mass photometry analysis to determine whether the more interesting results from SDS-PAGE analysis would be reproduced by an alternate means of analysis.

Based on the results of the SDS-PAGE, only two SpyTag libraries from the X3N group were selected for further analysis: X3L and X3M (where position 5 of SpyTag is fixed as leucine and methionine respectively), as these libraries showed some level of interaction with SpyTag, while the rest of the X3N libraries did not produce any detectable bands on the gels, indicating a lack of significant binding activity.

In contrast, four libraries from the NX5 group were selected for further investigation: IX5, VX5, LX5 and FX5. These libraries were prioritised because, like those selected from X3N libraries, they exhibited some interaction on the gels, making them suitable candidates for additional testing. The remaining NX5 libraries were not considered for further analysis as they did not show any detectable bands, suggesting they did not bind with the novel-binders SpyCatcher proteins under the experimental conditions.

Overall, these selected libraries (IX5, VX5, LX5 and FX5, X3L and X3M) then were subjected to examination using mass photometry, allowing for a closer evaluation of their interactions with each of the SpyCatcher proteins (2.2.10). The results are illustrated in Figure 7.7, Figure 7.8 and 7.9 (predicted binders) and Figure 7.10 (predicted non-binders).

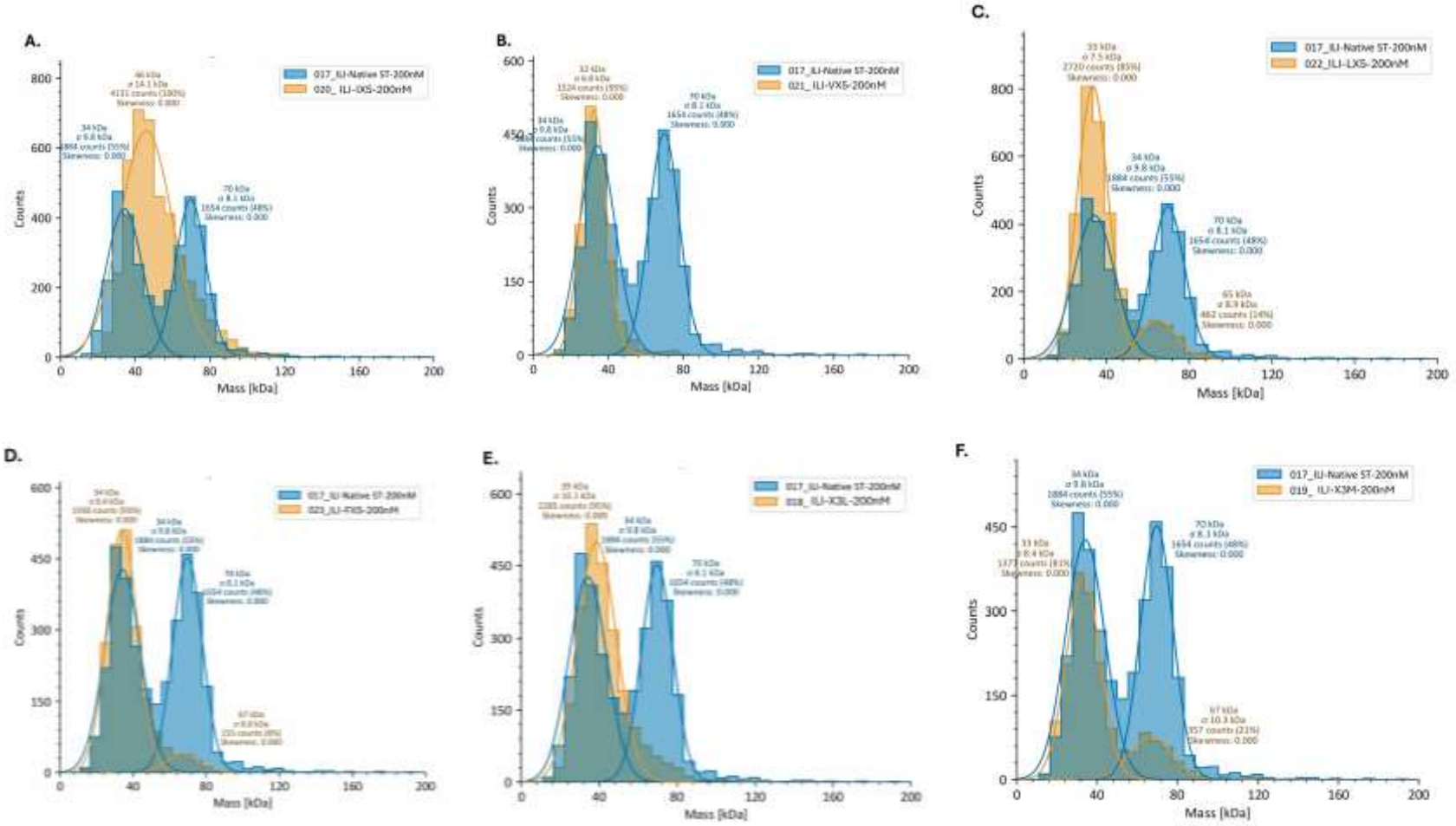


Figure 7.7. Mass photometry analysis of SpyCatcher ILI with selected-fixed position SpyTag libraries. *The spectra show the molecular mass distribution for the resulting complexes after the interaction between ILI SpyCatcher and native-SpyTag (ILI-Native ST, blue), and ILI with selected SpyTag-libraries (orange). Each dataset represents normalised counts plotted against molecular mass (kDa). Panels depict data for:* ***(A)*** *IX5,* ***(B)*** *VX5****, (C)*** *LX5,* ***(D)*** *FX5****, (E)*** *X3L and* ***(F)*** *X3M libraries.*

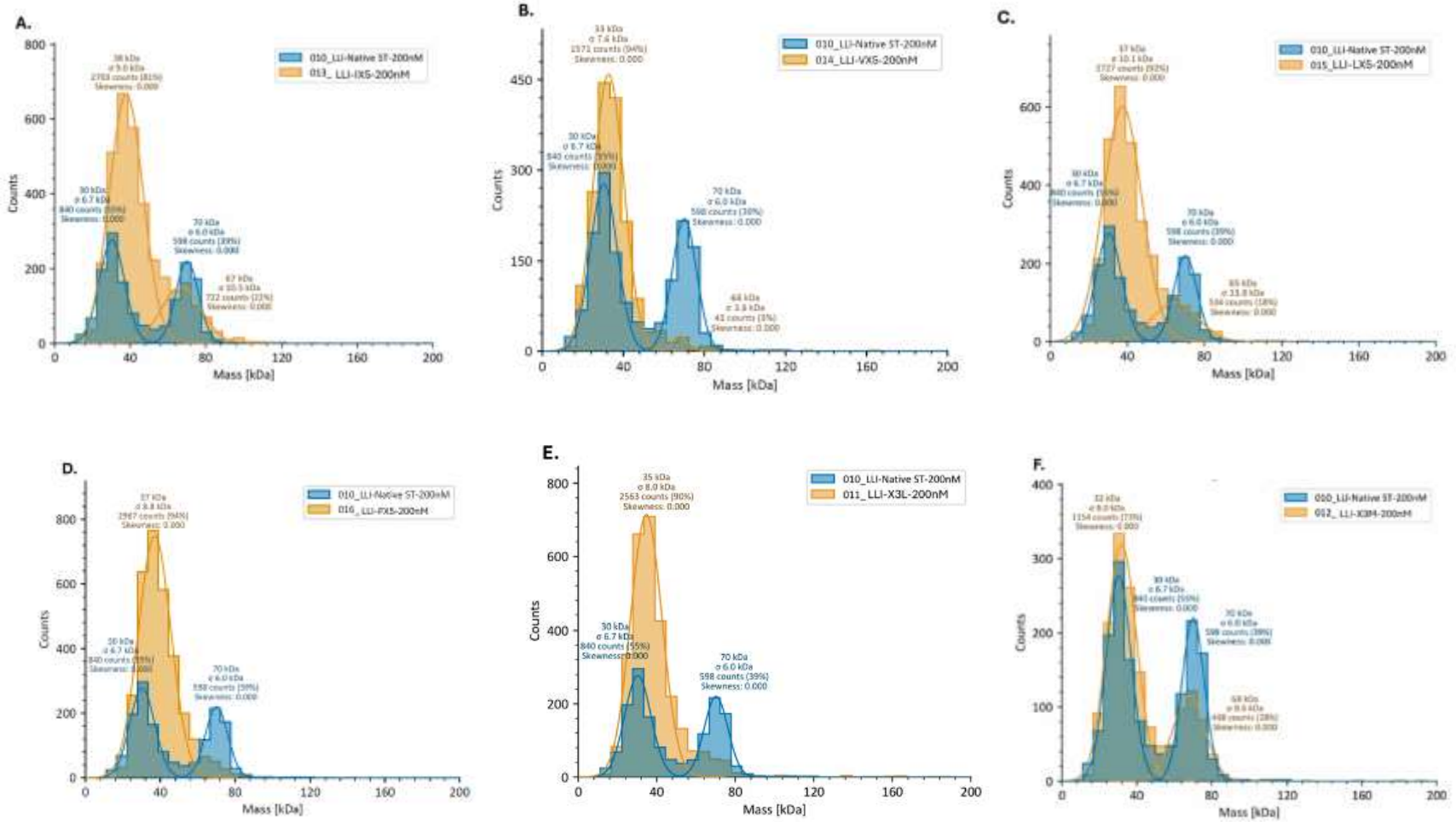


Figure 7.8. Mass photometry analysis of SpyCatcher LLI with selected-fixed position SpyTag libraries. *The spectra show the molecular mass distribution for the resulting complexes after the interaction between LLI SpyCatcher and native-SpyTag (LLI-Native ST, blue), and LLI with selected SpyTag-libraries (orange). Each dataset represents normalised counts plotted against molecular mass (kDa). Panels depict data for:* ***(A)*** *IX5,* ***(B)*** *VX5****, (C)*** *LX5****, (D)*** *FX5,* ***(E)*** *X3L and* ***(F)*** *X3M libraries****.***

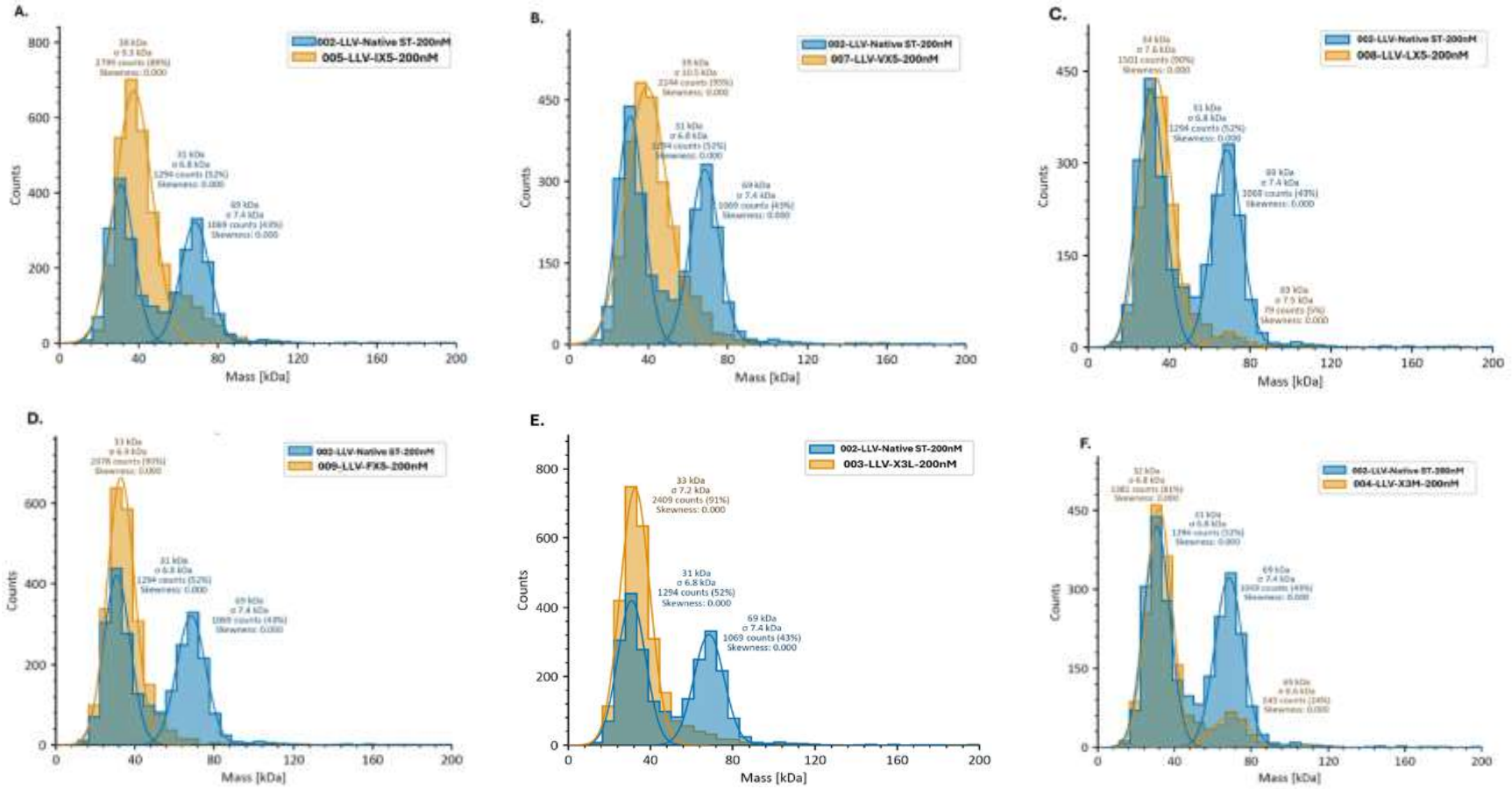


Figure 7.9. Mass photometry analysis of SpyCatcher LLV with selected-fixed position SpyTag libraries. *The spectra show the molecular mass distribution for the resulting complexes after the interaction between LLV SpyCatcher and native-SpyTag (LLV-Native ST, blue), and LLV with selected SpyTag-libraries (orange). Each dataset represents normalised counts plotted against molecular mass (kDa). Panels depict data for:* ***(A)*** *IX5,* ***(B)*** *VX5,* ***(C)*** *LX5,* ***(D)*** *FX5,* ***(E)*** *X3L and* ***(F)*** *X3M libraries.*

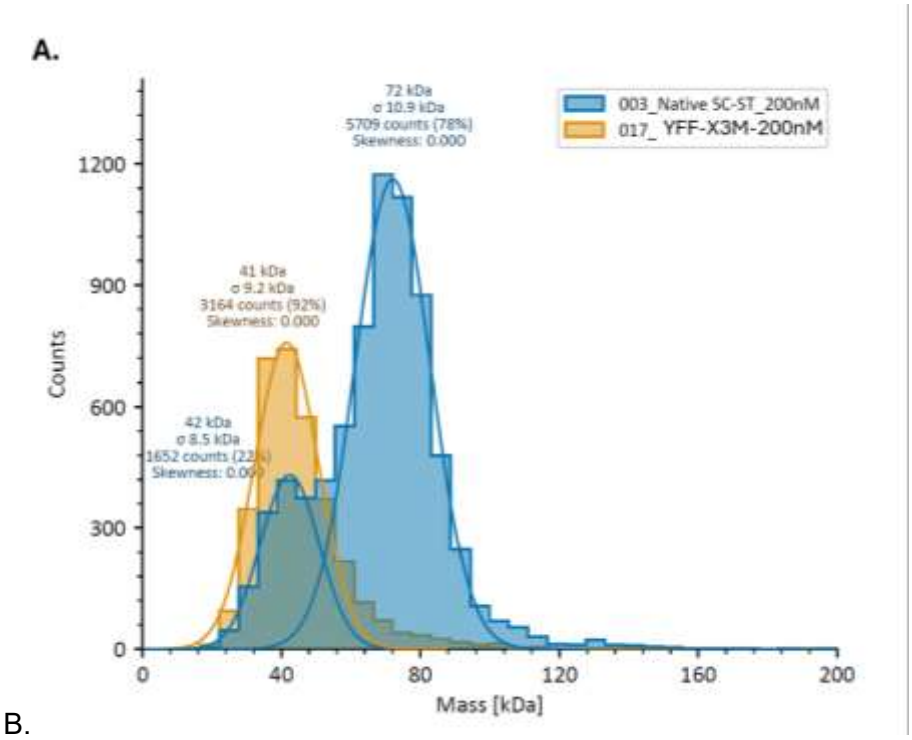


Figure 7.10. Mass photometry analysis of non-binders SpyCatcher with selected-fixed position SpyTag libraries. *The spectra show the molecular mass distribution for resulting complex after the interaction between native SpyCatcher and native SpyTag (Native SC-ST, blue), as well as non-binder SpyCatcher with selected SpyTag libraries (orange), Each dataset represents normalised counts plotted against molecular mass (kDa). Panels depict data for: (A) YFF-X3M, (B) MYF-LX5.*

Mass photometry analysis of SpyCatcher ILI demonstrated isopeptide bond formation with native SpyTag as evidenced by the peak in the 70 kDa region of the blue histogram traces in Figures 7.7A-F. Unexpectedly though, the clearest interaction at position 3 of SpyTag (as determined by the mass photometry software) is seen between the library in which position 3

of SpyTag is fixed as leucine (Figure 7.7C). However, closer inspection reveals a large shoulder on the trace for library IX5 which contains the native SpyTag (Figure 7.7A). In agreement with the SDS-PAGE results, phenylalanine is also accepted at position 3 of SpyTag to a lesser extent by SpyCatcher ILI (Figure 7.7D). While a bright band observed for the LX5 library on SDS-PAGE is also consistent with mass photometry results as described above, the weaker interaction with valine observed by SDS-PAGE analysis (library VX5, Figure 7.5A) is not evident in the mass photometry results (Figure 7.7B). In library FX5, the faint band observed by SDS-PAGE analysis (Figure 7.5 7.5A) has similarly been detected by mass photometry, albeit as a small shoulder (Figure 7.7D). On the other hand, with respect to position 5 of SpyTag (X3N libraries), as expected, the best interaction is seen between the library in which position 5 of SpyTag is fixed as the native methionine (Figure 7.7F), even though the band observed on the gel was weak. Meanwhile, in this set of results, manual inspection of the spectra is required in order to see the shoulder that represents the interaction of SpyCatcher ILI with library X3L (Figure 7.7E), meaning that SDS-PAGE analysis might be more sensitive than mass photometry for analysing SpyTag-SpyCatcher interactions.

Mass photometry analysis of SpyCatcher LLI again demonstrated isopeptide bond formation with native SpyTag as evidenced by the peak in the 70 kDa region of the blue histogram traces in Figures 7.8A-F. Among the SpyTag variants tested via mass photometry there was consistency with SDS-PAGE results. Specifically, the best interaction is seen between the library in which position 3 of SpyTag is fixed as the native isoleucine (Figure 7.8A), but leucine is also accepted at this position of SpyTag (Figure 7.8C). The weaker interactions with valine and phenylalanine observed by SDS-PAGE analysis (Figure 7.5C) were also detected by mass photometry as very small shoulders in the orange histogram trace around the 70-80 kDa range (Figure 7.8B&D respectively). Mass photometry analysis of LLI SpyCatcher with two libraries in which position 5 of SpyTag was fixed, as expected, showed the best interaction in X3M library in which position 5 is fixed as native methionine (Figure 7.8F) while X3L library which position 5 is fixed as leucine by faint band on SDS-PAGE (Figure 7.5D), also showed a small degree of complexation by mass photometry, again showing consistency between SDS-PAGE and the mass photometry analyses (Figure 7.8E).

Mass photometry analysis of SpyCatcher LLV again demonstrated isopeptide bond formation with native SpyTag as evidenced by the peak in the 70 kDa region of the blue histogram traces in Figures 7.9A-F. However, here the results between SDS-PAGE and mass photometry analyses are a little less consistent. Both analytical methods agree that isoleucine is preferred at position 3 of SpyTag (compare Figure 7.5D with the large shoulder in the histogram of Figure 7.9A) and both agree that methionine is preferred at position 5 of SpyCatcher (compare Figures 7.5F with 7.9F). These figures also both support that SpyCatcher LLV can bind either leucine or valine at position 3 of SpyTag. However, mass photometry suggests a very weak

interaction with libraries FX5 and X3L (as indicated by small shoulders - Figures 7.9D&E), neither of which are detectable by SDS-PAGE (Figure 7.5E&F).

SDS-PAGE analysis of SpyCatcher protein YFF showed weak interactions with libraries IX5 and X3M (Figures 7.6C&D) and mass photometry of SpyCatcher YFF with library X3M, if only by the presence of a shoulder in the 70 kDa region of the yellow histogram trace (Figure 7.10A), supported this finding. Conversely, while SpyCatcher MYF showed very weak interaction with library LX5 by SDS-PAGE (Figure 7.6E), there is no clear evidence of this interaction by mass photometry (Figure 7.10B).

7.4. Discussion

The findings from this chapter extend the exploration of SpyCatcher-SpyTag protein engineering probing the specificity of both native SpyCatcher and newly designed SpyCatcher proteins. The outcomes highlight the intricate interplay between specific residue substitutions in SpyCatcher protein and their influence on binding affinity towards SpyTag.

The SDS-PAGE analysis of the new SpyCatcher proteins towards the SpyTag variants offered a practical and rapid assessment of SpyCatcher-SpyTag interactions. It revealed differences in binding behaviour across the designed proteins to check for any orthogonality. However, the mass photometry analysis, as well as largely confirming results from SDS-PAGE is more exciting. We believe that this is the first time that mass photometry has been used to examine protein libraries and certainly to aid in the deconvolution of such libraries.

Returning to SpyTag-SpyCatcher interactions, predicted SpyCatcher binders (ILI, LLI, LLV), exhibited distinct binding patterns. ILI and LLI demonstrated broader binding profiles with NX5 SpyTag libraries (where position 3 of SpyTag is fixed), reflecting tolerance for multiple aliphatic, hydrophobic residues. However, their interactions with X3N libraries (where position 5 of SpyTag fixed) were more specific, primarily favouring native variants or leucine substitutions at this position. LLV SpyCatcher protein, in contrast, showed a more restricted binding profile, with reduced cross-reactivity, indicating moderately increased specificity. Analysing these proteins revealed that by replacing the native residues in SpyCatcher (I27M44I90) from one (ILI) to three (LLV) substitutions, the cross-reactivity decreased slightly, in addition LLV in which all three positions were substituted from the native SpyCatcher showed binding towards native-SpyTag. On the other hand, predicted non-binders (YFY, YFF, MYF) as expected, exhibited little to no binding toward native SpyTag but unfortunately, did not bind SpyTag variants, either meaning that no orthogonal pairs were identified. However, the results did support the hypothesis that certain substitutions, such the bulky aromatic residue phenylalanine diminish or prevent the interaction and polar aromatic residues (tyrosine), abolish binding interactions.

Overall, multiple positions within the SpyCatcher protein have been investigated to evaluate their specificity with the SpyTag variants and the results indicate that substitutions involving larger, hydrophobic, aliphatic residues at these critical positions are generally well-tolerated, maintaining or enhancing binding affinity. In contrast, the introduction of a hydroxyl group at these sites completely disrupted the interaction with the SpyTag, highlighting the sensitivity of the binding mechanism to changes in polarity and hydrophobicity. Despite testing of individual SpyCatcher variants, an orthogonal binding pair was not identified. Further investigations are required to explore alternative residue targets.

Finally, the promise that mass photometry can be used to deconvolute protein libraries may be particularly relevant when examining non-covalent protein-ligand interactions not amenable to analysis via SDS-PAGE.

Chapter 8 Conclusion

This study aimed to investigate whether the specificity of the SpyTag-SpyCatcher interaction could be modulated by introducing targeted substitutions within the hydrophobic binding pocket of SpyCatcher and the corresponding positions of SpyTag. Specifically, alternative hydrophobic residues were explored to assess their impact on the interaction dynamics. The ultimate goal was to determine whether such mutations could lead to the development of orthogonal SpyTag-SpyCatcher pairs.

Molecular modelling was employed to verify the proposed substitutions in both SpyCatcher and SpyTag. Following verification, positionally-fixed libraries for both SpyCatcher and SpyTag were designed, to make variations within both the protein and the peptide. Two complementary approaches, overlap PCR and MAX randomisation, were utilised to construct these libraries (SpyCatcher libraries and SpyTag libraries, respectively). The use of overlap PCR as a method for library construction was validated by the consistency in codon randomisation and amino acid distribution, as confirmed by both Sanger sequencing and Next-Generation Sequencing (NGS). The dual application of these sequencing methods provided complementary insights; Sanger sequencing accurately confirmed fixed positions, while NGS revealed the subtleties of library diversity.

A novel screening strategy was developed to evaluate the specificity of these fixed position libraries, with specific reference to the SpyTag-SpyCatcher interaction. SDS-PAGE was employed to provide semi-quantitative assessments of protein-peptide binding, offering a straightforward approach for initial analysis. In parallel, mass photometry was used as an innovative tool for evaluating binding interactions. Mass photometry's ability to detect interactions within complex mixtures proved invaluable, demonstrating its potential to identify interactions within a mixture. This study represents the first application of mass photometry for screening protein libraries, establishing it as a powerful method for evaluating protein-ligand specificity, providing that the ligand is of a suitable molecular mass.

Eighteen SpyCatcher libraries containing new SpyCatcher variants were expressed and their specificity was assessed for orthogonality. The overall comparative analysis of these newly synthesised SpyCatcher variants against twelve SpyTag libraries yielded almost consistent results across both methodologies as shown in Table 8.1.

SpyCatcher			Native	SpyTag library													
27	44	90	SpyTag	IX5	LX5	VX5	MX5	FX5	YX5	X3I	X3L	X3V	X3M	X3F	X3Y	Analysis method	Source
I	M	I							-	-		-		-	-	SDS-PAGE	Figure 6.2
						-	shoulder	shoulder	-	shoulder	shoulder	shoulder		-	-	Mass photometry	Figures 6.4 & 6.5
I	L	I							-	-				-	-	SDS-PAGE	Figure 7.5A&B
						-	nd		nd	nd	shoulder	nd		nd	nd	Mass photometry	Figure 7.7
L	L	I							-	-		-		-	-	SDS-PAGE	Figure 7.5C&D
							nd		nd	nd	shoulder	nd		nd	nd	Mass photometry	Figure 7.8
L	L	V					-	-	-	-	-	-		-	-	SDS-PAGE	Figure 7.5D&E
							nd	shoulder	nd	nd	shoulder	nd		nd	nd	Mass photometry	Figure 7.9
Y	F	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	SDS-PAGE	Figure 7.6 A&B
			nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	Mass photometry	
Y	F	F	-		-	-	-	-	-	-	-	-		-	-	SDS-PAGE	Figure 7.6C&D
			nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	shoulder	nd	nd	Mass photometry	Figure 7.10A
M	Y	F	-	-		-	-	-	-	-	-	-	-	-	-	SDS-PAGE	Figure 7.6E&F
			nd	nd	shoulder	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	Mass photometry	Figure 7.10B

Key

- moderate level of 72 kDa product
- wild type level of 72 kDa product
- low level of 72 kDa product
- faint level of 72 kDa product
- no detectable 72 kDa product
- nd not determined

Table 8.1. Comparative analysis of SpyCatcher variants interacting with SpyTag libraries, evaluated using SDS-PAGE and mass photometry. *Binding strength was assessed based on the detection of a 72 kDa product, categorized as wild-type, moderate, low, faint, or undetectable. Shoulder peaks indicate partial binding observed in specific libraries. Variants IMI (native SpyCatcher), ILI, LLI, and LLV displayed varying degrees of interaction, while YFF, YFY, and MYF showed no detectable binding.*

Table 8.1 provides comparison of the interactions between SpyCatcher variants and SpyTag libraries, highlighting differences in binding specificity due to mutations. The native SpyCatcher ($I_{27}M_{44}I_{90}$) demonstrates near consistency between the two analytical methods, achieving moderate levels of the 72 kDa product with IX5 and X3M SpyTags (native containing libraries), and wild-type levels of interaction with SpyTag LX5. Variants ILI and LLI exhibit better consistency between SDS-PAGE and MP, while LLV displays some minor discrepancies between the SDS-PAGE and molecular photometry. In contrast, SpyCatchers YFF, YFY, and MYF showed good consistency between the two analytical methods, though the levels of interaction with any SpyTag are minimal. These findings emphasise the structural importance of specific residues in maintaining interactions between SpyCatcher and SpyTag pairs.

The consistent findings between mass photometry and SDS-PAGE in this study confirm that both methods are reliable for evaluating specificity of interaction. Both techniques reliably identify binding and provide corroborative data on the relative interaction levels of different variants.

Among the SpyCatcher variants tested, the LLV mutant, with substitutions at all three native residues in its hydrophobic binding pocket, exhibited enhanced specificity. This variant interacted robustly with the native SpyTag, demonstrating a selective binding profile. When tested against SpyTag libraries containing alternative hydrophobic residues, LLV showed reduced cross-reactivity, binding preferentially to two libraries featuring native-like SpyTag variants (IX5 and X3M), compared to other newly synthesised SpyCatcher variants. The interaction of LLV with one exemplar SpyTag variant is shown in Figure 8.1.

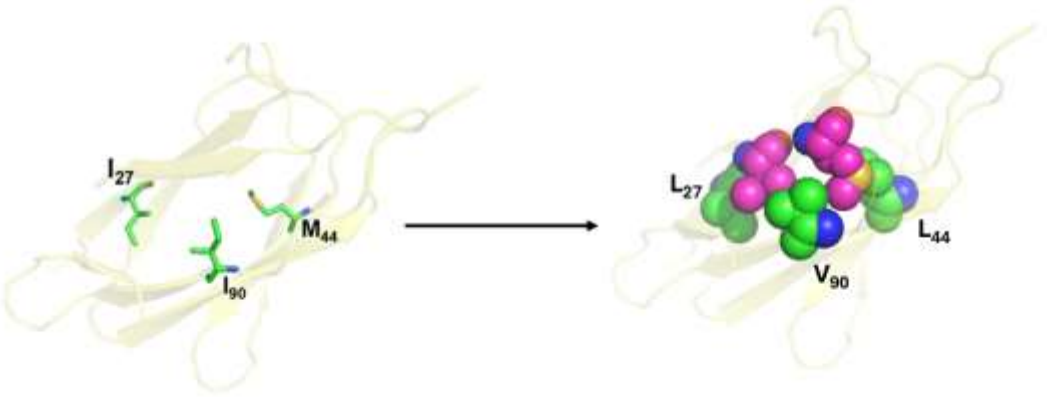


Figure 8.1. Schematic representation of targeted residues on native SpyCatcher ($I_{27}M_{44}I_{90}$) versus mutated SpyCatcher ($L_{27}L_{44}V_{90}$). *The selected residues are shown in stick format on SpyCatcher illustrating the replacement of all three native residues with LLV, the two residues on SpyTag illustrated by pink colour.*

Despite exploring numerous substitutions, no orthogonal SpyCatcher-SpyTag pair was identified, suggesting the need for either expanded substitutions or perhaps the possibility of targeting alternative residue(s) in the SpyTag-SpyCatcher pair. Comprising just 13 amino acids, SpyTag is short. The ideal residues within the hydrophobic cleft at positions 3 & 5 have already been examined within the current study. Meanwhile further downstream in the peptide, residue 7 forms the isopeptide bond with SpyCatcher. To determine the feasibility of targeting alternative residues, the C-terminal end of SpyTag was therefore visualised using PyMOL to see if there might be alternative targets for future mutagenesis. Two such putative targets were identified (Figure 8.2).

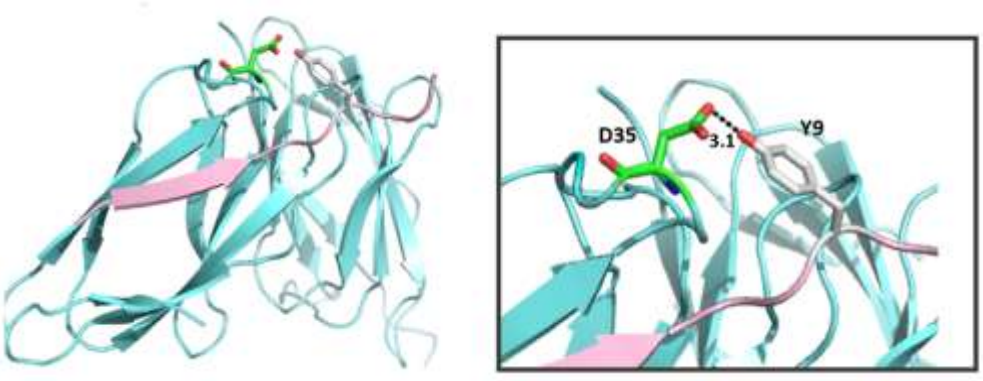


Figure 8.2. Interaction between tyrosine (Y9) of SpyTag and aspartic acid (D35) of SpyCatcher. Tyrosine (Y9 shown in grey and aspartic acid (D35) shown in green. The hydrogen bond distance is approximately 3.1 Å, indicating the bonding interaction between these two residues.

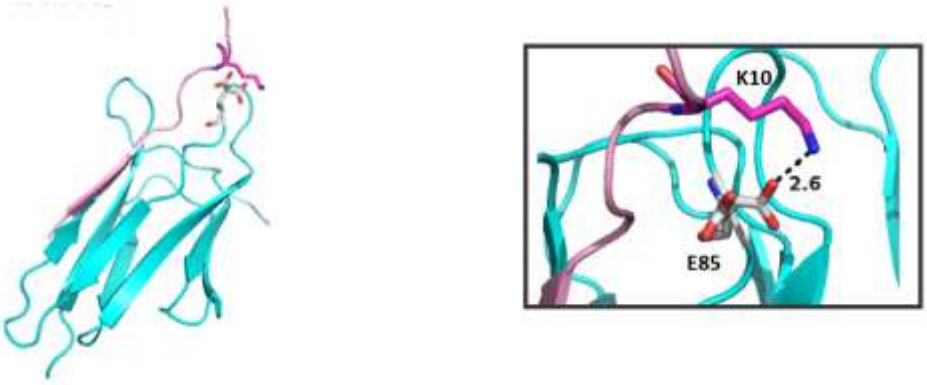


Figure 8.3. Interaction between lysine (K10) of SpyTag and glutamic acid (E85) of SpyCatcher. *Lysine (K10) shown in magenta and glutamic acid (E85) shown in grey. The hydrogen bond distance is approximately 2.6 Å, highlighting the close interaction between these two residues.*

As shown in Figure 8.2, the hydroxyl group of tyrosine 9 (Y_9) in SpyTag is reasonably well-placed to form a hydrogen bond with aspartic acid 35 (Asp_{35}) in SpyCatcher, with an inter-atomic distance of 3.1Å. Meanwhile, the lysine residue (K_{10}) in SpyTag is better-placed still to form an electrostatic interaction with glutamic acid 85 (E_{85}) in SpyCatcher with an inter-atomic distance of approximately 2.6 Å, as shown in Figure 8.3. Interestingly, both of these residues are described as part of the general set of interactions between SpyTag and SpyCatcher (Li et al., 2014).

Thus in summary, future directions for this work are twofold. Firstly, MAX randomisation could be used to systematically explore the amino acid sequence space at existing positions 27, 44 and 90 of SpyCatcher plus positions 3 and 5 of SpyTag by replacing each residue with all 20 amino acids instead of six hydrophobic residues examined in this study. This approach would ensure complete coverage of potential variants. Alternatively, positions 9 and 10 of SpyTag and corresponding positions 35 and 85 of SpyCatcher might be targeted for mutagenesis. In practice, a combination of both strategies may be required.

In conclusion this study represents a significant advancement by introducing mass photometry as an analytical tool for screening combinatorial libraries, marking its first known application in this context since no one has previously used mass photometry to examine molecular interactions within libraries. Previous methods for analysing protein libraries, such as phage display and fluorescence-based techniques, often require repetitive procedures and larger protein sample quantities, while lacking the simplicity offered by mass photometry. The application of mass photometry in this study not only validates the quality and specificity of the constructed libraries but also opens new avenues for its use in protein engineering. By enabling label-free analysis of protein interactions in their native state and in real time, mass photometry addresses key limitations of existing screening techniques.

This breakthrough positions mass photometry as a valuable tool for the field, offering unparalleled insight into the screening and characterisation of engineered protein systems. The findings demonstrate that mass photometry is not just a supplement to existing methods but has the potential to redefine how protein libraries are analysed.

References

Amelung, S., Nerlich, A., Rohde, M., Spellerberg, B., Cole, J. N., Nizet, V., Chhatwal, G. S., & Talay, S. R. (2011). The FbaB-type fibronectin-binding protein of Streptococcus pyogenes promotes specific invasion into endothelial cells. *Cellular Microbiology*, *13*(8), 1200–1211. https://doi.org/10.1111/J.1462-5822.2011.01610.X

Andersson, A. M. C., Buldun, C. M., Pattinson, D. J., Draper, S. J., & Howarth, M. (2019). SnoopLigase peptide-peptide conjugation enables modular vaccine assembly. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-40985-w

Ashraf, M., Frigotto, L., Smith, M. E., Patel, S., Hughes, M. D., Poole, A. J., Hebaishi, H. R. M., Ullman, C. G., & Hine, A. V. (2013). ProxiMAX randomization: A new technology for non-degenerate saturation mutagenesis of contiguous codons. *Biochemical Society Transactions*, *41*(5), 1189–1194. https://doi.org/10.1042/BST20130123

Asor, R., & Kukura, P. (2022). Characterising biomolecular interactions and dynamics with mass photometry. In *Current Opinion in Chemical Biology* (Vol. 68). Elsevier Ltd. https://doi.org/10.1016/j.cbpa.2022.102132

Baumgarten, T., Ytterberg, A. J., Zubarev, R. A., & de Gier, J. W. (2018). Optimizing recombinant protein production in the Escherichia coli periplasm alleviates stress. *Applied and Environmental Microbiology*, *84*(12). https://doi.org/10.1128/AEM.00270-18

Becker, J., Peters, J. S., Crooks, I., Helmi, S., Synakewicz, M., Schuler, B., & Kukura, P. (2023). *A quantitative description for optical mass measurement of single biomolecules*. https://doi.org/10.1101/2023.03.28.534430

Claasen, M., Kofinova, Z., Contino, M., & Struwe, W. B. (2024). Analysis of Protein Complex Formation at Micromolar Concentrations by Coupling Microfluidics with Mass Photometry. *Journal of Visualized Experiments*, *2024*(203). https://doi.org/10.3791/65772

Cooley, R. B., Feldman, J. L., Driggers, C. M., Bundy, T. A., Stokes, A. L., Karplus, P. A., & Mehl, R. A. (2014). Structural basis of improved second-generation 3-nitro-tyrosine tRNA synthetases. *Biochemistry*, *53*(12), 1916–1924. https://doi.org/10.1021/bi5001239

Dovala, D., Sawyer, W. S., Rath, C. M., & Metzger, L. E. (2016a). Rapid analysis of protein expression and solubility with the SpyTag-SpyCatcher system. *Protein Expression and Purification*, *117*, 44–51. https://doi.org/10.1016/j.pep.2015.09.021

Dovala, D., Sawyer, W. S., Rath, C. M., & Metzger, L. E. (2016b). Rapid analysis of protein expression and solubility with the SpyTag-SpyCatcher system. *Protein Expression and Purification*, *117*, 44–51. https://doi.org/10.1016/j.pep.2015.09.021

Duda, R. L. (1998). Protein Chainmail: Catenated Protein in Viral Capsids. *Cell*, *94*(1), 55–60. https://doi.org/10.1016/S0092-8674(00)81221-0

Ferreira Amaral, M. M., Frigotto, L., & Hine, A. V. (2017a). Beyond the Natural Proteome: Nondegenerate Saturation Mutagenesis—Methodologies and Advantages. In *Methods in Enzymology* (Vol. 585, pp. 111–133). Academic Press Inc. https://doi.org/10.1016/bs.mie.2016.10.005

Ferreira Amaral, M. M., Frigotto, L., & Hine, A. V. (2017b). Beyond the Natural Proteome: Nondegenerate Saturation Mutagenesis—Methodologies and Advantages. In *Methods in Enzymology* (Vol. 585, pp. 111–133). Academic Press Inc. https://doi.org/10.1016/bs.mie.2016.10.005

Fierer, J. O., Veggiani, G., & Howarth, M. (2014). SpyLigase peptide-peptide ligation polymerizes affibodies to enhance magnetic cancer cell capture. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(13). https://doi.org/10.1073/pnas.1315776111

Foley, E. D. B., Kushwah, M. S., Young, G., & Kukura, P. (2021). Mass photometry enables label-free tracking and mass measurement of single proteins on lipid bilayers. *Nature Methods*, *18*(10), 1247–1252. https://doi.org/10.1038/s41592-021-01261-w

Furka, Á. (2022). Forty years of combinatorial technology. *Drug Discovery Today*, *27*(10), 103308. https://doi.org/10.1016/J.DRUDIS.2022.06.008

Graciani, G., & Yoon, T. Y. (2023). The New Kid on the Block: Mass Photometry. *Molecules and Cells*, *46*(3), 187–189. https://doi.org/10.14348/molcells.2023.0017

Hae, J. K., Coulibaly, F., Clow, F., Proft, T., & Baker, E. N. (2007). Stabilizing isopeptide bonds revealed in gram-positive bacterial pilus structure. *Science (New York, N.Y.)*, *318*(5856), 1625–1628. https://doi.org/10.1126/SCIENCE.1145806

Hagan, R. M., Björnsson, R., McMahon, S. A., Schomburg, B., Braithwaite, V., Bühl, M., Naismith, J. H., & Schwarz-Linek, U. (2010). NMR Spectroscopic and Theoretical Analysis of a Spontaneously Formed Lys–Asp Isopeptide Bond. *Angewandte Chemie International Edition*, *49*(45), 8421–8425. https://doi.org/10.1002/ANIE.201004340

Hartzell, E. J., Terr, J., & Chen, W. (2021). Engineering a Blue Light Inducible SpyTag System (BLISS). *Journal of the American Chemical Society*, *143*(23), 8572–8577. https://doi.org/10.1021/jacs.1c03198

Hatlem, D., Trunk, T., Linke, D., & Leo, J. C. (2019). Catching a SPY: Using the SpyCatcher-SpyTag and related systems for labeling and localizing bacterial proteins. *International Journal of Molecular Sciences*, *20*(9). https://doi.org/10.3390/ijms20092129

Hendrickx, A. P. A., Budzik, J. M., Oh, S. Y., & Schneewind, O. (2011). Architects at the bacterial surface-sortases and the assembly of pili with isopeptide bonds. In *Nature Reviews Microbiology* (Vol. 9, Issue 3, pp. 166–176). https://doi.org/10.1038/nrmicro2520

Higuchi, R., Krummel, B., & Saiki, R. K. (1988a). A general method of in vitro preparation and specific rautageoesis of DNA fragments: study of protein and DNA interactions. In *Nucleic Acids Research* (Vol. 16). https://academic.oup.com/nar/article/16/15/7351/1073393

Higuchi, R., Krummel, B., & Saiki, R. K. (1988b). A general method of in vitro preparation and specific rautageoesis of DNA fragments: study of protein and DNA interactions. In *Nucleic Acids Research* (Vol. 16). https://academic.oup.com/nar/article/16/15/7351/1073393

Hilgarth, R. S., & Lanigan, T. M. (2020). Optimization of overlap extension PCR for efficient transgene construction. *MethodsX*, *7*. https://doi.org/10.1016/j.mex.2019.12.001

Houghten original Nature paper on positional fixing 1991 354084a0 (1). (Torrey Pines Institute for Molecular Studies, 3550 General Atomics Court,San Diego, California 92121, USA).

Houghten, R. A., Pinilla, C., Appel, J. R., Blondelle, S. E., Dooley, C. T., Eichler, J., Nefzi, A., & Ostresh, J. M. (1999). Mixture-based synthetic combinatorial libraries. In *Journal of Medicinal Chemistry* (Vol. 42, Issue 19, pp. 3743–3778). https://doi.org/10.1021/jm990174v

Huang, X., Xu, L., Bi, C., Zhao, L., Zhang, L., Chen, X., Qi, S., & Lin, S. (2022). Designing Overlap Extension PCR Primers for Protein Mutagenesis: A Programmatic Approach. *Methods in Molecular Biology*, *2461*, 1–7. https://doi.org/10.1007/978-1-0716-2152-3_1

Hughes, M. D., Nagel, D. A., Santos, A. F., Sutherland, A. J., & Hine, A. V. (2003a). Removing the redundancy from randomised gene libraries. *Journal of Molecular Biology*, *331*(5), 973–979. https://doi.org/10.1016/S0022-2836(03)00833-7

Hughes, M. D., Nagel, D. A., Santos, A. F., Sutherland, A. J., & Hine, A. V. (2003b). Removing the redundancy from randomised gene libraries. *Journal of Molecular Biology*, *331*(5), 973–979. https://doi.org/10.1016/S0022-2836(03)00833-7

Hughes, M. D., Zhang, Z. R., Sutherland, A. J., Santos, A. F., & Hine, A. V. (2005). Discovery of active proteins directly from combinatorial randomized protein libraries without display, purification or sequencing: Identification of novel zinc finger proteins. *Nucleic Acids Research*, *33*(3), 1–8. https://doi.org/10.1093/nar/gni031

Kang, H. J., & Baker, E. N. (2011a). Intramolecular isopeptide bonds: protein crosslinks built for stress? *Trends in Biochemical Sciences*, *36*(4), 229–237. https://doi.org/10.1016/J.TIBS.2010.09.007

Kang, H. J., & Baker, E. N. (2011b). Intramolecular isopeptide bonds: protein crosslinks built for stress? *Trends in Biochemical Sciences*, *36*(4), 229–237. https://doi.org/10.1016/J.TIBS.2010.09.007

Karimi Baba ahmadi, M., Mohammadi, S. A., Makvandi, M., Mamoueie, M., Rahmati, M., & Wood, D. (2020). Column-free purification and coating of SpyCatcher protein on ELISA wells generates universal solid support for capturing of SpyTag-fusion protein from the non-purified condition. *Protein Expression and Purification*, *174*. https://doi.org/10.1016/j.pep.2020.105650

Keeble, A. H., Banerjee, A., Ferla, M. P., Reddington, S. C., Anuar, I. N. A. K., & Howarth, M. (2017a). Evolving Accelerated Amidation by SpyTag/SpyCatcher to Analyze Membrane Dynamics. *Angewandte Chemie*, *129*(52), 16748–16752. https://doi.org/10.1002/ange.201707623

Keeble, A. H., & Howarth, M. (2020a). Power to the protein: Enhancing and combining activities using the Spy toolbox. In *Chemical Science* (Vol. 11, Issue 28, pp. 7281–7291). Royal Society of Chemistry. https://doi.org/10.1039/d0sc01878c

Keeble, A. H., & Howarth, M. (2020b). Power to the protein: Enhancing and combining activities using the Spy toolbox. In *Chemical Science* (Vol. 11, Issue 28, pp. 7281–7291). Royal Society of Chemistry. https://doi.org/10.1039/d0sc01878c

Keeble, A. H., Turkki, P., Stokes, S., Khairil Anuar, N. A., Rahikainen, R., Hytönen, V. P., & Howarth, M. (n.d.-a). *Approaching infinite affinity through engineering of peptide-protein interaction*. https://doi.org/10.1073/pnas.1909653116/-/DCSupplemental

Keeble, A. H., Turkki, P., Stokes, S., Khairil Anuar, N. A., Rahikainen, R., Hytönen, V. P., & Howarth, M. (n.d.-b). *Approaching infinite affinity through engineering of peptide-protein interaction*. https://doi.org/10.1073/pnas.1909653116/-/DCSupplemental

Khairil Anuar, I. N. A., Banerjee, A., Keeble, A. H., Carella, A., Nikov, G. I., & Howarth, M. (2019). Spy&Go purification of SpyTag-proteins using pseudo-SpyCatcher to access an oligomerization toolbox. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09678-w

Kille, S., Acevedo-Rocha, C. G., Parra, L. P., Zhang, Z. G., Opperman, D. J., Reetz, M. T., & Acevedo, J. P. (2013a). Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synthetic Biology*, *2*(2), 83–92. https://doi.org/10.1021/sb300037w

Kille, S., Acevedo-Rocha, C. G., Parra, L. P., Zhang, Z. G., Opperman, D. J., Reetz, M. T., & Acevedo, J. P. (2013b). Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synthetic Biology*, *2*(2), 83–92. https://doi.org/10.1021/sb300037w

Ladner-Keay, C. L., Turner, R. J., & Edwards, R. A. (2018). Fluorescent Protein Visualization Immediately After Gel Electrophoresis Using an In-Gel Trichloroethanol Photoreaction with Tryptophan. *Methods in Molecular Biology*, *1853*, 179–190. https://doi.org/10.1007/978-1-4939-8745-0_22

Li, L., Fierer, J. O., Rapoport, T. A., & Howarth, M. (2014). Structural analysis and optimization of the covalent association between SpyCatcher and a peptide tag. *Journal of Molecular Biology*, *426*(2), 309–317. https://doi.org/10.1016/j.jmb.2013.10.021

Li, R., Kang, G., Hu, M., & Huang, H. (2019). Ribosome Display: A Potent Display Technology used for Selecting and Evolving Specific Binders with Desired Properties. In *Molecular Biotechnology* (Vol. 61, Issue 1, pp. 60–71). Humana Press Inc. https://doi.org/10.1007/s12033-018-0133-0

Li, Y., Struwe, W. B., & Kukura, P. (2020). Single molecule mass photometry of nucleic acids. *Nucleic Acids Research*, *48*(17), E97. https://doi.org/10.1093/nar/gkaa632

Lin, Z., Lin, Q., Li, J., Pistolozzi, M., Zhao, L., Yang, X., & Ye, Y. (2020a). Spy chemistry-enabled protein directional immobilization and protein purification. *Biotechnology and Bioengineering*, *117*(10), 2923–2932. https://doi.org/10.1002/bit.27460

Lowe, G. (1995.). *Combinatorial Chemistry*. Chemical Society reviews. https://doi.org/10.1039/CS9952400309

Marty, M. T. (2021). Illuminating Individual Membrane Protein Complexes with Mass Photometry. *Chem*, *7*(1), 16–17. https://doi.org/10.1016/j.chempr.2020.12.012

Monti, A., Vitagliano, L., Caporale, A., Ruvo, M., & Doti, N. (2023). Targeting Protein–Protein Interfaces with Peptides: The Contribution of Chemical Combinatorial Peptide Library Approaches. In *International Journal of Molecular Sciences* (Vol. 24, Issue 9). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/ijms24097842

Ostresh, J. M., Husar, G. M., Blondelle, S. E., D6rner, B., Webert, P. A., & Houghten, R. A. (1994). "Libraries from libraries": Chemical transformation of combinatorial libraries to extend the range and

repertoire of chemical diversity (foI atnperakylated combinatorial Nbraries/pethylated peptdes/peptide brary/a ba). In *Proc. Nati. Acad. Sci. USA* (Vol. 91). https://www.pnas.org

Paul, S. S., Lyons, A., Kirchner, R., & Woodside, M. T. (2022). Quantifying Oligomer Populations in Real Time during Protein Aggregation Using Single-Molecule Mass Photometry. *ACS Nano*, *16*(10), 16462–16470. https://doi.org/10.1021/acsnano.2c05739

Pessino, V., Citron, Y. R., Feng, S., & Huang, B. (2017). Covalent Protein Labeling by SpyTag–SpyCatcher in Fixed Cells for Super-Resolution Microscopy. *ChemBioChem*, *18*(15), 1492–1495. https://doi.org/10.1002/cbic.201700177

Potyrailo, R., Rajan, K., Stoewe, K., Takeuchi, I., Chisholm, B., & Lam, H. (2011). Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS Combinatorial Science*, *13*(6), 579–633. https://doi.org/10.1021/CO200007W

Reddington, S. C., & Howarth, M. (2015a). Secrets of a covalent interaction for biomaterials and biotechnology: SpyTag and SpyCatcher. In *Current Opinion in Chemical Biology* (Vol. 29, pp. 94–99). Elsevier Ltd. https://doi.org/10.1016/j.cbpa.2015.10.002

Reddington, S. C., & Howarth, M. (2015b). Secrets of a covalent interaction for biomaterials and biotechnology: SpyTag and SpyCatcher. In *Current Opinion in Chemical Biology* (Vol. 29, pp. 94–99). Elsevier Ltd. https://doi.org/10.1016/j.cbpa.2015.10.002

Schoene, C., Fierer, J. O., Bennett, S. P., & Howarth, M. (2014a). SpyTag/Spycatcher cyclization confers resilience to boiling on a mesophilic enzyme. *Angewandte Chemie - International Edition*, *53*(24), 6101–6104. https://doi.org/10.1002/anie.201402519

Soltermann, F., Foley, E. D. B., Pagnoni, V., Galpin, M., Benesch, J. L. P., Kukura, P., & Struwe, W. B. (2020a). Quantifying Protein–Protein Interactions by Molecular Counting with Mass Photometry. *Angewandte Chemie - International Edition*, *59*(27), 10774–10779. https://doi.org/10.1002/anie.202001578

Song, H., Wang, Y., Dong, W., Chen, Q., Sun, H., Peng, H., Li, R., Chang, Y., & Luo, H. (2022). Effect of SpyTag/SpyCatcher cyclization on stability and refolding of green fluorescent protein. *Biotechnology Letters*, *44*(4), 613–621. https://doi.org/10.1007/s10529-022-03246-x

Sridharan, U., & Ponnuraj, K. (2016). Isopeptide bond in collagen- and fibrinogen-binding MSCRAMMs. *Biophysical Reviews*, *8*(1), 75. https://doi.org/10.1007/S12551-015-0191-5

Suay-García, B., Bueso-Bordils, J. I., Falcó, A., Antón-Fos, G. M., & Alemán-López, P. A. (2022). Virtual Combinatorial Chemistry and Pharmacological Screening: A Short Guide to Drug Design. *International Journal of Molecular Sciences 2022, Vol. 23, Page 1620*, *23*(3), 1620. https://doi.org/10.3390/IJMS23031620

Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M., & Jiang, R. (2012a). Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques*, *52*(3), 149–158. https://doi.org/10.2144/000113820

Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M., & Jiang, R. (2012b). Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques*, *52*(3), 149–158. https://doi.org/10.2144/000113820

Terrett, Nick. (1998). *Combinatorial chemistry*. 240. Oxford University Press,1998. https://books.google.com/books/about/Combinatorial_Chemistry.html?id=CVShtAEACAAJ

Tian, J., Jia, R., Wenge, D., Sun, H., Wang, Y., Chang, Y., & Luo, H. (2021). One-step purification and immobilization of recombinant proteins using SpyTag/SpyCatcher chemistry. *Biotechnology Letters*, *43*(5), 1075–1087. https://doi.org/10.1007/s10529-021-03098-x

Williams, D. M., Kaufman, G., Izadi, H., Gahm, A. E., Prophet, S. M., Vanderlick, K. T., Osuji, C. O., & Regan, L. (2018). Facile Protein Immobilization Using Engineered Surface-Active Biofilm Proteins. *ACS Applied Nano Materials*, *1*(6), 2483–2488. https://doi.org/10.1021/ACSANM.8B00520/SUPPL_FILE/AN8B00520_SI_001.PDF

Wu, D., & Piszczek, G. (2021). *Rapid Determination of Antibody-Antigen Affinity by Mass Photometry. 2021 Feb 8:(168):10.3791/61784.doi: 10.3791/61784.*

Wu, D., & Piszczek, G. (2020a). Measuring the affinity of protein-protein interactions on a single-molecule level by mass photometry. *Analytical Biochemistry*, *592*. https://doi.org/10.1016/j.ab.2020.113575

Wu, D., & Piszczek, G. (2021). Standard protocol for mass photometry experiments. *European Biophysics Journal*, *50*(3–4), 403–409. https://doi.org/10.1007/s00249-021-01513-9

Zakeri, B., Fierer, J. O., Celik, E., Chittock, E. C., Schwarz-Linek, U., Moy, V. T., & Howarth, M. (2012a). *Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin*. https://doi.org/10.1073/pnas.1115485109/-/DCSupplemental

Zakeri, B., & Howarth, M. (2010). Spontaneous intermolecular amide bond formation between side chains for irreversible peptide targeting. *Journal of the American Chemical Society*, *132*(13), 4526–4527. https://doi.org/10.1021/JA910795A

Annex 1 Oligonucletide sequences

Name	Oligoes(5' to 3')
SC GG Forward	TTTTTTGGTCTCAGAGCAAGGTCAGGTTACTGTAAATGGCAAAGCAACTAAAGG
SC GG Reverse	TTTTTTGGTCTCATATCTTCTTCAATTGTCATATCACCGGACTGACCTTGCTC
SC Forward	GCCATGGTTGATACCTTAT
SC Reverse	AATATGAGCGTCACCTTTAG
mNeongreen Forward	CTTGTACAGCTCGTCCATG
mNeongreen Reverse	ATGGTGAGCAAGGGCGAGG

Table 9.1. Oligos utilised to make SpyCatcher-mNeongreen plasmid

Name	Oligoes(5' to 3')
mCherry Forward	ggcggttcaGTGAGCAAGGGCGAGGAG
mCherry Reverse	gtgatggtgTTTGTACAGCTCGTCCATGC
mNeon knockout Forward	gctgtacaaaCACCATCACCATCACCATTAATAACTC
mNeon knockout Reverse	ccttgctcacTGAACCGCCACCACCGCT

Table 9.2. Oligos utilized to make Native SpyTag-mCherry plasmid

Name	SpyCatcher Oligoes (5’ to 3’)
27 I Forward	GATAGTGCTACCCATATTAAATTCTCAAAACG
27 V Forward	GATAGTGCTACCCATGTGAAATTCTCAAAACG
27 L Forward	GATAGTGCTACCCATCTGAAATTCTCAAAACG
27 F Forward	GATAGTGCTACCCATTTTAAATTCTCAAAACG
27 M Forward	GATAGTGCTACCCATATGAAATTCTCAAAACG
27 Y Forward	GATAGTGCTACCCATTATAAATTCTCAAAACG
44 I Reverse	CACGCAACTCAATAGTTGCACCAGC
44 V Reverse	CACGCAACTCCACAGTTGCACCAGC
44 L Reverse	CACGCAACTCCAGAGTTGCACCAGC
44 F Reverse	CACGCAACTCAAAAGTTGCACCAGC
44 M Reverse	CACGCAACTCCATAGTTGCACCAGC
44 Y Reverse	CACGCAACTCATAAGTTGCACCAGC
44 I Forward	GCTGGTGCAACTATTGAGTTGCGTGAT
44 V Forward	GCTGGTGCAACTGTGGAGTTGCGTGAT
44 L Forward	GCTGGTGCAACTCTGGAGTTGCGTGAT
44 F Forward	GCTGGTGCAACTTTTGAGTTGCGTGAT
44 M Forward	GCTGGTGCAACTATGGAGTTGCGTGAT
44 Y Forward	GCTGGTGCAACTTATGAGTTGCGTGAT
90 I Reverse	GCTCATTAACTGTAAAGGTAATAGCAGTTGCTACC
90 V Reverse	GCTCATTAACTGTAAAGGTAACAGCAGTTGCTACC
90 L Reverse	GCTCATTAACTGTAAAGGTCAGAGCAGTTGCTACC
90 F Reverse	GCTCATTAACTGTAAAGGTAAAAGCAGTTGCTACC
90 M Reverse	GCTCATTAACTGTAAAGGTCATAGCAGTTGCTACC
90 Y Reverse	GCTCATTAACTGTAAAGGTATAAGCAGTTGCTACC
27 Forward	TTTTTTGGTCTCAGATAGTGCTACCCAT
90 Reverse	TTTTTTGGTCTCAGCTCATTAACTGTAAAGGT

Table 9.3. Oligos utilised to construct 18 distinct SpyCatcher libraries

Annex 2:

SpyCatcher-mNeongreen plasmid sequence (6335bp)

CATCTAGTATTTCTCCTCTTTAATATGGATCCTAGAGGGGAATTGTTATCCGCTCACAATTCCCCTATAGTGAGTC
GTATTAATTTCGCGGGATCGAGATCTCGATCCTCTACGCCGGACGCATCGTGGCCGGCATCACCGGCGCCACA
GGTGCGGTTGCTGGCGCCTATATCGCCGACATCACCGATGGGGAAGATCGGGCTCGCCACTTCGGGCTCATGA
GCGCTTGTTTCGGCGTGGGTATGGTGGCAGGCCCCGTGGCCGGGGGACTGTTGGGCGCCATCTCCTTGCATG
CACCATTCCTTGCGGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTTCCTAATGCAGGAGTCGCAT
AAGGGAGAGCGTCGAGATCCCGGACACCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCG
GAAGAGAGTCAATTCAGGGTGGTGAATGTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTC
TTATCAGACCGTTTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCG
GCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTG
GCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCA
ACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAA
TCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAG
CTGCCTGCACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCA
TGAAGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGC
CCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCG
ATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCAT
CGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGG
CTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATCCCGCCGTT
AACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCC
AGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCA
AACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGG
CAGTGAGCGCAACGCAATTAATGTAAGTTAGCTCACTCATTAGGCACCGGGATCTCGACCGATGCCCTTGAGAG
CCTTCAACCCAGTCAGCTCCTTCCGGTGGGCGCGGGGCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTT
ATCATGCAACTCGTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGCG
CGACGATGATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTCGTCACTGGTCCC
GCCACCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATGGCGGCCCCACGGGTGCGCATGATCGTG
CTCCTGTCGTTGAGGACCCGGCTAGGCTGGCGGGGTTGCCTTACTGGTTAGCAGAATGAATCACCGATACGCG
AGCGAACGTGAAGCGACTGCTGCTGCAAAACGTCTGCGACCTGAGCAACAACATGAATGGTCTTCGGTTTCCGT
GTTTCGTAAAGTCTGGAAACGCGGAAGTCAGCGCCCTGCACCATTATGTTCCGGATCTGCATCGCAGGATGCTG
CTGGCTACCCTGTGGAACACCTACATCTGTATTAACGAAGCGCTGGCATTGACCCTGAGTGATTTTTCTCTGGTC
CCGCCGCATCCATACCGCCAGTTGTTTACCCTCACAACGTTCCAGTAACCGGGCATGTTCATCATCAGTAACCC
GTATCGTGAGCATCCTCTCTCGTTTCATCGGTATCATTACCCCCATGAACAGAAATCCCCCTTACACGGAGGCAT
CAGTGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAG
AAACTCAACGAGCTGGACGCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGATGAGCTTTA
CCGCAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACA
GCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGG
GGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATT
GTACTGAGAGTGCACCATATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGC
TCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAA
GGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAG
GCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAA
TCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCC
CTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGC
GCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACG
AACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGAC
TTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTT
GAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCT
TCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAG
CAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTG
GAACGAAAACTCACGTTAAGGGATTTTGGTCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAGGG
GTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCTAGGCCGCGATTAAATTCCAACATGGATGCTGATTTAT
ATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGAT
GCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACTAAA

CTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCATGGTTACTCAC
CACTGCGATCCCCGGGAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGC
GCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCG
TCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCT
GGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTG
ATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGC
AGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTT
TCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAAGAAT
TAATTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAA
AAGTGCCACCTGAAATTGTAAACGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCAGCTCATTTTTTA
ACCAATAGGCCGAAATCGGCAAAATCCCTTATAAATCAAAAGAATAGACCGAGATAGGGTTGAGTGTTGTTCCAG
TTTGGAACAAGAGTCCACTATTAAAGAACGTGGACTCCAACGTCAAAGGGCGAAAAACCGTCTATCAGGGCGAT
GGCCCACTACGTGAACCATCACCCTAATCAAGTTTTTTGGGGTCGAGGTGCCGTAAAGCACTAAATCGGAACCC
TAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAAGC
GAAAGGAGCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAACCACCACACCCGCCGCGCT
TAATGCGCCGCTACAGGGCGCGTCCCATTCGCCAATCCGGATATAGTTCCTCCTTTCAGCAAAAAACCCCTCAA
GACCCGTTTAGAGGCCCCAAGGGGTTATGCTAGTTATTGCTCAGCGGTGGCAGCAGCCAACTCAGCTTCCTTTC
GGGCTTTGTTAGCAGCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAGTTATTAATGGTGATGGTGATGGTGC
TTGTACAGCTCGTCCATGCCCATCACATCGGTAAAGGCCTTTTGCCACTCCTTGAAGTTGAGCTCGGTCTTGGAG
TGCTTGAGCTCCGTCTTACGGAACACGTACATCGGCTGGTTCTTCAGATAGTTAGCCGCCATTGGCTTGGCAAA
GGTGTAGGTGGTCCGCGCAGTGCTCCGGTAGCGCTTGCCATTTCCAGTGGTGTAACTCCACTTAAAGGTACTGA
TGATGGTTTTGTCGTTGGGGTAAGTCTTCTTCGACCTGCACCAGTCCGCAGCGGTCAGCGAGTTGGTCATCACA
GGACCGTCAGCAGGGAAACCAGTCCCCTTCACCTGGGCCTCTCCTTTGATGTGGCTTCCCTCGTAGGTGTAGCG
GTAGTTAACAGTAAGGGAGGCACCATCTTCAAACTGCATTGTGCGATGGACTTGGTAGCCGGAGCCATCTACCA
TGGCGGCCTGGAAAGGCGACATCCCGTCAGGGTAGGGCAGGTACTGATGGAAGCCATACCCGATATGAGGGA
CCAGAATCCAGGGGGAGAACTGGAGGTCACCCTTGGTGGACTTCAGGTTTAACTCCTCATAACCATCATTTGGA
TTGCCGGTGCCCTGACCCACCATGTCAAAGTCCACACCGTTGATGGAGCCAAAGATGTGTAACTCATGTGTCGC
TGGGAGAGAGGCCATGTTATCCTCCTCGCCCTTGCTCACCATTGAACCGCCACCACCGCTACCACCACCACCAA
TATGAGCGTCACCTTTAGTTGCTTTGCCATTTACAGTAACCTGACCTTGCTCATTAACTGTAAAGGTAATAGCAGT
TGCTACCTCATAACCGTCTGGTGCTGCGGTTTCGACAAATGTATATTTTCCTGGATACAGGTAGAAATCTTTCACT
TGTCCATCTGAAATCCATGTACTAATAGTTTTACCAGATGAATCACGCAACTCCATAGTTGCACCAGCTAACTCTT
TGCCGTCCTCATCACGTTTTGAGAATTTAATATGGGTAGCACTATCTTCTTCAATTGTCATATCACCGGACTGACC
TTGCTCACTTGATAAACCTGATAAGGTATCAACCATGGC

Annex 3
SpyTag-mCherry plasmid sequence (6026 bp)

TGAACCGCCACCACCGCTACCACCACCACCCTTCGTCGGCTTGTAGGCGTCTACCATTACTATGTGTGCCATCTA
GTATTTCTCCTCTTTAATATGGATCCTAGAGGGGAATTGTTATCCGCTCACAATTCCCCTATAGTGAGTCTATTAATTT
CGCGGGATCGAGATCTCGATCCTCTACGCCGGACGCATCGTGGCCGGCATCACCGGCGCCACAGGTGCGGTTG
CTGGCGCCTATATCGCCGACATCACCGATGGGGAAGATCGGGCTCGCCACTTCGGGCTCATGAGCGCTTGTTTC
GGCGTGGGTATGGTGGCAGGCCCCGTGGCCGGGGGACTGTTGGGCGCCATCTCCTTGCATGCACCATTCCTTG
CGGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTTCCTAATGCAGGAGTCGCATAAGGGAGAGCGT
CGAGATCCCGGACACCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAAT
TCAGGGTGGTGAATGTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTC
CCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCT
GAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCA
GTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGT
GGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGC
GTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTT
CCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAAGACGGTACGCGAC
TGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCG
GCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGG
CGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCT
GGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGAT
ATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATT
TTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCA
GCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGT

TGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAA
TGTAAGTTAGCTCACTCATTAGGCACCGGGATCTCGACCGATGCCCTTGAGAGCCTTCAACCCAGTCAGCTCCTT
CCGGTGGGCGCGGGGCATGACTATCGTCGCCGCACTTATGACGTCTTCTTTATCATGCAACTCGTAGGACAGGTG
CCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGCGCGACGATGATCGGCCTGTCGCTTG
CGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTCGTCACTGGTCCCGCCACCAAACGTTTCGGCGAGAAG
CAGGCCATTATCGCCGGCATGGCGGCCCCACGGGTGCGCATGATCGTGCTCCTGTCGTTGAGGACCCGGCTAG
GCTGGCGGGGTTGCCTTACTGGTTAGCAGAATGAATCACCGATACGCGAGCGAACGTGAAGCGACTGCTGCTGC
AAAACGTCTGCGACCTGAGCAACAACATGAATGGTCTTCGGTTTCCGTGTTTCGTAAAGTCTGGAAACGCGGAAG
TCAGCGCCCTGCACCATTATGTTCCGGATCTGCATCGCAGGATGCTGCTGGCTACCCTGTGGAACACCTACATCT
GTATTAACGAAGCGCTGGCATTGACCCTGAGTGATTTTTCTCTGGTCCCGCCGCATCCATACCGCCAGTTGTTTAC
CCTCACAACGTTCCAGTAACCGGGCATGTTCATCATCAGTAACCCGTATCGTGAGCATCCTCTCTCGTTTCATCGG
TATCATTACCCCCATGAACAGAAATCCCCCTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAAC
ATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCTGGACGCGGATGAACAGGC
AGACATCTGTGAATCGCTTCCGACCACGCTGATGAGCTTTACCGCAGCTGCCTCGCGCGTTTCGGTGATGACGG
TGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAA
GCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGC
GGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATATGCGGTGTGAAATA
CCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTC
GGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATA
ACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTT
TTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAG
GACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCG
GATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGG
TGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGT
AACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGC
AGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTA
TTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACC
ACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCT
TTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAACAATAAAA
CTGTCTGCTTACATAAACAGTAATACAAGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCTAGGCCGCG
ATTAAATTCCAACATGGATGCTGATTTATATGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGACAAT
CTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTT
ACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGTACTC
CTGATGATGCATGGTTACTCACCACTGCGATCCCCGGGAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTC
AGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTTTA
ACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGA
TGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCA
GTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGG
ACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTA
CAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAG
TTTTTCTAAGAATTAATTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCAC
ATTTCCCCGAAAAGTGCCACCTGAAATTGTAACGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCAGCT
CATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCTTATAAATCAAAAGAATAGACCGAGATAGGGTGAGTGTTG
TTCCAGTTTGGAACAAGAGTCCACTATTAAAGAACGTGGACTCCAACGTCAAAGGGCGAAAAACCGTCTATCAGG
GCGATGGCCCACTACGTGAACCATCACCCTAATCAAGTTTTTTGGGGTCGAGGTGCCGTAAAGCACTAAATCGGA
ACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAA
AGCGAAAGGAGCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAACCACCACACCCGCCGC
GCTTAATGCGCCGCTACAGGGCGCGTCCCATTCGCCAATCCGGATATAGTTCCTCCTTTCAGCAAAAAACCCCTC
AAGACCCGTTTAGAGGCCCCAAGGGGTTATGCTAGTTATTGCTCAGCGGTGGCAGCAGCCAACTCAGCTTCCTTT
CGGGCTTTGTTAGCAGCCGGATCTCAGTGGTGGTGGTGGTGGTGCTCGAGTTATTAATGGTGATGGTGATGGTGT
TTGTACAGCTCGTCCATGCCGCCGGTGGAGTGGCGGCCCTCGGCGCGTTCGTACTGTTCCACGATGGTGTAGTC
CTCGTTGTGGGAGGTGATGTCCAACTTGATGTTGACGTTGTAGGCGCCGGGCAGCTGCACGGGCTTCTTGGCCT
TGTAGGTGGTCTTGACCTCAGCGTCGTAGTGGCCGCCGTCCTTCAGCTTCAGCCTCTGCTTGATCTCGCCCTTCA
GGGCGCCGTCCTCGGGGTACATCCGCTCGGAGGAGGCCTCCCAGCCCATGGTCTTCTTCTGCATTACGGGGCC
GTCGGAGGGGAAGTTGGTGCCGCGCAGCTTCACCTTGTAGATGAACTCGCCGTCCTGCAGGGAGGAGTCCTGG
GTCACGGTCACCACGCCGCCGTCCTCGAAGTTCATCACGCGCTCCCACTTGAAGCCCTCGGGAAGGACAGCTT
CAAGTAGTCGGGGATGTCGGCGGGGTGCTTCACGTAGGCCTTGGAGCCGTACATGAACTGAGGGGACAGGATG
TCCCAGGCGAAGGGCAGGGGGCCACCCTTGGTCACCTTCAGCTTGGCGGTCTGGGTGCCCTCGTAGGGGCGG
CCCTCGCCCTCGCCCTCGATCTCGAACTCGTGGCCGTTCACGGAGCCCTCCATGTGCACCTTGAAGCGCATGAA
CTCCTTGATGATGGCCATGTTATCCTCCTCGCCCTTGCTCAC